"*Re-defining exacerbations of chronic obstructive pulmonary disease and developing tools to characterise heterogeneity and treatment response*"

Andreas Joseph Halner

St John's College, University of Oxford

Thesis submitted for the degree of Doctor of Philosophy in Clinical Medicine, University of Oxford, Trinity Term, 2022

Primary Supervisor: Professor Mona Bafadhel

Secondary Supervisors: Professor Philip Maini, Professor Dan Nicolau

Table of contents

Abstract

Chronic obstructive pulmonary disease (COPD) is the third leading cause of death worldwide. Progressive in nature, COPD is punctuated by acute periods of worsening termed ‘exacerbations’, a cause of morbidity and mortality. Exacerbations are heterogenous with respect to symptoms and inflammatory biology.

In this thesis, mathematical methods were researched for their appropriateness to characterise COPD heterogeneity and predict treatment response. Algorithms were developed to predict treatment outcome for a cohort of hospitalised patients with airways disease exacerbations. In a second patient cohort, cluster analysis was performed on longitudinal measurements at 'stable state’ and ‘exacerbation’ in order to define objective ‘states’ of COPD based on symptoms and inflammatory biology. Objective states included a low overall symptom burden state and crisis states with elevated symptoms with or without particular inflammation. Objective states were compared to conventional ‘stable state’ and ‘exacerbation’ definitions and found to better correspond to distinct patterns of symptoms and inflammation. Algorithms were trained and tested to identify to which objective state any given measurement of symptoms and inflammation from a patient belongs at a given point in time. Different treatment response trajectories or ‘trajectotypes’ of symptoms and inflammation were characterised for crisis states between the time of treatment and follow-up measurements. Trajectotypes included symptom resolution with or without different patterns of inflammatory resolution, as well as a trajectotype of no change between treatment and follow-up. Individualised symptom shape profile trajectories post-treatment were found from daily time series of self-reported symptom scores.

This thesis provides a new definition of COPD states as an alternative to ‘stable state’ and ‘exacerbation’ definitions. Together with the characterisation of distinct treatment

response trajectories based on symptom and inflammatory resolution, the findings of this thesis set the stage for future clinical studies evaluating enhanced personalisation of treatment allocation and monitoring for patients with COPD.

List of publications arising from this thesis

Publications

Halner et al. Usefulness of Biological Clustering Patterns in Chronic Obstructive Pulmonary Disease. *Barcelona Respir. Netw. Rev 2019.* 5, 201–14

Halner A, Beer S, Pullinger R, Bafadhel M, Russell REK (2021) Predicting treatment outcomes following an exacerbation of airways disease. PLoS ONE 16(8): e0254425. https://doi.org/10.1371/journal.pone.0254425

Oral Presentations

Halner, A., Brightling, C., Bafadhel, M. (2019) *Chronic obstructive pulmonary disease exacerbations – characterising the relationship between symptom severity and airway inflammation* British Thoracic Society

Abstracts

Halner, A., Beer, S., Bafadhel, M. and Russell, R. (2019) Predicting treatment outcomes following an exacerbation of airways disease European Respiratory Journal 2019 54: PA4291; DOI: 10.1183/13993003.congress-2019.PA4291

Halner, A., Beer, S., Bafadhel, M. and Russell, R. (2019) Predicting physician decision to prescribe systemic corticosteroids following an exacerbation of airways

disease European Respiratory Journal 2019 54: PA2726; DOI: 10.1183/13993003.congress-2019.PA2726

Halner, A., Beer, S., Bafadhel, M. and Russell, R. (2019) Identifying tools to investigate outcomes following an exacerbation of COPD European Respiratory Journal 2019 54: PA4300; DOI: 10.1183/13993003.congress-2019.PA4300

Halner, A., Nicolau, D., Bafadhel, M.(2020) Characterising symptom ‘trajectotypes’ in patients with COPD. European Respiratory Journal 2020 56: 2670; DOI: 10.1183/13993003.congress-2020.2670

Statement of work performed

In this thesis, analysis was performed by me on data from pre-existing patient cohorts and studies (see references in corresponding thesis chapters). I was responsible for researching mathematical methodologies and assessing their appropriateness in facilitating my thesis objectives. I performed the literature reviews of COPD studies discussed in this thesis. As part of skill development during my thesis, I developed machine learning expertise and learned how to code in ‘R’ and ‘Python’. I developed the code used for analysis in this thesis.

Acknowledgements

I would like to sincerely thank my primary supervisor Prof Mona Bafadhel for her guidance, enthusiasm and support throughout my thesis. I could not ask for a more positive DPhil supervisory experience! I would like to also thank my secondary supervisors Prof Philip Maini and Prof Dan Nicolau for their mentorship during my mathematical training. I remain mindful and fully appreciative of the particular contributions made by Prof Sandra Campbell, Prof Jaideep Pandit, Prof Zoltan Molnar, Prof Jason Schnell and Prof Philip Maini whilst I was a pre-clinical medical student.

I am grateful to all patients who kindly agreed to participate in the studies which generated the datasets I used for my thesis analyses. A special expression of my gratitude should also be directed to the nurses, scientists and clinicians involved in data collection and study oversight, including Sally Beer, Dr Richard Pullinger, Dr Richard Russell and Prof Christopher Brightling.

I am no less indebted to my father, mother and sisters for the meaningful, quality time they have shared with me. My academic development from childhood to this point would not have been possible without the support of my family, especially my father. Thank you mum, Mrs Vera Halner. Thank you dad, Dr Alan Joseph Halner. Eternally B”H!

List of abbreviations

AUC	Area under the receiver operating characteristic curve
BMI	Body Mass Index
CART	Classification and Regression Tree
CAT	COPD Assessment Tool score
COPD	Chronic obstructive pulmonary disease
CRP	C-reactive protein
CRQ	Chronic Respiratory Disease Questionnaire
CT	Computed Tomography
D0	Day of initial presentation to emergency department for exacerbating patients in Chapter 3 analysis
D1	One day after initial presentation to emergency department for exacerbating patients in Chapter 3 analysis
D5	Five days after initial presentation to emergency department for exacerbating patients in Chapter 3 analysis
D30	Thirty days after initial presentation to emergency department for exacerbating patients in Chapter 3 analysis
D90	Ninety days after initial presentation to emergency department for exacerbating patients in Chapter 3 analysis
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DO	Degree of overlap
ED	Emergency department
eGFR	Estimated glomerular filtration rate
FAMD	Factor analysis of mixed variables
FeNO	Fractional exhaled nitric oxide
FEV_1	Forced expiratory volume in 1 second
FVC	Forced vital capacity
GINA	Global Initiative for Asthma
GOLD	Global Initiative for Chronic Obstructive Lung Disease
HAD	Hospital Anxiety and Depression scale
HI	Haemophilus influenzae
ICS	Inhaled corticosteroids
K_{CO}	Carbon monoxide diffusing capacity of the lung/alveolar volume
KNN	K-nearest-neighbour
LABA	Long-acting beta agonist
LACH	Long-acting anticholinergic
LAMA	Long-acting muscarinic antagonist
LASSO	Least absolute shrinkage and selection operator
MCA	Multiple correspondence analysis

MC	Moxarella catarrhalis
NHS	National Health Service
NICE	National Institute for Health and Care Excellence
PAFI	Arterial partial pressure of oxygen /inspired fraction of oxygen ratio
P_aO2	Arterial partial pressure of oxygen
PCA	Principal component analysis
pCO_2	Partial pressure of carbon dioxide
PPI	Proton pump inhibitor
ROC	Receiver operating characteristic
SABA	Short-acting beta agonist
SACH	Short-acting anticholinergic
SAMA	Short-acting muscarinic antagonist
SA	Staphylococcus aureus;
SP	Streptococcus pneumoniae;
SCS	Systemic corticosteroids
SES	Statistically equivalent signature (algorithm)
SGRQ	St George's Respiratory Questionnaire
SVM	Support vector machine
VAS	Visual Analogue Scale
qPCR	Quantitative polymerase chain reaction

Word count

Abstract: 299 words.

Main body of thesis (from Chapter 1 to Chapter 6 inclusive but excluding tables, figures and associated legends): 36,839 words.

List of tables

List of figures

Chapter 1 – Introduction

1.1 Chronic Obstructive Pulmonary Disease

Chronic obstructive pulmonary disease (COPD) affects over 250 million people worldwide and is the 3rd leading cause of death globally[1–3]. COPD is characterised by chronic bronchitis, with elevated sputum production and thickening of the bronchiolar walls and emphysema, involving alveolar destruction. Loss of airways elastic recoil results along with increased airway resistance[4,5]. This leads to expiratory flow limitation and dyspnoea. COPD is a preventable condition, being caused by exposure to noxious particles or gases[6]. Airway inflammation due to tobacco smoking is the most common cause of COPD[4,7]. However, indoor air pollution due to the use of biomass fuel or occupational exposures to various dusts and chemical fumes are also risk factors for COPD development[6].

1.2 COPD clinical presentation

COPD is progressive and chronic in nature, as manifested in the long term deterioration of lung function of COPD patients, and is punctuated by acute periods of symptom worsening or 'exacerbations'[6]. Not all COPD features are present in all patients or at any one point in time in a given patient, a phenomenon termed COPD heterogeneity[8]. Recognition of COPD heterogeneity can be seen as early as 1955 when Dornhorst noted the 'pink puffer' and 'blue bloater' phenotypes[9]. Early classifications of COPD severity have been characterised by a focus on Forced Expiratory Volume in 1 second (FEV)[6,10]. While COPD diagnosis relies on FEV_1/forced vital capacity (FVC) < 0.7, the importance of other prognostic indicators has become

increasingly understood[6]. For example, exacerbation history is a better predictor of future exacerbation risk than is FEV_1 and dyspnoea is a better predictor of mortality[11,12]. Comorbidities are common in COPD and increase mortality risk[6,13,14]. However, an integrated understanding of COPD heterogeneity in terms of inflammation, symptomatology and clinical outcome is lacking. Even worse, drugs commonly used to treat COPD exacerbations such as systemic corticosteroids (SCS) are associated with considerable patient harm including osteoporosis[6,15]. Hence, there has been an acknowledgement by the research community that an urgent priority is the identification of different subgroups in the form of inflammatory, symptomatological and clinical characteristics[4,6]. Specifically, efforts have been made to identify COPD phenotypes, defined as: "*a single or combination of disease attributes that describe differences between individuals with COPD as they relate to clinically meaningful outcomes (symptoms, exacerbations, response to therapy, rate of disease progression, or death)*"[16]. A related concept, endotypes are defined as: "*subtype[s] of disease defined functionally and pathologically by a molecular mechanism or by treatment response*"[17].

Characterising COPD heterogeneity with the view to developing objective definitions of COPD exacerbations and developing clinical prediction tools for COPD exacerbation treatment outcomes necessitates the mathematical identification of patterns in data.

1.3 COPD data modelling

Appropriate mathematical techniques for the identification of patterns in data will now be discussed. Corresponding applications of such techniques in COPD research will then be evaluated, taking into account two criteria: 1) methodological soundness of the

application of the mathematical technique; 2) the clinical usefulness of the study, in terms of the potential to improve clinical care.

Cluster analysis – mathematical background

Cluster analysis is a form of unsupervised learning, in which data points are unlabelled and the objective is to identify natural structure in the data. The objective of clustering is to identify subgroups, or clusters, of data points which, in some way, are more similar to each other than to those data points in other clusters[18]. Cluster analysis is an ideal way to characterise heterogeneity since, for example, data points representing patients which are closely related to each other can be grouped into the same cluster and hence different clusters are representative of different, more homogenous subgroups within an overall heterogenous dataset. The choice of (dis)similarity measure or 'proximity measure' between data points can have significant implications for the clustering solution[19]. In recent decades, cluster analysis has received increasing interest in a number of fields, including clinical research, as more data have become available. Cluster analysis is especially useful in the context of high dimensional data which cannot be easily visualised and hence cannot easily be segmented into distinguishable subgroups using visual inspection alone. Rather, provided that the sample size is sufficiently large, robust clusters can be found in an n-dimensional feature space by a variety of algorithmic techniques. Research suggests that sample size should be much greater than the number of features (dimensions) or else the identified clusters may be unreliable[20,21]. This can be better understood in the context of the 'curse of dimensionality'[19]. As the number of dimensions in the input feature space increases, an exponential increase in the number of data points is required in order to maintain the same density of data points in the input size, which is important for performing

optimal clustering. For high dimensional data, it may hence be appropriate to first perform dimensionality reduction and then subsequently perform clustering on a smaller number of reduced dimensions[22]. Principal components analysis (PCA)[23,24] reduces dimensionality of numerical variables by creating new variables or components, each of which is a linear combination of the original variables, such that all new principal components are orthogonal to each other. Selecting the top principal components which explain most of the variation in the data enables most information about the data to be conserved while using a smaller number of dimensions. An approach similar to PCA for categorical data is multiple correspondence analysis (MCA)[25]. In the case of mixed data, dimensionality reduction may be achieved via an approach known as factor analysis of mixed variables (FAMD)[26].

Broadly, cluster analysis algorithms can be divided into partitional algorithms, agglomerative hierarchical algorithms and density-based algorithms. Table 1.1 summarises the procedure, strengths and limitations of the K-means, Ward's method and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithms, commonly used algorithms corresponding to partitional, agglomerative hierarchical and density-based methods, respectively[27].

Table 1.1 | Summary of clustering approaches. Adapted from Halner et al (2019)[27].

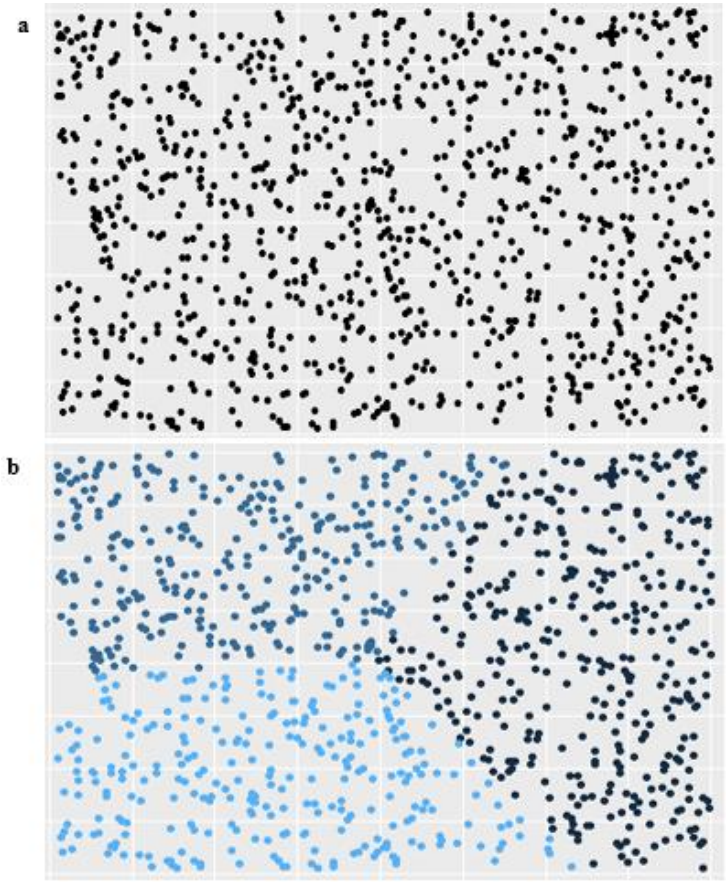
Clustering Approach	**K-means (partitional)[28-30]**	**Agglomerative hierarchical[28,31]**	**DBSCAN (Density-based method)[22,32]**
How it works	i) The user pre-specifies the number *K* of clusters which the algorithm is to find. ii) *K* initial centroids are chosen at random. iii) Based on a proximity measure, such as Euclidean distance, data points are assigned to the nearest centroid. iv) The position of data points belonging to each centroid's cluster is averaged in order to determine the new location of the centroids. v) The above steps are repeated until the centroids remain the same. The final grouping of data points around the centroids constitutes the final clusters.	i) Each data point represents an initial cluster. ii) A proximity measure, such as Euclidean distance, is used to determine which clusters are most similar to each other. iii) Clusters which are most similar (e.g. based on Ward's criterion) are merged with each other. iv) This step is repeated until only one cluster remains. A dendrogram enables geometrical interpretation of the tree of clusters.	i) Core data points are defined as those with at least *MinPts* points within a distance of *Eps* ii) Border points are defined as points which are not core but are located within the neighbourhood of core points. iii) Data points which are neither core nor border are labelled noise points. iv) Noise points are eliminated. v) Core points within a distance *Eps* of each other are connected with an edge. Each group of connected core points constitutes a separate cluster. Border points are assigned to the cluster of their corresponding core points.

Strengths	Computationally efficient.	Number of clusters does not need to be pre-specified by the user since all cluster-subcluster relationships are evident from dendrogram representation.	Clusters of arbitrary sizes and shapes can be identified, with cluster number automatically determined by the algorithm. Furthermore, algorithm is relatively insensitive to noise.
Limitations	i) Number of clusters must be pre-specified by the user. ii) Highly sensitive to position of initial centroids in step 1 of algorithm. iii) Highly sensitive to outliers. iv) Performs poorly if the true clusters are non-spherical, or if clusters vary considerably in size or density.	i) Although the number of clusters need not be pre-specified by the user in order to run the algorithm, a method is required for deciding on the most meaningful clusters to extract from the dendrogram. ii) Computationally expensive. iii) Cluster merging decisions in the early steps of the algorithm to maximise local optimization may lead to reduced optimization in downstream merging decisions, hence reducing global optimization.	i) Difficult to identify clusters of different densities. ii) Density is harder to define in high dimensions, potentially leading to poorer performance of the algorithm. iii) The pairwise proximity measure calculations needed for the density-based clustering procedure becomes computationally expensive in high dimensional situations. iv) Number of clusters identified is highly sensitive to the parameters *MinPts* and *Eps*.

Possible solutions to limitations	A number of cluster evaluation methods exist for deciding how many clusters from 2 to n, where n is the number of data points, are suitable e.g. pseudo F-statistic measures the ratio of between cluster variance to within-cluster variance (Calinski and Harabasz stopping rule). Bisecting K-means algorithm is less susceptible to initial centroid positions. Outliers can be removed before clustering or identified in a post-cluster processing step.	A number of cluster evaluation methods exist for deciding how many clusters from 2 to n, where n is the number of data points, are suitable. For example, pseudo-F statistic measures the ratio of between cluster variance to within-cluster variance (Calinski and Harabasz stopping rule).	Extensions to DBSCAN, such as the shared nearest neighbour (SNN) procedure, enable clusters of differing densities to be found.

Different clustering algorithms will find different clusters depending on the nature of the dataset at hand. There is no one clustering algorithm which one knows *a priori* will be superior in performance to all other clustering algorithms under all circumstances. This follows from the 'no free lunch' theorem of Wolpert and Macready for unsupervised optimisation problems[33,34]. Hence, for a given application, different clustering algorithms should be empirically compared with each other in terms of their performance. This includes visual inspection of the cluster solutions if the number of dimensions is 3 or less. For methods such as the K-means algorithm where the user must pre-specify the number of clusters, it is essential to confirm that the clusters are, in fact, real rather than resulting as an artefact of the clustering method which will find clusters if the data points have a uniform distribution (see Figure 1.1). This is especially important given that techniques such as the pseudo F-statistic (see Table 1.1) commonly used to determine optimal cluster number in cluster analysis studies assume the presence of at least two clusters in data and hence cannot ascertain whether the optimal solution is in fact one cluster i.e. a continuous spectrum rather than distinct subgroups[27]. Only if there is, *a priori*, a reason to suspect that at least two subgroups exist (e.g. due to different types of inflammation or other biological or clinical characteristics) is it valid to ignore the m=1 cluster case. However, the majority of cluster analysis studies consider clinical and epidemiological characteristics without consideration of biological subgroups of disease and do not justify why at least two clusters are to be expected.

Figure 1.1 | Illustration of cluster identification in a randomly distributed dataset. a: Randomly distributed two-dimensional data. b: K-means clustering algorithm identifies apparent clusters (highlighted light blue, dark blue and black). Adapted from Halner et al. (2019)[27].



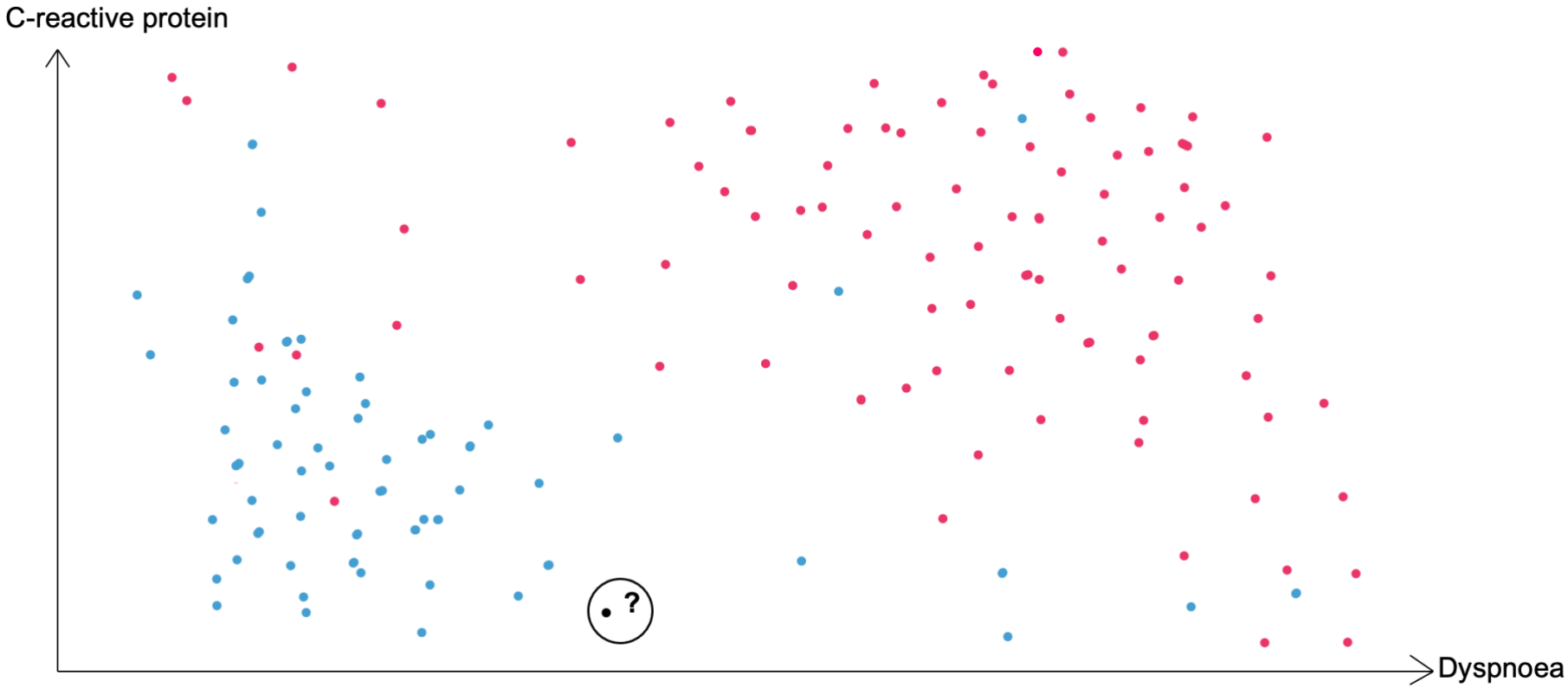
Spatial randomness tests, such as the Hopkins statistic, can be employed to confirm that cluster structure exists before one attempts to find clusters with a clustering algorithm[35,36]. Furthermore, for all cluster analysis methods, a variety of internal cluster validity indices have been developed to inform the user of the quality of clusters found by clustering algorithms (for example, refer to 'Possible solutions to limitations' of Table 1.1). However, different cluster validity indices measure different aspects of 'cluster quality' and there is no cluster validity index which is the 'best' under all circumstances. Hence, where data dimensionality is too high to enable visualisation of clusters found

by an algorithm, it is imperative that a variety of cluster validity indices suitable for the particular context in question are evaluated.

Classification to predict clinically relevant outcomes – mathematical background

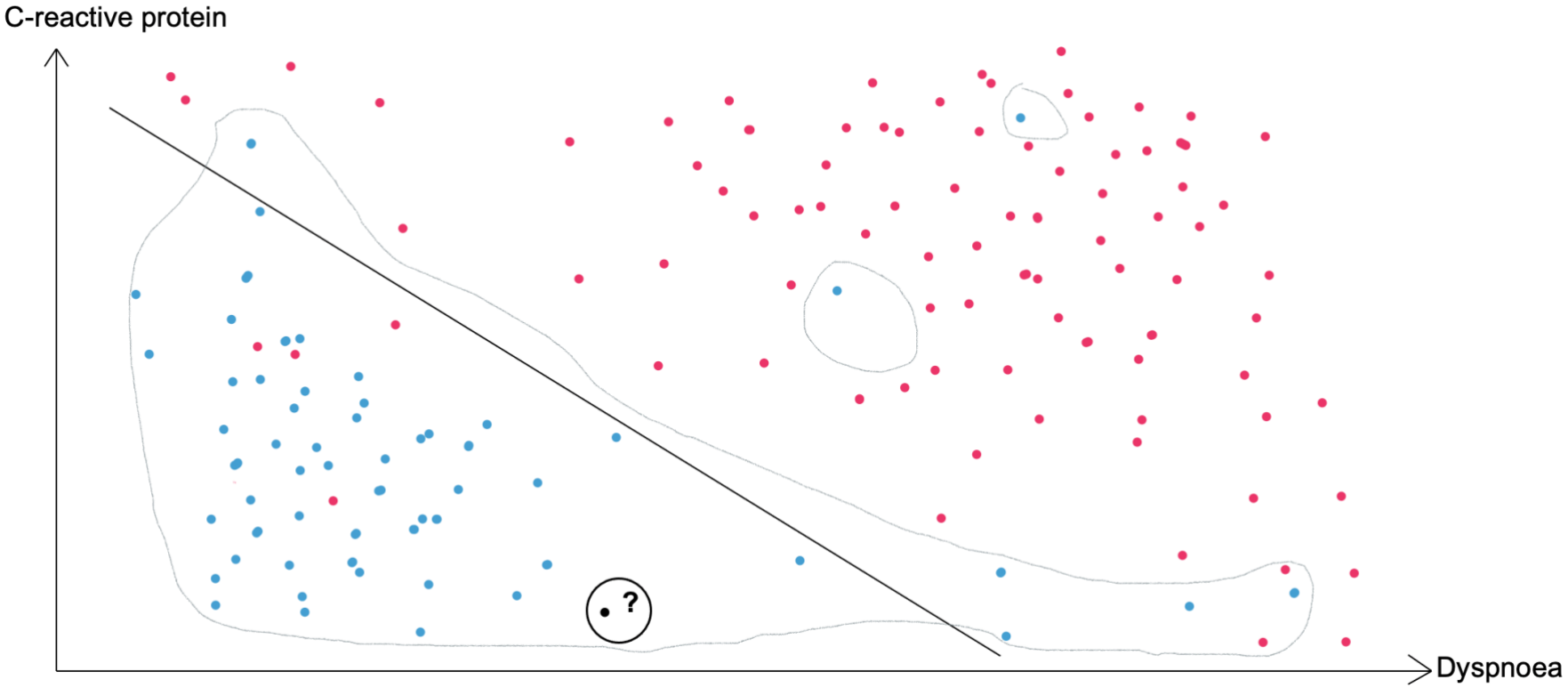
The goal in classification is to assign various data points (e.g. patients) to one or more classes or labels, hence the name supervised learning, based on the values of the variables associated with the data points (Figure 1.2).

Figure 1.2 | Illustration of the goal of classification. Consider a situation where patients are represented as data points described by two variables measured at exacerbation e.g. dyspnoea and inflammatory marker C-reactive protein (CRP) in the blood. Let blue and pink data points represent patients with a successful versus unsuccessful clinical outcome e.g. exacerbation treatment response, respectively. The goal of classification in this context would be to learn a pattern which separates the blue and pink data points on the basis of the values of dyspnoea and CRP. Therefore, for a new data point with given values for dyspnoea and CRP but unknown clinical outcome, the new data point (e.g. the patient represented by the encircled black data point with "?") is classified as belonging to either the successful or unsuccessful clinical outcome based on the learned pattern.



In most situations, the true classifier function is unknown and hence the goal in such analysis is to construct a function which can approximate the true classifier. The Bayes error is the lowest error obtainable when a true classifier classifies items in the case where the true class probabilities of all classes are known[37]. In classification problems, the Bayes error is, in reality, never reached because the actual nature of the true classifier is not known. The total error in addition to the Bayes error is comprised of two components, the variance and the bias[38]. These can be reduced by choosing an appropriate classifier model, training it with sufficient sample data representative of the population of interest and fine-tuning hyperparameters of the classifier model if there are any. Variance refers to the variation in the performance of the classifier based on the actual nature of the sample. Bias, on the other hand, is a systematic form of error due to the approximating function deviating systematically from the true classifier function. This normally occurs, for example, when the classifier is too simplistic a model. Therefore, a linear model where a linear decision boundary is used to separate different samples in the original feature space is likely to have higher bias than a non-linear classifier such as the K-nearest-neighbour (KNN) classifier which assigns new data points to whichever class the nearest K neighbours belong to[39,40]. However, the K-nearest-neighbour classifier is associated with a higher variance than a simplistic linear classifier model (Figure 1.3). Finding a suitable classifier hence involves a trade-off between variance and bias[38].

Figure 1.3 | Illustration of bias and variance trade-off for classifier models. Consider the situation previously depicted in Figure 1.2. The black decision boundary from a linear classifier model is too simple to fully separate the two patient groups, whereas the grey decision boundary from a KNN classifier model is effective at representing the complexity of separation. Therefore, the linear classifier model is characterised by greater bias. However, the grey decision boundary from a KNN classifier model is susceptible to slight changes in the values of dyspnoea and CRP of patients and the classifier error rate may increase in such cases. In contrast, slight changes in the sample of patients will have less effect on the classification error of the linear classifier model. Therefore, the linear classifier model is characterised by little variance whereas the KNN classifier model exhibits higher variance.



Linear classification approaches

The most commonly used linear classifier is logistic regression, in which the logarithm of the odds of belonging to one class is equal to the sum of an intercept term, an error term and the variable terms with their associated coefficients[38]. A key advantage of logistic regression, compared to non-linear techniques, is the interpretability of the model, where the size of the logistic regression coefficients indicates the importance of the associated variable in determining the class probability[38]. However, a notable disadvantage is that if the actual decision model is non-linear, then the logistic regression model must be modified (e.g. by including non-linear terms such as quadratic, cubic and higher order polynomial terms, as well as interaction terms between variables etc.) otherwise a different classifier which is non-linear must be used[38]. Furthermore, logistic regression is only effective when the number of variables is less than the number of samples[38]. In high dimensional low sample size problems where the number of predictor variables exceeds the sample size, a unique solution for the variable coefficients cannot be found and hence logistic regression in its most basic form is not usable[38]. One method of dealing with high dimensionality, whether the number of predictors is similar to or greater than the sample size, is to perform regularisation, such as via the least absolute shrinkage and selection operator (LASSO) in order to shrink coefficients towards zero[41]. In the case of lasso, fitting the classifier model and variable selection are performed simultaneously such that, in total, the number of variables included in the final model with non-zero coefficients will automatically not exceed the total sample size[41].

Non-linear classification approaches

Apart from the KNN classifier already described, there are a number of other non-linear classifier models. For example, the support vector machine (SVM) classifier finds the optimal hyperplane to maximise the distance between data points corresponding to the support vectors in order to construct what the decision boundary should be[42]. The support vectors are vectors from the origin to the particular data points closest to the decision boundary[42]. This can be performed in the case of linear decision boundaries. Even if the decision boundary is non-linear, it is possible to project the original feature space into a new, transformed, higher dimensional space and then in the higher dimensional space to construct a linear decision boundary which is equivalent to having a non-linear decision boundary in the original feature space before the transformation was applied. Furthermore, if the decision boundary is believed to be non-linear, then non-linear kernels, such as a polynomial or radial basis function, can be used instead. However, like KNN, the SVM classifier lacks straightforward interpretability. Similarly, neural network classifiers are capable of creating highly effective non-linear decision boundaries and also lack straightforward interpretability[43]. In a neural network, predictor variables are linearly combined and then passed through non-linear functions to create new variables which are located in a hidden layer[43]. The hidden variables are then recombined to form new hidden variables in further hidden layers[43]. After a specified number of hidden layers, the hidden variables in the final hidden layer are combined to determine the outcome class[43]. One non-linear classifier model which can maintain interpretability is the decision tree[38]. In a simple decision tree model, data are partitioned into separate groups based on variable cut-off values derived in order to maximise the separation of data points of different classes[38]. A clear potential limitation of such a decision tree is that a rectangular decision boundary must

be produced[38]. It would be impossible, for example, to produce an elliptical decision boundary using a simple decision tree[38]. Furthermore, variance of a simple tree-based classifier can be high since small changes in the data used to build the trees can lead to substantial changes in the decision tree structure[38]. There are a variety of modifications which can be made to improve tree-based models. A powerful variance-reducing extension of decision trees is the use of an ensemble of decision trees, the random forest classifier[44]. Whenever a new data point must be classified, each of the trees in the forest casts a vote and it is the total vote of the forest which determines the classification[44]. The reduction in variance errors in random forests compared to simple decision trees is due to the trees comprising the ensemble being decorrelated[44]. This decorrelation is achieved because each tree is built using only a randomly chosen subset of the total variables[44]. The result is that different trees use different subsets of variables, hence the decorrelation of trees and a reduction in the effect of noise variables[44]. As a result, classification results are much more stable. Both the SVM and random forest can be used in cases where the number of variables exceeds the number of data points, a notable advantage over the logistic regression classifier[38].

Selecting optimal classification approach

It is impossible to say *a priori* that a particular model will be the best one to use; instead, different models must be empirically compared with each other for the particular dataset at hand in order to decide which one is best for the task[33,34]. Hence, a variety of different models, all of which could be suitable for a certain task (bearing in mind, for example, the dimensionality of the data), should be selected and empirically compared following model training and evaluation[38]. For all machine learning

applications, it is essential that the dataset sample used for training represents the population of interest and that the signal to noise ratio is as high as possible[38]. If the population of interest is heterogenous but the sample is not, or if there is a lot of noise and the signal to noise ratio is relatively low, then it is highly likely that the classifier will have a high error rate associated with it when used for new members of the population (e.g. new patients). It is important to optimise the (hyper-)parameters of a classifier model[38]. This must be achieved based on the performance of the classifier. The dataset may be divided into a training set for estimating parameters and a validation set which is used to select the final model based on classifier performance[38].

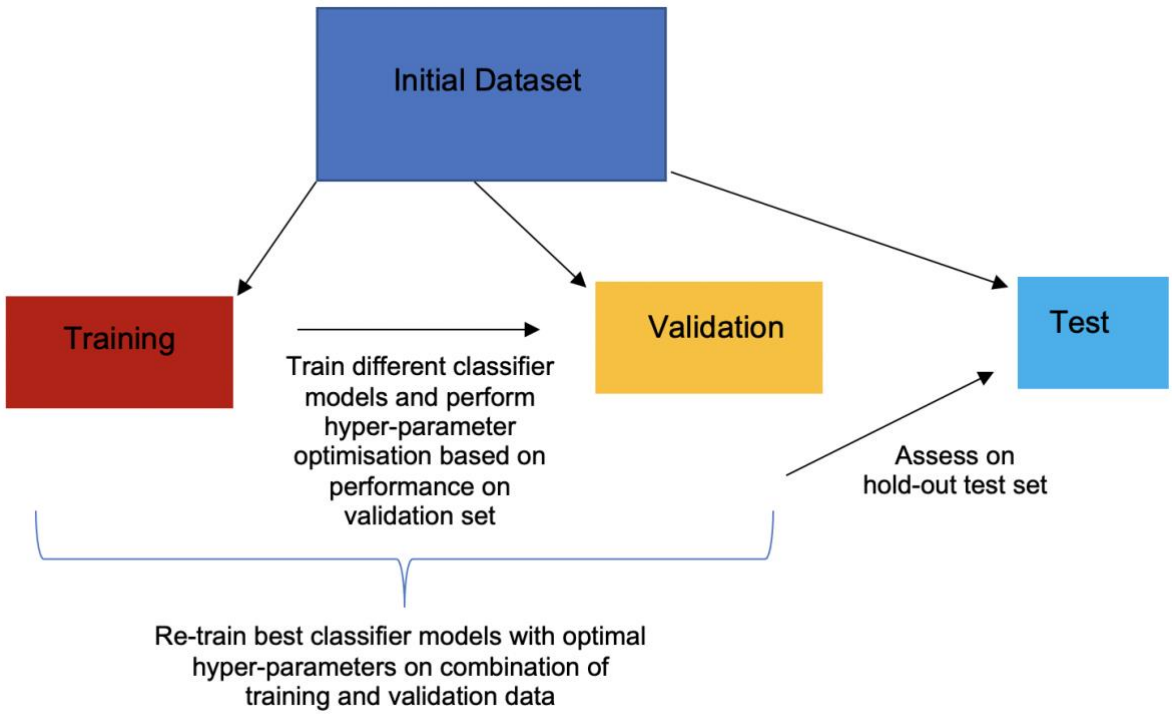
In addition to evaluating the performance of a classifier based on the overall proportion of correct classifications (accuracy) for all classes, it is sometimes, especially in clinical contexts, useful to assess the performance of the classifier in each of the classes. The sensitivity of a classifier in in detecting a particular class is equal to the proportion of true cases of this class which are correctly classified as belonging to that class. In contrast, the specificity of a classifier is the proportion of cases not belonging to a class which the classifier correctly classifies as not belonging to the class. In a purely balanced dataset in which there are two classes with equal number of data points, classifier accuracy is equal to the mean average of the classifier's sensitivity and specificity. A useful graphic to examine a classifier's performance is the receiver operating characteristic (ROC) curve, in which the classifier's sensitivity is plotted against the 1-specificty of the classifier for different thresholds of the classifier output probability[45]. A useful metric which summarises the performance of a binary classifier for different class probability thresholds is the area under the receiver operating characteristics curve (AUC). The AUC is defined as the probability that the classifier

will rank (e.g. assign a higher probability to) a randomly chosen data point belonging to the class of interest higher than a randomly chosen data point belonging to the other class[46]. As a result, a classifier with an AUC of 0.5 performs equally to using a coin toss in order to determine class membership[46]. Depending on the clinical application, it may be more important to maximise a classifier's sensitivity, or conversely, its specificity. For example, in the case of a screening test where the emphasis is on capturing all potential cases of a particular disease state, maximising sensitivity may be a priority. A highly sensitive test may in certain contexts be used to 'rule out' a disease for a patient with a negative test result since the vast majority of patients with the disease would be captured by the test. Such a screening test could then be followed by a highly specific diagnostic test for a final decision in relation to the patient having the disease. Implications of false negatives and false positives on patient psychology, the danger of a lack of/improper treatment intervention and health economics may all influence which sensitivity/specificity trade-off is most suitable for a particular clinical context. In contexts (such as in this thesis) in which classifiers are first developed to predict a particular target outcome, the AUC is a useful metric since the AUC of a model enables both sensitivity and specificity to be taken into account. However, in subsequent, future studies testing the performance of a classifier in widespread clinical use, a particular sensitivity and specificity threshold may be selected. This can be achieved by using a threshold for the classifier's output probability other than 0.5 to decide class membership if the ROC shows that a different probability threshold would yield a contextually more suitable sensitivity and specificity than that resulting from the 0.5 probability threshold. In addition, the prevalence of the target outcome in a relevant population of patients can be important when evaluating the usefulness of an already developed classifier model for informing clinical decisions.

Positive and negative predictive value of a classifier model may be employed in such contexts to reflect the proportion of classifier model predictions positive and negative respectively for the target outcome which correspond to true positive and true negative cases of the target outcome respectively.

Once a classifier has been entirely specified following selection of optimal hyper-parameters based on the classifier's performance in the validation set, the classifier is retrained on all training and validation data combined but using the same parameters as pre-specified and then assessing the classifier's performance in a hold-out test set which has not been used in any way before to tune hyper-parameters[38] (Figure 1.4). Training and assessing algorithm performance on the same data, rather than on a separate validation set, should never be done. This approach, often referred to as 'apparent validation', is fundamentally flawed since it does not provide an indication of the generalisability of the algorithm[38].

Figure 1.4 | Schematic to illustrate splitting of dataset into different portions to enable evaluation of different classifier models and hyperparameter optimisation.

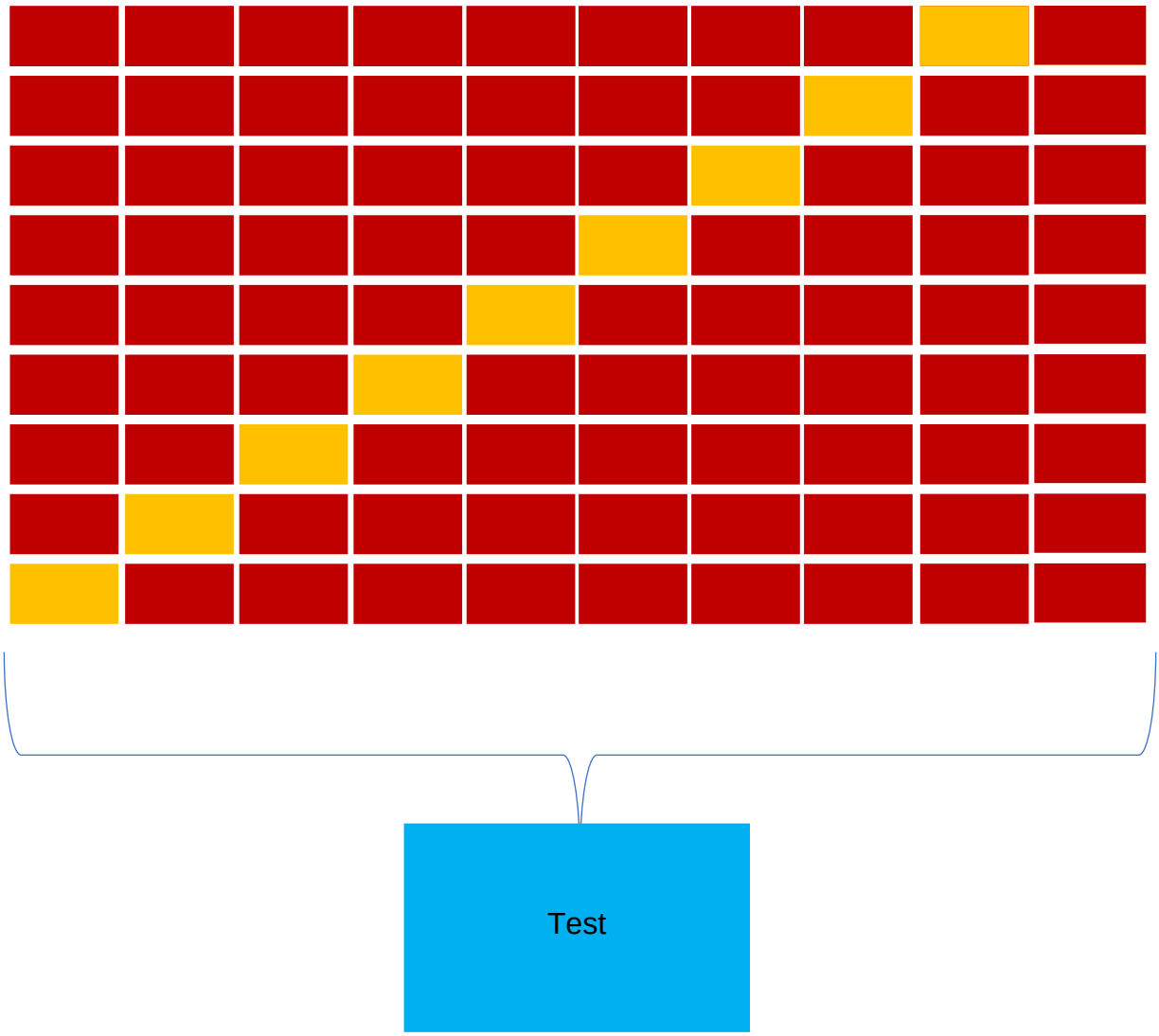


Cross-validation

In cases where the sample size is small, it may not be appropriate to have separate training, validation and hold-out test sets because one needs to maximise how much data is available for training the classifier in the first place[38]. One also needs sufficient data for testing, otherwise the estimated error rate of the classifier based on performance in the test set will be much more influenced by the particular data points in the test set rather than being a truly generalisable estimate. Instead of dividing a dataset into training, validation and hold-out test sets, a k-fold cross-validation approach may be used. In k-fold cross-validation, the entire dataset is used for training but is split into different k-folds which can be stratified by class to ensure that each of the training portions is balanced with respect to class membership[38]. In k-fold cross-validation, the classifier is trained on (k-1) folds of the data and tested on the remaining fold[38]. For example, in 10-fold cross validation, the data will be split into 10 equally

sized segments. This is repeated for each of the different folds. Based on the overall e.g. average error of cross-validation, hyper-parameter optimisation can be performed and the optimal model chosen[38]. Once this has been done, ideally the fully specified model should be assessed in a completely new test set not part of the cross-validation data (Figure 1.5).

Figure 1.5 | K-fold cross-validation illustrated for k = 10, followed by hold-out test set. A hold-out test set is first separated from the initial dataset. The remaining data not used for the test set are divided into 10 folds for the cross-validation scheme. Classifier training is performed on 9 of the 10 folds (shown in red) and assessed on the remaining validation fold (shown in yellow). The procedure is repeated another 9 times, each time using a different fold for validation and the remaining 9 folds for training. The optimal classifier model and hyperparameter values are selected based on average performance across the 10 cross-validation folds. This optimal model is retrained on the entire cross-validation data (both training and validation folds). Finally, the retrained model is assessed on the hold-out test set (shown in blue) to give an unbiased estimate of classification performance for unseen data, e.g. clinical outcome classification performance for new patients.



The cross-validated error has been shown theoretically and empirically not to be an unbiased estimate of the true error rate of the classifier[47]. A hold-out test error is still required. Hence, if sample size is very small and cross-validation is performed but no hold-out test set is used, then model performance metrics should be interpreted as exploratory in nature rather than constituting an unbiased estimate of classifier performance.

Feature selection

For classification problems, an important challenge lies in achieving appropriate variable selection. Including noise variables can significantly degrade the performance of a classifier. The sparsity of data in high dimensions due to the curse of dimensionality makes variable selection particularly challenging in high dimensional contexts, both due to the increased danger of selecting noise variables and due to the sparsity of data points in high dimensional space[19,38]. Hence, in high dimensional contexts it may be advisable to first perform dimensionality reduction before training the classifier. Some classifiers, such as the LASSO, automatically perform variable selection as part of regularisation procedures[41]. For classifiers in general, variable selection can be achieved using a wrapper approach or a filter approach[38]. In the latter case, the limitation is that variables may not individually predict the outcome of interest but may do so when combined with other variables[38]. Hence, performing univariate analysis in a filter approach will not necessarily select the best combined subset of variables for the classification task[38]. In the case of wrapper methods, such as (recursive) feature elimination, forward selection or a combination of these approaches, it is essential to ensure that the entire variable selection procedure forms part of the cross-validation bootstrap loop[38]. Ambroise and McLachlan have shown

that if variable selection is performed using the full dataset, rather than being performed on subsets of the dataset then, in the cross-validation loop, highly biased, erroneous results will follow[48].

1.4 Current definitions of COPD exacerbations

The outputs of cluster analysis and classification analysis are by definition influenced by which patient samples are included as well as, in the case of classification, which outcome label is assigned to each sample for algorithm training. Hence, the definition of exacerbations of COPD is foundational in any clustering studies and exacerbation treatment follow-up studies. Further background to COPD exacerbations and the strengths and limitations of current definitions of exacerbations of COPD will now be discussed.

Exacerbations of COPD cause accelerated lung function decline and negatively impact on patient health status[49-51]. Furthermore, exacerbation history is a good predictor of future exacerbation risk and exacerbations may require hospitalisation, increasing mortality risk and causing a large economic burden for the National Health Service[11,51-53]. According to GOLD: "*An exacerbation of COPD is an acute worsening of respiratory symptoms that results in additional therapy*"[6]. A strength of this definition is that it is based on 1) whatever respiratory symptoms an individual patient feels have worsened and 2) the extent of worsening of respiratory symptoms required to necessitate additional treatment and qualify as an exacerbation is determined by the patient. Hence, the sensitivity of the GOLD definition for capturing an acute episode of COPD worsening i.e. the proportion of true cases which are captured by the GOLD definition, is maximised. However, the GOLD definition obscures the fact that exacerbations in

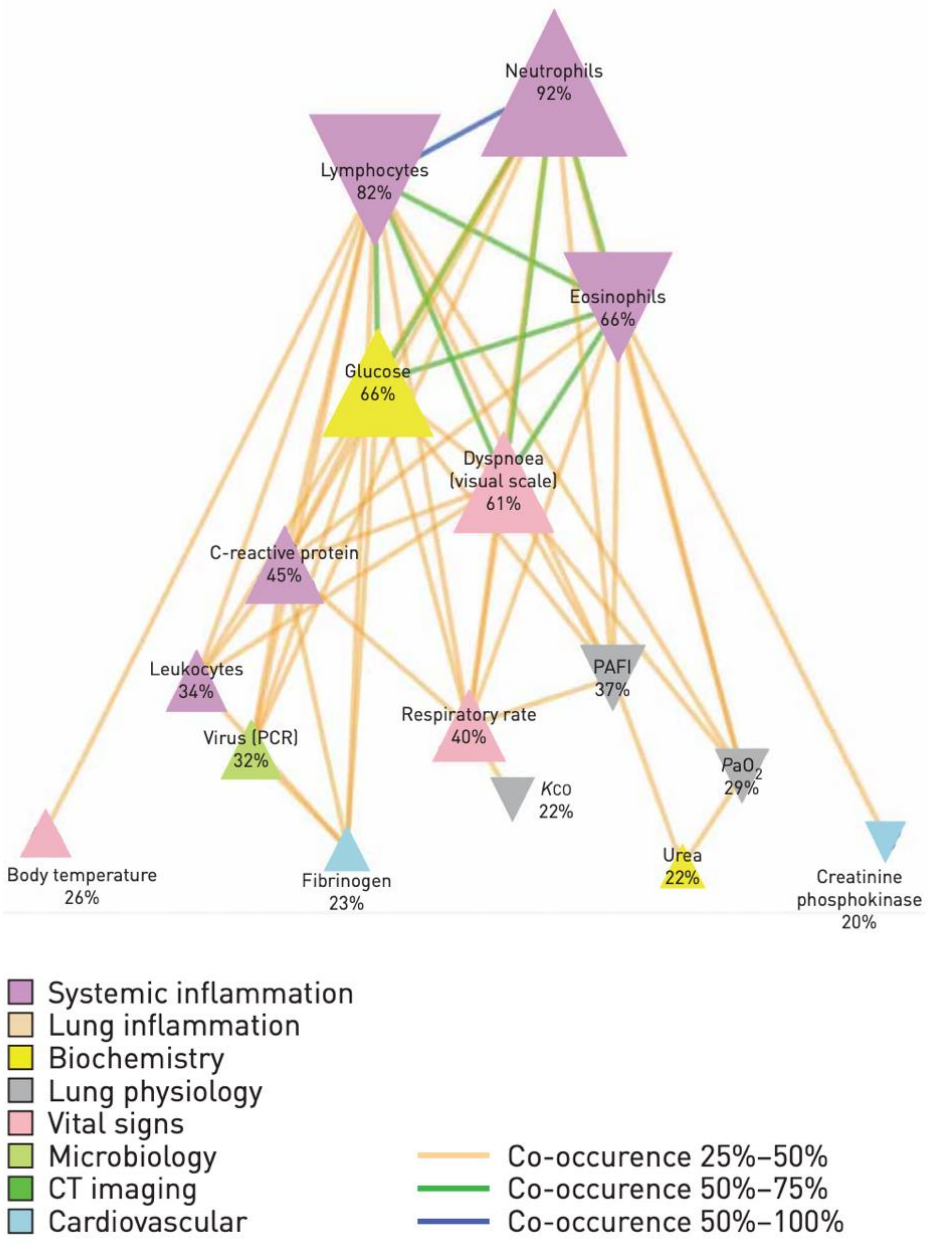
different patients or exacerbations at different time points in the life of a given patient may be characterised by different respiratory symptoms and different severity of increase in respiratory symptoms. The NICE (2010) definition is more detailed: "*An exacerbation is a sustained worsening of the patient's symptoms from their usual stable state which is beyond normal day-to-day variations, and is acute in onset. Commonly reported symptoms are worsening breathlessness, cough, increased sputum production and change in sputum colour. The change in these symptoms often necessitates a change in medication."*[54]. This is based on the definition of Rodriguez-Roisin: "*a sustained worsening of the patient's condition, from the stable state and beyond normal day-to-day variations, that is acute in onset and necessitates a change in regular medication in a patient with underlying COPD*"[55]. Both the NICE and the Rodriguez-Rosin definitions retain the strength of the GOLD definition of allowing a patient's perception of increase in symptom severity and the patient's perception of the need for additional medication to determine whether or not the symptom worsening classes as an 'exacerbation'. However, the NICE and Rodriguez-Rosin definitions have additional strengths compared to the GOLD definition in emphasising that: 1) the worsening of symptoms is sustained; 2) the worsening of symptoms is greater than worsening which may occur from day to day when the patient is in a stable COPD state; 3) examples of symptoms which commonly worsen at exacerbation are provided. Nevertheless, the NICE definition does not indicate the existence of different categories of exacerbation severity. Rodriguez-Rosin further subdivides exacerbations into the categories mild, moderate and severe according to whether the additional medication could be managed by the patient in their normal environment, requires additional medical assistance or experiences a rapid deterioration in condition necessitating hospitalisation, respectively[55]. However, all of the aforementioned

exacerbation definitions and severity categorisations are qualitative. No quantitative indication, e.g. via patient reported symptom scores of the magnitude of changes in respiratory symptoms, is provided. In addition, despite the strength of the above definitions in being simultaneously symptom- and event-based, they do not provide any insight into what constitutes stable state day-to-day variation. An even more significant shortcoming of these exacerbation definitions is that they do not take into account any biological information, such as blood and sputum inflammatory profiles. Recently, the ROME proposal emphasises that symptom worsening during exacerbations may be accompanied by tachypnoea and/or tachycardia and is often associated with increased local and systemic inflammation[56]. In the ROME proposal, a Delphi-based consensus of 17 experts agreed thresholds of dyspnoea, respiratory rate, heart rate, oxygen saturation and serum C-reactive protein to distinguish mild, moderate and severe exacerbations[56]. The ROME proposal's inclusion of quantitative thresholds of symptom, biological and physiological variables associated with different severities of exacerbations based on expert opinion and a thorough review of literature constitutes a step forward in the field. However, the definition of exacerbations compared to stable state used in the literature which was reviewed to derive the foundation for the ROME proposal Delphi-based consensus still corresponds to conventional clinical definitions. Hence, the ROME proposal is in part influenced by the opinions of physicians based on their current clinical practice and not entirely empirically derived.

On the surface, the Noell et al (2017) study may appear to be the closest study to-date in objectively defining exacerbations of COPD[57]. This was a multicentre trial including 86 patients hospitalised for exacerbations of COPD[57]. All patients were characterised

during the first 72 hours of their hospitalisation as well as at stable state at least three months after hospital discharge[57]. Multilevel Spearman correlation networks were used to examine the relationship between qualitative and quantitative clinical, functional, biological, microbiological and imaging variables at exacerbation compared to stable state[57]. Compared against stable state, the authors found that exacerbations were associated with a disruption of a multilevel correlation network with reduced links between variables[57]. The authors next identified all variables for which a significant proportion of values at exacerbation were outliers compared to the values at stable state (Figure 1.6)[57].

Figure 1.6 | Correlation network showing 16 'outlier' variables with values significantly higher (triangles pointing upward) or lower (triangles pointing downward) at exacerbation compared to stable state. Triangle size indicates the proportion of cases in which the exacerbation value was different from stable state value (exacerbation value > 95th percentile of stable state value or < 5th percentile of stable state value). Colour of edges connecting any two triangles corresponds to the extent to which the outliers (variable with exacerbation value outside stable state 5th-95th percentile range) co-occurred in the same patient. Abbreviations: P_aO2: arterial partial pressure of oxygen; PAFI: PaO2 (mmHg)/inspired fraction of oxygen ratio (%); K_{CO}: carbon monoxide diffusing capacity of the lung/alveolar volume (transfer factor). Adapted from Noell et al. (2017)[57].



All such variables which also individually achieved an AUC of 0.8 for detecting exacerbations were then entered into a general mixed linear model[57]. It was found that at exacerbation the combination of dyspnoea of ≥5 on a 0-10 visual analogue scale, circulating neutrophils ≥70% and CRP ≥3 $mg \cdot L^{-1}$ had a specificity of 96% and sensitivity of 90% for diagnosing exacerbations of COPD[57]. However, the Noell et al (2017) study exhibits several issues, such as the lack of any indication of validation of their exacerbation diagnostic model. The authors seem to have derived the specificity of 96% and sensitivity of 90% of their model from training data rather than a separate test set, rendering the conclusions of this study open to debate. Furthermore, the high proportion of circulating neutrophils during exacerbations and the low proportion of circulating eosinophils may be an artefact of the patients in the study receiving SCS before biological samples were taken for analysis.

To-date there is no validated high specificity and sensitivity definition of COPD exacerbations which is entirely empirically derived and incorporates inflammatory biomarkers together with symptoms.

1.5 Cluster analysis – COPD studies

The ideal application of cluster analysis for COPD would achieve one or both of the following: a) attempt to find clusters based on biological variables in order to identify biologically different subgroups of patients who may therefore benefit from different treatment even though these patients may be indistinguishable in terms of symptoms; b) provide superior prognostic information than currently established clinical prognostic indicators. I performed a literature review including cluster analysis studies which 1) compare different COPD exacerbation or stable state subgroups in terms of clinical

outcome measures or 2) compare different COPD exacerbation or stable state subgroups in terms of biological differences e.g. in inflammatory markers. In Appendix Table A1.1 and Appendix Table A1.2, I outline in detail the findings of each[58-69]. The vast majority of COPD cluster analysis studies either make serious mathematical methodological errors or are limited in their clinical usefulness due to the choice of variables included in the cluster analysis. Furthermore, cluster analysis studies examining exacerbations rely on the existing, subjective exacerbation definition. Lacking from all COPD studies performed to date is the use of cluster analysis to identify objective, mathematically-based COPD disease states rather than using the existing, subjective clinical definitions of exacerbations versus stable state as a starting point. In addition, cluster analysis has not been performed on post-exacerbation treatment time series of biologically and symptom-based data.

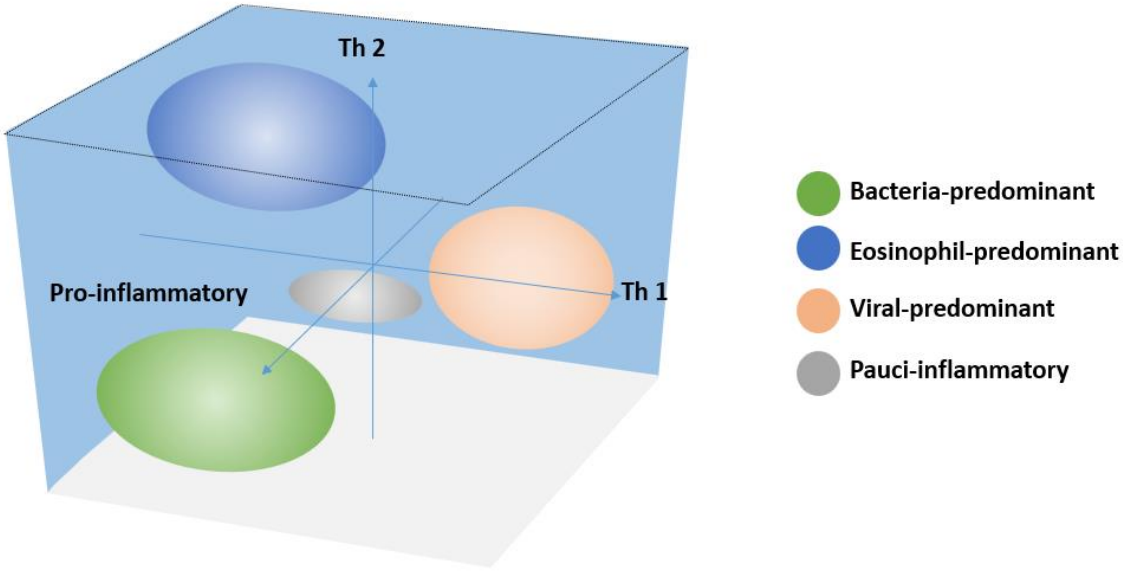
Some pertinent examples of the findings, strengths and limitations of COPD cluster analysis studies will now be discussed. Burgel *et al*. (2012) conducted a multi-centre study of 527 stable COPD patients from across all four GOLD severity categories[61]. Seven continuous variables (age, Body Mass Index (BMI), FEV_1, breathlessness, quality of life score and thoracic gas volume and diffusing capacity) in addition to numerous categorical variables relating to comorbidities and CT-based imaging data, were considered[61]. After separately performing PCA for the continuous variables and MCA for the categorical variables, 2 principal components and 14 MCA axes were selected[61]. Ward's method of hierarchical clustering was then performed on these new variables and a suitable cluster number was decided upon based on visual examination of the cluster dendrogram[61]. The authors identified 3 clusters: 1) mild to moderate airflow limitation and other respiratory disease features; 2) moderate to severe airflow limitation and other respiratory disease features, older patients, obesity

common, diabetes common, cardiovascular comorbidities common; 3) severe airflow limitation and other respiratory disease features, osteoporosis and muscle weakness common, cardiovascular comorbidities rare[61]. Longitudinal follow-up to determine all-cause mortality revealed that patients in clusters 2 and 3 had higher all-cause mortality than cluster 1 patients[61]. Although the Burgel *et al*. (2012) study is a seminal study which has been highly regarded by many in the COPD field, the limitations of this study are numerous. Firstly, the authors should not have separately applied PCA and MCA for continuous and categorical variables, respectively, but rather they should have used FAMD to simultaneously take into account the relationship between the continuous and categorical variables. This error casts doubt on the validity of the authors' choice of PCA- and MCA-derived new variables which they used for cluster analysis. This may reduce the reliability of the cluster solutions obtained. Secondly, the finding that patients in clusters 2 and 3 had higher all-cause mortality than in cluster 1 is not at all surprising since it is already known that dyspnoea (breathlessness) is a predictor of mortality[12]. Since the mortality associated with patients in the three clusters was not adjusted according to the differences in age and dyspnoea between the three clusters, it is far from clear whether the 3 clusters provide any prognostic value greater than simply using dyspnoea and age to predict mortality on a continuous scale. Thirdly, no biological variables, for example, inflammatory markers in the patients' blood or sputum were included in the analysis. This means that the clusters shed no light on different subgroups of patients at the level of biology. Finally, the authors did not assess whether the identified clusters in fact represent three natural substructures of the data or whether the patients are located on a continuum in terms of the distribution of the variables analysed in the study. The limitations discussed above are also seen in other COPD cluster analysis studies. Other studies appropriately applied

dimensionality reduction but were still limited in their clinical usefulness. For example, Rennard *et al* (2015) examined 2,164 stable state COPD patients across GOLD stages 2-4, measuring a total of 41 variables including clinical, physiological, biological and imaging data[63]. After performing factor analysis to reveal 13 factors which explained most of the variation in the data, the authors performed a cluster analysis on the variables corresponding to the highest loading for each of the 13 factors[63]. Five clusters were identified using silhouette width to assess cluster quality: A) patients with mild airways disease; B) patients with intermediate values for health status and emphysema but low levels of inflammatory markers; C) patients with high levels of systemic inflammation and multiple comorbidities; D) patients with severe respiratory disease in the form of low FEV_1 and severe emphysema; E) patients with intermediate values for most variables[63]. Longitudinal follow-up showed that clusters C and D had the highest all-cause mortality and cluster D had the shortest time to first exacerbation[63]. Unfortunately, as in the Burgel *et al*. (2012)[61] study, the Rennard *et al* (2015) study did not adjust the mortality and exacerbation risk differences between clusters by the differences in age, dyspnoea and lung function between clusters. Since the latter variables formed part of the cluster analysis and were not adjusted for, it is unclear whether the identified clusters provide any new clinical insights. Perhaps the most promising COPD cluster analysis study to-date is the Bafadhel *et al*. (2011) single-centre study which analysed 182 exacerbation events from 86 patients across all four GOLD severity categories[60]. Seventeen sputum inflammatory markers were considered for cluster analysis[60]. Importantly, all measurements were made before the patients received exacerbation treatment with oral corticosteroids and/or antibiotics[60]. Factor analysis revealed three factors which explained most of the variation in the data[60]. The authors then performed Ward's clustering on the sputum variables with

highest loading for the three factors and visually examined the cluster dendrogram[60]. Four clusters were visually identified from the dendrogram, after which the k-means algorithm was used to find 4 clusters[60]. The 4 biological clusters were characterised by: 1) a strong pro-inflammatory profile, with the exacerbations being associated with bacterial infection; 2) high levels of type 2 (Th2) immune mediators, with exacerbations associated with eosinophil inflammation; 3) high levels of type 1 (Th1) immune mediators, with exacerbations associated with viral infections; 4) low levels of all sputum variables included in the cluster analysis[60]. Importantly, the authors visualised the 4 clusters by plotting the data points in their clusters in a 3-dimensional space (see Figure 1.7) where the 3 axes corresponded to the 3 factors identified from the factor analysis[60].

Figure 1.7 | Biological COPD exacerbation clusters shown in three-dimensional space using three factors (proinflammatory, Th1 and Th2) as coordinate axes. Four distinct clusters are seen (bacteria-predominant, eosinophil-predominant, viral-predominant and pauci-inflammatory). Adapted from Bafadhel et al. (2011)[60].



The visual analysis confirmed that the identified clusters were real, not just artefactual. Furthermore, the clusters were not distinguishable in terms of clinical characteristics, making a compelling case for the existence of biological heterogeneity underlying COPD exacerbations which cannot be seen from purely clinical presentation[60]. Studies focusing on inflammatory data as input for cluster analysis are essential since they enable COPD heterogeneity to be investigated at the level of disease aetiology and inflammatory pathophysiology, unlike studies which focus on clinical and symptomatological data but neglect inflammatory biology[27].

Although Bafadhel *et al* (2011) identified several blood and sputum inflammatory biomarkers which increased at exacerbation compared to stable state, no individual biomarker achieved an AUC of more than 0.70 for distinguishing between exacerbation and stable state[60]. Multivariate classification models for distinguishing between

exacerbation and stable state were not investigated in the Bafadhel *et al* (2011) study, so whether a combination of blood, sputum and symptom biomarkers could be used to define a COPD exacerbation compared to stable state in the study remains unknown.

1.6 Classification to distinguish COPD exacerbations from stable state

The ability to identify an exacerbation in a COPD patient is key in order for patients to receive relevant treatment. An illustrative example of a study investigating the usefulness of biological variables for distinguishing between exacerbations and stable state is the Hurst et al (2006) study which investigated whether exacerbations could be distinguished from stable state in terms of plasma biomarkers and symptoms[70]. Symptoms were grouped as either major (dyspnoea, sputum volume or sputum purulence) or minor (cough, wheeze, sore throat or coryza) and exacerbations were defined as episodes of at least two consecutive days of new or increased patient symptoms, at least one symptom being a major symptom[70]. A total of 36 candidate plasma biomarkers were analysed in 90 paired baseline stable state and exacerbation samples from 90 COPD patients[70]. Importantly, as in the Bafadhel *et al* (2011) study, Hurst et al (2006) obtained the blood samples for analysis of plasma biomarkers at exacerbation before the exacerbation treatment was given to the patients[70]. It was shown that plasma C-reactive protein (CRP) concentration could distinguish exacerbations from stable state with area under the receiver operating characteristic curve (AUC) of 0.73, but this was lower than the AUC of 0.83 when using the presence of an increase in any one of the major symptoms on a day without considering CRP[70]. Hurst et al (2006) showed that a combination of CRP together with an increase in any one of the ‘major symptoms’ dyspnoea, sputum production volume and sputum

purulence may be more promising for distinguishing exacerbations from stable state, with the AUC increased to 0.88[70]. The presence of at least one major symptom recorded on a day together with plasma CRP 8 mg/L enabled detection of exacerbations with a specificity of 95% and a sensitivity of 57%[70]. In addition to analysing the ability of CRP alone, or with symptoms, to be used to detect exacerbations, Hurst *et al* (2006) also analysed pairs of the 36 plasma biomarkers, combinations of 3 of the 36 plasma biomarkers and all 36 plasma biomarkers in total[70]. This did not lead to higher performance compared to the use of a combination of CRP together with an increase in any major symptom[70]. Complementary analysis of other possible combinations of the 36 plasma biomarkers and symptoms as part of a variable importance ranking procedure may provide further insight. For example, it would be interesting to train a classifier model to use all 36 plasma biomarkers and major or minor symptoms to distinguish between exacerbation and stable state. This could be followed by using a statistical approach to identify the ideal combination of biomarkers for detecting exacerbations in a multivariable context. This CRP model by has yet to be established in clinical practice. This may be as a consequence of the absence of validation of this model in external cohorts. Nevertheless, the Hurst *et al* (2006) approach shows promise for the use of objective biomarkers for exacerbation detection[70]. A challenge of all studies to-date is that current clinical definitions of exacerbation versus stable state are used to train algorithms rather than using an objective, empirically derived and mathematically-based definition of exacerbations.

1.7 COPD exacerbation treatment outcome studies

Another important priority in COPD management is being able to classify patients into those who, after an exacerbation, respond successfully to exacerbation treatment versus those patients who do not respond successfully to exacerbation treatment. This

requires that a classifier is trained to discriminate between cases based on the values of relevant biomarkers measured at relevant time points, a biomarker being defined as any “defined characteristic that is measured as an indicator of normal biological processes, pathogenic processes or responses to an exposure or intervention.”[71]. Patient outcomes after standard exacerbation treatment with SCS and/or antibiotics, the aim of which is to reduce the impact of the present exacerbation and reduce the risk of future exacerbations, are variable[6,72]. Exacerbations may be managed in either an inpatient or an outpatient setting, according to how severe are the exacerbation and COPD in general[6]. Some patients require additional major treatment after the initial exacerbation treatment and in the case of inpatients, hospital readmission may occur[6]. At the same time, SCS are known to have a number of harmful side effects, including significantly increasing the risk of osteoporosis[6]. In a 2014 Cochrane review by Walters *et al*, whose primary outcome was treatment failure, the effectiveness of SCS versus placebo for treating acute exacerbations of COPD was compared[15]. Treatment failure was defined as the need for intensification of treatment and/or emergency department (ED)/hospital readmission during a period ranging from between 2 and 30 days, depending on the study (Table 1.2)[15]. There were 9 studies including 917 patients which contributed data for the comparison of SCS against placebo for treatment outcome (Table 1.2)[15].

Table 1.2 | Characteristics of 9 studies contributing data in the Cochrane analysis of the effectiveness of SCS in reducing treatment failure for acute exacerbations of COPD. Adapted from Walters et al (2014)[15].

Study	SCS			Placebo			Odds Ratio	Time Period for Treatment Failure Definition
	Treatment Failures	Total	Proportion of Treatment Failures (%)	Treatment Failures	Total	Proportion of Treatment Failures (%)		
Inpatient (treatment > 3 days)								
Bullard *et al* 1996[73]	5	60	8	14	53	26	0.25	< 14 days
Chen *et al* 2005[74]	5	43	12	6	43	14	0.81	Not known
Davies *et al* 1999[75]	1	29	3	5	27	19	0.16	< 14 days
Maltais *et al* 2002[76]	3	62	5	8	66	12	0.37	< 11 days
Niewoehner *et al* 1999[77]	34	160	21	35	111	32	0.59	< 30 days
Wood-Baker et al 1998[78]	1	13	8	4	13	31	0.19	< 7 days
Inpatient (treatment 1 day)								
Emerman *et al* 1989[79]	8	38	21	4	32	13	1.87	< 2 days
Outpatient								
Aaron *et al* 2003[80]	19	70	27	30	70	43	0.50	< 30 days
Thompson *et al* 1996[81]	0	13	0	8	14	57	0.00	< 14 days
Overall								
All Studies	76	488	16	114	429	27	0.48	N/A

Definition of abbreviations: SCS = systemic corticosteroids.

Although for the 9 studies SCS reduced treatment failure risk compared to placebo with an odds ratio of 0.48 (weighted according to the proportion of inpatient treatment > 3 days, inpatient treatment 1 day and outpatient studies out of the total 917 patients), 1 of the 9 studies (Emerman *et al*, 1989[79]) showed reduced treatment failure for placebo (odds ratio of 1.87). Strikingly, even for exacerbations treated with SCS, the overall percentage of exacerbations which end in treatment failure for the 9 studies is 16% and ranged from 0% to 27% depending on the individual study. Furthermore, nine exacerbations would need to be treated with SCS in order to prevent one treatment failure. Hence, it can be seen that there is an urgent need to understand a) overall, which exacerbations result in treatment failure, and b) for which exacerbations does the use of SCS enable treatment success which would not have been possible without SCS. This would enable, a) patients who are predicted to fail treatment after an exacerbation to be monitored more closely following treatment, and b) SCS to be presrcibed only for exacerbations in which it is predicted that treatment failure would occur if SCS had not been given.

There has been progress in directing systemic corticosteroid treatment at exacerbation according to the characteristics of patients at exacerbation, also demonstrating that the results of the Bafadhel *et al*. (2011) study have a direct application in treatment personalisation for COPD patients. For example, Bafadhel *et al* (2012) performed a randomized blood eosinophil-directed double-blind systemic corticosteroid versus standard therapy single centre study[82]. Treatment failure in this study was defined as occurring when new treatment was started/repeated within 30 days from randomisation, hospitalisation from any cause, or death[82]. In the standard treatment arm (80 exacerbations), all exacerbations were treated with systemic corticosteroids for 2 weeks[82]. In the blood eosinophil-directed treatment arm (86 exacerbations), only

exacerbations for which blood eosinophil percentage > 2% were given systemic corticosteroids, whereas for exacerbations for which blood eosinophil percentage was ≤ 2%, placebo was given[82]. It was shown that for exacerbations with blood eosinophil percentage ≤ 2%, treatment failure was more common for exacerbations treated with systemic corticosteroids than those treated with placebo (15% versus 2%, respectively; p = 0.04)[82]. Furthermore, although for exacerbations treated with systemic corticosteroids there was no difference in treatment failure rate between the > 2% and the ≤ 2% blood eosinophil exacerbations (8% versus 15%, p = 0.23), exacerbations with blood eosinophil percentage > 2% exhibited greater symptom improvement (determined by patient self-reported outcomes) over the 2-week period following the exacerbation[82]. Together, these findings suggest that it may be beneficial to prescribe systemic corticosteroids for exacerbation treatment only in cases where patients have a 'high' (in the Bafadhel *et al* (2012) study, > 2%) blood eosinophil percentage at the time of exacerbation[82]. The use of blood eosinophils to direct systemic corticosteroid therapy has also been assessed in a multicentre context. In the Sivapalan et al (2019) CORTICOsteroid reduction in COPD (CORTICO-COP) study, all patients received intravenous methylprednisolone on the first day and were randomly assigned to a standard therapy and an eosinophil-guided therapy group[83]. On subsequent days (up to a maximum of 4 days) the eosinophil-guided group were given oral prednisolone on days when their blood eosinophil count was $\geq 0.3 \times 10^9$ cells/L[83]. Importantly, the eosinophil-guided therapy was non-inferior compared to standard care in terms of the number of days alive and out of hospital[83]. However, there are important limitations of these studies. In the Bafadhel *et al* (2012) study, the 2% blood eosinophil cut-off was derived for ensuring, with high sensitivity, that exacerbations with sputum eosinophilia are treated with systemic corticosteroids on the basis of a previous study[82]. Similarly,

in the Sivapalan et al (2019) study the $\geq 0.3 \times 10^9$ cells/L threshold was determined based on another study (Yun et al, 2018[84]) in which the relationship between blood eosinophil levels and exacerbation risk was assessed[83]. However, the optimal cut-off of blood eosinophils for directing treatment with the purpose of minimising the treatment failure rate is not known. Secondly, in both Bafadhel *et al* (2012) and Sivapalan et al (2019), blood eosinophils were used as the only biomarker for directing oral systemic corticosteroid prescription at exacerbation. Both of these limitations could be addressed by developing empirically-based multivariate classifiers which are trained on exacerbation blood- and sputum-based biomarker data, with each exacerbation labelled as belonging to the treatment failure or treatment success class based on the definition of treatment failure in a study. Furthermore, different definitions of treatment failure should be explored, including the improvement of different kinds of symptoms as determined by patient self-reported outcome measures. An example of an effective machine learning strategy to develop multivariate treatment outcome prediction models is the Goto *et al* (2019) study, in which hospital readmission within 30 days of discharge of the initial hospitalisation constituted treatment failure[85]. A total of 44,929 unplanned COPD hospitalisations were analysed from a Japanese database, of which 3413 (7%) resulted in hospital readmission within 30 days of discharge of the initial hospitalisation[85]. Then, 70% of cases were randomly selected to constitute training data and the remaining 30% cases were set aside as test data[85]. Variables included patient age, sex, body mass index, smoking index, daily life indices, severity of respiratory condition, number of comorbidities, mode of arrival, number of admissions in the past 12 months prior to index hospitalisation, duration from last admission to index hospitalisation, in-hospital procedures, surgical interventions, and medications[85]. Logistic regression, LASSO regression and deep neural network

classifiers were trained on the 70% of hospitalisation cases and evaluated in the 30% test cases to produce AUC metrics[85]. Logistic regression had an AUC of 0.57 which was increased to an AUC of 0.61 for both lasso regression and deep neural network models, respectively[85]. Goto et al showed that male sex, along with hospital interventions such as tube feeding duration, blood transfusion, thoracentesis use, are predictors of 30-day readmission for COPD patient hospitalisations[85]. Unfortunately, these AUC results are poor, showing a discriminative ability of the models for predicting whether or not COPD hospitalisations result in hospital readmission within 30 days of discharge from the initial hospitalisation. Notably, neither blood-based nor sputum-based data were included in the Goto *et al* study. This may explain the low predictive performance of the classifiers, especially because of studies, such as the Bafadhel *et al* (2012) study, showing the usefulness of blood eosinophils for directing exacerbation treatment. A major limitation of the Goto *et al* study is that the hospital admissions were not necessarily all due to exacerbations defined in the same way and as a result there is likely a lot of heterogeneity in the data, which would increase the difficulty of the classification problem[85].

As it can be seen from the discussed studies, existing treatment response prediction remains poor. The relationship between treatment type and symptom and inflammation resolution are imprecise. Existing attempts at developing quantitative COPD exacerbation definitions rely on classifying patient profiles as exacerbation versus stable state based on the values of variables at exacerbation versus stable state according to existing clinical definitions. An objective, empirically derived definition of exacerbations will enable more suitable inclusion and exclusion criteria for COPD patients' eligibility in COPD monitoring and novel exacerbation treatment clinical trials.

In stark contrast to the current one-size-fits-all SCS- and antibiotics-based treatment approach, the development of an objective definition of COPD disease states and the identification of inflammation and symptom-oriented temporal trajectory clusters for COPD patients may inspire new, personalised, treatment approaches. In Chapter 2, I will outline the methodology at the core of the analysis in my thesis. Chapter 3, 4 and 5 will address my thesis hypothesis and aims.

1.8 Hypotheses and aims of this thesis

My hypotheses are as follows:

1) Treatment outcome for patients hospitalised for their exacerbations can be predicted from data known at, or before, the time of exacerbation.
2) Distinct biological and symptomatological subgroups of COPD disease states and post-treatment temporal trajectories exist.

My thesis aims are to:

1) Identify variables associated with, and develop predictive models for, exacerbation treatment failure following hospitalisation (Chapter 3).
2) Quantitatively redefine exacerbations of COPD, taking into account biological and symptom-based heterogeneity (Chapter 4).
3) Define subgroups of trajectories from exacerbation to post-treatment resolution (Chapter 5).

In Chapter 6, the conclusions of this thesis will be discussed.

Chapter 2 – Methods

The mathematical background for key methods will be outlined here. Further mathematical approaches specific to subsequent chapters will be discussed in the corresponding chapters. In addition, a description of the clinical study design underlying the datasets analysed in this thesis, variables available and included at each stage of analysis and the division of data into training, validation and testing portions for machine learning classification analysis will be provided in thesis Results chapters. Code used to perform all machine learning analyses will be provided in the Appendix Code.

2.1 Overview of methods used in thesis

I outline the key framework and order of steps for data analysis throughout my thesis below:

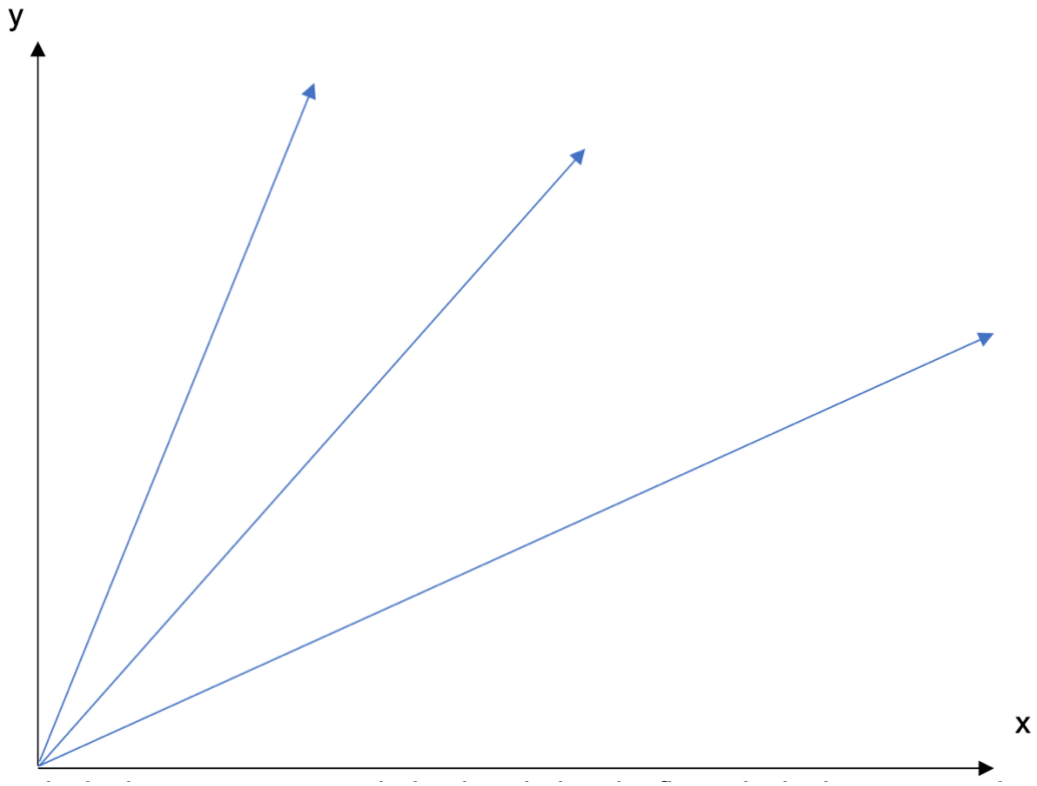
1) Summarise available variables for analysis.
2) Perform cleaning of dataset to ensure strange symbols or other erroneous values for variables are removed. This was performed both by eye, checking through variable columns in data files and by applying the 'fivenum' and 'table' functions from the coding language 'R'[86] to columns of the data files.
3) Hypothesis testing (in order to minimise bias, I performed hypothesis tests before data imputation).
4) Imputation of missing values.
5) Machine learning classification analysis e.g to predict exacerbation treatment failure **OR** dimensionality reduction followed by machine learning cluster analysis e.g. to characterise COPD heterogeneity.

This framework is applied in the context of different analyses depending on the objectives of each thesis chapter. The detailed analysis pipeline for each analysis is provided in corresponding chapters.

2.2 Principal component analysis

PCA[24] is a useful tool for achieving dimensionality reduction since it expresses a dataset of observations of inter-correlated dependent variables in terms of observations of new, orthogonal variables called principal components[87]. This concept is illustrated in Figure 2.1.

Figure 2.1 | Illustration of variables being representable by a smaller subset losing minimal or no information. The three blue variables can all represented by the two variables x and y via horizontal and vertical components.



The principal components are derived such that the first principal component has the greatest possible variance, the second principal component is orthogonal to the first and has the next largest possible variance and so on[87]. If desired, dimensionality reduction may be performed by selecting a number of principal components smaller than the number of original dependent variables which nevertheless explain a desired amount of cumulative variance of the data[87]. Let us consider the *I*×*J* matrix **X**, where *I* is the number of observations and *J* is the number of variables and *L* is the rank of **X** with[87]:

$$L \leq min\{I, J\}.$$

Each of the *J* variables may be centred to have zero mean and scaled to have unit variance (in order for the remaining PCA procedure not to be biased by virtue of differences in the original measurement units used for different variables)[87]. The singular value decomposition of **X** is given by[87]:

$$\mathbf{X} = \mathbf{P}\Sigma\mathbf{Q}^{\mathbf{T}}$$

where P is the *I*x*L* matrix of left singular vectors, Σ is the diagonal matrix of singular values and **Q** is the *J*x*L* matrix of right singular vectors[87]. This is equivalent to[87]:

$$\mathbf{X} = \mathbf{F}\Sigma\mathbf{Q}^{\mathbf{T}}$$

(**F = P)**

and **F** is therefore the *I*x*L* matrix of factor scores and **Q** is the *J*x*L* loading matrix[87]. In this thesis, PCA is performed with the 'prcomp' function using the package 'stats' which is part of 'R'[86].

In order to choose the number of principal components to keep when performing dimensionality reduction before cluster analysis, two approaches are used. Firstly, as a minimum the number of principal components which together account for at least 75% cumulative variance of the data are selected. Secondly, a scree plot is produced showing the percentage of variance accounted for by each principal component in order of decreasing explained variances i.e. principal component 1, followed by principal component 2, followed by principal component 3 etc[88]. The 'elbow' of the scree plot is identified as the point where there is a marked change in the gradient of the graph from steep to flat[91]. Only principal components prior to this elbow point are considered[87]. Scree plots are produced using the function 'fviz_eig' in the 'R' package 'factoextra'[89].

2.3 Machine learning cluster analysis

Hierarchical clustering is used in this thesis rather than a partitional method or a density-based method for a number of reasons:

1) Partitional methods require the number of clusters to be specified a priori[22,28]. In contrast, in hierarchical clustering a nested structure of clusters is produced without a priori cluster number specification[22,28].
2) In density-based clustering, core cluster points are defined as those with at least *MinPts* points within a distance of *Eps*, with remaining data points being categorised as either border points or noise as described in Table 1.1 of Chapter 1[22,32]. The choice of values for these parameters is not apparent a priori, yet is highly influential in determining the number of clusters, the size of clusters and the structure of clusters[22,32].
3) The hierarchical nested cluster relationships are geometrically interpretable via a dendrogram[22,28].

In general and also in this thesis, for computational efficiency agglomerative hierarchical clustering (in which one begins with each data point as a separate cluster and ends with a single cluster containing all data points) is used instead of divisive hierarchical clustering (in which one begins with a single cluster containing all data points and ends with each data point as a separate cluster)[22].

As shown in Table 2.1, agglomerative hierarchical clustering involves the following steps[22,28]:

1) Each data point represents an initial cluster.
2) A distance metric is used to determine proximity of data points to one another.
3) Clusters which are most similar based on a particular criterion to interpret the distance matrix are merged with each other.
4) This step is repeated iteratively until only one cluster remains.

In this thesis, the Euclidean distance metric is used since it is intuitive and enables a straightforward interpretation of the cluster results in terms of the original raw data[22]. The Euclidean distance d_{ij} between two points x_i and x_j is defined as[22]:

$$d_{ij} = [\sum_{k=1}^{p}(x_{ik} - x_{jk})^2]^{1/2}$$

Where the kth variable value of the points x_i and x_j in $\mathbf{E}^p$ is represented by x_{ik} and x_{jk} respectively.

Besides the distance metric between points, there are different ways to define distance between clusters. The single linkage[90] and complete linkage[91] methods (in which distance between clusters is defined as the distance between the closest and most distant pair of points respectively, each point in the pair belonging to another of the two

clusters in consideration) are not used in this thesis since they fail to take into account cluster structure[22]. Other methods which take into account cluster structure are summarised in Table 2.1 below.

Table 2.1 | Methods of defining inter-cluster distances to decide cluster merging during agglomerative hierarchical procedure.

Inter-cluster Distance Measure/Method	Description
Average Linkage[92]	Average distance between every possible pair of data points, each such pair consisting of one data point from each of the two clusters under consideration.
Weighted Average Linkage[93]	As with average linkage but inter-cluster distances are weighted inversely to the number of data points in a given cluster.
Ward's Method[31]	At each iteration of hierarchical clustering, the total increase in the within-cluster sum of squares error is minimised.
Centroid Linkage[92]	Squared Euclidean distance between cluster centroids.
Median Linkage[94]	As with centroid linkage but with centroids of constituent clusters equally weighted.

A number of rules can be used to determine the optimal number of clusters[22,28]. In this thesis, the Duda and Hart Je(2)/Je(1) index[95] and the related pseudo-t^2 index[95] are used since:

1) the Duda and Hart Je(2)/Je(1) index has shown to be one of the most reliable cluster number determination rules in simulation studies[96];
2) pseudo-t^2 is directly related to the hierarchical clustering procedure[95].

The derivation of these indices and the critical values of the indices for determining optimal cluster number are described by Duda & Hart (1973)[95] and Gordon (1999)[97].

Clustering was performed and optimal cluster number evaluated using the 'R' packages 'stats'[86] and 'NbClust'[98].

2.4 Machine learning classification analysis

LASSO

The LASSO[38,41] approach for classification is based on logistic regression but with a penalty function such that the number of total variables is reduced and does not exceed the total number of data samples for training the algorithm.

In logistic regression, the probability p of the outcome of interest (e.g. treatment failure having occurred) is given by[38]:

$$p = \frac{1}{1+\exp\left[-\left(\beta_0+\beta_1 x_1+\cdots+\beta_p x_p\right)\right]}$$

where β_0 is the intercept term and β_n is the coefficient of the nth variable x_n where n = 1, 2,3…p for a total p variables. Using the binomial distribution as the probability distribution for two target classes (e.g. treatment failure versus treatment success), the binomial likelihood function is[38]:

$L(p) = (nCr)(p^r)(1-p)^{n-r}$

where r = the number of events with the outcome of interest, n is the total number of events and nCr represents 'n Choose r', the number of possible combinations of r events with the outcome of interest (e.g. treatment failure) out of n total events. In logistic regression, values of β_n are chosen so that the binomial likelihood function L(p) is maximsed[38].

In the LASSO procedure, the log likelihood together with a penalty term are maximised[38,41]:

$$\text{logL(p)} \; - \lambda \sum_{n=1}^{P} |\beta_n|$$

where $|\beta_n|$ is the absolute value of a coefficient β_n and λ is a tuning parameter which determines the degree of penalisation[38,41]. Cross validation is used to select the best value of λ for optimal classification results. To enhance reliability, the lowest number of variables are chosen such that the cross-validation performance is within 1 standard error of the maximum performance[99]. For a given value of λ, the number of variables with $|\beta_n| \geq 0$ is fixed. Hence, LASSO simultaneously performs feature selection and classification and is especially well suited to high dimensional low sample size contexts. When applying LASSO for two class classification in my thesis, I used the function 'cv.glmnet' from the package 'glmnet' in 'R' and specified alpha = 1 to ensure lasso classification instead of other forms of regularisation such as ridge classification which do not perform feature selection[86,99]. I also specified NULL for λ parameter so that the optimal sequence of λ would be automatically calculated. Full code is shown in the Appendix Code.

Feature selection via Statistically Equivalent Signatures algorithm

For any classification task in my thesis, for classification methods which do not perform automatic feature selection, I needed to use a method to choose a subset of variables from the total possible variables which are important in the multivariable context. Such a method should not be dependent on the classification algorithm used if the selected variables represent a reliable, generalisable set of variables. Furthermore, it is important to employ a feature selection method which indicates not only an optimal set of variables but also whether and which other potential variables are equivalent to each

variable in the selected list. This enables a superior insight into the meaning of selected variables in the context of the clinical problem as well as the flexibility in the choice of variables to be measured to inform a classification model in future clinical scenarios. These desired qualities are fulfilled by the statistically equivalent signatures (SES) algorithm[100]. This algorithm consists of the following steps[100]:

1) Begin with an empty set S representing selected variables.
2) Consider all possible variables V for selection.
3) Alternately:
    a. select the variable maximally associated with the target outcome T (e.g. treatment failure) conditioned on any possible subsets of the variables already in S.
    b. exclude from S any variable X which is no longer associated with the target outcome T (e.g. treatment failure) given any subset of other variables in S. If a variable X is excluded from S, the exclusion is permanent.
4) Before a variable is eliminated, each other variable Y statistically equivalent to X is identified by assessing whether pYT.Z′ > alpha when $Z' \leftarrow (Z \cup \{X\}) \setminus \{Y\}$.

    All such variables Y are noted as being statistically equivalent to X.
5) A list of selected variables associated with T in the context of other variables and all equivalent variables to each selected variable is returned by the algorithm.

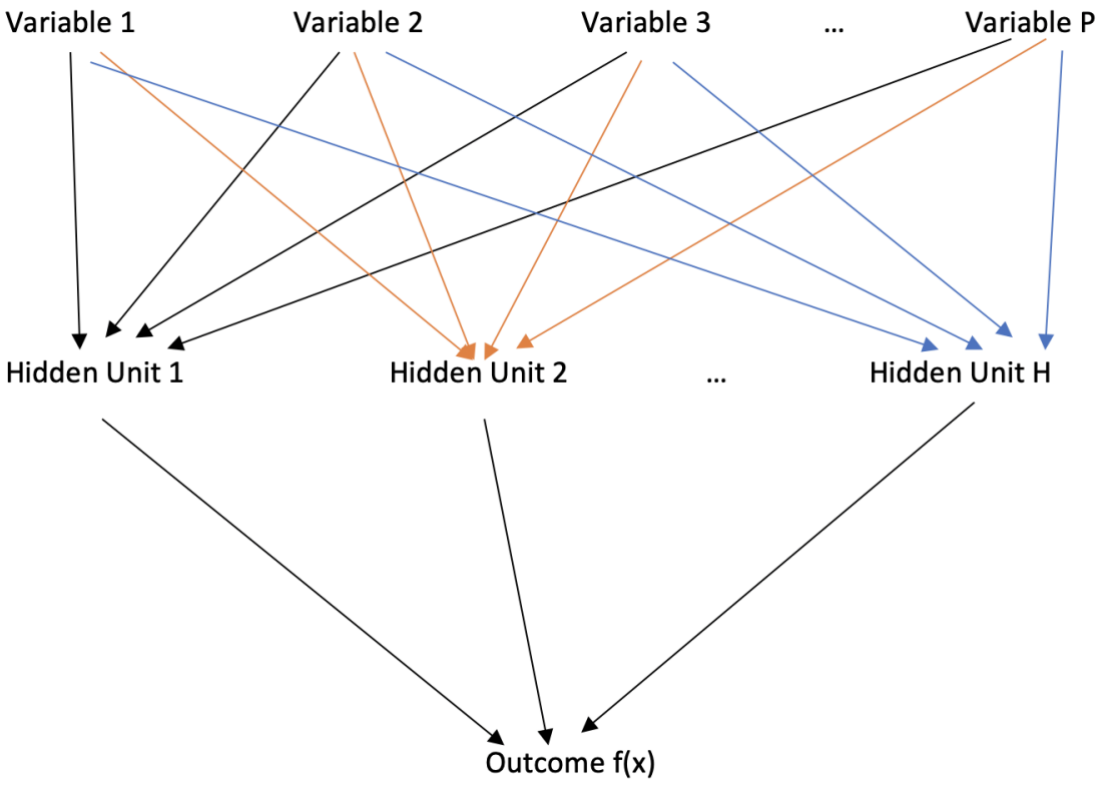
I used the threshold alpha < 0.1 for the conditional independence test probabilities to select variables associated with the target outcome T in the context of all other variables. Equivalent variables themselves have alpha ≥ 0.1 but nevertheless

equivalent to selected variables in the context of the elimination procedure outlines above. I used the logistic regression conditional independence test as suitable in the context of classification tasks with mixed continuous and categorical variables[100]. The logistic regression conditional independence approach for statistically equivalent signature algorithm is suitable in high dimensional low sample size contexts[100]. I used the 'R' package 'MXM' to run the SES algorithm[86, 100]. Full code is shown in thesis chapters in Appendix Code.

Neural Networks

A neural network consists of linear combinations of hidden units, each of which itself consist of a linear combination of variables similar to the linear combination of variables in logistic regression[38,43]. The output from different hidden layers is combined to produce the final output of the neural network for a sample[38,43]. The structure of a neural network is illustrated in Figure 2.2 below.

Figure 2.2 | Structure of a simple neural network.



In the case of an outcome for one particular class (e.g. treatment failure), the values for the intercept β_0 and the *P* other coefficients β_j of the P variables, as well as the intercept γ_0 and the H coefficients of the H hidden units are chosen in order to minimise the value of the following sum of squared errors[38,43]:

$$\sum_{i=1}^{n}(y_i - f_i(x))^2 + \lambda\sum_{k=1}^{H}\sum_{j=0}^{P}\beta_{j_k}^{\ 2} \quad + \quad \lambda\sum_{k=0}^{H}\gamma_k^2$$

where y_i is the true classification label of the training data point x_i, $f_i(x)$ is the predicted outcome of the neural network for the data point x_i and λ is a tuning parameter achieving 'weight decay'. The weight decay causes a reduction in variable and hidden unit coefficients to reduce the influence of coefficients with little impact[38,43]. This is an important procedure in order to counteract overfitting[38,43]. To ensure that for I classes

the outputs of $f_1(x)$, $f_2(x)$, $f_3(x)$,…, $f_l(x)$ sum to 1 so that f(x) is probability like, the following 'softmax' transformation is applied[38,43]:

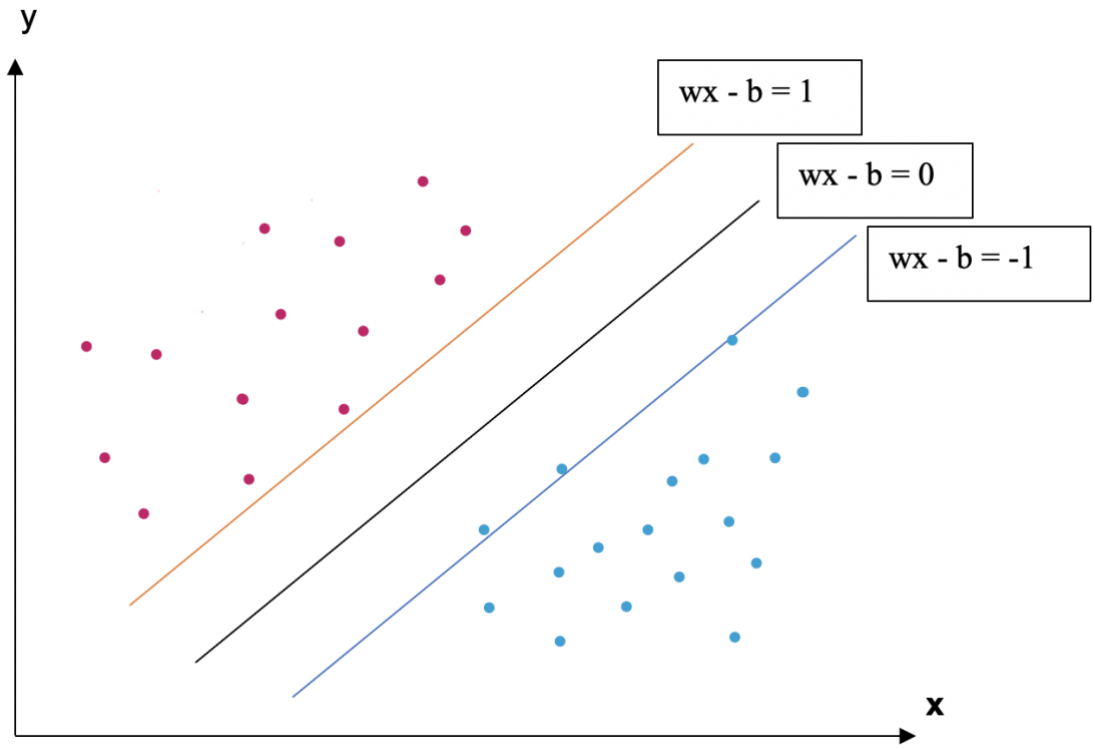
$$f_{il}^{*}(\mathbf{x}) = \frac{e^{f_{il}(x)}}{\sum_l e^{f_{il}(x)}}$$

where $f_{il}(x)$ is the neural network output probability that the ith sample belongs to the lth class. Minimising the sum of squared errors across classes as well as samples as above with the weight decay enables the final neural network variable and hidden layer coefficients and intercepts to be derived. When using neural networks in my thesis, I used the package 'caret' and 'nnet' in 'R'[86,101,102]. In keeping with recommendations and the cross-validation performance of my different neural network models in this thesis, I employed different numbers of hidden layers (size parameter ranging from 1 to 10 inclusive, in steps of 1) and different values of weight decay (λ ranging from 0 to 2, in steps of 0.1). Full code is shown in Appendix Code.

SVM

The SVM approach involves finding a hyperplane in n-dimensional space which separates the data points from both classes (e.g. treatment failure versus treatment success) as well as possible, with the data points most difficult to classify i.e. the data points closest to the decision boundary hyperplane being the most important in determining the hyperplane[38,42]. Consider Figure 2.3. For a decision boundary hyperplane wx – b = 0, any new data point which lies on wx – b = 1 or above will be assigned to one class (+1 e.g. representing treatment failure) whereas any data points lying on wx – b = -1 or below will be assigned to the other class (-1 e.g. treatment success).

Figure 2.3 | SVM hyperplane wx – b = 0 separating data points from two classes represented by pink and blue points.



For n variables, the values of the coefficients of the variables of the hyperplane wx – b = 0 are derived in order to minimise[38,42]:

$$C||\mathbf{w}||^2 + [\frac{1}{n}\sum_{i=1}^{n} max(0, 1 - y_i(\boldsymbol{w}^T\boldsymbol{x}_i - b))]$$

where C is a cost penalty which determines the size of the margin between the decision boundary hyperplane and the samples on the margin on either side of the decision boundary.

In order to enable non-linear decision boundaries, a non-linear transformation of space can be used so that even though the SVM hyperplane is linear relative to the transformed space, it is non-linear relative to the original space[38,42]. The radial basis function transformation is used in this thesis since it enables flexible non-linear

classification with the SVM approach[38,42]. I used the 'R' package 'kernlab' and 'caret' to perform SVM analysis[86,101,103]. Greater values of the cost penalty C enable increasing non-linear decision boundaries relative to the original space[38,42]. In keeping with recommendations and the cross-validation performance of my different SVM models in this thesis, I employed different values for C ($C= 2^x$, where x ranged from –4 to 4 inclusive in steps of 1) and different values of the inverse width parameter sigma (automatically calculated according to data). Full code is shown in Appendix Code.

KNN Algorithm

The KNN[39,40] methodology has been detailed during Introduction. In my thesis, I used different values of K as the number of nearest neighbours to use in a vote of which class a new data point should be assigned. I used the 'R' package 'caret' to perform KNN analysis[86,101.] I used values of K ranging from 1 to the total number of training samples using only odd numbers for K since even numbers would lead to a tie in the vote to decide classification. Full code is shown in Appendix Code.

Decision trees and random forest classification

In the case of classification where features or predictors are to be used to assign each sample to a particular class, a decision tree[104] consists of nested if-then statements which use the values of predictors to successively partition data into more homogenous subsets relating to a particular class. The Classification and Regression Trees (CART) decision tree methodology is as follows[38,104]:

1) Samples (data points) are sorted based on the values of predictors.

2) Split points are calculated as the midpoints of unique predictor values in the case of continuous predictors or, for categorical predictors, different ways of grouping/splitting categorical predictor values are considered.
3) Different split points are evaluated and the split point (for the best predictor) which minimises a tree node purity criterion is selected. For example, in a two class problem the Gini index to be minimised is defined as follows for a particular node:

   $p_1(1-p_1) + p_2(1-p_2)$

   where p_1 is the class 1 probability and p_2 is the class 2 probability. It can be seen that the Gini index is minimised when one of p_1 or p_2 are close to zero and the other is close to 1 i.e. when a tree node is 'pure' with respect to one class.
4) The splitting process continues iteratively within each of the new partitions and resulting new nodes of the tree.
5) A stopping criterion (e.g. maximum tree depth) terminates the process.

Trees which continue to be formed until maximum depth is reached are likely to overfit[38,104]. Therefore, a tuning parameter called the complexity parameter 'cp' can be used in which a penalty is associated with the number of terminal nodes in the tree[38,104]. I performed the 'R' package 'rpart' to perform classification decision tree analysis[86,105]. I used a tune length of 10 with the 10 values of cp calculated automatically by the algorithm according to the data. Full code is shown in Appendix Code.

The definition of a random forest is given by Breiman (2001)[44]:

"A random forest is a classifier consisting of a collection of tree-structured classifiers

$\{h(\mathbf{x},\Theta k\ ),\ k=1,\ ...\}$

where the

$\{\Theta k\}$

are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input **x**."

Typically and as done in this thesis, the trees comprising the random forest are Classification and Regression Trees (CART)[38,104]. In the case of Breiman's (2001) random forest, each classifier tree comprising the forest is constructed using a bootstrap sample T_K from the original training set T whereby on average the bootstrap sample leaves out approximately one third of training samples[38,44]. Thus, the classifier trees h(**x**, T_K) can vote to form the bagged random forest predictor[38,44]. As trees are grown, at any node where splitting is to be performed only a random subset of predictors may be considered in the splitting process[38,44]. The number of such predictors which may be randomly chosen is specified by the user[38,44]. The use of the bagging procedure improves the accuracy of random forest classifiers compared to simple, individual classification trees[38,44]. In addition, since only a random subset of predictors may be chosen for splitting at nodes, the trees comprising the random forest are decorrelated[38,44]. This enables random forests to achieve greater classification accuracy than that of individual trees in the ensemble[38,44]. Random forests are especially robust to outliers and noise[38,44]. It has been shown by Breiman (2001) that overfitting does not occur when more trees are added to the ensemble[44]. Other

attractive characteristics of random forests include that they are relatively insensitive to the number of randomly chosen subset predictor variables allowed for consideration for splitting nodes, with a recommendation that this parameter can in general be set to the square root of the total number of predictors in a classification problem[38,44]. In my thesis, I performed the 'R' package 'caret' and 'randomForest' to perform classification decision tree analysis[86,101,106]. I used different values 'ntree' for the number of trees in the random forest (ranging from 100 to 1,000, in steps of 100) and different values for the number of subset variables 'mtry' randomly chosen as candidates to be used for a split in a tree (ranging from 1 to the maximum number of candidate variables as selected by the statistically equivalent signatures algorithm).

2.5 Data pre-processing and imputation

Categorical variables with near zero variance were removed according to the default settings of the nearZeroVar function in the 'caret' package in 'R' in order to remove categorical variables which may serve as identifiers of individual participants rather than being informative with regard to the class label of interest in this thesis such as objective COPD state or exacerbation treatment outcome[38,86,101].

I used a random forest-based imputation method called 'missforest' for data imputation on the full set of patient/measurement results with data to be used in classification/clustering analysis[107]. This approach is suitable since it provides robust imputation without the need for user-specified and subjective assumptions regarding the data distributio[107]. Furthermore, this approach is also suitable for mixed continuous and categorical variables[107]. I used the 'missForest' package in 'R' to achieve imputation[108]. Full code is shown in Appendix Code.

Chapter 3 – Predicting emergency department exacerbation treatment outcomes

Abstract

Exacerbations of airways disease necessitating admission to the ED are associated with significant morbidity and increase the risk of mortality. Attempted treatment of exacerbations is not always successful, often leading to hospital re-admissions.

In this chapter, I first identify characteristics associated with the physician decision to prescribe SCS and with the physician decision to prescribe antibiotics in a cohort of exacerbating COPD and asthma patients. Next, I identify characteristics associated with and develop a tool to predict exacerbation treatment outcome at various follow-up time points using data available at the time of exacerbation. Treatment failure is defined in this chapter as the need for additional SCS and/or antibiotics and/or hospital re-admission within given time points from the initial ED visit. In addition, patient self-reported treatment outcome is explored. I show that prediction of exacerbation treatment outcome may be achieved via machine learning algorithms combining different predictors known at exacerbation.

3.1 Introduction

Exacerbations not only of COPD but also of asthma are disruptive to patients' lives, are associated with increased risk of future exacerbations, increased mortality risk and are a large economic burden; furthermore, each event may require hospital admission[6,109,110]. Asthma is defined in the Global Initiative for Asthma (GINA) as[111]:

"…a heterogeneous disease defined by the history of respiratory symptoms (e.g. wheeze, shortness of breath, chest tightness, and cough) that vary over time and in intensity, together with variable expiratory airflow limitation. Airflow limitation may later become persistent."

Airway hyper-responsiveness and chronic inflammation typically characterise asthma[111]. Whereas for COPD patients' post-bronchodilator $FEV_1/FVC < 0.7$, for asthmatic patients the FEV_1/FVC usually remains > 0.7, being either comparable to healthy patients in the case of intermittent or mild persistent asthma or reduced by up to about 5% or more in cases of moderate or severe persistent asthma[112,113]. However, it is acknowledged that both COPD and asthma are terms representing heterogeneous conditions and whose symptoms and diagnostic criteria may at times overlap[111]. Such 'asthma + COPD' features may be especially seen in older patients and smokers[111]. Furthermore, as is standard practice, often there is no access to lung function in ED from primary care. This makes it important to have tools for predicting exacerbation treatment outcomes effective for both asthma and COPD. Little is known about why some patients admitted to hospital succeed whereas others fail treatment. There is thus an urgent need to develop predictive tools to identify which patients at exacerbation will respond successfully to treatment and which patients will not. Such a tool could help identify patients at high risk of treatment failure who require closer

monitoring following their presentation to ED, as well as developing biomarker-driven treatment algorithms to optimise patient response to treatment. Since ED presentations are not treated by pulmonologists, a tool to help guide treatment of response in a personalised medicine way is needed for specialists and non-specialists.

In this chapter, my objectives are to:

1) Investigate factors, in patients who attend ED with an exacerbation of COPD or asthma, which may be associated with the physician SCS or antibiotic treatment prescription decision.
2) Characterise healthcare utilisation- and patient-defined treatment failure definitions for immediate, one month and three month follow-up time points.

I use a data-driven approach to train and assess various multivariable supervised learning algorithms to predict the physician prescription decision and which patients undergo treatment failure.

3.2 Methods

Participants

I analysed an existing dataset (AWARD) of a cohort of exacerbating COPD and asthma patients presenting to ED. Participant recruitment and data collection to create the dataset were previously performed by clinicians and nurses. Participants were recruited if they had a patient-reported primary or secondary care diagnosis of COPD or asthma and presented to the ED of the John Radcliffe Hospital, a large teaching hospital in Oxfordshire, with an exacerbation of COPD or asthma. Exacerbations of COPD and asthma were defined and treated as per GOLD[114] and BTS respectively[115]. Participants with a current history of active pulmonary tuberculosis or current primary malignancy or with any other clinically relevant lung disease judged to be the primary diagnosis were excluded. Any alternative causes for non-exacerbation related increase in symptoms were also excluded, including but not limited to pneumonia, pulmonary embolism, pneumothorax or primary ischemic event. All participants provided informed written consent and the study was approved by the North West London Research Ethics Committee (REC: 15/LO/2119).

Study Design and Measurements

Observational and routinely collected clinical data were prospectively collected by nurses and clinicians for a study duration of 12 months. On the day of exacerbation (D0), patient demographics, physiology, symptoms, quality of life, healthcare utilisation, blood data and medication use were measured. Table 3.1 summarises measured variables at D0. Participant symptoms, health status and quality of life were assessed using participant reported outcome measures. Physiological measurements were made and venous blood samples collected and analysed to characterise participants' inflammatory phenotype.

Table 3.1 | Summary of variables measured for patients at D0.

Demographics	Clinical and Questionnaires	Physiology	Biology	Medication
Gender	Previous respiratory diagnosis of patient	Presence of fever at exacerbation	Total blood leucocyte count	Whether the patient received antibiotics in the week preceding the exacerbation
Age	Age of patient at time of initial respiratory diagnosis	Blood oxygen saturation from pulse oximetry	Blood neutrophil count	Whether the patient received antibiotics at exacerbation
Ethnicity	Age of patient at onset of symptoms	Temperature	Blood eosinophil count	Whether the patient received SCS at exacerbation
Height	Allergen associated with atopy if patient has atopy	Systolic pressure	Blood basophil count	Whether the patient received SCS in the week preceding the exacerbation
Weight	Whether patient has nasal polyps	Diastolic pressure	Blood monocyte count	Whether the patient received intravenous hydrocortisone
Smoking status	Whether patient had previous nasal polyp surgery	Post-bronchodilator FEV_1	Blood lymphocyte count	Whether patient taking SABA
Pack Year History	Number of unscheduled primary care/accident and ED visits in the 12 months prior to the exacerbation	Post-bronchodilator FVC	Blood haemoglobin	Whether patient taking LABA
Current/Main occupation of patient	Number of hospital admissions in the 12 months prior to the exacerbation	Peak Flow	Blood platelet count	Whether patient taking LAMA
Working status of patient	Number of Intensive Therapy Unit admissions in the 12 months prior to the exacerbation	FeNO	Blood serum albumin	Whether patient taking ICS
	Presence and type of co-morbidity	Blood pCO_2	Blood CRP	Whether patient taking carbocystiene
	Number of days for which patient reported symptoms have been worse for in lead up to exacerbation	Heart rate	Microbiology	Whether patient taking mucolytic

	Whether the patient reported the exacerbation to be associated with increased dyspnoea	Blood pH	Blood glucose	Whether patient taking theophylline
	Whether the patient reported the exacerbation to be associated with increased sputum production	Respiratory Rate	Blood sodium	Whether patient on maintenance azithromycin
	Whether the patient reported the exacerbation to be associated with increased sputum purulence	eGFR	Blood potassium	Whether patient on maintenance prednisolone
	Whether the patient reported the exacerbation to be associated with increased cough		Blood urea	Whether patient taking monteleukast
	Whether the patient reported the exacerbation to be associated with increased wheeze		Blood creatinine	Whether patient taking anti-histamine
	EuroQol[116] mobility			Whether patient taking aspirin
	EuroQol[116] self-care			Whether patient taking nasal steroids
	EuroQol[116] usual activity			Nebuliser use
	EuroQol[116] pain/discomfort			Whether patient on Long Term Oxygen Therapy
	EuroQol[116] anxiety/depression			Whether patient taking proton pump inhibitor
	EuroQol[116] VAS score			Whether patient taking anti-diabetic Medication
	VAS[117] cough			Whether patient taking anti-hypertensive
	VAS[117] dyspnoea			Whether patient taking beta blockers
	VAS[117] sputum production			Whether patient taking bisphosphonates
	VAS[117] sputum purulence			Whether patient taking SAMA

	HAD[118] scale for depression			Whether oxygen was administered at exacerbation
	HAD[118] scale for anxiety			
	CAT score[119]			
	Medical research council dyspnoea scale[120]			
	Asthma Control Questionnaire[121]			
	Charlson Index Score[122]			
	GOLD stage			
	GINA stage			

Definition of abbreviations: CAT score = COPD Assessment Tool; CRP = C-reactive protein. eGFR = estimated glomerular filtration rate; FeNO: fractional exhaled nitric oxide. FEV_1 = Forced Expiratory Volume in 1 second; FVC = Forced Vital Capacity; GINA = Global Initiative for asthma; GOLD = Global Initiative for Chronic Obstructive Lung Disease; HAD = Hospital Anxiety and Depression scale; ICS = inhaled corticosteroids; LABA = Long-Acting Beta Agonist; LAMA = Long-Acting Muscarinic Antagonist; pCO_2 = partial pressure of carbon dioxide; PPI = Proton Pump Inhibitor; SABA = Short-Acting Beta Agonist; SAMA = Short-Acting Muscarinic Antagonist; SCS = systemic corticosteroids; VAS = Visual Analogue Scale.
It should be noted that the CAT score was collected for both asthma and COPD patients.
Variables highlighted grey were not analysed by me since for such variables data were either unavailable, unsuitable for analysis or there was concern associated with data collection quality as follows: **Current/main work occupation of patients** and **atopy data**: removed since there were a very large number of categories representing different unique qualitative inputs; **Asthma Control Questionnaire**: specificity to asthma, as well as concerns regarding quality of data collection; **Medical research council dyspnoea scale**: incorrect data entry made in study where the Medical Research Dyspnoea 1-5 scale was interchanged with the modified Medical Research Dyspnoea 0-4 scale. Consistency and accuracy of data input was not confirmed; **GINA score**: removed due to concerns regarding quality of data collection; **FEV_1, FVC, microbiology**, **FeNO** and **GOLD**: removed since data were ultimately not collected; **Whether patient was taking bisphosphonates**: not analysed further since none of the patients were taking bisphosphonates and so this variable was un-informative.
Quantitative dosage of medication were not analysed since there were very few unique values for dose, rendering the dosage variable un-informative.

For patients admitted to hospital directly from ED, physiology, blood data and symptoms were also measured at day 1 (D1) and day (D5) (Table 3.2).

Table 3.2 | Summary of variables measured for patients at D1/D5.

How the patient felt at time of exacerbation*	Blood urea
How the patient felt at a follow-up time point*	Blood creatinine
How the patient felt after treatment ~	Total blood leucocyte count
VAS cough	Blood neutrophil count
VAS dyspnoea	Blood eosinophil count
VAS sputum production	Blood basophil count
VAS sputum purulence	Blood monocyte count
CAT score	Blood lymphocyte count
Peak flow	Blood haemoglobin
Blood oxygen saturation	Blood platelet count
Blood glucose	Blood serum albumin
Blood sodium	Blood CRP
Blood potassium	eGFR

*Self-reported on a scale from 1 (worse possible) to 10 (best possible).
~ Patients provided qualitative responses indicating whether they felt better.
Definition of abbreviations: CAT score = COPD Assessment Tool score; CRP = C-reactive protein. eGFR = estimated glomerular filtration rate; VAS = Visual Analogue Scale.

Telephone consultations with patients at day 30 (D30) and day 90 (D90) enabled follow-up symptom data to be collected (Table 3.3).

Table 3.3 | Summary of variables measured for patients at D30/D90.

How the patient felt at a follow-up time point *
How the patient felt after treatment ~
VAS cough
VAS dyspnoea
VAS sputum production
VAS sputum purulence
CAT score

*Self-reported on a scale from 1 (worse possible) to 10 (best possible).
~ Patients provided qualitative responses indicating whether they felt better.
Definition of abbreviations: CAT score = COPD Assessment Tool score; VAS = Visual Analogue Scale.

Data cleaning and missing value imputation

I inspected the lower limit and upper limit of numerical variables and the different categories for binary/categorical variables to ensure that all values were sensical. Where non-sensical values or missing values occurred, I attempted to obtain the correct data via consultation with the study nurse and clinicians to retrieve data from patient electronic health records. For blood cell count data, if for a patient just one blood cell count variable was missing, I calculated the value by subtracting the sum of the individual cell counts for that patient from the total white blood cell count. If for a patient just the total white cell count was missing, I calculated its value by summing all of the individual blood cell counts for that patient. Where missing data still remained in the dataset, I performed imputation using the missForest algorithm as described and justified in Chapter 2, using the 'R' package 'missForest'[86,107,108].

Treatment failure definitions

I used 5 definitions of treatment failure:

D1/D5 hospitalisation: patients who were transferred from ED to hospital as determined by D1 and/or D5 follow-up.

D30 healthcare utilisation: additional SCS and/or antibiotics and/or hospital re-admission between D0 and D30. Hence, any patients with D30 healthcare utilisation treatment failure include patients with immediate hospital admission after D0 as well as patients with any other re-admission/additional major treatment up to D30.

D30 patient-defined treatment outcome: any patient with a D30 healthcare utilisation treatment failure or whose D30 self-reported symptoms are not better compared to D0 (based on how patient felt at D30).

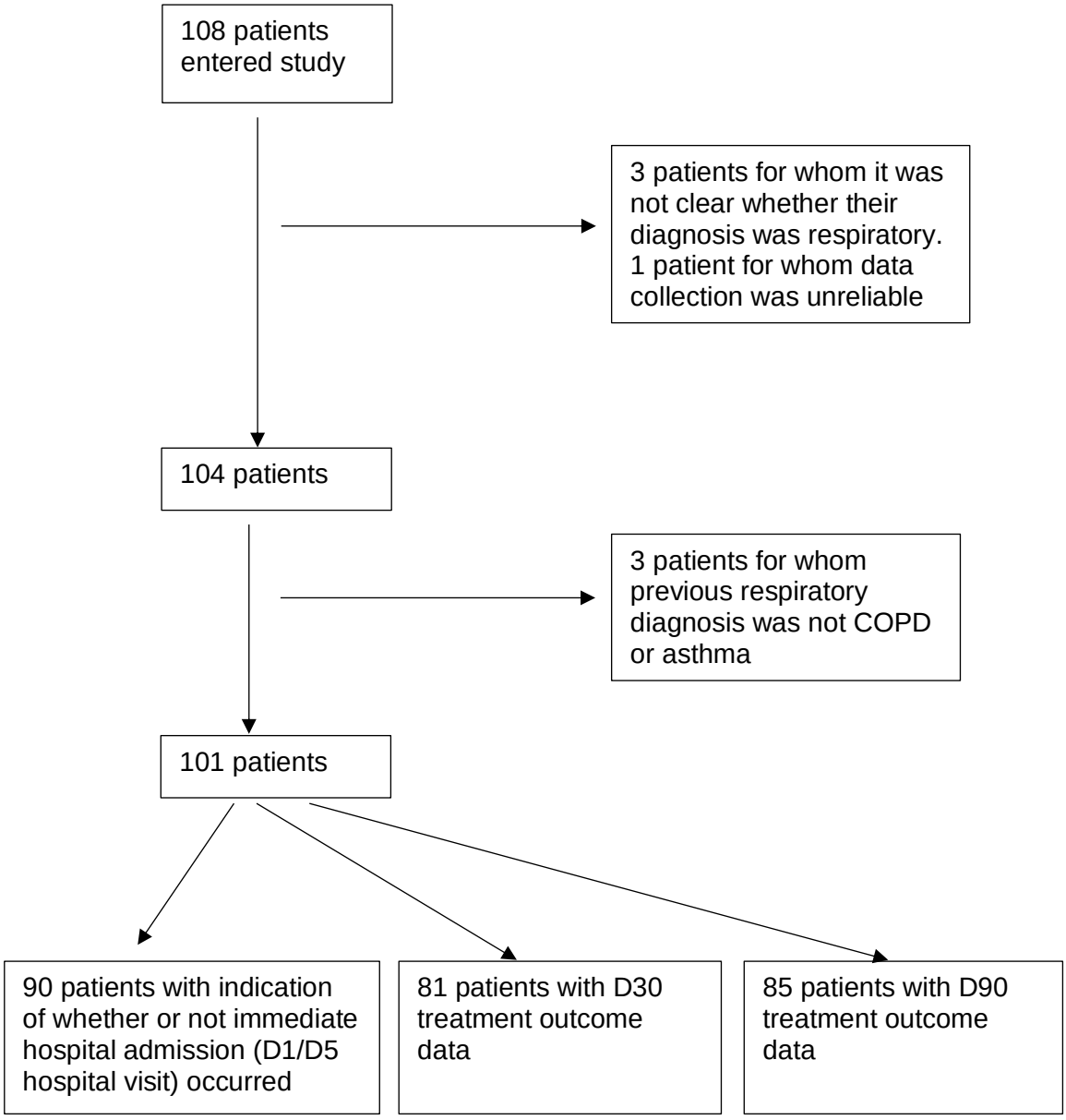
D90 healthcare utilisation: additional SCS and/or antibiotics and/or hospital re-admission between D0 and D90. Hence, any patients with D90 healthcare utilisation treatment failure include patients with D1/D5 hospital admission, patients with D30 healthcare utilisation treatment failure or patients with any other re-admission/additional major treatment up to D90.

D90 patient-defined treatment outcome: any patient with a D90 healthcare utilisation treatment failure or whose D90 self-reported symptoms are not better compared to D0 (based on how patient felt at D90). This definition was not analysed separately from the D90 healthcare utilisation definition since the number of treatment failures was the same.

Participant cohort data availability at different follow-up points

Out of 108 initial study participants, analysis was conducted on data from 101 participants to compare D0 characteristics of participants according to their exacerbation treatment. Immediate hospitalisation, D30 and D90 treatment outcome data were available for 90, 81 and 85 patients respectively (Figure 3.1).

Figure 3.1 | Diagram indicating participant cohort size and availability of treatment outcome data for different follow-up points.



It should be noted that some participants had D90 but not D30 treatment outcome data available since there may be an indication of hospitalisation and/or additional major treatment having occurred between D0 and D90 but not whether this occurred before or after D30. Although 81 and 85 participants had D30 and D90 treatment outcome data respectively, the number of participants with D30 and D90 patient-defined outcome data was 53 and 59 respectively.

Univariate statistical analysis

Univariate analyses were performed using 'R' and Prism Version 9 for macOS[86,123]. For the 101 participants analysed, numerical data were first assessed for normality using the Shapiro-Wilk test (Appendix Table A3.1).

I performed the following univariate analyses to compare available participant characteristics:

- D0 characteristics of participants receiving SCS and antibiotics versus SCS alone versus antibiotics alone versus neither treatment at D0;
- D0 characteristics of participants with D1/D5 hospital admission versus participants without D1/D5 hospital admission;
- Paired D0 versus D1 or D5 characteristics for participants with D1/D5 hospital admission;
- Paired D0 versus D30 characteristics for participants;
- D0 characteristics of participants with D30 healthcare utilisation treatment failure versus participants with successful D30 healthcare utilisation treatment outcome;
- Paired D0 versus D90 characteristics for participants;

- D0 characteristics of participants with D90 healthcare utilisation treatment failure versus participants with successful D90 healthcare utilisation treatment outcome.

Normally-distributed data are shown as mean (standard error). For non-normally distributed data, if a log transform of the variable is normally-distributed, data are shown as geometric mean (95% confidence interval). Otherwise, for non-normally distributed data, median (lower quartile to upper quartile) is used. For numerical variables, the t-test or wilcox test were used for comparing two groups depending on whether the data were normally (original or log-transformed data) or non-normally distributed respectively. When comparing more than two groups, either the one-way ANOVA test (and Tukey post hoc test) or the Kruskal-Wallis test (and post-hoc Kruskal Dunn test) was used, depending on whether the normality assumption was met. Bartlett's test for homoscedasticity was performed for variables with normal distributions or log-transformed variables with normal distributions. The chi-squared test was used for binary or categorical data. When comparing paired samples, the paired t-test (on original or log-transformed variable) or Wilcoxon signed-rank test (depending on normality as above) are used. Performing a hypothesis test multiple times simultaneously can result in a statistical multiplicity problem whereby there is an increased chance of false discoveries. For example, if $p<0.05$ is used as a threshold for statistical significance, performing a given hypothesis test for 100 different variables would mean that a false discovery (i.e. null hypothesis rejected although the null hypothesis is in fact true) would be expected for 5 out of the 100 variables. However, it should be noted that the aforementioned univariate analyses (e.g. participant characteristic comparison for different types of exacerbation treatment and different treatment outcomes) are hypothesis-generating rather than being used to make a

conclusive decision in relation to a particular variable truly being associated with exacerbation treatment type and treatment outcomes. In this hypothesis generation context, performing multiple comparison corrections could jeopardise the identification of variables potentially associated with treatment prescription decision and treatment outcomes. Therefore, a multiple comparison correction was not applied in this context. Furthermore, a distinction should be made between variables which meet the $p<0.05$ statistical significance threshold and variables which are clinically significant. A variable may statistically differ between two groups (e.g. patients receiving different exacerbation treatment types or between patients with different treatment outcomes) but without reflecting true biological meaning and clinical relevance. Hence, when interpreting results of this chapter in the Discussion section, the focus will be on variables which are both statistically and potentially also clinically significant.

In order to ascertain whether bias may be introduced into data analysis due to missing follow-up treatment outcome data, I compared the characteristics of the participants with and without treatment outcome data for the treatment outcome time point with the most missing follow-up data (81 participants with available D30 treatment outcome versus 20 participants with missing D30 treatment outcome). In this context where the emphasis was on confirmation that bias was not present due to missing follow-up data rather than a focus on hypothesis-generation, the Benjamini-Hochberg multiple comparison correction was applied. My analysis showed that none of the characteristics differed between the 81 patients with and the 20 patients without D30 treatment outcome (Appendix Table A3.2).

Supervised machine learning analyses to predict treatment outcomes

In this study, I developed a variety of supervised machine learning models to predict the different aforementioned treatment failure outcomes, the mathematics of which I have outlined in Chapter 2:

- LASSO, a l1-penalty regularised logistic regression model which simultaneously performs feature selection and treatment outcome classification prediction.
- Multivariable feature selection using forward-backward filter approach implemented by the Statistically Equivalent Signature (SES) algorithm outlined in Chapter 2[100]. I used a $p<0.1$ threshold rather than $p<0.05$ threshold for selecting features with multivariable importance in order to capture a greater number of potentially useful predictors. After identifying the multivariable combinations (with or without alternative statistically equivalent feature combinations), I selected the variables with a SES $p<0.1$ for use by a variety of non-linear machine learning algorithms:
    - Decision tree
    - Random forest
    - KNN
    - SVM with radial basis function
    - Neural network

As a first step when developing treatment outcome prediction algorithms, I used the 'R' package 'caret' to separate the participant data (number of participants depending on type of treatment outcome as described above) into a hold-out test set including 20% of participants and a training and validation portion containing 80% participants[86,101]. I next divided the 80% portion into four folds as part of four-fold cross-validation. In this procedure, each machine learning algorithm is trained on three of the

folds and validated on the fourth. This is repeated using each of the possible four folds as the validation portion and the average AUC across all four folds is then calculated for the algorithm. The cross-validation performance of a given machine learning algorithm type is used to select the best hyper-parameter values for each algorithm. Finally, I identified the best machine learning algorithm with optimal hyper-parameters, retrained this model on the entire 80% training and validation portion and then assessed the retrained model on the 20% hold-out test set. The hold-out test set AUC of this model provides a generalisable estimate of the ability of the machine learning model to predict treatment outcome for new patients. For each treatment outcome prediction analysis, I produced ROC curve figures for the best model on the hold-out test set using the 'R' package 'MLeval'[86,124]. When running code, I always set the programme 'seed' to a constant value in order to ensure consistent sample selection and shuffling when selecting participants for training and validation versus testing and producing different cross-validation folds. This is important for fair comparison of performance of different machine learning algorithms and for reproducibility of results.

Code

My code for multivariable feature selection and all of the above supervised machine learning algorithm training, validation (including different hyperparameters for each algorithm) and testing is provided in Appendix Chapter 3 Code.

3.3 Results

I analysed data from 101 participants. Normality tests were performed for all numerical variables analysed (Appendix Table A3.1).

Hospitalised COPD patients were older and had greater smoking history than asthma patients but did not differ in symptoms

A diagnosis of asthma and COPD was found in 79 (78%) and 22 (22%) participants respectively (Table 3.4). The participants with COPD were older (median 67 versus 42 years; $p<0.01$) and at exacerbation had a lower % oxygen saturation (median oxygen saturation in COPD 93% versus median in asthma 96%, $p<0.01$). COPD participants had a greater pack year history than asthma patients (median 34 pack years versus 0 pack years). A greater proportion of COPD than asthma participants were male (64% versus 32%). At exacerbation, there was no difference in VAS symptoms of cough, dyspnoea, sputum production or purulence between asthma or COPD at the time of exacerbation nor peripheral blood counts.

Table 3.4 | Characteristics of the asthma and COPD participants at exacerbation.

Characteristic	Primary Respiratory Diagnosis		P-value
	Asthma (n = 79)	COPD (n = 22)	
Male, n (%)	**25 (32)**	**14 (64)**	**<0.01**
Current smokers, n (%)	**17 (22)**	**6 (27)**	**<0.01**
Ex-smokers, n (%)	**26 (33)**	**16 (73)**	
Never smoker, n (%)	**36 (46)**	**0 (0)**	
Number taking ICS, n (%)	**40 (61)**	**7 (32)**	**0.04**
Increased wheeze at exacerbation, n (%)	68 (86)	19 (86)	1.00
Age (years)	**42 (28 – 57)**	**67 (60 – 72)**	**<0.01**
Pack year history	**0 (0 – 8)**	**34 (16 – 57)**	**<0.01**
Number of hospital admissions in previous 12 months	0 (0-1)	0 (0-1)	0.80
Oxygen saturation (%)	**96 (95 – 98)**	**93 (91 – 95)**	**<0.01**
VAS cough (mm)	55 (25-71)	42 (25-53)	0.26
VAS dyspnoea (mm)	69 (48-83)	76 (44-97)	0.20
VAS sputum production (mm)	22 (2-50)	21 (1-41)	0.75
VAS sputum purulence (mm)	10 (0-48)	23 (0-53)	0.35
Total leucocyte count, (x10^{9}cells/L)*	10.2 (9.5-11.0)	12.0 (10.4 -14.0)	0.18
Blood neutrophils (x10^{9}cells/L) *	7.1 (6.5-7.9)	8.6 (7.1-10.4)	0.15
Blood eosinophils (x10^{9}cells/L)	0.19 (0.04-0.41)	0.12 (0.03-0.31)	0.66
CRP (mg/L)	6.1 (0.9-21.0)	17.3 (1.8-58.1)	0.06

Definition of abbreviations: CRP = C-reactive protein; ICS = inhaled corticosteroids; VAS = Visual Analogue Scale.
Variable presentation and hypothesis tests: all numerical variables are presented as median (lower quartile to upper quartile) (wilcox test used for asthma versus COPD comparison) unless marked with * in which case the variable is presented as geometric mean (95% confidence interval) (t-test on log-transformed variable for asthma versus COPD comparison). Binary and categorical variables are presented as number (%) of instances (chi-squared test used for asthma versus COPD comparison).
Variables in bold font differ significantly between asthma and COPD.
Correction for multiple comparisons is not made since the analysis is hypothesis-generating.

Antibiotic prescription at exacerbation was associated with sputum production and purulence, while eosinophilic inflammation was associated with patients receiving SCS alone

Treatment allocation showed that at the time of an exacerbation 39 participants (39%) received both SCS and antibiotics; 42 participants (42%) received SCS alone; 7 participants (7%) received antibiotics alone; and 13 participants (13%) received neither. For the 101 participants, a comparison of characteristics between participants

prescribed both SCS and antibiotic, SCS alone, antibiotics alone, or neither treatment for the exacerbation at D0 are summarised in Table 3.5. VAS sputum production and purulence differed between treatment allocations (p=0.03 and p=0.04 respectively). Participants prescribed antibiotics alone had higher VAS sputum production than participants not prescribed antibiotics and participants prescribed both antibiotics and SCS than participants prescribed SCS alone. VAS sputum purulence was greater for patients prescribed both SCS and antibiotics than for participants prescribed SCS alone. Heart rate and diastolic pressure also differed across treatment allocations ($p<0.05$ and p=0.02 respectively). In particular, participants prescribed both SCS and antibiotics had higher heart rate than those prescribed antibiotics alone. Participants prescribed antibiotics alone had lower diastolic pressure than those prescribed either SCS alone or neither treatment. Neutrophilic and eosinophilic indices (both cell counts and percentage) differed between treatment allocations, as did lymphocyte percentage. For example, neutrophil cell count and percentage were higher and lymphocyte percentage lower in participants given SCS and antibiotics compared to participants given SCS alone. Eosinophilic cells and percentage were greater for participants given SCS alone compared to participants given antibiotics (with or without SCS). Participants given antibiotics alone at D0 were more likely than participants with other treatment allocations to be on LAMA and participants given antibiotics at D0 (with or without SCS) are more likely than the other treatment allocations to have received antibiotics during the week preceding D0.

Table 3.5 | Comparison of D0 characteristics of 101 participants according to type of exacerbation treatment (SCS and antibiotics, SCS alone, antibiotics alone or neither treatment).

Variable		SCS and antibiotics, (n=39)	SCS alone, (n=42)	Antibiotics alone, (n=7)	Neither treatment, (n=13)	P-Value
Male, n (%)		14 (36)	16 (38)	3 (43)	6 (46)	0.92
Age (years)		55 (36 – 68)	35 (27 – 63)	54 (46 – 76)	47 (33 – 58)	0.08
Weight (kg)*		84 (76 – 93)	79 (72 – 87)	85 (67 – 107)	84 (70 – 102)	0.77
Pack Year History		8 (0 – 24)	0 (0 – 14)	0 (0 – 28)	0 (0 – 27)	0.11
Smoking Status	Current smoker, n (%)	10 (26)	8 (19)	1 (14)	4 (31)	0.11
	Ex-smoker, n (%)	22 (56)	14 (33)	3 (43)	3 (23)	
	Non-smoker, n (%)	7 (18)	20 (48)	3 (43)	6 (46)	
Previous Respiratory Diagnosis	COPD, n (%)	11 (28)	7 (17)	1 (14)	3 (23)	0.61
	Asthma, n (%)	28 (72)	35 (83)	6 (86)	10 (77)	
Age at diagnosis (years)		29 (9 – 52)	17 (5 – 51)	15 (5 – 48)	39 (15 – 51)	0.48
Age at onset of symptoms (years)		20 (9 – 49)	16 (4 – 50)	15 (4 – 42)	40 (13 – 53)	0.49
Comorbidity	Cardiovascular, n (%)	9 (26)	2 (5)	1 (14)	4 (33)	0.25
	Other, n (%)	15 (43)	20 (54)	4 (57)	4 (33)	
	None, n (%)	11 (31)	15 (41)	2 (29)	4 (33)	
Caucasian, n (%)		35 (95)	37 (93)	6 (86)	11 (85)	0.65
Working Status	Full-time, n (%)	18 (49)	23 (56)	2 (29)	5 (50)	0.58
	Part-time, n (%)	5 (14)	7 (17)	1 (14)	1 (10)	
	Retired, n (%)	11 (30)	7 (17)	4 (57)	2 (20)	
	Unemployed, n (%)	3 (8)	4 (10)	0 (0)	2 (20)	
Nasal polyps, n (%)		0 (0)	6 (14)	0 (0)	1 (8)	0.07
Previous nasal polyp surgery, n (%)		0 (0)	2 (5)	0 (0)	0 (0)	0.41
Number of days for which patient reported symptoms have been worse for in lead up to exacerbation		3 (1 – 6)	4 (2 – 7)	3 (2 – 5)	3 (1 – 5)	0.29
VAS cough (mm)		54 (30 – 72)	39 (17 – 72)	66 (50 – 78)	53 (26 – 68)	0.40

VAS dyspnoea (mm)		74 (49 – 85)	70 (32 – 81)	67 (61 – 89)	71 (11 – 89)	0.97
VAS sputum production (mm)		**28 (5 – 53)**	**14 (0 – 31)**	**31 (24 – 75)**	**11 (0 – 33)**	**0.03**
VAS sputum purulence (mm)		**32 (3 – 66)**	**4 (0 – 19)**	**31 (0 – 45)**	**9 (0 – 42)**	**0.04**
Increased cough, n (%)		32 (82)	33 (79)	7 (100)	11 (85)	0.58
Increased dyspnoea, n (%)		**39 (100)**	**42 (100)**	**6 (86)**	**13 (100)**	**<0.01**
Increased sputum production, n (%)		20 (51)	19 (45)	6 (86)	6 (46)	0.26
Increased sputum purulence, n (%)		**22 (56)**	**10 (24)**	**4 (57)**	**2 (15)**	**<0.01**
Increased wheeze, n (%)		36 (92)	36 (86)	5 (71)	10 (77)	0.33
Fever, n (%)		24 (62)	17 (40)	4 (57)	3 (23)	0.06
EuroQol VAS Score		35 (10 – 50)	37 (19 – 50)	30 (5 – 35)	35 (22 – 58)	0.72
EuroQol mobility	1, n (%)	14 (36)	15 (36)	1 (14)	5 (42)	0.11
	2, n (%)	18 (46)	26 (62)	4 (57)	7 (58)	
	3, n (%)	7 (18)	1 (2)	2 (29)	0 (0)	
EuroQol self care	1, n (%)	29 (74)	30 (71)	4 (57)	10 (83)	0.90
	2, n (%)	9 (23)	11 (26)	3 (43)	2 (17)	
	3, n (%)	1 (3)	1 (3)	0 (0)	0 (0)	
EuroQol usual activity	1, n (%)	10 (26)	11 (26)	1 (14)	6 (50)	0.35
	2, n (%)	14 (36)	20 (48)	3 (43)	5 (42)	
	3, n (%)	15 (38)	11 (26)	3 (43)	1 (8)	
EuroQol pain discomfort	1, n (%)	10 (26)	13 (32)	1 (14)	5 (42)	0.74
	2, n (%)	27 (69)	27 (66)	6 (86)	6 (50)	
	3, n (%)	2 (5)	1 (2)	0 (0)	1 (8)	
EuroQol anxiety depression	1, n (%)	22 (56)	26 (62)	3 (43)	8 (67)	0.24
	2, n (%)	10 (26)	15 (36)	2 (29)	3 (25)	
	3, n (%)	7 (18)	1 (2)	2 (29)	1 (8)	
HAD anxiety		8 (4 – 10)	6 (3 – 9)	7 (4 – 10)	5 (2 – 7)	0.36
HAD depression		7 (1 – 9)	3 (1 -5)	7 (2 – 10)	4 (2 – 7)	0.13
CAT score		16 (0 – 30)	18 (0 – 28)	20 (0 – 25)	19 (0 – 26)	1.00

Charlson Index Score	0 (0 – 1)	0 (0 – 0)	0 (0 – 3)	1 (0 – 2)	0.36
Number of unscheduled primary care and emergency department visits in previous 12 months	1 (0 – 3)	0 (1 – 4)	1 (0 – 2)	1 (0 – 1)	0.52
Number Of Hospital Admissions In Previous 12 Months	0 (0 – 2)	0 (0 – 1)	0 (0 – 1)	0 (0 – 5)	0.85
Number Of Intensive Therapy Unit Admissions In Previous 12 Months	0 (0 – 0)	0 (0 – 0)	0 (0 – 2)	0 (0 – 0)	0.14
Oxygen saturation (%)	95 (92 – 97)	96 (94 – 98)	96 (93 – 99)	96 (94 – 99)	0.22
Blood pH	**7.4 (7.4 – 7.4)**	**7.4 (7.4 – 7.4)**	**7.5 (7.4 – 7.5)**	**7.4 (7.4 – 7.5)**	**0.03**
Blood pCO2 (kPa)*	5.4 (4.9 – 5.9)	5.3 (5.0 – 5.7)	4.5 (3.6 – 5.6)	5.2 (4.5 – 6.0)	0.30
Heart Rate (beats/minute)~	**102 (3)**	**99 (3)**	**82 (6)**	**96 (6)**	**<0.05**
Temperature (°C)	36.8 (36.2 – 37.5)	36.4 (36.0 – 37.0)	36.5 (35.8 – 37.2)	36.8 (36.4 – 36.9)	0.21
Systolic (mmHg)*	134 (128 – 140)	138 (132 – 145)	123 (111 – 127)	139 (121 – 158)	0.31
Diastolic (mmHg)~	**79 (3)**	**81 (2)**	**64 (3)**	**86 (5)**	**0.02**
Respiratory Rate (breaths per minute)*	21 (19 – 22)	20 (18 – 22)	20 (16 – 25)	20 (18 – 23)	0.96
Peak Flow (litres/minute)*	199 (171 – 231)	238 (205 – 276)	180 (132 – 247)	224 (145 – 347)	0.27
Total leucocyte count (x10⁹cells/L)*	11.0 (10.0 – 12.2)	9.9 (8.8 – 11.1)	12.3 (9.2 – 16.2)	10.9 (8.7 – 13.7)	0.29
Blood neutrophil count (x10⁹cells/L)*	**8.4 (7.4 – 9.5)**	**6.3 (5.5 – 7.3)**	**8.8 (6.1 – 12.6)**	**8.1 (6.3 – 10.4)**	**0.02**
Blood eosinophil count (x10⁹cells/L)	**0.12 (0.02 – 0.36)**	**0.32 (0.13 – 0.53)**	**0.03 (0.01 – 0.12)**	**0.10 (0.01 – 0.15)**	**<0.01**
Blood basophil count (x10⁹cells/L)	0.05 (0.04 - 0.07)	0.05 (0.03 - 0.08)	0.03 (0.03 – 0.06)	0.05 (0.02 -0.08)	0.82
Blood monocyte count (x10⁹cells/L)	0.75 (0.48 – 1.01)	0.76 (0.59 – 0.87)	0.98 (0.53 – 1.71)	0.80 (0.46 – 1.12)	0.41
Blood lymphocyte count (x10⁹cells/L)*	1.3 (1.0 – 1.6)	1.8 (1.5 – 2.2)	1.6 (0.7 – 3.3)	1.6 (1.1 – 2.2)	0.08
Blood neutrophil percentage~	**77.0 (1.9)**	**65.4 (55.9 – 75.9)**	**72.7 (4.6)**	**74.8 (3.3)**	**<0.01**
Blood eosinophil percentage	**1.2 (0.3 – 3.4)**	**3.2 (1.4 – 5.6)**	**0.3 (0.1 – 1.1)**	**0.8 (0.2 – 1.2)**	**<0.01**
Blood basophil percentage*	0.4 (0.3 – 0.5)	0.5 (0.4 – 0.6)	0.3 (0.2 – 0.4)	0.4 (0.3 – 0.6)	0.36
Blood monocyte percentage	6.5 (4.5 – 7.7)	7.4 (6.6 – 8.8)	7.9 (5.8 – 8.5)	6.4 (4.4 – 10.3)	0.13
Blood lymphocyte percentage	**11.8 (7.2 – 15.6)**	**20.7 (13.5 – 26.8)**	**15.1 (8.8 – 27.6)**	**16.1 (8.8 – 24.4)**	**<0.01**
Haemoglobin (g/L)~	140 (2)	144 (2)	143 (4)	150 (3)	0.09
Platelets (x10⁹cells/L)	257 (217 – 315)	276 (228 – 364)	315 (252 – 374)	304 (259 – 376)	0.33
Sodium (mmol/L)	138 (136 – 140)	139 (138 – 141)	140 (134 – 142)	140 (137 – 141)	0.16
Potassium (mmol/L)	3.8 (3.5 – 4.2)	3.8 (3.5 – 4.0)	3.4 (3.1 – 3.8)	3.8 (3.6 – 4.3)	0.29
Urea (mmol/L)	4.9 (3.8 – 6.4)	4.3 (3.3 – 5.9)	4.6 (3.9 – 6.9)	4.6 (3.7 – 5.7)	0.61

Creatinine (µmol/L)	68 (59 – 86)	66 (59 – 79)	70 (58 – 79)	64 (55 – 73)	0.75
eGFR (mL/min)	90 (80 – 90)	90 (80 – 90)	72 (69 – 90)	90 (82 – 90)	0.30
Blood glucose (mmol/L)	7.2 (6.1 – 8.6)	6.6 (5.7 – 7.2)	7.1 (5.7 – 8.8)	5.9 (5.6 – 8.8)	0.17
Blood serum albumin (g/L)	**35 (33 – 38)**	**38 (36 – 39)**	**38 (37 – 39)**	**38 (34 – 39)**	**0.03**
CRP (mg/L)	16.8 (3.7 – 42.6)	4.2 (0.5 – 16.9)	7.2 (0.1 – 19.8)	3.4 (0.5 – 18.5)	0.06
Oxygen administered at exacerbation, n (%)	16 (42)	8 (21)	2 (33)	1 (8)	0.06
Oral steroids in the last week, n (%)	22 (56)	20 (48)	5 (71)	10 (77)	0.24
Antibiotics in the last week, n (%)	**22 (56)**	**10 (24)**	**3 (43)**	**3 (23)**	**0.01**
Intravenous hydrocortisone, n (%)	8 (21)	4 (10)	2 (29)	3 (23)	0.39
SABA, n (%)	31 (79)	36 (86)	6 (86)	10 (77)	0.84
LABA, n (%)	3 (8)	2 (5)	1 (14)	0 (0)	0.57
LAMA, n (%)	**10 (26)**	**5 (12)**	**4 (57)**	**6 (46)**	**0.01**
SAMA, n (%)	1 (3)	1 (2)	0 (0)	0 (0)	0.92
ICS, n (%)	15 (45)	23 (61)	4 (67)	5 (45)	0.51
Carbocystiene, n (%)	3 (8)	1 (2)	0 (0)	2 (15)	0.30
Maintenance prednisolone, n (%)	3 (8)	2 (5)	0 (0)	2 (15)	0.51
Monteleukast, n (%)	5 (13)	8 (19)	1 (14)	3 (23)	0.80
Antihistamine, n (%)	3 (8)	5 (12)	0 (0)	0 (0)	0.45
PPI, n (%)	**6 (15)**	**0 (0)**	**2 (29)**	**3 (23)**	**0.02**
Antihypertensive, n (%)	7 (18)	1 (2)	1 (14)	2 (15)	0.14
Theophylline, n (%)	**0 (0)**	**1 (2)**	**0 (0)**	**2 (15)**	**0.04**
Maintenance azithromycin, n (%)	1 (3)	0 (0)	0 (0)	0 (0)	0.66
Aspirin, n (%)	2 (5)	1 (2)	0 (0)	1 (8)	0.76
Nasal steroids use, n (%)	1 (3)	1 (2)	0 (0)	1 (8)	0.73
Nebuliser use, n (%)	2 (5)	0 (0)	0 (0)	1 (8)	0.37
Long Term Oxygen Therapy, n (%)	1 (3)	0 (0)	0 (0)	0 (0)	0.66
Anti-diabetic medication, n (%)	3 (8)	0 (0)	0 (0)	0 (0)	0.18
Beta-blockers, n (%)	1 (3)	0 (0)	1 (14)	1 (8)	0.14

Definition of abbreviations: CAT score = COPD Assessment Tool; CRP = C-reactive protein; eGFR = estimated glomerular filtration rate; HAD = Hospital Anxiety and Depression scale; ICS = inhaled corticosteroids; LABA = Long-Acting Beta Agonist; LAMA = Long-Acting Muscarinic Antagonist; PPI = Proton Pump Inhibitor; SABA = Short-Acting Beta Agonist; SAMA = Short-Acting Muscarinic Antagonist; SCS = systemic corticosteroids; VAS = Visual Analogue Scale.

Variable presentation and hypothesis tests: For numerical variables following a normal distribution and where the homoscedasticity assumption is fulfilled, ANOVA is used. These variables are indicated by ~. For variables whose log-transformed version follow a normal distribution and homoscedasticity is met, ANOVA is performed on the log-transformed data

and indicated by *. For other variables, the Kruskal-Wallis test is used. Variables following a normal distribution are presented as mean (standard error of the mean); log-transformed variables following a normal distribution are presented as geometric mean (95% confidence interval); other variables are presented as median (lower quartile to upper quartile). Binary and categorical variables are presented as number (%) of instances (chi-squared test used for treatment type comparison).
Variables in bold font differ significantly between exacerbation treatment allocation type.
Correction for multiple comparisons is not made since the analysis is hypothesis-generating.

Classifier models using variables including blood eosinophil percentage and heart rate and classifier models using variables including exacerbation sputum purulence and blood neutrophil percentage can predict physician decision to prescribe SCS and to prescribe antibiotics respectively

I sought to develop supervised learning models to predict the physician decision to prescribe SCS and the physician decision to predict antibiotics. Due to the small number of participants who were given antibiotics alone or neither antibiotics nor SCS at exacerbation, I developed algorithms to predict a) SCS prescription irrespective of whether or not antibiotics were also given; b) antibiotics irrespective of whether SCS were also given. I first employed a linear approach using the LASSO method to enable selection of key variables to predict exacerbation treatment. The final LASSO model for predicting SCS prescription included heart rate (positive coefficient), LAMA use (negative coefficient) and PPI use (negative coefficient) and achieved an average cross-validation AUC of 0.61. The final lasso model for predicting antibiotic prescription included just one variable, namely, whether the participant reported to have increased sputum purulence at exacerbation (positive coefficient) and achieved an average cross-validation AUC of 0.68.

I next performed non-linear approaches for predicting exacerbation treatment. In order to select variables for use by the non-linear classification algorithms, I used the SES algorithm described in Chapter 2. The SES algorithm identified the following variables as being important for SCS prescription prediction: heart rate (SES $p<0.01$), haemoglobin (SES $p=0.07$), participant LAMA use (SES $p=0.03$) and blood eosinophil percentage (SES $p<0.01$). Blood eosinophil count was identified as being statistically equivalent to blood eosinophil percentage in the multivariable context for predicting SCS prescription.

For antibiotic prescription prediction, the SES algorithm identified increased sputum purulence reported at exacerbation (SES $p<0.01$), fever at exacerbation (SES $p=0.04$), whether antibiotics were given to the participant in the week preceding the exacerbation (SES $p=0.07$), presence of nasal polyps (SES $p=0.09$) and blood neutrophil percentage (SES $p<0.01$) as being important. Lymphocyte percentage (SES $p=0.80$) was identified as being statistically equivalent to blood neutrophil percentage in the multivariable context for predicting antibiotic prescription. Table 3.6 shows the best average cross-validation performance of each of the classification methods for SCS and antibiotic prescription.

Table 3.6 | Average cross-validation AUC performance of optimised algorithms to predict SCS prescription and antibiotic prescription.

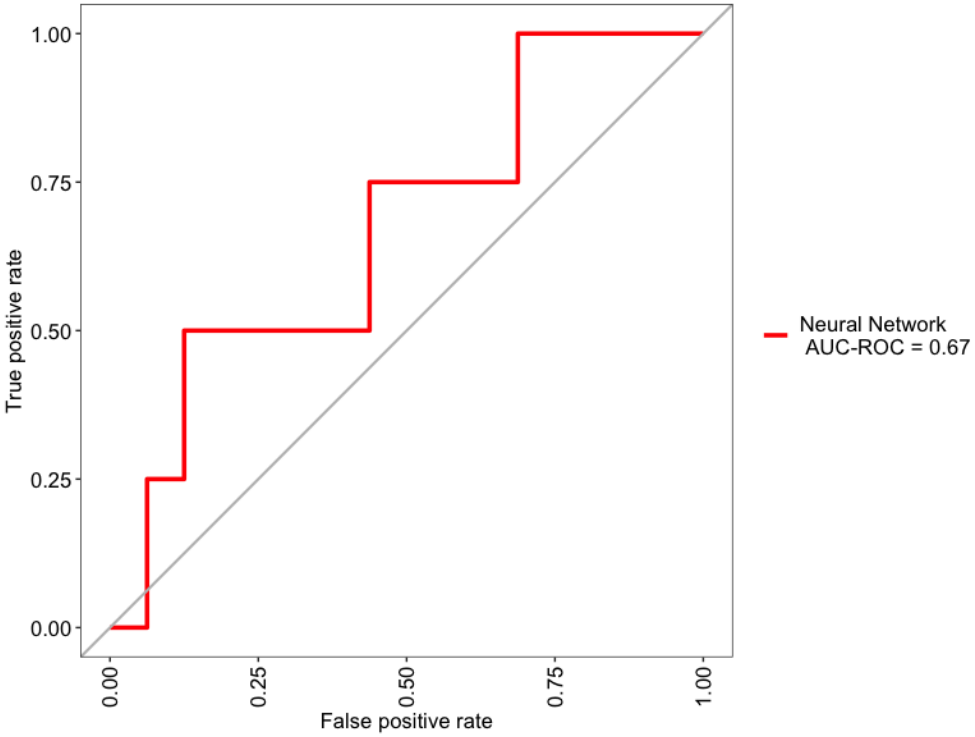
Predicting SCS prescription		
Model	**Average cross-validation AUC curve**	**Best hyper-parameter values**
LASSO	0.61	λ = 0.11
Random forest	0.77	800 trees, mtry parameter = 3
KNN	0.75	K=5
Decision tree	0.55	Complexity parameter = 0.55
SVM	0.82	Sigma=0.050, cost penalty=0.063
Neural network	**0.85**	**Size=1, decay=0.1**
Predicting antibiotic prescription		
LASSO	0.68	λ = 0.19
Random forest	0.86	100 trees, mtry parameter = 1
KNN	0.75	K=29
Decision tree	0.71	Complexity parameter = 0.033
SVM	**0.86**	**Sigma=0.052, cost penalty=0.063**
Neural network	0.85	Size=2, decay=1.2

Classifier model in bold font has best performance. Although random forest and support vector machine achieved 0.86 cross-validation AUC to 2 decimal places for predicting antibiotic prescription, the SVM had a higher performance to the third decimal place.

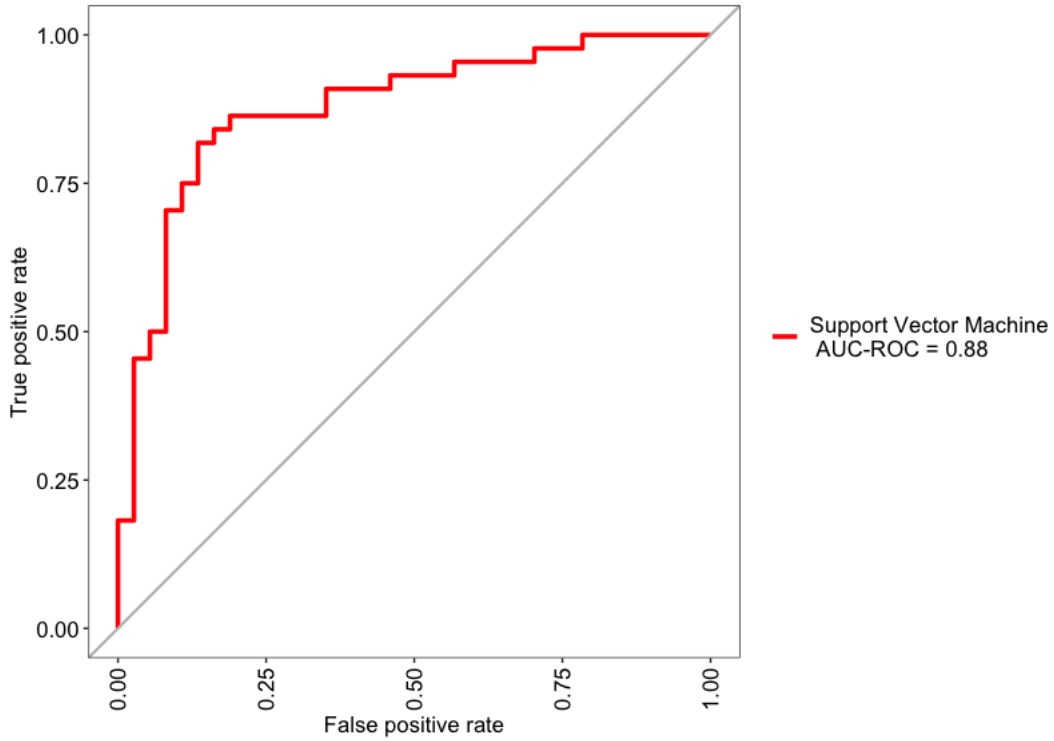
Based on the cross-validation results, the best models for predicting SCS prescription and antibiotic prescription were the neural network and SVM, respectively. I applied these models on the test set to obtain a generalisable estimate of performance. SCS prescription could be predicted with a test set AUC of 0.67 (Figure 3.2A). Antibiotic prescription could be predicted with a test set AUC of 0.88 (Figure 3.2B).

Figure 3.2 | Test set ROC curve for predicting exacerbation physician treatment prescription. A: Neural network predicting SCS prescription. B: SVM predicting antibiotic prescription.

A



B



Classifier model using variables including past healthcare utilisation, medication usage, exacerbation sputum purulence and whether oxygen was administered at exacerbation can predict D1/D5 hospitalisation

90 participants had data available for whether hospitalisation occurred within the first 5 days after D0. Out of these 90 participants, 29 (32%) were hospitalised (D1 and/or D5 hospital visit) and 61 (68%) were not hospitalised (Table 3.7). 28 of the 29 hospitalised participants had D1 measurements available (Appendix Table A3.3). 11 participants hospitalised at D5 had data available (Appendix Table A3.4). VAS cough and dyspnoea, basophil levels (blood basophil cell count and percentage) and haemoglobin decreased between D0 and D1 ($p<0.01$ in all cases). VAS dyspnoea at D5 was also lower than at D0 ($p<0.01$). In addition, peak flow at D5 had improved compared to D0 ($p<0.01$).

Participants who were hospitalised at D1/D5 had more frequent unscheduled primary care and emergency department visits during the preceding 12 months (Table 3.7) compared to participants who were not hospitalised immediately after D0 ($p<0.01$). D0 symptom burden was greater for participants who were immediately hospitalised after exacerbation than participants not immediately hospitalised (VAS cough $p <0.05$, VAS dyspnoea tending towards significance at $p = 0.05$, VAS sputum production and purulence both $p < 0.01$, increased sputum production and purulence reported at exacerbation both $p <0.01$) and EuroQol mobility score higher ($p<0.01$). Participants who were immediately hospitalised also had lower lymphocytes ($p=0.02$) and platelets ($p=0.01$). Participants who were immediately hospitalised were more likely to have had oxygen administered at D0 ($p<0.01$). In addition, participants who were immediately hospitalised were more likely to be on carbocystiene ($p=0.01$) and less likely to be on

inhaled corticosteroids (ICS) (p=0.01) compared to participants who were not immediately hospitalised (Appendix Table A3.5).

Table 3.7 | D0 exacerbation characteristics of 90 patients with known D1/5 healthcare utilisation treatment outcome. Characteristics of patients re-hospitalised versus not re-hospitalised within 5 days of exacerbation (hospital visit on D1 or D5) are compared.

Variable		D1/5 Hospitalisation (n=29)	No D1/5 Hospitalisation (n=61)	P-Value
Number of days for which patient reported symptoms have been worse for in lead up to exacerbation		1 (2 – 7)	3 (1 – 7)	0.64
VAS cough (mm)		**63 (36 – 79)**	**47 (21 – 69)**	**<0.05**
VAS dyspnoea (mm)		78 (56 – 89)	66 (20 – 80)	0.05
VAS sputum production (mm)		**40 (15 – 56)**	**19 (0 – 31)**	**<0.01**
VAS sputum purulence (mm)		**45 (10 – 68)**	**4 (0 – 32)**	**<0.01**
Increased cough, n (%)		25 (86)	49 (80)	0.50
Increased dyspnoea, n (%)		29 (100)	60 (98)	0.49
Increased sputum production, n (%)		**22 (76)**	**24 (39)**	**<0.01**
Increased sputum purulence, n (%)		**18 (62)**	**18 (30)**	**<0.01**
Increased wheeze, n (%)		24 (83)	54 (89)	0.45
Fever, n (%)		15 (52)	28 (46)	0.61
EuroQol VAS score		30 (10 – 40)	38 (18 – 58)	0.14
EuroQol mobility	**1, n (%)**	**8 (28)**	**22 (37)**	**<0.01**
	2, n (%)	**13 (45)**	**36 (60)**	
	3, n (%)	**8 (28)**	**2 (3)**	
EuroQol self care	1, n (%)	18 (62)	45 (75)	0.09
	2, n (%)	9 (31)	15 (25)	
	3, n (%)	2 (7)	0 (0)	
EuroQol usual activity	1, n (%)	5 (17)	19 (32)	0.15
	2, n (%)	12 (41)	27 (45)	
	3, n (%)	12 (41)	14 (23)	
EuroQol pain discomfort	1, n (%)	6 (21)	19 (32)	0.50
	2, n (%)	22 (76)	38 (63)	

	3, n (%)	1 (3)	3 (5)	
EuroQol anxiety depression	1, n (%)	20 (69)	29 (48)	0.18
	2, n (%)	7 (24)	23 (38)	
	3, n (%)	2 (7)	8 (13)	
HAD anxiety		7 (4 – 9)	7 (4 – 11)	0.54
HAD depression		7 (3 – 9)	4 (2 – 9)	0.46
CAT Score		18 (0 – 31)	19 (0 – 27)	0.81
Charlson Index Score		0 (0 – 2)	0 (0 – 1)	0.59
Oxygen saturation (%)		95 (92 – 97)	95 (94 – 97)	0.19
Blood pH		7.4 (7.4 – 7.4)	7.4 (7.4 – 7.4)	0.64
Blood pCO2 (kPa)*		5.4 (4.9 – 5.8)	5.3 (5.0 – 5.6)	0.84
Heart rate (beats/minute)~		102 (4)	97 (2)	0.29
Temperature (°C)		36.6 (36.1 – 37.2)	36.5 (36.2 – 37.0)	0.70
Systolic (mmHg)*		132 (125 – 138)	138 (132 – 144)	0.12
Diastolic (mmHg)*		80 (3)	79 (2)	0.70
Respiratory rate (breaths per minute)*		20 (18 – 22)	21 (19 – 22)	0.57
Peak flow (litres/minute)*		192 (164 – 225)	225 (196 – 259)	0.13
Total leucocyte count, (x10⁹cells/L)*		11.5 (10.2 – 13.0)	10.7 (9.9 – 11.6)	0.30
Blood neutrophil count (x10⁹cells/L)*		8.4 (7.1 – 10.1)	7.3 (6.5 – 8.1)	0.15
Blood eosinophil count (x10⁹cells/L)		0.11 (0.03 – 0.29)	0.21 (0.08 – 0.49)	0.06
Blood basophil count (x10⁹cells/L)		0.05 (0.03 – 0.06)	0.05 (0.03 – 0.08)	0.58
Blood monocyte count (x10⁹cells/L)		0.86 (0.64 – 1.14)	0.76 (0.57 – 0.91)	0.20
Blood lymphocyte count (x10⁹cells/L)*		**1.29 (0.99 – 1.67)**	**1.77 (1.52 – 2.06)**	**0.02**
Blood neutrophil percentage~		75.0 (2.6)	70 (2)	0.09
Blood eosinophil percentage		1.2 (0.2 – 2.5)	2.0 (0.6 – 5.3)	0.07
Blood basophil percentage*		0.4 (0.3 – 0.5)	0.44 (0.39 – 0.52)	0.29
Blood monocyte percentage		7.2 (5.4 – 9.6)	7.0 (5.2 – 8.3)	0.37
Blood lymphocyte percentage		**12.1 (7.5 – 18.2)**	**17.6 (11.8 – 26.1)**	**0.01**
Haemoglobin (g/L)~		143 (2)	144 (2)	0.87
Platelets (x10⁹cells/L)		**250 (217 – 289)**	**309 (239 – 376)**	**0.01**
Sodium (mmol/L)		140 (138 – 141)	139 (137 – 141)	0.76
Potassium (mmol/L)		3.7 (3.2 – 4.1)	3.8 (3.5 – 4.1)	0.19
Urea (mmol/L)		4.8 (3.3 – 6.8)	4.6 (3.7 – 6.0)	0.65

Creatinine (µmol/L)		64 (58 – 88)	69 (57 – 78)	0.86
eGFR (mL/min)		90 (73 – 90)	90 (80 – 90)	0.38
Blood glucose (mmol/L)		6.6 (6.0 – 8.4)	6.9 (5.7 – 7.6)	0.71
Serum albumin (g/L)		36 (33 – 39)	37 (34 – 38)	0.60
CRP (mg/L)		10.5 (2.9 – 39.1)	7.2 (0.9 – 24.3)	0.22
Intravenous hydrocortisone, n (%)		5 (17)	11 (18)	0.93
Oxygen administered at exacerbation, n (%)		**14 (48)**	**11 (20)**	**<0.01**
Exacerbation treatment, n (%)	SCS and antibiotics, n (%)	16 (55)	19 (31)	0.10
	SCS Only, n (%)	8 (28)	28 (46)	
	Antibiotics only, n (%)	3 (10)	4 (7)	
	Neither, n (%)	2 (7)	10 (16)	

Definition of abbreviations: CAT score = COPD Assessment Tool score; CRP = C-reactive protein; eGFR = estimated glomerular filtration rate; HAD = Hospital Anxiety and Depression scale; SCS = systemic corticosteroids; VAS = Visual Analogue Scale.
Variable presentation and hypothesis tests: Numerical variables following a normal distribution are presented as mean (standard error of the mean); log-transformed variables following a normal distribution are presented as geometric mean (95% confidence interval); other variables are presented as median (lower quartile to upper quartile). For variables following a normal distribution, unpaired two-sided t-tests are used. These variables are represented as mean and marked ~. For log-transformed variables following a normal distribution unpaired two-sided t-tests on log-transformed data are used and marked *. For other numerical variables, the Mann-Whitney U test is used. Binary and categorical variables are presented as number (%) of instances (chi-squared test used for treatment outcome comparison).
Variables in bold font differ significantly between D1/5 healthcare utilisation treatment outcomes.
Correction for multiple comparisons is not made since the analysis is hypothesis-generating.

A LASSO model used 19 variables for predicting D1/D5 hospitalisation and achieved an average cross-validation AUC of 0.79. The 19 variables selected by LASSO associated with D1/D5 hospitalisation included the number of unscheduled primary care and emergency department visits, total number of hospital admissions during the preceding 12 months, whether oxygen was administered at exacerbation, whether SCS were given at exacerbation and whether antibiotics were given at exacerbation.

Negatively associated with D1/D5 hospitalisation in the LASSO model was whether the participant was on ICS.

The SES algorithm identified the following variables as significant for use in non-linear classifier models to predict D1/D5 hospitalisation: number of unscheduled primary care and emergency department visits during previous 12 months (SES $p<0.01$), EuroQol anxiety and depression score (SES $p<0.01$; participants who are immediately hospitalised were more likely to have a lower EuroQol anxiety and depression score); whether the participant was on ICS (SES $p=0.08$), whether oxygen was administered at D0 (SES $p=0.03$) and VAS sputum purulence (SES $p=0.03$). VAS sputum production and whether the participant reported that the exacerbation was associated with increased sputum purulence were identified as statistically equivalent to VAS sputum purulence in the multivariable context for predicting D1/D5 hospitalisation. The average cross-validation performance of the LASSO, random forest algorithm, KNN algorithm, decision tree, SVM and neural network algorithm with optimal hyper-parameter values for predicting D1/D5 hospitalisation is shown in Table 3.8. The best non-linear algorithms outperformed the LASSO approach for the prediction.

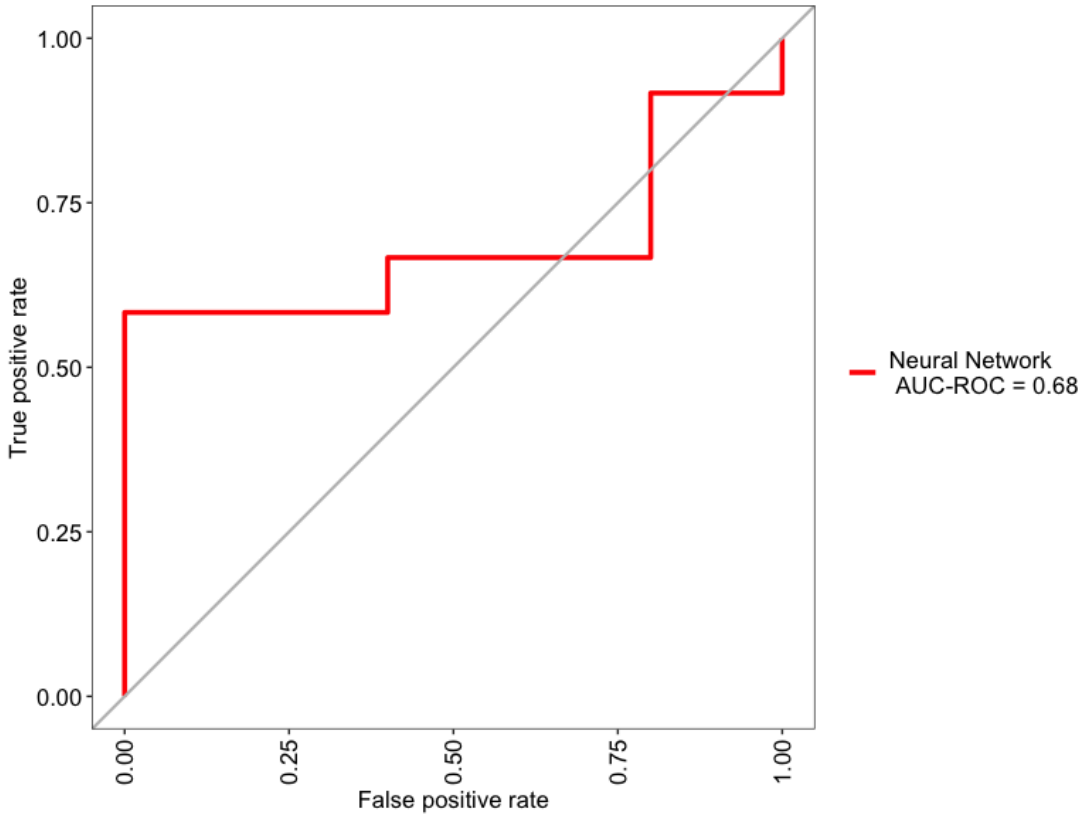
Table 3.8 | Average cross-validation AUC performance of optimised algorithms to predict D1/D5 hospitalisation.

Model	**Average cross-validation AUC**	**Best hyper-parameter values**
LASSO	0.79	λ = 0.045
Random forest	0.81	500 trees, mtry parameter = 1
KNN	0.73	K=17
Decision tree	0.66	Complexity parameter = 0.028
SVM	0.85	Sigma=0.051 , cost penalty=2
Neural network	**0.89**	**Size=3, decay=0.9**

Classifier model in bold font has best performance.

Based on the cross-validation results, the best model for predicting D1/D5 hospitalisation was the neural network. I applied the neural network model on the test set to obtain a generalisable estimate of performance. D1/D5 hospitalisation could be predicted with a test set AUC of 0.68 (Figure 3.3).

Figure 3.3 | Test set ROC curve for D1/D5 hospitalisation prediction model.



Classifier models using variables including past healthcare utilisation, medication usage and exacerbation dyspnoea can predict D30 healthcare utilisation and patient-defined treatment failure

D30 treatment outcome data were available for 81 participants. Paired data for D0 and D30 were available for 53 participants (Appendix Table A3.6). Between D0 and D30, participants showed improvement in VAS cough ($p<0.01$), dyspnoea ($p<0.01$), VAS sputum purulence ($p<0.05$) and a reduction in COPD Assessment Tool (CAT) score ($p<0.01$).

Healthcare utilisation treatment failure by D30 occurred for 43 (53%) participants. Exacerbation characteristics of participants with D30 healthcare utilisation failure versus success are shown in Table 3.9, while baseline characteristics are compared in Appendix Table A3.7. An additional 6 participants who did not experience healthcare utilisation treatment failure nevertheless did not report an improvement in symptoms by D30 relative to D0. Hence, 49 (60%) participants experienced patient-defined treatment failure by D30.

Participants who went on to experience D30 healthcare utilisation treatment failure had lower D0 blood oxygenation than participants with D30 healthcare utilisation treatment success ($p=0.03$). As in the case of D1/D5 hospitalisation, participants with D30 healthcare utilisation treatment failure had more frequent unscheduled primary care and emergency department visits in the preceding 12 months compared to participants with D30 treatment success ($p<0.01$). Also similar to the case of D1/D5 hospitalisation, D30 healthcare utilisation treatment failure participants had greater D0 symptom burden than participants with D30 treatment success. For example, D30 healthcare

utilisation treatment failure participants had greater VAS dyspnoea ($p<0.01$), sputum production ($p<0.01$) and purulence ($p=0.01$) at D0 and were also more likely to report that their exacerbation was associated with an increase in sputum production and purulence ($p=0.01$ and $p=0.02$ respectively) than participants with D30 treatment success. EuroQol mobility score was likelier to be in a higher category for participants with compared to without D30 healthcare utilisation treatment failure ($p=0.01$). Interestingly, D30 healthcare utilisation treatment failure participants were on average diagnosed with asthma/COPD later than participants with D30 treatment success (median age 49 years versus 19 years; $p=0.02$), a feature not apparent in the D1/D5 hospitalisation participant group compared to participants without D1/D5 hospitalisation. Again in keeping with the characteristics of participants with D1/D5 hospitalisation versus without D1/D5 hospitalisation, D30 healthcare utilisation treatment failure participants were more likely than D30 treatment success participants to have received oxygen at D0 ($p=0.02$) and less likely to be on ICS ($p=0.01$). Whereas for participants with versus without D1/D5 hospitalisation the difference in proportion of SABA use did not reach significance, participants with D30 healthcare utilisation treatment failure were significantly less likely to be on SABA than participants with D30 treatment success ($p=0.04$).

Table 3.9 | D0 exacerbation characteristics of 81 patients with known D30 healthcare utilisation treatment outcome. Patients who experienced healthcare utilisation treatment failure by D30 are compared against patients who had treatment success by D30.

Variable		**D30 healthcare utilisation treatment failure (n=43)**	**D30 healthcare utilisation treatment success (n=38)**	**P-Value**
Number of days for which patient reported symptoms have been worse for in lead up to exacerbation		4 (2 – 7)	3 (1 – 7)	0.47
VAS cough (mm)		57 (30 – 76)	42 (18 – 67)	0.11
VAS dyspnoea (mm)		**78 (58 – 89)**	**60 (5 – 79)**	**<0.01**
VAS sputum production (mm)		**28 (6 – 57)**	**10 (0 – 26)**	**<0.01**
VAS sputum purulence (mm)		**40 (0 – 70)**	**3 (0 – 20)**	**0.01**
Increased cough, n (%)		34 (79)	32 (84)	0.55
Increased dyspnoea, n (%)		43 (100)	37 (97)	0.28
Increased sputum production, n (%)		**28 (65)**	**13 (34)**	**0.01**
Increased sputum purulence, n (%)		**22 (51)**	**10 (26)**	**0.02**
Increased wheeze, n (%)		37 (86)	33 (87)	0.92
Fever, n (%)		22 (51)	16 (42)	0.42
EuroQol VAS score		35 (10 – 45)	30 (5 – 64)	0.68
EuroQol mobility	**1, n (%)**	**11 (26)**	**14 (38)**	**0.01**
	2, n (%)	**22 (51)**	**23 (62)**	
	3, n (%)	**10 (23)**	**0 (0)**	
EuroQol self care	1, n (%)	25 (58)	30 (81)	0.06
	2, n (%)	16 (37)	7 (19)	
	3, n (%)	2 (5)	0 (0)	
EuroQol usual activity	1, n (%)	8 (19)	12 (32)	0.10
	2, n (%)	18 (42)	18 (49)	
	3, n (%)	17 (40)	7 (19)	
EuroQol pain discomfort	1, n (%)	11 (26)	11 (30)	0.90
	2, n (%)	30 (70)	24 (65)	
	3, n (%)	2 (5)	2 (5)	
EuroQol anxiety depression	1, n (%)	27 (63)	17 (46)	0.18
	2, n (%)	13 (30)	13 (35)	
	3, n (%)	3 (7)	7 (19)	
HAD anxiety		7 (4 – 10)	7 (4 – 10)	0.93

HAD depression	5 (3 – 9)	4 (1 – 9)	0.54
CAT score	18 (0 – 30)	19 (0 – 27)	0.83
Charlson Index Score	0 (0 – 1)	0 (0 – 2)	0.73
Oxygen saturation (%)	**94 (92 – 97)**	**95 (95 – 97)**	**0.03**
Blood pH	7.4 (7.4 – 7.4)	7.4 (7.4 – 7.4)	0.70
Blood pCO2 (kPa)*	5.6 (5.2 – 6.0)	5.2 (4.8 – 5.6)	0.15
Heart rate (beats/minute)~	102 (3)	96 (3)	0.15
Temperature (°C)	36.6 (36.1 – 37.2)	36.6 (36.0 – 37.0)	0.88
Systolic (mmHg)*	133 (128 – 139)	137 (130 – 145)	0.40
Diastolic (mmHg)~	81 (2)	79 (3)	0.62
Respiratory rate (breaths per minute)*	21 (19 – 22)	20 (19 – 22)	0.70
Peak flow (litres/minute)*	192 (163 – 227)	214 (183 – 249)	0.35
Total leucocyte count, (x10⁹cells/L)*	11.1 (10.1 – 12.4)	10.6 (9.6 – 11.7)	0.51
Blood neutrophil count (x10⁹cells/L)*	7.8 (6.8 – 9.1)	7.4 (6.5 – 8.4)	0.54
Blood eosinophil count (x10⁹cells/L)	0.13 (0.03 – 0.36)	0.24 (0.06 – 0.47)	0.30
Blood basophil count (x10⁹cells/L)	0.05 (0.03 – 0.06)	0.05 (0.03 – 0.08)	0.37
Blood monocyte count (x10⁹cells/L)	0.76 (0.57 – 1.01)	0.76 (0.56 – 0.91)	0.41
Blood lymphocyte count (x10⁹cells/L)*	1.5 (1.2 – 1.9)	1.6 (1.3 – 1.9)	0.89
Blood neutrophil percentage~	72.1 (2.2)	71.0 (2.3)	0.72
Blood eosinophil percentage	1.5 (0.2 – 3.4)	1.7 (0.4 – 5.4)	0.33
Blood basophil percentage*	0.4 (0.3 – 0.5)	0.4 (0.4 – 0.6)	0.37
Blood monocyte percentage	7.1 (5.1 – 9.3)	7.0 (5.1 – 8.2)	0.44
Blood lymphocyte percentage	13.5 (8.8 – 24.8)	15.5 (10.4 – 25.9)	0.71

Haemoglobin (g/L)~		142 (2)	146 (2)	0.27
Platelets (x10⁹cells/L)		261 (221 – 302)	276 (230 – 369)	0.23
Sodium (mmol/L)		139 (138 – 141)	139 (137 – 141)	0.99
Potassium (mmol/L)		3.7 (3.5 – 4.1)	3.8 (3.5 – 4.2)	0.35
Urea (mmol/L)		4.7 (3.6 – 6.3)	4.4 (3.7 – 6.0)	0.47
Creatinine (µmol/L)		66 (59 -79)	70 (58 – 79)	0.90
eGFR (mL/min)		90 (72 – 90)	90 (80 – 90)	0.25
Blood glucose (mmol/L)		6.6 (6.0 – 8.3)	7.1 (5.6 – 8.1)	0.95
Serum albumin (g/L)		36 (34 – 38)	37 (33 – 38)	0.41
CRP (mg/L)		10.3 (2.4 – 25.4)	7.9 (1.0 – 32.9)	0.59
Intravenous hydrocortisone, n (%)		7 (16)	5 (13)	0.69
Oxygen administered at exacerbation, n (%)		**17 (41)**	**6 (17)**	**0.02**
Exacerbation treatment, n (%)	SCS and Antibiotics, n (%)	19 (44)	12 (32)	0.09
	SCS, n (%)	19 (44)	14 (37)	
	Antibiotics, n (%)	3 (7)	3 (8)	
	Neither, n (%)	2 (5)	9 (24)	

Definition of abbreviations: CAT score = COPD Assessment Tool; CRP = C-reactive protein; eGFR = estimated glomerular filtration rate; HAD = Hospital Anxiety and Depression scale; SCS = systemic corticosteroids; VAS = Visual Analogue Scale.
Variable presentation and hypothesis tests: Numerical variables following a normal distribution are presented as mean (standard error of the mean); log-transformed variables following a normal distribution are presented as geometric mean (95% confidence interval); other variables are presented as median (lower quartile to upper quartile).
For numerical variables following a normal distribution, unpaired two-sided t-tests are used. These variables are represented as mean and marked ~. For log-transformed variables following a normal distribution unpaired two-sided t-tests on log-transformed data are used and marked *. For other numerical variables, the Mann-Whitney U test is used.
Binary and categorical variables are presented as number (%) of instances (chi-squared test used for treatment outcome comparison).
Variables in bold font differ significantly between D30 healthcare utilisation treatment outcomes.
Correction for multiple comparisons is not made since the analysis is hypothesis-generating.

The LASSO model for predicting D30 healthcare utilisation treatment failure selected only VAS dyspnoea as a relevant variable and achieved a cross-validation AUC of 0.66. For predicting patient-defined D30 treatment outcome, the LASSO model selected 7 variables including, for example, VAS dyspnoea, blood oxygen saturation,

unscheduled primary care and emergency department visits during the preceding 12 months, ICS use and PPI use. This LASSO model achieved a cross-validation AUC of 0.65. Using the SES algorithm, I identified the following variables as important for predicting D30 healthcare utilisation treatment failure in the multivariable context: number of unscheduled primary care and emergency department visits in preceding 12 months (SES p=0.01), VAS dyspnoea (SES p<0.01), whether the participant reported that their exacerbation was associated with an increase in sputum production (SES p=0.08), participant ICS use (p=0.05) and whether oxygen was administered at D0 (SES p=0.06). For predicting D30 patient-defined treatment failure, the SES algorithm identified the following variables as significant in the multivariable context: blood oxygen saturation (SES p=0.06), number of unscheduled primary care and emergency department visits in preceding 12 months (SES p=0.01), VAS dyspnoea (SES p=0.08), participant ICS use (SES p<0.01) and blood monocyte percentage (SES p=0.04). Haemoglobin and participant proton pump inhibitor use were identified as being statistically equivalent to blood monocyte percentage. The performance of the various machine learning approaches for predicting D30 treatment outcome are shown in Table 3.10. For both healthcare utilisation and patient-defined D30 treatment failure, the neural network was the best model was the best predictive model.

Table 3.10 | Average cross-validation AUC performance of optimised algorithms to predict D30 treatment failure.

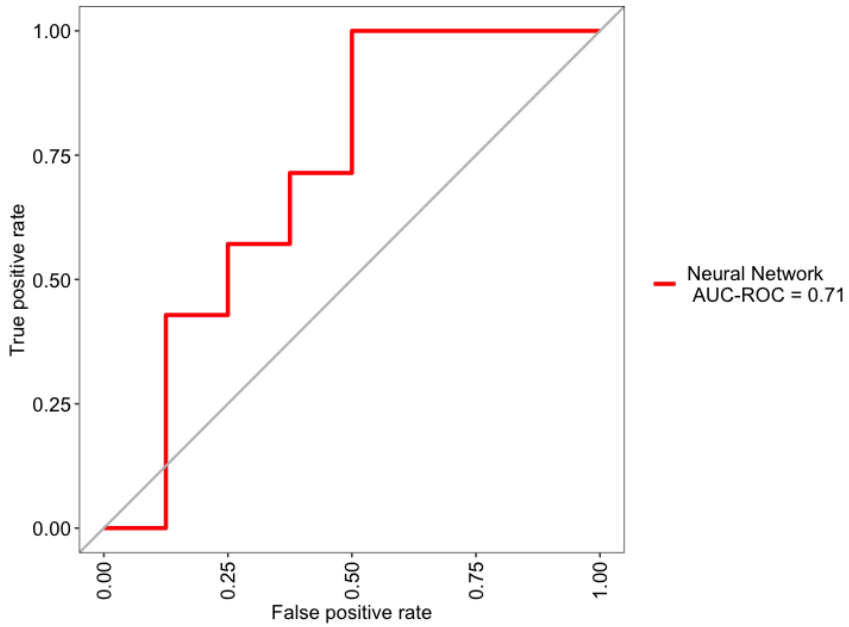
Predicting D30 healthcare utilisation treatment failure		
Model	**Average cross-validation AUC**	**Best hyper-parameter values**
LASSO	0.66	λ = 0.21
Random forest	0.82	400 trees, mtry parameter = 1
KNN	0.75	K=7
Decision tree	0.65	Complexity parameter = 0.00
SVM	0.86	Sigma = 0.042 , cost penalty= 0.063
Neural network	**0.87**	**Size=1, decay=2.0**
Predicting D30 patient-defined treatment failure		
LASSO	0.65	λ = 0.11
Random forest	0.81	300 trees, mtry parameter = 1
KNN	0.75	K=33
Decision tree	0.52	Complexity parameter = 0.27
SVM	0.90	Sigma = 0.046 , cost penalty= 4
Neural network	**0.91**	**Size=2, decay=0.4**

Classifier model in bold font has best performance.

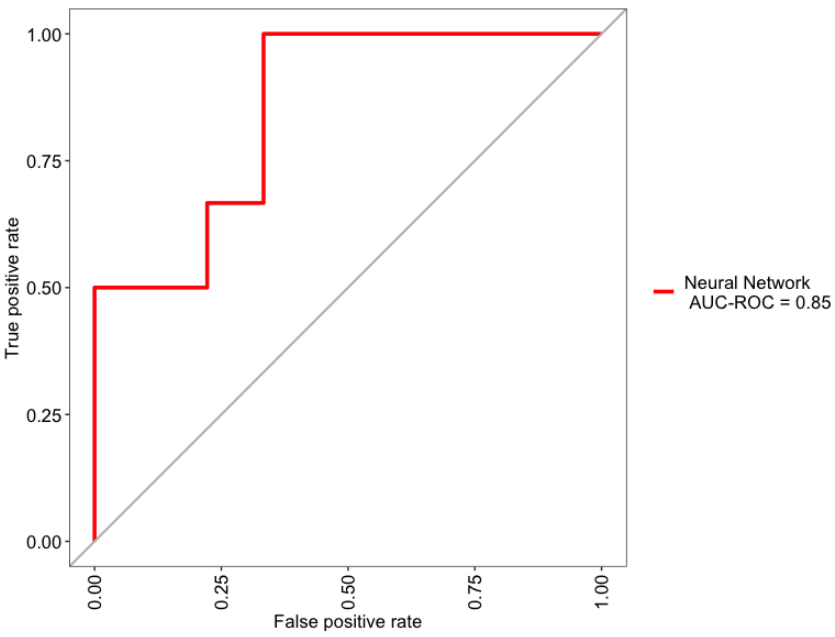
The test set performance of the neural network for predicting D30 healthcare utilisation treatment failure and for predicting D30 patient-defined treatment failure was an AUC of 0.71 and 0.85, respectively (Figure 3.4A and 3.4B).

Figure 3.4 | Test set ROC curve for D30 treatment failure prediction models. A: Neural network predicting D30 healthcare utilisation treatment failure. B: Neural network predicting D30 patient-defined treatment failure.

A



B



Classifier model using variables including medication usage and exacerbation dyspnoea can predict D90 healthcare utilisation treatment outcome

85 participants had D90 treatment outcome data available. Of these, 59 participants had paired D0 and D90 follow-up symptom data available (Appendix Table A3.8). All VAS scores and the CAT score improved between D0 and D90 ($p<0.01$ in all cases).

53 (62%) participants experienced healthcare utilisation treatment failure by D90. Exacerbation characteristics of participants with D90 healthcare utilisation failure versus success are shown in Table 3.11, while baseline characteristics are compared in Appendix Table A3.9. The characteristics differing between D90 healthcare utilisation treatment failure participants and D90 treatment success participants are similar to those identified when comparing D30 treatment failure versus success. The number of unscheduled primary care and emergency department visits in the preceding 12 months were greater for D90 healthcare utilisation treatment failure versus success ($p=0.02$). D0 VAS dyspnoea and sputum scores ($p<0.01$ in all cases) were also higher for participants with D90 healthcare utilisation treatment failure. The EuroQol mobility score is shifted towards higher levels in the D90 healthcare utilisation treatment failure compared to treatment success group ($p=0.03$). A greater proportion of D90 healthcare utilisation treatment failure participants than success participants had oxygen administered at D0 ($p=0.02$), while a lower proportion were on aspirin and ICS ($p<0.01$ in both cases). A higher proportion of participants with D90 healthcare utilisation treatment failure were on carbocystiene compared to D90 treatment success participants ($p<0.01$). However, the D90 healthcare utilisation treatment failure group had a greater proportion of non-smokers than the treatment success group ($p=0.03$).

Furthermore, the proportion of males was lower in the D90 healthcare utilisation treatment failure group (30% versus 56%; p=0.02).

Table 3.11 | D0 exacerbation characteristics of 85 patients with known D90 healthcare utilisation treatment outcome. Patients who experienced D90 healthcare utilisation treatment failure by D90 are compared against patients who had treatment success by D90.

Variable		D90 healthcare utilisation treatment failure (n=53)	D90 healthcare utilisation treatment success (n=32)	P-Value
Number of days for which patient reported symptoms have been worse for in lead up to exacerbation		4 (2 – 7)	3 (1 – 7)	0.18
VAS cough (mm)		57 (30 – 74)	36 (16 – 64)	0.06
VAS dyspnoea (mm)		**75 (57 – 89)**	**32 (2 – 78)**	**<0.01**
VAS sputum production (mm)		**25 (6 – 56)**	**12 (0 – 30)**	**<0.01**
VAS sputum purulence (mm)		**18 (1 – 58)**	**2 (0 – 25)**	**<0.01**
Increased cough, n (%)		44 (83)	26 (81)	0.84
Increased dyspnoea, n (%)		53 (100)	31 (97)	0.20
Increased sputum production, n (%)		**32 (60)**	**12 (38)**	**0.04**
Increased sputum purulence, n (%)		25 (47)	9 (28)	0.08
Increased wheeze, n (%)		46 (87)	28 (88)	0.93
Fever, n (%)		25 (47)	14 (44)	0.76
EuroQol VAS score		35 (15 – 43)	30 (3 – 67)	0.83
EuroQol mobility	**1, n (%)**	**14 (26)**	**12 (39)**	**0.03**
	2, n (%)	**29 (55)**	**19 (61)**	
	3, n (%)	**10 (19)**	**0 (0)**	
EuroQol self care	1, n (%)	34 (64)	25 (81)	0.21
	2, n (%)	17 (32)	6 (19)	
	3, n (%)	2 (4)	0 (0)	
	1, n (%)	11 (21)	11 (35)	0.11

EuroQol usual activity	2, n (%)	23 (43)	15 (48)	
	3, n (%)	19 (36)	5 (16)	
EuroQol pain discomfort	1, n (%)	13 (25)	9 (29)	0.74
	2, n (%)	38 (72)	20 (65)	
	3, n (%)	2 (4)	2 (6)	
EuroQol anxiety depression	1, n (%)	30 (57)	16 (52)	0.27
	2, n (%)	19 (36)	9 (29)	
	3, n (%)	4 (8)	6 (19)	
HAD anxiety		7 (4 – 10)	6 (3 – 9)	0.80
HAD depression		5 (3 – 9)	3 (1 – 9)	0.26
CAT score		19 (0 – 29)	19 (1 – 27)	0.82
Charlson Index Score		0 (0 – 1)	0 (0 – 2)	1.00
Oxygen saturation (%)		95 (92 – 97)	95 (95 – 98)	0.08
Blood pH		7.4 (7.4 – 7.4)	7.4 (7.4 – 7.4)	0.78
Blood pCO2 (kPa)*		5.4 (5.1 – 5.7)	5.3 (4.9 – 5.7)	0.60
Heart rate (beats/minute)~		102 (3)	96 (3)	0.15
Temperature (°C)		36.6 (36.2 – 37.2)	36.5 (36.0 – 37.0)	0.82
Systolic (mmHg)*		134 (129 – 140)	136 (128 – 145)	0.71
Diastolic (mmHg)~		81 (2)	79 (3)	0.60
Respiratory rate (breaths per minute)*		21 (19 – 22)	20 (18 – 22)	0.74
Peak Flow (litres/minute)*		208 (180 – 239)	213 (177 – 255)	0.84
Total leucocyte count, (x10⁹cells/L)*		11.3 (10.4 – 12.3)	10.3 (9.2 – 11.6)	0.21
Blood neutrophil count (x10⁹cells/L)*		7.8 (6.9 – 8.9)	7.3 (6.3 – 8.5)	0.48
Blood eosinophil count (x10⁹cells/L)		0.12 (0.04 – 0.38)	0.26 (0.09 – 0.51)	0.24
Blood basophil count (x10⁹cells/L)		0.05 (0.03 – 0.06)	0.05 (0.03 – 0.08)	0.93
Blood monocyte count (x10⁹cells/L)		0.80 (0.59 – 1.04)	0.72 (0.53 – 0.89)	0.10
Blood lymphocyte count (x10⁹cells/L)*		1.8 (1.1 – 2.4)	1.4 (1.1 – 1.8)	0.27
Blood neutrophil Percentage~		70.8 (2.0)	72.0 (2.5)	0.72
Blood eosinophil percentage		1.4 (0.3 – 3.4)	2.7 (0.5 – 5.5)	0.25
Blood basophil percentage*		0.4 (0.4 – 0.5)	0.4 (0.3 – 0.6)	0.64

Blood monocyte percentage		7.2 (5.8 – 8.9)	7.0 (4.9 – 8.2)	0.35
Blood lymphocyte percentage		15.3 (9.6 – 25.9)	15.3 (9.5 – 25.3)	0.66
Haemoglobin (g/L)~		142 (2)	146 (2)	0.28
Platelets (x10⁹cells/L)		273 (225 – 344)	267 (224 – 364)	0.87
Sodium (mmol/L)		140 (138 – 141)	139 (137 – 141)	0.47
Potassium (mmol/L)		3.7 (3.5 – 4.1)	3.8 (3.5 – 4.3)	0.38
Urea (mmol/L)		4.8 (3.7 – 6.3)	4.4 (3.7 – 5.9)	0.25
Creatinine (µmol/L)		66 (59 – 79)	70 (55 – 82)	0.84
eGFR (mL/min)		90 (73 – 90)	90 (79 – 90)	0.65
Blood glucose (mmol/L)		6.5 (5.9 – 8.3)	7.2 (5.6 – 7.8)	0.80
Serum albumin (g/L)		36 (34 – 38)	38 (33 – 38)	0.45
CRP (mg/L)		10.3 (2.1 – 24.8)	7.1 (0.9 – 33.8)	0.54
Intravenous hydrocortisone, n (%)		9 (17)	4 (13)	0.58
Oxygen Administered, n (%)		**20 (40)**	**5 (16)**	**0.02**
Exacerbation treatment, n (%)	SCS and antibiotics, n (%)	23 (43)	10 (31)	0.48
	SCS only, n (%)	22 (42)	13 (41)	
	Antibiotics only, n (%)	3 (6)	3 (9)	
	Neither, n (%)	5 (9)	6 (19)	

Definition of abbreviations: CAT score = COPD Assessment Tool; CRP = C-reactive protein; eGFR = estimated glomerular filtration rate; HAD = Hospital Anxiety and Depression scale; SCS = systemic corticosteroids; VAS = Visual Analogue Scale.
Variable presentation and hypothesis tests: Numerical variables following a normal distribution are presented as mean (standard error of the mean); log-transformed variables following a normal distribution are presented as geometric mean (95% confidence interval); other variables are presented as median (lower quartile to upper quartile).
For numerical variables following a normal distribution, unpaired two-sided t-tests are used. These variables are represented as mean and marked ~. For log-transformed variables following a normal unpaired two-sided t-tests on log-transformed data are used and marked *. For other numerical variables, the Mann-Whitney U test is used.
Binary and categorical variables are presented as number (%) of instances (chi-squared test used for treatment outcome comparison).
Variables in bold font differ significantly between D90 healthcare utilisation treatment outcomes.
Correction for multiple comparisons is not made since the analysis is hypothesis-generating.

The LASSO model for predicting D90 treatment outcome selected as many as 29 variables including VAS symptom scores, baseline medication usage and exacerbation treatment, EuroQol and demographic variables such as sex and smoking status. An average cross-validation AUC of 0.70 was achieved using the 29-variable LASSO model. In contrast, the SES algorithm selects a much more parsimonious set of 4 variables: VAS dyspnoea (SES p<0.01), sex (SES p<0.01), ICS use (SES p=0.06) and carbocystiene use (SES p=0.09). The best performing D90 treatment failure prediction classifier based on average cross-validation AUC was the random forest (Table 3.12).

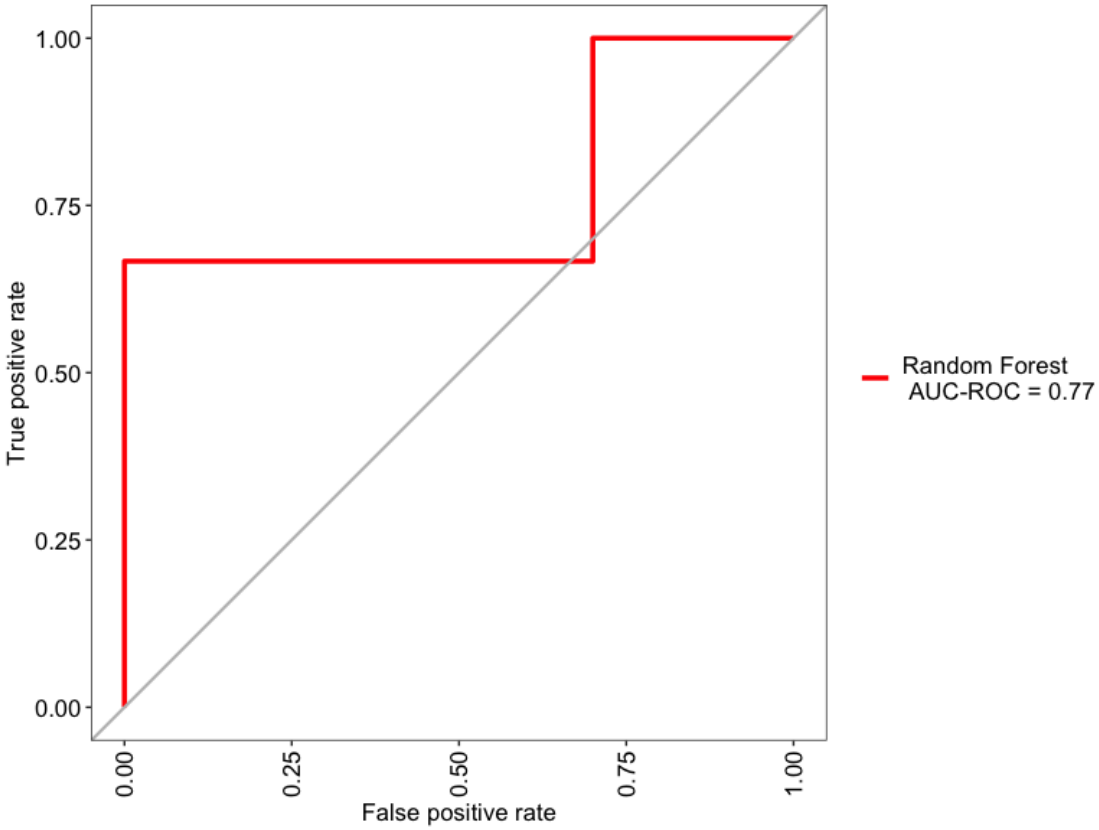
Table 3.12 | Average cross-validation AUC performance of optimised algorithms to predict D90 healthcare utilisation treatment failure.

Model	**Average cross-validation AUC**	**Best hyper-parameter values**
LASSO	0.70	λ = 0.0098
Random forest	**0.85**	**700 trees, mtry parameter = 1**
KNN	0.75	K=21
Decision tree	0.72	Complexity parameter = 0.085
SVM	0.85	Sigma = 0.21, cost penalty = 16
Neural network	0.85	Size = 9, decay = 0.1

Classifier model in bold font has best performance. Although random forest, support vector machine and neural network achieved 0.85 area under curve for cross validation to 2 decimal places, the random forest had a higher performance to the third decimal place.

The test set area under the receiver operating characteristic curve of the random forest D90 treatment outcome prediction model was 0.77 (Figure 3.5).

Figure 3.5 | Test set ROC curve for D90 treatment failure prediction model.



3.4 Discussion

Overall findings

In this chapter, I have described the characteristics of patients with asthma and COPD who present to the ED with an exacerbation. I have shown that treatment failure with hospital readmission immediately after the initial exacerbation occurs in 32% of patients. Treatment failure as hospital readmission or additional SCS and/or antibiotic therapy within 30 days from the initial exacerbation occurs in 53% of patients and within 90 days from the initial exacerbation occurs in 62% of patients. Depending on whether treatment outcome is analysed at the D1/5, D30 or D90 time point after initial exacerbation, features such us history of unscheduled primary care and emergency department visits, self-reported symptoms such as VAS dyspnoea, sputum production and purulence, whether oxygen is administered at exacerbation and patient use of ICS appear to be good predictors of a treatment failure following an exacerbation of airways disease. 7% of patients report that symptoms are not better by 30 days after exacerbation despite the fact the healthcare utilisation treatment failure outcome in the form of readmission to hospital or additional major treatment has not occurred. The clinician decision to prescribe SCS and the decision to prescribe antibiotics is related to several factors. Peripheral blood eosinophil percentage is associated with SCS prescription at the onset of an exacerbation of asthma and/or COPD. Although statistical significance was not reached, increased wheeze accompanying an exacerbation was reported more frequently for participants prescribed SCS compared to those not prescribed SCS.

Further interpretation of findings, strengths of analysis and comparison with other studies

My analysis is the first to assess a combination of VAS-based symptoms, biomarkers, clinical characteristics and demographics as predictors of treatment failure in a mixed cohort of hospitalised asthma and COPD patients. In this study I have used a variety of linear and non-linear machine learning algorithms to identify factors associated with exacerbation treatment failure at different time points from the initial exacerbation. I showed that patients who went on to fail treatment had greater symptom burden such as dyspnoea, sputum production and sputum purulence at exacerbation. Exacerbations are according to current definitions symptom-defined events and my findings illustrate this. Additionally, my results potentially allow for quantitative qualification of the symptoms measured at the onset of the event that are associated with exacerbations that fail to respond to management at the onset. Furthermore, patients who were on inhaled therapy (e.g. ICS) were less likely to have a treatment failure following an exacerbation, reiterating that undertreated COPD is associated with a higher chance of a patient requiring further healthcare utilisation[125]. I showed that in the case of D1/D5 hospitalisation as well as in the case of readmission or additional treatment with SCS and/or antibiotics within 30 days of initial exacerbation, the administration of oxygen at exacerbation is associated with treatment failure. This may reflect the fact that patients with more severe exacerbations, who hence are more likely to receive oxygen, are more likely to fail treatment. I also showed that the use of carbocystiene at exacerbation appeared to be associated with readmission to hospital or additional treatment with SCS and/or antibiotics by D90 after exacerbation. This may be due to patients being more likely to prescribed mucolytics such as carbocystiene if exacerbations are characterised by increased sputum production. The proportion of

patients with treatment failure by D90 who were on carbocystiene was indeed greater than that of the patients with D90 treatment success.

There are few validated, well-performing tools for predicting or stratifying patients' risk of COPD exacerbation treatment failure in the form of hospital readmission. The best performing such tool is the PEARL score (includes 5 predictors: previous admissions, extended Medical Research Council dyspnoea score, age, right-sided heart failure, left-sided heart failure) which stratifies patients' 90-day risk of hospital readmission or death without readmission[126,127]. The PEARL score was developed by Echevarria et al (2017) using 2,417 patients overall of whom 936 were readmitted or died within 90 days of discharge[127]. A cohort of 824 patients was used to train the PEARL logistic regression model, which was validated using an internal validation cohort of 802 patients (AUC 0.68 for 90-day hospital readmission or death without readmission) and an external validation cohort of 791 patients (AUC 0.70 for 90-day hospital readmission or death without readmission)[127]. The strengths of the PEARL score include the sizeable validation cohorts, the ease of measurement of the predictors and the multicentre nature of the cohorts used for PEARL score development[126,127]. However, it should be noted that the AUC of 0.70 in an external validation cohort is nevertheless not very high. This compares to my random forest model for predicting D90 treatment failure which achieved an AUC of 0.77 in a hold-out test set. My D90 treatment failure random forest model used patient self-reported dyspnoea via VAS scores, patient sex and medication use, in particular, ICS and carbocystiene. The lower performance of the PEARL score may in part be due to the fact that candidate predictors to include in this tool were chosen based on systematic literature review and 'clinical plausibility' rather than being selected mathematically from a wider pool of potential predictors[127].

The use of logistic regression rather than more sophisticated machine learning approaches discussed in Chapter 1 of this thesis may also limit performance. In terms of clinical context, the PEARL score is limited to COPD exacerbations and hence additional studies will be required before the PEARL score could be applied in context of exacerbations with ED admission in cases where lack of spirometry makes determination of COPD versus asthma diagnosis unclear. Finally, the PEARL score is focused on 90-day hospital readmission or death without readmission. This leaves open the task of developing a tool optimised for prediction of exacerbation treatment outcome within 30 days of initial discharge. The latter is also an important consideration since successful 30-day treatment failure prediction may translate into the possibility of increased monitoring of and interventions for patients shortly after initial exacerbation hospital admission in order to reduce the risk of early treatment failure. In the UK it is estimated that almost a quarter of patients hospitalised for COPD exacerbations are readmitted within 30 days of initial discharge[126,128]. Studies which attempt to use clinically widely used and easily measurable variables to predict 30-day/one-month treatment outcome following exacerbations requiring hospitalisation tend to either suffer from significant mathematical flaws or else fail to achieve high predictive performance. Matkovic et al performed a prospective, observational study in which data from 155 patients presenting to hospital for exacerbations of COPD were analysed[129]. Their study attempted to predict adverse outcomes following exacerbation, where adverse outcomes included: death during hospitalisation or 1-month follow-up, intensive care unit admission, invasive or non-invasive mechanical ventilation, hospitalisation of more than 11 days and COPD-related emergency visit or readmission within 1 month after discharge[129]. A three variable logistic regression model including arterial blood gas oxygen, CO_2 and the history of a COPD

exacerbation in the previous year was reported to achieve an AUC of 0.84[129]. While blood oxygenation and exacerbation history match predictors identified in my analysis and there is clear utility to a simple three variable model achieve successful prediction, unfortunately the results of the Matkovic et al study are likely significantly over-optimistic. This is because seemingly the performance of their logistic regression model is assessed using the same data as that used for model training. In the absence of cross validation or comparable methods and no hold-out test set, unfortunately the 0.84 AUC performance reported by their study may be unsubstantiated. At the opposite end of the spectrum, the Goto et al study described in Chapter 1 of my thesis in which 44,929 unplanned COPD hospitalisations were analysed from a Japanese database, used a robust machine learning framework but achieved a low AUC of 0.61 for predicting 30-day hospital readmission following an exacerbation of COPD[85]. This compares to my hospital readmission or additional systemic corticosteroids and/or antibiotics D30 treatment failure neural network model which achieved a test set area under receiver operating characteristic curve of 0.71. It is noteworthy that Goto et al did not include patient self-reported symptom scores nor were blood-based measurements included[85]. This may explain the low performance of Goto et al's algorithms especially because VAS symptom scores and blood oxygenation are key predictors of treatment failure according to my analysis. This reiterates the importance of including symptom-based variables in studies which seek to identify suitable predictors for predicting treatment failure for hospitalised exacerbations. Furthermore, none of the above studies compared various machine learning algorithms with optimisation of hyper-parameters and assessment of the best model from these steps on a separate test set. This highlights the strength of my analysis pipeline where such optimisation and rigorous testing of model performance is performed.

My finding that dyspnoea at exacerbation is a useful predictor of D30 and D90 treatment failure complements findings of other studies. Dyspnoea is already known to be a predictor of 5 year mortality for COPD patients[12]. Similarly, dyspnoea has been shown to be associated with exacerbation relapse risk in patients with acute exacerbations of COPD[130,131]. My finding that sputum purulence is a predictor of immediate rehospitalisation is interesting in the context of GOLD and NICE guidance that antibiotics should be given at exacerbation in patients with increased sputum purulence since this may reduce exacerbation relapse and treatment failure[6,132]. The relatively high prescription of antibiotics for asthma exacerbations in my analysis is not an uncommon occurrence, as has previously been shown[133]. This is likely due to patients not being seen by specialists in ED and reflects clinical practice, further emphasising the importance of developing tools to assist non-specialists in ED. My analysis revealed that sputum production and purulence at exacerbation were associated with patients receiving antibiotics at exacerbation. Therefore, it is unlikely that treatment failures in this study resulted from inappropriate or non- prescription of antibiotics at exacerbation. Rather, my finding that higher dyspnoea, sputum production and purulence are predictive of treatment failure may simply reflect that patients with greater disease burden and more severe exacerbations in terms of symptomatic presentation are more likely to fail treatment. It may also be the case that the patients with the highest symptom load always have a high symptom load, so the threshold to declare this is reached more frequently and treatment may not lower their symptoms enough for them to drop below their threshold. My conclusion that treatment failure prediction models should include exacerbation symptomatological presentation reaffirms the need for clinicians to pay close attention to patient symptom presentation at exacerbation. Furthermore, the ease of measuring patient symptoms through the

VAS score would make treatment failure prediction models like those in our study easy to implement in clinical practice.

Relationship between inflammatory markers, exacerbation treatment and treatment outcome

It is noteworthy that a high symptom burden for VAS sputum production and sputum purulence in addition to CRP were characteristic of patients given antibiotics in my analysis, although CRP did not reach statistical significance. It has been shown that using CRP as a biomarker to direct antibiotic treatment of severe COPD exacerbations is effective in achieving good clinical treatment outcomes as well as reducing antibiotic prescription[134]. The latter is especially important given the concerns associated with excessive antibiotic use for treating exacerbations promoting antibiotic resistance[135]. The only biological exacerbation characteristic which differed between patients given SCS and patients not given SCS was the degree of eosinophilic inflammation. Blood results would not have been available for clinicians prior to treatment initiation in this study. The association of SCS prescription with eosinophilic inflammation in this study and the higher levels (although not statistically significant) of CRP for patients prescribed antibiotics in this study may mask the relationship between biological indices of inflammation and treatment outcome. This may explain why symptoms, physiology, medication and medical history but not biological markers of inflammation could be identified as predictors of treatment outcome in my analysis. The association between eosinophilic inflammation and SCS prescription in this study is an interesting finding, as it suggests that there is something about eosinophilic inflammation that drives a clinician to prescribe SCS . In investigating this further, my univariate analysis showed that prescription of a LAMA and PPI was associated with reduced SCS

prescription. Wheeze was more frequently reported for participants associated with SCS prescription although statistical significance was not reached. However, there may be clinical significance to this result, which should be investigated further. Eosinophilic inflammation is related to increased luminal airway oedema and in asthma post-mortem narrowed airways, which could suggest that eosinophilic inflammation is related to wheeze and a greater degree of clinical severity of the presentation[136,137]. Studies have shown that patients with a high peripheral blood eosinophil count respond better to SCS than patients with a low blood eosinophil count[82] and my finding suggests the importance of measuring eosinophils in all exacerbations of airways disease.

Limitations of analysis

The analysis in my chapter has some limitations. A contextual limitation of my analysis may appear to be the combination of asthma and COPD which may seem to make interpretation difficult. However, as aforementioned this is a realistic clinical context since in standard practice there is not always access to lung function in emergency from primary care. In addition, where monitoring is impacted by the absence of spirometry in the community setting e.g due to the COVID pandemic in recent years, it is prudent to make the assumption that a clear-cut diagnosis of asthma or COPD is not always available. The combination of asthma and COPD can thus be seen as an advantage rather than a limitation. A second potential contextual limitation is that I did not perform exacerbation treatment prescription and treatment failure analyses for asthma compared to COPD patients. This is because out of the 101 patients, only 22 had a primary diagnosis of COPD, which would be too small a sample size for separate analysis. However, whether the primary diagnosis was asthma or COPD is a variable considered in my exacerbation treatment prescription and treatment failure prediction analyses and this variable was not selected by the multivariable SES approach. Hence,

I conclude that whether the diagnosis is asthma versus COPD is not related to SCS prescription, antibiotic prescription and treatment failure in the patient cohort analysed in this chapter.

A statistical limitation of my analysis is the small sample size. However, the use of the SES algorithm for multivariable feature selection also counters overfitting since this multivariable feature selection method is suitable even in high dimensional low sample size scenarios[100]. Therefore, features selected via this method are robust. Another useful property of SES algorithm is the indication of which variables if any are statistically equivalent to the each particular feature of the final feature set chosen by the algorithm[100]. For example, in my analysis, blood eosinophil count was indicated by the SES algorithm as being equivalent to blood eosinophil percentage in the multivariable context for predicting SCS prescription. Similarly, VAS sputum production and whether the participant reported that the exacerbation was associated with increased sputum purulence were statistically equivalent to VAS sputum purulence for predicting immediate hospitalisation. This provides flexibility for clinical context in which a particular variable may not be readily available but a statistically equivalent variable for the multivariable context of an exacerbation treatment prescription or treatment failure prediction task could instead be used. Small sample size can limit the performance of machine learning models. Powerful machine learning methods such as the random forest, SVM and neural network which I employed can undergo significant improvements in prediction performance as training data increases in size[38]. Small sample size can also lead to over-optimistic performance estimates of machine learning algorithms for predicting exacerbation treatment prescription and treatment failure[38]. However, my cross-validation and hold-out test set strategy was designed to

address this. In addition, although random forests can be effective in preventing overfitting, over-fitting is a risk for neural networks and SVM[38,44]. This is especially a risk in the low sample size high dimensional context of the data analysed in this chapter. However, this risk is accounted for by the use of a hold-out test set to assess performance of the best optimised machine learning algorithms based on cross-validation performance for predicting exacerbation treatment prescription and treatment failure at different time points. Although on one hand a small test set sample size could result in optimistic performance estimates of the models for unseen patients, a small test set could equally result in pessimistic estimates of model performance. An overfit machine learning model will not perform well on a hold-out test set of unseen data even if apparent performance based on cross-validation is high. My analysis remains the first in the treatment failure prediction field to compare and optimise a variety of different machine learning approaches and then assess performance of the final model on a separate test set.

Next steps

Following validation of the performance of the treatment failure prediction algorithms developed in my analysis in new patient cohorts, an important next step is the evaluation of the usefulness of prediction algorithms for different time points of treatment failure in helping to reduce treatment failure. This could be achieved via a prospective study in which exacerbating COPD and asthma patients who present to ED and are predicted to fail treatment receive greater clinician monitoring following discharge from the emergency department. Such monitoring could, for example, serve to increase LAMA, ICS and SABA use since lower use of these pharmacological treatments was associated with treatment failure in my analysis. However, the

importance of addressing psychological factors to reduce the risk of treatment failure should not be underestimated. Depression has been shown to be associated with increased exacerbation risk[138]. Depressive symptoms have also been shown to be associated with worse recovery following exacerbations. A noteworthy example is the Papaioannou et al prospective study of 230 patients hospitalised for acute exacerbations of COPD and without a previous diagnosis of depression[139]. Papaioannou et al's analysis showed that patients with depressive symptoms (evaluated using Beck's depression inventory completed on the first day of hospitalisation) required longer hospitalisation (mean±SD 11.6±3.7 versus 5.6±4.1 days, $p<0.001$)[139,140]. Furthermore, depressive symptoms on admission had a significant impact on dyspnoea ($p<0.001$) and CAT score ($p=0.012$) improvement[139]. Paradoxically, in my analysis a greater proportion of patients who were re-hospitalised within 5 days of initial exacerbation were in a lower number EuroQol anxiety and depression category than patients not re-hospitalised. In addition, interestingly, for none of the treatment outcome time points in my analysis were depression and anxiety as evaluated using the Hospital Anxiety and Depression (HAD) scale scores measured at the time of ED admission for exacerbation associated with treatment failure. However, it should be noted that Papaioannou et al's analysis focused on length of hospital admission, dyspnoea and CAT score improvements from day 1 to day 3, day 10 and day 40 whereas my analysis focused on re-hospitalisation immediately after exacerbation and on treatment within a 30 day and a 90 day timeframe from the time of exacerbation[139]. It also cannot be ruled out that differences in depressive symptoms in the days and weeks following the occurrence of exacerbation hospitalisation may play an important role in determining treatment outcome. Hence, I propose that future studies evaluating treatment outcome for exacerbating COPD and asthma patients

admitted to ED should include measures of depressive symptoms on a daily basis starting from the time of ED admission and continuing until at least 30 days. It would be pertinent to identify whether such time series data representing depressive symptoms may enable more accurate identification of patients likely to fail treatment as well as potentially highlighting ideal time points following initial exacerbation treatment during which heightened communication of healthcare staff with patients, perhaps with accompanying psychological interventions, may best reduce the risk of treatment failure.

3.5 Conclusion

I have shown that over half of all asthma and COPD patients admitted to ED for their exacerbation require additional major medication and/or readmission to the emergency department within 30 days of their exacerbation. My analysis suggests that prescription of SCS at exacerbation by physicians may be related to eosinophilic inflammation, potentially associated with symptoms such as wheeze. I have shown that non-linear classifier models using variables including past healthcare utilisation, medication usage and exacerbation dyspnoea can effectively predict treatment failure as defined by healthcare utilisation at different time points. In particular, re-hospitalisation within 5 days of the initial exacerbation and healthcare utilisation treatment failure by D30 and D90 after initial exacerbation can be predicted with AUC of 0.68, 0.71 and 0.77 respectively in a hold-out test set of patients. In the next chapter of my thesis, I seek to newly define COPD exacerbations to enable an objective symptom- and inflammation-based foundation for identifying exacerbations and characterising treatment response.

Chapter 4 – Defining the mathematical equilibrium and crisis states in chronic obstructive pulmonary disease

Abstract

Management of exacerbations of COPD is hampered by the subjective definition currently used in the clinic. In this chapter, I mathematically define objective COPD states to replace existing ‘exacerbation’ and ‘stable’ definitions. I identify the following states: **low overall symptom burden state** representing **equilibrium** with low symptoms and inflammation; **symptomatic crisis state** with high symptoms but low inflammation; **eosinophilic crisis state** with high symptoms accompanied by eosinophilic inflammation; **neutrophilic crisis state** with high symptoms accompanied by neutrophilic inflammation. These states also differ in levels of serum and sputum cytokines corresponding to neutrophilic- and eosinophil-predominant immune responses. Using VAS cough, dyspnoea, sputum production and sputum purulence scores, blood CRP, blood neutrophils, blood eosinophils and blood eosinophil percentage, I use a dataset of COPD patients to develop and compare a variety of classification algorithms to determine whether a patient at a given time point is in the low overall symptom state or a particular crisis state. Using a hold-out test group of patients, I demonstrate that the COPD state detection algorithm can accurately assign patients at a given time point to the correct state. I also validate the COPD state detection algorithm in a separate patient cohort using the ED patient cohort from Chapter 3. My new COPD state definitions and state assignment algorithm pave the way for tackling both underdiagnosis of COPD crises as well as enabling more accurate treatment personalisation.

4.1 Introduction

Unfortunately, patients with COPD remain to have poor outcomes following exacerbations[49]. The GOLD definition is[6]: *'an exacerbation of COPD is defined as an acute worsening of respiratory symptoms that results in additional therapy*'. This definition is subjective, lacks any indication of what magnitude of symptoms characterises an exacerbation and does not take into account any biological markers. This is a substantial limitation since exacerbations of COPD are heterogeneous with respect to biology[60]. Various proposals have been made for defining 'exacerbations' including the ROME proposal discussed in Chapter 1 of my thesis[56]. However, as previously discussed, these proposals are either directly or indirectly still based on/influenced by current exacerbation definitions. The concept of a 'crisis' has been suggested as 'a point in time in a new situation when identified stresses threaten to exceed the ability to cope and there is a possibility of suddenly getting better or worse'[141]. The 'crisis' concept takes into account the temporal context of disease worsening[141]. However, to-date a fully objective, empirically derived definition of what constitutes an exacerbation and stable state is lacking. An empirically derived, mathematically based definition has the potential to identify events which constitute true crises based on symptomatological and biological deviation from the norm without any reliance on current definitions of exacerbation versus stable state. In this chapter, I seek to quantitatively define objective COPD states, using data variables from previously published studies[60,82]. I also develop an algorithm to identify the COPD state a patient may be exhibiting at any given time point based on symptomatological and biologically variables easily measurable in the clinic.

The aim of this chapter is to quantitatively redefine exacerbations of COPD, taking into account biological and symptom-based heterogeneity.

4.2 Methods

Dataset

I analysed a dataset derived from two previously published studies[60,82]. These studies were observational, with data having been prospectively collected for 12-36 months. Both exacerbations and stable state measurements were included, defined according to physician diagnosis and national guidelines[142]. Stable state measurements were either from study baseline or from longitudinal 3-month follow-up timepoints. For a study measurement to count as stable, at least 6 weeks had to pass after any exacerbation before the measurement was made. Exacerbation measurements and data collection from participants were conducted before treatment with SCS and/or antibiotics[60,82]. All participants had provided informed consent and the studies were approved by the East Midlands Leicester Central ethics committee.

Data collection

In order to provide context for the dataset which I analysed in this chapter, I will outline data collection in the studies from which the dataset I analysed was derived. Baseline demographic data were available for all patients, as well as baseline data on comorbidities and medication usage. At all study visits made by patients, lung function measurements were made[60,82]. These included post-bronchodilator FEV_1, FVC and percentage predicted FEV_1[60,82]. Patient self-reported data included the VAS[117] to record symptoms of cough, dyspnoea, sputum production and sputum purulence in diary cards, measured on a 100mm line. Venous blood samples were collected from participants and analysed for CRP and blood cell counts . Serum cytokine data was

obtained from mesoscale discovery platform-based measurements. In addition, sputum samples were collected and similarly analysed for cytokines and cell counts. Sputum data also included bacteriology (colony forming units and qPCR for Staphylococcus aureus, Streptococcus pneumoniae, Haemophilus influenzae and Moraxella catarrhalis) and virology (presence or absence of virus based on reverse transcription PCR). Table 4.1 summarises the data variables which I analysed in this chapter.

Table 4.1 | Summary of data variables analysed in this chapter.

Demographics, comorbidities and medication	Clinical and questionnaires	Physiology	Blood tests	Sputum tests
Age	VAS[117] cough	FEV_1 post-bronchodilator	Total blood cell count	Sputum total cell count
Pack Year History	VAS[117] dyspnoea	Percentage predicted FEV_1	Blood neutrophils	Sputum neutrophils
Sex	VAS[117] sputum production	FEV_1 post-bronchodilator/FVC ratio	Blood eosinophils	Sputum eosinophils
Smoking status	VAS[117] sputum purulence	FE_{NO}	Blood basophils	Sputum lymphocytes
Cardiovascular comorbidity	CRQ[143] emotion	Haemoglobin	Blood monocytes	Sputum macrophages
Diabetes	CRQ[143] fatigue	Platelets	Blood lymphocytes	Sputum neopterin
Endocrine comorbidity	CRQ[143] mastery	Temperature	CRP	Sputum IL1B
Depression	CRQ[143] dyspnoea	Heart rate	Creatine	Sputum IL6R
Osteopenia/Osteoporosis	SGRQ[144] symptoms	Oxygen saturation	Urea	Sputum IL8
Anaemia	SGRQ[144] activity	Systolic blood pressure	Procalcitonin	Sputum CXCL10
SABA	SGRQ[144] impacts	Diastolic blood pressure	Surfactant protein D	Sputum CCL4
SACH		eGFR	Blood neopterin	Sputum TNFA
LACH			Serum IL1B	Sputum TNFB
LABA			Serum IL8	Sputum VEGFA
ICS			Serum CXCL10	Sputum IL5
Maintenance prednisolone			Serum CCL4	Sputum IL6
Prophylactic antibiotic			Serum TNF1A	Sputum CXCL11
Theophylline			Serum TNF1B	Sputum CCL2
Mucolytic			Serum VEGFA	Sputum CCL5
Oxygen			Serum IL13	Sputum CCL17

For exacerbation measurements, whether SCS given at exacerbation			Serum IL5	Sputum TNFalpha
For exacerbation measurements, whether antibiotics given at exacerbation			Serum IL6	Sputum CCL13
			Serum CXCL11	Sputum CCL3
			Serum CCL2	Sputum MMP8
			Serum CCL17	Sputum MMP9
			Serum TNFalpha	Sputum MMP1
			Serum IFNgamma	Sputum MMP2
			Serum IL10	Sputum MMP3
			Serum CCL3	Sputum MMP7
			Serum CCL13	Sputum MMP12
			Serum Amyloid A_1	sTREM
				Viral presence
				Colony Forming Units
				HI positive
				MC positive
				SA positive
				SP positive

Definition of abbreviations: SABA: short acting beta agonist; LABA: long acting beta agonist; SACH; short acting anticholinergic; LACH = long acting anticholinergic; SGRQ: St George's Respiratory Questionnaire; CRQ: Chronic Respiratory Disease Questionnaire; eGFR: estimated glomerular filtration rate; FEV_1 : forced expiratory volume after 1 second; FVC: forced vital capacity; HI: Haemophilus influenzae; SA: Staphylococcus aureus; SP: Streptococcus pneumoniae; MC: Moxarella catarrhalis; FE_{NO} : Fraction of exhaled nitric oxide; SCS = systemic corticosteroids; VAS = Visual Analogue Scale.

Statistical analysis

Univariate analyses were performed using 'R' and Prism Version 9 for macOS[86,123]. Normally distributed data were presented as mean (standard error of mean). The Shapiro-Wilk test revealed that the vast majority of variables did not follow a normal distribution. Such variables were log-transformed but if the log-transformed data was likewise not normally distributed, I presented data for hypothesis tests as median (interquartile range). Categorical data are shown as percentages of instances belonging to a specified category. Appendix Table A4.1 shows the results of the Shapiro-Wilk test for each numerical variable and where relevant the log-transformed variable. For two-class comparisons, I used the wilcox test. For multiclass comparisons, I used either the Kruskall-Wallis test with post-hoc Dunn test (continuous data) or chi-squared test (binary/categorical data). Since my analysis was exploratory, p-values correspond to two-sided hypothesis tests. The focus of the analyses was hypothesis generation, therefore, a multiple comparison correction was not applied in this context. When interpreting results of this chapter in the Discussion section, the focus will be on variables which are both statistically and potentially also clinically significant.

In order to minimise bias from imputation, missing data imputation was not performed prior to hypothesis testing but only where necessary before machine learning methodologies could be applied. The imputation methodology employed the missForest algorithm is described in Chapter 2.

In order to define objective COPD states, I first performed PCA on symptoms, blood and sputum inflammatory indices to identify principal components which are

independent from each other and explain most of the variance of the original data. The PCA and clustering were performed on the total dataset (irrespective of whether a set of measurements were labelled as ‘exacerbation’ or ‘stable state’ (baseline measurements and longitudinal 3 monthly follow-up measurements for participants were included) according to current clinical definitions using just the selected principal components. The first 5 principal components representing the original variables for cluster analysis were chosen since these were able to explain over 75% cumulative variance. I employed an objective, label-free approach of hierarchically clustering the measurements using the selected 5 principal components selected. I performed clustering using the R function ‘hclust’ (Euclidean distance metric with methods ‘average linkage’, ‘weighted average linkage’, ‘Ward’s method’, ‘centroid linkage’ and ‘median linkage’), a versatile clustering function from the R package ‘stats’[86]. I produced a dendrogram of the clustering results using the R package ‘dendextend’, a package which enables visualisation of hierarchical trees[145]. Due to the statistical considerations discussed in Chapter 2, I used visual inspection of the dendrogram and the use of the Duda and Hart index as well as the pseudo-t^2 index to identify a suitable cut-point to choose clusters to represent the newly defined COPD equilibrium and crisis states[95]. Different numbers of clusters from 2 to 16 were evaluated (16 was the maximum number of clusters considered representing possible combinations of high versus low values of overall symptoms, eosinophil levels, neutrophil levels and CRP). The most stable clustering solution according to the Duda and Hart index is the smallest number of clusters such that index > critical values[95]. The most stable clustering solution according to the pseudo-t^2 index is the smallest number of clusters such that index < critical values[95]. To decide between the methods ‘average linkage’, ‘weighted average linkage’, ‘Ward’s method’, ‘centroid linkage’ and ‘median linkage’, I

compared all methods for which the number of optimal clusters was greater than or equal to 3 since the dataset included stable state and exacerbation measurements, with the latter containing at least two different inflammatory subgroups as known from previous studies[95]. Hence, a method which detects only 2 clusters would not have been appropriate for defining objective COPD state to reflect symptomatological and inflammatory heterogeneity. I named the newly defined objective states as follows: low overall symptom burden state defined as the cluster with both low symptoms and low inflammation; a crisis state was defined as any high symptom cluster with or without either neutrophilic or eosinophilic inflammation. Depending on context, the low overall symptom state may represent a true ‘equilibrium’ in which a patient is asymptomatic or may represent a crisis in which overall symptoms are low although an individual symptom may be high.

I sought to visualise the overlap between current definitions of exacerbations and stable state in contrast to the separation between my newly defined COPD states. Therefore, I plotted data points in 3-dimensional space with principal components representing symptoms and blood inflammatory indices as axes. For PCA of symptom and blood variables, I chose the first 3 principal components representing the original variables for cluster analysis since this enables visualisation. I used the ‘scatter3d’ function from the ‘car’ and ‘rgl’ packages in ‘R’ to plot data since this function enables 3-dimensional visualisation with colour coding for specified distinct subgroups, allowing my newly defined COPD states to be easily visualised in 3 dimensions[146,147]. Concentration ellipsoids for exacerbation versus stable state and for the newly defined COPD states were plotted to aid visualisation.

In order to provide a measure of the extent of overlap of the 3-dimensional symptom- and blood inflammatory index-based principal component exacerbation versus stable state distributions and of the equilibrium and crisis states with each other, I employed the Degree of Overlap (DO) methodology[148]. This was the only suitable measure of overlap which I identified as yielding both mathematically and intuitively consistent results from empirical studies[148]. I used the 'crossmatchtest' function from the 'R' package 'crossmatch' which was designed to enable crossmatch calculations. Degree of Overlap (DO) between two groups is calculated as follows[148,149]:

$$DO = \min\left[1, \frac{n(n_1 - m_1)}{n_1 n_2}\right]$$

Where n (= $n_1 + n_2$) is the total number of data points under consideration for the groups being considered for overlap, n_1 is the number of data points in the first group, n_2 is the number of data pints in the second group and $n_1 - m_1$ (= $n_2 - m_2$) is the number of crossmatches. $n_1 - m_1$ was calculated using the function 'crossmatchtest' from the 'R' package 'crossmatch'[149]. $0 \leq DO \leq 1$ where DO = 0 indicates no overlap and DO = 1 indicates complete overlap of a pair of multivariate distributions[148].

Supervised machine learning analyses to distinguish between COPD states

In this chapter, I first compared the suitability of each individual VAS score and the blood inflammatory indices (CRP, neutrophil and eosinophil cell counts and percentages) by assessing the AUC for each individual variable to distinguish between the low overall symptom burden state and a crisis state. In keeping with the methodology described in Chapter 2 and applied in Chapter 3, the first step when developing objective COPD state prediction algorithms was the use of the 'R' package

'caret' to separate the participant data (number of participants depending on type of treatment outcome as described above) into a hold-out test set including 20% of participants and a training and validation portion containing 80% participants[86,101]. I ensured that in the case of multiple events from the same patient, in a given training and validation fold split all events from a given patient occur in either the test set of the 80% for training/validation but never in both test and training/validation. In order to comply with this criterion, the division of events into training/validation and testing corresponded to approximately 80% and 20% respectively but could not be exactly 80% and 20%. ROC curves were produced (using 80% of available participant measurements) for each biomarker associated with a particular state. Based on AUC for the ability of each biomarker to distinguish a particular crisis state from the low overall symptom burden state in 80% of participant measurements, I chose the most discriminative biomarker for each crisis state. I determined a suitable threshold value for the biomarker using the value corresponding to the top left-hand corner for the ROC curve. I evaluated the sensitivity and specificity of the chosen biomarker and threshold value for accurately distinguishing the crisis state from the low overall symptom burden state for the remaining 20% participant measurements which represented unseen data.

In order to distinguish any of the newly defined COPD states from each other rather than distinguishing just a particular crisis state from the low overall symptom burden state, I next developed a variety of supervised machine learning models using a combination of different symptom and blood inflammatory indices. The machine learning models I used (the rationale behind and mathematics of which I have outlined in Chapter 2) included:

-Decision tree

-Random forest

-KNN

-SVM

-Neural network

Using the 80% of participant data, I selected variables using the SES algorithm with a $p<0.1$ threshold as discussed in preceding chapters[100]. I divided the 80% portion into four folds as part of four-fold cross-validation. In this procedure, each machine learning algorithm was trained on three of the folds and validated on the fourth. The average cross-validation performance across all four folds is then calculated for the algorithm. The cross-validation performance of a given machine learning algorithm type is used to select the best hyper-parameter values for each algorithm. I used mean balanced accuracy as the metric to evaluate classification algorithm performance since in a multiclass context of more than two COPD states this enables algorithm performance to take into account all COPD states even in the face of data imbalance. Finally, I identified the best machine learning algorithm with optimal hyper-parameters, retrained this model on the entire 80% training and validation portion and then assessed the retrained model on the 20% hold out test set. The hold out test set performance of this model provides a generalisable estimate of the ability of the machine learning model to predict which objective COPD state a new patient is in at a particular point in time based on symptoms and inflammatory biology.

In addition, in order to explore the performance of the objective disease state prediction algorithm for a different patient cohort in a different clinical context, I retrained the optimal COPD state prediction algorithm on all participant data from the Chapter 4 datasets and applied the algorithm to the ED mixed COPD and asthma dataset from

Chapter 3. In the context of the ED Chapter 3 dataset, a low overall symptom burden state represents a crisis in which overall symptoms are low which may include events with a high VAS score for an individual symptom but low overall scores when combining symptoms. I determined which of the 101 participant measurements of the ED Chapter 3 dataset is assigned to each type of crisis state. I then compared the VAS symptom scores, blood CRP, blood cell counts and percentages across the different crisis states to assess whether the pattern of differences between the characteristics of participants assigned to different states in the Chapter 3 dataset is as expected based on the differences between COPD states in the Chapter 4 COPD dataset.

When running code, I always set the programme ‘seed’ to a constant value in order to ensure consistent sample selection and shuffling when selecting participants for training and validation versus testing and producing different cross-validation folds. This is important for fair comparison of performance of different machine learning algorithms and for reproducibility of results.

Code

Code for machine learning analysis is provided in Appendix Chapter 4 Code.

4.3 Results

Number of measurements analysed and characteristics of dataset

The dataset included 218 participants. The median age was 69 years (63 to 76 years) and 142 participants (65%) were male. 35% of participants were smokers at study entry, with an average pack year history of 46 years (33 to 60 years). From the 218 participants, 1,384 measurements were made including 352 exacerbations and 1,032 at stable state.

For defining objective COPD states it was important for each measurement to have at least one non-missing data value for one of the four VAS symptoms and at least one non-missing data value for a measure of blood/sputum based neutrophilic or eosinophilic inflammation (cell count/percentage or CRP). 1,348 of the 1,384 measurements fulfilled this criterion and were analysed further in this chapter. 347 measurements were at exacerbation and 1,001 measurements were at stable state.

Exacerbations were associated with reduction in lung function, greater symptom burden and greater inflammation compared to stable state

Age and pack year history did not differ between exacerbation and stable state (Table 4.2). FEV_1 post-bronchodilator was lower in exacerbations (0.96L) compared to stable state (1.25L) ($p<0.01$). Similarly, FEV_1 percentage of predicted was lower in exacerbations ($p<0.01$). VAS cough, dyspnoea, sputum production and sputum purulence were all greater at exacerbation than at stable state ($p<0.01$ in all cases). Exacerbations showed greater neutrophilic inflammation as shown by blood neutrophil cell counts (6.0 versus $5.2x10^9$cells/L, $p<0.01$) as well as sputum neutrophil cell counts (6.6 versus $2.1x10^3$ /mg, $p<0.01$) and CRP (10.0 versus 2.5mg/L, $p<0.01$). Overall

blood and sputum cell count, as well as blood and sputum neutrophil percentage were also greater at exacerbation (p<0.01 in all cases).

Table 4.2 | Participant demographic, physiological and inflammatory biological characteristics at stable state compared to exacerbation.

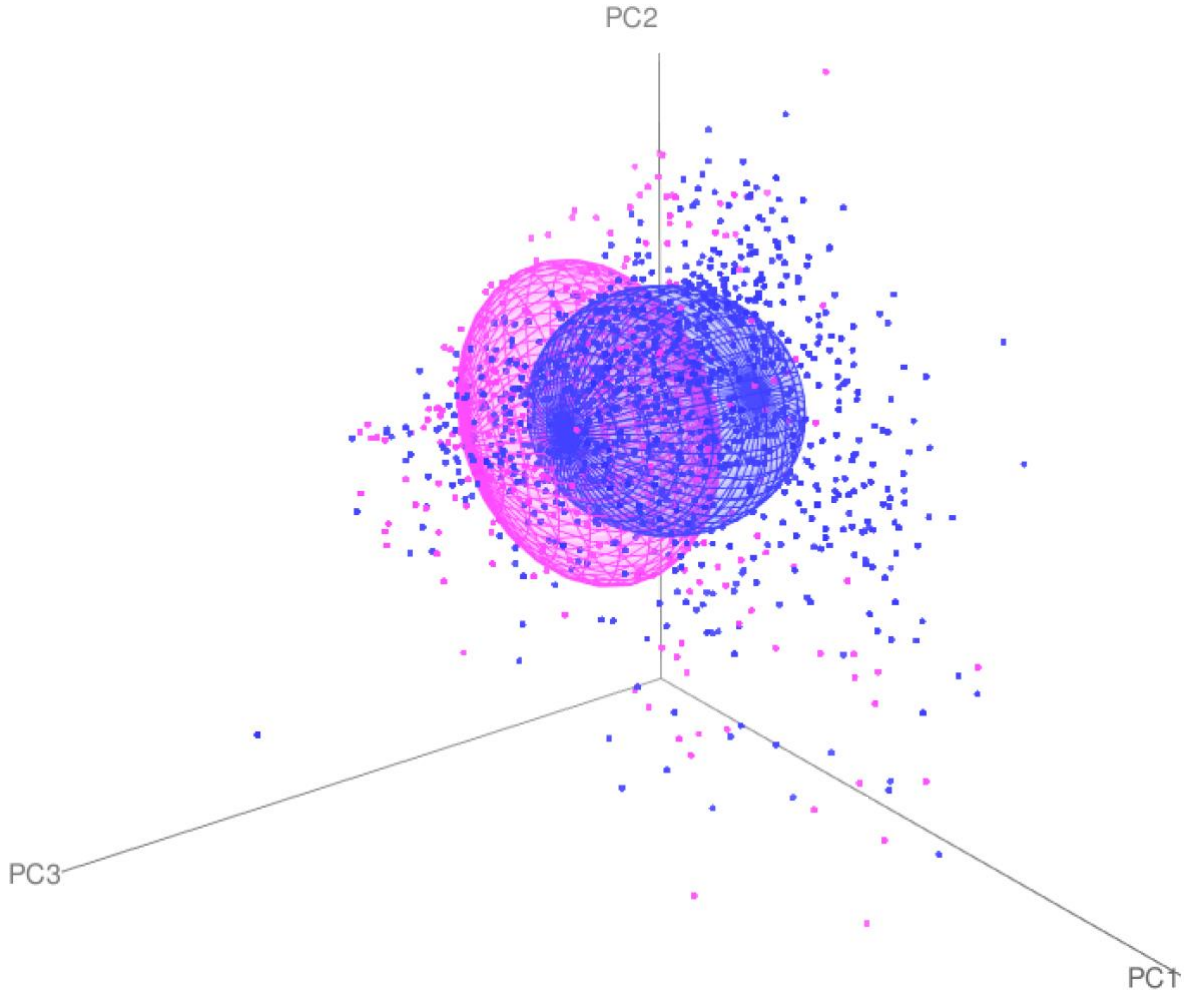
Characteristic	Stable state (n=1,001)	Exacerbation (n=347)	P-Value
Age (Years)	70 (64 – 77)	68 (63 – 77)	0.18
Pack Year History	47 (33 – 62)	46 (33 – 58)	0.37
FEV_1 post-bronchodilator (L)	**1.25 (0.90 - 1.74)**	**0.96 (0.70 - 1.37)**	**<0.01**
Percentage predicted FEV_1 (%)	**52 (36 – 65)**	**39 (27 – 56)**	**<0.01**
FEV_1 post-bronchodilator/FVC Ratio	0.51 (0.41 – 0.62)	0.49 (0.39 – 0.62)	0.10
VAS cough (mm)	**30 (11 – 52)**	**65 (45 – 78)**	**<0.01**
VAS dyspnoea (mm)	**49 (25 – 67)**	**74 (60 – 85)**	**<0.01**
VAS sputum production (mm)	**28 (10 – 52)**	**60 (35 – 82)**	**<0.01**
VAS sputum purulence (mm)	**24 (7 – 50)**	**59 (33 – 77)**	**<0.01**
Total blood cell count ($x10^9$cells/L)	**8.2 (6.8 - 9.6)**	**9.2 (7.8 - 11.3)**	**<0.01**
Blood neutrophils count ($x10^9$cells/L)	**5.2 (4.2 - 6.3)**	**6.0 (5.0 - 8.0)**	**<0.01**
Blood eosinophil count ($x10^9$cells/L)	0.2 (0.1 – 0.3)	0.2 (0.1 – 0.3)	0.44
Percentage blood neutrophils (%)	**64.6 (58.4 – 70.5)**	**68.6 (62.0 – 74.6)**	**<0.01**
Percentage blood eosinophils (%)	**2.5 (1.6 – 4.0)**	**2.1 (1.3 – 3.7)**	**<0.01**
Blood CRP (mg/L)	**2.5 (2.5 – 9.0)**	**10.0 (2.5 – 32.0)**	**<0.01**
Total sputum cell count ($x10^3$ cells/mg)	**3.0 (1.2 – 6.9)**	**8.4 (2.2 – 21.1)**	**<0.01**
Sputum neutrophil count ($x10^3$ cells/mg)	**2.1 (0.7 – 5.7)**	**6.6 (1.3 – 19.1)**	**<0.01**
Sputum eosinophils count ($x10^3$ cells/mg)	**0.0 (0.0 – 0.1)**	**0.1 (0.0 – 0.2)**	**<0.01**
Percentage sputum neutrophils (%)	**75.1 (55.1 – 89.7)**	**86.6 (67.0 – 94.8)**	**<0.01**
Percentage sputum Eosinophils (%)	1.0 (0.3 – 2.8)	0.6 (0.3 – 2.3)	0.05

Definition of abbreviations: CRP = C-reactive protein.
Measurements: variables are represented as median (lower quartile to upper quartile).
P-Values: All p-values are wilcox p-values. Rows in bold font correspond to variables differing significantly across states according to p<0.05.
Correction for multiple comparisons is not made since the analysis is hypothesis-generating.

Lack of multivariable mathematical separation of symptoms and inflammation in conventional exacerbation compared to stable state definition

The mathematical validity of the conventional exacerbation and stable state definitions depends on the multivariable separation of exacerbation versus stable state rather than only univariate differences. I therefore sought to assess the extent to which conventional exacerbation versus stable state labels correspond to a multivariable difference in symptoms and inflammatory biology. The inherent limitations of the conventional exacerbation and stable state clinical definitions are illustrated in Figure 4.1 which shows a high degree overlap of exacerbations and stable state data points (DO = 0.67) in 3-dimensional principal component space representing symptoms and blood inflammatory indices. The mathematical relationship between all principal components and the original symptomatological and blood inflammatory indices is shown is Appendix Table A4.2 and Appendix Table A4.3.

Figure 4.1 | Exacerbation and stable state events and concentration ellipsoids visualised in three-dimensional principal component space representing symptoms and blood inflammatory indices. Exacerbations shown in pink; stable state shown in dark blue. PC1 mostly reflects symptom variation (66% of PC1 is explained by symptoms; see **Appendix Table A4.3**), with events represented by data points further away from the axis origin showing lower symptoms. PC2 mostly reflects variation in eosinophilic inflammation (66% of PC2 is explained by eosinophilic inflammation; see **Appendix Table A4.3**), with events represented by data points further away from the axis origin showing lower eosinophilic inflammation. PC3 mostly reflects variation in neutrophilic inflammation (60% of PC3 is explained by neutrophilic inflammation as represented by CRP, neutrophil cell count and percentage; see **Appendix Table A4.3**), with events represented by data points further away from the axis origin showing lower neutrophilic inflammation.



Four mathematically defined objective COPD states exist in terms of symptoms and inflammatory biology

My first objective was to objectively define COPD disease states without reliance on conventional clinical labels of exacerbation and stable state. The mathematical relationship between all principal components and the original symptomatological, blood and sputum inflammatory indices is shown is Appendix Table A4.4. The Duda and Hart index and pseudo-t^2 index showed that a solution of 4 clusters represented the most mathematical stable clustering using Ward's method. Other methods besides Ward's only resolved 2 clusters and were not analysed further as discussed in Methods. Appendix Figure A4.1 shows a dendrogram of the relationship between the 1,348 events, where I identified 4 clusters. The 4 clusters represent 4 COPD states (Table 4.3), which differed across all symptoms as measured by VAS, SGRQ and CRQ scores ($p<0.01$ in all cases):

Low overall symptom burden state representing **equilibrium** includes 489 events and is characterised by low overall symptoms (VAS cough, dyspnoea, sputum production and sputum purulence of 17, 35, 14 and 11mm respectively), low inflammation and the highest lung function indices. This state may represent a condition of **equilibrium** in which a patient is free of all of cough, dyspnoea, sputum production and sputum purulence. Alternatively, this may represent a **low overall symptom burden crisis** in which a patient has a low combined burden of different symptoms but may still have a high burden of an individual symptom.

Symptomatic crisis includes 488 events and is characterised by high overall VAS symptoms (VAS cough, dyspnoea, sputum production and sputum purulence of 62, 67, 58 and 52mm respectively) but low inflammation and preserved lung function indices.

Eosinophilic crisis includes 277 events and is characterised by high VAS dyspnoea (52mm) and low to moderate VAS cough, sputum production and purulence (30, 25 and 27mm respectively), high eosinophilic inflammation indices and low colony-forming units.

Neutrophilic crisis includes 94 events and is characterised by high VAS cough, dyspnoea, sputum production and purulence (VAS cough, dyspnoea, sputum production and sputum purulence of 60, 72, 63 and 63mm respectively)**,** high CRP and high neutrophilic inflammation indices.

Objective COPD states differ in blood and sputum neutrophilic and eosinophilic inflammation and microbiology

Symptomatic crises had a comparable neutrophilic and eosinophilic inflammatory profile to the low overall symptom burden state. Neutrophilic crises were characterised by high blood and sputum neutrophil count compared to the low overall symptom burden state, symptomatic crises and eosinophilic crises (9.8 compared to 5.6, 5.2 and 4.9x10^{9}cells/L respectively, $p<0.01$; 32.9 compared to 2.6, 2.1 and 2.1x10^{3} cells/mg respectively, $p<0.01$), as well as high CRP levels (47.0mg/L in neutrophilic crises compared to 2.5mg/L for all other states, $p<0.01$). Neutrophil percentage in blood and sputum was also greatest for neutrophilic crises ($p<0.01$). Neutrophilic crises had higher levels of procalcitonin than the low overall symptom burden state, symptomatic crises and eosinophilic crises (0.09 versus 0.06, 0.06 and 0.05 ng/mL respectively, $p<0.01$). Neutrophil crises had greater sputum bacterial load as determined by colony forming units compared to the other states ($p<0.01$). However, the presence of particular types of bacteria and the presence of viruses did not differ significantly between states. Eosinophilic crises were characterised by high blood and sputum

eosinophil count compared to the low overall symptom burden state, symptomatic crises and neutrophilic crises (0.48 compared to 0.17, 0.17 and 0.14 x10⁹cells/L respectively, p<0.01; 0.2 compared to 0.0, 0.0 and 0.1x10^3 cells/mg respectively, p<0.01), as well as higher eosinophil percentage in blood and sputum (p<0.01). Eosinophilic crises were associated with higher FE_{NO} levels (26 parts per billion in eosinophilic crises compared to 16, 15 and 14 parts per billion respectively, p<0.01).

Objective COPD states differ in blood and sputum cytokine biology

Neutrophilic crises were associated with pro-inflammatory cytokine markers. Neutrophilic crises had higher levels of serum TNFA than the other states (p=0.02) and higher levels of serum IL6 and serum amyloid A_1 (p<0.01 in both cases). Sputum neopterin, IL1B, IL6R, IL8, TNFA, TNFB, TNFalpha and VEGFA were higher in neutrophilic crises compared to the other states. Both neutrophilic crises and symptomatic crises had higher levels of sputum CCL5 than eosinophilic crises and the low overall symptom burden state (p<0.01). Neutrophilic crises had the lowest levels of sputum CCL2 (p<0.01).

Serum IFNgamma was higher in eosinophilic compared to neutrophilic crises (p<0.01). Sputum CXCL10, CXCL11 and serum CXCL11 were higher in eosinophilic and symptomatic crises compared to neutrophilic crises (p<0.01, p<0.01 and p=0.02) but not compared to the low overall symptom burden state. Eosinophilic crises had higher serum CCL2 (p=0.04) and sputum CCL17 (p<0.01) than symptomatic and neutrophilic crises but not compared to the low overall symptom burden state. Sputum CCL13 was greater in eosinophilic crises than in symptomatic or neutrophilic crises (p<0.01).

Eosinophilic crises were differentiated by all other states with higher serum and sputum IL5 ($p<0.01$ in both cases).

Objective COPD states differ in lung function, physiology and baseline medication usage but not associated comorbidities

The states differed from each other with $p<0.01$ for all FEV_1 post-bronchodilator and percentage predicted FEV_1, with symptomatic crises and the low overall symptom burden states having comparable lung function but the eosinophilic and neutrophilic crises lower lung function. The proportion of males and smoking status did not differ between COPD states. Although 62% of participants with neutrophilic crises had cardiovascular comorbidity compared to 43% of participants with eosinophilic crises, the difference in cardiovascular and other comorbidities between COPD states did not reach statistical significance. The only baseline medication whose usage by participants differed between states was maintenance prednisolone ($p=0.04$). Neutrophilic crises in particular had the greatest proportion of maintenance prednisolone usage (15%). Neutrophilic crises had a lower estimated glomerular filtration rate ($p=0.02$) and lower blood oxygen saturation ($p<0.01$) compared to other COPD states.

Table 4.3 | Comparison of demographics, physiology, symptom burden, inflammatory biology, microbiology, ongoing medication usage and comorbidity across the 4 objective COPD states.

Characteristic	Low overall symptom burden state (n=489)	Symptomatic crisis (n=488)	Eosinophilic crisis (n=277)	Neutrophilic crisis (n= 94)	P-Value
Age (Years)	70 (64 – 77)	68 (63 – 77)	69 (63 – 76)	69 (65 – 76)	0.42
Pack Year History	46 (29 – 60)	46 (34 – 60)	47 (37 – 62)	47 (25 – 60)	0.39
FEV_1 post-bronchodilator (L)	**1.20 (0.85 – 1.71)**	**1.19 (0.85 – 1.61)**	**1.20 (0.80 – 1.74)**	**0.85 (0.64 – 1.20)**	**<0.01**
Percentage predicted FEV_1 (%)	**52 (37 – 66)**	**47 (34 – 62)**	**50 (31 – 62)**	**36 (29 – 52)**	**<0.01**
FEV_1 post-bronchodilator/ FVC ratio	0.52 (0.41 – 0.62)	0.50 (0.40 – 0.61)	0.50 (0.38 – 0.60)	0.50 (0.40 – 0.63)	0.35
VAS cough (mm)	**17 (5 – 31)**	**62 (46 – 75)**	**30 (9 – 60)**	**60 (45 – 72)**	**<0.01**
VAS dyspnoea (mm)	**35 (19 – 56)**	**67 (52 – 80)**	**52 (19 – 74)**	**72 (55 – 84)**	**<0.01**
VAS sputum production (mm)	**14 (4 – 28)**	**58 (42 – 75)**	**25 (7 – 56)**	**63 (45 – 79)**	**<0.01**
VAS sputum purulence (mm)	**11 (3 – 27)**	**52 (34 – 69)**	**27 (7 – 55)**	**63 (48 – 83)**	**<0.01**
SGRQ symptoms (Units)	**51.2 (39.0 – 64.7)**	**73.9 (59.9 – 84.3)**	**56.3 (35.3 – 74.4)**	**72.4 (63.0 – 80.7)**	**<0.01**
SGRQ activity (Units)	**66.3 (53.5 – 79.7)**	**79.1 (61.1 – 86.8)**	**66.3 (48.4 – 85.7)**	**70.2 (60.2 – 80.4)**	**<0.01**
SGRQ impacts (Units)	**30.3 (18.5 – 42.3)**	**42.8 (29.5 – 56.8)**	**31.6 (20.1 – 45.3)**	**40.8 (29.7 – 56.1)**	**<0.01**
CRQ emotion (Units)	**4.86 (3.86 - 5.71)**	**4.00 (3.14 - 5.29)**	**4.43 (3.14 - 5.71)**	**3.57 (2.57 - 4.86)**	**<0.01**
CRQ fatigue (Units)	**3.75 (3.00 - 4.75)**	**3.00 (2.25 - 4.00)**	**3.50 (2.50 – 4.50)**	**2.25 (1.50 - 3.25)**	**<0.01**
CRQ mastery (Units)	**5.50 (4.25 - 6.25)**	**4.25 (3.00 - 5.75)**	**4.75 (3.25 - 6.25)**	**4.13 (2.75 - 5.25)**	**<0.01**
Blood creatine (µmol/L)	76.0 (65.0 - 93.0)	76.0 (66.0 - 89.0)	79.0 (66.5 - 96.0)	78.0 (67.0 - 92.5)	0.22
Blood urea (mmol/L)	5.7 (4.5 - 6.9)	5.5 (4.2 - 6.9)	5.6 (4.7 - 7.0)	5.9 (4.9 - 7.6)	0.10
eGFR (mL/min)	**85 (70 - 100)**	**89 (72 - 100)**	**85 (69 - 100)**	**83 (67 - 100)**	**0.02**
FE_{NO}* (parts per billion)	**16 (14 – 18)**	**15 (13 – 17)**	**26 (20 – 32)**	**14 (8 – 23)**	**<0.01**
Temperature (°C)	36.6 (36.5 – 36.9)	36.7 (36.6 - 36.9)	36.7 (36.6 - 36.8)	36.8 (36.6 - 37.1)	0.27
Heart rate* (beats/minute)	87 (82 – 93)	87 (84 – 91)	93 (88 – 98)	95 (89 – 101)	0.05
Oxygen saturation (%)	**95 (93 - 97)**	**96 (94 – 97)**	**95 (94 - 97)**	**93 (90 – 96)**	**<0.01**
Systolic blood pressure~ (mmHg)	134 (5)	133 (2)	138 (3)	130 (3)	0.20
Diastolic blood pressure ~(mmHg)	80 (2)	78 (2)	81 (2)	77 (3)	0.61
Blood cell count (x10^9cells/L)	**8.5 (7.1 - 10.1)**	**8.0 (6.9 - 9.3)**	**8.2 (6.8 - 9.8)**	**12.1 (10.1 - 15.4)**	**<0.01**

Blood neutrophil count (x10⁹cells/L)	**5.6 (4.4 – 6.9)**	**5.2 (4.4 – 6.1)**	**4.9 (3.8 – 6.0)**	**9.8 (7.4 – 12.5)**	**<0.01**
Blood eosinophil count (x10⁹cells/L)	**0.17 (0.12 - 0.26)**	**0.17 (0.11 - 0.24)**	**0.48 (0.38 - 0.62)**	**0.14 (0.08 - 0.24)**	**<0.01**
Blood basophil count (x10⁹cells/L)	**0.04 (0.02 – 0.05)**	**0.04 (0.02 – 0.05)**	**0.05 (0.03 – 0.07)**	**0.04 (0.03 – 0.05)**	**<0.01**
Blood monocyte count (x10⁹cells/L)	**0.53 (0.42 - 0.67)**	**0.55 (0.46 – 0.70)**	**0.56 (0.47 - 0.70)**	**0.68 (0.58 – 0.800**	**<0.01**
Blood lymphocyte count (x10⁹cells/L)	**1.86 (1.51 – 2.40)**	**1.90 (1.53 - 2.32)**	**2.05 (1.71 - 2.62)**	**1.63 (1.16 - 2.20)**	**<0.01**
Percentage blood neutrophils (%)	**67.0 (60.8 – 72.6)**	**65.7 (60.1 – 70.4)**	**59.3 (53.1 – 65.0)**	**77.8 (71.8 – 85.4)**	**<0.01**
Percentage blood Eosinophils (%)	**2.1 (1.4 – 3.0)**	**2.1 (1.4 – 2.8)**	**5.7 (4.6 – 7.3)**	**1.2 (0.6 – 2.0)**	**<0.01**
Percentage blood basophils (%)	**0.4 (0.3 – 0.6)**	**0.4 (0.3 – 0.6)**	**0.6 (0.4 – 0.8)**	**0.3 (0.2 – 0.4)**	**<0.01**
Percentage blood monocytes (%)	**6.5 (5.4 – 7.6)**	**6.8 (5.8 – 8.3)**	**6.7 (5.9 – 7.8)**	**5.6 (4.7 – 7.1)**	**<0.01**
Percentage blood lymphocytes (%)	**22.9 (17.9 – 29.2)**	**24.1 (20.0 – 29.2)**	**26.6 (21.0 – 32.2)**	**13.4 (8.2 – 18.2)**	**<0.01**
Sputum cell count (x10^3 cells/mg)	**3.7 (1.2 – 8.8)**	**2.8 (1.3 – 7.7)**	**3.6 (1.5 – 7.5)**	**34.7 (18.2 – 52.8)**	**<0.01**
Sputum neutrophil count (x10^3 cells/mg)	**2.6 (0.8 – 6.9)**	**2.1 (0.9 – 6.3)**	**2.1 (0.7 – 5.4)**	**32.9 (17.6 – 51.5)**	**<0.01**
Sputum eosinophil count (x10^3 cells/mg)	**0.0 (0.0 – 0.1)**	**0.0 (0.0 – 0.1)**	**0.2 (0.0 – 0.7)**	**0.1 (0.1 – 0.2)**	**<0.01**
Sputum lymphocyte count (x10^3 cells/mg)	**0.0 (0.0 – 0.0)**	**0.0 (0.0 – 0.0)**	**0.0 (0.0 – 0.0)**	**0.1 (0.0 – 0.1)**	**<0.01**
Sputum macrophages count (x10^3 cells/mg)	**0.0 (0.0 – 0.4)**	**0.1 (0.0 – 0.4)**	**0.0 (0.0 – 0.3)**	**0.3 (0.0 – 1.2)**	**<0.01**
Percentage sputum neutrophils (%)	**79.0 (60.0 – 91.0)**	**79.0 (58.3 – 90.8)**	**68.0 (47.6 – 83.9)**	**96.0 (92.6 – 97.8)**	**<0.01**
Percentage sputum eosinophils (%)	**0.5 (0.3 – 2.0)**	**0.8 (0.3 – 1.8)**	**6.0 (1.3 – 17.3)**	**0.3 (.3 – 0.5)**	**<0.01**
Percentage sputum lymphocytes (%)	0.3 (0.3 – 0.5)	0.3 (0.3 – 0.3)	0.3 (0.3 – 0.4)	0.3 (0.3 – 0.3)	0.51
Percentage sputum macrophages (%)	**12.0 (4.8 – 29.3)**	**15.3 (5.8 – 30.0)**	**13.5 (7.1 – 28.3)**	**3.0 (2.0 – 6.5)**	**<0.01**
CRP (mg/L)	**2.5 (2.5 – 9.0)**	**2.5 (3.0 – 11.0)**	**2.5 (2.5 – 10.0)**	**47.0 (17.0 – 118.5)**	**<0.01**
Blood procalcitonin (ng/mL)	**0.06 (0.05 - 0.08)**	**0.06 (0.04 - 0.08)**	**0.05 (0.04 0.07)**	**0.09 (0.06 – 0.11)**	**<0.01**
Blood Surfactant Protein D (ng/L)	190 (117 – 294)	178 (121 – 264)	160 (123 – 242)	152 (99 – 236)	0.10
Blood neopterin (nmol/L)	9.9 (7.8 – 12.0)	10.2 (7.3 – 16.1)	10.4 (8.0 – 13.9)	12.2 (8.4 – 15.0)	0.39
Serum IL1B (pg/mL)	6.7 (5.7 – 8.1)	7.0 (5.3 – 8.7)	7.5 (5.8 – 10.5)	6.2 (5.5 – 7.9)	0.29
Serum IL8 (pg/mL)	6.5 (4.6 – 8.5)	5.6 (4.1 – 8.8)	5.8 (4.1 – 8.4)	4.7 (3.4 – 6.9)	0.15
Serum CXCL10 (pg/mL)	51.1 (40.6 – 65.8)	57.3 (41.1 – 89.0)	54.4 (38.8 – 75.2)	52.5 (44.2 – 63.0)	0.14
Serum CCL4 (pg/mL)	415 (305 – 506)	400 (306 – 522)	414 (301 – 522)	363 (255 – 449)	0.53

Serum TNF1A (pg/mL)	**7886 (6983 – 10067)**	**8866 (6678 – 11434)**	**7960 (6661 – 9111)**	**10651 (7828 – 13368)**	**0.02**
Serum TNF1B (pg/mL)	5179 (4194 – 5951)	5147 (4212 – 6471)	4975 (4136 – 5798)	5460 (4635 – 6129)	0.46
Serum VEGFA* (pg/mL)	725 (642 – 819)	770 (708 – 838)	742 (659 – 834)	861 (737 – 1007)	0.31
Serum IL13 (pg/mL)	28.2 (19.6 – 37.4)	29.9 (21.5 – 40.2)	26.7 (19.7 – 39.6)	32.5 (18.8 – 41.0)	0.81
Serum IL5 (pg/mL)	**2.2 (1.6 – 3.0)**	**2.2 (1.8 – 2.8)**	**2.9 (2.0 – 3.6)**	**2.3 (1.8 – 2.9)**	**<0.01**
Serum IL6 (pg/mL)	**13.2 (9.1 – 16.6)**	**14.4 (10.1 – 23.2)**	**14.9 (10.7 – 23.0)**	**48.4 (14.9 – 127.5)**	**<0.01**
Serum CXCL11 (pg/mL)	**59.9 (41.2 – 81.8)**	**62.2 (40.9 – 89.3)**	**68.5 (49.3 – 105.7)**	**50.7 (35.5 – 59.7)**	**0.02**
Serum CCL2 (pg/mL)	**691 (560 – 837)**	**619 (494 – 765)**	**749 (608 – 843)**	**560 (493 – 747)**	**0.04**
Serum CCL17 (pg/mL)	**344 (204 – 511)**	**382 (246 – 564)**	**601 (322 – 1134)**	**455 (243 – 705)**	**<0.01**
Serum TNFalpha (pg/mL)	3.4 (2.6 – 5.2)	3.3 (2.7 – 4.1)	3.6 (3.0 – 4.3)	3.3 (2.8 – 4.4)	0.18
Serum IFNgamma (pg/mL)	**0.1 (0.0 – 0.4)**	**0.2 (0.1 – 0.8)**	**0.2 (0.1 – 0.4)**	**0.1 (0.0 – 0.2)**	**<0.01**
Serum IL10 (pg/mL)	0.7 (0.0 – 2.1)	1.0 (0.0 – 2.3)	0.7 (0.0 – 1.5)	0.9(0.6 – 2.4)	0.49
Serum CCL3 (pg/mL)	20.2 (13.1 – 26.8)	18.8 (14.7 – 26.9)	20.5 (15.4 – 32.0)	15.8 (12.7 – 22.4)	0.14
Serum CCL13 (pg/mL)	642 (485 – 801)	694 (458 – 837)	819 (640 – 1023)	964 (659 – 1000)	0.16
Serum Amyloid A_1 (ng/mL)	**4804 (2337 – 9791)**	**5855 (2357 – 14462)**	**4642 (2137 – 15238)**	**106867 (18546 – 253955)**	**<0.01**
Sputum Neopterin (nmol/g)	**15.8 (12.4 – 28.8)**	**23.2 (12.5 – 54.3)**	**15.0 (11.5 – 23.1)**	**90.0 (23.7 – 152.5)**	**<0.01**
Sputum sTREM (pg/L)	4896 (1791 – 7659)	3796 (2008 – 7064)	3436 (1469 – 7146)	2821 (1425 – 6606)	0.49
Sputum IL1B (pg/mL)	**68.9 (19.6 – 384.1)**	**144.2 (31.3 – 813.5)**	**31.1 (11.7 – 276.5)**	**2170.7 (388.9 – 5451.1)**	**<0.01**
Sputum IL6R* (pg/mL)	**159 (114 – 221)**	**218 (174 – 272)**	**150 (117 – 192)**	**523 (344 – 793)**	**<0.01**
Sputum IL8 (pg/mL)	**4860 (1813 – 14980)**	**6493 (2035 – 15838)**	**3405 (1546 - 7891)**	**13828 (6125 – 18245)**	**<0.01**
Sputum CXCL10 (pg/mL)	**284 (81 – 765)**	**383 (124 – 2832)**	**218 (101 – 854)**	**122 (54 – 358)**	**<0.01**

Sputum CCL4 (pg/mL)		893 (447 - 2384)	1188 (563 – 3246)	1471 (808 – 2431)	963 (474 – 1686)	0.31
Sputum TNFA* (pg/mL)		**1341 (1017 – 1768)**	**1573 (1283 – 1929)**	**915 (715 – 1171)**	**4703 (3347 – 6608)**	**<0.01**
Sputum TNFB (pg/mL)		**387 (256 – 954)**	**691 (289 – 1431)**	**336 (158 – 966)**	**1661 (860. – 3247)**	**<0.01**
Sputum VEGFA* (pg/mL)		**1428 (1198 – 1702)**	**1332 (1191 – 1490)**	**1117 (957 – 1305)**	**2178 (1765 – 2687)**	**<0.01**
Sputum IL5 (pg/mL)		**1.4 (0.0 – 6.5)**	**0.0 (0.0 – 2.5)**	**3.0 (0.0 – 8.6)**	**0.0 (0.0 – 1.8)**	**<0.01**
Sputum IL6 (pg/mL)		610 (173 – 1003)	744 (331 – 2105)	634 (204 – 1328)	953 (398 – 2865)	0.13
Sputum CXCL11 (pg/mL)		**10.8 (3.1 – 32.0)**	**13.6 (3.9 – 185.4)**	**12.7 (3.8 – 34.7)**	**0.0 (0.0 – 8.2)**	**<0.01**
Sputum CCL2 (pg/mL)		**784 (318 – 1534)**	**449 (261 – 1227)**	**583 (265 – 922)**	**298 (138 – 527)**	**<0.01**
Sputum CCL5 (pg/mL)		**3.1 (0.0 – 6.1)**	**5.5 (2.8 – 19.2)**	**3.0 (1.7 – 7.5)**	**7.7 (4.5 – 17.1)**	**<0.01**
Sputum CCL17 (pg/mL)		**21.2 (8.7 – 57.0)**	**15.2 (7.0 – 26.8)**	**36.6 (15.8 – 62.2)**	**0.0 (0.0 – 11.7)**	**<0.01**
Sputum TNFalpha (pg/mL)		**6.0 (0.0 - 31.7)**	**25.2 (0.0 - 110.8)**	**0.0 (0.0 – 19.9)**	**85.0 (28.5 – 469.9)**	**<0.01**
Sputum CCL13 (pg/mL)		**31.1 (15.6 – 55.9)**	**24.2 (13.2 - 38.9)**	**36.0 (18.4 – 52.6)**	**0.0 (0.0 – 22.7)**	**<0.01**
Sputum CCL3 (pg/mL)		55.4 (33.1 – 196.8)	99.0 (41.0 – 246.2)	63.1 (31.5 – 262.7)	62.4 (30.3 – 121.5)	0.22
Virus Positive, n (%)		17 (10)	34 (18)	12 (11)	6 (18)	0.12
Colony Forming Units (x10^7 per ml)		**1.7 (0.3 – 9.1)**	**2.3 (0.4 – 16.7)**	**0.6 (0.3 – 5.8)**	**5.3 (0.5 – 37.4)**	**<0.01**
HI Positive, n (%)		98 (66)	167 (65)	73 (65)	46 (65)	1.00
SA Positive, n (%)		7 (7)	15 (7)	13 (14)	5 (7)	0.18
SP Positive, n (%)		69 (51)	137 (53)	54 (50)	29 (49)	0.47
MC Positive, n (%)		53 (47)	103 (45)	35 (38)	33 (48)	0.47
Characteristic		Participants with low Overall Symptom Burden State (n=139)	Participants with symptomatic Crisis (n=136)	Participants with eosinophilic Crisis (n=74)	Participants with neutrophilic Crisis (n= 55)	P-Value
Male, n (%)		87 (63)	89 (65)	53 (72)	37 (67)	0.61
Smoking Status	Current Smoker, n (%)	43 (31)	54 (40)	22 (31)	19 (36)	0.43
	Ex-Smoker, n (%)	93 (67)	78 (58)	47 (66)	31 (58)	
	Never Smoker, n (%)	3 (2)	2 (1)	2 (3)	3 (6)	
SABA, n (%)		137 (99)	126 (95)	67 (97)	49 (94)	0.14
SACH, n (%)		28 (20)	27 (20)	13 (19)	9 (17)	0.96

LACH, n (%)	82 (59)	75 (56)	39 (57)	33 (63)	0.82
LABA, n (%)	108 (78)	105 (79)	59 (86)	45 (87)	0.40
ICS, n (%)	113 (82)	112 (84)	61 (88)	46 (88)	0.54
Maintenance Prednisolone, n (%)	**10 (7)**	**11 (8)**	**1 (1)**	**8 (15)**	**0.04**
Theophylline, n (%)	16 (12)	18 (14)	13 (19)	10 (19)	0.39
Mucolytic, n (%)	21 (15)	22 (17)	6 (9)	11 (21)	0.28
Oxygen, n (%)	11 (8)	11 (8)	4 (6)	6 (12)	0.73
Prophylactic Antibiotic, n (%)	7 (5)	8 (6)	3 (4)	4 (8)	0.86
Cardiovascular Comorbidity, n (%)	82 (59)	79 (58)	32 (43)	34 (62)	0.17
Diabetes, n (%)	15 (11)	15 (12)	4 (6)	4 (8)	0.56
Endocrine Comorbidity, n (%)	7 (5)	6 (5)	1 (1)	3 (6)	0.62
Depression, n (%)	14 (10)	20 (15)	9 (13)	9 (17)	0.50
Osteopenia/Osteoporosis, n (%)	19 (14)	19 (15)	12 (18)	9 (17)	0.86
Anaemia, n (%)	3 (2)	4 (3)	3 (4)	0 (0)	0.47

Definition of abbreviations: SABA: short acting beta agonist; LABA: long acting beta agonist; SACH; short acting anticholinergic; long acting anticholinergic; SGRQ: St George's Respiratory Questionnaire; CRQ: Chronic Respiratory Disease Questionnaire; eGFR: estimated glomerular filtration rate; FEV_1 : forced expiratory volume after 1 second; FVC: forced vital capacity; HI: Haemophilus influenzae; SA: Staphylococcus aureus; SP: Streptococcus pneumoniae; MC: Moxarella catarrhalis; FE_{NO} : Fraction of exhaled nitric oxide.

Variable presentation and hypothesis tests: All numerical variables are presented as median (lower quartile to upper quartile) except when marked ~ or * in which case variables are presented as mean (standard error of the mean) and geometric mean (95% confidence interval) respectively. For numerical variables, all p-values are Kruskal-Wallis p-values except when variable is indicated with ~ or * in which case the p-values are ANOVA p-values. Binary and categorical variables are presented as number (%) of instances (chi-squared test used for COPD state comparison).

Correction for multiple comparisons is not made since the analysis is hypothesis-generating.

Explanation of binary/categorical variable presentation: percentages for each column may deviate slightly from expected percentage based on the total number of measurements given by the column header due to the presence of missing values.

Use of COPD state number of events or number of participants who had an event of a particular state: categorical variables reflecting demographics, baseline medication usage and comorbidities are presented based on the number of participants who had at least one event of a COPD state rather than on the number of events of a COPD state.

Rows in bold font correspond to variables differing significantly across states according to $p<0.05$.

Superiority of objective COPD states compared to conventional exacerbation and stable state definitions

Multivariable symptom and blood inflammatory index overlap between the objective COPD states shows good separation compared to the 0.67 DO between exacerbation and stable state labels. Table 4.4 summarises the overlap scores between the objective COPD states. The maximum DO occurs between the symptomatic crisis with the neutrophilic crisis (DO = 0.41). This is followed by a DO = 0.20 for overlap between eosinophilic crises and the low overall symptom burden state. Overlap between equilibrium and a crisis ranged from DO = 0.14 to DO = 0.20.

.

Table 4.4 | Degree of Overlap between each COPD state.

	Low Overall Symptom Burden State	Symptomatic Crisis	Eosinophilic Crisis	Neutrophilic Crisis
Low Overall Symptom Burden State	N/A			
Symptomatic Crisis	0.14	N/A		
Eosinophilic Crisis	0.20	0.07	N/A	
Neutrophilic Crisis	0.19	0.41	0.13	N/A

The separation between the objective COPD states is visualised in Figure 4.2 in a 3-dimensional principal component space representing VAS symptoms and blood inflammatory indices as described in Methods.

Figure 4.2 | Low overall symptom burden state and different crisis state events and concentration ellipsoids visualised in three-dimensional principal component space representing symptoms and blood inflammatory indices. Low overall symptom burden state shown in light blue; symptomatic crises shown in dark blue; eosinophilic crises shown in pink; neutrophilic crises shown in yellow. PC1 mostly reflects symptom variation (66% of PC1 is explained by symptoms; see **Appendix Table A4.3**), with events represented by data points further away from the axis origin showing lower symptoms. PC2 mostly reflects variation in eosinophilic inflammation (66% of PC2 is explained by eosinophilic inflammation; see **Appendix Table A4.3**), with events represented by data points further away from the axis origin showing lower eosinophilic inflammation. PC3 mostly reflects variation in neutrophilic inflammation (60% of PC3 is explained by neutrophilic inflammation as represented by CRP, neutrophil cell count and percentage; see **Appendix Table A4.3**), with events represented by data points further away from the axis origin showing lower neutrophilic inflammation.

Comparison of conventional exacerbation and stable state labels against objective COPD states reveals both potential under- and over-diagnosis of COPD worsening episodes

Out of the 1,001 conventionally labelled stable state events, the majority (56%, n=556) belonged to symptomatic crisis, eosinophilic crisis or neutrophilic crisis. Out of the 347 conventionally labelled exacerbations, 44 (13%) belonged to the low overall symptom burden state. A further 159 (46%) belonged to the symptomatic crisis and hence were characterised by little or no inflammation although with high symptom burden. A total of 203 (59%) events conventionally labelled exacerbation belonged to either low overall symptom burden state or the symptomatic crisis.

Table 4.5 summarises the correspondence and differences between conventional exacerbation and stable state labels and the 4 objective COPD states identified.

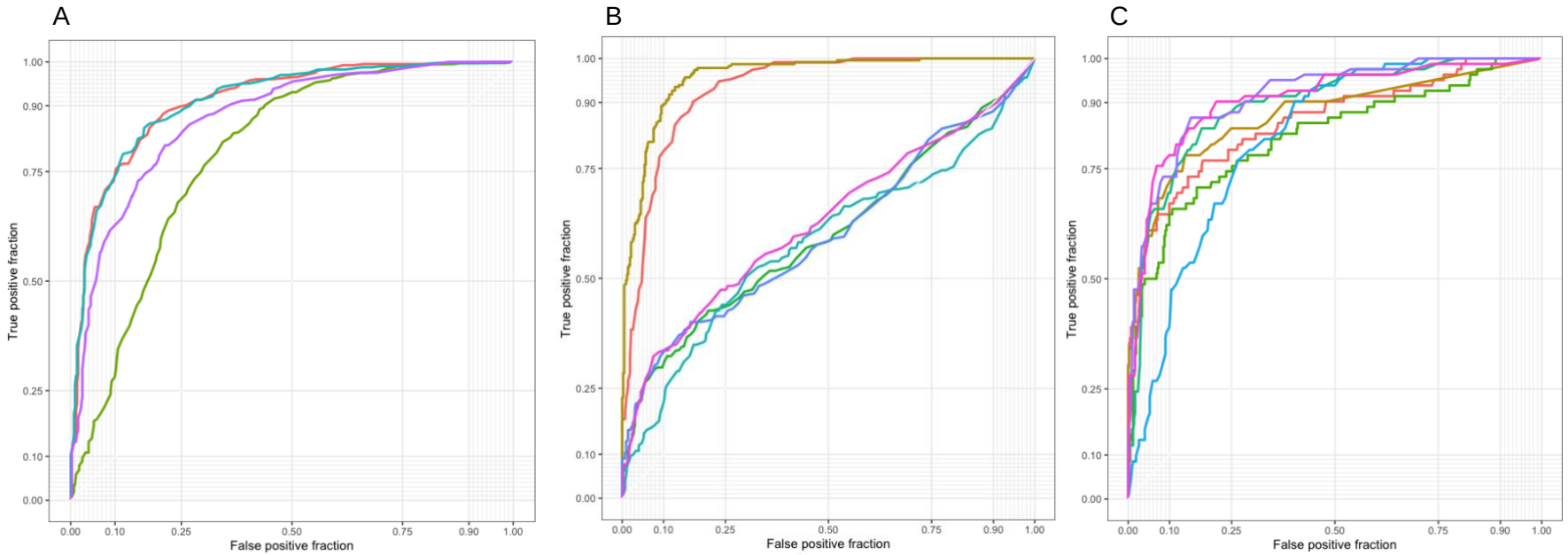
Table 4.5 | Correspondence between conventional exacerbation and stable state labels versus new, objective COPD disease states.

	Low overall symptom burden state (n=489)	Symptomatic crisis (n=488)	Eosinophilic crisis (n=277)	Neutrophilic crisis (n= 94)
Conventional Exacerbation Label, n (%)	44 (9)	159 (33)	73 (26)	71 (76)
Conventional Stable Label, n (%)	445 (91)	329 (67)	204 (74)	23 (25)

Identification of most discriminative biomarkers to distinguish crises from low overall symptom burden state

In order to determine which of the individual symptom or biological variables used to define the objective COPD states would be most appropriate to distinguish crises from the low overall symptom burden state in a clinical setting, I compared the AUC of each VAS symptom and blood inflammatory indices using 80% of participant training data (Figure 4.3A-C).

Figure 4.3 | ROC curve of each symptom and particular blood biological variables for discriminating between crises and the low overall symptom burden state using 80% of dataset samples. A: AUCs for low overall symptom burden state versus symptomatic crises were as follows: VAS cough (red) = 0.91; VAS dyspnoea (green) =0.79; VAS sputum production (light blue) = 0.91; VAS sputum purulence (purple) = 0.87. B: AUCs for low overall symptom burden state versus eosinophilic crises were as follows: blood eosinophil count (red) = 0.93; blood eosinophil percentage (brown) =0.96; VAS cough (green) =0.60; VAS dyspnoea (light blue) =0.58; VAS sputum production (dark blue) = 0.60; VAS sputum purulence (pink) = 0.63. C: AUCs for low overall symptom burden state versus neutrophilic crises were as follows: blood neutrophil count (red) = 0.86; blood neutrophil percentage (green) = 0.82; CRP (brown) = 0.87; VAS cough (light blue) = 0.89; VAS dyspnoea (dark blue)= 0.82; VAS sputum production (purple) = 0.91; VAS sputum purulence (pink) = 0.91.



The best biomarkers for distinguishing symptomatic crises from the low overall symptom state, the eosinophilic crises from the low overall symptom state and neutrophilic crises from the low overall symptom state were VAS cough (threshold ≥ 37mm), blood eosinophil percentage (threshold ≥3.9%) and VAS sputum production (threshold ≥ 35mm) respectively. In the dataset test set samples, I evaluated the sensitivity and specificity of these thresholds for detecting a crisis compared to low overall symptom burden state. The VAS cough threshold enabled 65 (78%) of the 83 test set symptomatic crises and 60 (90%) of the 67 test set measurements from the low overall symptom burden state to be identified. The blood eosinophil percentage threshold enabled 42 (93%) of the 45 test set eosinophilic crises and 65 (97%) of the 67 test set measurements from the low overall symptom burden state to be identified. The VAS sputum production threshold enabled all 12 (100%) of the 12 test set neutrophilic crises and 55 (82%) of the 67 test set measurements from the low overall symptom burden state to be identified.

Neural network algorithm using a combination of VAS symptom scores and blood inflammatory indices achieves optimal assignment of patient measurements to COPD state

Distinguishing each objective COPD state from each of the other states required a multivariable approach in contrast to the univariate approach of distinguishing a particular type of crisis from the low overall symptom burden state. In order to identify which COPD state a given event belongs based on patient symptom and blood inflammatory measurements, I first sought to identify which variables out of VAS symptoms, CRP, blood neutrophil and eosinophil count and percentages are significant in a multivariable context. Using the SES algorithm with a $p<0.1$ threshold

on data representing 80% of the 1,348 dataset measurements, I selected the following variables: VAS cough, VAS dyspnoea, VAS sputum production, VAS sputum purulence, blood CRP, blood neutrophil count, blood eosinophil count and blood eosinophil percentage. Using these variables, I trained and evaluated random forests, decision trees, KNN, SVM and neural network algorithms. The cross-validation mean balanced accuracy performance of the optimised algorithms for assigning any given measurement form a patient to the correct COPD state are shown in Table 4.6. Mean balanced accuracy of classifiers to identify a COPD state ranged from 72% (KNN) to 88% (neural network).

Table 4.6 | Mean balanced accuracy from cross-validation for each classifier type and accompanying optimal hyper-parameter values.

	Best Mean Balanced Accuracy	Best Hyperparameter Values
Random Forest	0.8630722	400 trees, mtry=1
KNN	0.7221363	K=7
Decision Tree	0.8260149	cp = 0.00
SVM	0.8601672	sigma = 0.30 and C = 0.13
Neural Network	0.8810098	size = 9 and decay = 0.2

COPD state detection algorithm validated on internal cohort test samples

I retrained the best performing COPD state detection algorithm, neural network, on all training and validation folds and assessed this algorithm on the hold out test set of the remaining 20% of dataset measurements (Table 4.7). In particular, 61 (91%) out of 67 test set low overall symptom burden events were correctly identified by the neural network algorithm; 69 (83%) out of the 83 symptomatic crisis events were correctly identified; 40 (89%) out of the 45 test set eosinophilic crisis events were correctly identified; 11 (92%) out of the 12 test set neutrophilic crisis events were correctly identified.

Only 9 (11%), 4 (9%) and none of the symptomatic, eosinophilic and neutrophilic crises respectively were falsely predicted by the neural network COPD state detection algorithm as belonging to the low overall symptom burden state.

Table 4.7 | For each COPD state, number of test set events which the neural network algorithm predicts belong to each COPD disease state.

	Actual low overall symptom burden (n=67)	Actual symptomatic crisis (n=83)	Actual eosinophilic crisis (n=45)	Actual neutrophilic crisis (n=12)
Predicted low overall symptom burden, n (%)	61 (91)	9 (11)	4 (9)	0 (0)
Predicted symptomatic crisis, n (%)	2 (3)	69 (83)	1 (2)	1 (8)
Predicted eosinophilic crisis, n (%)	1 (1)	3 (4)	40 (89)	0 (0)
Predicted neutrophilic crisis, n (%)	3 (4)	2 (2)	0 (0)	11 (92)

COPD state detection algorithm validated on external, separate patient cohort

I determined to assess the generalisability of the neural network COPD state detection algorithm in assigning patient measurements to suitable states using VAS cough, VAS dyspnoea, VAS sputum production, VAS sputum purulence, blood CRP, blood neutrophil count, blood eosinophil count and blood eosinophil percentage. To this end, I retrained the neural network COPD state detection algorithm on all 1,348 dataset measurements of this thesis Chapter and used the retrained algorithm to assign a COPD state to each of the 101 COPD/asthma participants from the ED dataset of Chapter 3 (Table 4.8). 47 of the 101 participants were assigned to the low symptom burden state, suggesting that these participants may have low inflammation and a high burden for an individual symptom but not overall for a combination of cough, dyspnoea, sputum production and sputum purulence. 18 participants admitted to the emergency department were assigned to the symptomatic crisis suggestive of high burden across

symptoms but nevertheless low inflammation. 20 participants were assigned to the eosinophilic crisis state and 16 participants to a neutrophilic crisis state. In keeping with the definition of the objective COPD states derived in this chapter, participants assigned to the low overall symptom burden state had lower VAS symptom scores compared to symptomatic, eosinophilic and neutrophilic crises ($p<0.01$). Blood neutrophil cell count and percentage was highest for participants assigned to the neutrophilic crisis state, as was blood CRP (25.0mg/L compared to 4.2, 7.0 and 6.5mg/L for participants assigned to the low overall symptom burden state, the symptomatic state and the eosinophilic state respectively) ($p<0.01$ in all cases). As expected, blood eosinophil count and percentage was highest for participants assigned to the eosinophilic crisis state (7.2% compared to 1.2%, 1.2% and 0.6% for participants assigned to the low overall symptom burden state, the symptomatic crisis state and the neutrophilic crisis state respectively) ($p<0.01$ in both cases).

Table 4.8 | Comparison of symptom and blood inflammatory characteristics of participants from chapter 3 emergency department dataset belonging to each COPD state predicted by COPD state detection algorithm.

Characteristic	Low Overall Symptom Burden Crisis (n=47)	Symptomatic Crisis (n=18)	Eosinophilic Crisis (n=20)	Neutrophilic Crisis (n= 16)	P-Value
VAS cough (mm)	29 (16 – 53)	74 (57 – 95)	61 (30 – 74)	65 (46 – 71)	<0.01
VAS dyspnoea (mm)	57 (6 – 75)	82 (68 - 89)	73 (29 – 84)	79 (57 – 89)	<0.01
VAS sputum production (mm)	3 (0 – 21)	45 (16 – 67)	28 (6 – 53)	52 (32 – 71)	<0.01
VAS sputum purulence (mm)	1 (0 – 10)	52 (14 – 76)	19 (0 – 50)	54 (43 – 68)	<0.01
Blood neutrophil count* (x10⁹cells/L)	7.9 (7.0 – 8.9)	5.8 (4.7 – 7.1)	5.8 (4.9 – 6.8)	11.2 (9.4 – 13.4)	<0.01
Blood eosinophil count (x10⁹cells/L)	0.1 (0.0 – 0.3)	0.1 (0.0 – 0.2)	0.6 (0.5 – 1.3)	0.1 (0.0 – 0.2)	<0.01
Percentage blood neutrophils (%)	75.0 (68.0 – 81.8)	74.9 (62.7 – 79.1)	57.2 (53.0 – 62.2)	80.6 (75.1 – 87.7)	<0.01
Percentage blood eosinophils (%)	1.2 (0.2 – 3.0)	1.2 (0.2 – 1.9)	7.2 (5.3 – 10.3)	0.6 (0.2 – 1.6)	<0.01
Blood CRP (mg/L)	4.2 (0.8 – 23.7)	7.0 (1.1 – 20.9)	6.5 (0.8 – 15.6)	25.0 (13.1 – 59.5)	0.04

Definition of abbreviations: VAS = Visual Analogue Scale; CRP = C-reactive protein.

4.4 Discussion

Overall findings

My mathematical definition of a low overall symptom burden state and of three crisis states is the first of its kind objective definition. It can be seen that 13% of conventionally labelled exacerbation events are in fact in the low overall symptom burden state. Furthermore, 59% of conventionally labelled exacerbations are objectively either low overall symptom burden state or symptomatic crisis events. This may suggest that up to 59% of conventionally labelled exacerbations of the dataset analysed in this chapter may not in fact require SCS or antibiotics. At the same time, I have shown that 56% of conventionally labelled stable state events in fact belong to one of the three crisis states. The greatest proportion of these potentially underdiagnosed events belong to the symptomatic crisis or eosinophilic crisis. My findings show that using these novel mathematical states, approximately three quarters of eosinophilic crisis events are missed according to conventional clinical definitions of exacerbations.

Interpretation of COPD state definitions, strengths of analysis and comparison with other studies

It is well known from previous studies that COPD is underdiagnosed[150] which may mean that exacerbations are not always captured. This may adversely impact patient outcomes since exacerbations are associated with increased morbidity and mortality[6]. Equally, overtreatment may be of concern since SCS use is associated with significant side effects while antibiotics should be used with care in view of the risk of contributing to antibiotic resistance amongst other factors[77,135]. These difficulties are amplified by

the limitation of current clinical definitions of COPD exacerbations. The GOLD definition and other attempted definitions discussed in Chapter 1 of my thesis lack any biological component and include no indication of the magnitude of symptoms at exacerbation[6]. The omission of biology from conventional clinical definitions is significant in view of the known biological heterogeneity of COPD exacerbations, namely, neutrophilic exacerbations, eosinophilic exacerbations and pauci-inflammatory exacerbations[60]. However, all studies to-date including Bafadhel *et al* (2011) have used existing clinical definitions of exacerbations as a starting point before characterising different types of exacerbations. Other studies have attempted to develop machine learning algorithms using variables including symptoms to distinguish between stable state and exacerbations as part of an automated approach but again these studies are circular in approach since they rely on existing stable state and exacerbation labels in order to train exacerbation detection algorithms. For example, Claxton *et al* (2021) used patient age, self-reported symptoms and cough audio data to detect exacerbations of COPD as part of a patient self-management approach[151]. As discussed in Chapter 1 of my thesis, the ROME proposal used a Delphi methodology to define exacerbations but the quantitative thresholds of symptom, biological and physiological variables in the ROME proposal map onto exacerbations as currently defined in the clinic rather than being entirely empirically derived[56]. The 'crisis concept' highlights the importance of acknowledging the potential or confirmed nature of a COPD worsening event, the temporal nature of the event and the aetiology of such an event with respect to inflammation, microbiology, psychology and comorbidities[141]. My analysis sought to begin to realise this goal by being the first to propose an objective mathematical definition of COPD disease states, including a low

overall symptom burden state and three crisis states which do not rely on current clinical definitions.

My analysis in this chapter showed an association of increased bacterial load (as measured by the number of sputum colony forming units) with neutrophilic crises. Neutrophilic crises also showed higher levels procalcitonin and of general pro-inflammatory cytokines in serum and/or sputum compared to other COPD states including TNFA, IL6, IL6R, IL1B and IL8. The higher levels of serum and sputum cytokines such as CCL13 and IL5 and sputum CCL17 in eosinophilic crises can be understood in the context of these cytokines in Th2-mediated immunity and recruiting eosinophils. It is interesting that CXCL10, CXCL11 and serum INFgamma, known to be associated with viral immune responses, were at a higher level in eosinophilic (CXCL10, CXCL11 and IFNgamma) and symptomatic crises (CXCL10 and CXCL11) compared to neutrophil crises even though viral presence in sputum did not differ between the groups. However, the presence versus absence of virus in sputum was assessed in 492 out of the 1,348 measurements in the dataset analysed in this chapter and qualitatively assessed. Future investigations should include quantitation of viral load in sputum to further ascertain possible viral aetiology for a subset of eosinophilic and symptomatic crises. This is especially advisable given that quantitative indicators of sputum bacterial load but not qualitative indicators were associated with neutrophilic crises in this chapter analysis. Of course, it should be noted that the associations of particular serum and sputum cytokines with particular COPD states do not prove causation. Studies should also be performed to elucidate potential mechanistic roles of the serum and sputum cytokines differing between COPD states in the progression of crisis events through emergence and resolution.

There are noteworthy similarities as well as some differences between the findings of this Chapter and those of Bafadhel et al (2011)[60]. In particular, Bafadhel et al revealed a cluster with the highest levels of sputum TNFalpha which was bacteria-associated, matching the characteristics of the neutrophilic crises identified in this chapter[60]. The eosinophilic crises identified by me have similar cytokine characteristics to the eosinophilic cluster of Bafadhel et al, for example, in terms of sputum CCL17[60]. However, whereas in Bafadhel et al an eosinophilc cluster and a virus predominant cluster could be distinguished by cytokines such as CXCL10 and IL5 being at higher levels in the virus-predominant cluster, this was not apparent in my findings[60]. Instead, the eosinophilic crises identified by me were characterised by high sputum CCL13, CCL17, CXCL10 and IL-5, yet viral infection did not differ between crisis type. Furthermore, only 16 (11%) of the 148 exacerbations with available sputum cytokine data in Bafadhel were in the pauci-inflammatory cluster whereas 57% of all crises identified in this thesis chapter were characterised by high symptoms but low inflammation[60]. The differences between my findings and those of Bafadhel et al correspond to the different objectives of analysis. I sought to build on the biological heterogeneity mapping of exacerbations by Bafadhel et al by providing new definitions of COPD states with a hybrid focus on symptoms and biology so that any patient with high symptoms can be identified and the presence and type of inflammation known without reliance on ‘exacerbation’ or ‘stable state’ definitions. I achieved this by including all events regardless of whether labelled as ‘stable state’ or as ‘exacerbation’ in my cluster analysis and by using both symptom and inflammatory variables as input for cluster analysis. In addition to the ‘stable state’ and ‘exacerbation’ measurements of the Bafadhel et al dataset[60], I expanded the sample size for my analysis in this thesis chapter by also including exacerbation and stable state measurements from a second

Bafadhel et al dataset[82]. In the light of the biological exacerbation heterogeneity found by Bafadhel et al[60], my findings add the following insights which may be important for COPD management:

1) A number of crisis events with high symptoms are currently missed when using conventional 'exacerbation' definitions.
2) A number events currently defined as 'stable state' in clinical practice are in fact characterised by significant eosinophilic inflammation and high symptoms.

My analysis in this thesis chapter provides a tool for distinguishing each COPD state so that symptom and blood inflammatory measurements from a patient at a given time point can be used to identify the COPD state a patient is in at that time point.

Significance of COPD state detection algorithm

My finding that VAS sputum production is a more effective biomarker than CRP for identifying neutrophilic crises is encouraging in view of the clinical ease of the patient self-reported VAS framework. A 20 mg/L ≤ CRP level ≤ 40 mg/L and a CRP level > 40 mg/ threshold have been previously used by Butler et al in a study in which 653 exacerbating patients were randomised and clinicians prescribed antibiotics based on point-of-care CRP testing[152]. The guidance for clinicians noted that antibiotics are unlikely to be beneficial if CRP < 20 mg/L, may be beneficial if 20 mg/L ≤ CRP level ≤ 40 mg/L especially if accompanied by sputum purulence and antibiotics are likely to be beneficial if CRP >40mg/L[152,153]. The primary outcomes were patient-reported use of antibiotics for acute exacerbations of COPD within 4 weeks after randomization and COPD-related health status at 2 weeks after randomization[152]. Similar conclusions were reached in the Prins et al study using a threshold of CRP ≥ 50 mg/L[134]. Here, 220 patients hospitalised for exacerbations of COPD were randomised to receive

antibiotics according to standard care or if CRP ≥ 50 mg/L[134]. The CRP-guided group had lower antibiotic use and no increase in adverse events including length of hospital stay, time to next exacerbation, 30 day treatment failure and 30 day symptom scores[134].

Ultimately, my analysis in this chapter shows that a multivariable approach is important if a clinician is to identify which of the four objective COPD states a patient is in at a given time because in clinical practice one does not know which of symptomatic, eosinophilic or neutrophilic crisis needs to be distinguished from the low overall symptom burden state. Rather, a clinically feasible approach must enable a set of measurements at any time point to be correctly assigned to a COPD state when it is unknown a priori which crisis state is most likely. Hence, the multivariable neural network COPD state detection algorithm developed in this chapter provides a foundation for exploring personalised multi-treatment directing management in real time.

The neural network COPD state algorithm developed in this chapter demonstrated high accuracy in a hold-out test set of patient measurements. Furthermore, I demonstrated that the neural network COPD state algorithm assigning patients in a separate, external cohort from chapter 3 to COPD states with characteristics consistent with the definition of the COPD states. The use of this algorithm could eventually even be extended to daily remote monitoring of patients. Based on the identified COPD disease state a patient is in at a given point in time, treatment could be tailored accordingly. The success of directing SCS therapy at exacerbation according to blood eosinophils[82,83] and of directing antibiotic therapy guided by CRP[145,163] has been

previously shown. In this thesis Chapter, I showed that a threshold of ≥37mm VAS cough, ≥3.9% blood eosinophil percentage and ≥35mm VAS sputum production could effectively distinguish symptomatic crises, eosinophilic crises and neutrophilic crises respectively from the low overall symptom burden state. The blood eosinophil percentage derived from my analysis is higher than that of other studies. For example, a threshold of blood eosinophil percentage > 2% was used in the Bafadhel et al (2012) study to direct systemic corticosteroid therapy, while a blood eosinophil count of ≥ $0.3x10^9$ cells/L was used in the multi-centre study by Sivapalan et al[82,83]. All patients in the Sivapalan et al study received intravenous methylprednisolone on the first day and were randomly assigned to a standard therapy and an eosinophil-guided therapy group[82]. On subsequent days (up to a maximum of 4 days) the eosinophil-guided group were given oral prednisolone on days when their blood eosinophil count was ≥$0.3x10^9$ cells/L[83]. Importantly, the eosinophil-guided therapy was non-inferior compared to standard care in terms of the number of days alive and out of hospital[83]. Hence, both the Bafadhel et al and Sivapalan et al studies enabled a reduction in SCS use while demonstrating beneficial or non-inferior outcomes compared to standard of care[82,83]. Based on the outcome of my analysis in this Chapter compared to existing studies, there may be potential for the use of a more stringent threshold for eosinophilia to further personalise SCS therapy for a smaller group of patients which could reduce side effects for other patients with less pronounced eosinophilia who may benefit less from SCS therapy. Whether such an approach would achieve non-inferiority in treatment outcomes and reduce side effects requires investigation. An exciting possibility is the use of treatment approaches targeting eosinophilic inflammation such as benralizumab, an IL5 receptor-α monoclonal antibody, without the side effects of SCS such as being investigated by Ramakrishnan et al[154]. My findings of symptomatic

and neutrophilic crises may suggest that prescribing SCS for symptomatic and neutrophilic crisis events may constitute an over-use of SCS. Furthermore, in keeping with the Bafadhel et al (2012) study conclusions, prescribing SCS for non-eosinophilic crises may even be expected to increase exacerbation treatment failure[82].

Possible clinical heterogeneity in terms of comorbidities in COPD events of other studies in comparison to the newly defined objective COPD states

Numerous studies have shown that cardiovascular comorbidities and psychological factors such as depression may be linked to COPD exacerbation risk and severity[138,139,155]. An example is the Mannino et al study in which data were analysed from 20,296 subjects in the Atherosclerosis Risk in Communities Study and the Cardiovascular Health Study[156]. It was found that after adjusting for age, sex, race, body mass index and education, patients with poorer lung function (GOLD stage 3 and stage 4) had a higher prevalence of cardiovascular disease, hypertension and diabetes[156]. Follow-up data for patients from baseline until either 5 years or death showed that risk of hospitalisation or death was higher for patients with comorbid disease, the effect on mortality being greater for cardiovascular disease and diabetes than with hypertension[156]. In a study by Patel et al, stable state data from 386 patients of the London COPD cohort were analysed and prospective exacerbations data collected for patients who completed at least a year worth of symptom diaries[157]. Patients with comorbid ischemic heart disease had worse health status, lower exercise capacity, more dyspnoea at stable state and longer exacerbations[157]. The link between depression and COPD exacerbation outcomes is illustrated by Papaioannou et al's analysis discussed earlier in this thesis[139]. Papaioannou et al showed that patients with depressive symptoms on the first day of hospitalisation required longer

hospitalisation[139]. Depressive symptoms on admission also had a significant impact on dyspnoea and CAT score improvement[139]. My analysis suggests that of 62% of patients with neutrophilic crises and 58% of patients with symptomatic crises have cardiovascular comorbidity compared to 43% of patients with eosinophilic crises. However, the difference did not reach statistical significance. Similarly, the proportion of patients with each COPD state who had depression did not differ in my analysis.

Limitations and next steps

A notable limitation of my analysis is that the temporal aspect of the 'crisis concept' was not addressed. It was not known when a transition occurs from the low overall symptom burden state to a symptomatic, eosinophilic or neutrophilic crisis and the magnitude of symptom and inflammatory marker change during this time interval of transition. A future step would be for participants to be tracked at very short intervals e.g. weekly or even daily monitoring of symptoms and blood-based inflammatory markers. An association between such a transition point and bacterial, viral and environmental exposure as potential aetiologies could also be assessed in such a protocol.

Another limitation of my analysis in this thesis Chapter is that although baseline comorbidity was known and analysed, the extent of comorbidity at the time of occurrence of each COPD objective disease state was not assessed. This is because only the presence or absence of cardiovascular comorbidity or psychological comorbidities such as depression was known. Hence, although depression at baseline was not significantly different between COPD states in this chapter analysis, associated with or other psychological comorbidities, it is a possibility that the level of

depression, cardiovascular and/or other comorbidities was higher at the time of a crisis compared to when the same patient was in the low overall symptom burden state. This merits future investigation. Future studies should include biomarkers for cardiovascular comorbidity such as natriuretic peptides and quantitative real time measures of depression and anxiety. Patel et al showed that in subsets of the stable state COPD patient sample, levels of N-terminal pro-brain natriuretic peptide (a validated and sensitive marker of myocardial dysfunction) were higher for patients with ischemic heart disease than for patients without[157]. However, not all patients in the Patel et al study had N-terminal pro-brain natriuretic peptide measured at recruitment, nor was N-terminal pro-brain natriuretic peptide measured at exacerbation. It would be pertinent to identify whether a subgroup of events in the symptomatic crisis state are associated with cardiovascular comorbidity and/or psychological comorbidity at the time of the crisis and whether this results in the high patient self-reported symptoms in the absence of significant inflammation. In addition, the analysis of real time potential biomarkers for comorbidities may enable further resolution in the neutrophilic crisis state group which in addition to the cardiovascular comorbidity association was also associated with lower estimated glomerular filtration rate, perhaps indicating general clinical unwellness beyond COPD crisis occurrence. This further emphasises the importance of assessing patients holistically when presenting with a crisis.

A further limitation of my analysis is that some biomarker with potential to identify COPD states were not assessed. The finding from my analysis revealed that eosinophilic crises were also characterised by higher levels of FE_{NO} compared to other COPD states supports the association of FE_{NO} with airways eosinophilia documented in existing literature[158]. However, in this chapter I did not further investigate the use of

FE_{NO} as a biomarker for algorithmically distinguishing eosinophilic crises from other states. This is because FE_{NO} was only measured in 243 out of the 1,348 dataset measurements. Future studies with widespread FE_{NO} measurements would be interesting to determine to what extent if at all FE_{NO} could enhance the performance of my multivariable COPD state detection algorithm.

As a next step, my COPD disease state definition and the ability of my algorithm to accurately identify them should be prospectively assessed in new patient cohorts. The efficacy of this approach could be further enhanced by monitoring COPD patients more frequently than the 3-month interval of the present study and identifying the trajectory in symptoms and biology from low overall symptom burden to a given crisis state on an individual patient level over the days leading up to the crisis state. Once further validated, there is the potential for a routine use of my COPD disease state identification algorithm by clinicians for patients who present with worsening symptoms as part of a point of care test. This should be feasible since the 8 variables informing my COPD disease state identification algorithm are easily measurable in clinic.

4.5 Conclusion

I mathematically identified four objective COPD disease states using an unsupervised machine learning technique. A low overall symptom burden state was characterised by low symptoms and low inflammation. Three crisis states were all characterised by high symptoms but accompanied either by low inflammation, eosinophilic inflammation or by neutrophilic inflammation. Importantly, in this chapter I also developed and validated a supervised learning algorithm which uses symptoms measured by VAS and systemic inflammatory markers (CRP, blood neutrophil count, blood eosinophil count and blood eosinophil percentage) to accurately identify to which COPD state a patient belongs at a given point in time. This has the potential to be used in clinical practice. In the next chapter of my thesis, I characterise symptom- and inflammation-based exacerbation treatment response and seek to identify distinct patterns of crises treatment response trajectories.

Chapter 5 – Mapping the temporal event course in COPD: crises and post-treatment response

Abstract

The reliance on existing subjective definitions of exacerbations and stable state may adversely impact attempts to personalise exacerbation treatment to achieve better treatment outcomes. In this chapter, I build upon my analysis in Chapter 4 in which I defined objective COPD states as ‘low overall symptom burden state’, ‘symptomatic crisis’, ‘eosinophilic crisis’ and the ‘neutrophilic crisis’. In this chapter I define different crisis treatment response trajectories, which I call ‘trajectotypes’, based on the difference in magnitude of patient symptoms and inflammatory biological indices between the time of the crisis and their magnitude on day 14 follow-up. These trajectotypes include: **no change**, a trajectotype showing **improvement in symptoms**, a trajectotype showing **improvements in symptoms and eosinophilic inflammation** and a trajectotype showing improvements in **symptoms and neutrophilic inflammation**. I also identify different patterns of symptom recovery based on patient self-reported VAS symptom diaries completed daily by patients from the time of exacerbation until day 14 follow-up. Notably, my analysis of the daily VAS diaries reveals patterns in daily symptom profiles over a two-week period which are not discernible by simply assessing day 0 and day 14 symptoms. The results of this chapter pave the way for personalised assessment of temporal treatment response pattern after crisis.

5.1 Introduction

The lack of an objective definition of what constitutes exacerbations of COPD compared to stable state hampers attempts to characterise the emergence of exacerbations from stable state. The subjectivity of the GOLD definition of exacerbations was previously discussed in this thesis[6]. The lack of biological inflammatory indices in the definition, despite COPD exacerbations being heterogeneous with regards to inflammatory biology, limits objective analyses of exacerbation treatment success and failure. Currently, assessment of exacerbation treatment outcome focusses on hospital readmission, retreatment or death[15]. Without knowing patterns of resolution of symptoms, physiology and inflammation and monitoring events for these patterns, personalised crisis management remains hampered.

In this chapter, perform cluster analysis to objectively define distinct crisis treatment response symptom and inflammation trajectories termed ‘trajectotypes’ irrespective of the type of exacerbation state. Finally, I identify different shape profiles of VAS symptom time series derived from patients completing daily symptom diaries between day 0 and day 14.

5.2 Methods

Participants, study design and measurements

The dataset which I analysed in this chapter was derived from previous studies[60,82]. All study participants were COPD patients who were followed up at 3-month intervals or whenever an exacerbation was reported. Patients with reported exacerbations were also followed up two weeks and six weeks post-exacerbation treatment. At each follow-up, measurements were made for physiological indices, blood and sputum inflammation and symptoms. The dataset consisted of two segments: phase 1 and phase 2. Serum and sputum cytokine data were only measured in phase 1 while certain physiological indices such as heart rate, temperature and blood pressure were only measured in phase 2. Whereas in Chapter 4 I could combine these measurements for exacerbation and stable state data inclusion as a foundation for developing new COPD state definitions, this chapter required a different approach. For defining post-treatment trajectories, it was necessary to compare the values of variables measured at both ‘exacerbation’ and post-treatment time points. These variables included symptom scores such as but not limited to VAS[117], lung function, blood and sputum cell counts and percentages, bacteriology and standard blood sample collection clinical reports. With a few exceptions, serum and sputum cytokines were not available at post-treatment follow-up, nor were physiological indices such as heart rate and blood pressure. Table 5.1 summarises the variables available at post-treatment follow-up time points as well as variables available at crisis or baseline measurements (irrespective of whether available at treatment follow-up) measured across both phase 1 and phase 2 of the datasets.

Table 5.1 | Available variables measured at exacerbation and post-treatment follow-up or available for both phase 1 and phase 2 at crisis/baseline time points.

Demographics, comorbidities and medication	Clinical and questionnaires	Physiology	Blood tests	Sputum tests
Age	VAS[117] cough	FEV_1 post-bronchodilator	Blood total cell count	Sputum total cell count
Pack Year History	VAS[117] dyspnoea	Percentage predicted FEV_1	Blood neutrophils	Sputum neutrophils
Sex	VAS[117] sputum production	FEV_1 post-bronchodilator/FVC Ratio	Blood eosinophils	Sputum eosinophils
Smoking status	VAS[117] sputum purulence	eGFR	Blood basophils	Sputum lymphocytes
Cardiovascular comorbidity	CRQ[143] emotion		Blood monocytes	Sputum macrophages
Diabetes	CRQ[143] fatigue		Blood lymphocytes	Sputum neopterin
Endocrine comorbidity	CRQ[143] mastery		CRP	Sputum sTREM
Depression	CRQ[143] dyspnoea		Creatine	
Osteopenia/Osteoporosis			Urea	Colony Forming Units
Anaemia			Procalcitonin	HI positive
SABA				MC positive
SACH				SA positive
LACH				SP positive
LABA				
ICS				
Maintenance prednisolone				
Prophylactic antibiotic				

Theophylline				
Mucolytic				
Oxygen				
For exacerbation measurements, whether SCS given at exacerbation				
For exacerbation measurements, whether antibiotics given at exacerbation				

Definition of abbreviations: CRQ: Chronic Respiratory Disease Questionnaire; eGFR: estimated glomerular filtration rate; FEV_1 : forced expiratory volume after 1 second; FVC: forced vital capacity; HI: Haemophilus influenzae; LABA: long acting beta agonist; LACH = long acting anticholinergic; MC: Moxarella catarrhalis; SA: Staphylococcus aureus; SABA: short acting beta agonist; SACH; short acting anticholinergic; SP: Streptococcus pneumoniae; SCS = systemic corticosteroids; VAS = Visual Analogue Scale.

The goal in defining post-treatment trajectories was to reflect resolution or lack of resolution of symptoms and inflammation characterising a particular crisis state a participant was in. I therefore used VAS symptoms, blood and sputum neutrophil and eosinophil cell counts and percentages and blood CRP as the basis for characterising post-treatment trajectories. I then compared how other available variables measured at the two-week post-treatment time point differed depending on 'trajectotype'. Week six-post treatment follow-up had a lot of missing data, comprising sample size. I therefore did not analyse as part of the chapter aims.

Statistical analyses

Univariate analyses were performed using 'R' and Prism Version 9 for macOS[86,123]. The focus of the analyses was hypothesis generation, therefore, multiple comparison corrections were not applied in this context. When interpreting results of this chapter in the Discussion section, the focus will be on variables which are both statistically and potentially also clinically significant.

I estimated **treatment response trajectories** for any given variable as exacerbation value minus two-week post-treatment follow-up value. I performed PCA on the exacerbation minus two-week follow-up values of VAS symptoms, blood and sputum inflammation to identify principal components which explained >75% cumulative variance. Next, I produced a dendrogram and performed hierarchical ward's clustering (using Euclidean distance metric) on identified principal components using the R package 'stats' and 'dendextend'[86,145]. The optimal number of clusters representing objectively defined treatment response 'trajectotypes' irrespective of crisis state was identified via optimal value of the Duda and Hart index and the pseudo-t^2 index[95]. This

enabled identification of a suitable cut-point to choose independent clusters to represent the newly defined post-treatment trajectotypes. Different numbers of clusters from 2 to 11 were evaluated (11 was the maximum number of clusters considered representing every possible option of a trajectotype representing no change from crisis, resolving in one or both of symptoms and inflammation as relevant, or worsening in one or both of inflammation as relevant).

When comparing differences in exacerbation treatment response according to type of exacerbation treatment and when comparing independent post-treatment trajectotypes, Kruskal-Wallis one way analysis of variance test was used for numerical data (since variables analysed did not follow a normal distribution based on the Shapiro-Wilk test results on variables and log-transformed variables of the same dataset in Chapter 4). Binary/categorical variables were compared using chi-squared.

In order to investigate whether a given post-treatment trajectotype represented a favourable resolution of the crisis, I determined which crises were followed by which post-treatment trajectotypes and whether the symptom and type of inflammatory improvement corresponded to the crisis state. Next, I investigated whether variables apart from symptoms and inflammatory biology can help to distinguish crisis events which are followed by an appropriate post-treatment trajectotype versus crises which are not followed by an appropriate resolution trajectotype. Symptom and inflammatory variables were not appropriate for this comparison since these variables were directly related to the definition of the trajectotypes. Hence including symptom and inflammatory variables in these comparisons could have confounded my analysis. The multivariable SES introduced in Chapter 2 fulfilled the criteria of a test for multivariable

context importance of potential biomarkers for predicting post-treatment resolution[100]. I applied SES ($p<0.1$ threshold) to the demographic, physiology, bacteriology, comorbidity and medication data included in Table 5.1. Exacerbation treatment data were also included to examine the relationship between SCS and/or antibiotic use at exacerbation with appropriate post-treatment resolution trajectotype. I then attempted to train and validate a machine learning classifiers (decision tree, random forest, KNN, SVM, neural network) to use significant variables to distinguish crisis events which would be followed by an appropriate resolution trajectotype from crisis events which would not. As in previous chapters, variable selection using SES and the development of machine learning classifiers was performed using approximately 80% of available data. In this chapter, variable selection and machine learning classifier training were performed separately for each crisis state (using 80% of crisis events of the particular state). The crisis-tailored approach of model development enabled a greater resolution for identifying discriminative variables. I divided the 80% portion into four folds as part of four-fold cross-validation. The cross-validation performance of the machine learning classifiers was used to select the best classifier type and hyper-parameter values for the algorithm. I used mean AUC as the metric to evaluate classification algorithm performance since the appropriate resolution versus non-appropriate resolution was a two-class problem. With the optimal classifier identified, I retrained this model on the entire 80% data used up until this point and then assessed the retrained model on the remaining approximately 20% data used as hold-out test set.

In addition to identifying exacerbation minus post-treatment trajectotypes based on measurements of symptoms and inflammatory indices on day 0 and day 14, I investigated different shape profiles of daily VAS score data entered by patients into a

symptom diary from exacerbation day 0 up until day 14. For each of cough, dyspnoea, sputum production and sputum purulence, for time series in which at least 7 of the 15 days (from exacerbation to 14 days post-exacerbation inclusive) had diary entries, I imputed the remaining VAS scores using a Kalman approach using the 'R' package "imputeTS"[159,160]. Subsequently, I standardized each 15-day time series. The k-Shape algorithm which uses a shape-based distance metric was used to identify a pre-specified number of clusters of similarly shaped time series since the shape profile of the time series data is prioritized using this approach[161]. For each symptom type, I compared centroids (representing the shape of time series representative of a given cluster) for the different shape-based time series clusters. I considered a range of pre-specified time series cluster numbers from 2 onwards for each symptom type until further clusters could not reveal any new shape profiles. For each symptom, I chose the most informative time series cluster shapes as the final symptom time series profiles, referred to as symptom shape profile trajectotypes.

When running code, I always set the programme 'seed' to a constant value in order to ensure consistent sample selection and shuffling when selecting participants for training and validation versus testing and producing different cross-validation folds. This is important for fair comparison of performance of different machine learning algorithms and for reproducibility of results.

Code

Code for machine learning analysis is provided in the Appendix Chapter 5 Code.

5.3 Results

The same 347 ‘exacerbations’ analysed in Chapter 4 were analysed in this chapter. I previously determined that the 347 ‘exacerbations’ could be more objectively understood as 44 low overall symptom burden state events, 159 symptomatic crises, 73 eosinophilic crises and 71 neutrophilic crises.

Majority of events were treated with both SCS and antibiotics

244 events (70%) were treated with both SCS and antibiotics. 39 events (11%) were treated with SCS alone, 60 events (17%) with antibiotics alone and 4 events (1%) received neither treatment. There was a significant association between treatment and COPD objective state (Table 5.2, p<0.01).

Table 5.2 | Treatment given for different COPD objective states.

		Low Overall Symptom Burden State (n=44)	**Symptomatic Crisis (n=159)**	**Eosinophilic Crisis (n=73)**	**Neutrophilic Crisis (n= 71)**	**P-Value**
Exacerbation Treatment Allocation, n (%)	SCS and Antibiotics	27 (61)	110 (69)	58 (79)	49 (69)	<0.01
	SCS Alone	9 (20)	18 (11)	11 (15)	1 (1)	
	Antibiotics Alone	7 (16)	29 (18)	3 (4)	21 (30)	
	Neither	1 (2)	2 (1)	1 (1)	0 (0)	

Definition of abbreviations: SCS = systemic corticosteroids.

Eosinophilic crises had a higher proportion of events treated with SCS alone (15% compared to 11% of symptomatic crises and only 1% of neutrophilic crises). Neutrophilic crises had the highest proportion of antibiotics alone treatments (30% compared to 18% for symptomatic and 4% for eosinophilic crises). There was no difference in lung function improvement for events according to treatment type (Table

5.3). Improvement in VAS sputum production and purulence was greatest for events treated with antibiotics alone and least for participants treated with SCS alone ($p<0.01$ in both cases). CRQ dyspnoea score also improved the most for events treated with antibiotics alone ($p=0.04$). Improvement in CRP, blood and sputum total cell count and neutrophilic inflammatory indices was most marked for events treated with antibiotics alone (CRP, blood total cell count and neutrophilic indices all $p<0.01$; sputum neutrophil count and percentage $p=0.01$ and $p=0.02$ respectively). Improvement in blood and sputum eosinophilic indices was greatest for events treated with SCS whether or not antibiotics were also received (blood eosinophil count and percentage both $p<0.01$; sputum eosinophil count and percentage $p=0.02$ and $p<0.01$ respectively).

Table 5.3 | Comparison of exacerbation minus two-week post-treatment follow-up characteristic values according to type of treatment received at exacerbation. Negative values indicate a lower value of the variable at exacerbation e.g. -0.10L FEV_1 post-bronchodilator indicates that the value of FEV_1 post bronchodilator was 0.10L lower at exacerbation than after two-week follow up.

Characteristic	**Both (n=244)**	**SCS alone (n=39)**	**Antibiotics alone (n =60)**	**P-Value**
FEV_1 post-bronchodilator (L)	-0.10 (-0.26 to 0.02)	-0.04 (-0.25 to 0.07)	-0.06 (-0.18 to -0.01)	0.33
Percentage predicted FEV_1 (%)	-4.1 (-9.5 to 0.5)	-2.5 (-10.3 to 2.1)	-3.0 (-7.7 to -0.2)	0.50
FEV_1 post-bronchodilator/FVC ratio	-0.0 (-0.1 to 0.1)	0.0 (-0.0 to 0.1)	0.0 (-0.1 to 0.1)	0.09
VAS cough (mm)	29 (6 to 50)	28 (1 to 49)	32 (13 to 47)	0.60
VAS dyspnoea (mm)	21 (6 to 43)	27 (2 to 57)	26 (12 to 46)	0.52
VAS sputum production (mm)	**23 (3 to 50)**	**9 (-4 to 25)**	**37 (17 to 54)**	**<0.01**
VAS sputum purulence (mm)	**28 (4 to 49)**	**3 (-4 to 19)**	**32 (17 to 58)**	**<0.01**
CRQ emotion (Units)	-0.71 (-1.43 to 0.00)	-0.71 (-1.22 to -0.14)	-0.57 (-1.86 to 0.00)	0.69
CRQ fatigue (Units)	-0.75 (-1.75 to 0.00)	-0.50 (-1.50 to -0.25)	-1.00 (-2.25 to -0.25)	0.42
CRQ mastery (Units)	-0.50 (-1.38 to 0.00)	-0.50 (-1.25 to 0.00)	-0.50 (-1.75 to 0.00)	0.84

CRQ dyspnoea (Units)	**-0.80 (-1.73 to -0.20)**	**-0.60 (-1.60 to 0.00)**	**-1.20 (-2.20 to -0.60)**	**0.04**
Blood procalcitonin (ng/mL)	-0.00 (-0.02 to 0.02)	0.00 (-0.01 to 0.02)	0.02 (-0.01 to 0.03)	0.29
Sputum neopterin (nmol/g)	8.2 (-10.7 to 66.9)	-1.4 (-7.4 to 1.8)	7.6 (-5.5 to 34.8)	0.48
Sputum sTREM (pg/L)	-90 (-3711 to 5482)	-3423 (-3983 to -2863)	-2399 (-5209 to 2050)	0.96
Blood creatine (µmol/L)	0.0 (-6.0 to 6.0)	-2.0 (-5.0 to 4.0)	2.0 (-4.0 to 10.0)	0.13
Blood urea (mmol/L)	-0.5 (-1.5 to 0.3)	-0.4 (-1.5 to 0.2)	-0.5 (-1.2 to 0.3)	0.13
eGFR (mL/min)	0 (-4 to 4)	0 (-7 to 4)	-2 (-10 to 1)	0.06
CRP (mg/L)	**1.0 (-3.5 to 22.0)**	**0.0 (-4.5 to 3.0)**	**10.5 (0.0 to 58.5)**	**<0.01**
Blood cell count (x10⁹cells/L)	**-0.9 (-3.5 to 0.7)**	**-1.1 (-3.7 to 0.3)**	**1.2 (-0.5 to 3.2)**	**<0.01**
Blood neutrophil count (x10⁹cells/L)	**-1.0 (-2.8 to 0.9)**	**-1.2 (-2.6 to 0.4)**	**1.1 (-0.4 to 2.8)**	**<0.01**
Blood eosinophil count (x10⁹cells/L)	**0.02 (-0.04 to 0.14)**	**0.02 (-0.04 to 0.11)**	**-0.01 (-0.06 to 0.02)**	**<0.01**
Percentage blood neutrophils (%)	**-1.6 (-7.4 to 4.4)**	**-2.7 (-8.1 to 0.6)**	**4.0 (-1.5 to 8.9)**	**<0.01**
Percentage blood eosinophils (%)	**0.5 (-0.4 to 1.7)**	**0.6 (-0.1 to 2.1)**	**-0.4 (-1.0 to 0.1)**	**<0.01**

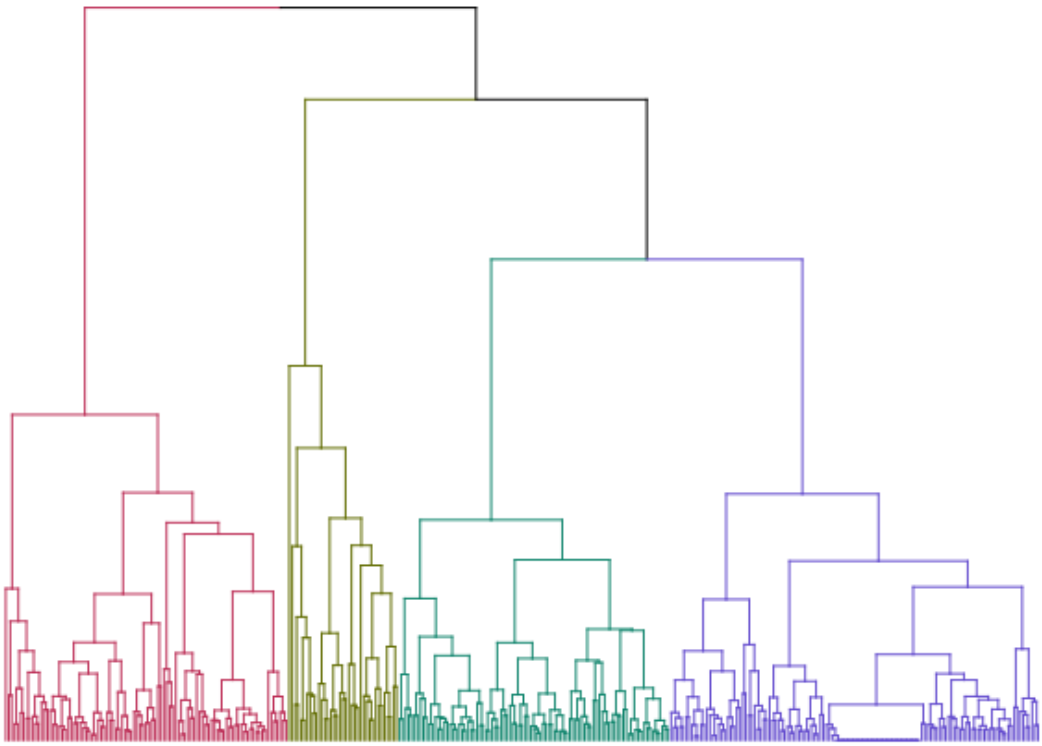
Sputum cell count ($x10^3$ cells/mg)	**3.1 (-0.4 to 14.2)**	**0.3 (-2.2 to 1.6)**	**5.2 (0.1 to 20.7)**	**0.01**
Sputum neutrophil count ($x10^3$ cells/mg)	**2.2 (-0.9 to 13.6)**	**0.3 (-2.0 to 1.1)**	**5.4 (0.3 to 21.1)**	**0.01**
Sputum eosinophil count ($x10^3$ cells/mg)	**0.05 (-0.00 to 0.21)**	**0.03 (-0.01 to 0.10)**	**0.01 (-0.01 to 0.08)**	**0.02**
Percentage sputum neutrophils (%)	**2.5 (-9.3 to 16.5)**	**-0.6 (-17.5 to 16.4)**	**10.8 (0.8 to 20.3)**	**0.02**
Percentage sputum eosinophils (%)	**0.0 (-0.3 to 1.0)**	**0.5 (-0.1 to 8.5)**	**0.0 (-0.8 to 0.2)**	**<0.01**

Definition of abbreviations: CRP = C-reactive protein; CRQ = Chronic Respiratory Disease Questionnaire; eGFR = estimated glomerular filtration rate; FEV_1 = forced expiratory volume after 1 second; FVC: forced vital capacity; SCS = systemic corticosteroids; VAS = Visual Analogue Scale.
Variable presentation and hypothesis tests: Measurements: variables are represented as median (lower quartile to upper quartile). All p-values for numerical variables are Kruskal-Wallis p-values.
Rows in bold font correspond to variables differing significantly across treatment types according to $p<0.05$.
Correction for multiple comparisons is not made since the analysis is hypothesis-generating.

Defining exacerbation minus two-week post-treatment follow-up trajectotypes

Five principal components (based on VAS cough, dyspnoea, sputum production, purulence, blood CRP, neutrophil and eosinophil counts and percentages in blood and sputum) explained 75% of cumulative variation (Appendix Table A5.1). I did not include the type of treatment received at the crisis event as a variable when defining post-treatment trajectotypes since I wished to capture symptomatological and inflammatory resolution patterns irrespective of treatment. This is because some crisis events treated differently may nevertheless resolve in a similar patter post-treatment. Similarly, not all crises receiving the same treatment respond the same way. I identified four clusters representing post-treatment trajectotypes (Figure 5.1).

Figure 5.1 | Dendrogram showing 4 exacerbation minus two-week post-treatment trajectotype clusters for 347 events.



These four trajectotypes corresponded to three patterns of resolution in symptoms and inflammation appropriate for particular crisis states and one pattern of no change (Table 5.4). Demographics such as age, sex and pack year history did not differ between trajectotypes. The four trajectotypes are:

Symptomatic and Eosinophilic Resolution Trajectotype followed 37 (11%) events. Lung function after treatment was 0.20L higher than at crisis. VAS cough and dyspnoea each improved by 36mm, VAS sputum production by 19mm and sputum purulence by 15mm. All eosinophilic inflammatory indices showed improvement ($p<0.01$ in all cases compared to no eosinophilic change for other trajectotypes). All of the events which were followed by this trajectotype had been treated with SCS.

No Change Trajectotype followed 91 (26%) events. Lung function after treatment, symptoms, neutrophilic and eosinophilic inflammatory indices were not improved. 69% of the events followed by this trajectotype had been treated with SCS and antibiotics.

Symptomatic Resolution Trajectotype followed 124 (36%) events. Lung function after treatment was on average 0.10L higher than at crisis. VAS cough improved by 37mm, dyspnoea by 27mm and sputum production as well as purulence each improved by 34mm and 37mm. Changes in inflammatory indices after treatment were either negligible or showed slight neutrophilic or eosinophilic inflammation reduction. 73% of the events followed by this trajectotype had been treated with both SCS and antibiotics.

Symptomatic and Neutrophilic Resolution followed 95 (27%) events. Lung function after treatment was also 0.10L higher than at crisis. VAS cough improved by 41mm, dyspnoea by 31mm, sputum production by 42mm and sputum purulence by 49mm. Treatment was associated with a substantial reduction in CRP and neutrophilic inflammation ($p<0.01$ in all cases). 61% of the events followed by this trajectotype had been treated with SCS and antibiotics and 36% of events with antibiotics alone.

Table 5.4 | Comparison of characteristics of crisis minus two-week post-treatment trajectotypes. Negative values indicate a lower value of the variable at crisis e.g. -0.20L FEV_1 post-bronchodilator indicates that the value of FEV_1 post bronchodilator was 0.20L lower at crisis than after two-week follow up. Age, sex, smoking status, pack year history, comorbidities and medication data are based on baseline measurements unless otherwise indicated. Other continuous variables are crisis minus two-week post-treatment measurement values at unless otherwise indicated.

Characteristic	Demographic Values or Crisis Minus Week 2 Post-Treatment Values				P-Value
	Symptomatic and Eosinophilic Resolution (n=37)	No change (n=91)	Symptomatic Resolution (n= 124)	Neutrophilic and Symptomatic Resolution (n= 95)	
Age (years)	68 (64 to 75)	67 (61 to 77)	69 (63 to 77)	69 (65 to 77)	0.45
Pack Year History	47 (37 to 63)	45 (31 to 51)	46 (33 to 58)	46 (34 to 60)	0.33
FEV_1 post-bronchodilator (L)	**-0.20 (-0.40 to -0.10)**	**-0.04 (-0.16 to 0.05)**	**-0.10 (-0.23 to 0.05)**	**-0.10 (-0.25 to -0.02)**	**<0.01**
Percentage predicted FEV_1 (%)	**-8 (-14 to -3)**	**-2 (-6 to 2)**	**-4 (-9 to 1)**	**-4 (-10 to -1)**	**<0.01**
FEV_1 post-bronchodilator/FVC Ratio	-0.03 (-0.06 to 0.06)	-0.02 (-0.10 to 0.04)	0.01 (-0.04 to 0.06)	-0.01 (-0.06 to 0.08)	0.12
VAS cough (mm)	**36 (8 to 52)**	**4 (-3 to 21)**	**37 (23 to 51)**	**41 (22 to 60)**	**<0.01**
VAS dyspnoea (mm)	**36 (13 to 59)**	**5 (-8 to 19)**	**27 (13 to 47)**	**31 (13 to 51)**	**<0.01**
VAS sputum production (mm)	**19 (2 to 50)**	**0 (-9 to 10)**	**34 (18 to 52)**	**42 (23 to 64)**	**<0.01**
VAS sputum purulence (mm)	**15 (0 to 40)**	**2 (-8 to 10)**	**37 (24 to 54)**	**49 (28 to 67)**	**<0.01**
CRQ emotion (Units)	**-0.86 (-1.85 to -0.29)**	**-0.14 (-0.72 to 0.14)**	**-0.86 (-1.71 to -0.14)**	**-0.86 (-1.71**	**<0.01**
CRQ fatigue (Units)	**-1.50 (-2.88 to -0.50)**	**-0.25 (-1.00 to 0.00)**	**-1.00 (-2.00 to -0.25)**	**-1.00 (-2.00 to -0.50)**	**<0.01**
CRQ mastery (Units)	**-1.00 (-2.00 to -0.25)**	**-0.25 (-0.75 to 0.25)**	**-0.50 (-1.75 to 0.00)**	**-0.50 (-1.75 to 0.00)**	**<0.01**
CRQ dyspnoea (Units)	**-1.40 (-2.80 to -0.40)**	**-0.40 (-1.00 to 0.20)**	**-0.80 (-1.80 to -0.20)**	**-1.20 (-2.20 to -0.40)**	**<0.01**
Total blood cell count ($x10^9$cells/L)	**-4.1 (-7.0 to -1.4)**	**-1.1 (-3.1 to 0.3)**	**-2.0 (-4.0 to -0.4)**	**2.2 (0.5 to 3.9)**	**<0.01**
Blood neutrophil count ($x10^9$cells/L)	**-3.0 (-7.5 to -1.2)**	**-1.0 (-2.2 to 0.2)**	**-1.8 (-3.5 to -0.7)**	**2.1 (0.8 to 4.0)**	**<0.01**

Blood eosinophils Count (x10⁹cells/L)		**0.4 (0.2 to 0.5)**	**0.0 (-0.0 to 0.1)**	**0.1 (-0.0 to 0.1)**	**-0.0 (-0.1 to 0.0)**	**<0.01**
Percentage blood neutrophils (%)		**-9.8 (-17.6 to -1.2)**	**-1.2 (-7.0 to 3.3)**	**-5.7 (-10.7 to -2.0)**	**6.1 (3.1 to 12.9)**	**<0.01**
Percentage blood eosinophils (%)		**5.5 (2.7 to 7.1)**	**0.4 (-0.2 to 0.8)**	**0.8 (0.2 to 1.6)**	**-0.6 (-1.4 to -0.2)**	**<0.01**
CRP (mg/L)		**0.0 (0.0 to 5.5)**	**0.0 (-10.0 to 5.3)**	**0.0 (-5.5 to 10.5)**	**27.5 (4.0 to 88.5)**	**<0.01**
Total sputum cell count (x10^3 cells/mg)		**1.0 (-1.8 to 3.7)**	**0.2 (-4.2 to 5.0)**	**2.4 (0.0 to 10.6)**	**13.8 (3.2 to 27.2)**	**<0.01**
Sputum neutrophil count (x10^3 cells/mg)		**-0.4 (-2.6 to 0.6)**	**-0.2 (-3.4 to 2.2)**	**2.2 (0.2 to 10.7)**	**15.0 (3.5 to 29.3)**	**<0.01**
Sputum eosinophil count (x10^3 cells/mg)		**1.0 (0.2 to 1.9)**	**0.0 (-0.0 to 0.0)**	**0.0 (-0.0 to 0.1)**	**0.1 (0.0 to 0.1)**	**<0.01**
Percentage sputum neutrophils (%)		**-22.4 (-35.8 to -16.5)**	**-3.2 (-20.5 to 3.5)**	**7.8 (-1.8 to 18.6)**	**15.3 (4.3 to 24.5)**	**<0.01**
Percentage sputum eosinophils (%)		**23.8 (11.0 to 30.9)**	**0.0 (-0.5 to 0.8)**	**0.0 (-0.3 to 0.8)**	**0.0 (-0.5 to 0.0)**	**<0.01**
Colony Forming Units (x10^7 per ml)		-0.1 (-0.6 to 0.3)	0.0 (-3.7 to 5.9)	0.0 (-1.7 to 9.7)	2.6 (-0.6 to 17.6)	0.15
HI Positive, n (%)		**10 (38)**	**34 (50)**	**61 (67)**	**46 (61)**	**0.03**
SA Positive, n (%)		3 (12)	4 (6)	10 (11)	2 (3)	0.16
SP Positive, n (%)		10 (38)	27 (40)	40 (44)	36 (48)	0.73
MC Positive, n (%)		7 (27)	30 (44)	35 (38)	38 (51)	0.15
Exacerbation Treatment, n (%)	**SCS and Antibiotics**	**32 (86)**	**63 (69)**	**91 (73)**	**58 (61)**	**<0.01**
	SCS alone	**5 (14)**	**17 (19)**	**14 (11)**	**3 (3)**	
	Antibiotics alone	**0 (0)**	**11 (12)**	**15 (12)**	**34 (36)**	
	Neither	**0 (0)**	**0 (0)**	**4 (3)**	**0 (0)**	
Objective COPD state, n (%)	**Low overall symptom burden state**	**1 (3)**	**26 (29)**	**12 (10)**	**5 (5)**	**<0.01**

	Symptomatic crisis	**0 (0)**	**46 (51)**	**75 (60)**	**38 (40)**	
	Eosinophilic crisis	**36 (97)**	**12 (13)**	**19 (15)**	**6 (6)**	
	Neutrophilic crisis	**0 (0)**	**7 (8)**	**18 (15)**	**46 (48)**	
		Number of participants who had an event followed by symptomatic and eosinophilic resolution (n=25)	**Number of participants who had an event followed by no change (n=65)**	**Number of participants who had an event followed by symptomatic resolution (n= 85)**	**Number of participants who had an event followed by neutrophilic and symptomatic resolution (n= 66)**	**P-Value**
Male, n (%)		21 (84)	42 (65)	55 (65)	41 (62)	0.24
Smoking status, n (%)	Current smoker, n (%)	8 (33)	23 (37)	34 (41)	25 (39)	0.78
	Ex-smoker, n (%)	16 (67)	38 (60)	47 (57)	36 (56)	
	Never smoker, n (%)	0 (0)	2 (3)	1 (1)	3 (5)	
SABA, n (%)		23 (100)	60 (95)	79 (96)	58 (94)	0.61
SACH, n (%)		4 (17)	8 (13)	16 (20)	10 (16)	0.75
LACH, n (%)		18 (78)	46 (73)	50 (61)	38 (61)	0.22
LABA, n (%)		21 (91)	50 (79)	67 (82)	50 (81)	0.64
ICS, n (%)		22 (96)	59 (94)	72 (88)	55 (89)	0.50
Maintenance prednisolone, n (%)		3 (13)	4 (6)	7 (9)	4 (6)	0.73
Theophylline, n (%)		5 (22)	10 (16)	17 (21)	10 (16)	0.81
Mucolytic, n (%)		4 (17)	11 (17)	14 (17)	9 (15)	0.97
Oxygen, n (%)		1 (4)	6 (10)	11 (13)	4 (6)	0.42
Prophylactic antibiotic, n (%)		2 (9)	3 (5)	7 (9)	3 (5)	0.72
Cardiovascular comorbidity, n (%)		11 (50)	34 (54)	47 (57)	42 (68)	0.33
Diabetes, n (%)		2 (9)	6 (10)	8 (10)	8 (13)	0.91
Endocrine comorbidity, n (%)		1 (5)	2 (3)	4 (5)	3 (5)	0.96

Depression, n (%)	1 (5)	11 (17)	16 (20)	12 (19)	0.40
Osteopenia/Osteoporosis, n (%)	5 (23)	8 (13)	14 (17)	12 (19)	0.66
Anaemia, n (%)	0 (0)	3 (5)	3 (4)	1 (2)	0.60

Definition of abbreviations: CRP = C-reactive protein. SABA: short acting beta agonist; LABA: long acting beta agonist; SACH; short acting anticholinergic; LACH: long acting anticholinergic; FEV_1 : forced expiratory volume after 1 second; FVC: forced vital capacity; HI: Haemophilus influenzae; MC: Moxarella catarrhalis; SA: Staphylococcus aureus; SCS = systemic corticosteroids; SP: Streptococcus pneumoniae;
Variable presentation and hypothesis tests: variables are represented as median (lower quartile to upper quartile). Binary and categorical variables are presented as number (%) of instances. For categorical variables, percentages for each column may deviate slightly from expected percentage based on the total number of measurements given by the column header due to the presence. All p-values for numerical variables are Kruskal-Wallis p-values. P-values for binary/categorical variables are chi-squared p-values.
Correction for multiple comparisons is not made since the analysis is hypothesis-generating.
Use of trajectotype number of events or number of participants who had a particular trajectotype: categorical variables reflecting demographics, baseline medication usage and comorbidities are presented based on the number of participants who had at least one event followed by a particular trajectotype rather than on the number of events followed by the trajectotype.
Rows in bold font correspond to variables differing significantly across trajectotypes according to $p<0.05$.

Post-treatment trajectotypes correspond to related crisis states but not all crises undergo appropriate resolution in symptoms and/or inflammatory biology

I determined the proportion of events followed by a trajectotype which belonged to each crisis state (Table 5.5). The associations were statistically significant ($p<0.01$). All but one event which were followed by the 'symptomatic and eosinophilic resolution' trajectotype had belonged to the eosinophilic crisis state. The majority (51%) of events which were followed by the 'no change' trajectotype had belonged to the symptomatic crisis state. Similarly, the majority (60%) of events which were followed by the 'symptomatic resolution' trajectotype had also belonged to the symptomatic crisis state.

Table 5.5 | Crisis minus two week post-treatment trajectotype following each objective COPD state.

		Low overall symptom burden state (n=44)	Symptomatic crisis (n=159)	Eosinophilic crisis (n=73)	Neutrophilic crisis (n= 71)	P-Value
Crisis minus two week post-treatment trajectotype	No change	26 (59)	46 (29)	12 (16)	7 (10)	<0.01
	Symptomatic resolution	12 (27)	75 (47)	19 (26)	18 (25)	
	Symptomatic and eosinophilic resolution	1 (2)	0 (0)	36 (49)	0 (0)	
	Symptomatic and neutrophilic resolution	5 (11)	38 (24)	6 (8)	46 (65)	

P-value is chi-squared p-value.

I wished to characterise the proportion of events of each crisis state which were not followed by appropriate treatment response trajectotypes:

- For symptomatic crises, an appropriate resolution could be represented by any symptomatic resolution trajectotype irrespective of whether inflammatory change is involved. According to this definition, 113 (71%) symptomatic crises achieved resolution.
- For eosinophilic crises, an appropriate resolution would be represented by the symptomatic and eosinophilic resolution trajectotype to account for both symptoms and inflammation. According to this definition, 36 (49%) eosinophilic crises achieved resolution.
- For neutrophilic crises, an appropriate resolution would be represented by the symptomatic and neutrophilic resolution trajectotype to account for both symptoms and inflammation. According to this definition, 46 (65%) neutrophilic crises achieved resolution.
- For the low overall symptom burden state, none of the trajectotypes including the ‘no change’ trajectotype would be inappropriate.

Lung function, bacteriology, exacerbation treatment and baseline medication differ between crises followed by appropriate resolution trajectotype and crises not followed by appropriate resolution trajectotype

Demographics (see variable list in Table 5.1) were not associated with whether a crisis followed the correct resolution. For symptomatic crises, a lower lung function at crisis (SES p=0.07) and the presence of Haemophilus influenzae (HI) at crisis (SES p=0.01)

were associated with a symptomatic resolution trajectotype (Table 5.6). Eosinophilic crises were more likely to be followed by resolution in both eosinophilia and symptoms if treated with SCS, if the crisis was associated with poorer lung function and if the participant was taking short acting anticholinergic (SACH) medication (SES p=0.09, 0.05 and 0.08 respectively). The SES algorithm identified the absence of depression to be statistically equivalent to whether the participant was taking SACH in the multivariate context of predicting that an eosinophilic crisis would be followed by both symptomatic and eosinophilic resolution. Finally, neutrophilic crises were more likely to be followed by both inflammatory and symptomatological resolution if the crisis was associated with Moxarella catarrhalis (MC) (SES p=0.07). Usage of baseline medication such as SACH and maintenance prednisolone was associated with lack of appropriate resolution trajectotype (SES p=0.08 and SES p=0.06 respectively).

Table 5.6 | Difference between crises followed by appropriate trajectotypes versus crises not followed by an appropriate trajectotype for variables identified by SES algorithm. Variables shown either passed the p<0.1 threshold in the SES test or were identified as potentially equivalent to a variable which passed the p<0.1 threshold.

	Crisis Type		**Statistically Equivalent Signature P-Value**
Characteristic	**Appropriate resolution achieved**	**Did not achieve appropriate resolution**	
Symptomatic crisis			
Characteristic	Appropriate resolution achieved, n = 113	Did not achieve appropriate resolution, n = 46	
FEV_1 post bronchodilator (L)	**1.10 (0.79 – 1.44)**	**1.24 (0.84 – 1.71)**	**0.07**
Haemophilus influenza, n (%)*	**58 (67)**	**15 (39)**	**0.01**
Eosinophilic crisis			
Characteristic	Appropriate resolution achieved, n = 36	Did not achieve appropriate resolution, n = 37	
SCS given at exacerbation, n (%)	**36 (100)**	**33 (89)**	**0.09**
Percentage predicted FEV_1 post-bronchodilator	**27 (23 – 35)**	**38 (26 – 56)**	**0.05**
FEV_1 post-bronchodilator (L)	0.77 (0.60 – 1.02)	1.10 (0.73 – 1.38)	0.99
FEV_1 post-bronchodilator/ FVC ratio	0.40 (0.31 – 0.49)	0.48 (0.36 – 0.57)	0.67
	Number of patients with eosinophilic crisis who achieved appropriate resolution (n=24)	Number of patients with eosinophilic crisis who did not achieve appropriate resolution (n=22)	
SACH, n (%)*	**4 (18)**	**2 (10)**	**0.08**
Depression, n (%)*	1 (5)	4 (20)	0.13
Neutrophilic crisis			
	Appropriate resolution achieved, n = 46	Did not achieve appropriate resolution, n = 25	
MC, n (%)*	**20 (54)**	**6 (32)**	**0.07**

	Number of patients with neutrophilic crisis who achieved appropriate resolution (n=32)	Number of patients with neutrophilic crisis who did not achieve appropriate resolution (n=24)	
SACH, n (%)*	**2 (7)**	**5 (22)**	**0.08**
Maintenance prednisolone, n (%)*	**3 (10)**	**5 (22)**	**0.06**

Definition of abbreviations: SACH = short acting anticholinergic; FEV_1 = forced expiratory volume after 1 second; FVC = forced vital capacity; HI: Haemophilus influenzae; MC = Moxarella catarrhalis; SCS = systemic corticosteroids; VAS = Visual Analogue Scale.
Variable presentation and hypothesis tests: numerical variables are presented as median (lower quartile to upper quartile). Binary variables are presented as number and percentage of total. For SACH, maintenance prednisolone and depression, the number and proportions of instances shown in the table correspond to the number of patients with a particular crisis resolution rather than the number of crisis resolution events. Note that the p-values are from the SES test and correspond to the 80% of the 347 samples used for classifier training to predict appropriate resolution i.e. 133 symptomatic crises, 51 eosinophilic crises and 60 neutrophilic crises.
Note that percentages for each column may deviate slightly from expected percentage based on the total number of measurements given by the column header due to the presence of missing values.
***Data availability:**
Symptomatic crisis resolution: There were 87 available Haemophilus influenza measurements and 38 available Haemophilus influenza measurements for symptomatic appropriate resolution achieved and appropriate resolution not achieved respectively.
Neutrophilic crisis resolution: There were 37 available Moxarella catarrhalis measurements and 19 available Moxarella catarrhalis measurements for neutrophilic appropriate resolution achieved and appropriate resolution not achieved respectively. There were 30 available patient short acting anticholinergic as well as maintenance prednisolone measurements and 23 available patient short acting anticholinergic as well as maintenance prednisolone measurements for neutrophilic appropriate resolution achieved and appropriate resolution not achieved respectively.
Eosinophilic crisis resolution: There were 22 available patient short acting anticholinergic measurements and 20 available patient short acting anticholinergic measurements for eosinophilic appropriate resolution achieved and appropriate resolution not achieved respectively. There were 21 available patient depression measurements and 20 available patient depression measurements for eosinophilic appropriate resolution achieved and appropriate resolution not achieved respectively.

Rows in bold font correspond to variables with SES algorithm $p<0.1$.

Using FEV_1 post-bronchodilator together with the presence of HI at crisis, the best classifier for predicting the occurrence of an appropriate resolution trajectotype following a symptomatic crisis was the KNN with cross-validation AUC of 0.68 (Table 5.7). A KNN classifier using SCS prescription at crisis, percentage predicted FEV_1 post-bronchodilator and SACH medication achieved a cross-validation AUC of 0.72 for predicting the occurrence of an appropriate resolution trajectotype following an eosinophilic crisis (Table 5.7). The best classifier for predicting the occurrence of an appropriate resolution trajectotype following a neutrophilic crisis using the presence of MC and whether a participant was taking SACH and maintenance prednisolone was the SVM (Table 5.7). A cross-validation AUC of 0.79 was achieved.

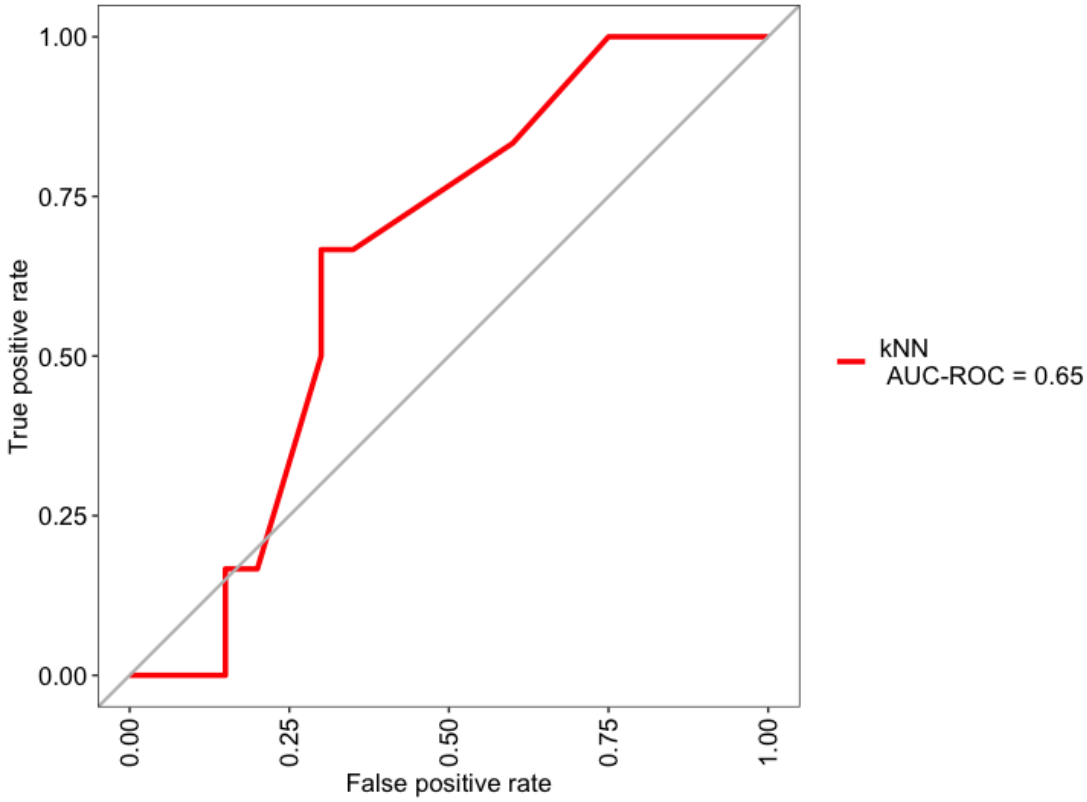
Table 5.7 | Average cross-validation AUC performance of optimised algorithms to predict whether appropriate resolution was achieved by particular crisis state.

Model	Average cross-validation AUC	Best hyper-parameter values
Symptomatic crisis		
Random Forest	0.62	500 trees, mtry parameter = 1
KNN	**0.68**	**K=23**
Decision Tree	0.56	Complexity parameter = 0.00
SVM	0.62	Sigma = 0.13, cost penalty = 16
Neural Network	0.65	Size = 9, decay = 0.1
Eosinophilic crisis		
Random Forest	0.71	200 trees, mtry parameter = 1
KNN	**0.72**	**K=25**
Decision Tree	0.70	Complexity parameter = 0.037
SVM	0.65	Sigma = 6.0, cost penalty = 8
Neural Network	0.69	Size = 6, decay = 0.1
Neutrophilic crisis		
Random Forest	0.75	200 trees, mtry parameter = 1*
KNN	0.68	K=19
Decision Tree	0.49	Complexity parameter = 0.10
SVM	**0.79**	**Sigma = 0.25, cost penalty = 2**
Neural Network	0.73	Size = 1, decay = 0.2

Classifier models in bold font have best performance.
*Random forests with 200 trees up to 1,000 trees all achieved area under receiver operating characteristic curve of 0.75.

Applying the classifiers on the hold-out test set events, it was revealed that the occurrence of an appropriate resolution trajectotype was only possible in the case of symptomatic crises (AUC of 0.65; Figure 5.2). In contrast, the occurrence of an appropriate resolution trajectotype following an eosinophilic crisis and following a neutrophilic crisis could not be achieved on the test set (AUCs of 0.47 and 0.48 respectively).

Figure 5.2 | Test set ROC curve performance of KNN classifier to predict occurrence of appropriate resolution trajectotype following a symptomatic crisis.



Shape profiles of daily VAS symptom time series reveal greater post-treatment resolution personalisation than a single two-week follow-up

The fact that a crisis is followed by a 'no change' trajectotype as determined by comparing day 0 to day 14 symptoms does not rule out that symptom improvement may temporarily occur during some phases between day 0 and day 14 e.g. at the beginning of treatment. Similarly, if crisis events are followed by one of the symptom improvement trajectotypes (with or without improvement in inflammation), it is possible that symptom improvement begins immediately after treatment, or that there is a delay before an adequate treatment response occurs. I therefore set out to investigate the shapes of symptom time series profiles from VAS diary cards completed daily from the time of treatment to follow-up at day 14. Data were available for 97 exacerbations from a total of 71 patients, treated with antibiotics as well as either SCS (70%) or placebo (30%). For each symptom type, I identified different shape profiles among the time series for the 97 exacerbations. Of the 97 exacerbations, there were 21 exacerbations had a 'no change' trajectotype, of which 14 corresponded to crises rather than low overall symptom burden/equilibrium. From the 97 VAS score time series available post-exacerbation, I identified the following time series shape profiles (Figure 5.3):

Fast resolution: cough: n = 24; dyspnoea: n=26; sputum production: n=28; sputum purulence: n=34.

Gradual resolution: cough: n=43; dyspnoea: n=54; sputum production: n=69; sputum purulence: n=63.

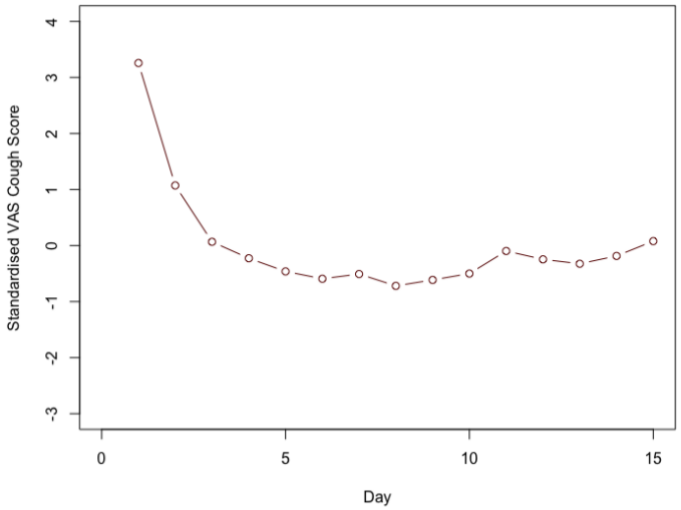
Oscillatory resolution/short term improvement followed by worsening: cough: n=7; dyspnoea: n=17.

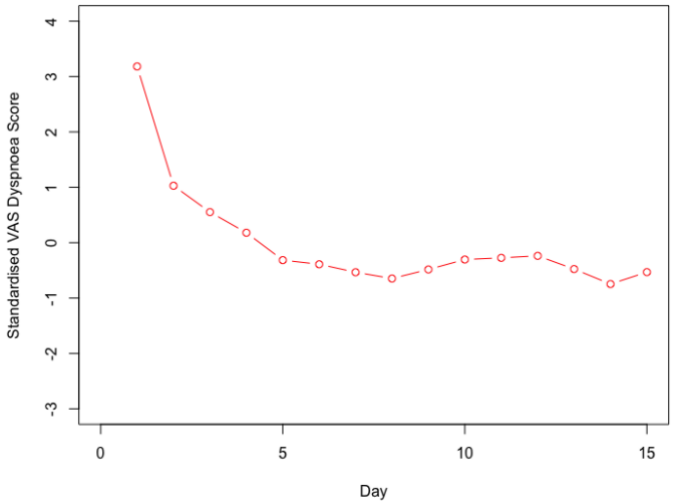
Delayed resolution: cough n=23.

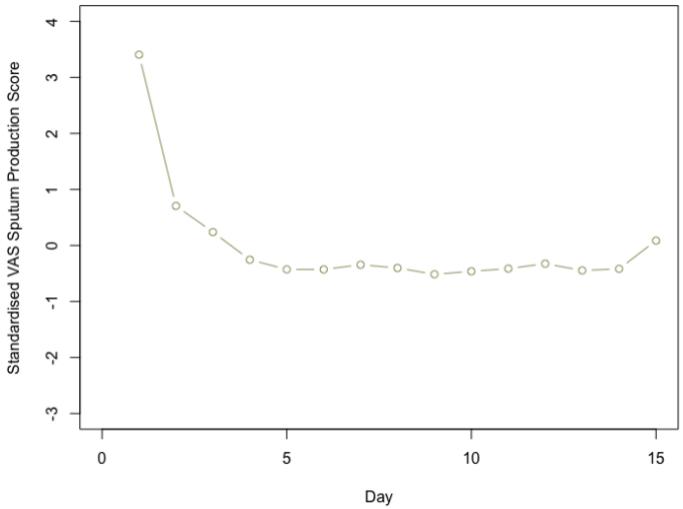
The oscillatory pattern for VAS cough shape profiles seems to correspond to eventual resolution. However, for oscillatory VAS dyspnoea shape profiles, the VAS dyspnoea standardised score is almost the same at the end of the 14 days post-treatment as the score at the time of exacerbation. Hence, the oscillatory pattern for VAS dyspnoea shape profiles may be indicative of a significant short term improvement followed by worsening with or without resolution beyond the two week post-treatment follow-up. Of the 17 short term improvement followed by worsening dyspnoeas shape profiles, 4 corresponded to ‘no change’ trajectotypes following a crisis. This indicated that at least 4 crisis events with no resolution by day 14 post-treatment did in fact have a short-lived improvement in dyspnoea during the initial days post-treatment.

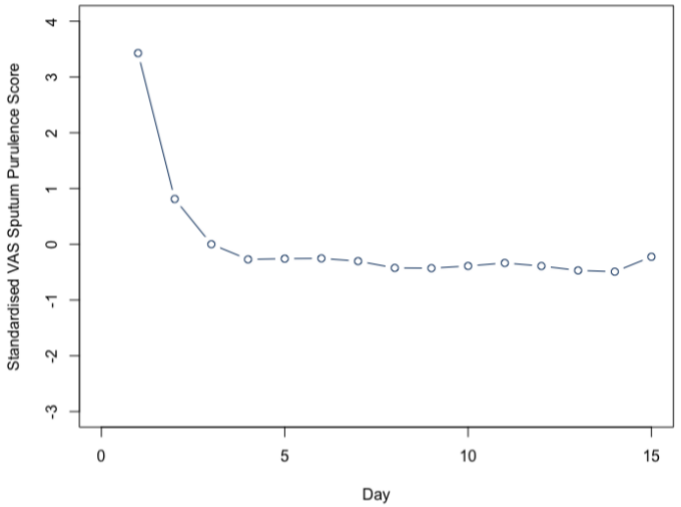
Figure 5. 3 | Centroids depicting characteristic time series shape profile trajectotypes for different VAS symptom scores. A: Fast resolution. B: Gradual resolution. C: Oscillatory resolution/short term improvement followed by worsening. D: Delayed resolution.

A

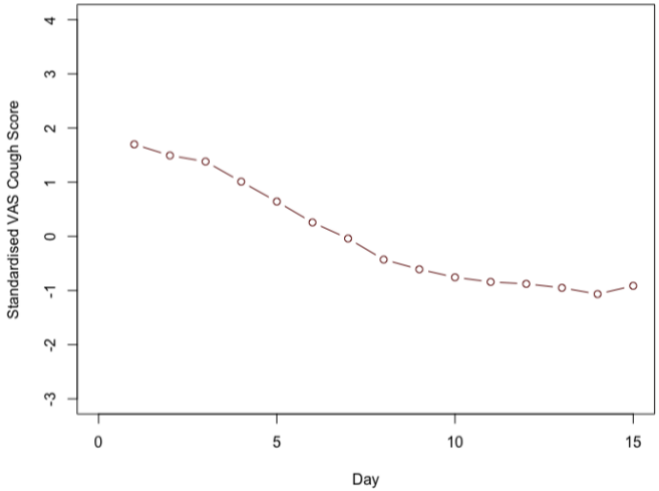


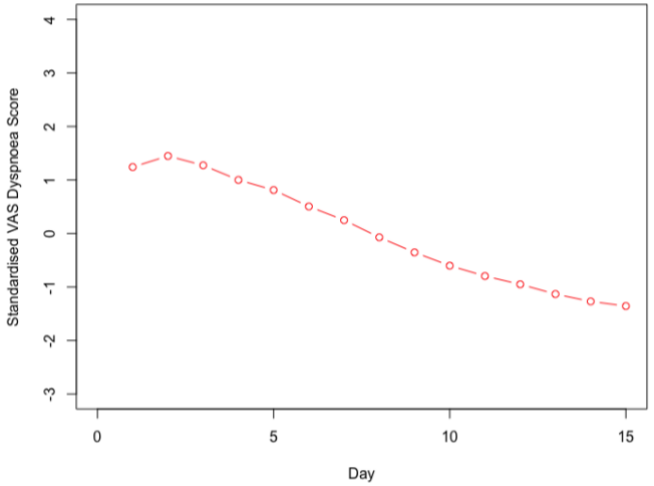


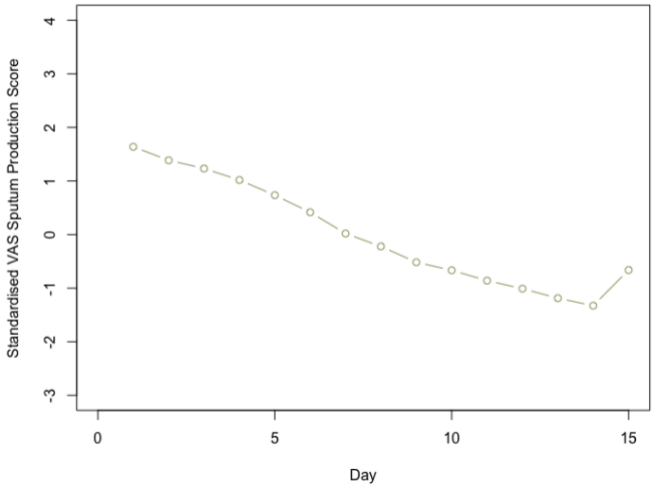


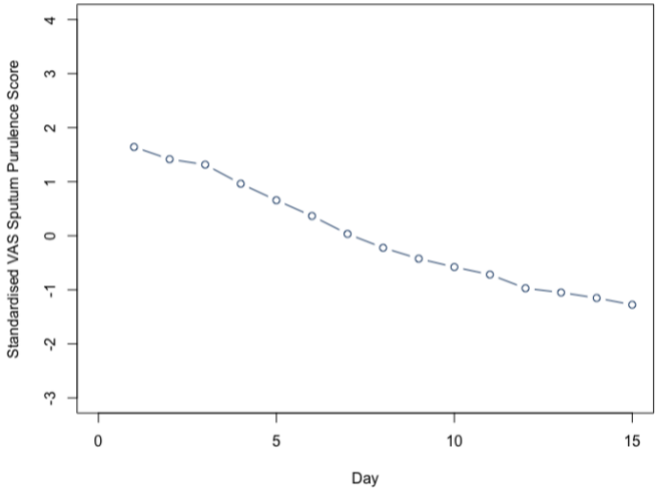


B

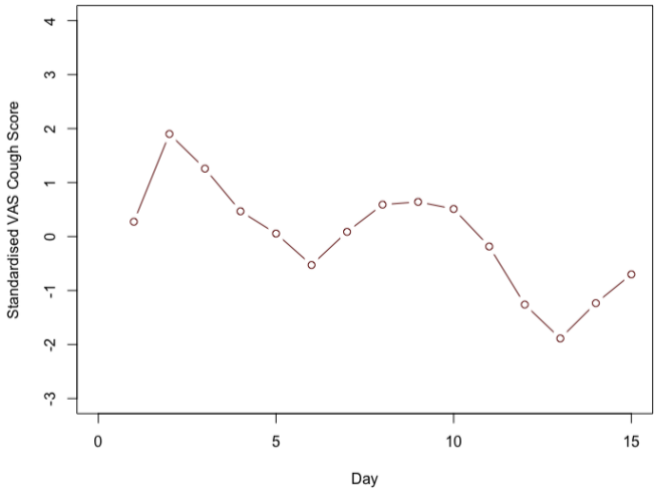

Standardised VAS Cough Score
4
3
2
1
0
-1
-2
-3
0
5
10
15
Day

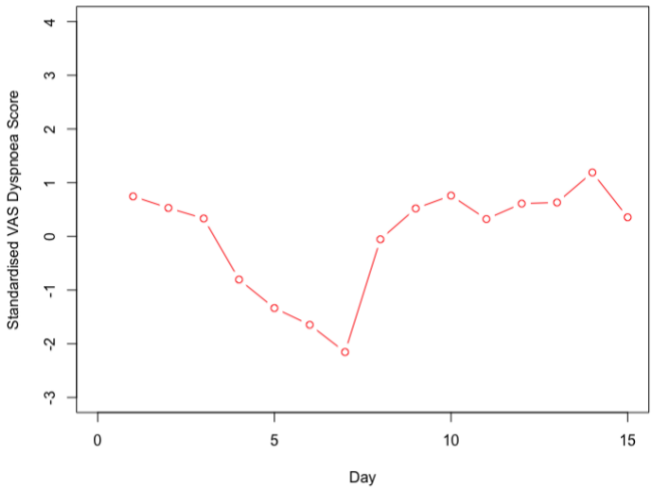

Standardised VAS Dyspnoea Score
4
3
2
1
0
-1
-2
-3
0
5
10
15
Day


Standardised VAS Sputum Production Score
4
3
2
1
0
-1
-2
-3
0
5
10
15
Day

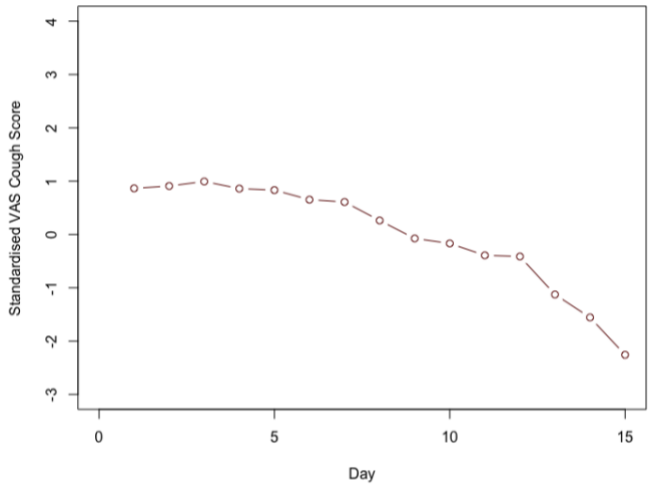

Standardised VAS Sputum Purulence Score
4
3
2
1
0
-1
-2
-3
0
5
10
15
Day

C


Standardised VAS Cough Score
-3
-2
-1
0
1
2
3
4
0
5
10
15
Day


Standardised VAS Dyspnoea Score
-3
-2
-1
0
1
2
3
4
0
5
10
15
Day

D

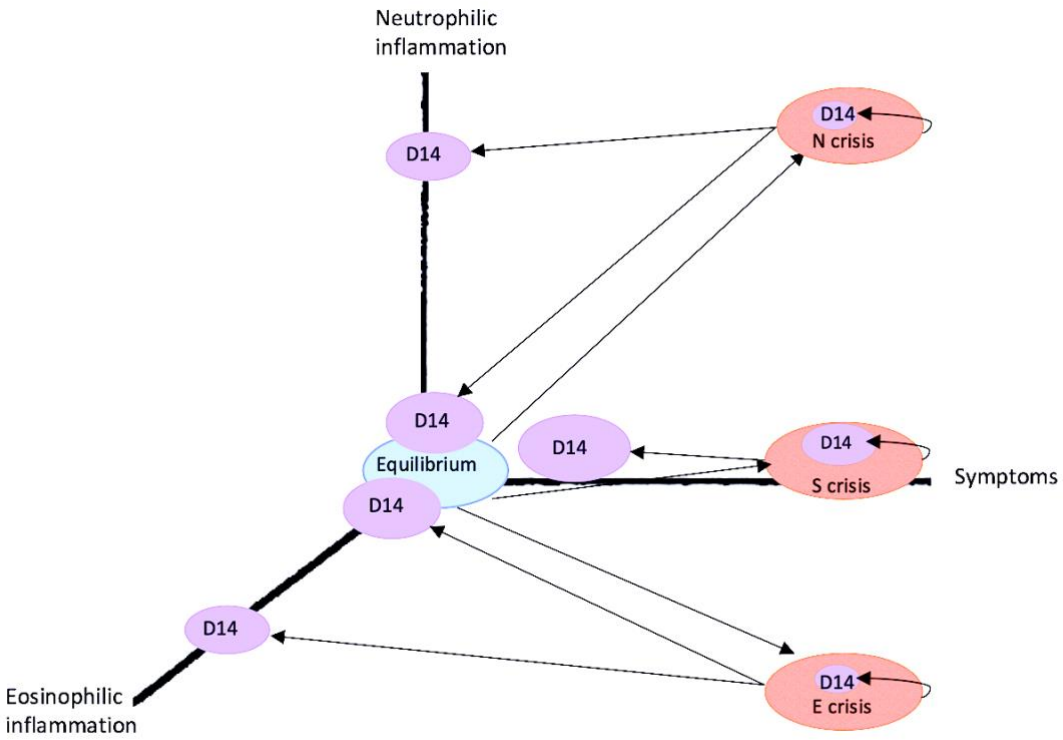

Standardised VAS Cough Score
-3
-2
-1
0
1
2
3
4
0
5
10
15
Day

5.4 Discussion

Overall findings

In this chapter I have mapped out the progression from each crisis state to different post-treatment trajectotypes based on two-week post-treatment follow-up data. This chapter also serves to identify characteristics associated with the correspondence of a particular crisis state to the appropriate post-treatment trajectotype to work towards predicting appropriate post-treatment trajectotype based on crisis treatment type, bacteriology, lung function and baseline medication usage. Figure 5.4 summarises the conceptual findings of this chapter.

Figure 5.4 | Schematic showcasing the transition of patients from equilibrium/low overall symptom burden state to one of symptomatic crisis (S crisis), eosinophilic crisis (E crisis) or neutrophilic crisis (N crisis) and subsequently at day 14 post-crisis (D14) in a three-dimensional space of symptoms, eosinophilic and neutrophilic inflammation. Equilibrium/low overall symptom burden state, crisis states and D14 post-crisis follow-up time points are represented as blue, red and purple respectively.



Further interpretation, strengths of analysis and comparison with existing studies

I found four trajectotypes: ‘no change’; ‘symptomatic resolution’; ‘symptomatic and eosinophilic resolution’; ‘symptomatic and neutrophilic resolution’. 26% of eosinophilic crises showed ‘symptomatic resolution’ after treatment but featured little improvement in blood eosinophilia, while 16% of eosinophilic crises events in the eosinophilic exacerbation state showed neither symptomatological nor inflammatory improvement. The 26% of eosinophilic crises which underwent symptom improvement but little eosinophilic inflammatory improvement necessitate further investigation. It is known from other studies that patients with eosinophilic ‘exacerbations’ tend to have a higher risk of ‘exacerbations’ recurring. For example, Chan et al performed a retrospective cohort analysis of 247 patients and showed that patients with stable state blood eosinophil percentage ≥2% had a greater frequency of exacerbations compared to patients with a blood eosinophil percentage < 2% during the 12 months following the index complete blood count (mean exacerbation 1.07 vs 0.34, $p < 0.001$)[162]. In Chapter 3 and Chapter 4 I discussed evidence in support of the benefits of directing SCS therapy according to whether exacerbations are characterised by eosinophilia, as shown by the studies of Bafadhel et al and Sivapalan et al[82,83]. However, my analysis is the first to define cluster-based treatment response subgroups to reveal eosinophilic crises where symptom recovery occurs following treatment but without eosinophilic inflammatory recovery. This chapter’s finding that eosinophilic crises treated with SCS are more likely to be followed by both symptom and inflammatory resolution provides further support. Future studies should be assessed whether patients with eosinophilic crises which improve in terms of symptoms but not eosinophilic inflammation are at a higher risk of future crises compared to patients whose eosinophilic crises improve in

terms of both symptoms and eosinophilic inflammation. The association of depression with eosinophilic crises which did not undergo complete resolution compared to eosinophilic crises which were followed by appropriate resolution merits further investigation. My finding that eosinophilic crisis with poorer long function are more likely to resolve may reflect the fact that current treatments are more successful in addressing airflow limitation but not all eosinophilic crises have marked airflow limitation which may be less pronounced than in neutrophilic crises.

The majority (65%) of neutrophilic crises were followed by symptomatic and neutrophilic resolution trajectotype, with 25% of neutrophilic crises followed by symptom improvement without neutrophilic inflammation improvement and only 10% showed neither symptomatological nor inflammatory improvement. Interestingly, studies indicate that non-recovery in CRP levels after an exacerbation is associated with an increased risk of future exacerbations. An example is the Perera et al (2007) study in which systemic markers of inflammation including CRP were measured for 73 patients at stable state, exacerbation and at 7, 14 and 35 days post-exacerbations[163]. It was observed that 22% of patients who experienced recurrent exacerbations within 50 days had significantly higher levels of serum CRP ($p=0.007$) than patients without exacerbation recurrence within 50 days of the index exacerbation[163]. Furthermore, higher CRP levels were also correlated with a shorter time to next exacerbation ($p<0.01$)[163]. However, non-recovery of symptoms (determined by diary cards to reveal an increase in respiratory symptoms) by day 35 was also associated with high CRP levels compared to patients who had recovered ($p=0.03$)[163]. It remains to be seen whether my analysis' identification of a subgroup of crises with symptom recovery but without neutrophilic inflammation recovery (as determined by blood CRP, blood and

sputum neutrophil count and percentage) by day 14 corresponds to a higher risk group of patients for future crises. The association of bacterial infection at neutrophilic crisis with better post-treatment trajectotype may be due to the efficacy of antibiotics in this context. This should be investigated in a separate patient cohort. It is possible that the association of baseline medication usage with poorer outcomes in my analysis reflects that patients who have poorer respiratory health in general are both more likely to take such medication as well as have more severe crises.

It is noteworthy that 29% of symptomatic crises had no symptom resolution. This could be due to a number of reasons. The association with poorer lung function and the presence of bacterial infection at symptomatic crisis with superior outcomes found in my analysis further supports the view that it is the conventional respiratory and microbiological aspects of these crises which current treatment methods address. An open question is what mechanistically underpins symptomatic crises with little or know identifiable infection and little or know reduced lung function. However, my multivariable analysis did not reveal comorbidities such as cardiovascular and/or psychological comorbidity to be predictive of whether symptomatic resolution would occur. This needs to be further studied in future patient cohorts in real time where biomarkers of cardiovascular morbidity such as natriuretic peptides and quantitative real time measures of depression are included. The intricate patterns of VAS shape profile time series trajectotypes corresponding to daily VAS scores between the time of crisis and 14 days post-crisis may also explain the difficulty in achieving successful prediction of which symptomatic crises would be followed by a ‘no change trajectotype derived from a single follow-up measurement at day 14 post-crisis. The shape profile trajectotypes enable insight into the speed and consistency of symptom improvement

and/or worsening over a 14 day period following the time of crisis. This adds a layer of temporal and individualised information when assessing crisis treatment response. For each of cough, dyspnoea, sputum production and sputum purulence, two clear improvement shape profile trajectotypes were seen – rapid improvement within a few days to 1 week after crisis versus a gradual improvement over the course of the two weeks after crisis. It is of particular note that some shape profile trajectotypes for cough and dyspnoea included phases of both worsening and improvement over the course of the two-week post-crisis follow-up period. For example, one cough shape profile trajectotype was characterised by repeated cycles of improvement followed by worsening over the two weeks after crisis. Another cough shape profile trajectotype featured little improvement during the first week of crisis follow-up but gradual improvement during the second week. In the case of dyspnoea, complexity in the time series post-crisis was also manifest. Two dyspnoea shape profile trajectotypes featured clear improvement (in one case rapid, in the other case gradual) but a third dyspnoea shape profile trajectotype was characterised by initial improvement followed by worsening. In this context, it is possible that the symptomatic crisis 'no change' post-treatment trajectotype versus symptom resolution trajectotype prediction algorithm may be confounded by the nuances of different phases of symptom improvement and worsening during the two-week post-crisis follow-up period. The fact that some crises exhibit a worsening pattern or no change soon after beginning treatment but with improvements later during the two-week period, further complicates prediction of two-week treatment response for symptomatic crises. The number of available daily time series available was too small to attempt development of a classification algorithm to predict which exact symptomatic shape profile a symptomatic crisis two-timepoint trajectotype would correspond to.

SCS and antibiotic prescriptions for exacerbating patients in the study whose data I analysed in this Chapter and in Chapter 4 were for 14 days and 7 days respectively. It is possible that stopping antibiotic medication on day 7 may explain a shift from improvement to worsening for the corresponding dyspnoea shape profile trajectotype. It is also possible that patient exposure to environmental factors during the post-crisis two-week follow-up period plays a role. Sama et al conducted a case cross-over study of 168 COPD patients[164]. Baseline health surveys were completed as well as exposure surveys (addressing home, community and workplace activities and exposures during a one-week period) at the time of exacerbation as well as at three randomly chosen control points when a patient was not exacerbating[164]. In a comparison between exposures in the week preceding exacerbations versus when exacerbations were not reported to occur, variations in outdoor temperature, self-reported respiratory infections, exhaust fumes and scented laundry products were positively associated with exacerbation risk[164]. On the other hand, self-reported exercise was negatively associated with exacerbation risk[164]. Although this study assessed the association between activities and exposures with exacerbation risk, it is possible that similar activities and exposures could impact crisis treatment response. A future study in which activity and exposure surveys similar to those used by Sama et al but in the context of post-crisis follow-up would enable assessment of association between patient activities and exposures with shape profile trajectotypes.

Many studies have characterised the length of time from exacerbation to recovery and the degree to which lung function recovery occurs at different follow-up time points. The Seemungal et al (2000) study has a number of strengths in analysing exacerbation emergence and treatment response[165]. A cohort of 101 patients with moderate and

severe COPD patients were studied over 2.5 years including stable state measurements and 504 exacerbations[165]. Daily morning peak expiratory flow rate and changes in respiratory symptoms were recorded on diary cards, with 34 patients also reporting daily spirometry[165]. The advantage of this study set-up is that an increase in symptoms and reduction in lung function could be detected in the days leading up to an exacerbation as well as followed up post-exacerbation treatment[165]. Symptoms were grouped as major (dyspnoea, sputum purulence, sputum amount) or minor (wheeze, sore throat, cough, common cold symptoms)[165]. The median total symptom score recovery time after exacerbation was 7 days (lower quartile to upper quartile 4 – 14)[165]. An extent of individualisation was enabled by this study and exacerbations which would normally go unreported were captured due to the use of the daily diaries and the exacerbation definition used (at least two consecutive days of increase in any two major symptoms or increase in one major and one minor symptom according to criteria modified from Anthonisen et al[166]). However, this seminal study also reflects the limitations common in the field when assessing exacerbation treatment response. In Chapter 4 I discussed the problems associated with the reliance on conventional exacerbation versus stable state definitions when assessing response to treatment. The exacerbation definition used my Seemungal et al used binary-coded symptom variables instead of allowing a magnitude of change of individual symptoms to be recorded[165]. Biological inflammatory heterogeneity is also not considered. Studies which do examine changes in systemic inflammatory markers in during exacerbation treatment follow-up typically examine an entire study cohort or subgroups of a study cohort according to particular treatment interventions. For example, Bafadhel et al and Perera et al measured values of blood eosinophils and CRP respectively during treatment follow-up[82,163]. However, these and other studies do not define exacerbation

treatment response groups according to a multivariate change from the time of exacerbation to a follow-up time point. Therefore, patients could not be followed up according to their multivariate symptomatological and biological inflammation at crisis. Overall treatment response across different events for any given crisis state and treatment type may be dominated by events with a large treatment response despite the potential existence of events which do not exhibit improvement after treatment. This makes it difficult for existing studies to objectively define subgroups of treatment responses which reflect symptoms and biology rather than labelling a given exacerbation as showing improvement in individual variables according to a priori cut-offs. These points re-iterate the importance of Chapter 4 and Chapter 5 analyses.

Many studies e.g. the Cochrane assessment of the effectiveness of SCS and antibiotics for ‘exacerbations’ as well as ‘exacerbation’ treatment failure prediction studies such as Halner et al (2021) rely on medication usage and/or (re-)hospitalisation in order to define treatment outcome[15,167,168]. However, the advantages of my treatment response trajectotype approach in this Chapter of my thesis are three-fold:

1) it can be seen whether a patient in a particular crisis state experiences a relevant treatment response in terms of an improvement in the symptoms and inflammatory features which characterised the crisis in the first place;
2) I reveal a subgroup of crises which require symptom improvement but not necessarily inflammatory improvement and hence may be in receipt of unnecessary SCS and/or antibiotic therapy;
3) I unmask a subgroup of crises which do not in fact undergo symptom or inflammatory improvement.

The latter consideration is potentially important since crises followed by the 'no change' trajectotype after treatment might have been assumed to undergo improvement if: a) a univariable cut-off method had been used to assess post-treatment biomarkers; b) patient treatment response had solely been analysed according to treatment type; c) patient treatment response had been analysed stratified according to biomarkers at the time of exacerbation or stratified according to patient demographic characteristics. This again emphasises the importance of using objective methods to profile patients according to their multivariable symptomatological and inflammatory profile as well as taking into account the temporal aspect of this profile before and after crisis treatment.

Limitations and additional next steps

A notable limitation of my analysis is that since the majority of participants received antibiotics, it is not possible for my analytical methodology to determine the extent to which antibiotics were necessary versus superfluous for crisis outcomes. On the other hand, a positive relationship between SCS use at crisis and appropriate post-treatment trajectotype could be determined. This was especially possible because approximately half of the crises analysed were part of a clinical randomisation framework in which SCS was directed based on blood eosinophil count[82]. This set-up ensured that there were sufficient participants not treated with SCS (if blood eosinophils not above 2%) for a fair comparison in eosinophilic crisis outcomes to be made. This is contrast to neutrophilic crises where the no-antibiotics sample size (n=1) was too small for predictive tools assessing utility of treatment to be developed.

Another limitation of my analysis is the small sample size for the comparison of characteristics of eosinophilic crises and neutrophilic crises followed by an appropriate

resolution trajectotype compared to those not followed by an appropriate resolution trajectotype. In contrast to symptomatic crises, prediction of which eosinophilic crises and neutrophilic crises would be followed by an appropriate resolution trajectotype was unsuccessful as revealed by the poor test set performance. This can perhaps be understood in the context of the test set size. Out of a test set of 22 eosinophilic events, all events were treated with SCS and all participants taking SACH. Out of a test set of 11 neutrophilic events, a single participant was taking SACH and one other participant was on maintenance prednisolone. This would undermine the usefulness of the exacerbation treatment and baseline medication data for a classifier to make an accurate prediction of resolution for test set crises, even if a relation between these variables and crisis post-treatment trajectotype does in fact exist. It would not have been appropriate to select crises for the test set to intentionally include more variation in exacerbation treatment and baseline medication data since random selection of events for a test set is a key component of unbiased performance assessment of prediction algorithms. Determination of whether the occurrence of an appropriate resolution trajectotype following an eosinophilic crisis and following a neutrophilic crisis can in fact be predicted using the classifiers developed in this chapter will require a more substantial sample size in a future study setting.

An important next step is to validate the trajectotypes which I have identified in this chapter. Unfortunately, the number of participants with post-treatment longitudinal symptom and inflammatory follow up data in the emergency exacerbations asthma and COPD cohort analysed in Chapter 3 is too small for performing cluster analysis and assessing which combination of exacerbation treatment (SCS and antibiotics, one of these or neither) and crisis type yield optimal outcomes. As previously discussed,

biomarkers associated with cardiovascular comorbidity and depression, as well as patient lifestyle activities and environmental exposures should also be measured at crisis and during the two week post-crisis follow-up period as discussed. This may be especially insightful given the association between depression and incomplete eosinophilic crisis resolution found by my analysis. Hopefully this will assist in enabling a high-performing supervised learning algorithm to predict a crisis event's post-treatment trajectotypes. Future studies could also include inflammatory index measurements daily or, if more feasible, weekly at low overall symptom burden state and daily during crisis follow-up to help identify trajectories from the low overall symptom burden state to crisis in the days leading up to a crisis based on inflammation in addition to symptoms. This could be achieved via VAS symptom diaries to be completed daily for a suitable time period such as year during which patients are predicted to experience at least one crisis. Finger-prick point of care tests for CRP, neutrophils and eosinophils would enable inflammatory measurements in the weeks and days leading up to a crisis and during the two-week post-crisis follow-up period to be obtained.

5.5 Conclusion

For the first time in the field, I have defined personalised post-treatment patterns based on symptom and inflammatory change. The variation in magnitude of inflammatory change between crisis events demonstrate the importance of assessing patients undergoing a crisis in the context of their multivariable inflammatory profile. My analysis is the first to identify objective cluster-derived multivariable treatment response patterns based on differences in patient symptoms and biological inflammatory indices between the time of crisis and a follow-up treatment response time point at two weeks after crisis. In Chapter 6, I will summarise the findings of my thesis and potential next steps to translate my research to clinical practice.

Chapter 6 – Thesis summary

Overall findings

In my thesis, I have:

1) developed tools to predict healthcare utilisation treatment outcome for patients hospitalised with exacerbations of airways disease;
2) mathematically developed the concept of and defined COPD equilibrium/low overall symptom burden state, symptomatic crisis, eosinophilic crisis and neutrophilic crisis states based on symptoms and biology to replace the conventional exacerbation and stable state definitions;
3) characterised symptom- and biology-based treatment response at two weeks post-crisis compared to the time of crisis, referred to as: no change trajectotype, symptomatic resolution trajectotype, symptomatic and eosinophilic resolution trajectotype and symptomatic and neutrophilic resolution trajectotype;
4) characterised different symptom time series shape profile trajectotypes (based on post-crisis daily VAS symptom score diaries completed from crisis until day 14 post-crisis), revealing different phases of symptom improvement within the course of a two-week period post-crisis.

Predicting which patients admitted to ED with exacerbations of airways disease will require additional medication and/or re-hospitalisation within a one month window from the initial exacerbation may enable better personalised treatment and patient monitoring. This could translate into better health outcomes and also reduce economic costs for the healthcare system. The use of a robust multivariable feature selection approach and machine learning classification enabled performance of the treatment failure prediction tools to be maximised. My finding that variables including VAS

dyspnoea, patient exacerbation history and medication usage can predict treatment failure is encouraging since these data are readily available or easily measured in the clinical context.

One of the most important results of my thesis is the new definition of COPD states. Since ‘exacerbations’ of COPD are acknowledged to be heterogenous and the definition used forms the foundation of clinical trials assessing treatment outcomes, the subjectivity of existing definitions of ‘exacerbations’ and lack of biological components of these definitions have posed a serious concern in COPD research. My new definition of an equilibrium state reflecting an empirically derived, mathematically based subgroup of events with low symptoms and inflammation, compared to symptomatic, eosinophilic and neutrophilic crisis events constitutes a milestone in solving this concern. The multivariate cluster analysis approach used to define these new COPD states avoided reliance on existing definitions of ‘exacerbations’ and ‘stable state’. This is a much needed step beyond studies in the field to-date which either use patient and/or clinician consensus to define thresholds of variables to identify ‘exacerbations’, or attempt to train machine learning models to recognise clinically-defined exacerbations and are hence circular in nature. I showed that my COPD state definitions are mathematically superior in separating out events according to symptoms and inflammatory biology compared to conventional ‘exacerbations’ and ‘stable state’ definitions. I developed an algorithm to identify whether a given set of measurements from a patient corresponds to equilibrium, symptomatic crisis, eosinophilic crisis or neutrophilic crisis. The mathematical advantages of the machine learning approach have been previously discussed in my thesis. However, the choice of variables to inform the algorithm was equally as important from a clinical

perspective. The COPD state identification algorithm uses variables easily measurable in the clinic: VAS scores for cough, dyspnoea, sputum production and sputum purulence, blood CRP, blood neutrophil count, blood eosinophil count and blood eosinophil percentage. My multivariate COPD state identification algorithm may enable a greater sensitivity and specificity for detecting events which can optimally benefit from SCS therapy and antibiotics than currently used eosinophil- or CRP-based thresholds respectively used in the research community. My analysis enabled crisis treatment response to be assessed based on resolution of the type of symptoms and inflammation which characterised the crisis in the first place. By day 14 post-crisis, approximately one third of symptomatic crisis events do not undergo symptomatic resolution, while about one quarter of eosinophilic and neutrophilic crisis events undergo symptomatic but not inflammatory resolution. Approximately 16% of eosinophilic and 10% of neutrophilic crisis events show neither symptom nor inflammatory resolution. Over half of all crises move toward back toward the equilibrium/low overall symptom burden state, showing resolution in both symptoms and the corresponding type of inflammation in the case of eosinophilic and neutrophilic crises.

My comparison of two-week post-treatment single time point trajectotypes with shape profile trajectotypes showed that crisis symptomatological treatment response is not as simple as showing ‘improvement’ or ‘non-improvement’. The shape-based clustering approach applied to each of cough, dyspnoea, sputum production and sputum purulence revealed a pattern of improvement followed by worsening or vice versa as well as different rates of improvement and possible cycles of improvement and worsening within a two-week post-crisis follow-up period.

Limitations and future work

I have previously discussed the limitations of my analyses in the results chapters of my thesis. An overriding limitation is sample size e.g. for the ED exacerbation treatment failure prediction algorithm and the need to prospectively validate the performance of this algorithm, as well as the performance of my COPD state identification algorithm in addition to patient cohorts. This should be done using patient cohorts derived from multiple centres and at different time points to ensure generalisability of my algorithms across centres and across time. Nevertheless, the cross-validation and hold-out test set approaches used in my analysis for treatment outcome and COPD state prediction algorithms definitions indicate high performance for unseen patients. Similarly, the multivariate DO calculations performed to compare COPD states to conventional ‘exacerbation’ and ‘stable state’ definitions indicate that the states identified in my analysis are robust.

Only a small proportion of the crisis events in my analysis had sufficient data points in the daily VAS time series post-crisis follow available. Therefore, although the sample size for was adequate for defining COPD states and characterising post-treatment trajectotypes, the sample size was too small to demonstrate efficacy of algorithms to predict whether eosinophilic and neutrophilic crises will be followed by an appropriate resolution trajectotype. Future studies should aim to maximise the availability of post-crisis follow-up data for eosinophilic and neutrophilic crises, as well as daily time series VAS score data.

Although the sophistication of the mathematical approaches which I used to predict ED treatment outcome, define COPD states and identify treatment response trajectotypes

is an advantage from a performance perspective, these approaches are less interpretable for clinicians than single variable threshold-based approaches and more conventional algorithms such as logistic regression. However, a web application could be built to implement my emergency department exacerbation treatment outcome and COPD state identification algorithm as well as future treatment response trajectotype algorithms. This would enable feasibility of these approaches in the context of wide scale clinical use.

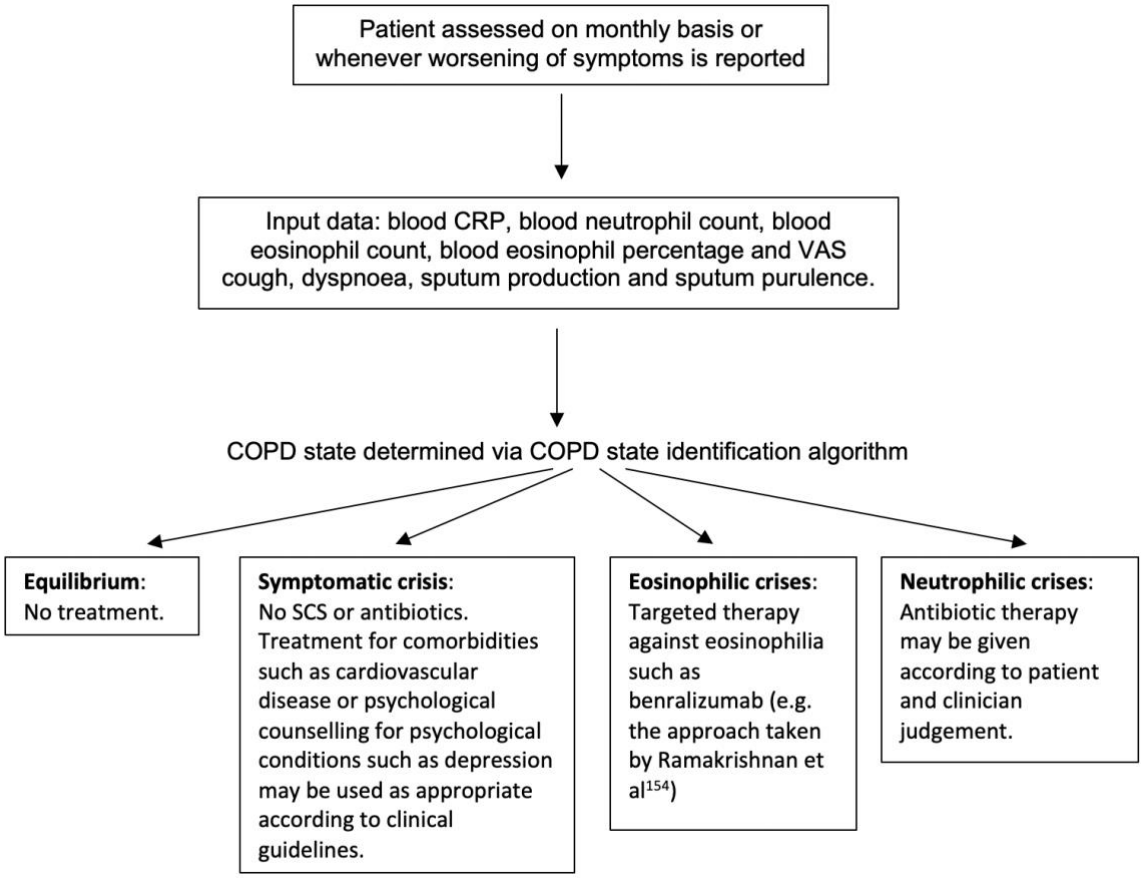
It should be noted that although my COPD state identification approach avoided the circularity of existing ‘exacerbation’ and ‘stable state’ definitions, the data used to develop the new definitions nevertheless did not include all severities of COPD ‘exacerbations’ to equal extent. Only a minority of the ‘exacerbations’ used to develop my definitions were severe. Nevertheless, validation in Chapter 4 showed that the COPD state identification also effectively separated out patients from the ED dataset from Chapter 3 based on symptoms and inflammation. A future study could include large patient cohorts with mild, moderate and severe COPD to be observed prospectively. Patient daily diary symptom cards to reveal increase in respiratory symptoms over consecutive days could increase reporting of exacerbations. My multivariate cluster COPD definition framework could then be applied to symptom and inflammation measurements made at one single time point i.e. the time of these ‘exacerbations’ as well as at ‘stable state’ without considering whether an event is labelled ‘exacerbation’ or ‘stable state’. Furthermore, the use of daily symptom VAS score diaries by patients even when at equilibrium and the use of daily finger-prick point of care blood tests for CRP and cell counts from the moment that the patient reports that his/her symptoms begin to worsen up until post-crisis follow-up could

enable greater temporal resolution as to the magnitude and time course of change of different symptoms and inflammatory indices between equilibrium and crisis and post-crisis treatment.

The crucial next step for translating my COPD state identification algorithm into potential improved clinical interventions for patients would be clinical trial in which patients are randomised to receive either:

1) biomarker-directed therapy according to currently used single variable-based thresholds e.g. blood eosinophil percentage >2% or blood eosinophil count $\geq 0.3 \times 10^9$ cells/L for therapy targeting eosinophilic inflammation such as benralizumab (versus placebo if biomarker negative) and ≥50mg/L blood CRP for antibiotics;
2) COPD state-directed therapy, for example, according to the schematic presented in **Figure 6.1**.

Figure 6.1 | Treatment allocation according to COPD state in future clinical trial representing next steps based on my findings.



If such a clinical trial shows superior performance for the COPD state-directed therapy group, as measured by treatment response trajectotypes for the symptoms and inflammatory indices which worsened from equilibrium to crisis, a COPD state-directed therapy approach for treating crises should be considered for widespread clinical use. The shape profile trajectotypes should be correlated with inflammatory biological profiles of recovery on a daily/weekly level and future crisis risk and treatment outcome in relation to shape profile patterns should be elucidated. This temporal shape of personalised patterns may represent a new biomarker itself for future crisis risk and treatment response. The mathematical frameworks used in this thesis taking into account symptom and biological inflammatory biomarkers and crisis to post-crisis

follow-up trajectories could enable superior management of COPD for patients, with improved health outcome and lower healthcare system costs.

A – Appendix

Appendix tables and figures

Appendix Table A1.1 | Detailed summary of COPD phenotyping and endotyping studies. Adapted from Halner et al (2019)[27].

Study	Burgel *et al* (2010)[58]	Garcia-Aymerich *et al* (2011)[59]
Study hypotheses/ research questions	COPD subjects can be grouped into clinical phenotypes.	COPD includes various clinically relevant subtypes.
Size and characteristics of patient cohort	n = 322 Multi-centre study. Recruited patients were in a stable condition (defined as having no history of exacerbation requiring medication during the previous 4 weeks) and had a COPD diagnosis based on post-bronchodilator FEV1/FVC < 0.7. Patients from all four GOLD severity categories were included. All patients in the cohort were positive for smoking history. Patients with a main diagnosis of asthma, bronchiectasis or other significant respiratory condition were excluded.	N = 342 Multi-centre study. Recruited patients had been hospitalised for a first time as a result of a COPD exacerbation. Measurements of variables were performed three months after discharge, at which point patients were clinically stable and COPD diagnosis was confirmed as post-bronchodilator FEV1/FVC < 0.7.
Variables selected for dimension reduction and/or cluster analysis	A total of 8 variables were chosen for their relevance to pulmonary and extrapulmonary features of COPD: patient age, tobacco-smoking, FEV1, number of exacerbations per patient per year, nutritional status, dyspnoea, health status and depressive symptoms.	A total of 536 variables, of which up to 150 variables were chosen for cluster analysis, were obtained relating to patient symptoms and quality of life, lung function tests, exercise capacity, patient nutritional status, biomarkers of systemic and bronchial inflammation, sputum microbiology, and imaging including CT of the thorax and echocardiography.

Method of analysis	PCA was performed to generate new independent variables – components – for cluster analysis. Ward's method of clustering was performed on 3 principal components since these components explained most of the variation in the data. Number of clusters was determined using the pseudo F and pseudo t^2 statistic.	K-means clustering was performed on the selected variables. Number of clusters was determined using the pseudo F statistic.
Cluster characteristics as described by author	Four clusters of patients were identified: 1: Younger patients with severe to very severe airflow limitation and respiratory symptoms but few comorbidities. 2: Older patients with mild respiratory disease and mild age-related comorbidities. 3: Younger patients with moderate to severe airflow limitation but milder symptoms and few comorbidities. 4: Older patients with moderate to severe airflow limitation, severe symptoms and significant cardiovascular comorbidity.	Three clusters of patients were identified: 1: Severe airflow limitation and very poor performance in other respiratory function tests. 2: Milder airflow limitation. 3: Milder airflow limitation with high proportion of comorbidity (obesity, cardiovascular disease, diabetes, systemic inflammation). Age of patients did not differ significantly between clusters.
Outcome measure for cluster validation and comparison	Longitudinal follow-up to determine all-cause mortality. Group 1 had the highest mortality rate.	Longitudinal follow-up to determine COPD-specific hospitalisation rate and all-cause mortality. Group 1 had the highest hospitalisation rate and all-cause mortality.

Study	Bafadhel *et al* (2011)[60]	Burgel *et al* (2012)[61]
Study hypotheses/ research questions	There are definable biological COPD clusters, which can be identified using biomarkers.	Subgroups of COPD patients differ in mortality.
Size and characteristics of patient cohort	n = 145 Single-centre study. Out of 145 patients entered into the study, 182 exacerbations were captured from 86 patients, and the desired variables were measured in 75 of these patients. All patients had post-bronchodilator FEV1/FVC < 0.7 and COPD diagnosis. Patients from all four GOLD severity categories were included. Patients were all aged 40 years or older. Asthma or other lung disease apart from COPD was an exclusion criterion.	N = 527 Multi-centre study. Recruited patients had a COPD diagnosis based on post-bronchodilator FEV1/FVC < 0.7. All patients were clinically stable. Patients from all four GOLD severity categories were included.
Variables selected for dimension reduction and/or cluster analysis	17 sputum biomarkers were considered for cluster analysis. Measurements were made for patients in the stable state and during exacerbations during the course of one year. Measurements of biomarkers at exacerbation were performed if the patients had not received prior oral corticosteroids or antibiotics.	7 continuous variables (patient age, BMI, FEV1, dyspnoea, quality of life scale, thoracic gas volume and diffusing capacity) and numerous categorical variables (relating to comorbidities and imaging data e.g. CT analysis for emphysema).

Method of analysis	Factor analysis (a technique similar to PCA) was performed on the biomarkers, and 3 factors were selected for subsequent analysis since they accounted for the majority of the variation in the data. For each factor, the biomarker with the highest loading was used for cluster analysis. Ward's method was used to generate a dendrogram for visual inspection to select an appropriate number of clusters (4 was the value chosen). K-means clustering, pre-specified to identify 4 biological clusters, was then applied to the highest loading biomarkers. ROC curves were used to determine suitable biomarkers for identification of clinical phenotypes.	PCA and MCA were separately performed on continuous variables and on categorical variables, respectively. 2 principal components and 14 MCA axes were retained for Ward's clustering. Visual assessment of dendrogram was used to decide upon a suitable number of clusters.

Cluster characteristics as described by author	4 biological exacerbation clusters were identified: Cluster 1: A proinflammatory profile. Cluster 2: High levels of type 2 mediators. Cluster 3: High levels of type 1 mediators. Cluster 4: Low sputum mediator concentrations cluster which exhibited few changes in inflammatory profile. Clinically, clusters 1, 2 and 3 were bacteria-predominant, eosinophil-predominant and virus-predominant, respectively. Bacterial and eosinophilic clinical exacerbation phenotypes could be predicted from stable state.	Three clusters of patients were identified: 1: Airflow limitation and other respiratory disease features were mild to moderate, with few comorbidities. 2: Airflow limitation and other respiratory disease features were severe. Variable comorbidities – osteoporosis and muscle weakness common but cardiovascular comorbidities rare. 3: Airflow limitation and other respiratory disease features moderate to severe. Older patients. Obesity, diabetes and cardiovascular comorbidities common.
Outcome measure for cluster validation and comparison	Sensitive and specific biomarkers for clusters 1-3 were identified. Sputum IL1-β, percentage peripheral eosinophils and serum CXCL10 best identified the bacteria-, eosinophil- and virus-predominant subgroups, respectively. Validation of these biomarkers was performed in an independent cohort of 89 patients.	Longitudinal follow-up to determine all-cause mortality. Group 2 and 3 patients were at a significantly higher risk of mortality than group 1 patients.

Study	Castaldi *et al* (2014)[62]	Rennard *et al* (2015)[63]
Study hypotheses/ research questions	Distinct subtypes of pulmonary damage occur in smokers and these subtypes are strongly associated with relevant clinical outcome measures and COPD-associated genetic variants.	Clinically relevant subgroups of COPD exist and exhibit differences in relevant clinical outcomes when evaluated longitudinally.
Size and characteristics of patient cohort	n = 10192 Multi-centre study. Patients were examined at stable state (at least one month after last exacerbation). Patients from all four GOLD categories were included. Patients with a respiratory disease diagnosis other than COPD, asthma or emphysema were excluded. All recruited patients were between the ages of 45 and 80 and have a smoking history.	N = 2164 Multi-centre study. All recruited patients had a COPD diagnosis. Patients from GOLD categories 2 to 4 were included. Patients were all exacerbation-free for at least 4 weeks before inclusion in the study. Patients with a respiratory disease diagnosis other COPD and severe alpha1-antitrypsin deficiency were excluded. All recruited patients were between the ages of 40 and 75 and have a smoking history.
Variables selected for dimension reduction and/or cluster analysis	Different variables were used for different cluster models. The variables chosen for the final model were FEV1, airway wall thickness and measures of emphysema from CT imaging.	41 variables relating to clinical, physiologic, imaging and biomarker parameters.

Method of analysis	Half of the patients constituted a training set to be included for cluster analysis, while the other half constituted a validation set to test the clusters. Various approaches were used to select variables for K-Means clustering. The resulting clusters from these different models, as well as the different options for a suitable number of clusters, were compared against each other using normalised mutual information (NMI), a cluster stability measure. The different models were also compared for discriminatory power between relevant clinical outcomes. The best model consisted of 4 clusters, using only those variables which were uncorrelated (Pearson's correlation < 0.7).	Factor analysis was performed on the 41 variables. 13 factors were selected since they accounted for most of the variability in the data; variables with the highest loading for the 13 factors were chosen for cluster analysis. A random forest-based clustering approach was used and number of clusters chosen based on silhouette width and clinical relevance of the clusters.
Cluster characteristics as described by author	4 subgroups of patients were identified, all of which were reproduced in the validation group of patients: 1: Smoking-resistant patients with few symptoms of airways disease. 2: Patients with mild upper zone -predominant emphysema and airflow obstruction. 3: Patients with airway-predominant disease. 4: Patients with severe airway obstruction and emphysema.	5 subgroups of patients were identified: A: Patients with mild airways disease. B: Patients with intermediate values for health status and emphysema but low levels of inflammatory markers. C: Systemic inflammation, multiple comorbidities. D: Low FEV1 and severe emphysema. E: Intermediate values for most variables.

Outcome measure for cluster validation and comparison	Clinical measures (including dyspnoea, exacerbation rates in previous year, hospitalisation rates in previous year) and COPD-associated gene variants. Subgroup 2, 3 and 4 had greater disease severity (subgroup 4 the worst) than subgroup 1 in terms of the validation clinical measures. After adjustment for the GOLD categories the patients in each subgroup were in, the association with clinical outcome measures remained significant. There was a strong association between subgroup 2 and the single nucleotide polymorphism (SNP) rs1980057 near the gene HHIP.	Longitudinal follow-up to determine all-cause mortality and time-to-first exacerbation. Subgroups C and D had the highest mortality and D had the shortest time-to-first exacerbation.

Study	Esteban *et al* (2016)[64]	Chang *et al* (2016)[65]
Study hypotheses/ research question	Assess the stability of cluster-based COPD subgroups in patients with stable disease over 1 year.	Systemic inflammatory signals in peripheral blood gene expression data can identify clinically important COPD-related disease subtypes.
Size and characteristics of patient cohort	n = 543 Multicentre study. All patients had a previous COPD diagnosis for at least six months and were free from an exacerbation for at least 6 weeks prior to enrolling. For all patients, FEV1/FVC < 0.7 and FEV1/FEV1 predicted < 0.8. Patients with an asthma diagnosis were excluded. One year after inclusion in the study, survivors were followed up.	N = 364 Multi-centre study. Data were captured at stable state. 229 former smokers (of whom 141 met the spirometric FEV1/FVC < 0.7 criterion for COPD diagnosis) from the ECLIPSE cohort constituted the training set. Clusters obtained based on the gene expression data of these patients were then validated in a separate cohort of 135 smokers (of whom 76 met the spirometric FEV1/FVC < 0.7 criterion for COPD diagnosis) from the COPDGene study.
Variables selected for dimension reduction and/or cluster analysis	A number of sociodemographic and clinical variables were included: age, BMI, number of previous hospitalisations, FEV1%, hand strength, walking test, physical activity, dyspnoea, Charlson comorbidity index scores and occurrence of various comorbidities.	Gene expression patterns as determined from 1812 probesets associated with FEV1 and FEV1/FVC in the ECLIPSE study were used as input for cluster analysis.

Method of analysis	MCA was performed on the variables. 4 MCA factors were selected for hierarchical clustering. An optimal number of clusters was decided upon based on minimum loss of inertia. The aforementioned analyses were performed for patients at first inclusion in the study and then again after a year for surviving patients.	Network-based stratification (NBS) was used to achieve clustering of gene expression data. This was compared against a non-network based method (non-negative matrix factorisation); the former displayed superior performance to the latter as determined by stability indices associated with different numbers of latent factors. After clustering, subtype-specific gene expression for each cluster was analysed for gene ontology enrichment of known biological processes. The cluster model developed from gene expression of ECLIPSE cohort patients was used to generate clusters from the COPDGene cohort.

Cluster characteristics as described by author	4 clusters were identified: Cluster A: Less dyspnoea, better health-related quality of life and lower Charlson comorbidity scores. Cluster B: Intermediate between A and C. Cluster C: Most severe dyspnoea, and poorer pulmonary function and quality of life. Cluster D: Higher rates of hospitalization during the previous year; higher comorbidity scores. Whereas clusters A, B, and C, had marked respiratory profiles with a continuum in severity of several variables, cluster D was associated with a more systemic profile with intermediate respiratory disease severity.	4 clusters were identified in the ECLIPSE cohort via NBS and could be reproduced in the COPDGene cohort: Cluster 1: Enrichment for wound healing and inflammatory processes. Cluster 2: Enrichment for cytoskeletal and actin filament organization. Cluster 3: Enrichment for protein catabolism and ubiquitination. Cluster 4: Lymphocyte activation and protein synthesis.

Outcome measure for cluster validation and comparison	Clusters at baseline were compared with clusters after 1 year to assess stability. Clusters remained stable after 1 year, with only 28% patients migrating between the clusters. In addition, clusters were compared in terms of clinical outcomes such as mortality. Cluster D had the highest mortality after 1 year of follow-up, followed by cluster C, B and A respectively.	The gene-expression-based clusters were compared in terms of clinical characteristics and peripheral blood cell counts: Cluster 1: Most severe lung function impairment, respiratory symptoms, and emphysema. Higher levels of neutrophils than the other clusters. Cluster 2: Intermediate levels of lung function impairment, emphysema, and respiratory symptoms. Higher levels of eosinophils than the other clusters. Cluster 3 and Cluster 4: Relatively preserved lung function. Cluster 3 had more emphysema, more respiratory symptoms, and a higher percentage of women than cluster 4. Cluster 3 and cluster 4 had higher levels of lymphocytes than the other clusters. Cell count differentials alone were not enough to predict COPD cluster membership. Smoking exposure did not differ significantly between the groups, suggesting that biological variability rather than cumulative smoke exposure, is more likely to explain the cluster patterns.

Study	Burgel *et al* (2017)[66]	Zarei *et al* (2017)[67]
Study hypotheses/ research questions	An algorithm can be developed for allocating COPD patients into cluster analysis-derived clinical phenotypes.	Peripheral blood proteomic data can be used to find subtypes of COPD within clinically similar individuals at stable state.
Size and characteristics of patient cohort	n = 6060 Multi-centre study. All recruited patients had a COPD diagnosis. Patients from all four GOLD categories were included. Patients were recruited in a stable state or at the time of a hospitalisation due to an exacerbation.	N = 396 Multi-centre study. All recruited patients were former smokers with at least 10 pack years but abstinence from smoking for at least 12 months before study. All patients had moderate to severe COPD (post-bronchodilator FEV1/FVC < 0.7; FEV1/FEV1 predicted < 70% predicted; Diffusing capacity of the Lung for Carbon Dioxide (DL_{CO}) < 0.7) and were at stable state. All patients had emphysema based on visual examination of CT scans.
Variables selected for dimension reduction and/or cluster analysis	Clinical variables including age, BMI, FEV1, dyspnoea, number of exacerbations in previous 12 months, presence of comorbidities (cardiovascular and diabetes).	87 protein biomarkers measured from peripheral blood were used as input for cluster analysis.

Method of analysis	FAMD was performed on the variables of 2409 patients and the highest loading variables for the selected factors were used as input for Ward's method of clustering. Classification and regression trees (CARTs) were then trained to develop an algorithm for allocating patients into the clusters obtained from Ward's method in the group of 2409 patients. The algorithm was then tested on a separate cohort of 3651 patients.	Agglomerative McQuitty hierarchical clustering was performed on the biomarker dataset. The optimal number of clusters was determined through the R package NbClust which takes into account a variety of indices for this purpose. After clustering, enrichment analysis was performed for the biomarkers which had different mean values among the clusters in order to identify the molecular pathways associated with these biomarkers.
Cluster characteristics as described by author	Five subgroups were identified: I: Severe respiratory disease, older age patients, high prevalence of comorbidities. II: Moderate to severe respiratory disease, younger patients, few comorbidities. III: Older patients, high prevalence of comorbidities. IV: Very severe respiratory disease, younger patients, few comorbidities. V: Mild respiratory disease, younger patients, few comorbidities.	4 stable state biological clusters were identified. The biomarkers which distinguished cluster 3 from cluster 1 and cluster 2 mapped to platelet alpha granule and cell chemotaxis pathways.

Outcome measure for cluster validation and comparison	Longitudinal follow-up to determine all-cause mortality. Subgroups I and IV have highest mortality rate. The CART-developed algorithm successfully assigned patients in the validation cohort to five classes which corresponded to the clusters obtained in the training cohort in terms of the relative clinical characteristics and mortality rates between the classes.	The clusters were compared in terms of clinical and physiological characteristics. Compared to cluster 1 and cluster 2, cluster 3 had less emphysema on quantitative analysis of chest CT scans and worse disease-related quality of life based on the St. George's Respiratory Questionnaire.

Appendix Table A1.2 | Detailed examples of cluster analysis studies including both asthma and COPD. Adapted from Halner et al (2019)[27].

Study	Ghebre *et al* (2015)[68]	Ghebre *et al* (2018)[69]
Study hypotheses/ research questions	Determine the extent to which COPD and asthma at stable state represent distinct or overlapping conditions in terms of sputum cellular and mediator profiles.	Investigate the sputum cellular, mediator, and microbiome profiles of asthma and COPD exacerbations via a cluster analysis-based approach.
Size and characteristics of patient cohort	n = 385 Single centre study. 161 patients (86 with severe asthma and 75 with moderate to severe COPD) were recruited to constitute the training set for cluster analysis. Patients were examined at stable state (defined here as at least 8 weeks free from an exacerbation). 224 patients (166 with severe asthma and 58 with COPD) were used for validation.	N = 105 Single-centre study. 32 asthmatic patients and 73 patients with COPD were recruited. Patients were assessed at exacerbation.
Variables selected for dimension reduction and/or cluster analysis	21 sputum mediators were analysed in the training study and 14 in the validation study.	19 sputum mediators were analysed.

Method of analysis	Factor analysis was performed and 4 factors selected based on scree plots and having an eigenvalue > 1. K-Means analysis was then performed on the factor scores. The optimal number of clusters was selected based on a scree plot and the biological/clinical interpretability of the resulting clusters.	Factor analysis was performed and 3 factors selected based on scree plots and having an eigenvalue > 1. K-means analysis was then performed on the factor scores. The optimal number of clusters was selected based on a scree plot.
Cluster characteristics as described by author	5 stable state biological clusters were identified and reproduced in the validation group of patients: Cluster 1:Asthma-predominant, eosinophilic, high TH2 cytokines. Cluster 2: Asthma and COPD overlap, neutrophilic; this cluster was associated with the highest overall inflammatory cell count. Cluster 3: COPD-predominant, mixed eosinophilic and neutrophilic.	6 exacerbation state biological clusters were identified: Cluster 1: COPD predominant. 27 patients with COPD and 7 asthmatic patients exhibited increased blood and sputum neutrophil counts, proinflammatory, and proportions of the bacterial phylum Proteobacteria. Cluster 2: 10 asthmatic patients and 17 patients with COPD with increased blood and sputum eosinophil counts, type 2 mediators, and proportions of the bacterial phylum Bacteroidetes. Cluster 3: 15 asthmatic patients and 29 patients with COPD with increased type 1 mediators and proportions of the phyla Actinobacteria and Firmicutes.

Outcome measure for cluster validation and comparison	The clusters were compared in terms of clinical, physiological and biological characteristics. Linear discriminant analysis was performed on the inflammatory mediators and a model developed to group patients into the training set clusters. This model was then used to categorise patients of the validation set into inflammatory mediator-based clusters. Patients in the asthma-COPD overlap cluster had higher symptoms of cough than patients of the other two clusters.	The clusters were compared in terms of clinical, physiological and biological characteristics. Linear discriminant analysis was performed on the inflammatory mediators to develop a model to separate patients out by cluster. However, this model was not validated in a separate patient test set.

Appendix Table A3.1 | Normality tests for Chapter 3 patient cohort variables and where relevant log-transformed data variables. For variables with zero/negative values, then normality assumption for the log-transformed version of the variable was not performed.

Variable	Shapiro-Wilk test p-value	Shapiro-Wilk test p-value for log-transformed variable
Age (years)	<0.01	<0.01
Age at diagnosis (years)	<0.01	<0.01
Age at onset of symptoms (years)	<0.01	<0.01
Height (cm)*	0.02	0.06
Weight (kg)*	<0.01	0.23
Pack Year History	<0.01	N/A
Number of days for which patient reported symptoms have been worse for in lead up to exacerbation	<0.01	N/A
VAS cough (mm)	0.01	N/A
VAS dyspnoea (mm)	<0.01	N/A
VAS sputum production (mm)	<0.01	N/A
VAS sputum purulence (mm)	<0.01	N/A
EuroQol VAS Score	<0.01	N/A
HAD anxiety	<0.01	N/A
HAD depression	<0.01	N/A
CAT Score	<0.01	N/A
Charlson Index Score	<0.01	N/A
Number of unscheduled primary care and emergency department visits in previous 12 months	<0.01	N/A
Number of hospital admissions in previous 12 months	<0.01	N/A
Number of intensive therapy unit admissions in previous 12 months	<0.01	N/A
Oxygen saturation (%)	0.02	0.01
Blood pH	<0.01	<0.01
Blood pCO2 (kPa)*	<0.01	0.18
Heart rate (beats/minute)~	0.54	N/A
Temperature (°C)	<0.01	0.02
Systolic (mmHg)*	0.01	0.44
Diastolic (mmHg)~	0.23	N/A
Respiratory rate (breaths per minute)*	<0.01	0.16
Peak flow (litres/minute)*	0.01	0.06
Total leucocyte count (x10^9cells/L)*	<0.01	0.58
Blood neutrophil count (x10^9cells/L)*	<0.01	0.92

Blood eosinophil count (x10⁹cells/L)	<0.01	<0.01
Blood basophil count (x10⁹cells/L)	<0.01	0.03
Blood monocyte count (x10⁹cells/L)	<0.01	<0.01
Blood lymphocyte count (x10⁹cells/L)*	<0.01	0.31
Blood neutrophil percentage~	0.06	N/A
Blood eosinophil percentage	<0.01	<0.01
Blood basophil percentage*	<0.01	0.56
Blood monocyte percentage	<0.01	<0.01
Blood lymphocyte percentage	<0.01	<0.01
Haemoglobin (g/L)~	0.18	N/A
Platelets (x10⁹cells/L)	<0.01	0.02
Sodium (mmol/L)	<0.01	<0.01
Potassium (mmol/L)	<0.01	0.01
Urea (mmol/L)	<0.01	<0.01
Creatinine (µmol/L)	0.01	<0.01
eGFR (mL/min)	<0.01	<0.01
Blood glucose (mmol/L)	<0.01	<0.01
Serum albumin (g/L)	<0.01	<0.01
CRP (mg/L)	<0.01	<0.01

Definition of abbreviations: CAT score = COPD Assessment Tool score; CRP = C-reactive protein; eGFR = estimated glomerular filtration rate; HAD = Hospital Anxiety and Depression scale; VAS = Visual Analogue Scale.
Symbols: ~indicated that a variable is normally distributed. * Indicates that the log-transformed version of the variable is normally distributed.

Appendix Table A3.2 | P-values showing D0 characteristics of 81 patients with known D30 healthcare utilisation treatment outcome compared to 20 patients without available D30 treatment outcome data.

Variable	Benjamini-Hochberg P-Value
Age (years)	0.64
Height (cm)*	0.93
Weight (kg)*	0.58
Pack Year History	0.79
Number of days for which patient reported symptoms have been worse for in lead up to exacerbation	0.78
Oxygen saturation (%)	0.36
Blood pH	0.58
Blood pCO2 (kPa)*	0.57
Heart rate (beats/minute)~	0.64
Temperature (°C)	0.58
Systolic (mmHg)*	0.92
Diastolic (mmHg)*	0.87
Respiratory Rate (breaths per minute)*	0.71
Peak flow (litres/minute)*	0.33
Number of unscheduled primary care and emergency department visits in previous 12 months	0.78
Number of hospital admissions in previous 12 months	0.83
Number of Intensive Therapy Unit admissions in previous 12 months	0.95
EuroQol VAS Score	0.58
VAS cough (mm)	0.72
VAS dyspnoea (mm)	0.72
VAS sputum production (mm)	0.79
VAS sputum purulence (mm)	0.93
HAD anxiety	0.58
HAD depression	0.58
Charlson Index Score	0.83
Total leucocyte count (x10^{9}cells/L)*	0.58
Blood neutrophil count (x10^{9}cells/L)*	0.59
Blood eosinophil count (x10^{9}cells/L)	0.90
Blood basophil count (x10^{9}cells/L)	0.93
Blood monocyte count (x10^{9}cells/L)	0.85
Blood lymphocyte count (x10^{9}cells/L)*	0.95
Blood neutrophil percentage~	0.93
Blood eosinophil percentage	0.82
Blood basophil percentage*	0.79
Blood monocyte percentage	0.87
Blood lymphocyte percentage	0.93

Haemoglobin (g/L)~	0.36
Platelets (x10⁹cells/L)	0.59
Sodium (mmol/L)	0.92
Potassium (mmol/L)	0.83
Urea (mmol/L)	0.87
Creatinine (µmol/L)	0.58
eGFR (mL/min)	0.59
Blood glucose (mmol/L)	0.58
Serum albumin (g/L)	0.24
CRP (mg/L)	0.58
Age at diagnosis (years)	0.58
Age at onset of symptoms (years)	0.58
CAT Score	0.58
EuroQol mobility	0.58
EuroQol self care	0.58
EuroQol usual activity	0.72
EuroQol pain discomfort	0.79
EuroQol anxiety depression	0.63
Sex	0.58
Smoking status	0.58
Increased dyspnoea	0.84
Increased sputum production	0.96
Increased sputum purulence	0.78
Increased cough	0.87
Increased wheeze	0.93
Fever	0.93
Oral steroids in the last week	0.93
Antibiotics in the last week	0.58
Intravenous hydrocortisone	0.64
Previous respiratory diagnosis	0.24

Nasal polyps	0.87
Comorbidity	0.24
SABA	0.92
LABA	0.58
LAMA	0.58
SAMA	0.79
ICS	0.24
Carbocystiene	0.58
Maintenance prednisolone	0.58
Monteleukast	0.58
Antihistamine	0.58
PPI	0.93
Antihypertensive	0.79
Ethnicity	0.87
Comorbidity	0.24
Previous nasal polyp surgery	0.79
Theophylline	0.72
Maintenance azithromycin	0.84
Aspirin	0.70
Nasal steroids use	0.72
Nebuliser use	0.72
Long Term Oxygen Therapy	0.84
Anti-diabetic medication	0.72
Beta-blockers	0.82
Oxygen administered	0.79
Exacerbation treatment (SCS and antibiotics, SCS only, antibiotics only, neither)	0.96

Definition of abbreviations: CAT score = COPD Assessment Tool; CRP = C-reactive protein; eGFR = estimated glomerular filtration rate; HAD = Hospital Anxiety and Depression scale; ICS = inhaled corticosteroids; LABA = Long-Acting Beta Agonist; LAMA = Long-Acting Muscarinic Antagonist; PPI = Proton Pump Inhibitor; SABA = Short-Acting Beta Agonist; SAMA = Short-Acting Muscarinic Antagonist; SCS = systemic corticosteroids; VAS = Visual Analogue Scale.
Hypothesis tests: For numerical variables following a normal distribution, unpaired two-sided t-tests are used. These variables are marked ~. For log-transformed variables following a normal unpaired two-sided t-tests on log-transformed data are used and marked *. For other numerical variables, the Mann-Whitney U test is used. For binary and categorical variables, the chi-squared test is used.
The Benjamini-Hochberg correction for multiple comparisons was used.

Appendix Table A3.3 | Paired comparison of characteristics of participants between D0 and D1.

Variable		**D0 (n= 28 paired participants)**	**D1 (n=28 paired participants)**	**P-Value**
Rating of how patient felt on the day		N/A	5 (4 – 7)	N/A
Whether patient felt better compared to D0	Better, n (%)	N/A	22 (81)	N/A
	No improvement, n (%)	N/A	5 (19)	
VAS cough (mm)		**63 (35 – 77)**	**49 (17 – 65)**	**<0.01**
VAS dyspnoea (mm)		**77 (55 – 89)**	**49 (33 – 70)**	**<0.01**
VAS sputum production (mm)		41 (14 – 57)	20 (1 – 54)	0.16
VAS sputum purulence (mm)		45 (8 – 65)	19 (0 – 55)	0.48
CAT score		15 (0 – 31)	0 (0 – 24)	0.09
Peak flow (litres/minute)*		198 (171 – 230)	223 (176 – 281)	0.21
Oxygen saturation (%)		95 (92 – 97)	95 (93 – 97)	0.32
Total leucocyte count (x10⁹cells/L)*		11.6 (10.3 – 13.1)	11.3 (9.3 – 13.8)	0.39
Blood neutrophil count (x10⁹cells/L)*		8.6 (7.1 – 10.3)	9.7 (7.8 – 11.9)	0.33
Blood eosinophil count (x10⁹cells/L)		0.12 (0.02 – 0.29)	0.01 (0.01 – 0.13)	0.14
Blood basophil count (x10⁹cells/L)		**0.50 (0.03 – 0.06)**	**0.03 (0.01 – 0.05)**	**0.01**
Blood monocyte count (x10⁹cells/L)		0.84 (0.60 – 1.06)	0.41 (0.27 – 0.86)	0.26
Blood lymphocyte count (x10⁹cells/L)*		1.27 (0.97 – 1.67)	0.72 (0.42 – 1.22)	0.13
Blood neutrophil Percentage~		75.4 (2.7)	85.9 (2.3)	0.11

Blood eosinophil percentage		1.2 (0.2 – 2.8)	0.1 (0.1 – 0.1)	0.17
Blood basophil percentage*		**0.4 (0.3 – 0.5)**	**0.2 (0.2 – 0.3)**	**<0.01**
Blood monocyte percentage		7.1 (5.3 – 9.4)	4.2 (2.5 – 6.7)	0.19
Blood lymphocyte percentage		12.0 (7.4 – 16.6)	9.5 (3.6 – 11.7)	0.09
Haemoglobin (g/L)~		**143 (3)**	**137 (4)**	**<0.01**
Platelets (x10^{9}cells/L)		254 (216 – 289)	257 (219 – 305)	0.31
Sodium (mmol/L)		140 (138 – 141)	139 (139 – 142)	0.57
Potassium (mmol/L)		3.7 (3.2 – 4.1)	3.7 (3.3 – 4.0)	0.23
Urea (mmol/L)		4.9 (3.6 – 6.8)	4.5 (3.6 – 6.2)	1.00
Creatinine (µmol/L)		65 (58 – 89)	61 (50 – 71)	0.59
eGFR (mL/min)		90 (73 – 90)	90 (84 – 90)	0.42
Blood glucose (mmol/L)		6.6 (6.1 – 8.4)	7.1 (6.1 – 11.2)	0.53
Serum albumin (g/L)		36 (33 – 39)	33 (29 – 39)	0.18
CRP (mg/L)		10.4 (2.7 – 24.8)	0.0 (0.0 – 5.9)	0.08
Exacerbation treatment	SCS and antibiotics, n (%)	16 (57)		N/A
	SCS alone, n (%)	8 (29)		
	Antibiotics alone, n (%)	3 (11)		
	Neither treatment, n (%)	1 (4)		

Definition of abbreviations: CAT score = COPD Assessment Tool; CRP = C-reactive protein; eGFR = estimated glomerular filtration rate; SCS = systemic corticosteroids; VAS = Visual Analogue Scale.
Variable presentation and hypothesis tests: For numerical variables following a normal distribution, the paired t-test is used. These variables are indicated by ~. For variables whose log-transformed version follow a normal distribution the paired t-test is performed on the log-transformed data and indicated by *. For other numerical variables, the paired wilcox test is used. Variables following a normal distribution are presented as mean (standard error of the mean); log-transformed variables following a normal distribution are presented as geometric mean (95% confidence interval); other variables are presented as median (lower quartile to upper quartile). Binary and categorical variables are presented as number (%) of instances.
Rows in bold font correspond to variables which differed significantly between D0 and D1.

Appendix Table A3.4 | Paired comparison of characteristics of participants between D0 and D5.

Variable		**D0 (n= 11 paired participants)**	**D5 (n=11 paired participants)**	**P-Value**
Rating of how patient felt on the day		N/A	7 (4 – 7)	N/A
Whether patient felt better compared to D0	Better, n (%)	N/A	9 (90)	N/A
	No improvement, n (%)	N/A	1 (10)	
VAS cough (mm)		67 (35 – 78)	19 (11 – 68)	0.08
VAS dyspnoea (mm)		**89 (73 – 89)**	**24 (8 – 55)**	**<0.01**
VAS sputum production (mm)		41 (24 – 76)	20 (10 – 53)	0.31
VAS sputum purulence (mm)		53 (5 – 77)	19 (4 – 51)	0.22
CAT score		18 (0 – 31)	4 (0 – 27)	0.24
Peak flow (litres/minute)*		**151 (128 - 179)**	**182 (135 - 245)**	**<0.01**
Oxygen saturation (%)		94 (91 – 97)	96 (95 – 97)	0.35
Total leucocyte count (x10⁹cells/L)*		11.2 (8.8 – 14.1)	10.0 (8.4 – 11.0)	0.09
Blood neutrophil count (x10⁹cells/L)*		7.8 (5.4 – 11.2)	6.2 (4.6 – 8.5)	0.15
Blood eosinophil count (x10⁹cells/L)		0.12 (0.03 – 0.30)	0.09 (0.03 – 0.18)	0.84
Blood basophil count (x10⁹cells/L)		0.05 (0.03 – 0.06)	0.05 (0.03 – 0.07)	0.50
Blood monocyte count (x10⁹cells/L)		0.90 (0.50 – 1.70)	0.83 (0.54 – 1.22)	0.11
Blood lymphocyte count (x10⁹cells/L)*		1.28 (0.77 – 2.10)	2.09 (1.20 – 3.62)	0.06
Blood neutrophil Percentage~		72.2 (5.3)	64.6 (5.7)	0.27

Blood eosinophil percentage		1.2 (0.2 – 3.0)	0.9 (0.3 – 2.0)	0.74
Blood basophil percentage*		0.4 (0.3 – 0.5)	0.5 (0.3 – 0.7)	0.18
Blood monocyte percentage		7.7 (5.8 – 14.7)	7.8 (6.8 – 9.7)	0.64
Blood lymphocyte percentage		12.0 (7.6 – 19.4)	20.1 (14.7 – 44.1)	0.11
Haemoglobin (g/L)~		147 (4)	140 (5)	0.13
Platelets (x10⁹cells/L)		244 (214 - 341)	276 (229 - 354)	0.38
Sodium (mmol/L)		140 (137 - 141)	140 (138 - 142)	0.52
Potassium (mmol/L)		3.7 (3.2 – 3.7)	3.7 (3.5 – 3.9)	0.60
Urea (mmol/L)		5.7 (2.6 – 7.4)	4.8 (3.8 – 8.5)	0.94
Creatinine (µmol/L)		64 (52 - 93)	56 (54 - 81)	0.40
eGFR (mL/min)		83 (63 – 90)	68 (56 - 90)	0.36
Blood glucose (mmol/L)		6.5 (5.6 – 8.4)	9.7 (7.5 – 10.0)	0.88
Serum albumin (g/L)		36 (34 – 38)	34 (33 - 34)	1.00
CRP (mg/L)		16.0 (1.7 – 25.4)	2.9 (0.0 – 9.7)	0.08
Exacerbation treatment	SCS and antibiotics, n (%)	5 (45)		N/A
	SCS alone, n (%)	2 (18)		
	Antibiotics alone, n (%)	3 (27)		
	Neither treatment, n (%)	1 (9)		

Definition of abbreviations: CAT score = COPD Assessment Tool score; CRP = C-reactive protein; eGFR = estimated glomerular filtration rate; SCS = systemic corticosteroids; VAS = Visual Analogue Scale.
Variable presentation and hypothesis tests: For numerical variables following a normal distribution, the paired t-test is used. These variables are indicated by ~. For variables whose log-transformed version follow a normal distribution the paired t-test is performed on the log-transformed data and indicated by *. For other variables, the paired wilcox test is used. Numerical variables following a normal distribution are presented as mean (standard error of the mean); log-transformed variables following a normal distribution are presented as geometric mean (95% confidence interval); other variables are presented as median (lower quartile to upper quartile). Binary and categorical variables are presented as number (%) of instances.
Rows in bold font correspond to variables which differed significantly between D0 and D5.

Appendix Table A3.5 | D0 baseline characteristics of 90 patients for whom follow-up data was available within 5 days of initial exacerbation. Characteristics of patients re-hospitalised versus not re-hospitalised within 5 days of exacerbation (hospital visit on D1 or D5) are compared.

Variable		**D1/5 Hospitalisation (n=29)**	**No D1/5 Hospitalisation (n=61)**	**P-Value**
Age (years)		51 (30 - 61)	53 (29 – 67)	0.84
Weight (kg)*		81 (71 - 93)	82 (76 - 88)	0.99
Pack Year History		4 (0 – 17)	1 (0 – 28)	0.95
Number of unscheduled primary care and emergency department visits in previous 12 months		**3 (1 – 7)**	**1 (0 – 2)**	**<0.01**
Number of hospital admissions in previous 12 months		**1 (0 – 4)**	**0 (0 – 1)**	**0.04**
Number of Intensive Therapy Unit admissions in previous 12 months		0 (0 - 0)	0 (0 - 0)	0.16
Age at diagnosis (years)		32 (10 – 52)	20 (8 – 50)	0.48
Age at onset of symptoms (years)		24 (9 – 48)	20 (9 – 50)	0.98
Male, n (%)		25 (41)	11 (38)	0.78
Smoking Status	Current Smoker, n (%)	8 (28)	15 (25)	0.95
	Ex-Smoker, n (%)	12 (41)	26 (43)	
	Non-Smoker, n (%)	9 (31)	20 (33)	

Oral steroids in the last week, n (%)		20 (69)	29 (48)	0.06
Antibiotics in the last week, n (%)		15 (52)	19 (31)	0.06
Previous respiratory diagnosis	COPD, n (%)	7 (24)	15 (25)	0.96
	Asthma, n (%)	22 (76)	46 (75)	
Nasal polyps, n (%)		1 (3)	6 (10)	0.29
Comorbidity	Cardiovascular, n (%)	5 (17)	10 (18)	1.00
	Other, n (%)	13 (0.45)	25 (45)	
	None, n (%)	11 (38)	21 (38)	
SABA, n (%)		21 (72)	54 (89)	0.06
LABA, n (%)		3 (10)	3 (5)	0.33
LAMA, n (%)		7 (24)	18 (30)	0.60
SAMA, n (%)		1 (3)	1 (2)	0.59
ICS, n (%)		**9 (31)**	**35 (63)**	**0.01**
Carbocystiene, n (%)		**5 (17)**	**1 (2)**	**0.01**
Maintenance prednisolone, n (%)		4 (14)	3 (5)	0.14

Monteleukast, n (%)		6 (21)	11 (18)	0.76
Antihistamine, n (%)		2 (7)	6 (10)	0.65
PPI, n (%)		2 (7)	8 (13)	0.38
Antihypertensive, n (%)		4 (14)	5 (8)	0.41
Caucasian, n (%)		27 (93)	53 (91)	0.78
Working Status	Full-Time, n (%)	13 (48)	28 (49)	0.93
	Part-Time, n (%)	3 (11)	9 (0.16)	
	Retired, n (%)	8 (30)	14 (25)	
	Unemployed, n (%)	3 (11)	6 (11)	
Previous nasal polyp surgery, n (%)		0 (0)	2 (3)	0.32
Theophylline, n (%)		2 (7)	1 (2)	0.19
Maintenance azithromycin, n (%)		1 (3)	0 (0)	0.14
Aspirin, n (%)		0 (0)	4 (7)	0.16
Nasal steroids use, n (%)		2 (7)	1 (2)	0.19
Nebuliser use, n (%)		2 (7)	1 (2)	0.19
Long Term Oxygen Therapy, n (%)		0 (0)	1 (2)	0.49
Anti-diabetic medication, n (%)		0 (0)	3 (5)	0.22
Beta-blockers, n (%)		1 (3)	2 (3)	0.97

Definition of abbreviations: ICS = inhaled corticosteroids; LABA = Long-Acting Beta Agonist; LAMA = Long-Acting Muscarinic Antagonist; PPI = Proton Pump Inhibitor; SABA = Short-Acting Beta Agonist; SAMA = Short-Acting Muscarinic Antagonist.
Variable presentation and hypothesis tests: For log-transformed numerical variables following a normal unpaired two-sided t-tests on log-transformed data are used and marked *. For other variables, the Mann-Whitney U test is used. Log-transformed variables following a

normal distribution are presented as geometric mean (95% confidence interval); other numerical variables are presented as median (lower quartile to upper quartile).
Binary and categorical variables are presented as number (%) of instances (chi-squared test used for treatment outcome comparison).
Rows in bold font correspond to variables which differed significantly between D1/5 healthcare utilisation treatment outcomes.

Appendix Table A3.6 | Paired comparison of characteristics of participants between D0 and D30.

Variable		**D0 (n= 53 paired participants)**	**D30 (n=53 paired participants)**	**P-Value**
Rating of how patient felt on the day		N/A	7 (5 – 8)	N/A
Whether patient felt better compared to D0	Better, n (%)	N/A	40 (75)	N/A
	No improvement, n (%)	N/A	13 (25)	
VAS cough (mm)		**48 (27 – 70)**	**15 (3 – 40)**	**<0.01**
VAS dyspnoea (mm)		**70 (27 – 88)**	**35 (8 – 62)**	**<0.01**
VAS sputum production (mm)		19 (1 – 33)	5 (0 – 20)	0.06
VAS sputum purulence (mm)		**3 (0 – 51)**	**0 (0 – 18)**	**<0.05**
CAT score		**21 (0 – 29)**	**8 (0 – 20)**	**<0.01**
Exacerbation treatment	SCS and antibiotics, n (%)	22 (42)		N/A
	SCS, n (%)	20 (38)		
	Antibiotics alone, n (%)	3 (6)		
	Neither treatment, n (%)	8 (15)		

Definition of abbreviations: CAT score = COPD Assessment Tool score; SCS = systemic corticosteroids; VAS = Visual Analogue Scale.
Variable presentation and hypothesis tests: The paired wilcox test is used. Variables are presented as median (lower quartile to upper quartile). Binary and categorical variables are presented as number (%) of instances.
Rows in bold font correspond to variables which differed significantly between D0 and D30.

Appendix Table A3.7 | D0 baseline characteristics of 81 patients with known D30 healthcare utilisation treatment outcome. Patients who experienced healthcare utilisation treatment failure by D30 are compared against patients who had treatment success by D30.

Variable	**D30 healthcare utilisation treatment failure (n=43)**	**D30 healthcare utilisation treatment success (n=38)**	**P-Value**
Age (years)	54 (29 - 69)	51 (33 – 64)	0.53
Weight (kg)*	82 (74 – 91)	78 (72 - 85)	0.46
Pack Year History	2 (0 – 17)	4 (0 – 27)	0.49
Number of unscheduled primary care and emergency department visits in previous 12 months	**2 (1 – 5)**	**0 (0 – 3)**	**<0.01**
Number of hospital admissions in previous 12 months	0 (0 – 1)	0 (0 – 1)	0.34
Number of Intensive Therapy Unit admissions in previous 12 months	0 (0 – 0)	0 (0 – 0)	0.82
Age at diagnosis (years)	**49 (11 – 55)**	**19 (5 – 41)**	**0.02**
Age at onset of symptoms (years)	32 (11 - 53)	20 (5 – 43)	0.22

Male, n (%)		14 (33)	20 (53)	0.07
Smoking Status	Current smoker, n (%)	9 (21)	12 (32)	0.27
	Ex-smoker, n (%)	17 (40)	17 (45)	
	Non-smoker, n (%)	17 (40)	9 (24)	
Oral steroids in the last week, n (%)		25 (58)	21 (55)	0.79
Antibiotics in the last week, n (%)		19 (44)	14 (37)	0.50
Previous respiratory diagnosis	COPD, n (%)	12 (28)	10 (26)	0.87
	Asthma, n (%)	31 (72)	28 (74)	
Nasal polyps, n (%)		2 (5)	4 (11)	0.31
Comorbidity	Cardiovascular, n (%)	7 (16)	8 (21)	0.64
	Other, n (%)	17 (40)	17 (45)	
	None, n (%)	19 (44)	13 (34)	

SABA, n (%)		**32 (74)**	**35 (92)**	**0.04**
LABA, n (%)		4 (9)	2 (5)	0.49
LAMA, n (%)		**8 (19)**	**15 (39)**	**0.04**
SAMA, n (%)		1 (2)	1 (3)	0.93
ICS, n (%)		**15 (35)**	**25 (66)**	**0.01**
Carbocystiene, n (%)		5 (12)	1 (3)	0.12
Maintenance prednisolone, n (%)		4 (9)	3 (8)	0.82
Monteleukast, n (%)		10 (23)	6 (16)	0.40
Antihistamine, n (%)		3 (7)	5 (13)	0.35
PPI, n (%)		2 (5)	7 (18)	0.05
Antihypertensive, n (%)		4 (9)	4 (11)	0.85
Caucasian, n (%)		39 (93)	3 (8)	0.84
Working status	Full-time, n (%)	20 (49)	16 (46)	0.92
	Part-time, n (%)	5 (12)	6 (17)	

	Retired, n (%)	12 (29)	9 (26)	
	Unemployed, n (%)	4 (10)	4 (11)	
Previous nasal polyp surgery, n (%)		0 (0)	2 (5)	0.13
Theophylline, n (%)		2 (5)	1 (3)	0.63
Maintenance azithromycin, n (%)		1 (2)	0 (0)	0.34
Aspirin, n (%)		**0 (0)**	**4 (11)**	**0.03**
Nasal steroids use, n (%)		2 (5)	1 (3)	0.63
Nebuliser use, n (%)		2 (5)	1 (3)	0.63
Long Term Oxygen Therapy, n (%)		0 (0)	1 (3)	0.28
Anti-diabetic medication, n (%)		0 (0)	3 (8)	0.06
Beta-blockers, n (%)		1 (2)	1 (3)	0.93

Definition of abbreviations: ICS = inhaled corticosteroids; LABA = Long-Acting Beta Agonist; LAMA = Long-Acting Muscarinic Antagonist; PPI = Proton Pump Inhibitor; SABA = Short-Acting Beta Agonist; SAMA = Short-Acting Muscarinic Antagonist.
Variable presentation and hypothesis tests: Numerical variables following a normal distribution are presented as mean (standard error of the mean); log-transformed variables following a normal distribution are presented as geometric mean (95% confidence interval); other variables are presented as median (lower quartile to upper quartile).
For numerical variables following a normal distribution, unpaired two-sided t-tests are used. These variables are represented as mean and marked ~. For log-transformed variables following a normal unpaired two-sided t-tests on log-transformed data are used and marked *. For other numerical variables, the Mann-Whitney U test is used.
Binary and categorical variables are presented as number (%) of instances (chi-squared test used for treatment outcome comparison).
Rows in bold font correspond to variables which differed significantly between D30 healthcare utilisation treatment outcomes.

Appendix Table A3.8 | Paired comparison of characteristics of participants between D0 and D90.

Variable		D0 (n= 59 paired participants)	D90 (n=59 paired participants)	P-Value
Rating of how patient felt on the day		N/A	8 (5 – 9)	N/A
Whether patient felt better compared to D0	Better, n (%)	N/A	49 (88)	N/A
	No improvement, n (%)	N/A	7 (13)	
VAS cough (mm)		**51 (28 – 70)**	**8 (0 – 30)**	**<0.01**
VAS dyspnoea (mm)		**72 (49 – 89)**	**10 (0 – 50)**	**<0.01**
VAS sputum production (mm)		**25 (3 – 51)**	**2 (0 – 30)**	**<0.01**
VAS sputum purulence (mm)		**16 (0 – 54)**	**0 (0 – 15)**	**<0.01**
CAT score		**18 (0 – 26)**	**2 (0 – 13)**	**<0.01**
Exacerbation treatment	SCS and antibiotics, n (%)	25 (42)		N/A
	SCS alone, n (%)	21 (36)		
	Antibiotics alone, n (%)	6 (10)		
	Neither treatment, n (%)	7 (12)		

Definition of abbreviations: CAT score = COPD Assessment Tool; SCS = systemic corticosteroids; VAS = Visual Analogue Scale.
Variable presentation and hypothesis tests: The paired wilcox test is used. Variables are presented as median (lower quartile to upper quartile). Binary and categorical variables are presented as number (%) of instances.
Variables in bold font differ significantly between D0 and D90.

Appendix Table A3.9 | D0 baseline characteristics of 85 patients with known D90 healthcare utilisation treatment outcome. Patients who experienced D90 healthcare utilisation treatment failure by D90 are compared against patients who had treatment success by D90.

Variable	**D90 healthcare utilisation treatment failure (n=53)**	**D90 healthcare utilisation treatment success (n=32)**	**P-Value**
Age (years)	52 (29 – 68)	49 (33 – 65)	0.84
Weight (kg)*	83 (76 – 92)	76 (70 – 82)	0.16
Pack Year History	3 (0 – 17)	4 (0 – 28)	0.42
Number of unscheduled primary care and emergency department visits in previous 12 months	**2 (0 – 5)**	**1 (0 – 3)**	**0.02**
Number of hospital admissions in previous 12 months	0 (0 – 1)	0 (0 – 1)	0.09
Number of Intensive Therapy Unit admissions in previous 12 months	0 (0 – 0)	0 (0 - 0)	0.73
Age at diagnosis (years)	35 (11 – 53)	18 (4 – 43)	0.09
Age at onset of symptoms (years)	24 (9 – 50)	20 (4 – 47)	0.68
Male, n (%)	**16 (30)**	**18 (56)**	**0.02**

Smoking status	**Current smoker, n (%)**	**12 (23)**	**10 (31)**	**0.03**
	Ex-smoker, n (%)	**20 (38)**	**15 (47)**	
	Non-smoker, n (%)	**21 (40)**	**7 (22)**	
Oral steroids in the last week, n (%)		32 (60)	16 (50)	0.35
Antibiotics in the last week, n (%)		21 (40)	13 (41)	0.93
Previous respiratory diagnosis	COPD, n (%)	14 (26)	8 (25)	0.89
	Asthma, n (%)	39 (74)	24 (75)	
Nasal polyps, n (%)		2 (4)	4 (13)	0.13
Comorbidity	Cardiovascular, n (%)	8 (16)	7 (22)	0.71
	Other, n (%)	21 (42)	14 (34)	
	None, n (%)	21 (42)	11 (44)	
SABA, n (%)		42 (79)	29 (91)	0.17
LABA, n (%)		4 (8)	2 (6)	0.82
LAMA, n (%)		13 (25)	11 (34)	0.33

SAMA, n (%)		1 (2)	1 (3)	0.72
ICS, n (%)		**19 (37)**	**23 (72)**	**<0.01**
Carbocystiene, n (%)		**6 (11)**	**0 (0)**	**<0.01**
Maintenance prednisolone, n (%)		5 (9)	2 (6)	0.60
Monteleukast, n (%)		13 (25)	4 (13)	0.18
Antihistamine, n (%)		4 (8)	4 (13)	0.45
PPI, n (%)		5 (9)	5 (16)	0.39
Antihypertensive, n (%)		6 (11)	2 (6)	0.44
Caucasian, n (%)		48 (92)	27 (90)	0.72
Working status	Full-time, n (%)	25 (52)	13 (42)	0.77
	Part-time, n (%)	6 (13)	6 (19)	
	Retired, n (%)	12 (25)	9 (29)	
	Unemployed, n (%)	5 (10)	3 (10)	
Previous nasal polyp surgery, n (%)		0 (0)	2 (6)	0.07
Theophylline, n (%)		3 (6)	0 (0)	0.17
Maintenance azithromycin, n (%)		1 (2)	0 (0)	0.43

Aspirin, n (%)	**0 (0)**	**4 (13)**	**<0.01**
Nasal steroids use, n (%)	2 (4)	1 (3)	0.88
Nebuliser use, n (%)	2 (4)	1 (3)	0.88
Long Term Oxygen Therapy, n (%)	0 (0)	1 (3)	0.20
Anti-diabetic medication, n (%)	2 (4)	1 (3)	0.88
Beta-blockers, n (%)	1 (2)	1 (3)	0.72

Definition of abbreviations: ICS = inhaled corticosteroids; LABA = Long-Acting Beta Agonist; LAMA = Long-Acting Muscarinic Antagonist; PPI = Proton Pump Inhibitor; SABA = Short-Acting Beta Agonist; SAMA = Short-Acting Muscarinic Antagonist.
Variable presentation and hypothesis tests: Numerical variables following a normal distribution are presented as mean (standard error of the mean); log-transformed variables following a normal distribution are presented as geometric mean (95% confidence interval); other variables are presented as median (lower quartile to upper quartile). For numerical variables following a normal distribution, unpaired two-sided t-tests are used. These variables are represented as mean and marked ~. For log-transformed variables following a normal unpaired two-sided t-tests on log-transformed data are used and marked *. For other numerical variables, the Mann-Whitney U test is used.
Binary and categorical variables are presented as number (%) of instances (chi-squared test used for treatment outcome comparison).
Rows in bold font correspond to variables which differed significantly between D90 healthcare utilisation treatment outcomes.

Appendix Table A4.1 | Normality tests for Chapter 4 and Chapter 5 patient cohort data variables and where relevant log-transformed data variables. For variables following the normal distribution as determined by Shapiro-Wilk test $p<0.05$ or for variables with zero/negative values, then normality assumption for the log-transformed version of the variable was not performed.

Variable	Shapiro-Wilk test p-value	Shapiro-Wilk test p-value for log-transformed Variable
FEV_1 post-bronchodilator	<0.001	<0.001
Percentage predicted FEV_1	<0.001	<0.001
FEV_1 post-bronchodilator/FVC ratio	<0.001	<0.001
Age	<0.001	<0.001
Pack Year History	<0.001	<0.001
Blood cell count	<0.001	0.002
Blood neutrophil count	<0.001	<0.001
Blood eosinophil count	<0.001	<0.001
Blood monocyte count	<0.001	<0.001
Blood basophil count	<0.001	<0.001
Blood lymphocyte count	<0.001	<0.001
Percentage blood neutrophils	<0.001	<0.001
Percentage blood eosinophils	<0.001	<0.001
Percentage blood monocytes	<0.001	<0.001
Percentage blood basophils	<0.001	<0.001
Percentage blood lymphocytes	<0.001	<0.001
Colony Forming Units	<0.001	<0.001
Sputum cell count	<0.001	<0.001
Sputum neutrophil count	<0.001	<0.001
Sputum eosinophil count	<0.001	N/A
Sputum lymphocyte count	<0.001	N/A
Sputum macrophage count	<0.001	N/A
Percentage sputum neutrophils	<0.001	<0.001
Percentage sputum eosinophils	<0.001	<0.001
Percentage sputum lymphocytes	<0.001	N/A
Percentage sputum macrophages	<0.001	<0.001
CRP	<0.001	<0.001
Creatine	<0.001	<0.001
Urea	<0.001	<0.001
eGFR	<0.001	<0.001
SGRQ symptoms	<0.001	N/A
SGRQ activity	<0.001	N/A
SGRQ impacts	<0.001	N/A
CRQ emotion	<0.001	<0.001
CRQ fatigue	<0.001	<0.001
CRQ mastery	<0.001	<0.001
CRQ dyspnoea	<0.001	<0.001
VAS cough	<0.001	N/A
VAS dyspnoea	<0.001	N/A
VAS sputum production	<0.001	N/A
VAS sputum purulence	<0.001	N/A
Procalcitonin	<0.001	<0.001
Surfactant Protein D	<0.001	<0.001
Blood neopterin	<0.001	<0.001

Serum IL1B	<0.001	<0.001
Serum IL8	<0.001	<0.001
Serum CXCL10	<0.001	<0.001
Serum CCL4	<0.001	0.03
Serum TNF1A	<0.001	<0.001
Serum TNF1B	<0.001	0.01
Serum VEGFA*	<0.001	0.80
Serum IL13	<0.001	N/A
Serum IL5	<0.001	0.003
Serum IL6	<0.001	<0.001
Serum CXCL11	<0.001	<0.001
Serum CCL2	<0.001	<0.001
Serum CCL17	<0.001	<0.001
Serum TNFalpha	<0.001	<0.001
Serum IFNgamma	<0.001	N/A
Serum IL10	<0.001	N/A
Serum CCL3	<0.001	N/A
Serum CCL13	<0.001	0.002
Serum Amyloid A_1	<0.001	<0.001
sTREM	<0.001	N/A
Sputum neopterin		
Sputum IL1B	<0.001	N/A
Sputum IL6R*	<0.001	0.77
Sputum IL8	<0.001	<0.001
Sputum CXCL10	<0.001	N/A
Sputum CCL4	<0.001	N/A
Sputum TNFA*	<0.001	0.25
Sputum TNFB	<0.001	N/A
Sputum VEGFA*	<0.001	0.58
Sputum IL5	<0.001	N/A
Sputum IL6	<0.001	0.01
Sputum CXCL11	<0.001	N/A
Sputum CCL2	<0.001	0.005
Sputum CCL5	<0.001	N/A
Sputum CCL17	<0.001	N/A
Sputum TNFalpha	<0.001	N/A
Sputum CCL13	<0.001	N/A
Sputum CCL3	<0.001	N/A
Sputum MMP8	<0.001	N/A
Sputum MMP9*	<0.001	0.08
Sputum MMP1	<0.001	N/A
Sputum MMP2	<0.001	N/A
Sputum MMP3	<0.001	N/A
Sputum MMP7*	<0.001	0.26
Sputum MMP12	<0.001	N/A
FeNO*	<0.001	0.87
Haemoglobin	<0.001	<0.001
Platelets	<0.001	<0.001
Temperature	<0.001	<0.001
Heart rate*	0.01	0.38
Oxygen saturation	<0.001	<0.001
Systolic blood pressure~	0.85	N/A
Diastolic blood pressure~	0.31	N/A

~Indicates that a variable is normally distributed. * Indicates that the log-transformed version of the variable is normally distributed.

Appendix Table A4.2 | PCA output for symptoms and blood inflammatory indices. Loading coefficients showing relationship between original variables and each principal component are shown, as well as the cumulative proportion of variance explained by principal components. The first three principal components subsequently used as axes for visualisation could explain 74% of cumulative variance.

	PC1	**PC2**	**PC3**	PC4	PC5	PC6	PC7	PC8	PC9
Blood neutrophil count	**-0.33**	**0.19**	**-0.56**	0.17	-0.07	0.00	0.64	-0.11	0.29
Blood eosinophil count	**0.13**	**-0.57**	**-0.43**	0.09	-0.03	0.03	0.13	-0.03	-0.67
Blood neutrophil percentage	**-0.31**	**0.30**	**-0.41**	0.41	-0.12	0.01	-0.64	0.16	-0.13
Blood eosinophil percentage	**0.21**	**-0.58**	**-0.30**	0.02	-0.01	0.00	-0.28	0.06	0.67
CRP	**-0.29**	**0.06**	**-0.33**	-0.84	0.23	0.08	-0.19	-0.00	-0.05
VAS cough	**-0.41**	**-0.24**	**0.23**	0.09	0.05	0.73	0.10	0.40	0.02
VAS dyspnoea	**-0.36**	**-0.19**	**0.12**	0.25	0.80	-0.33	-0.01	-0.11	0.00
VAS sputum production	**-0.43**	**-0.24**	**0.20**	0.00	-0.36	0.08	-0.15	-0.75	0.01
VAS sputum purulence	**-0.41**	**-0.24**	**0.16**	-0.13	-0.40	-0.58	0.09	0.47	-0.02
Cumulative proportion of variance	**0.36**	**0.59**	**0.74**	0.82	0.89	0.93	0.97	1.00	1.00

Definition of abbreviations: VAS = Visual Analogue Scale; CRP = C-reactive protein.
Principal components in bold font were subsequently used as axes for visualisation.

Appendix Table A4.3 | Proportion of each of first three principal components (representing symptoms and blood inflammatory indices) explained by each symptom and blood inflammatory index. For each principal component, the sum of the proportions of each symptom and blood inflammatory index explaining the principal component is equal to 1.00 but may deviate slightly in this table due to decimal rounding.

	PC1	PC2	PC3
Blood neutrophil count	0.11	0.04	0.32
Blood eosinophil count	0.02	0.32	0.19
Blood neutrophil percentage	0.10	0.09	0.17
Blood eosinophil percentage	0.04	0.34	0.09
CRP	0.09	0.00	0.11
VAS cough	0.17	0.06	0.05
VAS dyspnoea	0.13	0.04	0.01
VAS sputum Production	0.19	0.06	0.04
VAS sputum purulence	0.17	0.06	0.03

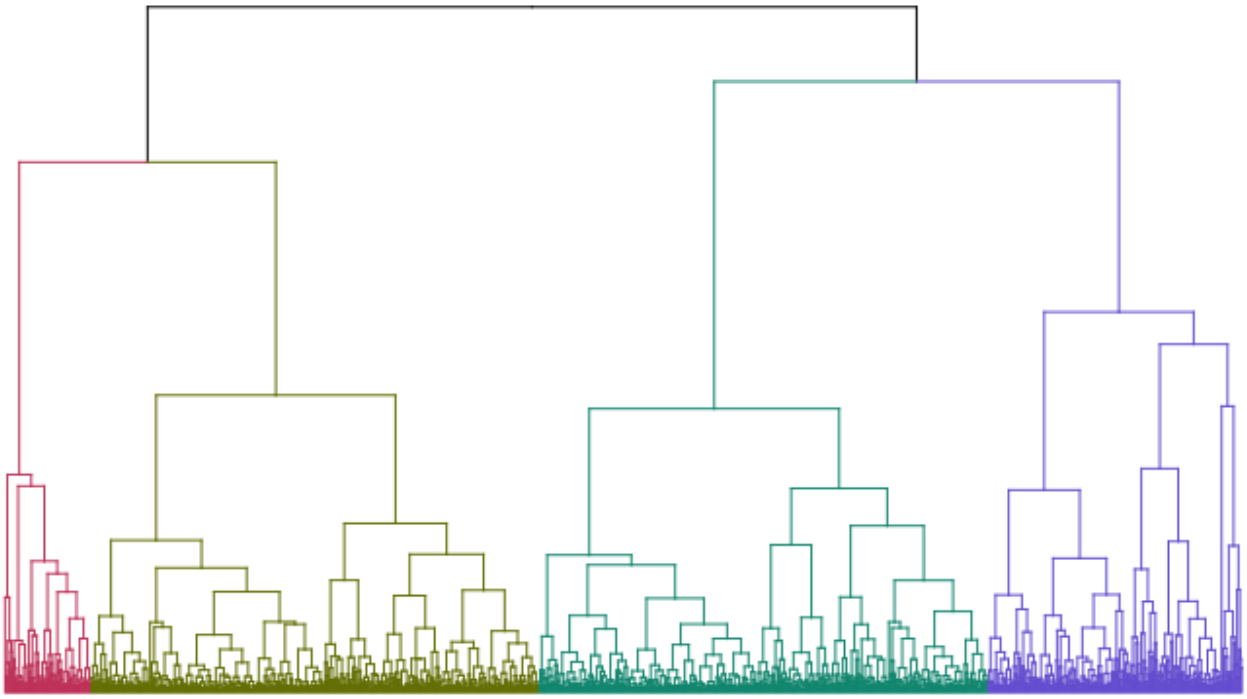
Definition of abbreviations: VAS = Visual Analogue Scale; CRP = C-reactive protein.

Appendix Table A4.4 | PCA output for symptoms, blood and sputum inflammatory indices. Loading coefficients showing relationship between original variables and each principal component are shown, as well as the cumulative proportion of variance explained by principal components. The first five principal components explain 75% of cumulative variance.

	PC1	**PC2**	**PC3**	**PC4**	**PC5**	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
Blood neutrophil count	**-0.31**	**0.00**	**-0.45**	**0.25**	**0.35**	-0.07	0.06	-0.08	-0.02	-0.08	-0.64	-0.08	-0.28
Blood eosinophil count	**0.20**	**-0.46**	**-0.20**	**-0.25**	**0.42**	-0.09	0.01	-0.04	-0.09	-0.10	-0.08	-0.04	0.66
Blood neutrophil percentage	**-0.30**	**0.10**	**-0.35**	**0.40**	**0.25**	-0.29	0.14	0.04	0.04	0.14	0.63	0.14	0.12
Blood eosinophil percentage	**0.27**	**-0.45**	**-0.11**	**-0.27**	**0.29**	-0.04	-0.00	0.02	-0.01	0.01	0.30	0.05	-0.68
CRP	**-0.29**	**-0.07**	**-0.34**	**-0.11**	**-0.05**	0.72	-0.27	0.36	-0.19	-0.04	0.15	-0.00	0.04
VAS cough	**-0.32**	**-0.29**	**0.30**	**0.06**	**-0.04**	-0.05	-0.05	-0.20	-0.62	0.34	-0.09	0.41	-0.02
VAS dyspnoea	**-0.28**	**-0.26**	**0.19**	**0.12**	**0.01**	-0.25	-0.78	-0.00	0.30	-0.14	0.02	-0.11	-0.00
VAS sputum production	**-0.35**	**-0.30**	**0.27**	**0.04**	**-0.00**	0.04	0.35	0.06	-0.04	0.12	0.09	-0.75	-0.01
VAS sputum purulence	**-0.33**	**-0.29**	**0.23**	**-0.03**	**-0.03**	0.12	0.40	0.14	0.45	-0.38	-0.05	0.47	0.02
Sputum neutrophil percentage	**-0.25**	**0.14**	**-0.12**	**-0.56**	**-0.12**	-0.47	0.02	0.54	-0.01	0.23	-0.12	0.05	-0.01
Sputum eosinophil percentage	**0.23**	**-0.37**	**-0.19**	**0.29**	**-0.28**	0.08	0.02	0.16	0.37	0.64	-0.17	0.08	0.06
Sputum neutrophil count	**-0.27**	**-0.03**	**-0.34**	**-0.42**	**-0.31**	0.04	0.02	-0.68	0.21	0.11	0.10	-0.03	0.02
Sputum eosinophil count	**0.11**	**-0.30**	**-0.31**	**0.20**	**-0.60**	-0.27	0.07	0.09	-0.32	-0.45	0.02	-0.06	-0.02
Cumulative proportion of variance	**0.28**	**0.47**	**0.59**	**0.68**	**0.76**	0.81	0.86	0.89	0.93	0.96	0.98	1.00	1.00

Definition of abbreviations: VAS = Visual Analogue Scale; CRP = C-reactive protein.
Principal components in bold font explain ≥ 75% cumulative variance.

Appendix Figure 4A.1 | Dendrogram showing four clusters within the 1,348 events.



Appendix Table A5.1 | PCA output representing exacerbation minus two-week post-treatment follow-up symptoms, blood and sputum inflammatory indices. Loading coefficients showing relationship between original variables and each principal component are shown. The first five principal components explain 75% of cumulative variance.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
VAS cough	**-0.06**	**-0.44**	**0.06**	**0.28**	**-0.05**	0.13	-0.21	0.66	-0.41	0.11	0.20	-0.05	-0.03
VAS dyspnoea	**-0.00**	**-0.41**	**-0.03**	**0.31**	**0.29**	0.23	-0.57	-0.39	0.24	-0.26	-0.08	0.07	-0.03
VAS sputum production	**-0.12**	**-0.46**	**0.11**	**0.20**	**-0.15**	-0.16	0.37	-0.11	-0.04	0.18	-0.70	0.03	0.02
VAS sputum purulence	**-0.14**	**-0.44**	**0.07**	**0.06**	**-0.23**	-0.29	0.34	-0.27	0.11	-0.19	0.63	0.00	0.02
Blood neutrophils	**-0.41**	**-0.00**	**-0.31**	**0.03**	**0.35**	0.17	0.22	-0.04	-0.07	-0.07	0.00	-0.72	0.11
Blood eosinophils	**0.36**	**-0.22**	**0.09**	**-0.27**	**0.40**	0.20	0.33	0.02	-0.09	-0.14	0.01	0.00	-0.63
Percentage blood neutrophils	**-0.38**	**0.06**	**-0.43**	**0.09**	**0.17**	0.28	0.27	-0.01	-0.17	-0.07	0.04	0.67	0.04
Percentage blood eosinophils	**0.42**	**-0.21**	**0.10**	**-0.21**	**0.32**	0.14	0.16	-0.01	-0.10	0.02	0.05	0.08	0.75
CRP	**-0.22**	**-0.26**	**-0.22**	**-0.43**	**0.26**	-0.36	-0.14	0.38	0.50	0.13	-0.05	0.11	0.00
Sputum neutrophil count	**-0.25**	**-0.18**	**0.01**	**-0.62**	**-0.15**	-0.03	-0.32	-0.32	-0.54	0.04	-0.05	0.00	-0.02
Sputum eosinophil count	**0.24**	**-0.15**	**-0.44**	**-0.20**	**-0.54**	0.29	0.03	0.18	0.16	-0.47	-0.16	-0.10	0.05
Percentage sputum neutrophils	**-0.28**	**-0.07**	**0.40**	**-0.20**	**-0.20**	0.66	0.08	0.01	0.36	0.31	0.09	-0.00	0.01
Percentage sputum eosinophils	**0.32**	**-0.13**	**-0.53**	**0.07**	**-0.10**	0.07	-0.04	-0.24	0.03	0.70	0.16	-0.07	-0.11
Cumulative proportion of variance	**0.30**	**0.51**	**0.61**	**0.69**	**0.76**	0.81	0.86	0.90	0.93	0.96	0.98	1.00	1.00

Definition of abbreviations: VAS = Visual Analogue Scale; CRP = C-reactive protein.
Principal components in bold font explain ≥ 75% cumulative variance.

Appendix chapter 3 code

R version

```
R.version

_
platform       x86_64-apple-darwin17.0
arch           x86_64
os             darwin17.0
system         x86_64, darwin17.0
status
major          4
minor          2.1
year           2022
month          06
day            23
svn rev        82513
language       R
version.string R version 4.2.1 (2022-06-23)
nickname       Funny-Looking Kid
```

R libraries used for analysis

```
library(MXM)
library(glmnet)
library(caret)
library(kernlab)
library(nnet)
library(randomForest)
library(rpart)
library(gbm)
library(MLeval)
library(missForest)
```

Creation of hold-out test set of patients as well as cross-validation folds in remaining data for treatment outcome analysis and physician treatment prescription analysis

```
set.seed(5782)
split_0.8_train_0.2_test_81_Patients_D30_Hard <-
createDataPartition(Imputed_81_Patients_D30_With_Hard_Treatment_Outcome$D3
0_Treatment_Outcome_Based_on_Readmission_or_Need_of_Additional_Steroids_o
r_Antibiotics, p=0.8, list = FALSE)
Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN <-
Imputed_81_Patients_D30_With_Hard_Treatment_Outcome[split_0.8_train_0.2_test
_81_Patients_D30_Hard,]
Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TEST <-
Imputed_81_Patients_D30_With_Hard_Treatment_Outcome[-
split_0.8_train_0.2_test_81_Patients_D30_Hard,]
```

```
Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.2_TEST <-
Imputed_81_Patients_D30_With_Hard_Treatment_Outcome[-
split_0.8_train_0.2_test_81_Patients_D30_Hard,]
Imputed_81_Patients_D30_Without_Hard_Treatment_Outcome_0.8_TRAIN$D30_Tr
eatment_Outcome_Based_on_Readmission_or_Need_of_Additional_Steroids_or_A
ntibiotics <- NULL

set.seed(5782)
split_0.8_train_0.2_test_81_Patients_D30_Hard_Soft <-
createDataPartition(Imputed_81_Patients_With_D30_Hard_and_Soft_Treatment_Ou
tcomes_SCS_and_or_Abx_at_exacerbation$D30_Hard_and_Soft_Outcome, p=0.8,
list = FALSE)
set.seed(5782)
split_0.8_train_0.2_test_85_Patients_D90_Hard <-
createDataPartition(Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_wit
h_Exacerbation_Treatment$D90_Hard_Outcome, p=0.8, list = FALSE)
set.seed(5782)
split_0.8_train_0.2_test_90_Patients_D1_D5_Hard <-
createDataPartition(Imputed_90_Patients_With_Different_Treatment_Outcomes_Exa
cerbation_Treatment$D1_D5_Hard_Treatment_Outcome, p=0.8, list = FALSE)
set.seed(5782)
split_0.8_train_0.2_test_101_Patients_D0_Antibiotic_Prescription_Prediction <-
createDataPartition(Imputed_101_Patients_Antibiotic_Prescription_Prediction$Antibi
oticsToday, p=0.8, list = FALSE)
set.seed(5782)
split_0.8_train_0.2_test_101_Patients_D0_SCS_Prescription_Prediction <-
createDataPartition(Imputed_101_Patients_SCS_Prescription_Prediction$OralPredni
solone, p=0.8, list = FALSE)

Imputed_81_Patients_With_D30_Hard_and_Soft_Treatment_Outcomes_SCS_and_
or_Abx_at_exacerbation_0.8_TRAIN <-
Imputed_81_Patients_With_D30_Hard_and_Soft_Treatment_Outcomes_SCS_and_
or_Abx_at_exacerbation[split_0.8_train_0.2_test_81_Patients_D30_Hard_Soft,]
Imputed_81_Patients_With_D30_Hard_and_Soft_Treatment_Outcomes_Exacerbatio
n_Treatment_0.8_TRAIN <-
Imputed_81_Patients_With_D30_Hard_and_Soft_Treatment_Outcomes_Exacerbatio
n_Treatment[split_0.8_train_0.2_test_81_Patients_D30_Hard_Soft,]
Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_Exacerbation_Tre
atment_0.8_TRAIN <-
Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_Exacerbation_Tre
atment[split_0.8_train_0.2_test_85_Patients_D90_Hard,]
Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_SCS_and_or_Abx_at_
exacerbation_0.8_TRAIN <-
Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_SCS_and_or_Abx_at_
exacerbation[split_0.8_train_0.2_test_85_Patients_D90_Hard,]
Imputed_90_Patients_With_Different_Treatment_Outcomes_Exacerbation_Treatme
nt_0.8_TRAIN <-
Imputed_90_Patients_With_Different_Treatment_Outcomes_Exacerbation_Treatme
nt[split_0.8_train_0.2_test_90_Patients_D1_D5_Hard,]
```

```
Imputed_90_Patients_With_Different_Treatment_Outcomes_SCS_and_or_Abx_at_exacerbation_0.8_TRAIN <-
Imputed_90_Patients_With_Different_Treatment_Outcomes_SCS_and_or_Abx_at_exacerbation[split_0.8_train_0.2_test_90_Patients_D1_D5_Hard,]
Imputed_101_Patients_Antibiotic_Prescription_Prediction_0.8_TRAIN <-
Imputed_101_Patients_Antibiotic_Prescription_Prediction[split_0.8_train_0.2_test_101_Patients_D0_Antibiotic_Prescription_Prediction,]
Imputed_101_Patients_SCS_Prescription_Prediction_0.8_TRAIN <-
Imputed_101_Patients_SCS_Prescription_Prediction[split_0.8_train_0.2_test_101_Patients_D0_SCS_Prescription_Prediction,]
```

Identification of key variables for predicting treatment failure and predicting physician treatment prescription decision using statistically equivalent signatures algorithm

```
set.seed(5782)
Draft_81_Patients_D30_Hard_TRAIN_0.8_sesObject_max_k_91_threshold_0.1_Including_Exacerbation_Treatment <-
SES(Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN$D30_Treatment_Outcome_Based_on_Readmission_or_Need_of_Additional_Steroids_or_Antibiotics,
Imputed_81_Patients_D30_Without_Hard_Treatment_Outcome_0.8_TRAIN_Including_Exacerbation_Treatment, max_k = 91, threshold = 0.1, test = NULL, hash = TRUE, hashObject = NULL)
D30_Hard_SES_logged_p_values_predictor_set_including_exacerbation_treatment <- c(-0.31086087, -1.58155033, -1.32628086, -0.41832687, -0.50678303, -1.82503911, -0.82462834, -1.87240402, -0.95918590, -0.72188424, -0.57674735, -0.06212775, -0.22097363, -4.19979353, -1.15472695, -0.23203974, -2.16148485, -1.22790522, -1.26793825, -0.18649449, -0.66357949, -2.18003713, -1.28224410, -6.18593228, -2.28705796, -1.92159415, -0.16445415, -0.40985430, -0.22938262, -1.83296458, -2.00814196, -1.87724909, -1.03344858, -1.82356927, -0.02582173, -0.93781423, -0.97184116, -0.02082967, -1.97407686, -0.02897822, -0.48402814, -1.52285616, -0.42961016, -1.01088290, -0.09387345, -1.86785303, -1.55773057, -2.54673667, -1.16473051, -1.08343399, -0.20177490, -0.52980934, -0.06216859, -0.06216859, -1.98724963, -0.03346766, -1.17674450, -0.56295955, -1.79483202, -0.73840009, -1.51890744, -2.96299043, -1.64962407, -0.13222363, -1.68256121, -1.05856716, -1.17674450, -1.04587403, -0.60971270, -0.05705113, -1.00248625, -0.50306909, -1.53106507, -2.23148337, -1.35679578, -2.22520270, -0.07195263, -0.07195263, -1.53106507, -1.90816777, -1.53106507, -0.84885570, -0.27579304, -1.53106507, -2.82984446, -0.07195263, -0.62771320, -1.03818929, -2.29979051, -1.09195886, -0.15409216, -2.26384803)
D30_Hard_SES_p_values_predictor_set_including_exacerbation_treatment <-
exp(D30_Hard_SES_logged_p_values_predictor_set_including_exacerbation_treatment)


set.seed(5782)
Draft_81_Patients_D30_Hard_Soft_TRAIN_0.8_sesObject_max_k_91_threshold_0.1_Including_Exacerbation_Treatment <-
```

```
SES(Imputed_81_Patients_With_D30_Hard_and_Soft_Treatment_Outcomes_Exacerbation_Treatment_0.8_TRAIN$D30_Hard_and_Soft_Treatment_Outcome, Imputed_81_Patients_With_D30_Hard_and_Soft_Treatment_Outcomes_Exacerbation_Treatment_0.8_TRAIN[,1:92], max_k = 91, threshold = 0.1, test = NULL, hash = TRUE, hashObject = NULL)

set.seed(5782)
Draft_85_Patients_D90_Hard_TRAIN_0.8_sesObject_max_k_91_threshold_0.1_Including_Exacerbation_Treatment <-
SES(Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_Exacerbation_Treatment_0.8_TRAIN$D90_Hard_Outcome,
Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_Exacerbation_Treatment_0.8_TRAIN[,1:92], max_k = 91, threshold = 0.1, test = NULL, hash = TRUE, hashObject = NULL)


Imputed_90_Patients_D1_D5_Hospitalisation_Outcomes_Exacerbation_Treatment_0.8_TRAIN <-
Imputed_90_Patients_With_Different_Treatment_Outcomes_Exacerbation_Treatment_0.8_TRAIN

set.seed(5782)
Draft_90_Patients_D1_D5_Hospitalisation_TRAIN_0.8_sesObject_max_k_91_threshold_0.1_Including_Exacerbation_Treatment <-
SES(Imputed_90_Patients_D1_D5_Hospitalisation_Outcomes_Exacerbation_Treatment_0.8_TRAIN$D1_D5_Hard_Treatment_Outcome,
Imputed_90_Patients_D1_D5_Hospitalisation_Outcomes_Exacerbation_Treatment_0.8_TRAIN[,1:92], max_k = 91, threshold = 0.1, test = NULL, hash = TRUE, hashObject = NULL)

set.seed(5782)
Draft_101_Antibiotic_Prescription_Prediction_0.8_sesObject_max_k_90_threshold_0.1 <-
SES(Imputed_101_Patients_Antibiotic_Prescription_Prediction_0.8_TRAIN$AntibioticsToday,
Imputed_101_Patients_Antibiotic_Prescription_Prediction_0.8_TRAIN[,1:91], max_k = 90, threshold = 0.1, test = NULL, hash = TRUE, hashObject = NULL)

set.seed(5782)
Draft_101_SCS_Prescription_Prediction_0.8_sesObject_max_k_90_threshold_0.1 <-
SES(Imputed_101_Patients_SCS_Prescription_Prediction_0.8_TRAIN$OralPrednisolone, Imputed_101_Patients_SCS_Prescription_Prediction_0.8_TRAIN[,1:91], max_k = 90, threshold = 0.1, test = NULL, hash = TRUE, hashObject = NULL)
```

Lasso classifier for predicting treatment failure and predicting physician treatment prescription decision using statistically equivalent signatures algorithm

```
set.seed(5782)
cv.glmnet_Matrix_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_
TRAIN <-
cv.glmnet(Matrix_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_
TRAIN[,2:92],
Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN$D30_Treat
ment_Outcome_Based_on_Readmission_or_Need_of_Additional_Steroids_or_Antibi
otics_Failure_Alphabetically_Last, weights = NULL, offset = NULL, lambda = NULL,
type.measure = "auc", nfolds = 4, foldid = NULL, alignment = "fraction", grouped =
TRUE, keep = TRUE, parallel = FALSE, gamma = c(0, 0.25, 0.5, 0.75, 1), relax =
TRUE, trace.it = 0, alpha = 1, family = "binomial")
cv.glmnet_Matrix_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_
TRAIN
cv.glmnet_Matrix_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_
TRAIN$lambda.min
cv.glmnet_Matrix_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_
TRAIN$lambda.1se
cv.glmnet_Matrix_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_
TRAIN$lambda
cv.glmnet_Matrix_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_
TRAIN$cvm
coef(cv.glmnet_Matrix_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_
0.8_TRAIN, s=0.08347783)
coef(cv.glmnet_Matrix_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_
0.8_TRAIN, s=0.21165)

set.seed(5782)
cv.glmnet_Matrix_Imputed_81_Patients_With_D30_Hard_and_Soft_Treatment_Outc
omes_SCS_and_or_Abx_at_exacerbation_0.8_TRAIN_gamma_1 <-
cv.glmnet(Matrix_Imputed_81_Patients_With_D30_Hard_and_Soft_Treatment_Outc
omes_SCS_and_or_Abx_at_exacerbation_0.8_TRAIN[,1:93],
Imputed_81_Patients_With_D30_Hard_and_Soft_Treatment_Outcomes_SCS_and_
or_Abx_at_exacerbation_0.8_TRAIN$D30_Hard_and_Soft_Outcome_Failure_Alpha
betically_Last, weights = NULL, offset = NULL, type.measure = "auc", nfolds = 4,
foldid = NULL, alignment = "fraction", grouped = TRUE, keep = TRUE, parallel =
FALSE, gamma = 1, relax = TRUE, trace.it = 0, alpha = 1, family = "binomial")
set.seed(5782)
cv.glmnet_Matrix_Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_SCS
_and_or_Abx_at_exacerbation_0.8_TRAIN_gamma_1 <-
cv.glmnet(Matrix_Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_SCS
_and_or_Abx_at_exacerbation_0.8_TRAIN[,1:93],
Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_SCS_and_or_Abx_at_
exacerbation_0.8_TRAIN$D90_Hard_Outcome_Failure_Alphabetically_Last, weights
= NULL, offset = NULL, type.measure = "auc", nfolds = 4, foldid = NULL, alignment =
"fraction", grouped = TRUE, keep = TRUE, parallel = FALSE, gamma = 1, relax =
TRUE, trace.it = 0, alpha = 1, family = "binomial")
set.seed(5782)
cv.glmnet_Matrix_Imputed_90_Patients_With_Different_Treatment_Outcomes_SCS
_and_or_Abx_at_exacerbation_0.8_TRAIN_gamma_1 <-
```

```
cv.glmnet(Matrix_Imputed_90_Patients_With_Different_Treatment_Outcomes_SCS_
and_or_Abx_at_exacerbation_0.8_TRAIN[,1:93],
Imputed_90_Patients_With_Different_Treatment_Outcomes_SCS_and_or_Abx_at_e
xacerbation_0.8_TRAIN$D1_D5_Hard_Treatment_Outcome_Failure_Alphabetically_
Last, weights = NULL, offset = NULL, type.measure = "auc", nfolds = 4, foldid =
NULL, alignment = "fraction", grouped = TRUE, keep = TRUE, parallel = FALSE,
gamma = 1, relax = TRUE, trace.it = 0, alpha = 1, family = "binomial")
set.seed(5782)
cv.glmnet_Matrix_Imputed_101_Patients_Antibiotic_Prescription_Prediction_0.8_TR
AIN_gamma_1 <-
cv.glmnet(Matrix_Imputed_101_Patients_Antibiotic_Prescription_Prediction_0.8_TR
AIN[,1:91],
Imputed_101_Patients_Antibiotic_Prescription_Prediction_0.8_TRAIN$AntibioticsTo
day_Abx_Given_Alphabetically_Last, weights = NULL, offset = NULL, type.measure
= "auc", nfolds = 4, foldid = NULL, alignment = "fraction", grouped = TRUE, keep =
TRUE, parallel = FALSE, gamma = 1, relax = TRUE, trace.it = 0, alpha = 1, family =
"binomial")
set.seed(5782)
cv.glmnet_Matrix_Imputed_101_Patients_SCS_Prescription_Prediction_0.8_TRAIN_
gamma_1 <-
cv.glmnet(Matrix_Imputed_101_Patients_SCS_Prescription_Prediction_0.8_TRAIN[,
1:91],
Imputed_101_Patients_SCS_Prescription_Prediction_0.8_TRAIN$OralPrednisolone
_SCS_Given_Alphabetically_Last, weights = NULL, offset = NULL, type.measure =
"auc", nfolds = 4, foldid = NULL, alignment = "fraction", grouped = TRUE, keep =
TRUE, parallel = FALSE, gamma = 1, relax = TRUE, trace.it = 0, alpha = 1, family =
"binomial")
```

Non-linear classifiers for predicting treatment failure and predicting physician treatment prescription decision

```
set.seed(5782)
training_control <- trainControl(method = "cv",
                         summaryFunction = twoClassSummary,
                         classProbs = TRUE,
                         number = 4, repeats = NA, savePredictions = "all",
set.seed(5782))
set.seed(5782)
knn_cv_4_folds_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_T
RAIN <-
train(D30_Treatment_Outcome_Based_on_Readmission_or_Need_of_Additional_St
eroids_or_Antibiotics ~ .,
                                                    data =
Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN_Selected_
Variables_from_SES_Threshold_0.1_k_91_Categories_Coded_As_Numbers,
                                                    method = "knn",
                                                    trControl = training_control,
                                                    metric = "ROC",
```

```
                                                    tuneGrid = data.frame(k =
seq(1,49,by = 2)))

knn_cv_4_folds_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_T
RAIN
plot(knn_cv_4_folds_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.
8_TRAIN)


set.seed(5782)
sigmaRange_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRA
IN <-
sigest(Matrix_Imputed_81_Patients_D30_Without_Hard_Treatment_Outcome_0.8_T
RAIN_Selected_Variables_from_SES_Threshold_0.1_k_91)
set.seed(5782)
svmRGRid_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN
<- expand.grid(.sigma =
seq(sigmaRange_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_
TRAIN[1],sigmaRange_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome
_0.8_TRAIN[3],by =
((sigmaRange_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TR
AIN[3] -
sigmaRange_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRA
IN[1])/8)), .C = 2^(seq(-4,4)))

set.seed(5782)
SVM_Radial_cv_4_folds_Imputed_81_Patients_D30_With_Hard_Treatment_Outcom
e_0.8_TRAIN <-
train(D30_Treatment_Outcome_Based_on_Readmission_or_Need_of_Additional_St
eroids_or_Antibiotics ~ .,
                                                    data =
Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN_Selected_
Variables_from_SES_Threshold_0.1_k_91_Categories_Coded_As_Numbers,
                                                    method = "svmRadial",
                                                    trControl =
training_control,
                                                    metric = "ROC",
                                                    preProcess = c("center",
"scale"),
                                                    fit = FALSE,
                                                    tuneGrid =
svmRGRid_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN
)

SVM_Radial_cv_4_folds_Imputed_81_Patients_D30_With_Hard_Treatment_Outcom
e_0.8_TRAIN
plot(SVM_Radial_cv_4_folds_Imputed_81_Patients_D30_With_Hard_Treatment_Ou
tcome_0.8_TRAIN)

set.seed(5782)
```

```
nnetRGRid_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN <- expand.grid(.size = 1:10, .decay = seq(0, 2, by = 0.1))
maxSize_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN <-
max(nnetRGRid_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN$.size)
numWts_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN <- 1*(maxSize_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN *
(ncol(Imputed_81_Patients_D30_Without_Hard_Treatment_Outcome_0.8_TRAIN_Selected_Variables_from_SES_Threshold_0.1_k_91_Categories_Coded_As_Numbers) + 1) +
maxSize_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN + 1)
set.seed(5782)
nnet_cv_4_folds_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN <-
train(D30_Treatment_Outcome_Based_on_Readmission_or_Need_of_Additional_Steroids_or_Antibiotics ~ .,
                                                    data = Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN_Selected_Variables_from_SES_Threshold_0.1_k_91_Categories_Coded_As_Numbers,
                                                    method = "nnet",
                                                    trControl = training_control,
                                                    metric = "ROC",
                                                    preProcess = c("center", "scale", "spatialSign"),
                                                    tuneGrid = nnetRGRid_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN,
                                                    trace = FALSE,
                                                    maxit = 2000,
                                                    MaxNWts = numWts_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN)


nnet_cv_4_folds_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN
plot(nnet_cv_4_folds_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN)


set.seed(5782)
RandomForestGRid_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN <- expand.grid(.mtry = c(1:4))
set.seed(5782)
randomforest_cv_4_folds_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN_ntree_100 <-
train(D30_Treatment_Outcome_Based_on_Readmission_or_Need_of_Additional_Steroids_or_Antibiotics ~ .,
```

```
                                                    data =
Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN_Selected_
Variables_from_SES_Threshold_0.1_k_91,
                                                    method = "rf",
                                                    trControl =
training_control,
                                                    metric = "ROC",
                                                    ntree = 100,
                                                    tuneGrid =
RandomForestGRid_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.
8_TRAIN)
set.seed(5782)
randomforest_cv_4_folds_Imputed_81_Patients_D30_With_Hard_Treatment_Outco
me_0.8_TRAIN_ntree_200 <-
train(D30_Treatment_Outcome_Based_on_Readmission_or_Need_of_Additional_St
eroids_or_Antibiotics ~ .,
                                                    data =
Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN_Selected_
Variables_from_SES_Threshold_0.1_k_91,
                                                    method = "rf",
                                                    trControl =
training_control,
                                                    metric = "ROC",
                                                    ntree = 200,
                                                    tuneGrid =
RandomForestGRid_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.
8_TRAIN)


set.seed(5782)
randomforest_cv_4_folds_Imputed_81_Patients_D30_With_Hard_Treatment_Outco
me_0.8_TRAIN_ntree_300 <-
train(D30_Treatment_Outcome_Based_on_Readmission_or_Need_of_Additional_St
eroids_or_Antibiotics ~ .,
                                                    data =
Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN_Selected_
Variables_from_SES_Threshold_0.1_k_91,
                                                    method = "rf",
                                                    trControl =
training_control,
                                                    metric = "ROC",
                                                    ntree = 300,
                                                    tuneGrid =
RandomForestGRid_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.
8_TRAIN)


set.seed(5782)
randomforest_cv_4_folds_Imputed_81_Patients_D30_With_Hard_Treatment_Outco
me_0.8_TRAIN_ntree_400 <-
```

```
train(D30_Treatment_Outcome_Based_on_Readmission_or_Need_of_Additional_Steroids_or_Antibiotics ~ .,
                                                                    data =
Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN_Selected_Variables_from_SES_Threshold_0.1_k_91,
                                                                    method = "rf",
                                                                    trControl =
training_control,
                                                                    metric = "ROC",
                                                                    ntree = 400,
                                                                    tuneGrid =
RandomForestGRid_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN)

set.seed(5782)
randomforest_cv_4_folds_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN_ntree_500 <-
train(D30_Treatment_Outcome_Based_on_Readmission_or_Need_of_Additional_Steroids_or_Antibiotics ~ .,
                                                                    data =
Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN_Selected_Variables_from_SES_Threshold_0.1_k_91,
                                                                    method = "rf",
                                                                    trControl =
training_control,
                                                                    metric = "ROC",
                                                                    ntree = 500,
                                                                    tuneGrid =
RandomForestGRid_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN)


set.seed(5782)
randomforest_cv_4_folds_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN_ntree_600 <-
train(D30_Treatment_Outcome_Based_on_Readmission_or_Need_of_Additional_Steroids_or_Antibiotics ~ .,
                                                                    data =
Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN_Selected_Variables_from_SES_Threshold_0.1_k_91,
                                                                    method = "rf",
                                                                    trControl =
training_control,
                                                                    metric = "ROC",
                                                                    ntree = 600,
                                                                    tuneGrid =
RandomForestGRid_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN)

set.seed(5782)
```

```
randomforest_cv_4_folds_Imputed_81_Patients_D30_With_Hard_Treatment_Outco
me_0.8_TRAIN_ntree_700 <-
train(D30_Treatment_Outcome_Based_on_Readmission_or_Need_of_Additional_St
eroids_or_Antibiotics ~ .,
                                                                    data =
Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN_Selected_
Variables_from_SES_Threshold_0.1_k_91,
                                                                    method = "rf",
                                                                    trControl =
training_control,
                                                                    metric = "ROC",
                                                                    ntree = 700,
                                                                    tuneGrid =
RandomForestGRid_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.
8_TRAIN)


set.seed(5782)
randomforest_cv_4_folds_Imputed_81_Patients_D30_With_Hard_Treatment_Outco
me_0.8_TRAIN_ntree_800 <-
train(D30_Treatment_Outcome_Based_on_Readmission_or_Need_of_Additional_St
eroids_or_Antibiotics ~ .,
                                                                    data =
Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN_Selected_
Variables_from_SES_Threshold_0.1_k_91,
                                                                    method = "rf",
                                                                    trControl =
training_control,
                                                                    metric = "ROC",
                                                                    ntree = 800,
                                                                    tuneGrid =
RandomForestGRid_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.
8_TRAIN)

set.seed(5782)
randomforest_cv_4_folds_Imputed_81_Patients_D30_With_Hard_Treatment_Outco
me_0.8_TRAIN_ntree_900 <-
train(D30_Treatment_Outcome_Based_on_Readmission_or_Need_of_Additional_St
eroids_or_Antibiotics ~ .,
                                                                    data =
Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN_Selected_
Variables_from_SES_Threshold_0.1_k_91,
                                                                    method = "rf",
                                                                    trControl =
training_control,
                                                                    metric = "ROC",
                                                                    ntree = 900,
                                                                    tuneGrid =
RandomForestGRid_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.
8_TRAIN)
```

```
set.seed(5782)
randomforest_cv_4_folds_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN_ntree_1000 <-
train(D30_Treatment_Outcome_Based_on_Readmission_or_Need_of_Additional_Steroids_or_Antibiotics ~ .,
                                                                    data =
Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN_Selected_Variables_from_SES_Threshold_0.1_k_91,
                                                                    method = "rf",
                                                                    trControl =
training_control,
                                                                    metric = "ROC",
                                                                    ntree = 1000,
                                                                    tuneGrid =
RandomForestGRid_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN)

set.seed(5782)
cart1_cv_4_folds_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN <-
train(D30_Treatment_Outcome_Based_on_Readmission_or_Need_of_Additional_Steroids_or_Antibiotics ~ .,
                                                                data =
Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN_Selected_Variables_from_SES_Threshold_0.1_k_91,
                                                                method = "rpart",
                                                                trControl = training_control,
                                                                metric = "ROC",
                                                                tuneLength = 10)

svmFuncs <- caretFuncs
fiveStats <- function(...) c(twoClassSummary(...),defaultSummary(...))
svmFuncs$summary <- fiveStats
set.seed(5782)
index <-
createMultiFolds(Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN_Selected_Variables_from_SES_Threshold_0.1_k_91_Categories_Coded_As_Numbers$D30_Treatment_Outcome_Based_on_Readmission_or_Need_of_Additional_Steroids_or_Antibiotics, times=4)
set.seed(5782)
ctrl <- rfeControl(method="cv", verbose = TRUE, functions = svmFuncs, index =
index, number = 4, repeats = 1)
varSeq <- seq(1,4, by=1)
set.seed(5782)
seeds <- vector(mode = "list", length = 5)
for(i in 1:4) seeds[[i]] <- sample.int(1000, 5)
seeds[[5]] <- sample.int(1000, 1)

set.seed(5782)
```

```
rfe_SVM_Radial_cv_4_folds_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN_rfe_starting_with_5_variables_from_SES <- rfe(y = Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN_Selected_Variables_from_SES_Threshold_0.1_k_91_Categories_Coded_As_Numbers$D30_Treatment_Outcome_Based_on_Readmission_or_Need_of_Additional_Steroids_or_Antibiotics,
x = Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN_Selected_Variables_from_SES_Threshold_0.1_k_91_Categories_Coded_As_Numbers[,c("UnscheduledGP_and_AandEVisits","VAS_Breathlessness","IncreasedSputumProduction","ICS","OxygenAdministered" )],

sizes = varSeq,

method = "svmRadial",

tuneLength = 12,

rfeControl = ctrl,

metric = "ROC",

preProcess = c("center", "scale"),

fit = FALSE,

trControl = trainControl(method="cv", verboseIter = FALSE, classProbs = TRUE, number = 4, repeats = 1))

set.seed(5782)
randomforest_cv_4_folds_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN_ntree_400_all_possible_variables <- train(D30_Treatment_Outcome_Based_on_Readmission_or_Need_of_Additional_Steroids_or_Antibiotics ~ .,
data = Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN_all_possible_variables,

method = "rf",

trControl = training_control,

metric = "ROC",

ntree = 400,

tuneGrid =
```

```
RandomForestGRid_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.
8_TRAIN)

set.seed(5782)
randomforest_cv_4_folds_Imputed_81_Patients_D30_With_Hard_Treatment_Outco
me_0.8_TRAIN_ntree_1000_all_possible_variables <-
train(D30_Treatment_Outcome_Based_on_Readmission_or_Need_of_Additional_St
eroids_or_Antibiotics ~ .,

data =
Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN_all_possibl
e_variables,

method = "rf",

trControl = training_control,

metric = "ROC",

ntree = 1000,

tuneGrid =
RandomForestGRid_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.
8_TRAIN)

set.seed(5782)
randomforest_cv_4_folds_Imputed_81_Patients_D30_With_Hard_Treatment_Outco
me_0.8_TRAIN_ntree_5000_all_possible_variables <-
train(D30_Treatment_Outcome_Based_on_Readmission_or_Need_of_Additional_St
eroids_or_Antibiotics ~ .,

data =
Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.8_TRAIN_all_possibl
e_variables,

method = "rf",

trControl = training_control,
                                                                    #
metric = "ROC",

ntree = 5000,

tuneGrid =
RandomForestGRid_Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.
8_TRAIN)
```

```
test_set_D30_hard_treatment_failure_prediction <-
predict(nnet_cv_4_folds_Imputed_81_Patients_D30_With_Hard_Treatment_Outcom
e_0.8_TRAIN,
Dataframe_Imputed_81_Patients_D30_Without_Hard_Treatment_Outcome_0.2_TE
ST_Selected_Variables_from_SES_Threshold_0.1_k_90, type = "prob")
test_set_D30_hard_treatment_failure_prediction_ROC_Neural_Network <-
evalm(data.frame(test_set_D30_hard_treatment_failure_prediction,
Imputed_81_Patients_D30_With_Hard_Treatment_Outcome_0.2_TEST_Selected_V
ariables_from_SES_Threshold_0.1_k_90$D30_Treatment_Outcome_Based_on_Re
admission_or_Need_of_Additional_Steroids_or_Antibiotics, Group = "Neural
Network"))
```

```
RandomForestGRid_Imputed_81_Patients_With_D30_Hard_and_Soft_Treatment_O
utcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES <-
expand.grid(.mtry = c(1:4))
set.seed(5782)
randomforest_cv_4_folds_Imputed_81_Patients_With_D30_Hard_and_Soft_Treatme
nt_Outcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES_ntree_1
00 <- train(D30_Hard_and_Soft_Treatment_Outcome ~ .,
                                                    data =
Imputed_81_Patients_With_D30_Hard_and_Soft_Treatment_Outcomes_Exacerbatio
n_Treatment_0.8_TRAIN_5_variables_from_SES,
                                                    method = "rf",
                                                    trControl =
training_control,
                                                    metric = "ROC",
                                                    ntree = 100,
                                                    tuneGrid =
RandomForestGRid_Imputed_81_Patients_With_D30_Hard_and_Soft_Treatment_O
utcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES)

set.seed(5782)
randomforest_cv_4_folds_Imputed_81_Patients_With_D30_Hard_and_Soft_Treatme
nt_Outcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES_ntree_2
00 <- train(D30_Hard_and_Soft_Treatment_Outcome ~ .,

data =
Imputed_81_Patients_With_D30_Hard_and_Soft_Treatment_Outcomes_Exacerbatio
n_Treatment_0.8_TRAIN_5_variables_from_SES,
```

```
method = "rf",

trControl = training_control,

metric = "ROC",

ntree = 200,

tuneGrid =
RandomForestGRid_Imputed_81_Patients_With_D30_Hard_and_Soft_Treatment_O
utcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES)


set.seed(5782)
randomforest_cv_4_folds_Imputed_81_Patients_With_D30_Hard_and_Soft_Treatme
nt_Outcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES_ntree_3
00 <- train(D30_Hard_and_Soft_Treatment_Outcome ~ .,

data =
Imputed_81_Patients_With_D30_Hard_and_Soft_Treatment_Outcomes_Exacerbatio
n_Treatment_0.8_TRAIN_5_variables_from_SES,

method = "rf",

trControl = training_control,

metric = "ROC",

ntree = 300,

tuneGrid =
RandomForestGRid_Imputed_81_Patients_With_D30_Hard_and_Soft_Treatment_O
utcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES)

set.seed(5782)
randomforest_cv_4_folds_Imputed_81_Patients_With_D30_Hard_and_Soft_Treatme
nt_Outcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES_ntree_4
00 <- train(D30_Hard_and_Soft_Treatment_Outcome ~ .,

data =
Imputed_81_Patients_With_D30_Hard_and_Soft_Treatment_Outcomes_Exacerbatio
n_Treatment_0.8_TRAIN_5_variables_from_SES,

method = "rf",

trControl = training_control,

metric = "ROC",
```

```
ntree = 400,

tuneGrid =
RandomForestGRid_Imputed_81_Patients_With_D30_Hard_and_Soft_Treatment_O
utcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES)

set.seed(5782)
randomforest_cv_4_folds_Imputed_81_Patients_With_D30_Hard_and_Soft_Treatme
nt_Outcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES_ntree_5
00 <- train(D30_Hard_and_Soft_Treatment_Outcome ~ .,

data =
Imputed_81_Patients_With_D30_Hard_and_Soft_Treatment_Outcomes_Exacerbatio
n_Treatment_0.8_TRAIN_5_variables_from_SES,

method = "rf",

trControl = training_control,

metric = "ROC",

ntree = 500,

tuneGrid =
RandomForestGRid_Imputed_81_Patients_With_D30_Hard_and_Soft_Treatment_O
utcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES)

set.seed(5782)
randomforest_cv_4_folds_Imputed_81_Patients_With_D30_Hard_and_Soft_Treatme
nt_Outcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES_ntree_6
00 <- train(D30_Hard_and_Soft_Treatment_Outcome ~ .,

data =
Imputed_81_Patients_With_D30_Hard_and_Soft_Treatment_Outcomes_Exacerbatio
n_Treatment_0.8_TRAIN_5_variables_from_SES,

method = "rf",

trControl = training_control,

metric = "ROC",

ntree = 600,

tuneGrid =
RandomForestGRid_Imputed_81_Patients_With_D30_Hard_and_Soft_Treatment_O
utcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES)
```

```
set.seed(5782)
randomforest_cv_4_folds_Imputed_81_Patients_With_D30_Hard_and_Soft_Treatment_Outcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES_ntree_700 <- train(D30_Hard_and_Soft_Treatment_Outcome ~ .,

data =
Imputed_81_Patients_With_D30_Hard_and_Soft_Treatment_Outcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES,

method = "rf",

trControl = training_control,

metric = "ROC",

ntree = 700,

tuneGrid =
RandomForestGRid_Imputed_81_Patients_With_D30_Hard_and_Soft_Treatment_Outcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES)


set.seed(5782)
randomforest_cv_4_folds_Imputed_81_Patients_With_D30_Hard_and_Soft_Treatment_Outcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES_ntree_800 <- train(D30_Hard_and_Soft_Treatment_Outcome ~ .,

data =
Imputed_81_Patients_With_D30_Hard_and_Soft_Treatment_Outcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES,

method = "rf",

trControl = training_control,

metric = "ROC",

ntree = 800,

tuneGrid =
RandomForestGRid_Imputed_81_Patients_With_D30_Hard_and_Soft_Treatment_Outcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES)


set.seed(5782)
randomforest_cv_4_folds_Imputed_81_Patients_With_D30_Hard_and_Soft_Treatment_Outcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES_ntree_900 <- train(D30_Hard_and_Soft_Treatment_Outcome ~ .,

data =
```

```
Imputed_81_Patients_With_D30_Hard_and_Soft_Treatment_Outcomes_Exacerbatio
n_Treatment_0.8_TRAIN_5_variables_from_SES,

method = "rf",

trControl = training_control,

metric = "ROC",

ntree = 900,

tuneGrid =
RandomForestGRid_Imputed_81_Patients_With_D30_Hard_and_Soft_Treatment_O
utcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES)


set.seed(5782)
randomforest_cv_4_folds_Imputed_81_Patients_With_D30_Hard_and_Soft_Treatme
nt_Outcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES_ntree_1
000 <- train(D30_Hard_and_Soft_Treatment_Outcome ~ .,

data =
Imputed_81_Patients_With_D30_Hard_and_Soft_Treatment_Outcomes_Exacerbatio
n_Treatment_0.8_TRAIN_5_variables_from_SES,

method = "rf",

trControl = training_control,

metric = "ROC",

ntree = 1000,

tuneGrid =
RandomForestGRid_Imputed_81_Patients_With_D30_Hard_and_Soft_Treatment_O
utcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES)


set.seed(5782)
cart1_cv_4_folds_Imputed_81_Patients_With_D30_Hard_and_Soft_Treatment_Outc
omes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES <-
train(D30_Hard_and_Soft_Treatment_Outcome ~ .,
                                                    data =
Imputed_81_Patients_With_D30_Hard_and_Soft_Treatment_Outcomes_Exacerbatio
n_Treatment_0.8_TRAIN_5_variables_from_SES,
                                                    method = "rpart",
                                                    trControl = training_control,
                                                    metric = "ROC",
                                                    tuneLength = 10)
```

```
set.seed(5782)
knn_cv_4_folds_Imputed_81_Patients_With_D30_Hard_and_Soft_Treatment_Outcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES <-
train(D30_Hard_and_Soft_Treatment_Outcome ~ .,

data =
Dataframe_Imputed_81_Patients_With_D30_Hard_and_Soft_Treatment_Outcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES,

method = "knn",

trControl = training_control,

metric = "ROC",

tuneGrid = data.frame(k = seq(1,49,by = 2)))



set.seed(5782)
sigmaRange_Imputed_81_Patients_Without_D30_Hard_and_Soft_Treatment_Outcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES <-
sigest(Matrix_Imputed_81_Patients_Without_D30_Hard_and_Soft_Treatment_Outcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES)
set.seed(5782)
svmRGRid_Imputed_81_Patients_Without_D30_Hard_and_Soft_Treatment_Outcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES <-
expand.grid(.sigma =
seq(sigmaRange_Imputed_81_Patients_Without_D30_Hard_and_Soft_Treatment_Outcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES[1],sigmaRange_Imputed_81_Patients_Without_D30_Hard_and_Soft_Treatment_Outcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES[3],by =
((sigmaRange_Imputed_81_Patients_Without_D30_Hard_and_Soft_Treatment_Outcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES[3] -
sigmaRange_Imputed_81_Patients_Without_D30_Hard_and_Soft_Treatment_Outcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES[1])/8)), .C =
2^(seq(-4,4)))

set.seed(5782)
SVM_Radial_cv_4_folds_Imputed_81_Patients_Without_D30_Hard_and_Soft_Treatment_Outcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES <-
train(D30_Hard_and_Soft_Treatment_Outcome ~ .,
                                                    data =
Dataframe_Imputed_81_Patients_With_D30_Hard_and_Soft_Treatment_Outcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES,
                                                    method = "svmRadial",
```

```
                                                    trControl =
training_control,
                                                    metric = "ROC",
                                                    preProcess = c("center",
"scale"),
                                                    fit = FALSE,
                                                    tuneGrid =
svmRGRid_Imputed_81_Patients_Without_D30_Hard_and_Soft_Treatment_Outcom
es_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES)

set.seed(5782)
nnetRGRid_Imputed_81_Patients_Without_D30_Hard_and_Soft_Treatment_Outcom
es_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES <-
expand.grid(.size = 1:10, .decay = seq(0, 2, by = 0.1))
maxSize_Imputed_81_Patients_Without_D30_Hard_and_Soft_Treatment_Outcomes
_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES <-
max(nnetRGRid_Imputed_81_Patients_Without_D30_Hard_and_Soft_Treatment_O
utcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES$.size)
numWts_Imputed_81_Patients_Without_D30_Hard_and_Soft_Treatment_Outcomes
_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES <-
1*(maxSize_Imputed_81_Patients_Without_D30_Hard_and_Soft_Treatment_Outco
mes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES *
(ncol(Dataframe_Imputed_81_Patients_Without_D30_Hard_and_Soft_Treatment_O
utcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES) + 1) +
maxSize_Imputed_81_Patients_Without_D30_Hard_and_Soft_Treatment_Outcomes
_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES + 1)
set.seed(5782)
nnet_cv_4_folds_Imputed_81_Patients_Without_D30_Hard_and_Soft_Treatment_O
utcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES <-
train(D30_Hard_and_Soft_Treatment_Outcome ~ .,
                                                    data =
Dataframe_Imputed_81_Patients_With_D30_Hard_and_Soft_Treatment_Outcomes_
Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES,
                                                    method = "nnet",
                                                    trControl = training_control,
                                                    metric = "ROC",
                                                    preProcess = c("center",
"scale", "spatialSign"),
                                                    tuneGrid =
nnetRGRid_Imputed_81_Patients_Without_D30_Hard_and_Soft_Treatment_Outcom
es_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES,
                                                    trace = FALSE,
                                                    maxit = 2000,
                                                    MaxNWts =
numWts_Imputed_81_Patients_Without_D30_Hard_and_Soft_Treatment_Outcomes
_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES)
```

```
nnet_cv_4_folds_Imputed_81_Patients_With_D30_Hard_and_Soft_Treatment_Outc
omes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES <-
nnet_cv_4_folds_Imputed_81_Patients_Without_D30_Hard_and_Soft_Treatment_O
utcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES

Imputed_81_Patients_With_D30_Hard_and_Soft_Treatment_Outcomes_Exacerbatio
n_Treatment_0.8_TEST <-
Imputed_81_Patients_With_D30_Hard_and_Soft_Treatment_Outcomes_Exacerbatio
n_Treatment[-split_0.8_train_0.2_test_81_Patients_D30_Hard_Soft,]


test_set_D30_Hard_and_Soft_Treatment_failure_prediction <-
predict(nnet_cv_4_folds_Imputed_81_Patients_With_D30_Hard_and_Soft_Treatmen
t_Outcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES,
Dataframe_Imputed_81_Patients_Without_D30_Hard_and_Soft_Treatment_Outcom
es_Exacerbation_Treatment_0.8_TEST_5_variables_from_SES, type = "prob")
test_set_D30_Hard_and_Soft_Treatment_failure_prediction_ROC_Neural_Network
<- evalm(data.frame(test_set_D30_Hard_and_Soft_Treatment_failure_prediction,
Imputed_81_Patients_With_D30_Hard_and_Soft_Treatment_Outcomes_Exacerbatio
n_Treatment_0.8_TEST_5_variables_from_SES$D30_Hard_and_Soft_Treatment_O
utcome, Group = "Neural Network"))
test_set_D30_Hard_and_Soft_Treatment_failure_prediction_ROC_Neural_Network$
roc



Imputed_90_Patients_With_D1_D5_Hospitalisation_Data_Treatment_Outcomes_Ex
acerbation_Treatment_0.8_TRAIN <-
Imputed_90_Patients_With_Different_Treatment_Outcomes_Exacerbation_Treatme
nt_0.8_TRAIN
RandomForestGRid_Imputed_90_Patients_With_D1_D5_Hospitalisation_Data_Treat
ment_Outcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES <-
expand.grid(.mtry = c(1:4))
set.seed(5782)
randomforest_cv_4_folds_Imputed_90_Patients_With_D1_D5_Hospitalisation_Exac
erbation_Treatment_0.8_TRAIN_5_variables_from_SES_ntree_100 <-
train(D1_D5_Hard_Treatment_Outcome ~ .,

data =
Imputed_90_Patients_With_D1_D5_Hospitalisation_Data_Treatment_Outcomes_Ex
acerbation_Treatment_0.8_TRAIN_SES_5_variables,

method = "rf",

trControl = training_control,

metric = "ROC",

ntree = 100,
```

```
tuneGrid =
RandomForestGRid_Imputed_90_Patients_With_D1_D5_Hospitalisation_Data_Treat
ment_Outcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES)

set.seed(5782)
randomforest_cv_4_folds_Imputed_90_Patients_With_D1_D5_Hospitalisation_Exac
erbation_Treatment_0.8_TRAIN_5_variables_from_SES_ntree_200 <-
train(D1_D5_Hard_Treatment_Outcome ~ .,

data =
Imputed_90_Patients_With_D1_D5_Hospitalisation_Data_Treatment_Outcomes_Ex
acerbation_Treatment_0.8_TRAIN_SES_5_variables,

method = "rf",

trControl = training_control,

metric = "ROC",

ntree = 200,

tuneGrid =
RandomForestGRid_Imputed_90_Patients_With_D1_D5_Hospitalisation_Data_Treat
ment_Outcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES)


set.seed(5782)
randomforest_cv_4_folds_Imputed_90_Patients_With_D1_D5_Hospitalisation_Exac
erbation_Treatment_0.8_TRAIN_5_variables_from_SES_ntree_300 <-
train(D1_D5_Hard_Treatment_Outcome ~ .,

data =
Imputed_90_Patients_With_D1_D5_Hospitalisation_Data_Treatment_Outcomes_Ex
acerbation_Treatment_0.8_TRAIN_SES_5_variables,

method = "rf",

trControl = training_control,

metric = "ROC",

ntree = 300,

tuneGrid =
RandomForestGRid_Imputed_90_Patients_With_D1_D5_Hospitalisation_Data_Treat
ment_Outcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES)


set.seed(5782)
```

```
randomforest_cv_4_folds_Imputed_90_Patients_With_D1_D5_Hospitalisation_Exac
erbation_Treatment_0.8_TRAIN_5_variables_from_SES_ntree_400 <-
train(D1_D5_Hard_Treatment_Outcome ~ .,

data =
Imputed_90_Patients_With_D1_D5_Hospitalisation_Data_Treatment_Outcomes_Ex
acerbation_Treatment_0.8_TRAIN_SES_5_variables,

method = "rf",

trControl = training_control,

metric = "ROC",

ntree = 400,

tuneGrid =
RandomForestGRid_Imputed_90_Patients_With_D1_D5_Hospitalisation_Data_Treat
ment_Outcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES)

set.seed(5782)
randomforest_cv_4_folds_Imputed_90_Patients_With_D1_D5_Hospitalisation_Exac
erbation_Treatment_0.8_TRAIN_5_variables_from_SES_ntree_500 <-
train(D1_D5_Hard_Treatment_Outcome ~ .,

data =
Imputed_90_Patients_With_D1_D5_Hospitalisation_Data_Treatment_Outcomes_Ex
acerbation_Treatment_0.8_TRAIN_SES_5_variables,

method = "rf",

trControl = training_control,

metric = "ROC",

ntree = 500,

tuneGrid =
RandomForestGRid_Imputed_90_Patients_With_D1_D5_Hospitalisation_Data_Treat
ment_Outcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES)



set.seed(5782)
randomforest_cv_4_folds_Imputed_90_Patients_With_D1_D5_Hospitalisation_Exac
erbation_Treatment_0.8_TRAIN_5_variables_from_SES_ntree_600 <-
train(D1_D5_Hard_Treatment_Outcome ~ .,

data =
```

```
Imputed_90_Patients_With_D1_D5_Hospitalisation_Data_Treatment_Outcomes_Exacerbation_Treatment_0.8_TRAIN_SES_5_variables,

method = "rf",

trControl = training_control,

metric = "ROC",

ntree = 600,

tuneGrid = RandomForestGRid_Imputed_90_Patients_With_D1_D5_Hospitalisation_Data_Treatment_Outcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES)


set.seed(5782)
randomforest_cv_4_folds_Imputed_90_Patients_With_D1_D5_Hospitalisation_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES_ntree_700 <- train(D1_D5_Hard_Treatment_Outcome ~ .,

data = Imputed_90_Patients_With_D1_D5_Hospitalisation_Data_Treatment_Outcomes_Exacerbation_Treatment_0.8_TRAIN_SES_5_variables,

method = "rf",

trControl = training_control,

metric = "ROC",

ntree = 700,

tuneGrid = RandomForestGRid_Imputed_90_Patients_With_D1_D5_Hospitalisation_Data_Treatment_Outcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES)


set.seed(5782)
randomforest_cv_4_folds_Imputed_90_Patients_With_D1_D5_Hospitalisation_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES_ntree_800 <- train(D1_D5_Hard_Treatment_Outcome ~ .,

data = Imputed_90_Patients_With_D1_D5_Hospitalisation_Data_Treatment_Outcomes_Exacerbation_Treatment_0.8_TRAIN_SES_5_variables,

method = "rf",

trControl = training_control,
```

```
metric = "ROC",

ntree = 800,

tuneGrid =
RandomForestGRid_Imputed_90_Patients_With_D1_D5_Hospitalisation_Data_Treat
ment_Outcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES)


set.seed(5782)
randomforest_cv_4_folds_Imputed_90_Patients_With_D1_D5_Hospitalisation_Exac
erbation_Treatment_0.8_TRAIN_5_variables_from_SES_ntree_900 <-
train(D1_D5_Hard_Treatment_Outcome ~ .,

data =
Imputed_90_Patients_With_D1_D5_Hospitalisation_Data_Treatment_Outcomes_Ex
acerbation_Treatment_0.8_TRAIN_SES_5_variables,

method = "rf",

trControl = training_control,

metric = "ROC",

ntree = 900,

tuneGrid =
RandomForestGRid_Imputed_90_Patients_With_D1_D5_Hospitalisation_Data_Treat
ment_Outcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES)


set.seed(5782)
randomforest_cv_4_folds_Imputed_90_Patients_With_D1_D5_Hospitalisation_Exac
erbation_Treatment_0.8_TRAIN_5_variables_from_SES_ntree_1000 <-
train(D1_D5_Hard_Treatment_Outcome ~ .,

data =
Imputed_90_Patients_With_D1_D5_Hospitalisation_Data_Treatment_Outcomes_Ex
acerbation_Treatment_0.8_TRAIN_SES_5_variables,

method = "rf",

trControl = training_control,

metric = "ROC",

ntree = 1000,

tuneGrid =
```

```
RandomForestGRid_Imputed_90_Patients_With_D1_D5_Hospitalisation_Data_Treat
ment_Outcomes_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES)


set.seed(5782)
cart1_cv_4_folds_Imputed_90_Patients_With_D1_D5_Hospitalisation_Exacerbation
_Treatment_0.8_TRAIN_5_variables_from_SES <-
train(D1_D5_Hard_Treatment_Outcome ~ .,

data =
Imputed_90_Patients_With_D1_D5_Hospitalisation_Data_Treatment_Outcomes_Ex
acerbation_Treatment_0.8_TRAIN_SES_5_variables,

method = "rpart",

trControl = training_control,

metric = "ROC",

tuneLength = 10)

set.seed(5782)
knn_cv_4_folds_Imputed_90_Patients_With_D1_D5_Hospitalisation_Exacerbation_
Treatment_0.8_TRAIN_5_variables_from_SES <-
train(D1_D5_Hard_Treatment_Outcome ~ .,

data =
Dataframe_Imputed_90_Patients_With_D1_D5_Hospitalisation_Data_Treatment_Ou
tcomes_Exacerbation_Treatment_0.8_TRAIN_SES_5_variables,

method = "knn",

trControl = training_control,

metric = "ROC",

tuneGrid = data.frame(k = seq(1,53,by = 2)))


set.seed(5782)
sigmaRange_Imputed_90_Patients_With_D1_D5_Hospitalisation_Exacerbation_Tre
atment_0.8_TRAIN_5_variables_from_SES <-
sigest(Matrix_Imputed_90_Patients_Without_D1_D5_Hospitalisation_Data_Treatme
nt_Outcomes_Exacerbation_Treatment_0.8_TRAIN_SES_5_variables)
set.seed(5782)
svmRGRid_Imputed_90_Patients_With_D1_D5_Hospitalisation_Exacerbation_Treat
ment_0.8_TRAIN_5_variables_from_SES <- expand.grid(.sigma =
seq(sigmaRange_Imputed_90_Patients_With_D1_D5_Hospitalisation_Exacerbation
_Treatment_0.8_TRAIN_5_variables_from_SES[1],sigmaRange_Imputed_90_Patien
ts_With_D1_D5_Hospitalisation_Exacerbation_Treatment_0.8_TRAIN_5_variables_f
```

```
rom_SES[3],by =
((sigmaRange_Imputed_90_Patients_With_D1_D5_Hospitalisation_Exacerbation_Tr
eatment_0.8_TRAIN_5_variables_from_SES[3] -
sigmaRange_Imputed_90_Patients_With_D1_D5_Hospitalisation_Exacerbation_Tre
atment_0.8_TRAIN_5_variables_from_SES[1])/8)), .C = 2^(seq(-4,4)))

set.seed(5782)
SVM_Radial_cv_4_folds_Imputed_90_Patients_With_D1_D5_Hospitalisation_Exace
rbation_Treatment_0.8_TRAIN_5_variables_from_SES <-
train(D1_D5_Hard_Treatment_Outcome ~ .,

data =
Dataframe_Imputed_90_Patients_With_D1_D5_Hospitalisation_Data_Treatment_Ou
tcomes_Exacerbation_Treatment_0.8_TRAIN_SES_5_variables,

method = "svmRadial",

trControl = training_control,

metric = "ROC",

preProcess = c("center", "scale"),

fit = FALSE,

tuneGrid =
svmRGRid_Imputed_90_Patients_With_D1_D5_Hospitalisation_Exacerbation_Treat
ment_0.8_TRAIN_5_variables_from_SES)


set.seed(5782)
nnetRGRid_Imputed_90_Patients_With_D1_D5_Hospitalisation_Exacerbation_Treat
ment_0.8_TRAIN_5_variables_from_SES <- expand.grid(.size = 1:10, .decay =
seq(0, 2, by = 0.1))
maxSize_Imputed_90_Patients_With_D1_D5_Hospitalisation_Exacerbation_Treatm
ent_0.8_TRAIN_5_variables_from_SES <-
max(nnetRGRid_Imputed_90_Patients_With_D1_D5_Hospitalisation_Exacerbation_
Treatment_0.8_TRAIN_5_variables_from_SES$.size)
numWts_Imputed_90_Patients_With_D1_D5_Hospitalisation_Exacerbation_Treatme
nt_0.8_TRAIN_5_variables_from_SES <-
1*(maxSize_Imputed_90_Patients_With_D1_D5_Hospitalisation_Exacerbation_Treat
ment_0.8_TRAIN_5_variables_from_SES *
(ncol(Dataframe_Imputed_90_Patients_Without_D1_D5_Hospitalisation_Data_Treat
ment_Outcomes_Exacerbation_Treatment_0.8_TRAIN_SES_5_variables) + 1) +
maxSize_Imputed_90_Patients_With_D1_D5_Hospitalisation_Exacerbation_Treatm
ent_0.8_TRAIN_5_variables_from_SES + 1)
set.seed(5782)
nnet_cv_4_folds_Imputed_90_Patients_With_D1_D5_Hospitalisation_Exacerbation_
Treatment_0.8_TRAIN_5_variables_from_SES <-
train(D1_D5_Hard_Treatment_Outcome ~ .,
```

```
data =
Dataframe_Imputed_90_Patients_With_D1_D5_Hospitalisation_Data_Treatment_Outcomes_Exacerbation_Treatment_0.8_TRAIN_SES_5_variables,

method = "nnet",

trControl = training_control,

metric = "ROC",

preProcess = c("center", "scale", "spatialSign"),

tuneGrid =
nnetRGRid_Imputed_90_Patients_With_D1_D5_Hospitalisation_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES,

trace = FALSE,

maxit = 2000,

MaxNWts =
numWts_Imputed_90_Patients_With_D1_D5_Hospitalisation_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES)


Imputed_90_Patients_With_D1_D5_Hospitalisation_Exacerbation_Treatment_0.2_TEST <-
Imputed_90_Patients_With_D1_D5_Hospitalisation_Exacerbation_Treatment[-
split_0.8_train_0.2_test_90_Patients_D1_D5_Hard,]

test_set_D1_D5_Hospitalisation_prediction <-
predict(nnet_cv_4_folds_Imputed_90_Patients_With_D1_D5_Hospitalisation_Exacerbation_Treatment_0.8_TRAIN_5_variables_from_SES,
Dataframe_Imputed_90_Patients_Without_D1_D5_Hospitalisation_Outcome_Exacerbation_Treatment_0.2_TEST_SES_5_variables, type = "prob")
test_set_D1_D5_Hospitalisation_prediction_ROC_Neural_Network <-
evalm(data.frame(test_set_D1_D5_Hospitalisation_prediction,
Imputed_90_Patients_With_D1_D5_Hospitalisation_Exacerbation_Treatment$D1_D5_Hard_Treatment_Outcome, Group = "Neural Network"))
test_set_D1_D5_Hospitalisation_prediction_ROC_Neural_Network$roc



RandomForestGRid_Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_Exacerbation_Treatment_0.8_TRAIN_SES_4_variables <- expand.grid(.mtry =
c(1:3))
set.seed(5782)
```

```
randomforest_cv_4_folds_Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_Exacerbation_Treatment_0.8_TRAIN_SES_4_variables_ntree_100 <-
train(D90_Hard_Outcome ~ .,

data =
Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_Exacerbation_Treatment_0.8_TRAIN_SES_4_variables,

method = "rf",

trControl = training_control,

metric = "ROC",

ntree = 100,

tuneGrid =
RandomForestGRid_Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_Exacerbation_Treatment_0.8_TRAIN_SES_4_variables)

set.seed(5782)
randomforest_cv_4_folds_Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_Exacerbation_Treatment_0.8_TRAIN_SES_4_variables_ntree_200 <-
train(D90_Hard_Outcome ~ .,

data =
Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_Exacerbation_Treatment_0.8_TRAIN_SES_4_variables,

method = "rf",

trControl = training_control,

metric = "ROC",

ntree = 200,

tuneGrid =
RandomForestGRid_Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_Exacerbation_Treatment_0.8_TRAIN_SES_4_variables)

set.seed(5782)
randomforest_cv_4_folds_Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_Exacerbation_Treatment_0.8_TRAIN_SES_4_variables_ntree_300 <-
train(D90_Hard_Outcome ~ .,

data =
Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_Exacerbation_Treatment_0.8_TRAIN_SES_4_variables,
```

```
method = "rf",

trControl = training_control,

metric = "ROC",

ntree = 300,

tuneGrid =
RandomForestGRid_Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_Exacerbation_Treatment_0.8_TRAIN_SES_4_variables)

set.seed(5782)
randomforest_cv_4_folds_Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_Exacerbation_Treatment_0.8_TRAIN_SES_4_variables_ntree_400 <-
train(D90_Hard_Outcome ~ .,

data =
Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_Exacerbation_Treatment_0.8_TRAIN_SES_4_variables,

method = "rf",

trControl = training_control,

metric = "ROC",

ntree = 400,

tuneGrid =
RandomForestGRid_Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_Exacerbation_Treatment_0.8_TRAIN_SES_4_variables)

set.seed(5782)
randomforest_cv_4_folds_Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_Exacerbation_Treatment_0.8_TRAIN_SES_4_variables_ntree_500 <-
train(D90_Hard_Outcome ~ .,

data =
Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_Exacerbation_Treatment_0.8_TRAIN_SES_4_variables,

method = "rf",

trControl = training_control,

metric = "ROC",

ntree = 500,
```

```
tuneGrid =
RandomForestGRid_Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_w
ith_Exacerbation_Treatment_0.8_TRAIN_SES_4_variables)

set.seed(5782)
randomforest_cv_4_folds_Imputed_85_Patients_With_D90_Hard_Treatment_Outco
me_with_Exacerbation_Treatment_0.8_TRAIN_SES_4_variables_ntree_600 <-
train(D90_Hard_Outcome ~ .,

data =
Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_Exacerbation_Tre
atment_0.8_TRAIN_SES_4_variables,

method = "rf",

trControl = training_control,

metric = "ROC",

ntree = 600,

tuneGrid =
RandomForestGRid_Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_w
ith_Exacerbation_Treatment_0.8_TRAIN_SES_4_variables)


set.seed(5782)
randomforest_cv_4_folds_Imputed_85_Patients_With_D90_Hard_Treatment_Outco
me_with_Exacerbation_Treatment_0.8_TRAIN_SES_4_variables_ntree_700 <-
train(D90_Hard_Outcome ~ .,

data =
Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_Exacerbation_Tre
atment_0.8_TRAIN_SES_4_variables,

method = "rf",

trControl = training_control,

metric = "ROC",

ntree = 700,

tuneGrid =
RandomForestGRid_Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_w
ith_Exacerbation_Treatment_0.8_TRAIN_SES_4_variables)


set.seed(5782)
```

```
randomforest_cv_4_folds_Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_Exacerbation_Treatment_0.8_TRAIN_SES_4_variables_ntree_800 <-
train(D90_Hard_Outcome ~ .,

data =
Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_Exacerbation_Treatment_0.8_TRAIN_SES_4_variables,

method = "rf",

trControl = training_control,

metric = "ROC",

ntree = 800,

tuneGrid =
RandomForestGRid_Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_Exacerbation_Treatment_0.8_TRAIN_SES_4_variables)


set.seed(5782)
randomforest_cv_4_folds_Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_Exacerbation_Treatment_0.8_TRAIN_SES_4_variables_ntree_900 <-
train(D90_Hard_Outcome ~ .,

data =
Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_Exacerbation_Treatment_0.8_TRAIN_SES_4_variables,

method = "rf",

trControl = training_control,

metric = "ROC",

ntree = 900,

tuneGrid =
RandomForestGRid_Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_Exacerbation_Treatment_0.8_TRAIN_SES_4_variables)


set.seed(5782)
randomforest_cv_4_folds_Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_Exacerbation_Treatment_0.8_TRAIN_SES_4_variables_ntree_1000 <-
train(D90_Hard_Outcome ~ .,

data =
```

```
Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_Exacerbation_Tre
atment_0.8_TRAIN_SES_4_variables,

method = "rf",

trControl = training_control,

metric = "ROC",

ntree = 1000,

tuneGrid =
RandomForestGRid_Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_w
ith_Exacerbation_Treatment_0.8_TRAIN_SES_4_variables)

set.seed(5782)
cart1_cv_4_folds_Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with
_Exacerbation_Treatment_0.8_TRAIN_SES_4_variables <-
train(D90_Hard_Outcome ~ .,
                                                                    data
=
Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_Exacerbation_Tre
atment_0.8_TRAIN_SES_4_variables,

method = "rpart",

trControl = training_control,

metric = "ROC",

tuneLength = 10)

set.seed(5782)
knn_cv_4_folds_Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_
Exacerbation_Treatment_0.8_TRAIN_SES_4_variables <- train(D90_Hard_Outcome
~ .,
                                                                    data =
Dataframe_Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_Exace
rbation_Treatment_0.8_TRAIN_SES_4_variables,

method = "knn",

trControl = training_control,
                                                                    metric
= "ROC",

tuneGrid = data.frame(k = seq(1,51,by = 2)))


set.seed(5782)
```

```
sigmaRange_Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_Exa
cerbation_Treatment_0.8_TRAIN_SES_4_variables <-
sigest(Matrix_Imputed_85_Patients_Without_D90_Hard_Treatment_Outcome_with_
Exacerbation_Treatment_0.8_TRAIN_SES_4_variables)
set.seed(5782)
svmRGRid_Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_Exac
erbation_Treatment_0.8_TRAIN_SES_4_variables <- expand.grid(.sigma =
seq(sigmaRange_Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with
_Exacerbation_Treatment_0.8_TRAIN_SES_4_variables[1],sigmaRange_Imputed_8
5_Patients_With_D90_Hard_Treatment_Outcome_with_Exacerbation_Treatment_0.
8_TRAIN_SES_4_variables[3],by =
((sigmaRange_Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_Ex
acerbation_Treatment_0.8_TRAIN_SES_4_variables[3] -
sigmaRange_Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_Exa
cerbation_Treatment_0.8_TRAIN_SES_4_variables[1])/8)), .C = 2^(seq(-4,4)))

set.seed(5782)
SVM_Radial_cv_4_folds_Imputed_85_Patients_With_D90_Hard_Treatment_Outcom
e_with_Exacerbation_Treatment_0.8_TRAIN_SES_4_variables <-
train(D90_Hard_Outcome ~ .,

data =
Dataframe_Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_Exace
rbation_Treatment_0.8_TRAIN_SES_4_variables,

method = "svmRadial",

trControl = training_control,

metric = "ROC",

preProcess = c("center", "scale"),
                                                                    fit
= FALSE,

tuneGrid =
svmRGRid_Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_Exac
erbation_Treatment_0.8_TRAIN_SES_4_variables)



set.seed(5782)
nnetRGRid_Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_Exac
erbation_Treatment_0.8_TRAIN_SES_4_variables <- expand.grid(.size = 1:10,
.decay = seq(0, 2, by = 0.1))
maxSize_Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_Exacer
bation_Treatment_0.8_TRAIN_SES_4_variables <-
max(nnetRGRid_Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_
Exacerbation_Treatment_0.8_TRAIN_SES_4_variables$.size)
```

```
numWts_Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_Exacerb
ation_Treatment_0.8_TRAIN_SES_4_variables <-
1*(maxSize_Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_Exac
erbation_Treatment_0.8_TRAIN_SES_4_variables *
(ncol(Dataframe_Imputed_85_Patients_Without_D90_Hard_Treatment_Outcome_wit
h_Exacerbation_Treatment_0.8_TRAIN_SES_4_variables) + 1) +
maxSize_Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_Exacer
bation_Treatment_0.8_TRAIN_SES_4_variables + 1)
set.seed(5782)
nnet_cv_4_folds_Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_
Exacerbation_Treatment_0.8_TRAIN_SES_4_variables <- train(D90_Hard_Outcome
~ .,
                                                          data
=
Dataframe_Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_Exace
rbation_Treatment_0.8_TRAIN_SES_4_variables,

method = "nnet",

trControl = training_control,
                                                          metric
= "ROC",

preProcess = c("center", "scale", "spatialSign"),

tuneGrid =
nnetRGRid_Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_Exac
erbation_Treatment_0.8_TRAIN_SES_4_variables,
                                                          trace
= FALSE,
                                                          maxit
= 2000,

MaxNWts =
numWts_Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_Exacerb
ation_Treatment_0.8_TRAIN_SES_4_variables)


test_set_Imputed_85_Patients_With_D90_Hard_Treatment_failure_prediction <-
predict(randomforest_cv_4_folds_Imputed_85_Patients_With_D90_Hard_Treatment
_Outcome_with_Exacerbation_Treatment_0.8_TRAIN_SES_4_variables_ntree_700,
Imputed_85_Patients_Without_D90_Hard_Treatment_Outcome_with_Exacerbation_
Treatment_0.2_TEST_SES_4_variables, type = "prob")
test_set_Imputed_85_Patients_With_D90_Hard_Treatment_failure_prediction_ROC
_Random_Forest <-
evalm(data.frame(test_set_Imputed_85_Patients_With_D90_Hard_Treatment_failure
_prediction,
Imputed_85_Patients_With_D90_Hard_Treatment_Outcome_with_Exacerbation_Tre
atment_0.2_TEST_SES_4_variables$D90_Hard_Outcome, Group = "Random
Forest"))
```

```
test_set_Imputed_85_Patients_With_D90_Hard_Treatment_failure_prediction_ROC
_Random_Forest$roc
```

```
RandomForestGRid_Imputed_101_Patients_Antibiotic_Prescription_Prediction_0.8_
TRAIN_SES_5_variables <- expand.grid(.mtry = c(1:4))
set.seed(5782)
randomforest_cv_4_folds_Imputed_101_Patients_Antibiotic_Prescription_Prediction
_0.8_TRAIN_SES_5_variables_ntree_100 <- train(AntibioticsToday ~ .,

data =
Imputed_101_Patients_Antibiotic_Prescription_Prediction_Treatment_Revealed_0.8
_TRAIN_SES_5_variables,

method = "rf",

trControl = training_control,

metric = "ROC",

ntree = 100,

tuneGrid =
RandomForestGRid_Imputed_101_Patients_Antibiotic_Prescription_Prediction_0.8_
TRAIN_SES_5_variables)

set.seed(5782)
randomforest_cv_4_folds_Imputed_101_Patients_Antibiotic_Prescription_Prediction
_0.8_TRAIN_SES_5_variables_ntree_200 <- train(AntibioticsToday ~ .,
                                                                    data =
Imputed_101_Patients_Antibiotic_Prescription_Prediction_Treatment_Revealed_0.8
_TRAIN_SES_5_variables,

method = "rf",

trControl = training_control,
                                                                    metric
= "ROC",
                                                                    ntree
= 200,

tuneGrid =
RandomForestGRid_Imputed_101_Patients_Antibiotic_Prescription_Prediction_0.8_
TRAIN_SES_5_variables)
```

```
set.seed(5782)
randomforest_cv_4_folds_Imputed_101_Patients_Antibiotic_Prescription_Prediction
_0.8_TRAIN_SES_5_variables_ntree_300 <- train(AntibioticsToday ~ .,
                                                                    data =
Imputed_101_Patients_Antibiotic_Prescription_Prediction_Treatment_Revealed_0.8
_TRAIN_SES_5_variables,

method = "rf",

trControl = training_control,
                                                                    metric
= "ROC",
                                                                    ntree
= 300,

tuneGrid =
RandomForestGRid_Imputed_101_Patients_Antibiotic_Prescription_Prediction_0.8_
TRAIN_SES_5_variables)

set.seed(5782)
randomforest_cv_4_folds_Imputed_101_Patients_Antibiotic_Prescription_Prediction
_0.8_TRAIN_SES_5_variables_ntree_400 <- train(AntibioticsToday ~ .,
                                                                    data =
Imputed_101_Patients_Antibiotic_Prescription_Prediction_Treatment_Revealed_0.8
_TRAIN_SES_5_variables,

method = "rf",

trControl = training_control,
                                                                    metric
= "ROC",
                                                                    ntree
= 400,

tuneGrid =
RandomForestGRid_Imputed_101_Patients_Antibiotic_Prescription_Prediction_0.8_
TRAIN_SES_5_variables)


set.seed(5782)
randomforest_cv_4_folds_Imputed_101_Patients_Antibiotic_Prescription_Prediction
_0.8_TRAIN_SES_5_variables_ntree_500 <- train(AntibioticsToday ~ .,
                                                                    data =
Imputed_101_Patients_Antibiotic_Prescription_Prediction_Treatment_Revealed_0.8
_TRAIN_SES_5_variables,

method = "rf",
```

```
trControl = training_control,
                                                                    metric
= "ROC",
                                                                    ntree
= 500,

tuneGrid =
RandomForestGRid_Imputed_101_Patients_Antibiotic_Prescription_Prediction_0.8_
TRAIN_SES_5_variables)

set.seed(5782)
randomforest_cv_4_folds_Imputed_101_Patients_Antibiotic_Prescription_Prediction
_0.8_TRAIN_SES_5_variables_ntree_600 <- train(AntibioticsToday ~ .,
                                                                    data =
Imputed_101_Patients_Antibiotic_Prescription_Prediction_Treatment_Revealed_0.8
_TRAIN_SES_5_variables,

method = "rf",

trControl = training_control,
                                                                    metric
= "ROC",
                                                                    ntree
= 600,

tuneGrid =
RandomForestGRid_Imputed_101_Patients_Antibiotic_Prescription_Prediction_0.8_
TRAIN_SES_5_variables)




set.seed(5782)
randomforest_cv_4_folds_Imputed_101_Patients_Antibiotic_Prescription_Prediction
_0.8_TRAIN_SES_5_variables_ntree_700 <- train(AntibioticsToday ~ .,
                                                                    data =
Imputed_101_Patients_Antibiotic_Prescription_Prediction_Treatment_Revealed_0.8
_TRAIN_SES_5_variables,

method = "rf",

trControl = training_control,
                                                                    metric
= "ROC",
                                                                    ntree
= 700,

tuneGrid =
```

```
RandomForestGRid_Imputed_101_Patients_Antibiotic_Prescription_Prediction_0.8_
TRAIN_SES_5_variables)


set.seed(5782)
randomforest_cv_4_folds_Imputed_101_Patients_Antibiotic_Prescription_Prediction
_0.8_TRAIN_SES_5_variables_ntree_800 <- train(AntibioticsToday ~ .,
                                                                    data =
Imputed_101_Patients_Antibiotic_Prescription_Prediction_Treatment_Revealed_0.8
_TRAIN_SES_5_variables,

method = "rf",

trControl = training_control,
                                                                    metric
= "ROC",
                                                                    ntree
= 800,

tuneGrid =
RandomForestGRid_Imputed_101_Patients_Antibiotic_Prescription_Prediction_0.8_
TRAIN_SES_5_variables)


set.seed(5782)
randomforest_cv_4_folds_Imputed_101_Patients_Antibiotic_Prescription_Prediction
_0.8_TRAIN_SES_5_variables_ntree_900 <- train(AntibioticsToday ~ .,
                                                                    data =
Imputed_101_Patients_Antibiotic_Prescription_Prediction_Treatment_Revealed_0.8
_TRAIN_SES_5_variables,

method = "rf",

trControl = training_control,
                                                                    metric
= "ROC",
                                                                    ntree
= 900,

tuneGrid =
RandomForestGRid_Imputed_101_Patients_Antibiotic_Prescription_Prediction_0.8_
TRAIN_SES_5_variables)


set.seed(5782)
randomforest_cv_4_folds_Imputed_101_Patients_Antibiotic_Prescription_Prediction
_0.8_TRAIN_SES_5_variables_ntree_1000 <- train(AntibioticsToday ~ .,
                                                                    data =
Imputed_101_Patients_Antibiotic_Prescription_Prediction_Treatment_Revealed_0.8
_TRAIN_SES_5_variables,
```

```
method = "rf",

trControl = training_control,
                                                                    metric
= "ROC",
                                                                    ntree
= 1000,

tuneGrid =
RandomForestGRid_Imputed_101_Patients_Antibiotic_Prescription_Prediction_0.8_
TRAIN_SES_5_variables)


set.seed(5782)
cart1_cv_4_folds_Imputed_101_Patients_Antibiotic_Prescription_Prediction_0.8_TR
AIN_SES_5_variables <- train(AntibioticsToday ~ .,

data =
Imputed_101_Patients_Antibiotic_Prescription_Prediction_Treatment_Revealed_0.8
_TRAIN_SES_5_variables,

method = "rpart",

trControl = training_control,

metric = "ROC",

tuneLength = 10)


set.seed(5782)
knn_cv_4_folds_Imputed_101_Patients_Antibiotic_Prescription_Prediction_0.8_TRAI
N_SES_5_variables <- train(AntibioticsToday ~ .,

data =
Dataframe_Imputed_101_Patients_Antibiotic_Prescription_Prediction_Treatment_Re
vealed_0.8_TRAIN_SES_5_variables,

method = "knn",

trControl = training_control,

metric = "ROC",

tuneGrid = data.frame(k = seq(1,59,by = 2)))


set.seed(5782)
```

```
sigmaRange_Imputed_101_Patients_Antibiotic_Prescription_Prediction_0.8_TRAIN_
SES_5_variables <-
sigest(Matrix_Imputed_101_Patients_Antibiotic_Prescription_Prediction_Treatment_
Not_Revealed_0.8_TRAIN_SES_5_variables)
set.seed(5782)
svmRGRid_Imputed_101_Patients_Antibiotic_Prescription_Prediction_0.8_TRAIN_S
ES_5_variables <- expand.grid(.sigma =
seq(sigmaRange_Imputed_101_Patients_Antibiotic_Prescription_Prediction_0.8_TR
AIN_SES_5_variables[1],sigmaRange_Imputed_101_Patients_Antibiotic_Prescriptio
n_Prediction_0.8_TRAIN_SES_5_variables[3],by =
((sigmaRange_Imputed_101_Patients_Antibiotic_Prescription_Prediction_0.8_TRAI
N_SES_5_variables[3] -
sigmaRange_Imputed_101_Patients_Antibiotic_Prescription_Prediction_0.8_TRAIN_
SES_5_variables[1])/8)), .C = 2^(seq(-4,4)))

set.seed(5782)
SVM_Radial_cv_4_folds_Imputed_101_Patients_Antibiotic_Prescription_Prediction_
0.8_TRAIN_SES_5_variables <- train(AntibioticsToday ~ .,

data =
Dataframe_Imputed_101_Patients_Antibiotic_Prescription_Prediction_Treatment_Re
vealed_0.8_TRAIN_SES_5_variables,

method = "svmRadial",

trControl = training_control,

metric = "ROC",

preProcess = c("center", "scale"),

fit = FALSE,

tuneGrid =
svmRGRid_Imputed_101_Patients_Antibiotic_Prescription_Prediction_0.8_TRAIN_S
ES_5_variables)


set.seed(5782)
nnetRGRid_Imputed_101_Patients_Antibiotic_Prescription_Prediction_0.8_TRAIN_S
ES_5_variables <- expand.grid(.size = 1:10, .decay = seq(0, 2, by = 0.1))
maxSize_Imputed_101_Patients_Antibiotic_Prescription_Prediction_0.8_TRAIN_SE
S_5_variables <-
max(nnetRGRid_Imputed_101_Patients_Antibiotic_Prescription_Prediction_0.8_TRA
IN_SES_5_variables$.size)
numWts_Imputed_101_Patients_Antibiotic_Prescription_Prediction_0.8_TRAIN_SES
_5_variables <-
1*(maxSize_Imputed_101_Patients_Antibiotic_Prescription_Prediction_0.8_TRAIN_
SES_5_variables *
(ncol(Dataframe_Imputed_101_Patients_Antibiotic_Prescription_Prediction_Treatme
```

```
nt_Not_Revealed_0.8_TRAIN_SES_5_variables) + 1) +
maxSize_Imputed_101_Patients_Antibiotic_Prescription_Prediction_0.8_TRAIN_SE
S_5_variables + 1)
set.seed(5782)
nnet_cv_4_folds_Imputed_101_Patients_Antibiotic_Prescription_Prediction_0.8_TR
AIN_SES_5_variables <- train(AntibioticsToday ~ .,

data =
Dataframe_Imputed_101_Patients_Antibiotic_Prescription_Prediction_Treatment_Re
vealed_0.8_TRAIN_SES_5_variables,

method = "nnet",

trControl = training_control,

metric = "ROC",

preProcess = c("center", "scale", "spatialSign"),

tuneGrid =
nnetRGRid_Imputed_101_Patients_Antibiotic_Prescription_Prediction_0.8_TRAIN_S
ES_5_variables,

trace = FALSE,

maxit = 2000,

MaxNWts =
numWts_Imputed_101_Patients_Antibiotic_Prescription_Prediction_0.8_TRAIN_SES
_5_variables)


test_set_Imputed_101_Patients_Antibiotic_Prescription_Prediction_0.8_TRAIN_SES
_5_variables <-
predict(SVM_Radial_cv_4_folds_Imputed_101_Patients_Antibiotic_Prescription_Pre
diction_0.8_TRAIN_SES_5_variables,
Dataframe_Imputed_101_Patients_Antibiotic_Prescription_Prediction_Treatment_No
t_Revealed_0.8_TRAIN_SES_5_variables, type = "prob")
test_set_Imputed_101_Patients_With_Antibiotic_Prescription_Prediction_ROC_SVM
<-
evalm(data.frame(test_set_Imputed_101_Patients_Antibiotic_Prescription_Prediction
_0.8_TRAIN_SES_5_variables,
Dataframe_Imputed_101_Patients_Antibiotic_Prescription_Prediction_Treatment_Re
vealed_0.8_TRAIN_SES_5_variables$AntibioticsToday, Group = "Support Vector
Machine"))
test_set_Imputed_101_Patients_With_Antibiotic_Prescription_Prediction_ROC_SVM
$roc
```

```
RandomForestGRid_Imputed_101_Patients_SCS_Prescription_Prediction_Treatmen
t_Not_Revealed_0.8_TRAIN_SES_4_Variables <- expand.grid(.mtry = c(1:3))
set.seed(5782)
randomforest_cv_4_folds_Imputed_101_Patients_SCS_Prescription_Prediction_Tre
atment_Not_Revealed_0.8_TRAIN_SES_4_Variables_ntree_100 <-
train(OralPrednisolone ~ .,
                                                                    data =
Imputed_101_Patients_SCS_Prescription_Prediction_Treatment_Revealed_0.8_TR
AIN_SES_4_Variables,

method = "rf",

trControl = training_control,
                                                                    metric
= "ROC",
                                                                    ntree
= 100,

tuneGrid =
RandomForestGRid_Imputed_101_Patients_SCS_Prescription_Prediction_Treatmen
t_Not_Revealed_0.8_TRAIN_SES_4_Variables)


set.seed(5782)
randomforest_cv_4_folds_Imputed_101_Patients_SCS_Prescription_Prediction_Tre
atment_Not_Revealed_0.8_TRAIN_SES_4_Variables_ntree_200 <-
train(OralPrednisolone ~ .,

data =
Imputed_101_Patients_SCS_Prescription_Prediction_Treatment_Revealed_0.8_TR
AIN_SES_4_Variables,

method = "rf",

trControl = training_control,

metric = "ROC",

ntree = 200,

tuneGrid =
RandomForestGRid_Imputed_101_Patients_SCS_Prescription_Prediction_Treatmen
t_Not_Revealed_0.8_TRAIN_SES_4_Variables)

set.seed(5782)
```

```
randomforest_cv_4_folds_Imputed_101_Patients_SCS_Prescription_Prediction_Tre
atment_Not_Revealed_0.8_TRAIN_SES_4_Variables_ntree_300 <-
train(OralPrednisolone ~ .,

data =
Imputed_101_Patients_SCS_Prescription_Prediction_Treatment_Revealed_0.8_TR
AIN_SES_4_Variables,

method = "rf",

trControl = training_control,

metric = "ROC",

ntree = 300,

tuneGrid =
RandomForestGRid_Imputed_101_Patients_SCS_Prescription_Prediction_Treatmen
t_Not_Revealed_0.8_TRAIN_SES_4_Variables)

set.seed(5782)
randomforest_cv_4_folds_Imputed_101_Patients_SCS_Prescription_Prediction_Tre
atment_Not_Revealed_0.8_TRAIN_SES_4_Variables_ntree_400 <-
train(OralPrednisolone ~ .,

data =
Imputed_101_Patients_SCS_Prescription_Prediction_Treatment_Revealed_0.8_TR
AIN_SES_4_Variables,

method = "rf",

trControl = training_control,

metric = "ROC",

ntree = 400,

tuneGrid =
RandomForestGRid_Imputed_101_Patients_SCS_Prescription_Prediction_Treatmen
t_Not_Revealed_0.8_TRAIN_SES_4_Variables)


set.seed(5782)
randomforest_cv_4_folds_Imputed_101_Patients_SCS_Prescription_Prediction_Tre
atment_Not_Revealed_0.8_TRAIN_SES_4_Variables_ntree_500 <-
train(OralPrednisolone ~ .,

data =
Imputed_101_Patients_SCS_Prescription_Prediction_Treatment_Revealed_0.8_TR
AIN_SES_4_Variables,
```

```
method = "rf",

trControl = training_control,

metric = "ROC",

ntree = 500,

tuneGrid =
RandomForestGRid_Imputed_101_Patients_SCS_Prescription_Prediction_Treatmen
t_Not_Revealed_0.8_TRAIN_SES_4_Variables)


set.seed(5782)
randomforest_cv_4_folds_Imputed_101_Patients_SCS_Prescription_Prediction_Tre
atment_Not_Revealed_0.8_TRAIN_SES_4_Variables_ntree_600 <-
train(OralPrednisolone ~ .,

data =
Imputed_101_Patients_SCS_Prescription_Prediction_Treatment_Revealed_0.8_TR
AIN_SES_4_Variables,

method = "rf",

trControl = training_control,

metric = "ROC",

ntree = 600,

tuneGrid =
RandomForestGRid_Imputed_101_Patients_SCS_Prescription_Prediction_Treatmen
t_Not_Revealed_0.8_TRAIN_SES_4_Variables)



set.seed(5782)
randomforest_cv_4_folds_Imputed_101_Patients_SCS_Prescription_Prediction_Tre
atment_Not_Revealed_0.8_TRAIN_SES_4_Variables_ntree_700 <-
train(OralPrednisolone ~ .,

data =
Imputed_101_Patients_SCS_Prescription_Prediction_Treatment_Revealed_0.8_TR
AIN_SES_4_Variables,

method = "rf",

trControl = training_control,
```

```
metric = "ROC",

ntree = 700,

tuneGrid =
RandomForestGRid_Imputed_101_Patients_SCS_Prescription_Prediction_Treatmen
t_Not_Revealed_0.8_TRAIN_SES_4_Variables)



set.seed(5782)
randomforest_cv_4_folds_Imputed_101_Patients_SCS_Prescription_Prediction_Tre
atment_Not_Revealed_0.8_TRAIN_SES_4_Variables_ntree_800 <-
train(OralPrednisolone ~ .,

data =
Imputed_101_Patients_SCS_Prescription_Prediction_Treatment_Revealed_0.8_TR
AIN_SES_4_Variables,

method = "rf",

trControl = training_control,

metric = "ROC",

ntree = 800,

tuneGrid =
RandomForestGRid_Imputed_101_Patients_SCS_Prescription_Prediction_Treatmen
t_Not_Revealed_0.8_TRAIN_SES_4_Variables)

set.seed(5782)
randomforest_cv_4_folds_Imputed_101_Patients_SCS_Prescription_Prediction_Tre
atment_Not_Revealed_0.8_TRAIN_SES_4_Variables_ntree_900 <-
train(OralPrednisolone ~ .,

data =
Imputed_101_Patients_SCS_Prescription_Prediction_Treatment_Revealed_0.8_TR
AIN_SES_4_Variables,

method = "rf",

trControl = training_control,

metric = "ROC",

ntree = 900,

tuneGrid =
```

```
RandomForestGRid_Imputed_101_Patients_SCS_Prescription_Prediction_Treatmen
t_Not_Revealed_0.8_TRAIN_SES_4_Variables)


set.seed(5782)
randomforest_cv_4_folds_Imputed_101_Patients_SCS_Prescription_Prediction_Tre
atment_Not_Revealed_0.8_TRAIN_SES_4_Variables_ntree_1000 <-
train(OralPrednisolone ~ .,

data =
Imputed_101_Patients_SCS_Prescription_Prediction_Treatment_Revealed_0.8_TR
AIN_SES_4_Variables,

method = "rf",

trControl = training_control,

metric = "ROC",

ntree = 1000,

tuneGrid =
RandomForestGRid_Imputed_101_Patients_SCS_Prescription_Prediction_Treatmen
t_Not_Revealed_0.8_TRAIN_SES_4_Variables)



set.seed(5782)
cart1_cv_4_folds_Imputed_101_Patients_SCS_Prescription_Prediction_Treatment_
Not_Revealed_0.8_TRAIN_SES_4_Variables <- train(OralPrednisolone ~ .,
                                                data =
Imputed_101_Patients_SCS_Prescription_Prediction_Treatment_Revealed_0.8_TR
AIN_SES_4_Variables,
                                                method = "rpart",
                                                trControl =
training_control,
                                                metric = "ROC",
                                                tuneLength = 10)

set.seed(5782)
knn_cv_4_folds_Imputed_101_Patients_SCS_Prescription_Prediction_Treatment_N
ot_Revealed_0.8_TRAIN_SES_4_Variables <- train(OralPrednisolone ~ .,
                                               data =
Dataframe_Imputed_101_Patients_SCS_Prescription_Prediction_Treatment_Reveal
ed_0.8_TRAIN_SES_4_Variables,
                                               method = "knn",
                                               trControl =
training_control,
                                               metric = "ROC",
```

```
                                                                    tuneGrid =
data.frame(k = seq(1,59,by = 2)))



set.seed(5782)
sigmaRange_Imputed_101_Patients_SCS_Prescription_Prediction_Treatment_Not_
Revealed_0.8_TRAIN_SES_4_Variables <-
sigest(Matrix_Imputed_101_Patients_SCS_Prescription_Prediction_Treatment_Not_
Revealed_0.8_TRAIN_SES_4_Variables)
set.seed(5782)
svmRGRid_Imputed_101_Patients_SCS_Prescription_Prediction_Treatment_Not_R
evealed_0.8_TRAIN_SES_4_Variables <- expand.grid(.sigma =
seq(sigmaRange_Imputed_101_Patients_SCS_Prescription_Prediction_Treatment_
Not_Revealed_0.8_TRAIN_SES_4_Variables[1],sigmaRange_Imputed_101_Patient
s_SCS_Prescription_Prediction_Treatment_Not_Revealed_0.8_TRAIN_SES_4_Vari
ables[3],by =
((sigmaRange_Imputed_101_Patients_SCS_Prescription_Prediction_Treatment_Not
_Revealed_0.8_TRAIN_SES_4_Variables[3] -
sigmaRange_Imputed_101_Patients_SCS_Prescription_Prediction_Treatment_Not_
Revealed_0.8_TRAIN_SES_4_Variables[1])/8)), .C = 2^(seq(-4,4)))

set.seed(5782)
SVM_Radial_cv_4_folds_Imputed_101_Patients_SCS_Prescription_Prediction_Trea
tment_Not_Revealed_0.8_TRAIN_SES_4_Variables <- train(OralPrednisolone ~ .,
                                                                    data =
Dataframe_Imputed_101_Patients_SCS_Prescription_Prediction_Treatment_Reveal
ed_0.8_TRAIN_SES_4_Variables,
                                                                    method =
"svmRadial",
                                                                    trControl =
training_control,
                                                                    metric =
"ROC",
                                                                    preProcess =
c("center", "scale"),
                                                                    fit = FALSE,
                                                                    tuneGrid =
svmRGRid_Imputed_101_Patients_SCS_Prescription_Prediction_Treatment_Not_R
evealed_0.8_TRAIN_SES_4_Variables)



set.seed(5782)
nnetRGRid_Imputed_101_Patients_SCS_Prescription_Prediction_Treatment_Not_R
evealed_0.8_TRAIN_SES_4_Variables <- expand.grid(.size = 1:10, .decay = seq(0,
2, by = 0.1))
maxSize_Imputed_101_Patients_SCS_Prescription_Prediction_Treatment_Not_Rev
ealed_0.8_TRAIN_SES_4_Variables <-
max(nnetRGRid_Imputed_101_Patients_SCS_Prescription_Prediction_Treatment_N
ot_Revealed_0.8_TRAIN_SES_4_Variables$.size)
```

```
numWts_Imputed_101_Patients_SCS_Prescription_Prediction_Treatment_Not_Revealed_0.8_TRAIN_SES_4_Variables <-
1*(maxSize_Imputed_101_Patients_SCS_Prescription_Prediction_Treatment_Not_Revealed_0.8_TRAIN_SES_4_Variables *
(ncol(Dataframe_Imputed_101_Patients_SCS_Prescription_Prediction_Treatment_Not_Revealed_0.8_TRAIN_SES_4_Variables) + 1) +
maxSize_Imputed_101_Patients_SCS_Prescription_Prediction_Treatment_Not_Revealed_0.8_TRAIN_SES_4_Variables + 1)
set.seed(5782)
nnet_cv_4_folds_Imputed_101_Patients_SCS_Prescription_Prediction_Treatment_Not_Revealed_0.8_TRAIN_SES_4_Variables <- train(OralPrednisolone ~ .,
                                                                                                                           data =
Dataframe_Imputed_101_Patients_SCS_Prescription_Prediction_Treatment_Revealed_0.8_TRAIN_SES_4_Variables,
                                                                                                                           method = "nnet",
                                                                                                                           trControl =
training_control,
                                                                                                                           metric = "ROC",
                                                                                                                           preProcess =
c("center", "scale", "spatialSign"),
                                                                                                                           tuneGrid =
nnetRGRid_Imputed_101_Patients_SCS_Prescription_Prediction_Treatment_Not_Revealed_0.8_TRAIN_SES_4_Variables,
                                                                                                                           trace = FALSE,
                                                                                                                           maxit = 2000,
                                                                                                                           MaxNWts =
numWts_Imputed_101_Patients_SCS_Prescription_Prediction_Treatment_Not_Revealed_0.8_TRAIN_SES_4_Variables)


test_set_Imputed_101_Patients_SCS_Prescription_Prediction_Treatment_Not_Revealed_0.8_TRAIN_SES_4_Variables <-
predict(nnet_cv_4_folds_Imputed_101_Patients_SCS_Prescription_Prediction_Treatment_Not_Revealed_0.8_TRAIN_SES_4_Variables,
Dataframe_Imputed_101_Patients_SCS_Prescription_Prediction_0.2_TEST_SES_4_variables_Treatment_Not_Revealed, type = "prob")
test_set_Imputed_101_Patients_SCS_Prescription_Prediction_Treatment_Not_Revealed_0.8_TRAIN_SES_4_Variables_ROC_Neural_Networks <-
evalm(data.frame(test_set_Imputed_101_Patients_SCS_Prescription_Prediction_Treatment_Not_Revealed_0.8_TRAIN_SES_4_Variables,
Imputed_101_Patients_SCS_Prescription_Prediction_0.2_TEST_SES_4_variables$OralPrednisolone, Group = "Neural Network"))
test_set_Imputed_101_Patients_SCS_Prescription_Prediction_Treatment_Not_Revealed_0.8_TRAIN_SES_4_Variables_ROC_Neural_Networks$roc
```

Appendix chapter 4 code

R libraries used for analysis

```
library(missForest)
library(ggplot2)
library(dendextend)
library(dplyr)
library(sm)
library(NbClust)
library(factoextra)
library(rgl)
library(car)
library(crossmatch)
library(MXM)
library(caret)
library(MLmetrics)
library(plotROC)
library(kernlab)
library(nnet)
library(MLeval)
library(plotROC)
library(PMCMRplus)
```

Principal component analysis and cluster analysis

```
PCA_Analysis_of_Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_Samples <-
prcomp(Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_Samples, retx = TRUE, center = TRUE, scale = TRUE)
summary(PCA_Analysis_of_Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_Samples)
PCA_Analysis_of_Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_Samples$rotation
Principal_Components_Representing_Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_Samples <-
PCA_Analysis_of_Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_Samples$x
Five_Key_Principal_Components_Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_Samples <-
Principal_Components_Representing_Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_Samples[,1:5]
Dist_Five_Key_Principal_Components_Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_Samples <-
dist(Five_Key_Principal_Components_Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_Samples, method="euclidean")

duda_ward.D2_Five_Key_Principal_Components_Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_Samples <-
NbClust(Five_Key_Principal_Components_Just_VAS_CRP_Neutrophils_Eosinophils
```

```
_Sputum_Blood_Imputed_All_Samples, distance = "euclidean", method = "ward.D2",
max.nc = 16, index = "duda")
pseudot2_ward.D2_Five_Key_Principal_Components_Just_VAS_CRP_Neutrophils_
Eosinophils_Sputum_Blood_Imputed_All_Samples <-
NbClust(Five_Key_Principal_Components_Just_VAS_CRP_Neutrophils_Eosinophils
_Sputum_Blood_Imputed_All_Samples, distance = "euclidean", method = "ward.D2",
max.nc = 16, index = "pseudot2")
duda_ward.D2_Five_Key_Principal_Components_Just_VAS_CRP_Neutrophils_Eosi
nophils_Sputum_Blood_Imputed_All_Samples$Best.nc
pseudot2_ward.D2_Five_Key_Principal_Components_Just_VAS_CRP_Neutrophils_
Eosinophils_Sputum_Blood_Imputed_All_Samples$Best.nc

duda_average_Five_Key_Principal_Components_Just_VAS_CRP_Neutrophils_Eosi
nophils_Sputum_Blood_Imputed_All_Samples <-
NbClust(Five_Key_Principal_Components_Just_VAS_CRP_Neutrophils_Eosinophils
_Sputum_Blood_Imputed_All_Samples, distance = "euclidean", method = "average",
max.nc = 16, index = "duda")
pseudot2_average_Five_Key_Principal_Components_Just_VAS_CRP_Neutrophils_
Eosinophils_Sputum_Blood_Imputed_All_Samples <-
NbClust(Five_Key_Principal_Components_Just_VAS_CRP_Neutrophils_Eosinophils
_Sputum_Blood_Imputed_All_Samples, distance = "euclidean", method = "average",
max.nc = 16, index = "pseudot2")
duda_average_Five_Key_Principal_Components_Just_VAS_CRP_Neutrophils_Eosi
nophils_Sputum_Blood_Imputed_All_Samples$Best.nc
pseudot2_average_Five_Key_Principal_Components_Just_VAS_CRP_Neutrophils_
Eosinophils_Sputum_Blood_Imputed_All_Samples$Best.nc
duda_mcquitty_Five_Key_Principal_Components_Just_VAS_CRP_Neutrophils_Eosi
nophils_Sputum_Blood_Imputed_All_Samples <-
NbClust(Five_Key_Principal_Components_Just_VAS_CRP_Neutrophils_Eosinophils
_Sputum_Blood_Imputed_All_Samples, distance = "euclidean", method = "mcquitty",
max.nc = 16, index = "duda")
pseudot2_mcquitty_Five_Key_Principal_Components_Just_VAS_CRP_Neutrophils_
Eosinophils_Sputum_Blood_Imputed_All_Samples <-
NbClust(Five_Key_Principal_Components_Just_VAS_CRP_Neutrophils_Eosinophils
_Sputum_Blood_Imputed_All_Samples, distance = "euclidean", method = "mcquitty",
max.nc = 16, index = "pseudot2")
duda_mcquitty_Five_Key_Principal_Components_Just_VAS_CRP_Neutrophils_Eosi
nophils_Sputum_Blood_Imputed_All_Samples$Best.nc
pseudot2_mcquitty_Five_Key_Principal_Components_Just_VAS_CRP_Neutrophils_
Eosinophils_Sputum_Blood_Imputed_All_Samples$Best.nc
duda_median_Five_Key_Principal_Components_Just_VAS_CRP_Neutrophils_Eosin
ophils_Sputum_Blood_Imputed_All_Samples <-
NbClust(Five_Key_Principal_Components_Just_VAS_CRP_Neutrophils_Eosinophils
_Sputum_Blood_Imputed_All_Samples, distance = "euclidean", method = "median",
max.nc = 16, index = "duda")
pseudot2_median_Five_Key_Principal_Components_Just_VAS_CRP_Neutrophils_E
osinophils_Sputum_Blood_Imputed_All_Samples <-
NbClust(Five_Key_Principal_Components_Just_VAS_CRP_Neutrophils_Eosinophils
_Sputum_Blood_Imputed_All_Samples, distance = "euclidean", method = "median",
max.nc = 16, index = "pseudot2")
```

```
duda_median_Five_Key_Principal_Components_Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_Samples$Best.nc
pseudot2_median_Five_Key_Principal_Components_Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_Samples$Best.nc
duda_centroid_Five_Key_Principal_Components_Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_Samples <-
NbClust(Five_Key_Principal_Components_Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_Samples, distance = "euclidean", method = "centroid", max.nc = 16, index = "duda")
pseudot2_centroid_Five_Key_Principal_Components_Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_Samples <-
NbClust(Five_Key_Principal_Components_Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_Samples, distance = "euclidean", method = "centroid", max.nc = 16, index = "pseudot2")
duda_centroid_Five_Key_Principal_Components_Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_Samples$Best.nc
pseudot2_centroid_Five_Key_Principal_Components_Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_Samples$Best.nc

All_Samples_For_Equilibrium_Crisis_Definition_Datasheet_Ready_Clean_Full <-
read.csv(file="All_Samples_For_Equilibrium_Crisis_Definition_Datasheet_Ready_Clean_Full.csv", sep=",", header=T)
All_Samples_For_Equilibrium_Crisis_Definition_Datasheet_Ready_Clean_Full_With_Objective_States <-
All_Samples_For_Equilibrium_Crisis_Definition_Datasheet_Ready_Clean_Full
All_Samples_For_Equilibrium_Crisis_Definition_Datasheet_Ready_Clean_Full_With_Objective_States$Objective_State <-
duda_ward.D2_Five_Key_Principal_Components_Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_Samples$Best.partition
All_Samples_For_Equilibrium_Crisis_Definition_Datasheet_Ready_Clean_Full_With_Objective_States$Objective_State <-
as.factor(All_Samples_For_Equilibrium_Crisis_Definition_Datasheet_Ready_Clean_Full_With_Objective_States$Objective_State)
Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_Samples_With_Objective_States <-
Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_Samples
Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_Samples_With_Objective_States$Objective_State <-
duda_ward.D2_Five_Key_Principal_Components_Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_Samples$Best.partition
Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_Samples_With_Objective_States$Objective_State <-
as.factor(Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_Samples_With_Objective_States$Objective_State)
summary(Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_Samples_With_Objective_States[Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_Samples_With_Objective_States$Objective_State=="1",])
summary(Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_Samples_With_Objective_States[Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_Samples_With_Objective_States$Objective_State=="2",])
```

```
summary(Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_Samples_With_Objective_States[Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_Samples_With_Objective_States$Objective_State=="3",])
summary(Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_Samples_With_Objective_States[Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_Samples_With_Objective_States$Objective_State=="4",])
All_Samples_For_Equilibrium_Crisis_Definition_Datasheet_Ready_Clean_Full_With_Objective_States$Objective_State_Name <-
ifelse(All_Samples_For_Equilibrium_Crisis_Definition_Datasheet_Ready_Clean_Full_With_Objective_States$Objective_State=="1", "Equilibrium",
ifelse(All_Samples_For_Equilibrium_Crisis_Definition_Datasheet_Ready_Clean_Full_With_Objective_States$Objective_State=="2", "Symptomatic_Crisis",
ifelse(All_Samples_For_Equilibrium_Crisis_Definition_Datasheet_Ready_Clean_Full_With_Objective_States$Objective_State=="3", "Eosinophilic_Crisis",
"Neutrophilic_Crisis")))
Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples <-
Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_Samples
Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples[,6:9] <-
NULL
PCA_Analysis_of_Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples <-
prcomp(Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples, retx = TRUE, center = TRUE, scale = TRUE)
summary(PCA_Analysis_of_Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples)
PCA_Analysis_of_Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples$rotation
Principal_Components_Representing_Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples <-
PCA_Analysis_of_Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples$x
Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples <-
Principal_Components_Representing_Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples[,1:3]

Dataframe_Five_Key_Principal_Components_Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_Samples$Objective_State_Name_Based_on_Five_PCs_Without_Label <-
ifelse(Dataframe_Five_Key_Principal_Components_Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_Samples$Objective_State_Name_Based_on_Five_PCs=="Equilibrium",".",
ifelse(Dataframe_Five_Key_Principal_Components_Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_Samples$Objective_State_Name_Based_on_Five_PCs=="Symptomatic_Crisis", ",",
ifelse(Dataframe_Five_Key_Principal_Components_Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_Samples$Objective_State_Name_Based_on_Five_PCs=="Eosinophilic_Crisis", ":","'")))
Dataframe_Five_Key_Principal_Components_Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_Samples$Objective_State_Name_Based_on_Five
```

```
_PCs_Without_Label <-
as.factor(Dataframe_Five_Key_Principal_Components_Just_VAS_CRP_Neutrophils
_Eosinophils_Sputum_Blood_Imputed_All_Samples$Objective_State_Name_Based
_on_Five_PCs_Without_Label)
Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosi
nophils_Imputed_All_Samples <-
as.data.frame(Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_E
osinophils_Imputed_All_Samples)
Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosi
nophils_Imputed_All_Samples$Exacerbation_or_Stable <-
All_Samples_For_Equilibrium_Crisis_Definition_Datasheet_Ready_Clean_Full_With
_Objective_States$Exacerbation_or_Stable
Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosi
nophils_Imputed_All_Samples$Exacerbation_or_Stable <-
as.factor(Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrop
hils_Eosinophils_Imputed_All_Samples$Exacerbation_or_Stable)
Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosi
nophils_Imputed_All_Samples$Exacerbation_or_Stable_Without_Label_Name <-
ifelse(Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils
_Eosinophils_Imputed_All_Samples$Exacerbation_or_Stable=="Exacerbation", ".",
",")
Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosi
nophils_Imputed_All_Samples$Exacerbation_or_Stable_Without_Label_Name <-
as.factor(Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrop
hils_Eosinophils_Imputed_All_Samples$Exacerbation_or_Stable_Without_Label_Na
me)
scatter3d(x =
Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosi
nophils_Imputed_All_Samples$PC1, y =
Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosi
nophils_Imputed_All_Samples$PC2, z =
Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosi
nophils_Imputed_All_Samples$PC3, xlab = "PC1", ylab = "PC2", zlab = "PC3",
surface = FALSE, ellipsoid = TRUE, revolutions = 0, groups =
Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosi
nophils_Imputed_All_Samples$Exacerbation_or_Stable_Without_Label_Name, level
= 0.3, axis.scales = FALSE)
scatter3d(x =
Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosi
nophils_Imputed_All_Samples$PC1, y =
Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosi
nophils_Imputed_All_Samples$PC2, z =
Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosi
nophils_Imputed_All_Samples$PC3, xlab = "PC1", ylab = "PC2", zlab = "PC3",
surface = FALSE, ellipsoid = TRUE, revolutions = 0, groups =
Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosi
nophils_Imputed_All_Samples$Objective_State_Name_Based_on_Five_PCs_Witho
ut_Label, level = 0.3, axis.scales = FALSE)
```

```
hclust_ward.D2_Five_Key_Principal_Components_Just_VAS_CRP_Neutrophils_Eos
inophils_Sputum_Blood_Imputed_All_Samples <-
hclust(Dist_Five_Key_Principal_Components_Just_VAS_CRP_Neutrophils_Eosinop
hils_Sputum_Blood_Imputed_All_Samples, method = 'ward.D2')
plot(hclust_ward.D2_Five_Key_Principal_Components_Just_VAS_CRP_Neutrophils
_Eosinophils_Sputum_Blood_Imputed_All_Samples)
cut_ward.D2_Five_Key_Principal_Components_Just_VAS_CRP_Neutrophils_Eosin
ophils_Sputum_Blood_Imputed_All_Samples <-
cutree(hclust_ward.D2_Five_Key_Principal_Components_Just_VAS_CRP_Neutroph
ils_Eosinophils_Sputum_Blood_Imputed_All_Samples, k=4)
ward.D2_dend_obj_Five_Key_Principal_Components_Just_VAS_CRP_Neutrophils_
Eosinophils_Sputum_Blood_Imputed_All_Samples <-
as.dendrogram(hclust_ward.D2_Five_Key_Principal_Components_Just_VAS_CRP_
Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_Samples)
ward.D2_col_dend_Five_Key_Principal_Components_Just_VAS_CRP_Neutrophils_
Eosinophils_Sputum_Blood_Imputed_All_Samples <-
color_branches(ward.D2_dend_obj_Five_Key_Principal_Components_Just_VAS_C
RP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_Samples, k = 4)
plot(ward.D2_col_dend_Five_Key_Principal_Components_Just_VAS_CRP_Neutrop
hils_Eosinophils_Sputum_Blood_Imputed_All_Samples)
```

Degree of overlap analysis

```
Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosi
nophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_Objective
_States$Current_Clinical_Definition <-
All_Samples_For_Equilibrium_Crisis_Definition_Datasheet_Ready_Clean_Full_With
_Objective_States$Exacerbation_or_Stable
Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosi
nophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_Objective
_States$Current_Clinical_Definition <-
as.factor(Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrop
hils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_
Objective_States$Current_Clinical_Definition)
Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosi
nophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_Objective
_States$Equilibrium_and_Crisis_States <-
All_Samples_For_Equilibrium_Crisis_Definition_Datasheet_Ready_Clean_Full_With
_Objective_States$Objective_State_Name
Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosi
nophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_Objective
_States$Equilibrium_and_Crisis_States <-
as.factor(Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrop
hils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_
Objective_States$Equilibrium_and_Crisis_States)
```

```
Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosi
nophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_Objective
_States_Exacerbation_V_Stable <-
Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosi
nophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_Objective
_States
Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosi
nophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_Objective
_States_Equilibrium_V_Symptomatic <-
Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosi
nophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_Objective
_States[Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophi
ls_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_O
bjective_States$Equilibrium_and_Crisis_States=="Equilibrium" |
Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosi
nophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_Objective
_States$Equilibrium_and_Crisis_States=="Symptomatic_Crisis",]
Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosi
nophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_Objective
_States_Equilibrium_V_Eosinophilic <-
Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosi
nophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_Objective
_States[Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophi
ls_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_O
bjective_States$Equilibrium_and_Crisis_States=="Equilibrium" |
Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosi
nophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_Objective
_States$Equilibrium_and_Crisis_States=="Eosinophilic_Crisis",]
Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosi
nophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_Objective
_States_Equilibrium_V_Neutrophilic <-
Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosi
nophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_Objective
_States[Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophi
ls_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_O
bjective_States$Equilibrium_and_Crisis_States=="Equilibrium" |
Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosi
nophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_Objective
_States$Equilibrium_and_Crisis_States=="Neutrophilic_Crisis",]
Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosi
nophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_Objective
_States_Symptomatic_V_Eosinophilic <-
Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosi
nophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_Objective
_States[Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophi
ls_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_O
bjective_States$Equilibrium_and_Crisis_States=="Symptomatic_Crisis" |
Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosi
nophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_Objective
_States$Equilibrium_and_Crisis_States=="Eosinophilic_Crisis",]
```

```
Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_Objective_States_Symptomatic_V_Neutrophilic <-
Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_Objective_States[Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_Objective_States$Equilibrium_and_Crisis_States=="Symptomatic_Crisis" |
Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_Objective_States$Equilibrium_and_Crisis_States=="Neutrophilic_Crisis",]
Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_Objective_States_Eosinophilic_V_Neutrophilic <-
Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_Objective_States[Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_Objective_States$Equilibrium_and_Crisis_States=="Eosinophilic_Crisis" |
Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_Objective_States$Equilibrium_and_Crisis_States=="Neutrophilic_Crisis",]
```

```
Binary_Vector_Exacerbation_V_Stable <-
ifelse(Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_Objective_States_Exacerbation_V_Stable$Current_Clinical_Definition=='Stable', 0, 1)
Binary_Vector_Equilibrium_V_Symptomatic_Crisis <-
ifelse(Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_Objective_States_Equilibrium_V_Symptomatic$Equilibrium_and_Crisis_States=='Equilibrium', 0, 1)
Binary_Vector_Equilibrium_V_Eosinophilic_Crisis <-
ifelse(Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_Objective_States_Equilibrium_V_Eosinophilic$Equilibrium_and_Crisis_States=='Equilibrium', 0, 1)
Binary_Vector_Equilibrium_V_Neutrophilic_Crisis <-
ifelse(Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_Objective_States_Equilibrium_V_Neutrophilic$Equilibrium_and_Crisis_States=='Equilibrium', 0, 1)
Binary_Vector_Symptomatic_Crisis_V_Neutrophilic_Crisis <-
ifelse(Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_Objective_States_Symptomatic_V_Neutrophilic$Equilibrium_and_Crisis_States=='Symptomatic_Crisis', 0, 1)
```

```
Binary_Vector_Symptomatic_Crisis_V_Eosinophilic_Crisis <-
ifelse(Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils
_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_Obj
ective_States_Symptomatic_V_Eosinophilic$Equilibrium_and_Crisis_States=='Symp
tomatic_Crisis', 0, 1)
Binary_Vector_Neutrophilic_Crisis_V_Eosinophilic_Crisis <-
ifelse(Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils
_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_Obj
ective_States_Eosinophilic_V_Neutrophilic$Equilibrium_and_Crisis_States=='Neutro
philic_Crisis', 0, 1)

Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophil
s_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_O
bjective_States_Exacerbation_V_Stable <-
Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosi
nophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_Objective
_States_Exacerbation_V_Stable
Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophil
s_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_O
bjective_States_Exacerbation_V_Stable <-
Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophil
s_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_O
bjective_States_Exacerbation_V_Stable[,1:3]
Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophil
s_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_O
bjective_States_Equilibrium_V_Symptomatic <-
Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosi
nophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_Objective
_States_Equilibrium_V_Symptomatic
Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophil
s_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_O
bjective_States_Equilibrium_V_Symptomatic <-
Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophil
s_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_O
bjective_States_Equilibrium_V_Symptomatic[,1:3]
Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophil
s_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_O
bjective_States_Equilibrium_V_Eosinophilic <-
Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosi
nophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_Objective
_States_Equilibrium_V_Eosinophilic
Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophil
s_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_O
bjective_States_Equilibrium_V_Eosinophilic <-
Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophil
s_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_O
bjective_States_Equilibrium_V_Eosinophilic[,1:3]
Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophil
s_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_O
bjective_States_Equilibrium_V_Neutrophilic <-
```

```
Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosi
nophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_Objective
_States_Equilibrium_V_Neutrophilic
Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophil
s_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_O
bjective_States_Equilibrium_V_Neutrophilic <-
Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophil
s_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_O
bjective_States_Equilibrium_V_Neutrophilic[,1:3]
Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophil
s_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_O
bjective_States_Symptomatic_V_Eosinophilic <-
Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosi
nophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_Objective
_States_Symptomatic_V_Eosinophilic
Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophil
s_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_O
bjective_States_Symptomatic_V_Eosinophilic <-
Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophil
s_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_O
bjective_States_Symptomatic_V_Eosinophilic[,1:3]
Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophil
s_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_O
bjective_States_Symptomatic_V_Neutrophilic <-
Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosi
nophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_Objective
_States_Symptomatic_V_Neutrophilic
Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophil
s_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_O
bjective_States_Symptomatic_V_Neutrophilic <-
Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophil
s_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_O
bjective_States_Symptomatic_V_Neutrophilic[,1:3]
Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophil
s_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_O
bjective_States_Eosinophilic_V_Neutrophilic <-
Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophils_Eosi
nophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_Objective
_States_Eosinophilic_V_Neutrophilic
Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophil
s_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_O
bjective_States_Eosinophilic_V_Neutrophilic <-
Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophil
s_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_O
bjective_States_Eosinophilic_V_Neutrophilic[,1:3]
Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophil
s_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_O
bjective_States_Exacerbation_V_Stable <-
as.matrix(Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_
```

```
Neutrophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_De
finition_Objective_States_Exacerbation_V_Stable)
Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophil
s_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_O
bjective_States_Equilibrium_V_Symptomatic <-
as.matrix(Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_
Neutrophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_De
finition_Objective_States_Equilibrium_V_Symptomatic)
Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophil
s_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_O
bjective_States_Equilibrium_V_Eosinophilic <-
as.matrix(Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_
Neutrophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_De
finition_Objective_States_Equilibrium_V_Eosinophilic)
Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophil
s_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_O
bjective_States_Equilibrium_V_Neutrophilic <-
as.matrix(Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_
Neutrophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_De
finition_Objective_States_Equilibrium_V_Neutrophilic)
Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophil
s_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_O
bjective_States_Symptomatic_V_Eosinophilic <-
as.matrix(Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_
Neutrophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_De
finition_Objective_States_Symptomatic_V_Eosinophilic)
Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophil
s_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_O
bjective_States_Symptomatic_V_Neutrophilic <-
as.matrix(Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_
Neutrophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_De
finition_Objective_States_Symptomatic_V_Neutrophilic)
Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutrophil
s_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definition_O
bjective_States_Eosinophilic_V_Neutrophilic <-
as.matrix(Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_
Neutrophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_De
finition_Objective_States_Eosinophilic_V_Neutrophilic)
Dist_Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutr
ophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definitio
n_Objective_States_Exacerbation_V_Stable <-
dist(Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutr
ophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definitio
n_Objective_States_Exacerbation_V_Stable, method="euclidean")
Matrix_Dist_Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CR
P_Neutrophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_
Definition_Objective_States_Exacerbation_V_Stable <-
as.matrix(Dist_Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_C
RP_Neutrophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical
_Definition_Objective_States_Exacerbation_V_Stable)
```

```
Dist_Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutr
ophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definitio
n_Objective_States_Equilibrium_V_Symptomatic <-
dist(Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutr
ophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definitio
n_Objective_States_Equilibrium_V_Symptomatic, method="euclidean")
Matrix_Dist_Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CR
P_Neutrophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_
Definition_Objective_States_Equilibrium_V_Symptomatic <-
as.matrix(Dist_Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_C
RP_Neutrophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical
_Definition_Objective_States_Equilibrium_V_Symptomatic)
Dist_Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutr
ophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definitio
n_Objective_States_Equilibrium_V_Eosinophilic <-
dist(Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutr
ophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definitio
n_Objective_States_Equilibrium_V_Eosinophilic, method="euclidean")
Matrix_Dist_Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CR
P_Neutrophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_
Definition_Objective_States_Equilibrium_V_Eosinophilic <-
as.matrix(Dist_Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_C
RP_Neutrophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical
_Definition_Objective_States_Equilibrium_V_Eosinophilic)
Dist_Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutr
ophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definitio
n_Objective_States_Equilibrium_V_Neutrophilic <-
dist(Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutr
ophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definitio
n_Objective_States_Equilibrium_V_Neutrophilic, method="euclidean")
Matrix_Dist_Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CR
P_Neutrophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_
Definition_Objective_States_Equilibrium_V_Neutrophilic <-
as.matrix(Dist_Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_C
RP_Neutrophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical
_Definition_Objective_States_Equilibrium_V_Neutrophilic)
Dist_Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutr
ophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definitio
n_Objective_States_Symptomatic_V_Eosinophilic <-
dist(Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutr
ophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definitio
n_Objective_States_Symptomatic_V_Eosinophilic, method="euclidean")
Matrix_Dist_Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CR
P_Neutrophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_
Definition_Objective_States_Symptomatic_V_Eosinophilic <-
as.matrix(Dist_Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_C
RP_Neutrophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical
_Definition_Objective_States_Symptomatic_V_Eosinophilic)
Dist_Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutr
ophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definitio
```

```
n_Objective_States_Symptomatic_V_Neutrophilic <-
dist(Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutr
ophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definitio
n_Objective_States_Symptomatic_V_Neutrophilic, method="euclidean")
Matrix_Dist_Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CR
P_Neutrophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_
Definition_Objective_States_Symptomatic_V_Neutrophilic <-
as.matrix(Dist_Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_C
RP_Neutrophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical
_Definition_Objective_States_Symptomatic_V_Neutrophilic)
Dist_Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutr
ophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definitio
n_Objective_States_Eosinophilic_V_Neutrophilic <-
dist(Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CRP_Neutr
ophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_Definitio
n_Objective_States_Eosinophilic_V_Neutrophilic, method="euclidean")
Matrix_Dist_Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CR
P_Neutrophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_
Definition_Objective_States_Eosinophilic_V_Neutrophilic <-
as.matrix(Dist_Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_C
RP_Neutrophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical
_Definition_Objective_States_Eosinophilic_V_Neutrophilic)

crossmatchtest_Exacerbation_V_Stable <-
crossmatchtest(Binary_Vector_Exacerbation_V_Stable,
Matrix_Dist_Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CR
P_Neutrophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_
Definition_Objective_States_Exacerbation_V_Stable)
crossmatchtest_Exacerbation_V_Stable$a1
crossmatchtest_Equilibrium_V_Symptomatic_Crisis <-
crossmatchtest(Binary_Vector_Equilibrium_V_Symptomatic_Crisis,
Matrix_Dist_Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CR
P_Neutrophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_
Definition_Objective_States_Equilibrium_V_Symptomatic)
crossmatchtest_Equilibrium_V_Symptomatic_Crisis$a1
crossmatchtest_Equilibrium_V_Eosinophilic_Crisis <-
crossmatchtest(Binary_Vector_Equilibrium_V_Eosinophilic_Crisis,
Matrix_Dist_Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CR
P_Neutrophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_
Definition_Objective_States_Equilibrium_V_Eosinophilic)
crossmatchtest_Equilibrium_V_Eosinophilic_Crisis$a1
crossmatchtest_Equilibrium_V_Neutrophilic_Crisis <-
crossmatchtest(Binary_Vector_Equilibrium_V_Neutrophilic_Crisis,
Matrix_Dist_Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CR
P_Neutrophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_
Definition_Objective_States_Equilibrium_V_Neutrophilic)
crossmatchtest_Equilibrium_V_Neutrophilic_Crisis$a1
crossmatchtest_Symptomatic_Crisis_V_Neutrophilic_Crisis <-
crossmatchtest(Binary_Vector_Symptomatic_Crisis_V_Neutrophilic_Crisis,
Matrix_Dist_Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CR
```

```
P_Neutrophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_
Definition_Objective_States_Symptomatic_V_Neutrophilic)
crossmatchtest_Symptomatic_Crisis_V_Neutrophilic_Crisis$a1
crossmatchtest_Symptomatic_Crisis_V_Eosinophilic_Crisis <-
crossmatchtest(Binary_Vector_Symptomatic_Crisis_V_Eosinophilic_Crisis,
Matrix_Dist_Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CR
P_Neutrophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_
Definition_Objective_States_Symptomatic_V_Eosinophilic)
crossmatchtest_Symptomatic_Crisis_V_Eosinophilic_Crisis$a1
crossmatchtest_Neutrophilic_Crisis_V_Eosinophilic_Crisis <-
crossmatchtest(Binary_Vector_Neutrophilic_Crisis_V_Eosinophilic_Crisis,
Matrix_Dist_Matrix_Dataframe_Three_Principal_Components_Just_VAS_Blood_CR
P_Neutrophils_Eosinophils_Imputed_All_Samples_Together_With_Current_Clinical_
Definition_Objective_States_Eosinophilic_V_Neutrophilic)
crossmatchtest_Neutrophilic_Crisis_V_Eosinophilic_Crisis$a1
```

Creation of hold-out test set of events as well as cross-validation folds in remaining data for COPD state detection

```
Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples <-
read.csv(file="Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Sample
s.csv", sep=",", header = T)
Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples$X <- NULL
Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_Samples_Wit
h_Objective_States <-
read.csv(file="Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All
_Samples_With_Objective_States.csv", sep=",", header=T)
Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_Samples_Wit
h_Objective_States$X <- NULL
All_Samples_For_Equilibrium_Crisis_Definition_Datasheet_Ready_Clean_Full_With
_Objective_States <-
read.csv(file="All_Samples_For_Equilibrium_Crisis_Definition_Datasheet_Ready_Cl
ean_Full_With_Objective_States.csv", sep=",", header=T)
Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_Ready_for_
Data_Splitting <-
Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples
Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_Ready_for_
Data_Splitting$Objective_State <-
All_Samples_For_Equilibrium_Crisis_Definition_Datasheet_Ready_Clean_Full_With
_Objective_States$Objective_State_Name
Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_Ready_for_
Data_Splitting$Subject_Number <-
All_Samples_For_Equilibrium_Crisis_Definition_Datasheet_Ready_Clean_Full_With
_Objective_States$Subject_number
Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_Ready_for_
Data_Splitting$Objective_State <-
```

```
as.factor(Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_Ready_for_Data_Splitting$Objective_State)
Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_Ready_for_Data_Splitting$Subject_Number <-
as.factor(Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_Ready_for_Data_Splitting$Subject_Number)


set.seed(5783)
k.folds <- 5
set.seed(5783)
index_Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_Ready_for_Data_Splitting <-
groupKFold(Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_Ready_for_Data_Splitting$Subject_Number, k = k.folds)

length(index_Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_Ready_for_Data_Splitting$Fold5)

Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_0.8_TRAIN <-
Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_Ready_for_Data_Splitting[index_Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_Ready_for_Data_Splitting$Fold5,]
Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_0.2_TEST <-
Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_Ready_for_Data_Splitting[-
index_Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_Ready_for_Data_Splitting$Fold5,]
Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_No_Label_No_Subject_Number_0.8_TRAIN <-
Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_0.8_TRAIN
Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_No_Label_No_Subject_Number_0.8_TRAIN$Objective_State <- NULL
Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_No_Label_No_Subject_Number_0.8_TRAIN$Subject_Number <- NULL
```

Single state biomarker ROC curve analysis

```
setwd("~/OneDrive - Nexus365/DPhil Research/MRC Analysis/Nov 2022")

Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC <-
Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_0.8_TRAIN
Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC$Equilibrium_or_Crisis <-
ifelse(Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_T
```

```
RAIN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC$Objective_State=="Equilibrium", "Equilibrium", "Crisis")
Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibrium_Symptomatic <-
Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC[Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC$Objective_State=="Equilibrium" |
Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC$Objective_State=="Symptomatic_Crisis",]
Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibrium_Eosinophilic <-
Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC[Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC$Objective_State=="Equilibrium" |
Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC$Objective_State=="Eosinophilic_Crisis",]
Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibrium_Neutrophilic <-
Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC[Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC$Objective_State=="Equilibrium" |
Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC$Objective_State=="Neutrophilic_Crisis",]

Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Class_Name_An_Equilibrium_V_Crisis <-
Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC
Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Class_Name_An_Equilibrium_V_Crisis$An_Equilibrium_or_Crisis <-
ifelse(Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Class_Name_An_Equilibrium_V_Crisis$Equilibrium_or_Crisis=="Equilibrium",
"An_Equilibrium", "Crisis")
```

```
Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_
Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Class_Na
me_An_Equilibrium_V_Crisis$An_Equilibrium_or_Crisis <-
as.factor(Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8
_TRAIN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Cl
ass_Name_An_Equilibrium_V_Crisis$An_Equilibrium_or_Crisis)

set.seed(5783)
longtest_0.8_TRAIN_Equilibrium_Versus_Symptomatic_Crisis_Only_Symptoms <-
melt_roc(Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8
_TRAIN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_E
quilibrium_Symptomatic, "Objective_State", c("VAS.Cough", "VAS.Dypsonea",
"VAS.Sputum.Prod", "VAS.Sputum.Purulence"))
ggplot(longtest_0.8_TRAIN_Equilibrium_Versus_Symptomatic_Crisis_Only_Sympto
ms, aes(d = D, m = M, color = name)) + geom_roc(n.cuts = 0, increasing = TRUE) +
style_roc()

Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_
Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibrium
_Symptomatic_0_1 <-
Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_
Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibrium
_Symptomatic
Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_
Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibrium
_Symptomatic_0_1$Eq_0_Symp_1 <-
ifelse(Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_T
RAIN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equi
librium_Symptomatic_0_1$Objective_State=="Equilibrium", 0, 1)

auc_VAS.Cough_Equilibrium_Versus_Symptomatic_Crisis <-
auc(Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRA
IN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibr
ium_Symptomatic_0_1$Eq_0_Symp_1,
Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_
Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibrium
_Symptomatic$VAS.Cough)
auc_VAS.Dyspnoea_Equilibrium_Versus_Symptomatic_Crisis <-
auc(Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRA
IN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibr
ium_Symptomatic_0_1$Eq_0_Symp_1,
Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_
Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibrium
_Symptomatic$VAS.Dypsonea)
auc_VAS.Sputum_Production_Equilibrium_Versus_Symptomatic_Crisis <-
auc(Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRA
IN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibr
ium_Symptomatic_0_1$Eq_0_Symp_1,
Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_
```

```
Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibrium
_Symptomatic$VAS.Sputum.Prod)
auc_VAS.Sputum_Purulence_Equilibrium_Versus_Symptomatic_Crisis <-
auc(Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRA
IN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibr
ium_Symptomatic_0_1$Eq_0_Symp_1,
Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_
Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibrium
_Symptomatic$VAS.Sputum.Purulence)
roc_VAS.Cough_Equilibrium_Versus_Symptomatic_Crisis <-
pROC::roc(Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.
8_TRAIN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_
Equilibrium_Symptomatic_0_1$Eq_0_Symp_1,
Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_
Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibrium
_Symptomatic$VAS.Cough)
coords_VAS.Cough_Equilibrium_Versus_Symptomatic_Crisis_return_threshold <-
pROC::coords(roc_VAS.Cough_Equilibrium_Versus_Symptomatic_Crisis, "best",
ret="threshold", best.method="closest.topleft", transpose = FALSE)
coords_VAS.Cough_Equilibrium_Versus_Symptomatic_Crisis_return_True_Positive_
Rate <- pROC::coords(roc_VAS.Cough_Equilibrium_Versus_Symptomatic_Crisis,
"best", ret="tpr", best.method="closest.topleft", transpose = FALSE)
coords_VAS.Cough_Equilibrium_Versus_Symptomatic_Crisis_return_False_Positive
_Rate <- pROC::coords(roc_VAS.Cough_Equilibrium_Versus_Symptomatic_Crisis,
"best", ret="fpr", best.method="closest.topleft", transpose = FALSE)

Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_
Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibrium
_Eosinophilic$An_Equilibrium_V_Eosinophilic_Crisis <-
ifelse(Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_T
RAIN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equi
librium_Eosinophilic$Objective_State=="Equilibrium", "An_Equilibrium",
"Eosinophilic_Crisis")
Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_
Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibrium
_Eosinophilic$An_Equilibrium_0_Eosinophilic_1 <-
ifelse(Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_T
RAIN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equi
librium_Eosinophilic$An_Equilibrium_V_Eosinophilic_Crisis=="An_Equilibrium", 0,1)
set.seed(5783)
longtest_0.8_TRAIN_Equilibrium_Versus_Eosinophilic_Crisis_Symptoms_and_Eosin
ophilic_Indices <-
melt_roc(Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8
_TRAIN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_E
quilibrium_Eosinophilic, "An_Equilibrium_0_Eosinophilic_1", c("VAS.Cough",
"VAS.Dypsonea", "VAS.Sputum.Prod", "VAS.Sputum.Purulence", "B.Eosinophils",
"Percentage_Blood_Eosinophils"))
ggplot(longtest_0.8_TRAIN_Equilibrium_Versus_Eosinophilic_Crisis_Symptoms_and
_Eosinophilic_Indices, aes(d = D, m = M, color = name)) + geom_roc(n.cuts = 0,
increasing = TRUE) + style_roc()
```

```
auc_VAS.Cough_Equilibrium_Versus_Eosinophilic_Crisis <-
auc(Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRA
IN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibr
ium_Eosinophilic$An_Equilibrium_0_Eosinophilic_1,
Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_
Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibrium
_Eosinophilic$VAS.Cough)
auc_VAS.Dyspnoea_Equilibrium_Versus_Eosinophilic_Crisis <-
auc(Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRA
IN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibr
ium_Eosinophilic$An_Equilibrium_0_Eosinophilic_1,
Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_
Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibrium
_Eosinophilic$VAS.Dypsonea)
auc_VAS.Sputum_Production_Equilibrium_Versus_Eosinophilic_Crisis <-
auc(Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRA
IN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibr
ium_Eosinophilic$An_Equilibrium_0_Eosinophilic_1,
Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_
Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibrium
_Eosinophilic$VAS.Sputum.Prod)
auc_VAS.Sputum_Purulence_Equilibrium_Versus_Eosinophilic_Crisis <-
auc(Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRA
IN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibr
ium_Eosinophilic$An_Equilibrium_0_Eosinophilic_1,
Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_
Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibrium
_Eosinophilic$VAS.Sputum.Purulence)
auc_B.Eosinophils_Equilibrium_Versus_Eosinophilic_Crisis <-
auc(Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRA
IN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibr
ium_Eosinophilic$An_Equilibrium_0_Eosinophilic_1,
Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_
Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibrium
_Eosinophilic$B.Eosinophils)
auc_Percentage_Blood_Eosinophils_Equilibrium_Versus_Eosinophilic_Crisis <-
auc(Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRA
IN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibr
ium_Eosinophilic$An_Equilibrium_0_Eosinophilic_1,
Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_
Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibrium
_Eosinophilic$Percentage_Blood_Eosinophils)
roc_Percentage_Blood_Eosinophils_Equilibrium_Versus_Eosinophilic_Crisis <-
pROC::roc(Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.
8_TRAIN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_
Equilibrium_Eosinophilic$An_Equilibrium_0_Eosinophilic_1,
Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_
Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibrium
_Eosinophilic$Percentage_Blood_Eosinophils)
```

```
coords_Percentage_Blood_Eosinophils_Equilibrium_Versus_Eosinophilic_Crisis_return_threshold <-
pROC::coords(roc_Percentage_Blood_Eosinophils_Equilibrium_Versus_Eosinophilic_Crisis, "best", ret="threshold", best.method="closest.topleft", transpose = FALSE)
coords_Percentage_Blood_Eosinophils_Equilibrium_Versus_Eosinophilic_Crisis_return_True_Positive_Rate <-
pROC::coords(roc_Percentage_Blood_Eosinophils_Equilibrium_Versus_Eosinophilic_Crisis, "best", ret="tpr", best.method="closest.topleft", transpose = FALSE)
coords_Percentage_Blood_Eosinophils_Equilibrium_Versus_Eosinophilic_Crisis_False_Positive_Rate <-
pROC::coords(roc_Percentage_Blood_Eosinophils_Equilibrium_Versus_Eosinophilic_Crisis, "best", ret="fpr", best.method="closest.topleft", transpose = FALSE)

Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibrium_Neutrophilic$An_Equilibrium_V_Neutrophilic_Crisis <-
ifelse(Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibrium_Neutrophilic$Objective_State=="Equilibrium",
"An_Equilibrium","Neutrophilic_Crisis")
Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibrium_Neutrophilic$An_Equilibrium_0_Neutrophilic_Crisis_1 <-
ifelse(Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibrium_Neutrophilic$An_Equilibrium_V_Neutrophilic_Crisis=="An_Equilibrium",0,1)
set.seed(5783)
longtest_0.8_TRAIN_Equilibrium_Versus_Neutrophilic_Crisis_Symptoms_and_Neutrophilic_and_CRP_Indices <-
melt_roc(Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibrium_Neutrophilic, "An_Equilibrium_0_Neutrophilic_Crisis_1", c("VAS.Cough",
"VAS.Dypsonea", "VAS.Sputum.Prod", "VAS.Sputum.Purulence", "B.Neutrophils",
"Percentage_Blood_Neutrophils","CRP"))
ggplot(longtest_0.8_TRAIN_Equilibrium_Versus_Neutrophilic_Crisis_Symptoms_and_Neutrophilic_and_CRP_Indices, aes(d = D, m = M, color = name)) +
geom_roc(n.cuts = 0, increasing = TRUE) + style_roc()
auc_VAS.Cough_Equilibrium_Versus_Neutrophilic_Crisis <-
auc(Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibrium_Neutrophilic$An_Equilibrium_0_Neutrophilic_Crisis_1,
Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibrium_Neutrophilic$VAS.Cough)
auc_VAS.Dypsonea_Equilibrium_Versus_Neutrophilic_Crisis <-
auc(Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibrium_Neutrophilic$An_Equilibrium_0_Neutrophilic_Crisis_1,
Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_
```

```
Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibrium
_Neutrophilic$VAS.Dypsonea)
auc_VAS.Sputum.Prod_Equilibrium_Versus_Neutrophilic_Crisis <-
auc(Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRA
IN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibr
ium_Neutrophilic$An_Equilibrium_0_Neutrophilic_Crisis_1,
Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_
Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibrium
_Neutrophilic$VAS.Sputum.Prod)
auc_VAS.Sputum.Purulence_Equilibrium_Versus_Neutrophilic_Crisis <-
auc(Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRA
IN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibr
ium_Neutrophilic$An_Equilibrium_0_Neutrophilic_Crisis_1,
Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_
Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibrium
_Neutrophilic$VAS.Sputum.Purulence)
auc_B.Neutrophils_Equilibrium_Versus_Neutrophilic_Crisis <-
auc(Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRA
IN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibr
ium_Neutrophilic$An_Equilibrium_0_Neutrophilic_Crisis_1,
Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_
Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibrium
_Neutrophilic$B.Neutrophils)
auc_Percentage_Blood_Neutrophils_Equilibrium_Versus_Neutrophilic_Crisis <-
auc(Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRA
IN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibr
ium_Neutrophilic$An_Equilibrium_0_Neutrophilic_Crisis_1,
Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_
Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibrium
_Neutrophilic$Percentage_Blood_Neutrophils)
auc_CRP_Equilibrium_Versus_Neutrophilic_Crisis <-
auc(Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRA
IN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibr
ium_Neutrophilic$An_Equilibrium_0_Neutrophilic_Crisis_1,
Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_
Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibrium
_Neutrophilic$CRP)
roc_VAS.Sputum.Prod_Equilibrium_Versus_Neutrophilic_Crisis <-
pROC::roc(Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.
8_TRAIN_Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_
Equilibrium_Neutrophilic$An_Equilibrium_0_Neutrophilic_Crisis_1,
Just_VAS_CRP_Neutrophils_Eosinophils_Sputum_Blood_Imputed_All_0.8_TRAIN_
Samples_With_Objective_States_Two_Class_for_Single_Variable_ROC_Equilibrium
_Neutrophilic$VAS.Sputum.Prod)
coords_VAS.Sputum.Prod_Equilibrium_Versus_Neutrophilic_Crisis_return_threshold
<- pROC::coords(roc_VAS.Sputum.Prod_Equilibrium_Versus_Neutrophilic_Crisis,
"best", ret="threshold", best.method="closest.topleft", transpose = FALSE)
coords_VAS.Sputum.Prod_Equilibrium_Versus_Neutrophilic_Crisis_return_True_Po
sitive_Rate <-
```

```
pROC::coords(roc_VAS.Sputum.Prod_Equilibrium_Versus_Neutrophilic_Crisis,
"best", ret="tpr", best.method="closest.topleft", transpose = FALSE)
coords_VAS.Sputum.Prod_Equilibrium_Versus_Neutrophilic_Crisis_False_Positive_
Rate <-
pROC::coords(roc_VAS.Sputum.Prod_Equilibrium_Versus_Neutrophilic_Crisis,
"best", ret="fpr", best.method="closest.topleft", transpose = FALSE)
```

Identification of key variables for COPD state detection algorithm

```
set.seed(5783)
SES_Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_No_L
abel_No_Subject_Number_0.8_TRAIN <-
SES(Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_0.8_T
RAIN$Objective_State,
Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_No_Label_
No_Subject_Number_0.8_TRAIN, max_k = 9, threshold = 0.1, test = NULL, hash =
FALSE, hashObject = NULL)
SES_Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_No_L
abel_No_Subject_Number_0.8_TRAIN
SES_Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_No_L
abel_No_Subject_Number_0.8_TRAIN_pvalues <- c(-22.9595983, -6.2778871, -
0.9320295, -8.2324598, -29.4622027, -44.6647267, -6.2531863, -19.7806419, -
15.6010552)
SES_Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_No_L
abel_No_Subject_Number_0.8_TRAIN_pvalues <-
exp(SES_Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_N
o_Label_No_Subject_Number_0.8_TRAIN_pvalues)
```

Training, cross-validation and testing of COPD state detection algorithms

```
Just_Only_SES_Selected_Blood_and_Symptoms_0.8_TRAIN <-
Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_0.8_TRAIN
Just_Only_SES_Selected_Blood_and_Symptoms_0.8_TRAIN$Percentage_Blood_N
eutrophils <- NULL

set.seed(5783)
k.4_folds <- 4

length(index_Just_Only_SES_Selected_Blood_and_Symptoms_0.8_TRAIN_with_O
bjective_Disease_State_4_Fold_Cross_Validation_on_0.8_TRAIN$Fold1)
length(index_Just_Only_SES_Selected_Blood_and_Symptoms_0.8_TRAIN_with_O
bjective_Disease_State_4_Fold_Cross_Validation_on_0.8_TRAIN$Fold2)
length(index_Just_Only_SES_Selected_Blood_and_Symptoms_0.8_TRAIN_with_O
bjective_Disease_State_4_Fold_Cross_Validation_on_0.8_TRAIN$Fold3)
length(index_Just_Only_SES_Selected_Blood_and_Symptoms_0.8_TRAIN_with_O
bjective_Disease_State_4_Fold_Cross_Validation_on_0.8_TRAIN$Fold4)
```

```
set.seed(5783)
index_Just_Only_SES_Selected_Blood_and_Symptoms_0.8_TRAIN_with_Objective
_Disease_State_4_Fold_Cross_Validation_on_0.8_TRAIN <-
groupKFold(Just_Only_SES_Selected_Blood_and_Symptoms_0.8_TRAIN$Subject_
Number, k = k.4_folds)
set.seed(5783)
training_control_Just_Only_SES_Selected_Blood_and_Symptoms <-
trainControl(method = "cv",
                     summaryFunction = multiClassSummary,
                     classProbs = TRUE,
                     number = 4, repeats = NA, savePredictions = "all",
set.seed(5782), index =
index_Just_Only_SES_Selected_Blood_and_Symptoms_0.8_TRAIN_with_Objective
_Disease_State_4_Fold_Cross_Validation_on_0.8_TRAIN)

Just_Only_SES_Selected_Blood_and_Symptoms_0.8_TRAIN_No_Subject_Number
<- Just_Only_SES_Selected_Blood_and_Symptoms_0.8_TRAIN
Just_Only_SES_Selected_Blood_and_Symptoms_0.8_TRAIN_No_Subject_Number
$Subject_Number <- NULL

RandomForestGRid_Just_Only_SES_Selected_Blood_and_Symptoms_with_Objecti
ve_Disease_State_4_Fold_Cross_Validation_on_0.8_TRAIN <- expand.grid(.mtry =
c(1:8))

set.seed(5783)
randomforest_Just_Only_SES_Selected_Blood_and_Symptoms_with_Objective_Dis
ease_State_4_Fold_Cross_Validation_on_0.8_TRAIN_ntree_100 <-
train(Objective_State ~ .,

data =
Just_Only_SES_Selected_Blood_and_Symptoms_0.8_TRAIN_No_Subject_Number,

method = "rf",

trControl = training_control_Just_Only_SES_Selected_Blood_and_Symptoms,

metric = "Mean_Balanced_Accuracy",

ntree = 100,

tuneGrid =
RandomForestGRid_Just_Only_SES_Selected_Blood_and_Symptoms_with_Objecti
ve_Disease_State_4_Fold_Cross_Validation_on_0.8_TRAIN)

set.seed(5783)
```

```
randomforest_Just_Only_SES_Selected_Blood_and_Symptoms_with_Objective_Dis
ease_State_4_Fold_Cross_Validation_on_0.8_TRAIN_ntree_200 <-
train(Objective_State ~ .,

data =
Just_Only_SES_Selected_Blood_and_Symptoms_0.8_TRAIN_No_Subject_Number,

method = "rf",

trControl = training_control_Just_Only_SES_Selected_Blood_and_Symptoms,

metric = "Mean_Balanced_Accuracy",

ntree = 200,

tuneGrid =
RandomForestGRid_Just_Only_SES_Selected_Blood_and_Symptoms_with_Objecti
ve_Disease_State_4_Fold_Cross_Validation_on_0.8_TRAIN)

set.seed(5783)
randomforest_Just_Only_SES_Selected_Blood_and_Symptoms_with_Objective_Dis
ease_State_4_Fold_Cross_Validation_on_0.8_TRAIN_ntree_300 <-
train(Objective_State ~ .,

data =
Just_Only_SES_Selected_Blood_and_Symptoms_0.8_TRAIN_No_Subject_Number,

method = "rf",

trControl = training_control_Just_Only_SES_Selected_Blood_and_Symptoms,

metric = "Mean_Balanced_Accuracy",

ntree = 300,

tuneGrid =
RandomForestGRid_Just_Only_SES_Selected_Blood_and_Symptoms_with_Objecti
ve_Disease_State_4_Fold_Cross_Validation_on_0.8_TRAIN)

set.seed(5783)
randomforest_Just_Only_SES_Selected_Blood_and_Symptoms_with_Objective_Dis
ease_State_4_Fold_Cross_Validation_on_0.8_TRAIN_ntree_400 <-
train(Objective_State ~ .,

data =
Just_Only_SES_Selected_Blood_and_Symptoms_0.8_TRAIN_No_Subject_Number,

method = "rf",

trControl = training_control_Just_Only_SES_Selected_Blood_and_Symptoms,
```

```
metric = "Mean_Balanced_Accuracy",

ntree = 400,

tuneGrid =
RandomForestGRid_Just_Only_SES_Selected_Blood_and_Symptoms_with_Objecti
ve_Disease_State_4_Fold_Cross_Validation_on_0.8_TRAIN)


set.seed(5783)
randomforest_Just_Only_SES_Selected_Blood_and_Symptoms_with_Objective_Dis
ease_State_4_Fold_Cross_Validation_on_0.8_TRAIN_ntree_500 <-
train(Objective_State ~ .,

data =
Just_Only_SES_Selected_Blood_and_Symptoms_0.8_TRAIN_No_Subject_Number,

method = "rf",

trControl = training_control_Just_Only_SES_Selected_Blood_and_Symptoms,

metric = "Mean_Balanced_Accuracy",

ntree = 500,

tuneGrid =
RandomForestGRid_Just_Only_SES_Selected_Blood_and_Symptoms_with_Objecti
ve_Disease_State_4_Fold_Cross_Validation_on_0.8_TRAIN)

set.seed(5783)
randomforest_Just_Only_SES_Selected_Blood_and_Symptoms_with_Objective_Dis
ease_State_4_Fold_Cross_Validation_on_0.8_TRAIN_ntree_600 <-
train(Objective_State ~ .,

data =
Just_Only_SES_Selected_Blood_and_Symptoms_0.8_TRAIN_No_Subject_Number,

method = "rf",

trControl = training_control_Just_Only_SES_Selected_Blood_and_Symptoms,

metric = "Mean_Balanced_Accuracy",

ntree = 600,

tuneGrid =
RandomForestGRid_Just_Only_SES_Selected_Blood_and_Symptoms_with_Objecti
ve_Disease_State_4_Fold_Cross_Validation_on_0.8_TRAIN)
```

```
set.seed(5783)
randomforest_Just_Only_SES_Selected_Blood_and_Symptoms_with_Objective_Disease_State_4_Fold_Cross_Validation_on_0.8_TRAIN_ntree_700 <-
train(Objective_State ~ .,

data =
Just_Only_SES_Selected_Blood_and_Symptoms_0.8_TRAIN_No_Subject_Number,

method = "rf",

trControl = training_control_Just_Only_SES_Selected_Blood_and_Symptoms,

metric = "Mean_Balanced_Accuracy",

ntree = 700,

tuneGrid =
RandomForestGRid_Just_Only_SES_Selected_Blood_and_Symptoms_with_Objective_Disease_State_4_Fold_Cross_Validation_on_0.8_TRAIN)


set.seed(5783)
randomforest_Just_Only_SES_Selected_Blood_and_Symptoms_with_Objective_Disease_State_4_Fold_Cross_Validation_on_0.8_TRAIN_ntree_800 <-
train(Objective_State ~ .,

data =
Just_Only_SES_Selected_Blood_and_Symptoms_0.8_TRAIN_No_Subject_Number,

method = "rf",

trControl = training_control_Just_Only_SES_Selected_Blood_and_Symptoms,

metric = "Mean_Balanced_Accuracy",

ntree = 800,

tuneGrid =
RandomForestGRid_Just_Only_SES_Selected_Blood_and_Symptoms_with_Objective_Disease_State_4_Fold_Cross_Validation_on_0.8_TRAIN)


set.seed(5783)
randomforest_Just_Only_SES_Selected_Blood_and_Symptoms_with_Objective_Disease_State_4_Fold_Cross_Validation_on_0.8_TRAIN_ntree_900 <-
train(Objective_State ~ .,

data =
Just_Only_SES_Selected_Blood_and_Symptoms_0.8_TRAIN_No_Subject_Number,
```

```
method = "rf",

trControl = training_control_Just_Only_SES_Selected_Blood_and_Symptoms,

metric = "Mean_Balanced_Accuracy",

ntree = 900,

tuneGrid =
RandomForestGRid_Just_Only_SES_Selected_Blood_and_Symptoms_with_Objecti
ve_Disease_State_4_Fold_Cross_Validation_on_0.8_TRAIN)

set.seed(5783)
randomforest_Just_Only_SES_Selected_Blood_and_Symptoms_with_Objective_Dis
ease_State_4_Fold_Cross_Validation_on_0.8_TRAIN_ntree_1000 <-
train(Objective_State ~ .,

data =
Just_Only_SES_Selected_Blood_and_Symptoms_0.8_TRAIN_No_Subject_Number,

method = "rf",

trControl = training_control_Just_Only_SES_Selected_Blood_and_Symptoms,

metric = "Mean_Balanced_Accuracy",

ntree = 1000,

tuneGrid =
RandomForestGRid_Just_Only_SES_Selected_Blood_and_Symptoms_with_Objecti
ve_Disease_State_4_Fold_Cross_Validation_on_0.8_TRAIN)


set.seed(5783)
knn_cv_4_folds_Just_Only_SES_Selected_Blood_and_Symptoms_with_Objective_
Disease_State_4_Fold_Cross_Validation_on_0.8_TRAIN <- train(Objective_State ~
.,

data =
Just_Only_SES_Selected_Blood_and_Symptoms_0.8_TRAIN_No_Subject_Number,
method = "knn", trControl =
training_control_Just_Only_SES_Selected_Blood_and_Symptoms, metric =
"Mean_Balanced_Accuracy",tuneGrid = data.frame(k = seq(1,391,by = 2)))


set.seed(5783)
cart1_cv_4_folds_Just_Only_SES_Selected_Blood_and_Symptoms_with_Objective
_Disease_State_4_Fold_Cross_Validation_on_0.8_TRAIN <- train(Objective_State ~
.,
```

```
data =
Just_Only_SES_Selected_Blood_and_Symptoms_0.8_TRAIN_No_Subject_Number,

method = "rpart",

trControl = training_control_Just_Only_SES_Selected_Blood_and_Symptoms,

metric = "Mean_Balanced_Accuracy",

tuneLength = 10)

Just_Only_SES_Selected_Blood_and_Symptoms_0.8_TRAIN_No_Subject_Number
_No_Objective_State <-
Just_Only_SES_Selected_Blood_and_Symptoms_0.8_TRAIN_No_Subject_Number
Just_Only_SES_Selected_Blood_and_Symptoms_0.8_TRAIN_No_Subject_Number
_No_Objective_State$Objective_State <- NULL
Matrix_Just_Only_SES_Selected_Blood_and_Symptoms_0.8_TRAIN_No_Subject_
Number_No_Objective_State <-
as.matrix(Just_Only_SES_Selected_Blood_and_Symptoms_0.8_TRAIN_No_Subject
_Number_No_Objective_State)

set.seed(5783)
sigmaRange_Just_Only_SES_Selected_Blood_and_Symptoms_with_Objective_Dis
ease_State_4_Fold_Cross_Validation_on_0.8_TRAIN <-
sigest(Matrix_Just_Only_SES_Selected_Blood_and_Symptoms_0.8_TRAIN_No_Su
bject_Number_No_Objective_State)
set.seed(5783)
svmRGRid_Just_Only_SES_Selected_Blood_and_Symptoms_with_Objective_Disea
se_State_4_Fold_Cross_Validation_on_0.8_TRAIN <- expand.grid(.sigma =
seq(sigmaRange_Just_Only_SES_Selected_Blood_and_Symptoms_with_Objective
_Disease_State_4_Fold_Cross_Validation_on_0.8_TRAIN[1],sigmaRange_Just_Onl
y_SES_Selected_Blood_and_Symptoms_with_Objective_Disease_State_4_Fold_Cr
oss_Validation_on_0.8_TRAIN[3],by =
((sigmaRange_Just_Only_SES_Selected_Blood_and_Symptoms_with_Objective_Di
sease_State_4_Fold_Cross_Validation_on_0.8_TRAIN[3] -
sigmaRange_Just_Only_SES_Selected_Blood_and_Symptoms_with_Objective_Dis
ease_State_4_Fold_Cross_Validation_on_0.8_TRAIN[1])/8)), .C = 2^(seq(-4,4)))

set.seed(5783)
SVM_Radial_Just_Only_SES_Selected_Blood_and_Symptoms_with_Objective_Dis
ease_State_4_Fold_Cross_Validation_on_0.8_TRAIN <- train(Objective_State ~ .,

data =
Just_Only_SES_Selected_Blood_and_Symptoms_0.8_TRAIN_No_Subject_Number,

method = "svmRadial",

trControl = training_control_Just_Only_SES_Selected_Blood_and_Symptoms,
```

```
metric = "Mean_Balanced_Accuracy",

preProcess = c("center", "scale"),

fit = FALSE,

tuneGrid =
svmRGRid_Just_Only_SES_Selected_Blood_and_Symptoms_with_Objective_Disea
se_State_4_Fold_Cross_Validation_on_0.8_TRAIN)


set.seed(5783)
nnetRGRid_Just_Only_SES_Selected_Blood_and_Symptoms_with_Objective_Disea
se_State_4_Fold_Cross_Validation_on_0.8_TRAIN <- expand.grid(.size = 1:10,
.decay = seq(0, 2, by = 0.1))
maxSize_Just_Only_SES_Selected_Blood_and_Symptoms_with_Objective_Disease
_State_4_Fold_Cross_Validation_on_0.8_TRAIN <-
max(nnetRGRid_Just_Only_SES_Selected_Blood_and_Symptoms_with_Objective_
Disease_State_4_Fold_Cross_Validation_on_0.8_TRAIN$.size)
numWts_Just_Only_SES_Selected_Blood_and_Symptoms_with_Objective_Disease
_State_4_Fold_Cross_Validation_on_0.8_TRAIN <-
4*(maxSize_Just_Only_SES_Selected_Blood_and_Symptoms_with_Objective_Dise
ase_State_4_Fold_Cross_Validation_on_0.8_TRAIN *
(ncol(Just_Only_SES_Selected_Blood_and_Symptoms_0.8_TRAIN_No_Subject_Nu
mber_No_Objective_State) + 1) +
maxSize_Just_Only_SES_Selected_Blood_and_Symptoms_with_Objective_Disease
_State_4_Fold_Cross_Validation_on_0.8_TRAIN + 1)
set.seed(5783)
nnet_Just_Only_SES_Selected_Blood_and_Symptoms_with_Objective_Disease_St
ate_4_Fold_Cross_Validation_on_0.8_TRAIN <- train(Objective_State ~ .,

data =
Just_Only_SES_Selected_Blood_and_Symptoms_0.8_TRAIN_No_Subject_Number,

method = "nnet",

trControl = training_control_Just_Only_SES_Selected_Blood_and_Symptoms,

metric = "Mean_Balanced_Accuracy",

preProcess = c("center", "scale", "spatialSign"),

tuneGrid =
nnetRGRid_Just_Only_SES_Selected_Blood_and_Symptoms_with_Objective_Disea
se_State_4_Fold_Cross_Validation_on_0.8_TRAIN,

trace = FALSE,

maxit = 2000,
```

```
MaxNWts =
numWts_Just_Only_SES_Selected_Blood_and_Symptoms_with_Objective_Disease
_State_4_Fold_Cross_Validation_on_0.8_TRAIN)

Just_Only_SES_Selected_Blood_and_Symptoms_Imputed_All_Samples_0.2_TEST
<-
Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_0.2_TEST
Just_Only_SES_Selected_Blood_and_Symptoms_Imputed_All_Samples_0.2_TEST
$Percentage_Blood_Neutrophils <- NULL
Just_Only_SES_Selected_Blood_and_Symptoms_Imputed_All_Samples_0.2_TEST
$Subject_Number <- NULL

Just_Only_SES_Selected_Blood_and_Symptoms_Imputed_All_Samples_0.2_TEST
_No_Objective_State <-
Just_Only_SES_Selected_Blood_and_Symptoms_Imputed_All_Samples_0.2_TEST
Just_Only_SES_Selected_Blood_and_Symptoms_Imputed_All_Samples_0.2_TEST
_No_Objective_State$Objective_State <- NULL

set.seed(5783)
test_set_Just_Only_SES_Selected_Blood_and_Symptoms_Imputed_All_Samples_n
net <-
predict(nnet_Just_Only_SES_Selected_Blood_and_Symptoms_with_Objective_Dise
ase_State_4_Fold_Cross_Validation_on_0.8_TRAIN,
Just_Only_SES_Selected_Blood_and_Symptoms_Imputed_All_Samples_0.2_TEST
_No_Objective_State)
Just_Only_SES_Selected_Blood_and_Symptoms_Imputed_All_Samples_0.2_TEST
_No_Objective_State_predicted_and_actual_states <-
Just_Only_SES_Selected_Blood_and_Symptoms_Imputed_All_Samples_0.2_TEST
_No_Objective_State
Just_Only_SES_Selected_Blood_and_Symptoms_Imputed_All_Samples_0.2_TEST
_No_Objective_State_predicted_and_actual_states$Actual_State <-
Just_Only_SES_Selected_Blood_and_Symptoms_Imputed_All_Samples_0.2_TEST
$Objective_State
Just_Only_SES_Selected_Blood_and_Symptoms_Imputed_All_Samples_0.2_TEST
_No_Objective_State_predicted_and_actual_states$Predicted_State_nnet <-
test_set_Just_Only_SES_Selected_Blood_and_Symptoms_Imputed_All_Samples_n
net
Just_Only_SES_Selected_Blood_and_Symptoms_Imputed_All_Samples_0.2_TEST
_No_Objective_State_predicted_and_actual_states$Predicted_State <- NULL
```

Assessment of COPD state detection algorithm in AWARD cohort from chapter 3 of thesis

```
Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_Ready_for_
Data_Splitting_No_Subject_Number <-
Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_Ready_for_
Data_Splitting
```

```
Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_Ready_for_Data_Splitting_No_Subject_Number$Subject_Number <- NULL
Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_Ready_for_Data_Splitting_No_Subject_Number_Only_SES_Selected <- Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_Ready_for_Data_Splitting_No_Subject_Number
Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_Ready_for_Data_Splitting_No_Subject_Number_Only_SES_Selected$Percentage_Blood_Neutrophils <- NULL



set.seed(5783)
nnet_Just_Only_SES_Selected_Blood_and_Symptoms_with_Objective_Disease_State_4_Fold_Cross_Validation_on_Full_Data <- train(Objective_State ~ .,
                                                                                                                          data =
Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_Ready_for_Data_Splitting_No_Subject_Number_Only_SES_Selected,
                                                                                                                          method
= "nnet",
                                                                                                                          metric =
"Mean_Balanced_Accuracy",

preProcess = c("center", "scale", "spatialSign"),
                                                                                                                          trace =
FALSE,
                                                                                                                          maxit =
2000,

tuneGrid=expand.grid(.size=9,.decay=0.2))



setwd("~/OneDrive - Nexus365/Thesis Write-Up/A Halner DPhil Thesis Draft/2022/AWARD/AWARD/Week Beginning 11th September")
MissForest_Imputed_All_Variables_Complete_AWARD_101_Patients_Variables_Cleaned_to_Remove_Non_Sensical_Values_Superfluous_Variables_Removed_No_Treatment_Today_No_Treatment_Outcome_No_AWARD_ID_Corrected_To_Prevent_Empty_String_Categories <-
read.csv(file="MissForest_Imputed_All_Variables_Complete_AWARD_101_Patients_Variables_Cleaned_to_Remove_Non_Sensical_Values_Superfluous_Variables_Removed_No_Treatment_Today_No_Treatment_Outcome_No_AWARD_ID_Corrected_To_Prevent_Empty_String_Categories.csv", sep=",", header = T)
AWARD_101_Symptoms_and_Blood_Biology_8_Variables <-
MissForest_Imputed_All_Variables_Complete_AWARD_101_Patients_Variables_Cleaned_to_Remove_Non_Sensical_Values_Superfluous_Variables_Removed_No_Treatment_Today_No_Treatment_Outcome_No_AWARD_ID_Corrected_To_Prevent_Empty_String_Categories
AWARD_101_Symptoms_and_Blood_Biology_8_Variables <-
AWARD_101_Symptoms_and_Blood_Biology_8_Variables[,c("VAS_Cough",
```

```
"VAS_Breathlessness", "VAS_SputumProduction", "VAS_SputumDiscolouration",
"CRP", "Neutrophils", "Eosinophils", "Eosinophil_Percentage")]
names(Just_VAS_Blood_CRP_Neutrophils_Eosinophils_Imputed_All_Samples_Rea
dy_for_Data_Splitting_No_Subject_Number_Only_SES_Selected)
names(AWARD_101_Symptoms_and_Blood_Biology_8_Variables)
AWARD_101_Symptoms_and_Blood_Biology_8_Variables$B.Neutrophils <-
AWARD_101_Symptoms_and_Blood_Biology_8_Variables$Neutrophils
AWARD_101_Symptoms_and_Blood_Biology_8_Variables$B.Eosinophils <-
AWARD_101_Symptoms_and_Blood_Biology_8_Variables$Eosinophils
AWARD_101_Symptoms_and_Blood_Biology_8_Variables$Percentage_Blood_Eosi
nophils <-
AWARD_101_Symptoms_and_Blood_Biology_8_Variables$Eosinophil_Percentage
AWARD_101_Symptoms_and_Blood_Biology_8_Variables$VAS.Cough <-
AWARD_101_Symptoms_and_Blood_Biology_8_Variables$VAS_Cough
AWARD_101_Symptoms_and_Blood_Biology_8_Variables$VAS.Dypsonea <-
AWARD_101_Symptoms_and_Blood_Biology_8_Variables$VAS_Breathlessness
AWARD_101_Symptoms_and_Blood_Biology_8_Variables$VAS.Sputum.Prod <-
AWARD_101_Symptoms_and_Blood_Biology_8_Variables$VAS_SputumProduction
AWARD_101_Symptoms_and_Blood_Biology_8_Variables$VAS.Sputum.Purulence
<-
AWARD_101_Symptoms_and_Blood_Biology_8_Variables$VAS_SputumDiscoloura
tion
names(AWARD_101_Symptoms_and_Blood_Biology_8_Variables)
AWARD_101_Symptoms_and_Blood_Biology_8_Variables[,1:4] <- NULL
names(AWARD_101_Symptoms_and_Blood_Biology_8_Variables)
AWARD_101_Symptoms_and_Blood_Biology_8_Variables[,2:4] <- NULL
names(AWARD_101_Symptoms_and_Blood_Biology_8_Variables)

set.seed(5783)
Just_SES_8_Variable_Objective_State_Prediction_Neural_Network_Algorithm_Used
_for_AWARD_101_Patients <-
predict(nnet_Just_Only_SES_Selected_Blood_and_Symptoms_with_Objective_Dise
ase_State_4_Fold_Cross_Validation_on_Full_Data,
AWARD_101_Symptoms_and_Blood_Biology_8_Variables)

AWARD_101_Symptoms_and_Blood_Biology_8_Variables_with_Objective_State_L
abel_Based_on_MRC_1348_Events_Derived_Objective_State_Neural_Network <-
AWARD_101_Symptoms_and_Blood_Biology_8_Variables
AWARD_101_Symptoms_and_Blood_Biology_8_Variables_with_Objective_State_L
abel_Based_on_MRC_1348_Events_Derived_Objective_State_Neural_Network$Obj
ective_State <-
Just_SES_8_Variable_Objective_State_Prediction_Neural_Network_Algorithm_Used
_for_AWARD_101_Patients
table(AWARD_101_Symptoms_and_Blood_Biology_8_Variables_with_Objective_St
ate_Label_Based_on_MRC_1348_Events_Derived_Objective_State_Neural_Networ
k$Objective_State)

summary(AWARD_101_Symptoms_and_Blood_Biology_8_Variables_with_Objective
_State_Label_Based_on_MRC_1348_Events_Derived_Objective_State_Neural_Net
work[AWARD_101_Symptoms_and_Blood_Biology_8_Variables_with_Objective_Sta
```

```
te_Label_Based_on_MRC_1348_Events_Derived_Objective_State_Neural_Network
$Objective_State=="Equilibrium",])
summary(AWARD_101_Symptoms_and_Blood_Biology_8_Variables_with_Objective
_State_Label_Based_on_MRC_1348_Events_Derived_Objective_State_Neural_Net
work[AWARD_101_Symptoms_and_Blood_Biology_8_Variables_with_Objective_Sta
te_Label_Based_on_MRC_1348_Events_Derived_Objective_State_Neural_Network
$Objective_State=="Symptomatic_Crisis",])
summary(AWARD_101_Symptoms_and_Blood_Biology_8_Variables_with_Objective
_State_Label_Based_on_MRC_1348_Events_Derived_Objective_State_Neural_Net
work[AWARD_101_Symptoms_and_Blood_Biology_8_Variables_with_Objective_Sta
te_Label_Based_on_MRC_1348_Events_Derived_Objective_State_Neural_Network
$Objective_State=="Eosinophilic_Crisis",])
summary(AWARD_101_Symptoms_and_Blood_Biology_8_Variables_with_Objective
_State_Label_Based_on_MRC_1348_Events_Derived_Objective_State_Neural_Net
work[AWARD_101_Symptoms_and_Blood_Biology_8_Variables_with_Objective_Sta
te_Label_Based_on_MRC_1348_Events_Derived_Objective_State_Neural_Network
$Objective_State=="Neutrophilic_Crisis",])

AWARD_101_Patients_AWARD_ID_and_Different_Treatment_Outcomes <-
read.csv(file="AWARD_101_Patients_AWARD_ID_and_Different_Treatment_Outco
mes.csv", sep=",", header=T)
All_Variables_Complete_AWARD_101_Patients_Variables_Cleaned_to_Remove_N
on_Sensical_Values_Superfluous_Variables_Removed_No_Treatment_Today_No_
Treatment_Outcome_No_AWARD_ID <-
read.csv(file="All_Variables_Complete_AWARD_101_Patients_Variables_Cleaned_t
o_Remove_Non_Sensical_Values_Superfluous_Variables_Removed_No_Treatment
_Today_No_Treatment_Outcome_No_AWARD_ID.csv", sep=",", header = T)
AWARD_101_Patients_AWARD_ID_and_Different_Treatment_Outcomes_with_Obje
ctive_State_Label_Based_on_MRC_1348_Events_Derived_Objective_State_Neural
_Network <-
AWARD_101_Patients_AWARD_ID_and_Different_Treatment_Outcomes
All_Variables_Complete_AWARD_101_Patients_Variables_Cleaned_to_Remove_N
on_Sensical_Values_Superfluous_Variables_Removed_No_Treatment_Today_No_
Treatment_Outcome_No_AWARD_ID_with_Objective_State_Label_Based_on_MRC
_1348_Events_Derived_Objective_State_Neural_Network <-
All_Variables_Complete_AWARD_101_Patients_Variables_Cleaned_to_Remove_N
on_Sensical_Values_Superfluous_Variables_Removed_No_Treatment_Today_No_
Treatment_Outcome_No_AWARD_ID
All_Variables_AWARD_101_Patients_with_Objective_State_Label_Based_on_MRC
_1348_Events_Derived_Objective_State_Neural_Network <-
All_Variables_Complete_AWARD_101_Patients_Variables_Cleaned_to_Remove_N
on_Sensical_Values_Superfluous_Variables_Removed_No_Treatment_Today_No_
Treatment_Outcome_No_AWARD_ID_with_Objective_State_Label_Based_on_MRC
_1348_Events_Derived_Objective_State_Neural_Network
AWARD_101_Patients_AWARD_ID_and_Different_Treatment_Outcomes_with_Obje
ctive_State_Label_Based_on_MRC_1348_Events_Derived_Objective_State_Neural
_Network$Objective_State <-
AWARD_101_Symptoms_and_Blood_Biology_8_Variables_with_Objective_State_L
abel_Based_on_MRC_1348_Events_Derived_Objective_State_Neural_Network$Obj
ective_State
```

```
All_Variables_AWARD_101_Patients_with_Objective_State_Label_Based_on_MRC
_1348_Events_Derived_Objective_State_Neural_Network$Objective_State <-
AWARD_101_Symptoms_and_Blood_Biology_8_Variables_with_Objective_State_L
abel_Based_on_MRC_1348_Events_Derived_Objective_State_Neural_Network$Obj
ective_State
```

Appendix chapter 5 code

R libraries used in analysis

```
library(missForest)
library(ggplot2)
library(dendextend)
library(dplyr)
library(sm)
library(NbClust)
library(factoextra)
library(rgl)
library(car)
library(caret)
library(MLmetrics)
library(plotROC)
library(kernlab)
library(nnet)
library(MLeval)
library(plotROC)
library(PMCMRplus)
library(dtwclust)
library(TSclust)
library(imputeTS)
```

PCA analysis and exacerbation minus day 14 post-treatment trajectotype identification

```
PCA_Analysis_of_Imputed_Exacerbation_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events <-
prcomp(Imputed_Exacerbation_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events, retx = TRUE, center = TRUE, scale = TRUE)
summary(PCA_Analysis_of_Imputed_Exacerbation_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events)
PCA_Analysis_of_Imputed_Exacerbation_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events$rotation
Principal_Components_Representing_Imputed_Exacerbation_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events <-
PCA_Analysis_of_Imputed_Exacerbation_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events$x
Five_Key_Principal_Components_Representing_Imputed_Exacerbation_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events <-
Principal_Components_Representing_Imputed_Exacerbation_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events[,1:5]
Dist_Five_Key_Principal_Components_Representing_Imputed_Exacerbation_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events <-
dist(Five_Key_Principal_Components_Representing_Imputed_Exacerbation_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events,
method="euclidean")
```

```
duda_ward.D2_Five_Key_Principal_Components_Representing_Imputed_Exacerbation_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events <- NbClust(Five_Key_Principal_Components_Representing_Imputed_Exacerbation_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events, distance = "euclidean", method = "ward.D2", max.nc = 11, index = "duda")
pseudot2_ward.D2_Five_Key_Principal_Components_Representing_Imputed_Exacerbation_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events <- NbClust(Five_Key_Principal_Components_Representing_Imputed_Exacerbation_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events, distance = "euclidean", method = "ward.D2", max.nc = 11, index = "pseudot2")
duda_ward.D2_Five_Key_Principal_Components_Representing_Imputed_Exacerbation_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events$Best.nc
pseudot2_ward.D2_Five_Key_Principal_Components_Representing_Imputed_Exacerbation_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events$Best.nc
duda_ward.D2_Five_Key_Principal_Components_Representing_Imputed_Exacerbation_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events$Best.partition

Imputed_Exacerbation_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events_with_Exacerbation_Minus_Week_2_Post_Treatment_Trajectotype <- Imputed_Exacerbation_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events
Imputed_Exacerbation_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events_with_Exacerbation_Minus_Week_2_Post_Treatment_Trajectotype$Exacerbation_Minus_Week_2_Trajectotype <- duda_ward.D2_Five_Key_Principal_Components_Representing_Imputed_Exacerbation_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events$Best.partition
Imputed_Exacerbation_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events_with_Exacerbation_Minus_Week_2_Post_Treatment_Trajectotype$Exacerbation_Minus_Week_2_Trajectotype <- as.factor(Imputed_Exacerbation_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events_with_Exacerbation_Minus_Week_2_Post_Treatment_Trajectotype$Exacerbation_Minus_Week_2_Trajectotype)



duda_average_Five_Key_Principal_Components_Representing_Imputed_Exacerbation_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events <- NbClust(Five_Key_Principal_Components_Representing_Imputed_Exacerbation_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events, distance = "euclidean", method = "average", max.nc = 11, index = "duda")

duda_average_Five_Key_Principal_Components_Representing_Imputed_Exacerbation_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events$Best.nc

duda_mcquitty_Five_Key_Principal_Components_Representing_Imputed_Exacerbation_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events <- NbClust(Five_Key_Principal_Components_Representing_Imputed_Exacerbation_Mi
```

```
nus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events, distance =
"euclidean", method = "mcquitty", max.nc = 11, index = "duda")

duda_mcquitty_Five_Key_Principal_Components_Representing_Imputed_Exacerbat
ion_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events$Best.nc

duda_median_Five_Key_Principal_Components_Representing_Imputed_Exacerbati
on_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events <-
NbClust(Five_Key_Principal_Components_Representing_Imputed_Exacerbation_Mi
nus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events, distance =
"euclidean", method = "median", max.nc = 11, index = "duda")

duda_median_Five_Key_Principal_Components_Representing_Imputed_Exacerbati
on_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events$Best.nc

duda_centroid_Five_Key_Principal_Components_Representing_Imputed_Exacerbati
on_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events <-
NbClust(Five_Key_Principal_Components_Representing_Imputed_Exacerbation_Mi
nus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events, distance =
"euclidean", method = "centroid", max.nc = 11, index = "duda")

duda_centroid_Five_Key_Principal_Components_Representing_Imputed_Exacerbati
on_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events$Best.nc

pseudot2_average_Five_Key_Principal_Components_Representing_Imputed_Exace
rbation_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events <-
NbClust(Five_Key_Principal_Components_Representing_Imputed_Exacerbation_Mi
nus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events, distance =
"euclidean", method = "average", max.nc = 11, index = "pseudot2")

pseudot2_average_Five_Key_Principal_Components_Representing_Imputed_Exace
rbation_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events$Best
.nc

pseudot2_mcquitty_Five_Key_Principal_Components_Representing_Imputed_Exace
rbation_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events <-
NbClust(Five_Key_Principal_Components_Representing_Imputed_Exacerbation_Mi
nus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events, distance =
"euclidean", method = "mcquitty", max.nc = 11, index = "pseudot2")

pseudot2_mcquitty_Five_Key_Principal_Components_Representing_Imputed_Exace
rbation_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events$Best
.nc

pseudot2_median_Five_Key_Principal_Components_Representing_Imputed_Exacer
bation_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events <-
NbClust(Five_Key_Principal_Components_Representing_Imputed_Exacerbation_Mi
nus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events, distance =
"euclidean", method = "median", max.nc = 11, index = "pseudot2")
```

```
pseudot2_median_Five_Key_Principal_Components_Representing_Imputed_Exacerbation_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events$Best.nc

pseudot2_centroid_Five_Key_Principal_Components_Representing_Imputed_Exacerbation_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events <- NbClust(Five_Key_Principal_Components_Representing_Imputed_Exacerbation_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events, distance = "euclidean", method = "centroid", max.nc = 11, index = "pseudot2")

pseudot2_centroid_Five_Key_Principal_Components_Representing_Imputed_Exacerbation_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events$Best.nc


hclust_ward.D2_Five_Key_Principal_Components_Representing_Imputed_Exacerbation_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events <- hclust(Dist_Five_Key_Principal_Components_Representing_Imputed_Exacerbation_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events, method = 'ward.D2')
plot(hclust_ward.D2_Five_Key_Principal_Components_Representing_Imputed_Exacerbation_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events)
cut_ward.D2_Five_Key_Principal_Components_Representing_Imputed_Exacerbation_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events <- cutree(hclust_ward.D2_Five_Key_Principal_Components_Representing_Imputed_Exacerbation_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events, k=4)


ward.D2_dend_obj_Five_Key_Principal_Components_Representing_Imputed_Exacerbation_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events <- as.dendrogram(hclust_ward.D2_Five_Key_Principal_Components_Representing_Imputed_Exacerbation_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events)
ward.D2_col_dend_Five_Key_Principal_Components_Representing_Imputed_Exacerbation_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events <- color_branches(ward.D2_dend_obj_Five_Key_Principal_Components_Representing_Imputed_Exacerbation_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events, k = 4)
plot(ward.D2_col_dend_Five_Key_Principal_Components_Representing_Imputed_Exacerbation_Minus_Week_2_Symptoms_Blood_Sputum_Inflammation_347_Events)

indices_representing_variables_to_convert_to_factor_in_files_named_Condensed_Potential_Trajectotype_Predictors_WITHOUT_Trajectotype_Outcome_Subject_Number <- c(1,2,7:9,11:27,59:62,69)
Symptomatic_Crises_Condensed_Potential_Trajectotype_Predictors_WITHOUT_Trajectotype_Outcome_Subject_Number[,indices_representing_variables_to_convert_to_factor_in_files_named_Condensed_Potential_Trajectotype_Predictors_WITHOUT_Trajectotype_Outcome_Subject_Number] <- lapply(Symptomatic_Crises_Condensed_Potential_Trajectotype_Predictors_WITHO
```

```
UT_Trajectotype_Outcome_Subject_Number[,indices_representing_variables_to_co
nvert_to_factor_in_files_named_Condensed_Potential_Trajectotype_Predictors_WIT
HOUT_Trajectotype_Outcome_Subject_Number],factor)
Eosinophilic_Crises_Condensed_Potential_Trajectotype_Predictors_WITHOUT_Traj
ectotype_Outcome_Subject_Number[,indices_representing_variables_to_convert_to
_factor_in_files_named_Condensed_Potential_Trajectotype_Predictors_WITHOUT_
Trajectotype_Outcome_Subject_Number] <-
lapply(Eosinophilic_Crises_Condensed_Potential_Trajectotype_Predictors_WITHOU
T_Trajectotype_Outcome_Subject_Number[,indices_representing_variables_to_conv
ert_to_factor_in_files_named_Condensed_Potential_Trajectotype_Predictors_WITH
OUT_Trajectotype_Outcome_Subject_Number],factor)
Neutrophilic_Crisis_Condensed_Potential_Trajectotype_Predictors_WITHOUT_Traje
ctotype_Outcome_Subject_Number[,indices_representing_variables_to_convert_to_f
actor_in_files_named_Condensed_Potential_Trajectotype_Predictors_WITHOUT_Tr
ajectotype_Outcome_Subject_Number] <-
lapply(Neutrophilic_Crisis_Condensed_Potential_Trajectotype_Predictors_WITHOUT
_Trajectotype_Outcome_Subject_Number[,indices_representing_variables_to_conve
rt_to_factor_in_files_named_Condensed_Potential_Trajectotype_Predictors_WITHO
UT_Trajectotype_Outcome_Subject_Number],factor)

Symptomatic_Crises_Condensed_Potential_Trajectotype_Predictors_WITHOUT_Tra
jectotype_Outcome_Subject_Number[Symptomatic_Crises_Condensed_Potential_Tr
ajectotype_Predictors_WITHOUT_Trajectotype_Outcome_Subject_Number == ""] <-
NA
Eosinophilic_Crises_Condensed_Potential_Trajectotype_Predictors_WITHOUT_Traj
ectotype_Outcome_Subject_Number[Eosinophilic_Crises_Condensed_Potential_Tra
jectotype_Predictors_WITHOUT_Trajectotype_Outcome_Subject_Number == ""] <-
NA
Neutrophilic_Crisis_Condensed_Potential_Trajectotype_Predictors_WITHOUT_Traje
ctotype_Outcome_Subject_Number[Neutrophilic_Crisis_Condensed_Potential_Traje
ctotype_Predictors_WITHOUT_Trajectotype_Outcome_Subject_Number == ""] <- NA
```

Appropriate resolution trajectotype achieved versus not achieved analysis – identification of variables important in multivariable context via SES algorithm and development of classifiers for prediction

```
set.seed(5783)
k.folds <- 5
k.4_folds <- 4
set.seed(5783)

Imputed_Symptomatic_Crises_Condensed_Potential_Trajectotype_Predictors_No_Symptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$Sex <- as.factor(Imputed_Symptomatic_Crises_Condensed_Potential_Trajectotype_Predictors_No_Symptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$Sex)
Imputed_Symptomatic_Crises_Condensed_Potential_Trajectotype_Predictors_No_Symptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$Race <- as.factor(Imputed_Symptomatic_Crises_Condensed_Potential_Trajectotype_Predictors_No_Symptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$Race)
Imputed_Symptomatic_Crises_Condensed_Potential_Trajectotype_Predictors_No_Symptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$Smoking.Status <- as.factor(Imputed_Symptomatic_Crises_Condensed_Potential_Trajectotype_Predictors_No_Symptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$Smoking.Status)
Imputed_Symptomatic_Crises_Condensed_Potential_Trajectotype_Predictors_No_Symptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$qHI <- as.factor(Imputed_Symptomatic_Crises_Condensed_Potential_Trajectotype_Predictors_No_Symptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$qHI)
Imputed_Symptomatic_Crises_Condensed_Potential_Trajectotype_Predictors_No_Symptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$qSA <- as.factor(Imputed_Symptomatic_Crises_Condensed_Potential_Trajectotype_Predictors_No_Symptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$qSA)
Imputed_Symptomatic_Crises_Condensed_Potential_Trajectotype_Predictors_No_Symptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$qSP <- as.factor(Imputed_Symptomatic_Crises_Condensed_Potential_Trajectotype_Predictors_No_Symptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$qSP)
Imputed_Symptomatic_Crises_Condensed_Potential_Trajectotype_Predictors_No_Symptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$qMC <- as.factor(Imputed_Symptomatic_Crises_Condensed_Potential_Trajectotype_Predictors_No_Symptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$qMC)
Imputed_Symptomatic_Crises_Condensed_Potential_Trajectotype_Predictors_No_Symptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$Combination_of_Objective_State_and_Treatment_Allocation <- as.factor(Imputed_Symptomatic_Crises_Condensed_Potential_Trajectotype_Predict
```

```
ors_No_Symptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$Combination_of_Objective_State_and_Treatment_Allocation)
Imputed_Symptomatic_Crises_Condensed_Potential_Trajectotype_Predictors_No_Symptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$Personalised_Trajectotype_Verdict <- as.factor(Imputed_Symptomatic_Crises_Condensed_Potential_Trajectotype_Predictors_No_Symptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$Personalised_Trajectotype_Verdict)

index_Symptomatic_Crises_Trajectotype_Prediction_Ready_for_Data_Splitting <- groupKFold(Imputed_Symptomatic_Crises_Condensed_Potential_Trajectotype_Predictors_No_Symptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$Subject_Number, k = k.folds)

length(index_Symptomatic_Crises_Trajectotype_Prediction_Ready_for_Data_Splitting$Fold5)

Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_No_Symptoms_Inflammation <- Imputed_Symptomatic_Crises_Condensed_Potential_Trajectotype_Predictors_No_Symptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number[index_Symptomatic_Crises_Trajectotype_Prediction_Ready_for_Data_Splitting$Fold5,]
Symptomatic_Crises_Trajectotype_Prediction_0.2_TEST_No_Symptoms_Inflammation <- Imputed_Symptomatic_Crises_Condensed_Potential_Trajectotype_Predictors_No_Symptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number[-index_Symptomatic_Crises_Trajectotype_Prediction_Ready_for_Data_Splitting$Fold5,]
Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_No_Symptoms_Inflammation_No_Label_No_Subject_Number <- Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_No_Symptoms_Inflammation
Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_No_Symptoms_Inflammation_No_Label_No_Subject_Number$Personalised_Trajectotype_Verdict <- NULL
Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_No_Symptoms_Inflammation_No_Label_No_Subject_Number$Subject_Number <- NULL


set.seed(5783)

Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_No_Symptoms_Inflammation$Personalised_Trajectotype_Verdict <- as.factor(Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_No_Symptoms_Inflammation$Personalised_Trajectotype_Verdict)
SES_max_k_32_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_No_Symptoms_Inflammation_No_Label_No_Subject_Number <- SES(Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_No_Symptoms_Inflammation$Personalised_Trajectotype_Verdict,
Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_No_Symptoms_Inflamma
```

```
tion_No_Label_No_Subject_Number, max_k = 32, threshold = 0.1, test = NULL, hash
= FALSE, hashObject = NULL)
SES_max_k_32_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_No_Sy
mptoms_Inflammation_No_Label_No_Subject_Number

SES_max_k_32_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_No_Sy
mptoms_Inflammation_No_Label_No_Subject_Number_pvalues <- c(-2.71245718, -
4.35862951)
SES_max_k_32_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_No_Sy
mptoms_Inflammation_No_Label_No_Subject_Number_pvalues <-
exp(SES_max_k_32_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_No
_Symptoms_Inflammation_No_Label_No_Subject_Number_pvalues)


Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables <-
Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_No_Symptoms_Inflamma
tion
Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables <-
Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables[,c("FEV1.post.Bronchodilator", "qHI")]
Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables_No_Subject_Number <-
Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables
Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables_No_Subject_Number_No_Personalised_Trajectotype_Verdict <-
Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables_No_Subject_Number
Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables_No_Subject_Number_WIth_Personalised_Trajectotype_Verdict <-
Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables_No_Subject_Number_No_Personalised_Trajectotype_Verdict
Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables_No_Subject_Number_WIth_Personalised_Trajectotype_Verdict$Personalised
_Trajectotype_Verdict <-
Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_No_Symptoms_Inflamma
tion$Personalised_Trajectotype_Verdict
Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables_No_Subject_Number_WIth_Personalised_Trajectotype_Verdict$Personalised
_Trajectotype_Verdict <-
as.factor(Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Sele
cted_Variables_No_Subject_Number_WIth_Personalised_Trajectotype_Verdict$Pers
onalised_Trajectotype_Verdict)
Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables$Subject_Number <-
Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_No_Symptoms_Inflamma
tion$Subject_Number
```

```
set.seed(5783)
index_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables <-
groupKFold(Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables$Subject_Number, k = k.4_folds)

set.seed(5783)
training_control_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables <- trainControl(method = "cv",

summaryFunction = twoClassSummary,
                                                                    classProbs =
TRUE,
                                                                    number = 4,
repeats = NA, savePredictions = "all", set.seed(5783), index =
index_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables)

set.seed(5783)
cart1_cv_4_folds_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables <- train(Personalised_Trajectotype_Verdict ~ .,
                                                                    data =
Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_No_Subject_Number_WIth_Personalised_Trajectotype_Verdict,
                                                                    method = "rpart",
                                                                    trControl =
training_control_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables,
                                                                    metric = "ROC",
                                                                    tuneLength = 10)


Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_No_Subject_Number_Categories_Numerical_WIth_Personalised_Trajectotype_Verdict <-
Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_No_Subject_Number_WIth_Personalised_Trajectotype_Verdict
Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_No_Subject_Number_Categories_Numerical_WIth_Personalised_Trajectotype_Verdict$qHI <-
ifelse(Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_No_Subject_Number_Categories_Numerical_WIth_Personalised_Trajectotype_Verdict$qHI=="Positive",1,0)


set.seed(5783)
knn_cv_4_folds_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables <- train(Personalised_Trajectotype_Verdict ~ .,

data =
```

```
Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_No_Subject_Number_Categories_Numerical_WIth_Personalised_Trajectotype_Verdict, method = "knn", trControl = training_control_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables, metric = "ROC",tuneGrid = data.frame(k = seq(1,131,by = 2)))


RandomForestGRid_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables <- expand.grid(.mtry = c(1:2))

set.seed(5783)
randomforest_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_ntree_100 <- train(Personalised_Trajectotype_Verdict ~ .,
                                                                                                                  data = Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_No_Subject_Number_WIth_Personalised_Trajectotype_Verdict,
                                                                                                                  method = "rf",
                                                                                                                  trControl = training_control_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables,
                                                                                                                  metric = "ROC",
                                                                                                                  ntree = 100,
                                                                                                                  tuneGrid = RandomForestGRid_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables)
set.seed(5783)
randomforest_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_ntree_200 <- train(Personalised_Trajectotype_Verdict ~ .,
                                                                                                                  data = Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_No_Subject_Number_WIth_Personalised_Trajectotype_Verdict,
                                                                                                                  method = "rf",
                                                                                                                  trControl = training_control_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables,
                                                                                                                  metric = "ROC",
                                                                                                                  ntree = 200,
                                                                                                                  tuneGrid = RandomForestGRid_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables)
set.seed(5783)
randomforest_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_ntree_300 <- train(Personalised_Trajectotype_Verdict ~ .,
                                                                                                                  data = Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_No_Subject_Number_WIth_Personalised_Trajectotype_Verdict,
                                                                                                                  method = "rf",
```

```
                                                                    trControl      =
training_control_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SE
S_Selected_Variables,
                                                                    metric         =
"ROC",
                                                                    ntree = 300,
                                                                    tuneGrid       =
RandomForestGRid_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Onl
y_SES_Selected_Variables)
set.seed(5783)
randomforest_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES
_Selected_Variables_ntree_400 <- train(Personalised_Trajectotype_Verdict ~ .,
                                                                    data           =
Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables_No_Subject_Number_WIth_Personalised_Trajectotype_Verdict,
                                                                    method = "rf",
                                                                    trControl      =
training_control_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SE
S_Selected_Variables,
                                                                    metric         =
"ROC",
                                                                    ntree = 400,
                                                                    tuneGrid       =
RandomForestGRid_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Onl
y_SES_Selected_Variables)
set.seed(5783)
randomforest_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES
_Selected_Variables_ntree_500 <- train(Personalised_Trajectotype_Verdict ~ .,
                                                                    data           =
Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables_No_Subject_Number_WIth_Personalised_Trajectotype_Verdict,
                                                                    method = "rf",
                                                                    trControl      =
training_control_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SE
S_Selected_Variables,
                                                                    metric         =
"ROC",
                                                                    ntree = 500,
                                                                    tuneGrid       =
RandomForestGRid_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Onl
y_SES_Selected_Variables)

set.seed(5783)
randomforest_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES
_Selected_Variables_ntree_600 <- train(Personalised_Trajectotype_Verdict ~ .,
                                                                    data           =
Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables_No_Subject_Number_WIth_Personalised_Trajectotype_Verdict,
                                                                    method = "rf",
```

```
                                                                    trControl      =
training_control_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SE
S_Selected_Variables,
                                                                    metric         =
"ROC",
                                                                    ntree = 600,
                                                                    tuneGrid       =
RandomForestGRid_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Onl
y_SES_Selected_Variables)
set.seed(5783)
randomforest_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES
_Selected_Variables_ntree_700 <- train(Personalised_Trajectotype_Verdict ~ .,
                                                                    data           =
Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables_No_Subject_Number_WIth_Personalised_Trajectotype_Verdict,
                                                                    method = "rf",
                                                                    trControl      =
training_control_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SE
S_Selected_Variables,
                                                                    metric         =
"ROC",
                                                                    ntree = 700,
                                                                    tuneGrid       =
RandomForestGRid_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Onl
y_SES_Selected_Variables)
set.seed(5783)
randomforest_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES
_Selected_Variables_ntree_800 <- train(Personalised_Trajectotype_Verdict ~ .,
                                                                    data           =
Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables_No_Subject_Number_WIth_Personalised_Trajectotype_Verdict,
                                                                    method = "rf",
                                                                    trControl      =
training_control_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SE
S_Selected_Variables,
                                                                    metric         =
"ROC",
                                                                    ntree = 800,
                                                                    tuneGrid       =
RandomForestGRid_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Onl
y_SES_Selected_Variables)
set.seed(5783)
randomforest_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES
_Selected_Variables_ntree_900 <- train(Personalised_Trajectotype_Verdict ~ .,
                                                                    data           =
Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables_No_Subject_Number_WIth_Personalised_Trajectotype_Verdict,
                                                                    method = "rf",
```

```
                                                                    trControl =
training_control_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SE
S_Selected_Variables,
                                                                    metric =
"ROC",
                                                                    ntree = 900,
                                                                    tuneGrid =
RandomForestGRid_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Onl
y_SES_Selected_Variables)
set.seed(5783)
randomforest_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES
_Selected_Variables_ntree_1000 <- train(Personalised_Trajectotype_Verdict ~ .,
                                                                    data =
Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables_No_Subject_Number_WIth_Personalised_Trajectotype_Verdict,
                                                                    method = "rf",
                                                                    trControl =
training_control_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SE
S_Selected_Variables,
                                                                    metric =
"ROC",
                                                                    ntree = 1000,
                                                                    tuneGrid =
RandomForestGRid_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Onl
y_SES_Selected_Variables)

Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables_No_Subject_Number_No_Personalised_Trajectotype_Verdict_Categories_Nu
merical                                                             <-
Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables_No_Subject_Number_No_Personalised_Trajectotype_Verdict
Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables_No_Subject_Number_No_Personalised_Trajectotype_Verdict_Categories_Nu
merical$qHI                                                         <-
ifelse(Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selecte
d_Variables_No_Subject_Number_No_Personalised_Trajectotype_Verdict_Categori
es_Numerical$qHI=="Positive", 1,0)
Matrix_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Select
ed_Variables_No_Subject_Number_No_Personalised_Trajectotype_Verdict_Categor
ies_Numerical                                                       <-
as.matrix(Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Sel
ected_Variables_No_Subject_Number_No_Personalised_Trajectotype_Verdict_Cate
gories_Numerical)

set.seed(5783)
sigmaRange_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_
Selected_Variables                                                  <-
sigest(Matrix_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_
Selected_Variables_No_Subject_Number_No_Personalised_Trajectotype_Verdict_C
ategories_Numerical)
```

```
set.seed(5783)
svmRGRid_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables <- expand.grid(.sigma = seq(sigmaRange_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables[1],sigmaRange_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables[3],by = ((sigmaRange_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables[3] - sigmaRange_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables[1])/8)), .C = 2^(seq(-4,4)))

set.seed(5783)
SVM_Radial_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables <- train(Personalised_Trajectotype_Verdict ~ ., data = Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_No_Subject_Number_Categories_Numerical_WIth_Personalised_Trajectotype_Verdict,

method = "svmRadial",

trControl = training_control_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables,

metric = "ROC",

preProcess = c("center", "scale"), fit = FALSE,

tuneGrid = svmRGRid_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables)


set.seed(5783)
nnetRGRid_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables <- expand.grid(.size = 1:10, .decay = seq(0, 2, by = 0.1))
maxSize_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables <- max(nnetRGRid_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables$.size)
numWts_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables <- 1*(maxSize_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables * (ncol(Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_No_Subject_Number_No_Personalised_Trajectotype_Verdict) + 1) +
```

```
maxSize_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Sel
ected_Variables + 1)
set.seed(5783)
nnet_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selecte
d_Variables <- train(Personalised_Trajectotype_Verdict ~ .,
                                                                    data =
Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables_No_Subject_Number_Categories_Numerical_WIth_Personalised_Trajectotype
_Verdict,
                                                                    method
= "nnet",
                                                                    trControl
=
training_control_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SE
S_Selected_Variables,
                                                                    metric =
"ROC",

preProcess = c("center", "scale", "spatialSign"),
                                                                    tuneGrid
=
nnetRGRid_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_S
elected_Variables,
                                                                    trace =
FALSE,
                                                                    maxit =
2000,

MaxNWts                                                             =
numWts_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Sele
cted_Variables)

Symptomatic_Crises_Trajectotype_Prediction_0.2_TEST_No_Symptoms_Inflammati
on_Only_SES_Selected                                                <-
Symptomatic_Crises_Trajectotype_Prediction_0.2_TEST_No_Symptoms_Inflammati
on[,c("FEV1.post.Bronchodilator", "qHI")]
Symptomatic_Crises_Trajectotype_Prediction_0.2_TEST_No_Symptoms_Inflammati
on_Categories_Numerical_Only_SES_Selected                           <-
Symptomatic_Crises_Trajectotype_Prediction_0.2_TEST_No_Symptoms_Inflammati
on_Only_SES_Selected
Symptomatic_Crises_Trajectotype_Prediction_0.2_TEST_No_Symptoms_Inflammati
on_Categories_Numerical_Only_SES_Selected$qHI                       <-
ifelse(Symptomatic_Crises_Trajectotype_Prediction_0.2_TEST_No_Symptoms_Infla
mmation_Categories_Numerical_Only_SES_Selected$qHI=="Positive",1,0)
test_set_Symptomatic_Crises_Trajectotype_Prediction                 <-
predict(knn_cv_4_folds_Symptomatic_Crises_Trajectotype_Prediction_0.8_TRAIN_
Only_SES_Selected_Variables,
Symptomatic_Crises_Trajectotype_Prediction_0.2_TEST_No_Symptoms_Inflammati
on_Categories_Numerical_Only_SES_Selected, type = "prob")
```

```
test_set_Symptomatic_Crises_Trajectotype_Prediction_ROC_kNN <-
evalm(data.frame(test_set_Symptomatic_Crises_Trajectotype_Prediction,
Symptomatic_Crises_Trajectotype_Prediction_0.2_TEST$Personalised_Trajectotype
_Verdict, Group = "kNN"))
test_set_Symptomatic_Crises_Trajectotype_Prediction_ROC_kNN$roc
plot(test_set_Symptomatic_Crises_Trajectotype_Prediction_ROC_kNN$roc)



Imputed_Eosinophilic_Crises_Condensed_Potential_Trajectotype_Predictors_No_Sy
mptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$Sex <-
as.factor(Imputed_Eosinophilic_Crises_Condensed_Potential_Trajectotype_Predicto
rs_No_Symptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$Sex
)
Imputed_Eosinophilic_Crises_Condensed_Potential_Trajectotype_Predictors_No_Sy
mptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$Race <-
as.factor(Imputed_Eosinophilic_Crises_Condensed_Potential_Trajectotype_Predicto
rs_No_Symptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$Rac
e)
Imputed_Eosinophilic_Crises_Condensed_Potential_Trajectotype_Predictors_No_Sy
mptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$Smoking.Stat
us <-
as.factor(Imputed_Eosinophilic_Crises_Condensed_Potential_Trajectotype_Predicto
rs_No_Symptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$Sm
oking.Status)
Imputed_Eosinophilic_Crises_Condensed_Potential_Trajectotype_Predictors_No_Sy
mptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$qHI <-
as.factor(Imputed_Eosinophilic_Crises_Condensed_Potential_Trajectotype_Predicto
rs_No_Symptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$qHI)
Imputed_Eosinophilic_Crises_Condensed_Potential_Trajectotype_Predictors_No_Sy
mptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$qSA <-
as.factor(Imputed_Eosinophilic_Crises_Condensed_Potential_Trajectotype_Predicto
rs_No_Symptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$qSA
)
Imputed_Eosinophilic_Crises_Condensed_Potential_Trajectotype_Predictors_No_Sy
mptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$qSP <-
as.factor(Imputed_Eosinophilic_Crises_Condensed_Potential_Trajectotype_Predicto
rs_No_Symptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$qSP
)
Imputed_Eosinophilic_Crises_Condensed_Potential_Trajectotype_Predictors_No_Sy
mptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$qMC <-
as.factor(Imputed_Eosinophilic_Crises_Condensed_Potential_Trajectotype_Predicto
rs_No_Symptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$qM
C)
Imputed_Eosinophilic_Crises_Condensed_Potential_Trajectotype_Predictors_No_Sy
mptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$Combination_
of_Objective_State_and_Treatment_Allocation <-
as.factor(Imputed_Eosinophilic_Crises_Condensed_Potential_Trajectotype_Predicto
rs_No_Symptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$Co
mbination_of_Objective_State_and_Treatment_Allocation)
```

```
Imputed_Eosinophilic_Crises_Condensed_Potential_Trajectotype_Predictors_No_Symptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$Personalised_Trajectotype_Verdict <- as.factor(Imputed_Eosinophilic_Crises_Condensed_Potential_Trajectotype_Predictors_No_Symptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$Personalised_Trajectotype_Verdict)

set.seed(5783)
index_Eosinophilic_Crises_Trajectotype_Prediction_Ready_for_Data_Splitting <- groupKFold(Imputed_Eosinophilic_Crises_Condensed_Potential_Trajectotype_Predictors_No_Symptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$Subject_Number, k = k.folds)

length(index_Eosinophilic_Crises_Trajectotype_Prediction_Ready_for_Data_Splitting$Fold5)

Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_No_Symptoms_Inflammation <- Imputed_Eosinophilic_Crises_Condensed_Potential_Trajectotype_Predictors_No_Symptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number[index_Eosinophilic_Crises_Trajectotype_Prediction_Ready_for_Data_Splitting$Fold5,]
Eosinophilic_Crises_Trajectotype_Prediction_0.2_TEST_No_Symptoms_Inflammation <- Imputed_Eosinophilic_Crises_Condensed_Potential_Trajectotype_Predictors_No_Symptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number[-index_Eosinophilic_Crises_Trajectotype_Prediction_Ready_for_Data_Splitting$Fold5,]
Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_No_Symptoms_Inflammation_No_Label_No_Subject_Number <- Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_No_Symptoms_Inflammation
Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_No_Symptoms_Inflammation_No_Label_No_Subject_Number$Personalised_Trajectotype_Verdict <- NULL
Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_No_Symptoms_Inflammation_No_Label_No_Subject_Number$Subject_Number <- NULL

set.seed(5783)
Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_No_Symptoms_Inflammation$Personalised_Trajectotype_Verdict <- as.factor(Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_No_Symptoms_Inflammation$Personalised_Trajectotype_Verdict)
SES_max_k_32_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_No_Symptoms_Inflammation_No_Label_No_Subject_Number <- SES(Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_No_Symptoms_Inflammation$Personalised_Trajectotype_Verdict,
Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_No_Symptoms_Inflammation_No_Label_No_Subject_Number, max_k = 32, threshold = 0.1, test = NULL, hash = FALSE, hashObject = NULL)
SES_max_k_32_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_No_Symptoms_Inflammation_No_Label_No_Subject_Number
```

```
SES_max_k_32_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_No_Symptoms_Inflammation_No_Label_No_Subject_Number_pvalues <- c(-2.444876770, -3.062603702, -2.479312282)
SES_max_k_32_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_No_Symptoms_Inflammation_No_Label_No_Subject_Number_pvalues <- exp(SES_max_k_32_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_No_Symptoms_Inflammation_No_Label_No_Subject_Number_pvalues)
exp(-0.009900017)
exp(-0.397678378)
exp(-2.031132945)

Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_and_Subject_Number_and_Personalised_Trajectotype_Verdict <- Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_No_Symptoms_Inflammation
Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_and_Subject_Number_and_Personalised_Trajectotype_Verdict <- Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_and_Subject_Number_and_Personalised_Trajectotype_Verdict[,c("Systemic.Corticosteroids.at.Exacerbation","X..Pred.PB.FEV1","sACH","Personalised_Trajectotype_Verdict","Subject_Number")]
Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_Personalised_Trajectotype_Verdict_No_Subject_Number <- Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_and_Subject_Number_and_Personalised_Trajectotype_Verdict[,c("Systemic.Corticosteroids.at.Exacerbation","X..Pred.PB.FEV1","sACH","Personalised_Trajectotype_Verdict")]
Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_No_Personalised_Trajectotype_Verdict_No_Subject_Number <- Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_Personalised_Trajectotype_Verdict_No_Subject_Number[,c("Systemic.Corticosteroids.at.Exacerbation","X..Pred.PB.FEV1","sACH")]

set.seed(5783)
index_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables <- groupKFold(Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_and_Subject_Number_and_Personalised_Trajectotype_Verdict$Subject_Number, k = k.4_folds)

set.seed(5783)
training_control_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables <- trainControl(method = "cv",

summaryFunction = twoClassSummary,
                                                                    classProbs =
TRUE,
                                                                    number = 4,
repeats = NA, savePredictions = "all", set.seed(5783), index =
```

```
index_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selecte
d_Variables)

set.seed(5783)
cart1_cv_4_folds_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_S
ES_Selected_Variables <- train(Personalised_Trajectotype_Verdict ~ .,
                                                                data =
Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables_Personalised_Trajectotype_Verdict_No_Subject_Number,
                                                                method = "rpart",
                                                                trControl =
training_control_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SE
S_Selected_Variables,
                                                                metric = "ROC",
                                                                tuneLength = 10)
```

```
Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables_Categories_Numerical_and_Subject_Number_and_Personalised_Trajectotype
_Verdict                                                        <-
Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables_and_Subject_Number_and_Personalised_Trajectotype_Verdict
Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables_Categories_Numerical_Personalised_Trajectotype_Verdict_No_Subject_Num
ber                                                             <-
Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables_Personalised_Trajectotype_Verdict_No_Subject_Number
Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables_Categories_Numerical_No_Personalised_Trajectotype_Verdict_No_Subject_N
umber                                                           <-
Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables_No_Personalised_Trajectotype_Verdict_No_Subject_Number
Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables_Categories_Numerical_and_Subject_Number_and_Personalised_Trajectotype
_Verdict$Systemic.Corticosteroids.at.Exacerbation                <-
ifelse(Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected
_Variables_Categories_Numerical_and_Subject_Number_and_Personalised_Traject
otype_Verdict$Systemic.Corticosteroids.at.Exacerbation==TRUE,1,0)
Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables_Categories_Numerical_and_Subject_Number_and_Personalised_Trajectotype
_Verdict$sACH                                                   <-
ifelse(Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected
_Variables_Categories_Numerical_and_Subject_Number_and_Personalised_Traject
otype_Verdict$sACH==TRUE,1,0)
Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables_Categories_Numerical_Personalised_Trajectotype_Verdict_No_Subject_Num
ber$Systemic.Corticosteroids.at.Exacerbation                     <-
```

```
ifelse(Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected
_Variables_Categories_Numerical_Personalised_Trajectotype_Verdict_No_Subject_
Number$Systemic.Corticosteroids.at.Exacerbation==TRUE,1,0)
Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables_Categories_Numerical_Personalised_Trajectotype_Verdict_No_Subject_Num
ber$sACH <-
ifelse(Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected
_Variables_Categories_Numerical_Personalised_Trajectotype_Verdict_No_Subject_
Number$sACH==TRUE,1,0)
Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables_Categories_Numerical_No_Personalised_Trajectotype_Verdict_No_Subject_N
umber$Systemic.Corticosteroids.at.Exacerbation <-
ifelse(Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected
_Variables_Categories_Numerical_No_Personalised_Trajectotype_Verdict_No_Subj
ect_Number$Systemic.Corticosteroids.at.Exacerbation==TRUE,1,0)
Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables_Categories_Numerical_No_Personalised_Trajectotype_Verdict_No_Subject_N
umber$sACH <-
ifelse(Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected
_Variables_Categories_Numerical_No_Personalised_Trajectotype_Verdict_No_Subj
ect_Number$sACH==TRUE,1,0)



set.seed(5783)
knn_cv_4_folds_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SE
S_Selected_Variables <- train(Personalised_Trajectotype_Verdict ~ .,
                              data =
Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables_Categories_Numerical_Personalised_Trajectotype_Verdict_No_Subject_Num
ber, method = "knn", trControl =
training_control_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SE
S_Selected_Variables, metric = "ROC",tuneGrid = data.frame(k = seq(1,35,by = 2)))


RandomForestGRid_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only
_SES_Selected_Variables <- expand.grid(.mtry = c(1:3))

set.seed(5783)
randomforest_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_
Selected_Variables_ntree_100 <- train(Personalised_Trajectotype_Verdict ~ .,
                                      data =
Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables_Personalised_Trajectotype_Verdict_No_Subject_Number,
                                      method = "rf",
                                      trControl =
training_control_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SE
S_Selected_Variables,
                                      metric =
"ROC",
```

```
                                                                    ntree = 100,
                                                                    tuneGrid      =
RandomForestGRid_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only
_SES_Selected_Variables)
set.seed(5783)
randomforest_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_
Selected_Variables_ntree_200 <- train(Personalised_Trajectotype_Verdict ~ .,
                                                                    data          =
Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables_Personalised_Trajectotype_Verdict_No_Subject_Number,
                                                                    method = "rf",
                                                                    trControl     =
training_control_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SE
S_Selected_Variables,
                                                                    metric        =
"ROC",
                                                                    ntree = 200,
                                                                    tuneGrid      =
RandomForestGRid_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only
_SES_Selected_Variables)

set.seed(5783)
randomforest_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_
Selected_Variables_ntree_300 <- train(Personalised_Trajectotype_Verdict ~ .,
                                                                    data          =
Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables_Personalised_Trajectotype_Verdict_No_Subject_Number,
                                                                    method = "rf",
                                                                    trControl     =
training_control_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SE
S_Selected_Variables,
                                                                    metric        =
"ROC",
                                                                    ntree = 300,
                                                                    tuneGrid      =
RandomForestGRid_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only
_SES_Selected_Variables)

set.seed(5783)
randomforest_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_
Selected_Variables_ntree_400 <- train(Personalised_Trajectotype_Verdict ~ .,
                                                                    data          =
Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables_Personalised_Trajectotype_Verdict_No_Subject_Number,
                                                                    method = "rf",
                                                                    trControl     =
training_control_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SE
S_Selected_Variables,
                                                                    metric        =
"ROC",
```

```
                                                                    ntree = 400,
                                                                    tuneGrid      =
RandomForestGRid_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only
_SES_Selected_Variables)


set.seed(5783)
randomforest_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_
Selected_Variables_ntree_500 <- train(Personalised_Trajectotype_Verdict ~ .,
                                                                    data          =
Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables_Personalised_Trajectotype_Verdict_No_Subject_Number,
                                                                    method = "rf",
                                                                    trControl     =
training_control_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SE
S_Selected_Variables,
                                                                    metric        =
"ROC",
                                                                    ntree = 500,
                                                                    tuneGrid      =
RandomForestGRid_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only
_SES_Selected_Variables)


set.seed(5783)
randomforest_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_
Selected_Variables_ntree_600 <- train(Personalised_Trajectotype_Verdict ~ .,
                                                                    data          =
Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables_Personalised_Trajectotype_Verdict_No_Subject_Number,
                                                                    method = "rf",
                                                                    trControl     =
training_control_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SE
S_Selected_Variables,
                                                                    metric        =
"ROC",
                                                                    ntree = 600,
                                                                    tuneGrid      =
RandomForestGRid_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only
_SES_Selected_Variables)

set.seed(5783)
randomforest_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_
Selected_Variables_ntree_700 <- train(Personalised_Trajectotype_Verdict ~ .,
                                                                    data          =
Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables_Personalised_Trajectotype_Verdict_No_Subject_Number,
                                                                    method = "rf",
```

```
                                                                    trControl     =
training_control_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SE
S_Selected_Variables,
                                                                    metric        =
"ROC",
                                                                    ntree = 700,
                                                                    tuneGrid      =
RandomForestGRid_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only
_SES_Selected_Variables)


set.seed(5783)
randomforest_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_
Selected_Variables_ntree_800 <- train(Personalised_Trajectotype_Verdict ~ .,
                                                                    data          =
Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables_Personalised_Trajectotype_Verdict_No_Subject_Number,
                                                                    method = "rf",
                                                                    trControl     =
training_control_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SE
S_Selected_Variables,
                                                                    metric        =
"ROC",
                                                                    ntree = 800,
                                                                    tuneGrid      =
RandomForestGRid_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only
_SES_Selected_Variables)

set.seed(5783)
randomforest_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_
Selected_Variables_ntree_900 <- train(Personalised_Trajectotype_Verdict ~ .,
                                                                    data          =
Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables_Personalised_Trajectotype_Verdict_No_Subject_Number,
                                                                    method = "rf",
                                                                    trControl     =
training_control_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SE
S_Selected_Variables,
                                                                    metric        =
"ROC",
                                                                    ntree = 900,
                                                                    tuneGrid      =
RandomForestGRid_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only
_SES_Selected_Variables)

set.seed(5783)
randomforest_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_
Selected_Variables_ntree_1000 <- train(Personalised_Trajectotype_Verdict ~ .,
```

```
                                                                    data =
Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables_Personalised_Trajectotype_Verdict_No_Subject_Number,
                                                                    method = "rf",
                                                                    trControl =
training_control_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SE
S_Selected_Variables,
                                                                    metric =
"ROC",
                                                                    ntree = 1000,
                                                                    tuneGrid =
RandomForestGRid_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only
_SES_Selected_Variables)


Matrix_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selecte
d_Variables_Categories_Numerical_No_Personalised_Trajectotype_Verdict_No_Su
bject_Number                                                        <-
as.matrix(Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Sele
cted_Variables_Categories_Numerical_No_Personalised_Trajectotype_Verdict_No_
Subject_Number)

set.seed(5783)
sigmaRange_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_
Selected_Variables                                                  <-
sigest(Matrix_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_
Selected_Variables_Categories_Numerical_No_Personalised_Trajectotype_Verdict_
No_Subject_Number)
set.seed(5783)
svmRGRid_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Se
lected_Variables          <-          expand.grid(.sigma          =
seq(sigmaRange_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_S
ES_Selected_Variables[1],sigmaRange_Eosinophilic_Crises_Trajectotype_Predictio
n_0.8_TRAIN_Only_SES_Selected_Variables[3],by                       =
((sigmaRange_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES
_Selected_Variables[3]                                              -
sigmaRange_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_
Selected_Variables[1])/8)), .C = 2^(seq(-4,4)))

set.seed(5783)
SVM_Radial_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_
Selected_Variables <- train(Personalised_Trajectotype_Verdict ~ .,
                                                                    data =
Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables_Categories_Numerical_Personalised_Trajectotype_Verdict_No_Subject_Num
ber,
                                                                    method =
"svmRadial",
```

```
                                                    trControl =
training_control_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SE
S_Selected_Variables,
                                                    metric = "ROC",
                                                    preProcess =
c("center", "scale"),
                                                    fit = FALSE,
                                                    tuneGrid =
svmRGRid_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Se
lected_Variables)


set.seed(5783)
nnetRGRid_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Se
lected_Variables <- expand.grid(.size = 1:10, .decay = seq(0, 2, by = 0.1))
maxSize_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Sele
cted_Variables <-
max(nnetRGRid_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SE
S_Selected_Variables$.size)
numWts_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selec
ted_Variables <-
1*(maxSize_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_S
elected_Variables *
(ncol(Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected
_Variables_Categories_Numerical_No_Personalised_Trajectotype_Verdict_No_Subj
ect_Number) + 1) +
maxSize_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Sele
cted_Variables + 1)
set.seed(5783)
nnet_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected
_Variables <- train(Personalised_Trajectotype_Verdict ~ .,
                                                    data =
Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Vari
ables_Categories_Numerical_Personalised_Trajectotype_Verdict_No_Subject_Num
ber,
                                                    method = "nnet",
                                                    trControl =
training_control_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SE
S_Selected_Variables,
                                                    metric = "ROC",
                                                    preProcess = c("center",
"scale", "spatialSign"),
                                                    tuneGrid =
nnetRGRid_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Se
lected_Variables,
                                                    trace = FALSE,
                                                    maxit = 2000,
                                                    MaxNWts =
numWts_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selec
ted_Variables)
```

```
Eosinophilic_Crises_Trajectotype_Prediction_0.2_TEST_No_Symptoms_Inflammatio
n_Only_SES_Selected <-
Eosinophilic_Crises_Trajectotype_Prediction_0.2_TEST_No_Symptoms_Inflammatio
n[,c("Systemic.Corticosteroids.at.Exacerbation","X..Pred.PB.FEV1","sACH")]
Eosinophilic_Crises_Trajectotype_Prediction_0.2_TEST_No_Symptoms_Inflammatio
n_Only_SES_Selected_Categories_Numerical <-
Eosinophilic_Crises_Trajectotype_Prediction_0.2_TEST_No_Symptoms_Inflammatio
n_Only_SES_Selected
Eosinophilic_Crises_Trajectotype_Prediction_0.2_TEST_No_Symptoms_Inflammatio
n_Only_SES_Selected_Categories_Numerical$Systemic.Corticosteroids.at.Exacerb
ation <-
ifelse(Eosinophilic_Crises_Trajectotype_Prediction_0.2_TEST_No_Symptoms_Infla
mmation_Only_SES_Selected_Categories_Numerical$Systemic.Corticosteroids.at.E
xacerbation==TRUE,1,0)
Eosinophilic_Crises_Trajectotype_Prediction_0.2_TEST_No_Symptoms_Inflammatio
n_Only_SES_Selected_Categories_Numerical$sACH <-
ifelse(Eosinophilic_Crises_Trajectotype_Prediction_0.2_TEST_No_Symptoms_Infla
mmation_Only_SES_Selected_Categories_Numerical$sACH==TRUE,1,0)
test_set_Eosinophilic_Crises_Trajectotype_Prediction <-
predict(knn_cv_4_folds_Eosinophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_O
nly_SES_Selected_Variables,
Eosinophilic_Crises_Trajectotype_Prediction_0.2_TEST_No_Symptoms_Inflammatio
n_Only_SES_Selected_Categories_Numerical, type = "prob")
test_set_Eosinophilic_Crises_Trajectotype_Prediction_ROC_kNN <-
evalm(data.frame(test_set_Eosinophilic_Crises_Trajectotype_Prediction,
Eosinophilic_Crises_Trajectotype_Prediction_0.2_TEST_No_Symptoms_Inflammatio
n$Personalised_Trajectotype_Verdict, Group = "kNN"))
test_set_Eosinophilic_Crises_Trajectotype_Prediction_ROC_kNN$roc
plot(test_set_Eosinophilic_Crises_Trajectotype_Prediction_ROC_kNN$roc)
```

```
Imputed_Neutrophilic_Crisis_Condensed_Potential_Trajectotype_Predictors_No_Sy
mptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$Sex <-
as.factor(Imputed_Neutrophilic_Crisis_Condensed_Potential_Trajectotype_Predictor
s_No_Symptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$Sex)
Imputed_Neutrophilic_Crisis_Condensed_Potential_Trajectotype_Predictors_No_Sy
mptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$Race <-
as.factor(Imputed_Neutrophilic_Crisis_Condensed_Potential_Trajectotype_Predictor
s_No_Symptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$Rac
e)
Imputed_Neutrophilic_Crisis_Condensed_Potential_Trajectotype_Predictors_No_Sy
mptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$Smoking.Stat
us <-
as.factor(Imputed_Neutrophilic_Crisis_Condensed_Potential_Trajectotype_Predictor
```

```
s_No_Symptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$Smo
king.Status)
Imputed_Neutrophilic_Crisis_Condensed_Potential_Trajectotype_Predictors_No_Sy
mptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$qHI <-
as.factor(Imputed_Neutrophilic_Crisis_Condensed_Potential_Trajectotype_Predictor
s_No_Symptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$qHI)
Imputed_Neutrophilic_Crisis_Condensed_Potential_Trajectotype_Predictors_No_Sy
mptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$qSA <-
as.factor(Imputed_Neutrophilic_Crisis_Condensed_Potential_Trajectotype_Predictor
s_No_Symptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$qSA)
Imputed_Neutrophilic_Crisis_Condensed_Potential_Trajectotype_Predictors_No_Sy
mptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$qSP <-
as.factor(Imputed_Neutrophilic_Crisis_Condensed_Potential_Trajectotype_Predictor
s_No_Symptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$qSP)
Imputed_Neutrophilic_Crisis_Condensed_Potential_Trajectotype_Predictors_No_Sy
mptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$qMC <-
as.factor(Imputed_Neutrophilic_Crisis_Condensed_Potential_Trajectotype_Predictor
s_No_Symptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$qMC
)
Imputed_Neutrophilic_Crisis_Condensed_Potential_Trajectotype_Predictors_No_Sy
mptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$Combination_
of_Objective_State_and_Treatment_Allocation <-
as.factor(Imputed_Neutrophilic_Crisis_Condensed_Potential_Trajectotype_Predictor
s_No_Symptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$Com
bination_of_Objective_State_and_Treatment_Allocation)
Imputed_Neutrophilic_Crisis_Condensed_Potential_Trajectotype_Predictors_No_Sy
mptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$Personalised_
Trajectotype_Verdict <-
as.factor(Imputed_Neutrophilic_Crisis_Condensed_Potential_Trajectotype_Predictor
s_No_Symptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$Pers
onalised_Trajectotype_Verdict)

set.seed(5783)
index_Neutrophilic_Crises_Trajectotype_Prediction_Ready_for_Data_Splitting <-
groupKFold(Imputed_Neutrophilic_Crisis_Condensed_Potential_Trajectotype_Predic
tors_No_Symptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number$S
ubject_Number, k = k.folds)

length(index_Neutrophilic_Crises_Trajectotype_Prediction_Ready_for_Data_Splitting
$Fold5)

Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_No_Symptoms_Inflammati
on <-
Imputed_Neutrophilic_Crisis_Condensed_Potential_Trajectotype_Predictors_No_Sy
mptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number[index_Neutrop
hilic_Crises_Trajectotype_Prediction_Ready_for_Data_Splitting$Fold5,]
Neutrophilic_Crises_Trajectotype_Prediction_0.2_TEST_No_Symptoms_Inflammatio
n <-
Imputed_Neutrophilic_Crisis_Condensed_Potential_Trajectotype_Predictors_No_Sy
mptoms_Inflammation_with_Trajectotype_Outcome_Subject_Number[-
```

```
index_Neutrophilic_Crises_Trajectotype_Prediction_Ready_for_Data_Splitting$Fold5
,]
Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_No_Symptoms_Inflammati
on_No_Label_No_Subject_Number <-
Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_No_Symptoms_Inflammati
on
Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_No_Symptoms_Inflammati
on_No_Label_No_Subject_Number$Personalised_Trajectotype_Verdict <- NULL
Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_No_Symptoms_Inflammati
on_No_Label_No_Subject_Number$Subject_Number <- NULL

set.seed(5783)
Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_No_Symptoms_Inflammati
on$Personalised_Trajectotype_Verdict <-
as.factor(Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_No_Symptoms_I
nflammation$Personalised_Trajectotype_Verdict)
SES_max_k_32_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_No_Sym
ptoms_Inflammation_No_Label_No_Subject_Number <-
SES(Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_No_Symptoms_Infla
mmation$Personalised_Trajectotype_Verdict,
Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_No_Symptoms_Inflammati
on_No_Label_No_Subject_Number, max_k = 32, threshold = 0.1, test = NULL, hash
= FALSE, hashObject = NULL)
SES_max_k_32_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_No_Sym
ptoms_Inflammation_No_Label_No_Subject_Number

SES_max_k_32_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_No_Sym
ptoms_Inflammation_No_Label_No_Subject_Number_pvalues <- c(-2.55506831, -
2.73657435, -2.61424690)
SES_max_k_32_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_No_Sym
ptoms_Inflammation_No_Label_No_Subject_Number_pvalues <-
exp(SES_max_k_32_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_No_
Symptoms_Inflammation_No_Label_No_Subject_Number_pvalues)

Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Varia
bles_and_Subject_Number_and_Personalised_Trajectotype_Verdict <-
Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_No_Symptoms_Inflammati
on
Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Varia
bles_and_Subject_Number_and_Personalised_Trajectotype_Verdict <-
Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Varia
bles_and_Subject_Number_and_Personalised_Trajectotype_Verdict[,c("sACH","PO.
Pred","qMC","Personalised_Trajectotype_Verdict","Subject_Number")]
Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Varia
bles_Personalised_Trajectotype_Verdict_No_Subject_Number <-
Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Varia
bles_and_Subject_Number_and_Personalised_Trajectotype_Verdict[,c("sACH","PO.
Pred","qMC","Personalised_Trajectotype_Verdict")]
Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Varia
bles_No_Personalised_Trajectotype_Verdict_No_Subject_Number <-
```

```
Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_Personalised_Trajectotype_Verdict_No_Subject_Number[,c("sACH","PO.Pred","qMC")]


set.seed(5783)
index_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables <- groupKFold(Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_and_Subject_Number_and_Personalised_Trajectotype_Verdict$Subject_Number, k = k.4_folds)

set.seed(5783)
training_control_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables <- trainControl(method = "cv",

summaryFunction = twoClassSummary,
                                                                    classProbs = TRUE,
                                                                    number = 4, repeats = NA, savePredictions = "all", set.seed(5783), index = index_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables)

set.seed(5783)
cart1_cv_4_folds_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables <- train(Personalised_Trajectotype_Verdict ~ .,
                                                                    data = Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_Personalised_Trajectotype_Verdict_No_Subject_Number,
                                                                    method = "rpart",
                                                                    trControl = training_control_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables,
                                                                    metric = "ROC",
                                                                    tuneLength = 10)


Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_Categories_Numerical_and_Subject_Number_and_Personalised_Trajectotype_Verdict <- Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_and_Subject_Number_and_Personalised_Trajectotype_Verdict
Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_Categories_Numerical_Personalised_Trajectotype_Verdict_No_Subject_Number <- Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_Personalised_Trajectotype_Verdict_No_Subject_Number
Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_Categories_Numerical_No_Personalised_Trajectotype_Verdict_No_Subject_Nu
```

```
mber <-
Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_No_Personalised_Trajectotype_Verdict_No_Subject_Number

Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_Categories_Numerical_and_Subject_Number_and_Personalised_Trajectotype_Verdict$sACH <-
ifelse(Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_Categories_Numerical_and_Subject_Number_and_Personalised_Trajectotype_Verdict$sACH==TRUE,1,0)
Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_Categories_Numerical_and_Subject_Number_and_Personalised_Trajectotype_Verdict$PO.Pred <-
ifelse(Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_Categories_Numerical_and_Subject_Number_and_Personalised_Trajectotype_Verdict$PO.Pred==TRUE,1,0)
Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_Categories_Numerical_and_Subject_Number_and_Personalised_Trajectotype_Verdict$qMC <-
ifelse(Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_Categories_Numerical_and_Subject_Number_and_Personalised_Trajectotype_Verdict$qMC=="Positive",1,0)

Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_Categories_Numerical_Personalised_Trajectotype_Verdict_No_Subject_Number$sACH <-
ifelse(Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_Categories_Numerical_Personalised_Trajectotype_Verdict_No_Subject_Number$sACH==TRUE,1,0)
Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_Categories_Numerical_Personalised_Trajectotype_Verdict_No_Subject_Number$PO.Pred <-
ifelse(Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_Categories_Numerical_Personalised_Trajectotype_Verdict_No_Subject_Number$PO.Pred==TRUE,1,0)
Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_Categories_Numerical_Personalised_Trajectotype_Verdict_No_Subject_Number$qMC <-
ifelse(Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_Categories_Numerical_Personalised_Trajectotype_Verdict_No_Subject_Number$qMC=="Positive",1,0)

Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_Categories_Numerical_No_Personalised_Trajectotype_Verdict_No_Subject_Number$sACH <-
ifelse(Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_Categories_Numerical_No_Personalised_Trajectotype_Verdict_No_Subject_Number$sACH==TRUE,1,0)
Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_Categories_Numerical_No_Personalised_Trajectotype_Verdict_No_Subject_Nu
```

```
mber$PO.Pred <-
ifelse(Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected
_Variables_Categories_Numerical_No_Personalised_Trajectotype_Verdict_No_Subj
ect_Number$PO.Pred==TRUE,1,0)
Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Varia
bles_Categories_Numerical_No_Personalised_Trajectotype_Verdict_No_Subject_Nu
mber$qMC <-
ifelse(Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected
_Variables_Categories_Numerical_No_Personalised_Trajectotype_Verdict_No_Subj
ect_Number$qMC=="Positive",1,0)



set.seed(5783)
knn_cv_4_folds_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SE
S_Selected_Variables <- train(Personalised_Trajectotype_Verdict ~ .,
                              data =
Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Varia
bles_Categories_Numerical_Personalised_Trajectotype_Verdict_No_Subject_Numb
er, method = "knn", trControl =
training_control_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SE
S_Selected_Variables, metric = "ROC",tuneGrid = data.frame(k = seq(1,41,by = 2)))


RandomForestGRid_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only
_SES_Selected_Variables <- expand.grid(.mtry = c(1:3))

set.seed(5783)
randomforest_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_
Selected_Variables_ntree_100 <- train(Personalised_Trajectotype_Verdict ~ .,
                                      data =
Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Varia
bles_Personalised_Trajectotype_Verdict_No_Subject_Number,
                                      method = "rf",
                                      trControl =
training_control_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SE
S_Selected_Variables,
                                      metric =
"ROC",
                                      ntree = 100,
                                      tuneGrid =
RandomForestGRid_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only
_SES_Selected_Variables)
set.seed(5783)
randomforest_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_
Selected_Variables_ntree_200 <- train(Personalised_Trajectotype_Verdict ~ .,
                                      data =
Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Varia
bles_Personalised_Trajectotype_Verdict_No_Subject_Number,
                                      method = "rf",
```

```
                                                                    trControl     =
training_control_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SE
S_Selected_Variables,
                                                                    metric        =
"ROC",
                                                                    ntree = 200,
                                                                    tuneGrid      =
RandomForestGRid_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only
_SES_Selected_Variables)

set.seed(5783)
randomforest_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_
Selected_Variables_ntree_300 <- train(Personalised_Trajectotype_Verdict ~ .,
                                                                    data          =
Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Varia
bles_Personalised_Trajectotype_Verdict_No_Subject_Number,
                                                                    method = "rf",
                                                                    trControl     =
training_control_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SE
S_Selected_Variables,
                                                                    metric        =
"ROC",
                                                                    ntree = 300,
                                                                    tuneGrid      =
RandomForestGRid_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only
_SES_Selected_Variables)

set.seed(5783)
randomforest_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_
Selected_Variables_ntree_400 <- train(Personalised_Trajectotype_Verdict ~ .,
                                                                    data          =
Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Varia
bles_Personalised_Trajectotype_Verdict_No_Subject_Number,
                                                                    method = "rf",
                                                                    trControl     =
training_control_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SE
S_Selected_Variables,
                                                                    metric        =
"ROC",
                                                                    ntree = 400,
                                                                    tuneGrid      =
RandomForestGRid_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only
_SES_Selected_Variables)

set.seed(5783)
randomforest_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_
Selected_Variables_ntree_500 <- train(Personalised_Trajectotype_Verdict ~ .,
                                                                    data          =
Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Varia
bles_Personalised_Trajectotype_Verdict_No_Subject_Number,
```

```
                                                                    method = "rf",
                                                                    trControl     =
training_control_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SE
S_Selected_Variables,
                                                                    metric        =
"ROC",
                                                                    ntree = 500,
                                                                    tuneGrid      =
RandomForestGRid_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only
_SES_Selected_Variables)

set.seed(5783)
randomforest_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_
Selected_Variables_ntree_600 <- train(Personalised_Trajectotype_Verdict ~ .,
                                                                    data          =
Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Varia
bles_Personalised_Trajectotype_Verdict_No_Subject_Number,
                                                                    method = "rf",
                                                                    trControl     =
training_control_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SE
S_Selected_Variables,
                                                                    metric        =
"ROC",
                                                                    ntree = 600,
                                                                    tuneGrid      =
RandomForestGRid_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only
_SES_Selected_Variables)

set.seed(5783)
randomforest_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_
Selected_Variables_ntree_700 <- train(Personalised_Trajectotype_Verdict ~ .,
                                                                    data          =
Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Varia
bles_Personalised_Trajectotype_Verdict_No_Subject_Number,
                                                                    method = "rf",
                                                                    trControl     =
training_control_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SE
S_Selected_Variables,
                                                                    metric        =
"ROC",
                                                                    ntree = 700,
                                                                    tuneGrid      =
RandomForestGRid_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only
_SES_Selected_Variables)

set.seed(5783)
randomforest_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_
Selected_Variables_ntree_800 <- train(Personalised_Trajectotype_Verdict ~ .,
```

```
                                                                    data          =
Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_Personalised_Trajectotype_Verdict_No_Subject_Number,
                                                                    method = "rf",
                                                                    trControl     =
training_control_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables,
                                                                    metric        =
"ROC",
                                                                    ntree = 800,
                                                                    tuneGrid      =
RandomForestGRid_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables)

set.seed(5783)
randomforest_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_ntree_900 <- train(Personalised_Trajectotype_Verdict ~ .,
                                                                    data          =
Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_Personalised_Trajectotype_Verdict_No_Subject_Number,
                                                                    method = "rf",
                                                                    trControl     =
training_control_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables,
                                                                    metric        =
"ROC",
                                                                    ntree = 900,
                                                                    tuneGrid      =
RandomForestGRid_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables)

set.seed(5783)
randomforest_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_ntree_1000 <- train(Personalised_Trajectotype_Verdict ~ .,
                                                                    data          =
Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_Personalised_Trajectotype_Verdict_No_Subject_Number,
                                                                    method = "rf",
                                                                    trControl     =
training_control_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables,
                                                                    metric        =
"ROC",
                                                                    ntree = 1000,
                                                                    tuneGrid      =
RandomForestGRid_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables)
```

```
Matrix_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_Categories_Numerical_No_Personalised_Trajectotype_Verdict_No_Subject_Number <- as.matrix(Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_Categories_Numerical_No_Personalised_Trajectotype_Verdict_No_Subject_Number)

set.seed(5783)
sigmaRange_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables <- sigest(Matrix_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_Categories_Numerical_No_Personalised_Trajectotype_Verdict_No_Subject_Number)
set.seed(5783)
svmRGRid_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables <- expand.grid(.sigma = seq(sigmaRange_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables[1],sigmaRange_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables[3],by = ((sigmaRange_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables[3] - sigmaRange_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables[1])/8)), .C = 2^(seq(-4,4)))

set.seed(5783)
SVM_Radial_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables <- train(Personalised_Trajectotype_Verdict ~ .,
                                                                                                      data = Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_Categories_Numerical_Personalised_Trajectotype_Verdict_No_Subject_Number,
                                                                                                      method = "svmRadial",
                                                                                                      trControl = training_control_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables,
                                                                                                      metric = "ROC",
                                                                                                      preProcess = c("center", "scale"),
                                                                                                      fit = FALSE,
                                                                                                      tuneGrid = svmRGRid_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables)


set.seed(5783)
nnetRGRid_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables <- expand.grid(.size = 1:10, .decay = seq(0, 2, by = 0.1))
maxSize_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables <-
```

```
max(nnetRGRid_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables$.size)
numWts_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables <- 1*(maxSize_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables * (ncol(Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_Categories_Numerical_No_Personalised_Trajectotype_Verdict_No_Subject_Number) + 1) + maxSize_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables + 1)
set.seed(5783)
nnet_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables <- train(Personalised_Trajectotype_Verdict ~ .,
                                                                                                data = Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables_Categories_Numerical_Personalised_Trajectotype_Verdict_No_Subject_Number,
                                                                                                method = "nnet",
                                                                                                trControl = training_control_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables,
                                                                                                metric = "ROC",
                                                                                                preProcess = c("center", "scale", "spatialSign"),
                                                                                                tuneGrid = nnetRGRid_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables,
                                                                                                trace = FALSE,
                                                                                                maxit = 2000,
                                                                                                MaxNWts = numWts_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only_SES_Selected_Variables)


Neutrophilic_Crises_Trajectotype_Prediction_0.2_TEST_No_Symptoms_Inflammation_Only_SES_Selected <- Neutrophilic_Crises_Trajectotype_Prediction_0.2_TEST_No_Symptoms_Inflammation[,c("sACH","PO.Pred","qMC")]
Neutrophilic_Crises_Trajectotype_Prediction_0.2_TEST_No_Symptoms_Inflammation_Categories_Numerical_Only_SES_Selected <- Neutrophilic_Crises_Trajectotype_Prediction_0.2_TEST_No_Symptoms_Inflammation_Only_SES_Selected
Neutrophilic_Crises_Trajectotype_Prediction_0.2_TEST_No_Symptoms_Inflammation_Categories_Numerical_Only_SES_Selected$sACH <- ifelse(Neutrophilic_Crises_Trajectotype_Prediction_0.2_TEST_No_Symptoms_Inflammation_Categories_Numerical_Only_SES_Selected$sACH==TRUE,1,0)
Neutrophilic_Crises_Trajectotype_Prediction_0.2_TEST_No_Symptoms_Inflammation_Categories_Numerical_Only_SES_Selected$PO.Pred <-
```

```
ifelse(Neutrophilic_Crises_Trajectotype_Prediction_0.2_TEST_No_Symptoms_Infla
mmation_Categories_Numerical_Only_SES_Selected$PO.Pred==TRUE,1,0)
Neutrophilic_Crises_Trajectotype_Prediction_0.2_TEST_No_Symptoms_Inflammatio
n_Categories_Numerical_Only_SES_Selected$qMC <-
ifelse(Neutrophilic_Crises_Trajectotype_Prediction_0.2_TEST_No_Symptoms_Infla
mmation_Categories_Numerical_Only_SES_Selected$qMC=="Positive",1,0)

test_set_Neutrophilic_Crises_Trajectotype_Prediction <-
predict(SVM_Radial_Neutrophilic_Crises_Trajectotype_Prediction_0.8_TRAIN_Only
_SES_Selected_Variables,
Neutrophilic_Crises_Trajectotype_Prediction_0.2_TEST_No_Symptoms_Inflammatio
n_Categories_Numerical_Only_SES_Selected, type = "prob")
test_set_Neutrophilic_Crises_Trajectotype_Prediction_ROC_SVM <-
evalm(data.frame(test_set_Neutrophilic_Crises_Trajectotype_Prediction,
Neutrophilic_Crises_Trajectotype_Prediction_0.2_TEST_No_Symptoms_Inflammatio
n$Personalised_Trajectotype_Verdict, Group = "SVM"))
test_set_Neutrophilic_Crises_Trajectotype_Prediction_ROC_SVM$roc
test_set_Neutrophilic_Crises_Trajectotype_Prediction_ROC_SVM$roc
```

Identification of VAS symptom shape profiles

```
Phase_2_VAS_Dyspnoea_14_Day_Trajectory_with_Kalman_Imputed_Values_to_N
earest_Whole_VAS_Unit_Rows_as_Time_Series <-
read.csv(file="Phase_2_VAS_Dyspnoea_14_Day_Trajectory_with_Kalman_Imputed
_Values_to_Nearest_Whole_VAS_Unit_Rows_as_Time_Series.csv", sep=",",
header=TRUE)
Phase_2_VAS_Dyspnoea_14_Day_Trajectory_with_Kalman_Imputed_Values_to_N
earest_Whole_VAS_Unit_Rows_as_Time_Series_without_Subject_Names <-
Phase_2_VAS_Dyspnoea_14_Day_Trajectory_with_Kalman_Imputed_Values_to_N
earest_Whole_VAS_Unit_Rows_as_Time_Series
Phase_2_VAS_Dyspnoea_14_Day_Trajectory_with_Kalman_Imputed_Values_to_N
earest_Whole_VAS_Unit_Rows_as_Time_Series_without_Subject_Names$Phase_2
_VAS_Dyspnoea_14_Day_Trajectory_Excluding_Patients_With_Missing_Values <-
NULL
Phase_2_VAS_Dyspnoea_14_Day_Trajectory_with_Kalman_Imputed_Values_to_N
earest_Whole_VAS_Unit_Rows_as_Time_Series_without_Subject_Names$Day_Po
st_Exacerbation <- NULL
Phase_2_VAS_Dyspnoea_14_Day_Trajectory_with_Kalman_Imputed_Values_to_N
earest_Whole_VAS_Unit_Rows_as_Time_Series_without_Subject_Names_List_of_
Univariate_Time_Series <-
Phase_2_VAS_Dyspnoea_14_Day_Trajectory_with_Kalman_Imputed_Values_to_N
earest_Whole_VAS_Unit_Rows_as_Time_Series_without_Subject_Names
Phase_2_VAS_Dyspnoea_14_Day_Trajectory_with_Kalman_Imputed_Values_to_N
earest_Whole_VAS_Unit_Rows_as_Time_Series_without_Subject_Names_List_of_
Univariate_Time_Series <-
tslist(Phase_2_VAS_Dyspnoea_14_Day_Trajectory_with_Kalman_Imputed_Values_
to_Nearest_Whole_VAS_Unit_Rows_as_Time_Series_without_Subject_Names_List
_of_Univariate_Time_Series)
Phase_2_VAS_Dyspnoea_14_Day_Trajectory_with_Kalman_Imputed_Values_to_N
earest_Whole_VAS_Unit_Rows_as_Time_Series_without_Subject_Names_List_of_
```

```
Univariate_Time_Series_Normalized <-
zscore(Phase_2_VAS_Dyspnoea_14_Day_Trajectory_with_Kalman_Imputed_Value
s_to_Nearest_Whole_VAS_Unit_Rows_as_Time_Series_without_Subject_Names_Li
st_of_Univariate_Time_Series)
pc_ks_dyspnoea_experiment_2_clusters <-
tsclust(Phase_2_VAS_Dyspnoea_14_Day_Trajectory_with_Kalman_Imputed_Value
s_to_Nearest_Whole_VAS_Unit_Rows_as_Time_Series_without_Subject_Names_Li
st_of_Univariate_Time_Series_Normalized, k = 2, seed = 100, distance = "sbd",
centroid = "shape")
plot(unlist(pc_ks_dyspnoea_experiment_2_clusters@centroids[[1]]), type = "b")
plot(unlist(pc_ks_dyspnoea_experiment_2_clusters@centroids[[2]]), type = "b")
pc_ks_dyspnoea_experiment_3_clusters <-
tsclust(Phase_2_VAS_Dyspnoea_14_Day_Trajectory_with_Kalman_Imputed_Value
s_to_Nearest_Whole_VAS_Unit_Rows_as_Time_Series_without_Subject_Names_Li
st_of_Univariate_Time_Series_Normalized, k = 3, seed = 100, distance = "sbd",
centroid = "shape")
plot(unlist(pc_ks_dyspnoea_experiment_3_clusters@centroids[[1]]), type = "b")
plot(unlist(pc_ks_dyspnoea_experiment_3_clusters@centroids[[2]]), type = "b")
plot(unlist(pc_ks_dyspnoea_experiment_3_clusters@centroids[[3]]), type = "b")
pc_ks_dyspnoea_experiment_4_clusters <-
tsclust(Phase_2_VAS_Dyspnoea_14_Day_Trajectory_with_Kalman_Imputed_Value
s_to_Nearest_Whole_VAS_Unit_Rows_as_Time_Series_without_Subject_Names_Li
st_of_Univariate_Time_Series_Normalized, k = 4, seed = 100, distance = "sbd",
centroid = "shape")
plot(unlist(pc_ks_dyspnoea_experiment_4_clusters@centroids[[1]]), type = "b")
plot(unlist(pc_ks_dyspnoea_experiment_4_clusters@centroids[[2]]), type = "b")
plot(unlist(pc_ks_dyspnoea_experiment_4_clusters@centroids[[3]]), type = "b")
plot(unlist(pc_ks_dyspnoea_experiment_4_clusters@centroids[[4]]), type = "b")
pc_ks_dyspnoea_experiment_5_clusters <-
tsclust(Phase_2_VAS_Dyspnoea_14_Day_Trajectory_with_Kalman_Imputed_Value
s_to_Nearest_Whole_VAS_Unit_Rows_as_Time_Series_without_Subject_Names_Li
st_of_Univariate_Time_Series_Normalized, k = 5, seed = 100, distance = "sbd",
centroid = "shape")
plot(unlist(pc_ks_dyspnoea_experiment_5_clusters@centroids[[1]]), type = "b")
plot(unlist(pc_ks_dyspnoea_experiment_5_clusters@centroids[[2]]), type = "b")
plot(unlist(pc_ks_dyspnoea_experiment_5_clusters@centroids[[3]]), type = "b")
plot(unlist(pc_ks_dyspnoea_experiment_5_clusters@centroids[[4]]), type = "b")
plot(unlist(pc_ks_dyspnoea_experiment_5_clusters@centroids[[5]]), type = "b")
pc_ks_dyspnoea_experiment_6_clusters <-
tsclust(Phase_2_VAS_Dyspnoea_14_Day_Trajectory_with_Kalman_Imputed_Value
s_to_Nearest_Whole_VAS_Unit_Rows_as_Time_Series_without_Subject_Names_Li
st_of_Univariate_Time_Series_Normalized, k = 6, seed = 100, distance = "sbd",
centroid = "shape")
plot(unlist(pc_ks_dyspnoea_experiment_6_clusters@centroids[[1]]), type = "b")
plot(unlist(pc_ks_dyspnoea_experiment_6_clusters@centroids[[2]]), type = "b")
plot(unlist(pc_ks_dyspnoea_experiment_6_clusters@centroids[[3]]), type = "b")
plot(unlist(pc_ks_dyspnoea_experiment_6_clusters@centroids[[4]]), type = "b")
plot(unlist(pc_ks_dyspnoea_experiment_6_clusters@centroids[[5]]), type = "b")
plot(unlist(pc_ks_dyspnoea_experiment_6_clusters@centroids[[6]]), type = "b")
```

```
pc_ks_dyspnoea_experiment_7_clusters <- tsclust(Phase_2_VAS_Dyspnoea_14_Day_Trajectory_with_Kalman_Imputed_Values_to_Nearest_Whole_VAS_Unit_Rows_as_Time_Series_without_Subject_Names_List_of_Univariate_Time_Series_Normalized, k = 7, seed = 100, distance = "sbd", centroid = "shape")
plot(unlist(pc_ks_dyspnoea_experiment_7_clusters@centroids[[1]]), type = "b")
plot(unlist(pc_ks_dyspnoea_experiment_7_clusters@centroids[[2]]), type = "b")
plot(unlist(pc_ks_dyspnoea_experiment_7_clusters@centroids[[3]]), type = "b")
plot(unlist(pc_ks_dyspnoea_experiment_7_clusters@centroids[[4]]), type = "b")
plot(unlist(pc_ks_dyspnoea_experiment_7_clusters@centroids[[5]]), type = "b")
plot(unlist(pc_ks_dyspnoea_experiment_7_clusters@centroids[[6]]), type = "b")
plot(unlist(pc_ks_dyspnoea_experiment_7_clusters@centroids[[7]]), type = "b")
pc_ks_dyspnoea_experiment_2_to_7_clusters <- tsclust(Phase_2_VAS_Dyspnoea_14_Day_Trajectory_with_Kalman_Imputed_Values_to_Nearest_Whole_VAS_Unit_Rows_as_Time_Series_without_Subject_Names_List_of_Univariate_Time_Series_Normalized, k = 2L:7L, seed = 100, distance = "sbd", centroid = "shape")
names(pc_ks_dyspnoea_experiment_2_to_7_clusters) <- paste0("k_", 2L:7L)
pc_ks_dyspnoea_experiment_2_to_7_clusters_evaluation_metric_summary <- sapply(pc_ks_dyspnoea_experiment_2_to_7_clusters, cvi, type = "internal")
Phase_2_VAS_Cough_14_Day_Trajectory_with_Kalman_Imputed_Values_to_Nearest_Whole_VAS_Unit_Rows_as_Time_Series <- read.csv(file = "Cough_Day_0_to_Day_14_Events_by_Row.csv", sep = ",", header = FALSE)
Phase_2_VAS_Sputum_Production_14_Day_Trajectory_with_Kalman_Imputed_Values_to_Nearest_Whole_VAS_Unit_Rows_as_Time_Series <- read.csv(file = "Sputum_Production_Day_0_to_Day_14_Trajectories_by_Row.csv", sep = ",", header = FALSE)
Phase_2_VAS_Sputum_Purulence_14_Day_Trajectory_with_Kalman_Imputed_Values_to_Nearest_Whole_VAS_Unit_Rows_as_Time_Series <- read.csv(file = "Sputum_Purulence_Day_0_to_14_Trajectories_by_Row.csv", sep = ",", header = FALSE)
Phase_2_VAS_Cough_14_Day_Trajectory_with_Kalman_Imputed_Values_to_Nearest_Whole_VAS_Unit_Rows_as_Time_Series_without_Subject_Names_List_of_Univariate_Time_Series <- tslist(Phase_2_VAS_Cough_14_Day_Trajectory_with_Kalman_Imputed_Values_to_Nearest_Whole_VAS_Unit_Rows_as_Time_Series)
Phase_2_VAS_Cough_14_Day_Trajectory_with_Kalman_Imputed_Values_to_Nearest_Whole_VAS_Unit_Rows_as_Time_Series_without_Subject_Names_List_of_Univariate_Time_Series_Normalized <- zscore(Phase_2_VAS_Cough_14_Day_Trajectory_with_Kalman_Imputed_Values_to_Nearest_Whole_VAS_Unit_Rows_as_Time_Series_without_Subject_Names_List_of_Univariate_Time_Series)
Phase_2_VAS_Sputum_Production_14_Day_Trajectory_with_Kalman_Imputed_Values_to_Nearest_Whole_VAS_Unit_Rows_as_Time_Series_without_Subject_Names_List_of_Univariate_Time_Series <- tslist(Phase_2_VAS_Sputum_Production_14_Day_Trajectory_with_Kalman_Imputed_Values_to_Nearest_Whole_VAS_Unit_Rows_as_Time_Series)
Phase_2_VAS_Sputum_Production_14_Day_Trajectory_with_Kalman_Imputed_Values_to_Nearest_Whole_VAS_Unit_Rows_as_Time_Series_without_Subject_Names_List_of_Univariate_Time_Series_Normalized <-
```

```
zscore(Phase_2_VAS_Sputum_Production_14_Day_Trajectory_with_Kalman_Imput
ed_Values_to_Nearest_Whole_VAS_Unit_Rows_as_Time_Series_without_Subject_
Names_List_of_Univariate_Time_Series)
Phase_2_VAS_Sputum_Purulence_14_Day_Trajectory_with_Kalman_Imputed_Valu
es_to_Nearest_Whole_VAS_Unit_Rows_as_Time_Series_without_Subject_Names_
List_of_Univariate_Time_Series <-
tslist(Phase_2_VAS_Sputum_Purulence_14_Day_Trajectory_with_Kalman_Imputed
_Values_to_Nearest_Whole_VAS_Unit_Rows_as_Time_Series)
Phase_2_VAS_Sputum_Purulence_14_Day_Trajectory_with_Kalman_Imputed_Valu
es_to_Nearest_Whole_VAS_Unit_Rows_as_Time_Series_without_Subject_Names_
List_of_Univariate_Time_Series_Normalized <-
zscore(Phase_2_VAS_Sputum_Purulence_14_Day_Trajectory_with_Kalman_Impute
d_Values_to_Nearest_Whole_VAS_Unit_Rows_as_Time_Series_without_Subject_N
ames_List_of_Univariate_Time_Series)
pc_ks_cough_experiment_2_clusters <-
tsclust(Phase_2_VAS_Cough_14_Day_Trajectory_with_Kalman_Imputed_Values_to
_Nearest_Whole_VAS_Unit_Rows_as_Time_Series_without_Subject_Names_List_o
f_Univariate_Time_Series_Normalized, k = 2, seed = 100, distance = "sbd", centroid
= "shape")
plot(unlist(pc_ks_cough_experiment_2_clusters@centroids[[1]]), type = "b")
plot(unlist(pc_ks_cough_experiment_2_clusters@centroids[[2]]), type = "b")
pc_ks_cough_experiment_3_clusters <-
tsclust(Phase_2_VAS_Cough_14_Day_Trajectory_with_Kalman_Imputed_Values_to
_Nearest_Whole_VAS_Unit_Rows_as_Time_Series_without_Subject_Names_List_o
f_Univariate_Time_Series_Normalized, k = 3, seed = 100, distance = "sbd", centroid
= "shape")
plot(unlist(pc_ks_cough_experiment_3_clusters@centroids[[1]]), type = "b")
plot(unlist(pc_ks_cough_experiment_3_clusters@centroids[[2]]), type = "b")
plot(unlist(pc_ks_cough_experiment_3_clusters@centroids[[3]]), type = "b")
pc_ks_cough_experiment_4_clusters <-
tsclust(Phase_2_VAS_Cough_14_Day_Trajectory_with_Kalman_Imputed_Values_to
_Nearest_Whole_VAS_Unit_Rows_as_Time_Series_without_Subject_Names_List_o
f_Univariate_Time_Series_Normalized, k = 4, seed = 100, distance = "sbd", centroid
= "shape")
plot(unlist(pc_ks_cough_experiment_4_clusters@centroids[[1]]), type = "b")
plot(unlist(pc_ks_cough_experiment_4_clusters@centroids[[2]]), type = "b")
plot(unlist(pc_ks_cough_experiment_4_clusters@centroids[[3]]), type = "b")
plot(unlist(pc_ks_cough_experiment_4_clusters@centroids[[4]]), type = "b")
pc_ks_cough_experiment_5_clusters <-
tsclust(Phase_2_VAS_Cough_14_Day_Trajectory_with_Kalman_Imputed_Values_to
_Nearest_Whole_VAS_Unit_Rows_as_Time_Series_without_Subject_Names_List_o
f_Univariate_Time_Series_Normalized, k = 5, seed = 100, distance = "sbd", centroid
= "shape")
plot(unlist(pc_ks_cough_experiment_5_clusters@centroids[[1]]), type = "b")
plot(unlist(pc_ks_cough_experiment_5_clusters@centroids[[2]]), type = "b")
plot(unlist(pc_ks_cough_experiment_5_clusters@centroids[[3]]), type = "b")
plot(unlist(pc_ks_cough_experiment_5_clusters@centroids[[4]]), type = "b")
plot(unlist(pc_ks_cough_experiment_5_clusters@centroids[[5]]), type = "b")
pc_ks_cough_experiment_6_clusters <-
tsclust(Phase_2_VAS_Cough_14_Day_Trajectory_with_Kalman_Imputed_Values_to
```

```
_Nearest_Whole_VAS_Unit_Rows_as_Time_Series_without_Subject_Names_List_of_Univariate_Time_Series_Normalized, k = 6, seed = 100, distance = "sbd", centroid = "shape")
plot(unlist(pc_ks_cough_experiment_6_clusters@centroids[[1]]), type = "b")
plot(unlist(pc_ks_cough_experiment_6_clusters@centroids[[2]]), type = "b")
plot(unlist(pc_ks_cough_experiment_6_clusters@centroids[[3]]), type = "b")
plot(unlist(pc_ks_cough_experiment_6_clusters@centroids[[4]]), type = "b")
plot(unlist(pc_ks_cough_experiment_6_clusters@centroids[[5]]), type = "b")
plot(unlist(pc_ks_cough_experiment_6_clusters@centroids[[6]]), type = "b")
pc_ks_cough_experiment_2_to_6_clusters <- tsclust(Phase_2_VAS_Cough_14_Day_Trajectory_with_Kalman_Imputed_Values_to_Nearest_Whole_VAS_Unit_Rows_as_Time_Series_without_Subject_Names_List_of_Univariate_Time_Series_Normalized, k = 2L:6L, seed = 100, distance = "sbd", centroid = "shape")
names(pc_ks_cough_experiment_2_to_6_clusters) <- paste0("k_", 2L:6L)
pc_ks_cough_experiment_2_to_6_clusters_evaluation_metric_summary <- sapply(pc_ks_cough_experiment_2_to_6_clusters, cvi, type = "internal")
pc_ks_sputum_production_experiment_2_clusters <- tsclust(Phase_2_VAS_Sputum_Production_14_Day_Trajectory_with_Kalman_Imputed_Values_to_Nearest_Whole_VAS_Unit_Rows_as_Time_Series_without_Subject_Names_List_of_Univariate_Time_Series_Normalized, k = 2, seed = 100, distance = "sbd", centroid = "shape")
plot(unlist(pc_ks_sputum_production_experiment_2_clusters@centroids[[1]]), type = "b")
plot(unlist(pc_ks_sputum_production_experiment_2_clusters@centroids[[2]]), type = "b")
pc_ks_sputum_production_experiment_3_clusters <- tsclust(Phase_2_VAS_Sputum_Production_14_Day_Trajectory_with_Kalman_Imputed_Values_to_Nearest_Whole_VAS_Unit_Rows_as_Time_Series_without_Subject_Names_List_of_Univariate_Time_Series_Normalized, k = 3, seed = 100, distance = "sbd", centroid = "shape")
plot(unlist(pc_ks_sputum_production_experiment_3_clusters@centroids[[1]]), type = "b")
plot(unlist(pc_ks_sputum_production_experiment_3_clusters@centroids[[2]]), type = "b")
plot(unlist(pc_ks_sputum_production_experiment_3_clusters@centroids[[3]]), type = "b")
pc_ks_sputum_production_experiment_4_clusters <- tsclust(Phase_2_VAS_Sputum_Production_14_Day_Trajectory_with_Kalman_Imputed_Values_to_Nearest_Whole_VAS_Unit_Rows_as_Time_Series_without_Subject_Names_List_of_Univariate_Time_Series_Normalized, k = 4, seed = 100, distance = "sbd", centroid = "shape")
plot(unlist(pc_ks_sputum_production_experiment_4_clusters@centroids[[1]]), type = "b")
plot(unlist(pc_ks_sputum_production_experiment_4_clusters@centroids[[2]]), type = "b")
plot(unlist(pc_ks_sputum_production_experiment_4_clusters@centroids[[3]]), type = "b")
plot(unlist(pc_ks_sputum_production_experiment_4_clusters@centroids[[4]]), type = "b")
```

```
pc_ks_sputum_production_experiment_5_clusters <-
tsclust(Phase_2_VAS_Sputum_Production_14_Day_Trajectory_with_Kalman_Imput
ed_Values_to_Nearest_Whole_VAS_Unit_Rows_as_Time_Series_without_Subject_
Names_List_of_Univariate_Time_Series_Normalized, k = 5, seed = 100, distance =
"sbd", centroid = "shape")
plot(unlist(pc_ks_sputum_production_experiment_5_clusters@centroids[[1]]), type =
"b")
plot(unlist(pc_ks_sputum_production_experiment_5_clusters@centroids[[2]]), type =
"b")
plot(unlist(pc_ks_sputum_production_experiment_5_clusters@centroids[[3]]), type =
"b")
plot(unlist(pc_ks_sputum_production_experiment_5_clusters@centroids[[4]]), type =
"b")
plot(unlist(pc_ks_sputum_production_experiment_5_clusters@centroids[[5]]), type =
"b")
pc_ks_sputum_production_experiment_6_clusters <-
tsclust(Phase_2_VAS_Sputum_Production_14_Day_Trajectory_with_Kalman_Imput
ed_Values_to_Nearest_Whole_VAS_Unit_Rows_as_Time_Series_without_Subject_
Names_List_of_Univariate_Time_Series_Normalized, k = 6, seed = 100, distance =
"sbd", centroid = "shape")
plot(unlist(pc_ks_sputum_production_experiment_6_clusters@centroids[[1]]), type =
"b")
plot(unlist(pc_ks_sputum_production_experiment_6_clusters@centroids[[2]]), type =
"b")
plot(unlist(pc_ks_sputum_production_experiment_6_clusters@centroids[[3]]), type =
"b")
plot(unlist(pc_ks_sputum_production_experiment_6_clusters@centroids[[4]]), type =
"b")
plot(unlist(pc_ks_sputum_production_experiment_6_clusters@centroids[[5]]), type =
"b")
plot(unlist(pc_ks_sputum_production_experiment_6_clusters@centroids[[6]]), type =
"b")
pc_ks_sputum_production_experiment_2_to_6_clusters <-
tsclust(Phase_2_VAS_Sputum_Production_14_Day_Trajectory_with_Kalman_Imput
ed_Values_to_Nearest_Whole_VAS_Unit_Rows_as_Time_Series_without_Subject_
Names_List_of_Univariate_Time_Series_Normalized, k = 2L:6L, seed = 100, distance
= "sbd", centroid = "shape")
names(pc_ks_sputum_production_experiment_2_to_6_clusters) <- paste0("k_",
2L:6L)
pc_ks_sputum_production_experiment_2_to_6_clusters_evaluation_metric_summar
y <- sapply(pc_ks_cough_experiment_2_to_6_clusters, cvi, type = "internal")
pc_ks_sputum_purulence_experiment_2_clusters <-
tsclust(Phase_2_VAS_Sputum_Purulence_14_Day_Trajectory_with_Kalman_Imputed_
Values_to_Nearest_Whole_VAS_Unit_Rows_as_Time_Series_without_Subject_Nam
es_List_of_Univariate_Time_Series_Normalized, k = 2, seed = 100, distance =
"sbd", centroid = "shape")
plot(unlist(pc_ks_sputum_purulence_experiment_2_clusters@centroids[[1]]), type =
"b")
plot(unlist(pc_ks_sputum_purulence_experiment_2_clusters@centroids[[2]]), type =
"b")
```

```
pc_ks_sputum_purulence_experiment_3_clusters <-
tsclust(Phase_2_VAS_Sputum_Purulence_14_Day_Trajectory_with_Kalman_Imputed_Values_to_Nearest_Whole_VAS_Unit_Rows_as_Time_Series_without_Subject_Names_List_of_Univariate_Time_Series_Normalized, k = 3, seed = 100, distance = "sbd", centroid = "shape")
plot(unlist(pc_ks_sputum_purulence_experiment_3_clusters@centroids[[1]]), type = "b")
plot(unlist(pc_ks_sputum_purulence_experiment_3_clusters@centroids[[2]]), type = "b")
plot(unlist(pc_ks_sputum_purulence_experiment_3_clusters@centroids[[3]]), type = "b")
pc_ks_sputum_purulence_experiment_4_clusters <-
tsclust(Phase_2_VAS_Sputum_Purulence_14_Day_Trajectory_with_Kalman_Imputed_Values_to_Nearest_Whole_VAS_Unit_Rows_as_Time_Series_without_Subject_Names_List_of_Univariate_Time_Series_Normalized, k = 4, seed = 100, distance = "sbd", centroid = "shape")
plot(unlist(pc_ks_sputum_purulence_experiment_4_clusters@centroids[[1]]), type = "b")
plot(unlist(pc_ks_sputum_purulence_experiment_4_clusters@centroids[[2]]), type = "b")
plot(unlist(pc_ks_sputum_purulence_experiment_4_clusters@centroids[[3]]), type = "b")
plot(unlist(pc_ks_sputum_purulence_experiment_4_clusters@centroids[[4]]), type = "b")
pc_ks_sputum_purulence_experiment_5_clusters <-
tsclust(Phase_2_VAS_Sputum_Purulence_14_Day_Trajectory_with_Kalman_Imputed_Values_to_Nearest_Whole_VAS_Unit_Rows_as_Time_Series_without_Subject_Names_List_of_Univariate_Time_Series_Normalized, k = 5, seed = 100, distance = "sbd", centroid = "shape")
plot(unlist(pc_ks_sputum_purulence_experiment_5_clusters@centroids[[1]]), type = "b")
plot(unlist(pc_ks_sputum_purulence_experiment_5_clusters@centroids[[2]]), type = "b")
plot(unlist(pc_ks_sputum_purulence_experiment_5_clusters@centroids[[3]]), type = "b")
plot(unlist(pc_ks_sputum_purulence_experiment_5_clusters@centroids[[4]]), type = "b")
plot(unlist(pc_ks_sputum_purulence_experiment_5_clusters@centroids[[5]]), type = "b")
pc_ks_sputum_purulence_experiment_6_clusters <-
tsclust(Phase_2_VAS_Sputum_Purulence_14_Day_Trajectory_with_Kalman_Imputed_Values_to_Nearest_Whole_VAS_Unit_Rows_as_Time_Series_without_Subject_Names_List_of_Univariate_Time_Series_Normalized, k = 6, seed = 100, distance = "sbd", centroid = "shape")
plot(unlist(pc_ks_sputum_purulence_experiment_6_clusters@centroids[[1]]), type = "b")
plot(unlist(pc_ks_sputum_purulence_experiment_6_clusters@centroids[[2]]), type = "b")
plot(unlist(pc_ks_sputum_purulence_experiment_6_clusters@centroids[[3]]), type = "b")
```

```
plot(unlist(pc_ks_sputum_purulence_experiment_6_clusters@centroids[[4]]), type = "b")
plot(unlist(pc_ks_sputum_purulence_experiment_6_clusters@centroids[[5]]), type = "b")
plot(unlist(pc_ks_sputum_purulence_experiment_6_clusters@centroids[[6]]), type = "b")
pc_ks_sputum_purulence_experiment_2_to_6_clusters <- tsclust(Phase_2_VAS_Sputum_Purulence_14_Day_Trajectory_with_Kalman_Imputed_Values_to_Nearest_Whole_VAS_Unit_Rows_as_Time_Series_without_Subject_Names_List_of_Univariate_Time_Series_Normalized, k = 2L:6L, seed = 100, distance = "sbd", centroid = "shape")
names(pc_ks_sputum_purulence_experiment_2_to_6_clusters) <- paste0("k_", 2L:6L)
pc_ks_sputum_purulence_experiment_2_to_6_clusters_evaluation_metric_summary <- sapply(pc_ks_cough_experiment_2_to_6_clusters, cvi, type = "internal")

Vector_Cough_Centroids_of_4_Clusters <- as.vector(pc_ks_cough_experiment_4_clusters@centroids)
Vector_Dyspnoea_Centroids_of_3_Clusters <- as.vector(pc_ks_dyspnoea_experiment_3_clusters@centroids)
Vector_Sputum_Production_Centroids_of_2_Clusters <- as.vector(pc_ks_sputum_production_experiment_2_clusters@centroids)
Vector_Sputum_Purulence_Centroids_of_2_Clusters <- as.vector(pc_ks_sputum_purulence_experiment_2_clusters@centroids)

colnames(Vector_Cough_Centroids_of_4_Clusters) <- NULL
Matrix_Vector_Cough_Centroids_of_4_Clusters <- as.matrix(Vector_Cough_Centroids_of_4_Clusters)
plot(Vector_Cough_Centroids_of_4_Clusters[[1]],type = "b", xlab="Day", ylab="Standardised VAS Cough Score", main = "Centroids of Four Trajectotypes", xlim = c(0,15), ylim = c(-3,4), col = "#660000")
plot(Vector_Cough_Centroids_of_4_Clusters[[2]],type = "b", xlab="Day", ylab="Standardised VAS Cough Score", main = "Centroids of Four Trajectotypes", xlim = c(0,15), ylim = c(-3,4), col = "#660000")
plot(Vector_Cough_Centroids_of_4_Clusters[[3]],type = "b", xlab="Day", ylab="Standardised VAS Cough Score", main = "Centroids of Four Trajectotypes", xlim = c(0,15), ylim = c(-3,4), col = "#660000")
plot(Vector_Cough_Centroids_of_4_Clusters[[4]],type = "b", xlab="Day", ylab="Standardised VAS Cough Score", main = "Centroids of Four Trajectotypes", xlim = c(0,15), ylim = c(-3,4), col = "#660000")

colnames(Vector_Dyspnoea_Centroids_of_3_Clusters) <- NULL
Matrix_Vector_Dyspnoea_Centroids_of_3_Clusters <- as.matrix(Vector_Dyspnoea_Centroids_of_3_Clusters)
plot(Vector_Dyspnoea_Centroids_of_3_Clusters[[1]],type = "b", xlab="Day", ylab="Standardised VAS Dyspnoea Score", main = "Centroids of Three Trajectotypes", xlim = c(0,15), ylim = c(-3,4), col = "#FF0000")
plot(Vector_Dyspnoea_Centroids_of_3_Clusters[[2]],type = "b", xlab="Day", ylab="Standardised VAS Dyspnoea Score", main = "Centroids of Three Trajectotypes", xlim = c(0,15), ylim = c(-3,4), col = "#FF0000")
```

```
plot(Vector_Dyspnoea_Centroids_of_3_Clusters[[3]],type = "b", xlab="Day", ylab="Standardised VAS Dyspnoea Score", main = "Centroids of Three Trajectotypes", xlim = c(0,15), ylim = c(-3,4), col = "#FF0000")

colnames(Vector_Sputum_Production_Centroids_of_2_Clusters) <- NULL
plot(Vector_Sputum_Production_Centroids_of_2_Clusters[[1]],type = "b", xlab="Day", ylab="Standardised VAS Sputum Production Score", main = "Centroids of Two Trajectotypes", xlim = c(0,15), ylim = c(-3,4), col = "#999966")
plot(Vector_Sputum_Production_Centroids_of_2_Clusters[[2]],type = "b", xlab="Day", ylab="Standardised VAS Sputum Production Score", main = "Centroids of Two Trajectotypes", xlim = c(0,15), ylim = c(-3,4), col = "#999966")

colnames(Vector_Sputum_Purulence_Centroids_of_2_Clusters) <- NULL
plot(Vector_Sputum_Purulence_Centroids_of_2_Clusters[[1]],type = "b", xlab="Day", ylab="Standardised VAS Sputum Purulence Score", main = "Centroids of Two Trajectotypes", xlim = c(0,15), ylim = c(-3,4), col = "#003366")
plot(Vector_Sputum_Purulence_Centroids_of_2_Clusters[[2]],type = "b", xlab="Day", ylab="Standardised VAS Sputum Purulence Score", main = "Centroids of Two Trajectotypes", xlim = c(0,15), ylim = c(-3,4), col = "#003366")
```

References

1. Vos, T. *et al.* Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990-2016: A systematic analysis for the Global Burden of Disease Study 2016. *Lancet* 390, 1211–1259 (2017).

2. Abajobir, A. A. *et al.* Global, regional, and national age-sex specific mortality for 264 causes of death, 1980–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet* 390, 1151–1210 (2017).

3. The top 10 causes of death. *WHO* (2018).

4. Decramer, M. *et al*. Chronic obstructive pulmonary disease. *Lancet* 379, 1341–1351 (2012).

5. O'Donnell, D. E. Hyperinflation, Dyspnea, and Exercise Intolerance in Chronic Obstructive Pulmonary Disease. *Proc. Am. Thorac. Soc.* 3, 180–184 (2006).

6. Global Initiative for Chronic Obstructive Lung Disease. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease – 2022 report.
Available from: https://goldcopd.org/wpcontent/uploads/2021/11/ GOLD-REPORT-2022-v1.0- 12Nov2021_WMV.pdf
Date accessed: 11th April 2022.

7. Kohansal, R. *et al.* The natural history of chronic airflow obstruction revisited: An analysis of the Framingham Offspring Cohort. *Am. J. Respir. Crit. Care Med.* 180, 3–10 (2009).

8. Agusti, A. *et al*. Biomarkers, the control panel and personalized COPD medicine. *Respirology* 21, 24–33 (2016).

9. Dornhorst, A. C. Respiratory Insufficiency. *Lancet* 265, 1185–1187 (1955).

10. Pauwels, R. A. *et al*. NHLBI / WHO Workshop Summary Global Strategy for the Diagnosis , Management , and Prevention of Chronic Obstructive Pulmonary Disease NHLBI / WHO Global Initiative for Chronic Obstructive Lung Disease (GOLD) Workshop Summary. 163, 1256–1276 (2001).

11. Hurst, J. R. *et al.* Susceptibility to Exacerbation in Chronic Obstructive Pulmonary Disease. *N. Engl. J. Med.* 363, 1128–1138 (2010).

12. Nishimura, K. *et al*. Dyspnea is a better predictor of 5-year survival than airway obstruction in patients with COPD. *Chest* 121, 1434–1440 (2002).

13. Fabbri, L. M. *et al*. Complex chronic comorbidities of COPD. *Eur. Respir. J.* 31, 204–212 (2008).

14. Burgel, P.R. *et al*. Identification of Clinical Phenotypes Using Cluster Analyses

in COPD Patients with Multiple Comorbidities. *Biomed Res. Int.* 2014, 1–9 (2014).

15. Walters, J.A. *et al.* Systemic corticosteroids for acute exacerbations of chronic obstructive pulmonary disease (Review). *Cochrane Database Syst. Rev.* (2014) doi:10.1002/14651858.CD001288.pub4.www.cochranelibrary.com.

16. Han, M. K. *et al.* Chronic Obstructive Pulmonary Disease Phenotypes. *Am. J. Respir. Crit. Care Med.* 182, 598–604 (2010).

17. Anderson, G. P. Endotyping asthma: new insights into key pathogenic mechanisms in a complex, heterogeneous disease. *Lancet* 372, 1107–1119 (2008).

18. Wardlaw, A. J. *et al*. Multi-dimensional phenotyping: Towards a new taxonomy for airway disease. *Clin. Exp. Allergy* 35, 1254–1262 (2005).

19. Bellman, R.E. Adaptive Control Processes. Princeton University Press, Princeton, NJ, 1961.

20. Dolnicar, S.A. Review of Unquestioned Standards in Using Cluster Analysis for Data-driven Market Segmentation, CD Conference Proceedings of the Australian and New Zealand Marketing Academy Conference (2002).

21. Formann, A.K. Die Latent-Class-Analyse : Einführung in Theorie und Anwendung. Beltz; 1984.

22. Everitt, B.S. *et al*. Cluster Analysis. 5th ed. Wiley (2011).

23. Pearson, K. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2, 559-572 (1901).

24. Hotelling, H. Analysis of a complex of statistical variables into principal components. *J Educ Psychol,* 25:417–441 (1933).

25. Greenacre, M.J. Theory and applications of correspondence analysis. *London: Academic Press* (1984).

26. Pagès J. Analyse factorielle de données mixtes. *Rev. Stat. qppliquée*, 52: 93–111 (2004).

27. Halner, A. & Bafadhel, M. Usefulness of Biological Clustering Patterns in Chronic Obstructive Pulmonary Disease. *Barcelona Respir. Netw. Rev.* **5**, 201–14 (2019).

28. Tan P.N. *et al*. Introduction to Data Mining. 1st ed. Pearson Addison Wesley (2006).

29. Hartigan J.A. & Wong MA. Algorithm AS 136: A K-Means Clustering Algorithm. *Appl. Stat.* 28: 100 (1979).

30. Calinski T & Harabasz J. A dendrite method for cluster analysis. *Commun. Stat. - Theory Methods* 3: 1–27 (1974).

31. Ward, J.H. Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc.* 58: 236–244 (1963).

32. Ester, M *et al*. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD* 96: 226–231 (1996).

33. Wolpert, D.H. & Macready WG. No Free Lunch Theorems for Optimisation. *IEEE Trans. Evol. Comput.* 1: 67–82 (1997).

34. Wolpert, D.H. Coevolutionary Free Lunches. *IEEE Trans. Evol. Comput.* 9: 721–735 (2005).

35. Hopkins, B. A New Method for determining the Type of Distribution of Plant Individuals. *Ann Bot* 112: 213–227 (1954).

36. Fernández Pierna, J.A. & Massart, D.L. Improved algorithm for clustering tendency. *Anal. Chim. Acta* 408: 13–20 (2000).

37. Fukunaga, K. Introduction to Statistical Pattern Recognition. *Computer Science and Scientific Computing* (1990).

38. Kuhn, M. & Johnson, K. *Applied Predictive Modeling*. New York : Springer (2013).

39. Fix, E. & Hodges, J.L. Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties. *International Statistical Review*, 57: 238-247 (1951).

40. Altman, N.S. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*. 46: 175–185 (1992).

41. Tibshirani, R. Regression Shrinkage and Selection via the lasso". *Journal of the Royal Statistical Society*, 58: 267–88 (1996).

42. Cortes, C. & Vladimir Vapnik. Support-Vector Networks. *Machine Learning*, 20: 273–297 (1995).

43. Ripley, B. *Pattern Recognition and Neural Networks.* Cambridge University Press (1996).

44. Breiman, L. Random Forests. *Mach. Learn.* 45**,** 5–32 (2001).

45. Altman D.G. & Bland J.M. Diagnostic tests 3: receiver operating characteristic plots. *BMJ*. 309:188 (1994).

46. Fawcett, T. An Introduction to ROC analysis. *Pattern Recognition Letters*, 27, 861-874 (2006).

47. Varma, S. & Simon, R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* 7**,** 91 (2006).

48. Ambroise, C. & McLachlan G.J. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci*. 99:6562-6 (2002).

49. Seemungal, T.A.R. *et al.* Effect of Exacerbation on Quality of Life in Patients with Chronic Obstructive Pulmonary Disease. *Am. J. Respir. Crit. Care Med.* 157, 1418–1422 (1998).

50. Donaldson, G.C. *et al*. Relationship between exacerbation frequency and lungfunction decline in chronic obstructive pulmonary disease. *Thorax* 57, 847–852 (2002).

51. Connors Jr, A. *et al.* Outcomes following acute exacerbation of severe chronic obstructive lung disease. The SUPPORT investigators (Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments). *Am. J. Respir. Crit. Care Med.* 154, 959–967 (1996).

52. Burge, P. S. *et al*. Randomised, double blind, placebo controlled study of fluticasone propionate in patients with moderate to severe chronic obstructive pulmonary disease: the ISOLDE trial. *Br. Med. J.* 320, 1297–1303 (2000).

53. National Institute for Health and Care Excellence. Chronic obstructive disease in over 16s: diagnosis and management [H] Economic model report. *London: NICE Guideline Updates Team* (2018).

Available at: https://www.nice.org.uk/guidance/ng115/evidence/h-economic-model-report-pdf-6602768757

Last updated: 26 July 2019.

54. National Institute for Health and Care Excellence. Chronic obstructive pulmonary disease: management of chronic obstructive pulmonary disease in adults in primary and secondary care. *London: NICE Guideline Updates Team* (2010).
Available at:
http://guidance.nice.org.uk/CG101/Guidance/pdf/English
Last updated: 26 July 2019.

55. Rodriguez-Roisin, R. Toward a Consensus Definition for COPD Exacerbations. *Chest* 117, 398S-401S (2000).

56. Celli, B.R *et al*. An Updated Definition and Severity Classification of Chronic Obstructive Pulmonary Disease Exacerbations The Rome Proposal. *American Journal of Respiratory and Critical Care Medicine* 204: 1251-58 (2021).

57. Noell G. *et al*. Multi-level differential network analysis of COPD exacerbations. Eur Respir J, 50: 1700075 (2017).

58. Burgel P.R. *et al*. Clinical COPD phenotypes: A novel approach using principal component and cluster analyses. *Eur Respir J*. 36: 531–9 (2010).

59. Garcia-Aymerich, J. et al. Identification and prospective validation of clinically relevant chronic obstructive pulmonary disease (COPD) subtypes. *Thorax*, 66:430–37 (2011).

60. Bafadhel, M. *et al.* Acute exacerbations of chronic obstructive pulmonary disease: Identification of biologic clusters and their biomarkers. *Am. J. Respir. Crit. Care Med.* 184, 662–671 (2011).

61. Burgel, P.R. *et al*. Two Distinct Chronic Obstructive Pulmonary Disease (COPD) Phenotypes Are Associated with High Risk of Mortality. *PLoS One* 7:e51048 (2012).

62. Castaldi, P.J. *et al*. Cluster analysis in the COPDGene study identifies subtypes of smokers with distinct patterns of airway disease and emphysema. *Thorax*, 69:415–22 (2014).

63. Rennard, S.I. *et al*. Identification of five chronic obstructive pulmonary disease subgroups with different prognoses in the ECLIPSE cohort using cluster analysis. *Ann Am Thorac Soc.* 12:303–12 (2015).

64. Esteban, C. *et al*. Chronic obstructive pulmonary disease subtypes. transitions over time. *PLoS One.* 11:1–16 (2016).

65. Chang, Y. *et al*. COPD subtypes identified by network-based clustering of blood gene expression. *Genomics.* 107:51–8 (2016).

66. Burgel, P.R. *et al.* A simple algorithm for the identification of clinical COPD phenotypes. *Eur Respir J.* 50:1701034 (2017).

67. Zarei, S. *et al*. Subtyping Chronic Obstructive Pulmonary Disease Using Peripheral Blood Proteomics. *Chronic Obstr Pulm Dis.* 4:97–108 (2017).

68. Ghebre, M.A. *et al.* Biological clustering supports both “Dutch” and “British” hypotheses of asthma and chronic obstructive pulmonary disease. *J Allergy Clin Immunol.* 135:63–72 (2015).

69. Ghebre, M.A. *et al.* Biological exacerbation clusters demonstrate asthma and chronic obstructive pulmonary disease overlap with distinct mediator and microbiome profiles. *J Allergy Clin Immunol.* 141: 2027–36 (2018).

70. Hurst, J.R. *et al.* Use of plasma biomarkers at exacerbation of chronic obstructive pulmonary disease. *Am J Respir Crit Care Med,* 174: 867–874 (2006).

71. Atkinson, A. J. *et al*. Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. *Clin Pharmacol Ther,* 69:89-95 *(*2001).

72. Martinez, F.J. *et al*. Role of infection and antimicrobial therapy in acute exacerbations of chronic obstructive pulmonary disease. *Expert Rev. Anti Infect. Ther.* 4(1), 101-124 (2006).

73. Bullard, M. *et al*. Early corticosteroid use in acute exacerbations of chronic airflow obstruction. *American Journal of Emergency Medicine,* 14:139-43 (1996).

74. Chen, G. & Xie, C. The efficacy and treatment length of oral corticosteroids in patients with acute exacerbations of chronic obstructive pulmonary disease. *European Respiratory Journal,* 26(Suppl 49):Abstract No. 1966 (2005).

75. Davies, L. *et al*. Oral corticosteroids in patients admitted to hospital with exacerbations of chronic obstructive pulmonary disease: a prospective randomised controlled trial. *Lancet,* 354:456-60 (1999).

76. Maltais, F. *et al*. Comparison of nebulized budesonide and oral prednisolone with placebo in the treatment of acute exacerbations of chronic obstructive pulmonary disease: a randomized controlled trial. *American Journal of Respiratory and Critical Care Medicine,* 165(5):698-703 (2002).

77. Niewoehner, D.E. *et al*. Effect of systemic glucocorticoids on exacerbations of chronic obstructive pulmonary disease. *New England Journal of Medicine,* 340:1941-7 (1999).

78. Wood-Baker, R. *et al*. A double-blind, randomised, placebo-controlled trial of corticosteroids for acute exacerbations of chronic obstructive pulmonary disease [Abstract]. *Australian and New Zealand Journal of Medicine,* 28:262 (1998).

79. Emerman, C. *et al*. A randomized controlled trial of methylprednisolone in the emergency treatment of acute exacerbations of COPD. *Chest* 95:563-7 (1989).

80. Aaron, S.D. *et al*. Outpatient oral prednisone after emergency treatment of chronic obstructive pulmonary disease. *New England Journal of Medicine,* 348:2618-25 (2003).

81. Thompson, W. *et al*. Controlled trial of oral prednisolone in outpatients with acute COPD exacerbation. *American Journal of Respiratory and Critical Care Medicine* 1996;154:407-12.

82. Bafadhel, M. *et al.* Blood eosinophils to direct corticosteroid treatment of exacerbations of chronic obstructive pulmonary disease: A randomized placebo-controlled trial. *Am. J. Respir. Crit. Care Med.* 186, 48–55 (2012).

83. Sivapalan, P. *et al.* Eosinophil-guided corticosteroid therapy in patients admitted to hospital with COPD exacerbation (CORTICO-COP): a multicentre, randomised, controlled, open-label, non-inferiority trial. *Lancet Respir. Med*. 7(8), 699–709 (2019).

84. Yun, J.H. *et al*. COPDGene and ECLIPSE Investigators. Blood eosinophil count thresholds and exacerbations in patients with chronic obstructive pulmonary disease. *J Allergy Clin Immunol.* 141:2037-2047 (2018.

85. Goto, T. *et al*. Machine Learning-Based Prediction Models for 30-Day Readmission after Hospitalization for Chronic Obstructive Pulmonary Disease. *COPD: Journal of Chronic Obstructive Pulmonary Disease* 16.5-6: 338-43 (2019).

86. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2019). URL https://www.R-project.org/.

87. Abdi, H. & Williams, L.J. Principal Component Analysis. *Wiley Interdisciplinary Reviews. Computational Statistics* 2.4: 433-59 (2010).

88. Cattell, R.B. The scree test for the number of factors. *Multivariate Behav Res*, 1:245–276 (1966).

89. Kassambara, A. & Mundt , F. factoextra: Extract and Visualize the Results of Multivariate Data Analyses. *R package version 1.0.6.* (2019).

90. Sneath, P.H. Computers in taxonomy. *J. gen. Microbiol.,* 17, 201-226 (1957).

91. Sørensen, T. A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species and Its Application to Analyses of the Vegetation on Danish Commons. *Biologiske Skrifter/Kongelige Danske Videnskabernes Selskab*, 5, 1-34 (1948).

92. Sokal, R.R., & Michener, C.D. A statistical method for evaluating systematic relationships. *University of Kansas science bulletin, 38*, 1409-1438 (1958).

93. McQuitty, L.L. Similarity Analysis by Reciprocal Pairs for Discrete and Continuous Data. *Educational and Psychological Measurement*. 26:825-831 (1966).

94. Gower, J.C. A Comparison of Some Methods of Cluster Analysis. *Biometrics*. 23: 623-637 (1967).

95. Duda, R. O & Hart, P. E. Pattern classification and scene analysis. New York: Wiley (1973).

96. Milligan, G. W., & M. C. Cooper. An examination of procedures for determining the number of clusters in a dataset. *Psychometrika* 50: 159–179 (1985).

97. Gordon, A. D. Classification. 2nd ed. Boca Raton, FL: Chapman & Hall/CRC (1999).

98. Charrad, M. et al. NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software,* 61(6), 1-36.

99. Friedman, J. *et al*. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33, 1–22 (2010).

100. Lagani, V. et al. Feature Selection with the R Package MXM: Discovering Statistically Equivalent Feature Subsets. *Journal of Statistical Software*, *80*(7), 1–25 (2017).

101. Kuhn, M. (2008). Caret package. *Journal of Statistical Software*, 28(5).

102. Venables, W.N. & Ripley, B.D. Modern Applied Statistics with S. *Fourth edition. Springer, New York* (2002).

103. Karatzoglou, A. *et al*. kernlab: Kernel-Based Machine Learning Lab*. R package version 0.9-31* (2022).

104. Breiman, L. *et al*. Classification and Regression Trees. *Wadsworth* (1984).

105. Therneau, T. & Atkinson, B. rpart: Recursive Partitioning and Regression Trees. *R package version 4.1.16* (2022).

106. Liaw, A. & Wiener, M. Classification and Regression by randomForest. *R News*, 2(3), 18-22 (2002).

107. Stekhoven, D. J. & Bühlmann, P. Missforest-Non-parametric missing value imputation for mixed-type data. *Bioinformatics* 28, 112–118 (2012).

108. Stekhoven, D. J. Package 'missForest' (2016).

109. National Institute for Health and Care Excellence. Chronic obstructive pulmonary disease in over 16s: diagnosis and management | Guidance | NICE. (2018).
Available at:
https://www.nice.org.uk/guidance/ng115/chapter/Recommendations#systemic-corticosteroids
Last updated: 26 July 2019.
Date accessed: 8th August 2022.

110. Greening, N.J. *et al.* An early rehabilitation intervention to enhance recovery during hospital admission for an exacerbation of chronic respiratory disease: Randomised controlled trial. *BMJ*. 349(7967): G4315 (2014).

111. Global Initiative for Asthma. 2020 GINA main report, Global Strategy for Asthma Management and Prevention.

Available from:

https://ginasthma.org/gina-reports/

Date accessed: 8th August 2022.

112. Yawn, B.P. Differential assessment and management of asthma vs chronic obstructive pulmonary disease. *Medscape J Med.* 11(1):20 (2009).

113. Wells, C.D. & Joo, M.J. COPD and asthma: Diagnostic accuracy requires spirometry. *J Fam Pract.* 68(2):76-81 (2019).

114. Vestbo, J. *et al.* Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease GOLD executive summary. *Am J Respir Crit Care Med*. 187:347–365 (2013).

115. British guideline on the management of asthma (2014).

116. EuroQol Group. EuroQol - a new facility for the measurement of health-related quality of life. *Health Policy*. 16: 199–208 (1990).

117. Brightling, C.E. *et al.* Induced sputum and other outcome measures in chronic obstructive pulmonary disease: Safety and repeatability. *Respir. Med.* 95: 999–1002 (2001).

118. Zigmond, A.S. & Snaith, R.P. The Hospital Anxiety and Depression Scale. *Acta Psychiatr. Scand*. 67: 361–370 (1983).

119. Jones, P.W. *et al*. Development and first validation of the COPD Assessment Test. *Eur Respir J*. 34:648-54 (2009).

120. Fletcher, C.M. *et al*. Standardised questionnaire on respiratory symptoms : a statement prepared and approved by the MRC Committee on the Aetiology of Chronic Bronchitis (MRC breathlessness score). *BMJ*. 2: 1665 (1960).

121. Juniper, E.F. Identifying 'well-controlled' and 'not well-controlled' asthma using the Asthma Control Questionnaire. *Respiratory medicine*, 100:616-21 (2006).

122. Charlson, M.E. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of chronic diseases* 40:373-83 (1987).

123. Prism 9 for macOS. GraphPad Software, LLC.

124. John, C.R. MLeval: Machine Learning Model Evaluation. *R package version 0.3* (2020).

125. Ding, B. *et al*. COPD Symptom Burden: Impact on Health Care Resource Utilization, and Work and Activity Impairment. *Int J Chron Obstruct Pulmon Dis*. 12: 677-89 (2017).

126. Kong, C.W. & Wilkinson, T.M.A. Predicting and preventing hospital readmission for exacerbations of COPD. *ERJ Open Res.* 6:00325-2019 (2020).

127. Echevarria, C. *et al*. The PEARL score predicts 90-day readmission or death after hospitalisation for acute exacerbation of COPD. *Thorax* 72**:**686-693 (2017).

128. Stone, R.A. *et al.* COPD: Who Cares When it Matters Most? National Chronic Obstructive Pulmonary Disease (COPD) Audit Programme: Outcomes from the Clinical Audit of COPD Exacerbations Admitted to Acute Units in England 2014. *National Supplementary Report. London, Royal College of Physicians*, (2017).

129. Matkovic, Z. et al. Predictors of Adverse Outcome in Patients Hospitalised for Exacerbation of Chronic Obstructive Pulmonary Disease. *Respiration,* 84:17–26 (2012).

130. Steer, J. *et al*. Dyspnoea severity and pneumonia as predictors of in-hospital mortality and early readmission in acute exacerbations of COPD. *Thorax*. 67(2): 117-21 (2012).

131. Mantero, M. *et al.* Acute exacerbations of COPD: Risk factors for failure and relapse. *Int J Chron Obstruct Pulmon Dis*. 12: 2687–2693 (2017).

132. National Institute for Health and Care Excellence. Chronic obstructive pulmonary disease (acute exacerbation): antimicrobial prescribing (NG114) (2018).

    Available at: https://www.nice.org.uk/guidance/ng114/resources/chronic-obstructive-pulmonary-disease-acute-exacerbation-antimicrobial-prescribing-pdf-66141598418629

133. Bafadhel, M, *et al*. Procalcitonin and C-reactive protein in hospitalized adult patients with community-acquired pneumonia or exacerbation of asthma or COPD. *Chest*. 139:1410-1418 (2010).

134. Prins, H.J, *et al*. CRP-guided Antibiotic Treatment in acute exacerbations of COPD admitted to Hospital. *Eur Respir J*. 53(5):180201 (2019).

135. Boersma, W.G. Antibiotics in acute exacerbations of COPD: The good, the bad and the ugly. *Eur Respir J*. 40: 1–3 (2012).

136. Trivedi, S.G. & Lloyd, C.M. Eosinophils in the Pathogenesis of Allergic Airways Disease. *Cell Mol Life Sci.* 64: 1269-289 (2007).

137. Houston, J.C. *et al*. A Clinical and Pathological Study of Fatal Cases of Status Asthmaticus. *Thorax*. 8: 207-13 (1953).

138. Laurin, C. *et al*. Impact of Anxiety and Depression on Chronic Obstructive Pulmonary Disease Exacerbation Risk. *American Journal of Respiratory and Critical Care Medicine* 185(9): 918–23 (2011).

139. Papaioannou, A. *et al.* The Impact of Depressive Symptoms on Recovery and Outcome of Hospitalised COPD Exacerbations. *Eur Respir J*. 41.4: 815-23 (2013).

140. Richter, P. et al. On the validity of the Beck Depression Inventory. A review. *Psychopathology*, 31: 160–168 (1998).

141. Bafadhel *et al.* Exacerbations of chronic obstructive pulmonary disease: time to rename. *Lancet Respir Med*; 8: 133-135 (2020).

142. National Institute for Health and Care Excellence. Chronic obstructive pulmonary disease in over 16s: diagnosis and management | Guidance | NICE. (2010).

143. Guyatt, G. Measuring health status in chronic airflow limitation. *Eur.Respir.J.* 1:560-564 (1988).

144. Jones, P.W. *et al*. A self-complete measure of health status for chronic airflow limitation. The St. George's Respiratory Questionnaire. *Am.Rev.Respir.Dis.* 145:1321-1327 (1992).

145. Galili, T. dendextend: an R package for visualizing, adjusting, and comparing trees of hierarchical clustering. *Bioinformatics* (2015).

146. Fox, J. & Weisberg, S. An {R} Companion to Applied Regression. *Third Edition. Thousand Oaks CA: Sage* (2019).

147. Murdoch *et al*. rgl: 3D Visualization Using OpenGL. *R package version 0.100.30* (2019).

148. Gutiérrez-Peña, E. & Walker, S. An Efficient Method to Determine the Degree of Overlap of Two Multivariate Distributions. *Springer Proceedings in Mathematics & Statistics* (2019).

149. Heller, R. *et al*. crossmatch: The Cross-match Test. *R package version 1.3-1* (2012).

150. Almagro, P. & Soriano, J. Underdiagnosis in COPD: a battle worth fighting. *The Lancet Respiratory Medicine*. 5(5): 367-68 (2017).

151. Claxton, S. *et al*. Identifying acute exacerbations of chronic obstructive pulmonary disease using patient-reported symptoms and cough feature analysis. *npj Digit. Med.* 4, 107 (2021).

152. Butler, C. *et al.* C-Reactive Protein Testing to Guide Antibiotic Prescribing for COPD Exacerbations. N Engl J Med 381:111-120 (2019).

153. Miravitlles, M. *et al*. Is it possible to identify exacerbations of mild to moderate COPD that do not require antibiotic treatment? *Chest* 144:1571-7 (2013).

154. Ramakrishnan, S. & Bafadhel, M. The ABRA Study - Acute exacerbations treated with Benralizumab. *ClinicalTrials.gov Identifier: NCT04098718* (2017).

155. Rabe, K.F. *et al*. Cardiovascular disease and COPD: dangerous liaisons? *Eur Respir Rev*. 27: 180057 (2018).

156. Mannino, D.M. *et al*. Prevalence and outcomes of diabetes, hypertension and cardiovascular disease in COPD. *Eur Respir J;* 32: 962–969 (2008).

157. Patel, A.R. *et al*. The impact of ischemic heart disease on symptoms, health status, and exacerbations in patients with COPD. *Chest*, 141: 851–857 (2012).

158. Berry, M.A. *et al*. The use of exhaled nitric oxide concentration to identify eosinophilic airway inflammation: an observational study in adults with asthma. *Clin.Exp.Allergy,* 35:1175-1179 (2005).

159. Harvey, A.C. Forecasting, structural time series models and the Kalman filter. *Cambridge University Press* (1990).

160. Moritz, S. & Bartz-Beielstein, T. imputeTS: Time Series Missing Value Imputation in R. *The R Journal*, *9*(1), 207-218. (2017).

161. Paparrizos, J. & Luis, G. K-Shape: Efficient and Accurate Clustering of Time Series. *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data,* 2015: 1855-870 (2015).

162. Chan, M. C. *et al*. Blood Eosinophil and Risk of Exacerbation in Chronic Obstructive Pulmonary Disease Patients: A Retrospective Cohort Analysis. *International journal of chronic obstructive pulmonary disease,* 15: 2869-2877 (2020).

163. Perera, W.R. *et al*. Inflammatory changes, recovery and recurrence at COPD exacerbation. *Eur Respir J;* 29: 527-534 (2007).

164. Sama, R.S. *et al*. Environmental triggers of COPD symptoms: a case cross-over study. *BMJ Open Respir Res.* 4(1):e000179 (2017).

165. Seemungal, T.A. *et al.* Time course and recovery of exacerbations in patients with chronic obstructive pulmonary disease. *Am J Respir Crit Care Med*, 161(5):1608-13 (2000).

166. Anthonisen, N. R. *et al*. Antibiotic therapy in exacerbations of chronic obstructive pulmonary disease. *Ann. Intern. Med.* 106:196-204 (1987).

167. Vollenweider, D.J. *et al*. Antibiotics for exacerbations of chronic obstructive pulmonary disease. Cochrane Database Syst Rev. 10(10):CD010257 (2018).

168. Halner, A, *et al*. Predicting treatment outcomes following an exacerbation of airways disease. *PLoS One.* 16(8):e0254425 (2021).