# University of Southampton

Faculty of Medicine

Clinical and Experimental Sciences

**Exhaled Volatile Organic Compounds as Biomarkers for Airway Biology in Severe Asthma**

by

**Adnan Azim**

Thesis for the degree of Doctor of Philosophy

January 2023

# University of Southampton

## Abstract

Faculty of Medicine

Clinical and Experimental Sciences

Doctor of Philosopy

Exhaled Volatile Organic Compounds as Biomarkers for Airway Biology in Severe Asthma

by

Adnan Azim

Breathomics, the measurement of exhaled volatile organic compounds (VOCs), is an exciting new biomarker medium for airways disease. The greatest unmet need for biomarkers in severe asthma is in T2 low disease and so, in this thesis, I sought to identify a T2 low phenotype and assess whether breathomics could be used as a biomarker for this patient group.

A cohort of severe asthma patients was recruited and clinically characterised in parallel to sputum induction and exhaled breath collection. Though the T2 high phenotype was easy to recognise, T2 low disease was poorly defined by inflammatory cell counts alone. Measures of inflammatory cell activation provided were insufficient to describe new phenotypes.

16s rRNA sequencing of sputum samples identified a cohort of T2 low patients, characterised by airway colonisation with *Haemophilus*, sputum neutrophilia and ongoing disease burden. However, none of the clinically available biomarkers were able to identify this cohort of patents.

The exhaled VOC samples from this severe asthma cohort demonstrate a clear structure to the exhaled VOC matrix, however, sensitivity to underlying airway inflammation was weak. Repeated breath sampling identified heterogeneity in the stability of VOCs during an otherwise clinically stable state. Exclusion of VOCs with a high degree of within-subject variability resulted in less model overfitting and AUC an of 0.643 for predicting sputum eosinophilia ($\geq$2%). 2-pentanone, was identified as having the strongest feature importance. This ketone is thought to be generated in the airway epithelium.

Applying this newly established analytical framework, a model was built to predict the cluster of patients with heavy *Haemophilus* colonisation, potentially amenable to Azithromycin therapy. A model built on non-erratic VOCs predicting *Haemophilus* with an AUC of 0.857. Decane was identified as a possible biomarker, however further validation is required.

The findings from this thesis demonstrate sensitivity of exhaled VOCs to the airway biology of severe asthma patients but require validation.

# Table of Contents

# Table of Tables

Table of Tables

# Table of Figures

# Research Thesis: Declaration of Authorship

Print name: Adnan Azim

Title of thesis: Exhaled Volatile Organic Compounds as Biomarkers for Airway Biology in Severe Asthma

I declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:-

Azim A, Barber C, Dennison P, et al. Exhaled volatile organic compounds in adult asthma: a systematic review. Eur Respir J 2019;54(3).

Azim A, Mistry H, Freeman A, et al. Protocol for the Wessex AsThma CoHort of difficult asthma (WATCH): a pragmatic real-life longitudinal study of difficult asthma in the clinic. BMC Pulm Med 2019;19(1):99.

Azim A, Rezwan FI, Barber C, et al. Measurement of Exhaled Volatile Organic Compounds as a Biomarker for Personalised Medicine: Assessment of Short-Term Repeatability in Severe Asthma. J Pers Med 2022;12(10).

Signature: ...................................................................... Date: 14/01/23

# Acknowledgements

First and foremost, I would like to thank all the patients that I have had the pleasure to see in both clinical and research settings throughout the course of my fellowship. Being a part of patient centred research has been one of the most inspiring and rewarding times of my career. I am forever grateful to the patients that took part in my clinical research, the dedicated staff associated with the NIHR Southampton Biomedical Research Centre and the stellar collaborators that I have had the fortune and pleasure of working with. None of the work presented in this thesis would be possible without them.

*Undertaking a PhD is hard* – I wish to give thanks to my supervisors for their invaluable guidance. David's support has been steadfast and unwavering; I could not have completed my thesis without the technical and logistic assistance that he gave me and will be forever indebted to the expertise he imparted on me. I am similarly indebted to John and his dedication to me as a student; his patience and humility is nothing short of inspirational. Lastly, whilst I recognise that Peter has provided me with many wonderful opportunities, I will always be grateful for the freedom he gave me to find my own path. There are very few people who have impacted the trajectory of my life more than he has.

*Undertaking a PhD is lonely* – I wish also to give special thanks to my fellow students, but in particular to Clair, who I could not have completed this journey without. Our friendship has sustained me during the best and worst times of my fellowship. More than that, however, no-one has challenged me to aspire to the highest standards of ethical and scientific conduct like she has.

*Undertaking a PhD is selfish* – Words cannot express how grateful I am for the love and patient forbearance of my wife. This thesis is dedicated to her.

# Definitions and Abbreviations

| | |
|---|---|
| ACQ6 | Asthma Control Questionnaire (6 Point) |
| ANA | Anti Nuclear Antibody |
| ANCA | Antineutrophil Cytoplasmic Antibodies |
| ANOVA | Analysis of Variance |
| APC | Antigen Presenting Cells |
| ARTP | Association for Respiratory Technology and Physiology |
| ASV | Amplicon Sequence Variant |
| ATS | American Thoracic Society |
| AUC | Area Under the Curve |
| BAL | Bronchoalveolar Lavage |
| BD | “bis die” (twice daily) |
| BDPe | Beclomethasone Dipropriate equivalent |
| BLAST | Basic Local Alignment Search Tool |
| BMI | Body Mass Index |
| BRC | Biomedical Research Centre |
| BTS | British Thoracic Society |
| CD4 | Cluster of Differentiation 4 |
| CO2 | Carbon Dioxide |
| COPD | Chronic Obstructive Pulmonary Disease |
| CRF | Clinical Research Facility |
| CT | Computed Topography |
| $CV_{MAD}$ | Coefficient of Variation (Median Absolute Deviation) |
| DEXA | Dual-Energy X-ray Absorptiometry |
| DNA | Deoxyribonucleic acid |
| DTE | Dithioerythritol |
| ECP | Eosinophilic Cationic Protein |

Definitions and Abbreviations

| | |
|---|---|
| EDN | Eosinophil Derived Neurotoxin |
| EGAPP | Evaluation of Genomic Applications in Practice and Prevention |
| ENT | Ear Nose and Throat |
| ERS | European Respiratory Society |
| EVOC | Exhaled Volatile Organic Compound |
| EVOC4M | Exhaled Volatile Organic Compounds for Mepolizumab Study |
| F1 | F Score *(classification performance derived from precision and recall)* |
| FBC | Full Blood Count |
| FC | Fold Change |
| FDR | False Discovery Rate |
| FEF$_{25-75}$ | Forced Expiratory Flow at 25 and 75% of the Pulmonary Volume |
| FeNO | Fractional Exhaled Nitric Oxide |
| FEV$_1$ | Forced Expiratory Volume in 1 second |
| FID | Flame-Ionization Detection |
| FVC | Forced Vital Capacity |
| FWD | Forward |
| GC | Gas Chromatography |
| GCMS | Gas Chromatography Mass Spectrometry |
| GI | Gastrointestinal |
| GLI | Global Lung Initiative |
| GORD | Gastro-Oesophageal Reflux Disease |
| HADS | Hospital Anxiety and Depression Score |
| HADSTOT | Hospital Anxiety and Depression Score (Total) |
| HRCT | High Resolution Computed Tomography |
| ICS | Inhaled Corticosteroid |
| ID | Identification |
| IgE | Immunoglobulin E |
| IL | Interleukin |

| | |
|---|---|
| ILC2 | Type 2 Innate Lymphoid Cell |
| IMI-3TR | Innovative Medicines Initiative – Taxonomy, Treatments, Targets and Remission |
| ITS | Internal Transcribed Spacer |
| LOD | Limit of Detection |
| MDS | Multidimensional Scaling |
| MDT | Multidisciplinary Team |
| MF | Molecular Feature |
| MG | Milligram |
| MHC | Major Histocompatibility Complex |
| mOCS | Maintenance Oral Corticosteroids |
| MPO | Myeloperoxidase |
| mRNA | Messenger Ribonucleic acid |
| MS | Mass Spectrometry |
| MSD | Meso Scale Discovery |
| m/z | Mass to Charge Ratio |
| NE | Neutrophil Elastase |
| NHS | National Health Service |
| NHSE | National Health Service England |
| NIHR | National Institute for Health and Care Research |
| NIST | National Institute of Standards and Technology |
| NMDS | Non-Metric Multidimensional Scaling |
| NO | Nitric Oxide |
| NPV | Negative Predictive Value |
| NY | New York |
| OCS | Oral Corticosteroid |
| OGD | Oesophago-Gastro-Duodenoscopy |
| OR | Odds Ratio |

| | |
|---|---|
| PAM | Partition Around Medoids |
| PBS | Phosphate-Buffered Saline |
| PC | Principal Component |
| PCA | Principal Component Analysis |
| PCR | Polymerase Chain Reaction |
| PG | Paucigranular |
| ppb | Parts per Billion |
| PPV | Positive Predictive Value |
| PTH | Parathyroid Hormone |
| QAH | Queen Alexandra Hospital |
| QC | Quality Control |
| REC | Research Ethics Committee |
| REV | Reverse |
| RFE | Recursive Feature Elimination |
| RNA | Ribonucleic Acid |
| rRNA | Ribosomal Ribonucleic Acid |
| ROC | Reciever Operating Characteristic |
| SABA | Short Acting Beta Agonist |
| SD | Standard Deviation |
| SNOT20 | 20 Point SinoNasal Outcome Test |
| SOP | Standard Operating Procedure |
| STARD | Standards for Reporting of Diagnostic Accuracy Studies |
| TAC | Transcriptomic Associated Cluster |
| TD | Thermal Desorption |
| TDA | Topological Data Analysis |
| $Th_1$ | T Helper 1 |
| $Th_2$ | T Helper 1 |
| TOF | Time of Flight |

| | |
|---|---|
| TNF | Tumour Necrosis Factor |
| TRIPOD | Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis |
| U-BIOPRED | Unbiased Biomarkers for the Prediction of Respiratory Disease Outcomes |
| UHSFT | University Hospital Southampton Foundation Trust |
| UK | United Kingdom |
| USA | United States of America |
| VOC | Volatile Organic Compound |
| WATCH | Wessex Asthma Lifetime Cohort study |
| XGB | Extreme Gradient Boosting |

# Chapter 1 Introduction

## 1.1 Asthma

### 1.1.1 Definition of Asthma

The term “asthma” is derived from the Greek root “ασθμαινω”, which means “to gasp for breath”, and was originally used to describe non-specific respiratory symptoms before narrowing to its modern use as a diagnostic label [1]. In the mid nineteenth century, Henry Hyde Salter described “*paroxysmal dyspnoea of a peculiar character with intervals of healthy respiration between attacks*”[2], which is remarkably similar to the twenty-first century description from the Global Initiative for Asthma of “*a heterogenous disease, usually characterised by chronic airway inflammation. It is defined by the history of respiratory symptoms such as wheeze, shortness of breath, chest tightness and cough that vary over time and in intensity, together with variable expiratory airflow limitation*” [3]. Today, over 300 million people worldwide [4], including 1 in 12 adults in the UK [5], have asthma and its incidence appears to be rising [6,7].

Defining a disease is central to the philosophy of medicine and, traditionally, requires distinguishing the disease state from normal healthy state [8]. Value judgements on “normality” [9,10] aside, this is relatively straight-forward for the majority of diseases [11]. For asthma, however, definitions tend to avoid etiological implications [12]. This is in part due to the diverse clinical presentation of asthma, which have long been understood to reflect the complex interplay of genetic and environmental components [13] interacting to influence disease expression [14]. Instead, focus has been aimed at observable characteristics (clinical, biological, and physiological) and the description of distinct phenotypes [15,16], resulting in treatment paradigms that are based upon disease severity rather than their underlying mechanisms.

### 1.1.2 Burden of Asthma

The greatest burden of asthma comes from those with uncontrolled disease: a definition that usually captures one or more undesirable consequences: frequent exacerbation rate [17,18], poor lung function [19] poor quality of life [20-22] or death [23]. These patients account for the majority of asthma related healthcare expenditure [24,25], costing the National Health Service (NHS) an estimated £1 billion/year in addition to the hidden societal costs of disability, missed schooling and lost work days[26]. Most people with asthma respond well to standard preventer therapies (inhaled corticosteroids with or without long acting beta$_2$ agonists) [3] so the majority of uncontrolled asthma can be managed by addressing factors such as poor medication adherence, significant co-morbidities (e.g. rhinitis, gastro-oesophageal reflux, obesity and psychological co-morbidities) and external triggers (e.g. allergens and environmental factors) that may be contributing to poor disease control[27].

However, some 3-10% [28,29] of patients with asthma remain poorly controlled despite escalation of preventer therapies and addressing of the factors described above. These patients have severe asthma, as defined by the International ERS/ATS guidelines: "*asthma which requires treatment with high dose inhaled corticosteroids (ICS) plus a second controller (and/or systemic corticosteroids) to prevent it from becoming 'uncontrolled' or which remains 'uncontrolled' despite this therapy*"[28]. This definition of severe asthma highlights the limitations of ignoring the aetiology of asthma and of a severity-based approach to therapy; their recognition has heralded a shift towards understanding the mechanisms underlying asthma, complex though they may be.

### 1.1.3 Pathophysiology of Asthma

Before tackling the aetiology of asthma, it is worth considering the purpose of the organ of interest: the airways. One of the primary roles of the lung is to facilitate gas exchange between the circulatory system and the external environment. As such, the airways are organised into a branching configuration so as to maximise the surface area of respiratory bronchioles and alveoli, which participate in gas exchange [30]. However, exposing a large mucosal surface area to the environment leads to its constant exposure to external stimuli, which demands the robust discrimination of what is harmful and what is not [31]. Unsurprisingly, an orchestra of mechanisms is required to fulfil this mandate, commonly involving epithelial cells, fibroblasts, endothelial cells and smooth muscle cells, inflammatory mast cells, eosinophils and T lymphocytes.

None of these should be considered in isolation but T lymphocytes (T Cells) are widely recognised as central to the asthma immune response. Derived from pluripotent haematopoietic stem cells in the bone marrow, T cells mature in the thymus (hence their nomenclature) [32] and can be described by their gene or protein expression and associated functions. $CD4^+$ T cells, otherwise known as T Helper Cells ($T_H$ cells), are activated by peptide antigens presented by MHC class II molecules on antigen presenting cells (APCs) [33] and, based on their cytokine repertoire, can be further subdivided into two functionally distinct subsets: $T_H1$ (defence against intracellular bacteria, viruses and cancer) and $T_H2$ (defence against parasites and allergens) [34]. This dichotomy was the basis for traditional asthma dogma [35], in which asthma was understood to represent an exaggerated specific IgE mediated [36] $T_H2$ cell response [37-40] driving airway hyperresponsiveness [41] and responsive to steroid therapy.

However, the varied response of asthma patients to steroid treatment [42,43], the identification of persistent eosinophilia in patients treated with high dose inhaled steroids [44,45] and the identification of eosinophilia in patients without strong atopic features [46-49] illustrates that the $T_H2$ paradigm is incomplete. We now recognise that T2 cytokines (IL-4, IL-5, IL-13) can be produced by pathways not directly related to $T_H2$ cells [50,51] and the past few decades have seen the detailed description of the immunology of asthma [52].

### 1.1.4 Heterogeneity of Severe Asthma

Asthma is now widely appreciated to be an umbrella term encompassing a number of endotypes, disease entities defined by specific biologic mechanisms [53], manifesting with similar clinical presentations [54]. We can appreciate that phenotypes (e.g. the presence of eosinophils) can be driven by multiple mechanisms but also that an endotype (e.g. airway remodelling) can be associated with multiple phenotypes [55]. The current paradigm of phenotyping patients offers limited insight into the mechanisms driving poor disease control and contribute to the varied response to treatments given ubiquitously. Shifting focus to endotypes promises the potential that specific therapies can be appropriately prescribed to improve asthma control [56].

## 1.2 A Systems Biology Approach to Severe Asthma

### 1.2.1 Biomarkers

A biomarker is "*a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention*" [57]. In addition to its utility in drug discovery (e.g. target identification) and development (e.g. target engagement) [58], biomarkers can serve a number of clinical purposes: diagnostic (e.g. presence or absence of a disease), prognostics (e.g. indicate severity and outcomes of a disease) and theragnostics (e.g. predictors of response to treatments) [59].

Several biomarkers are described in asthma, most notably sputum eosinophilia, expressed as a percentage of inflammatory cells [60]. Airway eosinophilia can predict the response to inhaled steroids [61,62], is altered by steroid therapy [63,64] and can be used to titrate asthma therapy [42]. It is unsuitable, however, for routine clinical practice or large epidemiological studies due to the practical limitations of undertaking sputum induction in a clinical setting.

Correlation between blood and sputum eosinophil counts [65,66] has facilitated the emergence of blood eosinophilia as an alternative biomarker. Raised blood eosinophil counts are associated with poor asthma control [67], lung function decline [68] and exacerbations [69,70]; they have become central in defining a phenotype of severe asthma patients [71,72] that respond to anti-IL5/ IL-5 receptor alpha (IL-5 Rα) therapies [73-76] (even if selection by this criteria is not perfect [77]).

Nitric oxide, synthesised by NO synthetases [78], can be measured in exhaled breath by chemiluminescence [79,80]. The fractional exhaled NO concentration (FeNO) in exhaled breath, expressed as parts per billion (ppb), though only modestly accurate in predicting sputum eosinophilia [81], can predict the risk of exacerbations [82,83] and is altered by steroid therapy [84]. This allows serial FeNO measurements to be used as a measure of inhaled corticosteroid compliance [85].

Numerous other biomarkers are described in asthma but are generally biased towards T2 cytokine driven disease [86]. Non-T2 driven disease, defined by the absence of measurable T2 cytokine activity, is poorly described and has no robust biomarkers associated with it [87,88], despite these patient groups being most treatment-resistant and having the worst clinical outcomes [47,89,90].

### 1.2.2 Metabolomics

Unpicking the complexity of asthma pathophysiology requires a shift towards appreciation that dysregulation of physiological mechanisms occurs within a network of other closely related/regulated mechanisms [91]. This system level approach to biology was first proposed many decades ago [92] but has recently gained momentum in clinical research through advances in a number of inter-disciplinary fields: biology, mathematics, statistics and computer sciences [93]. “Systems biology” is difficult to define [94] but aims to identify general principles of a system, through comprehensive study of its molecular diversity [95]. In clinical research, molecular diversity can be described using high-throughput measurement of biomolecules: genomics for DNA, transcriptomics for RNA transcripts and proteomics for translated proteins from various biological samples (biofluids, cells or tissues).

However, it is not enough to simply list the components of a system: it is necessary to understand how these components fit together, how they behave under different conditions and what their regulatory mechanisms involve [96]. Systems biology therefore requires exploration of these multi-dimensional datasets through sophisticated mathematical and computational modelling [97,98]. Typically, this describes an unbiased approach that makes few a priori assumptions, allowing the data to generate new hypotheses. This approach is well suited to asthma due to its biological complexity and gene-environment interactions [99].

Metabolomics describes the systems biology approach to small (typically measuring 50-1500 daltons) molecules measured in biofluids, cells or tissues [100]. These small molecules (e.g. lipids and proteins) are the substrates and products of metabolism, and so metabolomics aims to provide a snapshot of the biochemical activity associated with a cellular state. Metabolomic profiling of serum and urine samples has previously demonstrated differences between healthy controls and asthma [101] as well as between clinical phenotypes [102-104] making it perfectly suited to biomarker discovery [105].

In identifying the need for more biomarkers for severe asthma, it is useful to consider how to judge the quality of a biomarker. The Evaluation of Genomic Applications in Practice and Prevention (EGAPP) Initiative of the United States Centers for Disease Control propose three measures, in addition to considering the legal and social issues around that test: analytical validity, clinical validity and clinical utility [106]. These can broadly be summarised by asking whether the biomarker test result is true, meaningful and useful, respectively [107].

Those biomarkers that originate from the organ of interest are more likely to offer direct insight into airway biology [108]. However, in order to have the highest translational potential, biomarkers need to be easily sampled at the point of care [86]. Where sputum and bronchoscopic sampling fall at this hurdle [109], exhaled breath offers the optimal combination of proximity to airways and minimal invasiveness [110]. The clinical adoption of $C^{13/14}$ urea for *Helicobacter pylori* [111]and FeNO for airway inflammation [112] demonstrates the utility of exhaled breath: this medium can be sampled safely, non-invasively and repeatedly, almost without exhaustion (e.g. capnography [113]).

## 1.3 Breathomics as a Biomarker for Asthma

### 1.3.1 Definition

Though primarily composed of water vapour and inert gases, exhaled breath also contains thousands of volatile organic compounds (VOCs) [114]. VOCs are the main molecular substrate triggering our sense of smell, and characteristic breath odours have been used to identify illnesses since Hippocrates[115]. The medical applications of this strategy were transformed in 1971 by the demonstration that exhaled breath contained more than 250 VOCs [116]. Though exhaled breath only constitutes a small proportion of metabolomics research, [117], it is recognised that exhaled VOC concentrations can reflect different disease states [118,119], suggesting a role for exhaled VOC analysis as a non-invasive metabolomic biomarker [115].

### 1.3.2 Methodological Overview

Modern VOC analysis can be considered under one of two broad methodological headings: pattern recognition-based sensors or chemical analytical techniques [120]. Pattern recognition-based sensors, synonymous with electronic noses, are modelled on the mammalian nose. These e-noses contain an array of cross-reactive sensors, which react promiscuously and non-selectively to VOCs.  The final “breathprint” reflects the differential signalling of multiple sensors to partially overlapping VOCs [121,122]. These breathprints then require analysis by pattern recognition algorithms [123,124]; similar to how our brain would interpret signals from the nose. Chemical analytical techniques typically refer to mass spectrometry (MS) [125]or MS Hybrid techniques [126-128], in which ions created by VOCs can be measured based upon their mass/charge (m/z) ratio.

Mass spectrometry continues to be the gold standard method for VOC analysis [129] due to its ability to identify composition and concentrations of individual VOCs. It is, however, expensive in terms of expertise and equipment. Though the e-nose sensor array is cheaper, quicker and easier to use [130,131], it sacrifices that ability to reliably trace back to analytes of interest [132]. If understanding the mechanisms behind a biomarker/biomarker profile is not necessary [133], sensor-based systems are ideally suited, but if the biomarker is being interrogated for mechanistic purposes [87] (e.g. drug target discovery), then the costs savings in pursuing pattern recognition-based sensors over chemical analytical techniques may represent a false economy. That is not to say that the two technologies are mutually exclusive: both can be successfully integrated into the same study [134], and improved understanding of the mechanisms and factors contributing to VOC profiles could inform future sensor-based technologies.

### 1.3.3 Breathomics in Asthma

Exhaled VOC analysis has demonstrated excellent accuracy for discriminating patients with asthma from healthy controls and other respiratory conditions [135]. More than diagnostics, however, a number of studies have successfully applied exhaled VOCs as a biomarker for inflammatory phenotyping, treatment stratification, treatment monitoring and exacerbation assessment {Azim, 2019 #293

Almost all the identified discriminatory VOCs identified in asthma studies are straight-chain, branched, or aromatic hydrocarbons [157] and whilst there are proposed endogenous origins to such compounds [158], they frequently occur exogenously [159] and may simply reflect differential uptake of environmental VOCs [160]. Consequently, those VOCs which occur commonly and in high concentrations or those that are common to many inflammatory states [161,162], are over-represented in the literature

For airways diseases, it is likely that, only a fraction of the VOCs identified in exhaled breath relate to airway events. While exhaled breath sampling is simple for patients, the analyte of interest is by definition, volatile, and each stage of exhaled breath analysis introduces additional sources of variation [163].

Just like exhaled nitric oxide concentrations [80,164], individual VOCs concentrations [156] and e-nose sensor deflections [123] can be influenced by flow rate and breath holding [123]. Similarly, most studies collect VOCs from the total expiratory phase, as it is logistically most simple to achieve. E-nose breathprints using this breath sample show more inferior diagnostic potential (asthma vs healthy controls) than when excluding air within the oropharynx [138], possibly reflecting the dilution of discriminatory VOCs by this contaminant air. There has been a range of strategies used to exclude air from this dead space, including valves and estimated volumes, which are inconsistent and unreliable [130,165], to a highly engineered system of pressure sensors [142,166,167], which is expensive and bulky. The optimal solution needs to balance practicality and precision, so as not to negate the clinical utility of breath sampling. The breath sampler developed by a broad consortium of breath researchers and engineers (http://www.breathe-free.org), represents one such solution. A similar pragmatism is likely necessary for the exclusion of exogenous VOCs. Subtracting ambient VOC concentrations from exhaled VOC concentrations (alveolar gradient) [168] risks the loss of salient signals and ignores VOC interactions within the airways [158,169]. Though not able to eliminate all exogenous VOCs, filters can at least reduce background contamination [131].

Once collected, unless performing online analysis, the storage medium bears consideration. Most early studies stored breath samples in an inert polymer (usually Tedlar) bag, but the concentrations of compounds stored in Tedlar bags show compound-specific decay rates and the bags themselves can introduce contaminants [157,170]. Van der Schee *et al*. found no variation when breath samples were stored for up to two weeks [155], and many studies try to minimise storage time [131,138]. Alternative or subsequent storage solutions: thermal desorption tubes containing some sort of adsorbent material: porous organic polymers, activated charcoal, carbon molecular sieves or graphitized carbon blacks do not guarantee against this decay [157]. No adsorptive material can completely capture all the VOCs in the breath without some degree of loss [171].

Moreover, different materials are vulnerable to breakthrough (non-quantitative adsorption of analytes) and memory effect (incomplete desorption resulting in interference with subsequent measures) [172]. Ideally, therefore, the choice of adsorbent materials and the duration of storage [155] should be determined by compounds of interest [157,173]. Most studies now adsorb onto Tenax TA (2,6-diphenyl-*p*-phenylene oxide) [174], due to its hydrophobicity, thermal stability and its ability to absorb a wide range of VOCs [175]. To add further complications, it is also appreciated that not all VOCs originate from the airways. Non-asthma-VOC research proposes a model of blood/gas coefficients [176] and quantification of regional lung ventilation and perfusion [177] to describe the delivery and migration of systemically generated VOCs to exhaled breath.

Even then, a robust and reliable sample collection is relatively straightforward when compared to data management, analysis, and interpretation [178]. The past few decades of asthma-VOC literature demonstrate the rapid technological evolution of this field. Early breath research was limited to a "bottom-up" approach: VOCs were targeted *a priori* and analysed using expensive and laborious chemical techniques. Consequently, these studies were limited in numbers and focussed on markers of oxidative stress [152] (inflammation not specific to asthma [149,162]). The modern parallel developments of improved separation techniques, improved lower limits of detection, electronic nose (e-nose) technology and high-throughput omics analysis platforms allow the full spectrum of exhaled VOCs to be analysed "top-down" as highly dimensional composite profiles: "breathomics".

In the absence of a clear consensus on the optimal statistical approach, studies are likely to publish highly internally valid results which, in the absence of external validation, likely overestimate real-world findings [179]. This influence of data handling on biomarker identification means that transparency is more important than ever. The TRIPOD recommendations on reporting multivariable prediction models [180] and STARD guidelines on reporting of diagnostics accuracy studies [181] provide useful frameworks for future publications.

The methodological heterogeneity of the breath analysis literature is well documented [163,182] and yearns for methodological standardisation [163,182]. Metabolomics studies demand standard operating protocols for both the analytical and computational workflows [157], including strategies to monitor within- and between-batch measurement variations [183]. Sharing meta-data relating to sample handling, processing and analysis is of paramount importance and represent a critical step in building reference libraries [184,185]with the ultimate intention for breathomics research being inter-laboratory and equipment comparison [186]. Only then will true external validation, where findings are replicated in a new study, be possible [187]. Nevertheless, expert consensus is optimistic for a role of exhaled VOCs in delivering precision medicine for asthma [188]: the right treatment for the right patient at the right time [189].

## 1.4 Knowledge Gap: Improving Breathomics Study Endpoints in Asthma

> “Late at night, a police officer finds a drunk man crawling around on his hands and knees under a streetlight. The drunk man tells the officer he’s looking for his wallet. When the officer asks if he’s sure this is where he dropped the wallet, the man replies that he

> thinks he more likely dropped it across the street. Then why are you looking over here? the befuddled officer asks. Because the light's better here, explains the drunk man."[190]

The chemical and data handling issues described above are of paramount importance to driving the field of breathomics forwards. However, parallel to these efforts, there is also a need to innovate the study designs of breathomics research.. One of the unique selling points for a breathomics based biomarker is its ease of sampling and, therefore, potential to translate to clinical practice [110]. It is highly conceivable that an e-nose type device could be used in an outpatient clinical setting but this ambition is undermined by the fact that easily accessible biomarkers are already established in routine clinical practice [3,81,196] , including an exhaled breath biomarker (FeNO). To augment or join FeNO and blood eosinophils in a clinical setting, a breathomics biomarker either needs to be more accurate at predicting a feature that can already be predicted or predict something that cannot currently be predicted.

### 1.4.1 Breathomics as a More Accurate Biomarker for Existing Phenotypes

In considering more accurate inflammatory phenotyping, there remains, of course a ceiling effect: the development of a breathomics biomarker in order to achieve an AUC of 0.9, when biomarkers with an AUC of 0.7 are already established [197], may not be financially or practically viable. In this aforementioned Belgian study [197], it is striking that, despite being a well conducted study in a large number of patients, exhaled VOCs only achieve an AUC of 0.7, comparable to blood eosinophil and FeNO alone, rather than the oft promised (albeit in studies lacking validation cohorts) AUCs exceeding 0.9 [136]. One conclusion is that there needs to be improvement in the breathomics technology (i.e., chemical and data handling).

Alternatively, it could be that an AUC of 0.9 is not biologically plausible for a metabolomics biomarker. It is well recognised that airway eosinophilia can arise from multiple mechanisms [50,52]. Indeed, omics analysis of the airways samples of severe asthma patients identify a number of eosinophilic and neutrophilic sub-phenotypes [198]. Biomarkers can only be as good as the gold standard against which they are assessed [199]; if VOCs are specifically related to only one of these distinct mechanisms (e.g. a non-allergic, ILC2 mediated eosinophilia rather than an allergic Th2 cell mediated eosinophilia [50]), using an umbrella phenotype (e.g. sputum eosinophilia) as the study endpoint would miss that relationship.

Reducing the application of VOC biomarkers to predicting inflammatory phenotypes does not comprehensively address the ability of breathomics to describe airway inflammation, which may be better described by more detailed analysis of induced sputum [47,198]. Exploratory analysis in the U-BIOPRED cohort, for example, suggest that e-nose, can discriminate clusters defined by sputum transcriptomics [153] more effectively than existing biomarkers. This finding underlines the notion that, though sputum eosinophils are clinically relevant endpoints in themselves [42], they are also imperfect biomarkers for characterisation efforts at a molecular level [198].

### 1.4.2 Breathomics as a Biomarker for Novel Phenotypes

Taking inspiration again from the Belgian study [197], the authors report an AUC of 0.73 for predicting airway neutrophilia, an AUC comparable to existing biomarkers used to predict airway eosinophilia. This is one example for how a breathomics biomarker may offer novel information to the clinician as sputum neutrophils represent one paradigm for describing T2 low asthma. T2 low asthma is poorly defined but describe patients who do not display T2 high signals; these patients are characterised by resistance to steroid therapy and ineligibility to currently licenced biologic therapies, with pathophysiology putatively ascribed to Th1 and/or Th17 cells [202].

Though sputum neutrophils are a candidate marker for non T2 driven disease [203], inconsistent cut-offs [46,192,204], concerns of possible confounding as a product of steroid therapy [205] and the finding that therapeutic reductions in circulating neutrophilia have not reduced the exacerbation frequency of patients with severe uncontrolled asthma [206] undermine its value as a robust phenotype [207]. Consequently, comparing breathomics to such a poorly defined gold-standard may once again undermine the value of breathomics.

As described in 1.4.1, one way to better assess breathomics is to once again, better define the T2 low phenotype. The association, between sputum neutrophilia and airway colonisation by potentially pathogenic bacteria [208] suggest that host-microbial interactions might be a better way to understand these poorly-characterised asthma endotypes [209]. The resident microbiome has been found to have a key role in the establishment and maintenance of healthy gastrointestinal tract and critical in understanding chronic inflammatory diseases of this organ [210], which like the respiratory tract is also a large mucosal surface area with exposure to external stimuli.

Microbiome analysis, whether clinically or academically, suffers from a reliance on airway samples, which as we have discussed in 1.2 is impractical and expensive. Beyond the obstacles in airway sampling issues, microbiome analysis of the airways suffers from the fact that sputum samples are not as rich in bacterial matter as samples from the GI tract [31]. However, bacteria are abundant producers of VOCs [211-213] and VOCs from bacteria are likely to contribute to the spectra of exhaled VOCs. As such it is plausible that exhaled VOCs could be used as a biomarker for describing the airway microbiome, which could in turn help better phenotype T2 low asthma patients, who have the greatest unmet need for biomarkers.

### 1.4.3 Breathomics to Define New Phenotypes

Finally, clustering has been at the heart of efforts to understand the heterogeneity of asthma for the past few decades. Originally applied to clinical variables [46,48,214], it is now commonly applied to omics type data [198,215]. Clustering is an unsupervised machine learning technique that seeks to identify the similarity of patients across a defined set of variables and grouping them on that basis [216]. When applied to omics data, it can group patients with similar biology together and separate patients with distinct biology [217].

Few studies in asthma have considered clustering on VOCs[218], however, metabolites, in addition to reflecting cellular genetic information and its mRNA expression (measured by genomics and transcriptomics respectively), can be influenced by or arise from exogenous sources: micro-organisms, xenobiotics (e.g. drug and environmental pollutants) and dietary sources [219,220]. Of the many thousands of VOCs in exhaled breath [114], a proportion of these relate to age [221], gender [222], diet [223], exercise [224] and smoking [225], features associated with T2 low disease [202]. Some of these factors, such as asthma therapy [154,226], the resident microbiome [120] and environmental exposures [133,157,227-229] may be highly relevant to understanding severe asthma.

These factors are likely to be highly salient to asthma, which is characteristically a product of gene-environment interactions. Clustering on this information may provide novel and meaningful insight, particularly to T2 low asthma which remains poorly defined using conventional techniques.

## 1.5 Aims and Objectives

The aim of this thesis is to understand whether exhaled volatile organic compounds can give an insight into biologic events in the airways of severe asthma patients.

The objectives of the thesis are to:

- Characterise an asthma cohort and confirm that it is representative of a severe asthma population in which a breathomics biomarker would be useful by applying descriptive on their clinical characteristics and sputum inflammatory cell counts.
- Assess the value of granulocyte activation markers as possible biomarkers of airway inflammation by using descriptive statistics across sputum inflammatory phenotypes and supervised machine learning approaches to predict those phenotypes in this cohort.

- Describe the airway microbiome using 16S rRNA sequencing on sputum samples from this cohort using multivariate statistical approaches to compare across inflammatory phenotypes and unsupervised machine learning to define new phenotypes.
- Assess the plausibility of the exhaled VOC measures collected in this cohort using multivariate statistical approaches to describe the relationship of VOCs to one another and to the eosinophilic phenotype and multivariate statistical approaches to assess the repeatability of VOCs
- Use supervised machine learning approaches to assess the exhaled VOCs for the prediction of microbially defined phenotypes and unsupervised machine learning to define new phenotypes.

# Chapter 2 Methods

## 2.1 Introduction

The analyses presented in this thesis have been captured via a substudy of the Wessex AsThma CoHort of difficult asthma (WATCH) cohort. As detailed below, the WATCH Cohort has been established in order to efficiently capture the clinical characteristics of difficult to treat asthma patients (PI Dr Ramesh Kurukulaaratchy). The study also provides the infrastructure for new studies to easily access this cohort (or subsets of this cohort) by incorporating them as sub-studies.

Airway sampling of severe asthma patients was established as a WATCH sub-study (PI Professor Peter Howarth), which included sputum induction as part of the characterisation/sampling process. Breath sampling of severe asthma patients was established as another WATCH sub-study (PI Dr Adnan Azim), which included breath sampling and was performed alongside patients providing induced sputum samples

Methodology relating to the Breath Sampling of Severe Asthma Patients is detailed in Chapter 6.

## 2.2 WATCH Study

The WATCH cohort is a dual centre study (University Hospital Southampton Foundation Trust (UHSFT) and Queen Alexandra Hospital (QAH)) which started recruitment at UHSFT in August 2015. The Difficult Asthma Clinic at UHSFT and QAH are formally commissioned Regional Centres for Severe Asthma, providing support for patients across the South Central, UK region. The Adult Asthma multidisciplinary team (MDT) at both sites comprise consultants, research fellows, specialist nurses, associate practitioners, clinical psychologists, physiotherapists and dietitians. The service is provided via dedicated regional severe asthma clinics, transitional/young asthma patient clinic, Isle of Wight outreach asthma clinic, biologics (omalizumab, mepolizumab, reslizumab and benralizumab) clinic, nurse-led asthma clinic, clinical psychology clinic and respiratory physiotherapy clinic. Patients attending the adult or transitional regional asthma clinics are assessed by a physician and referred onto further MDT members and investigations with clinical decisions supported by regular post-clinic MDTs and monthly biologics referral MDTs This facilitates the extensive characterisation of each patient, which, in turn, enables the clinic to meet the standards of care described by NHSE.

### 2.2.1 Study Design

WATCH is a prospective observational study of patients with Difficult Asthma attending the Difficult Asthma Clinics at UHSFT and QAH. The study design, protocol and paperwork was approved by West Midlands – Solihull Research Ethics Committee (REC reference: 14/WM/1226). Patients were recruited into the study by way of a discrete “Enrolment” Study Visit capturing core demographic and clinical information and the results of the characterisation process provided by the clinic. For new to clinic patients an additional 3-month follow-up visit was also undertaken. Thereafter, records are continuously updated through annual “Follow Up” Study Visits and extraction from electronic clinical records. This pragmatic, opportunistic approach to data collection takes full advantage of the broad multidisciplinary clinical approach to difficult asthma management without becoming too onerous for the patient, clinician or researcher (Figure 2). Patients can complete all their CRF’s during clinic appointments for convenience but are also given the opportunity to complete their longer enrolment visit on a separate day.

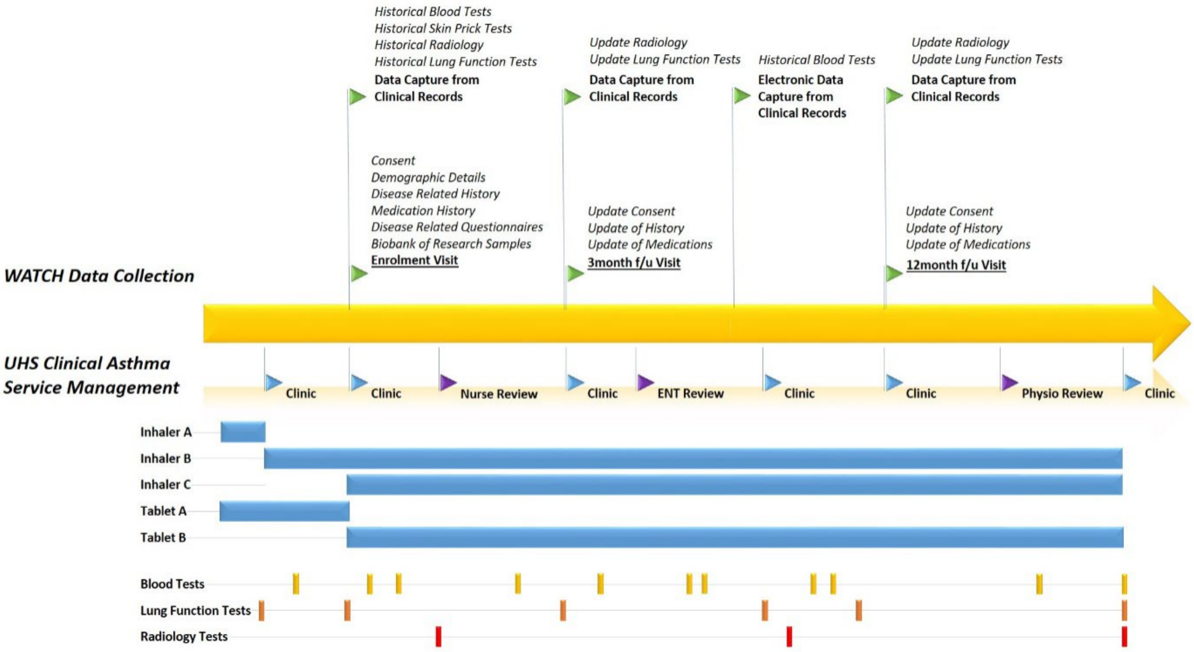


Figure 2.1 A schematic outline of the aligned clinical and research pathways/timelines followed by a patient under the UHSFT Difficult Asthma Clinic participating in the WATCH study.

Examples of how clinical tests (blood tests, lung function and radiology) and medication changes over time were captured for the study. *Ear Nose and throat (ENT), follow up (f/u).

### 2.2.2 Patient Recruitment

- All patients who attend the Adult or Transitional Regional Asthma Clinic managed with "high dose therapies" and/or "continuous or frequent use of oral steroids" according to the BTS Adult Asthma Management Guidelines 2016 are invited to the WATCH study
- Patients were excluded from the WATCH study if they attended the Adult or Transitional Regional Asthma Clinic at UHSFT but are not managed with "high dose therapies" and/or "continuous or frequent use of oral steroids" according to the BTS Adult Asthma Management Guidelines 2016 or if they lacked the ability to provide informed consent.

### 2.2.3 Data Collection

Data for the WATCH study was captured both through Case Record Forms (CRFs). The initial enrolment CRF contains a large suite of questionnaires that mirrors the extensive characterisation undertaken in clinical practice (Table 1 and 2). This was completed after the patient has received and read a patient study information sheet, a clinical or research member of the WATCH study team has received consent from the participant, and they have been assigned a study number. A clinical or research member of the WATCH team then proceeded to interview the participant, asking the questions from the Enrolment CRF (Table 2.1) and then filling in questionnaires with the participant (Table 2.2). In addition to the questionnaires, the enrolment visit collects objective measures and biological samples.

Table 2.1 Summary of Objective Clinical Measures Captured at Enrolment

* Immunoglobulin E (IgE), Fractional Exhaled Nitric Oxide (FeNO), Bronchoalveolar Lavage (BAL)

| **Investigations** | **Including** |
|---|---|
| Blood Tests | Full Blood Count, Serum Total IgE, (as well as any other clinically requested samples) |
| Lung Function Test | Spirometry +/- Reversibility (as well as any other clinically requested tests) |
| Exhaled Breath | FeNO |
| Anthropometry | Height, Weight, Bioelectrical Impedance |
| Biobank Samples | Blood, Urine, Induced Sputum, BAL |

Finally, additional objective clinical data from the hospital electronic systems were harnessed to provide retrospective and current investigation findings.

Table 2.2 Summary of Historical Clinical Record Data Collected During Participation in WATCH Study

| **Investigations** | **May Include** | **Limit of Data Retrieval** |
|---|---|---|
| Historical Blood Tests | Full blood count (FBC), Total Immunoglobulins E, G, M, A (IgE, IgG, IgM, IgA), Aspergillus precipitins (IgG), 25-hydroxy-vitamin D3, Anti-Neutrophil Cytoplasmic Antibody (ANCA), Antinuclear Antibody (ANA), Alpha-1-Antitrysin level (A1AT), Urea & Electrolytes, Liver Profile, Parathyroid Hormone (PTH), Thyroid function tests (Thyroid Stimulating Hormone & Free Thyroxine) | 10 years |
| Allergy Testing | Either Skin Prick Tests or Specific IgE Blood Tests to common aeroallergens [aspergillus fumigatus, alternaria tenius, cladosporium, penicillium, mixed moulds, grass mix, birch, weed mix, Dermatophagoides pteronyssinus and Dermatophagoides farinae, feathers, cat, dog, horse, and rabbit] | 10 years |
| Radiology | Computed Tomography (CT) or High Resolution CT Chest (HRCT), CT sinuses, dual energy X-ray absorptiometry (DEXA) scan | 10 years |
| Oesophageal Investigation Results | Oeosphagogastroduodenoscopy (OGD), Oesophageal Manometry, pH/Impedance Testing | 10 years |
| ENT Results | Nasoendoscopy | 10 years |
| Lung Function Tests | Spirometry +/- Bronchodilator Reversibility<br>Exhaled Nitric Oxide<br>Gas Transfer<br>Impulse Oscillometry<br>Static Lung Volumes<br>Multiple Nitrogen Breath Washout | 1 year<br>1 year<br>5 years<br>1 year<br>1 year<br>1 year |

### 2.2.4 Electronic Healthcare Data Collection

The WATCH database captured results from clinically requested and clinically processed investigations. These were performed by hospital departments in line with Standard Operating Procedures (SOPs) that conform to standards required of an NHS Hospital Service. Height and weight were measured by the study team from which BMI was calculated.

Clinically requested lung function was performed by the UHSFT Respiratory Physiology Department or Specialist Asthma Nurses, who operated in accordance with local department

SOPS and ARTP (Association of Respiratory Technology and Physiology) guidelines, described in 2.2.3 Spirometry with Reversibility.

### 2.2.5 Clinical Characterisation

For patients in the Airway Sampling in Severe Asthma Cohort, the full characterisation schedule was performed on the same morning, starting with breath collection and ending with sputum induction. In some cases, in order to obtain a viable sputum sample, it was necessary to repeat sputum induction on a second date; if so, this was performed within 7 days of the breath sample. Patients were excluded from the analysis if a viable sputum sample was not obtainable. A subset of patients were invited to provide breath samples on five consecutive days in addition to the characterisation schedule. Breath samples were collected in the same room, at the same time of day for each measure. Sputum induction was performed within 7 days prior to the first breath sample.

#### 2.2.5.1.1 Skin Prick Testing

Skin prick testing was usually performed at 1st clinical assessment to a standard panel of aeroallergens by the Asthma Specialist Nurses to common allergens using standard commercially available solutions. This included positive (histamine) and negative (saline) controls plus *Aspergillus Fumigatus*; *Altenaria Tenius*; Grass Mix Pollen; Birch Pollen; Weed Mix (Mugworth, Nettle, Pellion, Dandelion, English Plantain); Flower Mix (Aster, Chrysanthemum, Dahlia, Golden Rod, Marguerite); Rape Pollen; *Dermatophagoides pteronyssinus*; *Dermatophagoides farina*; Cockroach; Feathers; Cat Fur; Dog Fur; Horse. Antihistamines were omitted for 3 days prior to the test and Tricyclic Antidepressants for 7 days prior to the test. A positive skin prick test was defined as a mean wheal diameter ≥3mm than the negative control.

#### 2.2.5.1.2 Radiology Results

Clinically requested radiological investigations were usually performed by the UHSFT Radiology Department, which operates according to Royal College of Radiology guidelines and standards. Results were stored on the hospital’s local results server from which data was extracted (see data management). If the patient had imaging results at another site (e.g. local secondary care hospital), these were sought and imported to the study database.

#### 2.2.5.1.3 Spirometry with Reversibility

Subjects were asked to refrain from the following if possible before lung function testing.

- No smoking for 24 hours

- No alcohol consumption for 4 hours
- No vigorous exercise for 30 minutes
- No tight fitting clothing that could restrict full chest and/or abdominal expansion
- No food for 2 hours
- If clinically acceptable, no supplemental oxygen for 10 minutes

For spirometry, if patients were able to withhold their regular inhaler therapy then reversibility testing was performed using salbutamol (2.5mg nebulised or 400µg via spacer device). Restrictions before Reversibility Testing:

- Withhold short acting inhalers such as the ß-agonist salbutamol or the anticholinergic drug ipratropium bromide for at least 4 hours.
- Withhold long acting ß-agonist bronchodilators such as salmeterol for at least 12 hours.
- Withhold oral therapy with aminophylline or slow release ß-agonist for 12 hours

Spirometry was performed using either Carefusion® (Chatham, UK) or Nspire Health Ltd (Hertford, UK) equipment according to ERS/ATS guidelines [230]. If subjects were unable to withhold regular inhaler therapy then Spirometry results were recorded as "post-bronchodilator". Z Scores and percentage predictive values were calculated using GLI (Global Lung function Initiative) look up tables.

#### 2.2.5.1.4 Exhaled Nitric Oxide

Fraction of Exhaled Nitric Oxide (FeNO) was measured using the NIOX VERO® (Oxford, UK) or Bedfont NObreath® (Aylesford, UK) at a flow rate of 50ml/s according to ERS/ATS guidelines [80]. Exhaled nitric oxide was measured before any other lung function test due to the influence of breathing manoeuvres on FeNO readings. A minimum of 2 technically acceptable tests were recorded with the two highest values within 10% of each other and the mean value reported.

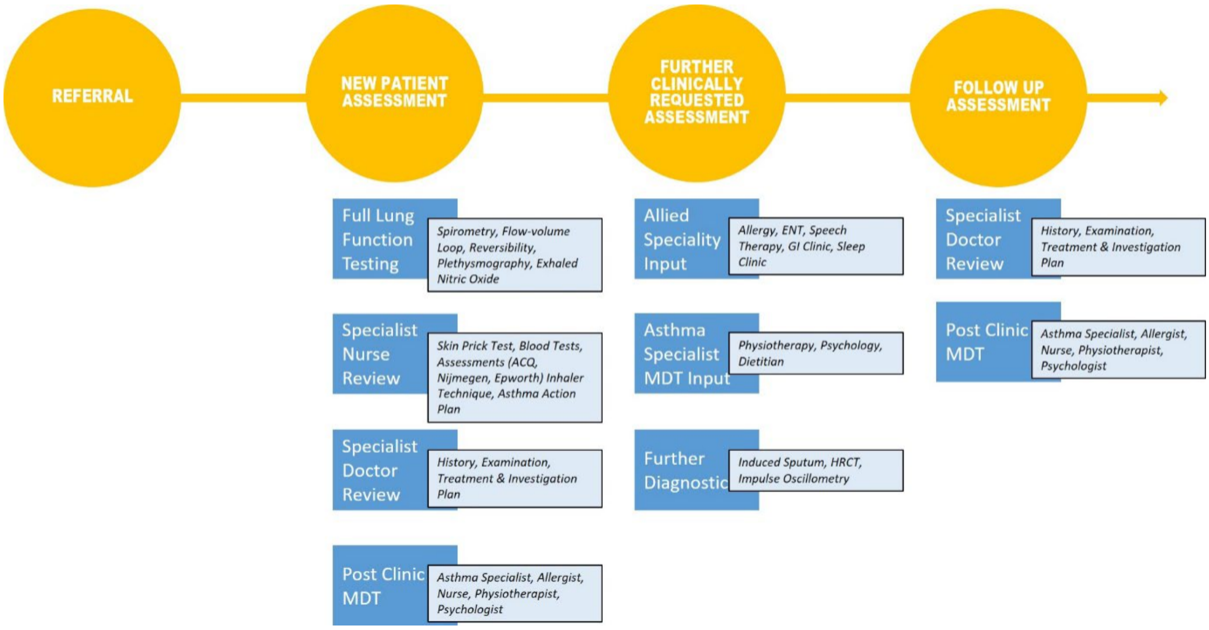

REFERRAL
NEW PATIENT ASSESSMENT
FURTHER CLINICALLY REQUESTED ASSESSMENT
FOLLOW UP ASSESSMENT
Full Lung Function Testing
Spirometry, Flow-volume Loop, Reversibility, Plethysmography, Exhaled Nitric Oxide
Specialist Nurse Review
Skin Prick Test, Blood Tests, Assessments (ACQ, Nijmegen, Epworth) Inhaler Technique, Asthma Action Plan
Specialist Doctor Review
History, Examination, Treatment & Investigation Plan
Post Clinic MDT
Asthma Specialist, Allergist, Nurse, Physiotherapist, Psychologist
Allied Speciality Input
Allergy, ENT, Speech Therapy, GI Clinic, Sleep Clinic
Asthma Specialist MDT Input
Physiotherapy, Psychology, Dietitian
Further Diagnostic
Induced Sputum, HRCT, Impulse Oscillometry
Specialist Doctor Review
History, Examination, Treatment & Investigation Plan
Post Clinic MDT
Asthma Specialist, Allergist, Nurse, Physiotherapist, Psychologist

Figure 2.2 Schematic Representation of the Assessments and Interactions Patients Undertake as Part of Their Assessment in the UHSFT Clinical Asthma Service

Abbreviations: Ear Nose and Throat (ENT), Gastrointestinal (GI) Multidisciplinary (MDT), High Resolution Computed Tomography (HRCT)

## 2.3 Airway Sampling of Severe Asthma Patients

### 2.3.1 Study Design

As per the parent WATCH study, biological sampling of severe asthma patients was performed entirely cross-sectionally.

### 2.3.2 Patient Recruitment

- Patients with severe asthma, confirmed by an asthma specialist in accordance with the BTS (British Thoracic Society) guidelines with alternative causes for symptoms excluded and treatment for co-morbidities optimised.
- Participants were aged between 18 and 80 years with no restrictions according to gender, race, or smoking status.

### 2.3.3 Sampling and Analyses

#### 2.3.3.1 Sputum Induction

Sputum was induced using a DeVilbiss® Ultraneb (DeVilbiss, NY, USA) following a standardised protocol based on the methods described by ten Brinke et al [231]. Patients were bronchodilated with short acting beta-agonist (SABA) medication prior to sputum induction and lung function (FEV1) was measured after each 5-minute nebulisation beginning with 0.9% saline followed by 3% and finally 4.5%, if tolerated, to check if a 20% drop from post bronchodilator FEV1 had been reached at which point the induction would be stopped. Lung function (FEV1) was measured after each 5-minute nebulisation and after 2 minutes of nebulisation if the subject's FEV1 <1.5L. Samples were stored on ice during collection and transport to the laboratory for processing.

#### 2.3.3.2 Sputum Processing

Sputum processing was performed in the NIHR Southampton Biomedical Research Centre BRC by Clair Barber. The concurrent method of sputum processing was performed providing PBS and DTE supernatant for analysis [232]. Sputum samples were processed as soon as possible and within 2 hours of expectoration with 8× volume of phosphate buffered saline (PBS) and a proportion of supernatant was then removed and the sample was further incubated with 0.2% dithioerythritol (DTE) giving a final concentration of 0.1% DTE. Cytospins were stained using by rapid Romanowski staining (Fisher Scientific, Loughborough, UK). The proportion of inflammatory cells were assessed by counting 800 respiratory cells plus squamous to give a mean percentage of respiratory cells.

#### 2.3.3.3 MSD Analysis

Inflammatory mediators were measured using a V-plex multiple cytokine immunoassay platform (Meso Scale Discovery, MSD) as per the manufacturer's instructions by Dr Laurie Lau. The assay use SULFO-TAG labelled Detection Antibody for electroemiluminescence.

### 2.3.4 Microbial 16S rRNA Sequencing

DNA was extracted by Professor James Chalmers' lab at Dundee University on the Qiacube DNA extraction machine using the DNeasy PowerSoil Pro Kit (250). DNA extraction was performed in batches. An extraction negative was performed for any new reagent used. Samples were analysed in two 16s rRNA sequencing runs: Run 1 (111 sputum samples samples) and Run 2 (108 samples – a mix of sputum and bronchoalevolar lavage). Each batch included a PCR negative control and a Qiagen elution buffer negative control as well as a sequencing positive control (Mock Community from Zymo – ZmyoBIOMICS Microbial Community DNA Standard – D6305). Stutter primers

(Nextera v2 indexes SET A & C) were used. Illumina sequenced paired-end fastq files were demultiplexed by sample and barcodes removed.

## 2.4 Breath Sampling of Severe Asthma Patients

Methodology relating to the Breath Sampling of Severe Asthma Patients is detailed in Chapter 6. Briefly, this sub-study was designed in order to pair breath samples to sputum samples collected in the Airway Sampling of Severe Asthma Patients sub-study and clinical characterisation of the WATCH Study. The sub-study had two arms: “cross sectional” and “repeatability”.

## 2.5 Multidisciplinary Contributions

All the work presented in this thesis was performed by myself except in the following instances in which work was performed by others or assisted by others:

- Ms Kim Bentley was the WATCH study nurse and assisted with many of study visits and procedures.
- Mr Matthew Harvey was the WATCH study co-ordinator and provided guidance on many aspects of sub-study management, including the various ethics submissions
- Mr Colin Newell was the WATCH data manager and assisted with interactions with the study database.
- Dr Hitasha Rupani’s team at Queen Alexandra Hospital contributed to characterisation of participants recruited at the Portsmouth site
- Dr Clair Barber performed the initial pre-processing of many of the sputum samples and performed the cell counts.
- Dr Laurie Lau performed the cytokine analyses of sputum samples
- Professor James Chalmers’ team at the University of Dundee, including (but not limited to) Dr Hollian Richardson and Dr Alison Dicker performed the 16S sequencing of sputum samples.
- The Owlstone Medical team performed the GCMS analysis of breath samples and pre-processed the data according to their pipelines.

- Professor John Langley, Dr Grielof Koster and Dr Paul Afolabi provided guidance on technical matters relating to organic chemistry and mass spectrometry
- Dr Faisal Rezwan provided guidance on machine learning approaches used throughout this thesis
- Dr David Cleary provided guidance on bioinformatic approaches used in the microbial analyses
- Dr Ramesh Kurukulaaratchy is PI for the WATCH Study
- Professor Peter Howarth is PI for the Airway Sampling of Severe Asthma Patients WATCH sub-study and provided guidance for the Breath Sampling of Severe Asthma Patients WATCH sub-study.

# Chapter 3 Characterising a Severe Asthma Cohort

## 3.1 Introduction

Traditional phenotyping efforts in severe asthma classify patients according to their clinical characteristics such as demographics or clinical history [46,233,234] but provide little insight into the nature of the underlying mechanisms. Stratification is thought to be more meaningful if done by measures of airway inflammation[191]. The inflammation can be objectively measured by quantitative cytometry of bronchial washes, bronchoalveolar lavage or bronchial biopsy [193] but is most commonly performed in induced sputum [235]. Patients are typically stratified into groups based upon the presence and absence of eosinophils and neutrophils: eosinophilic, neutrophilic, mixed granulocytic (eosinophils and neutrophils both present) and paucigranulocytic (eosinophils and neutrophils both absent) [191]. Though there is no consensus cut-offs for these granulocytes [46,192,204], these inflammatory phenotypes appear to have distinct characteristics [204,236] and indicate the delineation of underlying mechanisms.

Analysis from subsequent chapters of my thesis relates to this cohort and so the objectives of this chapter are to

- Describe the clinical characteristics of this cohort
- Confirm that the cohort is representative of the patients in whom a breathomics biomarker would be clinically useful
- Describe the clinical characteristics of sputum inflammatory phenotypes
- Assess the predictive values of currently available clinical biomarkers

## 3.2 Chapter Specific Methods

### 3.2.1 Patient Population

The patients described in this chapter were recruited from the Airway Sampling in Severe Asthma WATCH Sub-Study. Patients could be enrolled into study via two sources: patients already enrolled into the WATCH Study of Difficult Asthma were recruited if they met severity criteria. Alternatively, patients were recruited directly from the Regional Difficult to Treat Asthma Services at University Hospital Southampton and Queen Alexandra Hospital via enrolment into the WATCH

Study. Patients were excluded if they were unable to provide a viable sputum sample (≥50% viability).

### 3.2.2 Statistical Analysis

Statistical analysis was performed using Python scripting language (version 3.8.3) [237]. Clinical characteristics were described using median and 95% confidence intervals with between group comparisons by Mann Whitney U tests for continuous variables and absolute numbers with percentages within each group and Chi Squared tests for categorical variables. Correlations were calculated by spearman rank coefficients. Receiver Operating Characteristic (ROC) Curves were constructed from the false positive and true positive rate and the performance of each biomarker was reported by the Area Under the Curve (AUC).

In order to define the optimal cut-off for biomarkers predicting sputum eosinophils of >2% and sputum neutrophils of >40%, >61% and >76%, the precision, recall and threshold values were calculated for each biomarker value in the dataset for predicting the target. This process was iterated over every possible cut-off. The F1 score was extracted for each cut-off and ranked so as to find the optimal cut-off. The combined use of biomarkers refers to both criteria being satisfied.

## 3.3 Results

### 3.3.1 Patient Population Recruitment

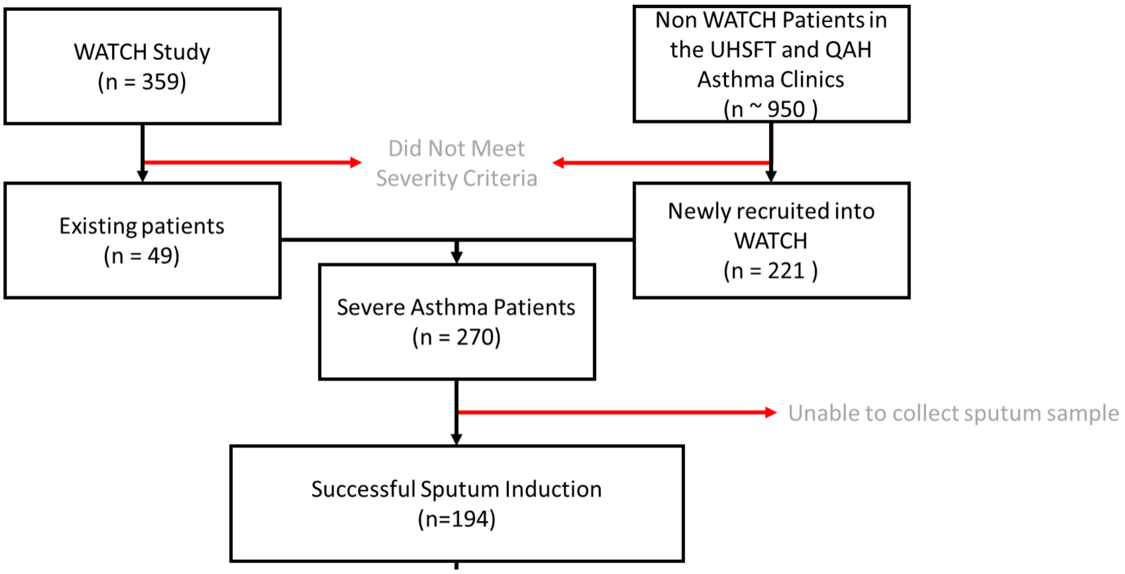

WATCH Study
(n = 359)
Non WATCH Patients in the UHSFT and QAH Asthma Clinics
(n ~ 950 )
Did Not Meet Severity Criteria
Existing patients
(n = 49)
Newly recruited into WATCH
(n = 221 )
Severe Asthma Patients
(n = 270)
Unable to collect sputum sample
Successful Sputum Induction
(n=194)

Figure 3.1 Consort Diagram of Patients Recruited to the Airway Sampling in Severe Asthma Cohort

Patient recruitment started in October 2017 at UHSFT. At the start of the study, there were 359 patients in the WATCH study, of which 49 met the inclusion criteria for this cohort and were invited for sputum induction. At the start of the recruitment, there were an estimated 550 patients in the asthma clinic, not already in the WATCH study. Over the recruitment period, a further 400 patients had been referred to the service. Two hundred and twenty one patients met the inclusion criteria for this cohort and were newly recruited to the WATCH study. Of the 270 patients undertaking sputum induction, 194 patients provided a viable sputum sample, representing a 71.8% success rate (which includes patients providing a viable sputum sample at repeat sputum induction attempts) (Figure 3.1).

#### 3.3.1.1 Airway Sampling in Severe Asthma Cohort compared to Parent WATCH Cohort

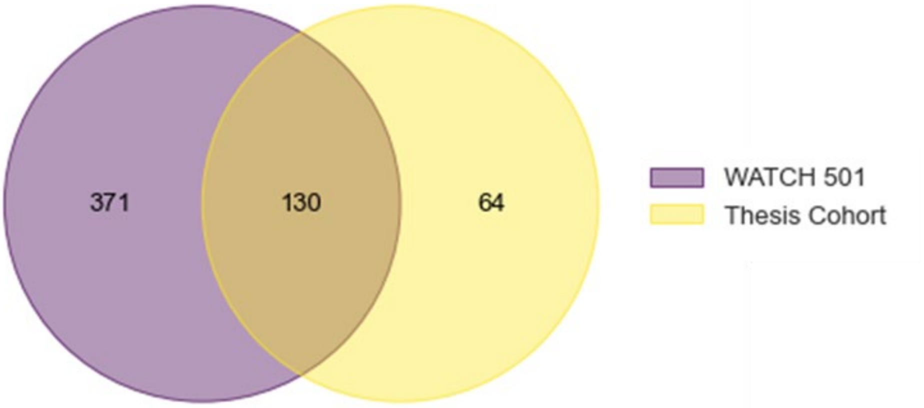


Figure 3.2 Overlap of Patients in the Airway Sampling in Severe Asthma Cohort of Severe Asthma and patients described in the WATCH Cohort of Difficult to Treat Asthma

All patients characterised in the Airway Sampling and Breath Sampling sub-studies were recruited into the parent WATCH study. Of the patients that were included in the analysis of 501 difficult to treat asthma patients [238], only one third of patients (130) were included in the Airway Sampling sub-study (Figure 3.2).

Table 3.1 Comparison of the Thesis Severe Asthma Cohort and WATCH Difficult to Treat Asthma Cohort (excluding 130 overlapping patients)

Continuous variables expressed as median [Q1, Q3] with differences measured by Mann-Whitney U test. Categorical variables expressed as n (%) with differences measured by chi-square test. Abbreviations: ICS, inhaled corticosteroid; BDPe, beclomethasone dose equivalent; FeNO, fraction of nitric oxide in exhaled breath; post BD, post bronchodilator; FEV1, forced expiratory volume in 1 second; FVC, forced vital capacity; FEF25-75%, forced expiratory flow at 25% to 75% of FVC; ACQ, asthma control questionnaire, HADSTOT, Hospital Anxiety and Depression Total Score; SNOT, SinoNasal Outcome Score

| | **Airway Sampling of Severe Asthma (n = 64)** | **WATCH Difficult to Treat Asthma (n = 371)** | **P-Value** |
|---|---|---|---|
| Sex (% Female) | 31 (48.4) | 255 (68.7) | 0.003 |
| Age | 54.5 [41.8,65.0] | 51.0 [36.0,62.0] | 0.024 |
| BMI | 27.7 [25.6,32.5] | 29.5 [25.4,35.6] | 0.358 |
| Smoker (% Never) | 39 (60.9) | 202 (54.6) | 0.640 |
| Atopy | 46 (71.9) | 213 (68.5) | 0.700 |
| Age of Onset | 23.0 [7.0,49.0] | 18.0 [3.0,38.0] | 0.121 |
| Exacerbations in Last 12 months | 2.0 [0.0,4.0] | 3.0 [1.0,5.0] | 0.019 |
| ICS (BDPe) | 2960.0 [2000.0,3585.0] | 3000.0 [2000.0,3000.0] | 0.420 |
| FeNO | 28.0 [16.5,52.0] | 17.1 [9.0,34.7] | <0.001 |
| Blood Eosinophil Count | 0.3 [0.1,0.4] | 0.2 [0.1,0.3] | 0.004 |
| PostBD FEV1 | 84.9 [63.0,97.2] | 77.8 [61.4,92.8] | 0.172 |
| PostBD FEV1/FVC | 69.0 [60.8,80.0] | 69.0 [58.0,78.0] | 0.553 |
| PostBD FEF25-75 %predicted | 55.9 [36.7,84.7] | 49.8 [27.7,80.5] | 0.239 |
| ACQ6 | 2.5 [1.2,3.0] | 2.5 [1.5,3.5] | 0.260 |
| HADSTOT | 9.0 [6.0,15.0] | 11.0 [6.0,18.5] | 0.202 |
| SNOT20 | 28.0 [18.5,44.0] | 29.0 [17.0,46.8] | 0.810 |

Compared to the WATCH Cohort of Difficult to Treat Asthma, the Airway Sampling in Severe Asthma Cohort had fewer females and was marginally older (Table 3.1). The Airway Sampling in Severe Asthma Cohort had higher blood eosinophil counts and FeNO despite similar levels of ICS therapy but statistically fewer exacerbations in the past 12 months (Table 3.1). There were no statistically significant differences between the two cohorts in terms of post-bronchodilator lung function or self-reported asthma scores (Table 3.1).

### 3.3.2 Sputum Inflammatory Phenotypes

Of the 194 patients that produced a viable sputum sample, 50% were found to have an eosinophilia (either eosinophilic (36.4%) or mixed granulocytic (13.9%), Figure 3.3). 32.3% of patients were paucigranular. Neutrophilia was seen in a third (either neutrophilic (17.4%) or mixed granulocytic (13.9%) of patients. The mixed granulocytic phenotype was the rarest of the four sputum inflammatory phenotypes.

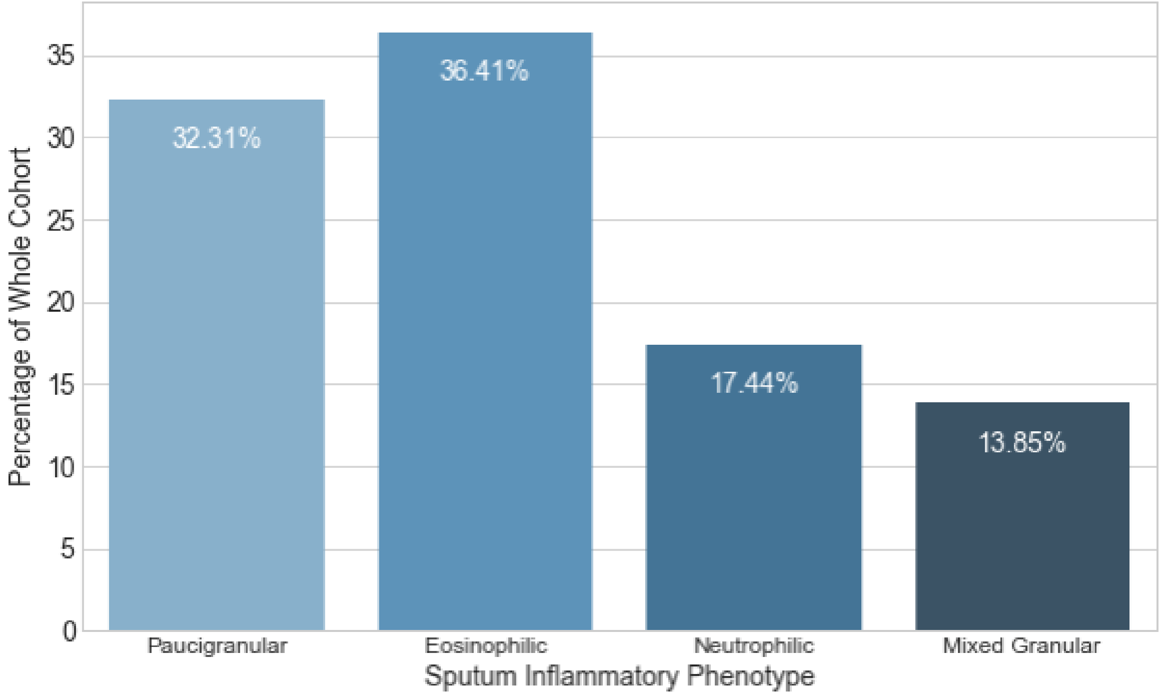


Figure 3.3 Proportion of Sputum Inflammatory Phenotypes

Paucigranular = sputum eosinophils <2% and sputum neutrophils <61%; Eosinophilic = sputum eosinophils >2% and sputum neutrophils <61%; Neutrophilic = sputum eosinophils <2% and sputum neutrophils >61%; Mixed Granular = sputum eosinophils>2% and sputum neutrophils >61%

There were no differences in terms of treatments (ICS dose, mOCS use or biologics use), co-morbidities (smoking status, atopic status, GORD) or self-report asthma burden (ACQ6, HADS, SNOT) between the inflammatory phenotypes (Table 3.4).

The eosinophilic and neutrophilic phenotypes were predominantly male (43.7% and 44.1% female respectively) whilst the mixed granulocytic and paucigranular phenotypes were predominantly female (59.3% and 66.1% respectively). The mixed and paucigranular phenotypes had BMIs greater than 30 (34.6 (27.2-36.2) and 30.8 (27.7-34.0) respectively).

The eosinophilic phenotype was associated with a late age of asthma onset (33.0 (14.2-54.0) years), nasal polyps (reported in 40.8%), frequent exacerbations (3 (1.0-5.0)), high FeNO (43.5 (27.0-71.8)), lung function reversibility (13.3 (4.3-20.8)) and poor post BD spirometry ($FEV_1$ 67.7 (54.8-83.5) %predicted).

The mixed granulocytic phenotype had the poorest post BD $FEV_1$ (62.6 (54.3,83.7) %predicted) but, consistent with the high BMI (34.6 (27.2-36.2)), there is a trend towards these patients having the lowest post BD FVC (77.4 (68.9-95.2) %predicted). Nevertheless, they continue to have an obstructive post BD $FEV_1$/FVC ratio (64.0 (54.0-77.5)) and evidence of small airways disease (post BD $FEF^{25-75\%}$ 38.3 (23.8-82.2) %predicted).

The neutrophilic phenotype had the youngest age of onset (7.0 (2.0-19.0)), and relatively preserved post BD spirometry: FEV1 (%predicted) 81.2 (60.1 - 92.6)), FVC (% predicted) 93.0 (77.8-101.3). The pauci-granulocytic group had the highest proportion of female patients (66.1%) and well preserved post BD spirometry: it is the only phenotype with a non-obstructive post BD $FEV_1$/FVC ratio 74.5 (64.5-81.0).

Table 3.2 Clinical Characteristics Across Sputum Inflammatory Phenotypes

Continuous variables expressed as median [Q1, Q3] with differences measured by Mann-Whitney U test.Categorical variables expressed as n (%) with differences measured by chi-square test. Abbreviations: GORD, gastro-oesophageal reflux disease; ICS, inhaled corticosteroid; BDPe, beclomethasone dose equivalent; OCS, oral corticosteroids; IgE, Immunoglobulin E; IL-5, Interleukin 5; ACQ, asthma control questionnaire, HADSTOT, Hospital Anxiety and Depression Total Score; SNOT, SinoNasal Outcome Score; FeNO, fraction of nitric oxide in exhaled breath;  post BD, post bronchodilator; FEV1, forced expiratory volume in 1 second; FVC, forced vital capacity; FEF25-75%, forced expiratory flow at 25% to 75% of FVC;

| | **Eosinophilic (n = 71)** | **Mixed Granular (n = 27)** | **Neutrophilic (n = 34)** | **Paucigranular N = 62)** | P-Value |
|---|---|---|---|---|---|
| **Female Sex** | 31<br>(43.7) | 16<br>(59.3) | 15<br>(44.1) | 41<br>(66.1) | **0.041** |
| **Age** | 60.0<br>[47.5,70.5] | 59.0<br>[48.5,66.0] | 54.0<br>[44.0,62.0] | 53.0<br>[42.5,64.8] | **0.093** |
| **BMI** | 27.6<br>[24.8,31.5] | 34.6<br>[27.2,36.2] | 28.9<br>[24.9,32.4] | 30.8<br>[27.7,34.0] | **0.017** |
| **Never Smoker** | 46<br>(64.8) | 16<br>(59.3) | 21<br>(61.8) | 41<br>(66.1) | **0.66** |
| **Atopic** | 46<br>(64.8) | 17<br>(63.0) | 19<br>(55.9) | 37<br>(59.7) | **0.829** |
| **GORD** | 45<br>(63.4) | 16<br>(59.3) | 23<br>(67.6) | 35<br>(56.5) | **0.812** |
| **Nasal Polyps** | 29<br>(40.8) | 5<br>(19.2) | 8<br>(24.2) | 6<br>(9.8) | **0.003** |
| **Age of Onset** | 33.0<br>[14.2,54.0] | 18.0<br>[5.0,28.0] | 7.0<br>[2.0,19.0] | 13.0<br>[4.0,38.0] | **0.002** |
| **Exacerbations in the Last 12 Months** | 3.0<br>[1.0,5.0] | 1.0<br>[0.0,3.0] | 1.0<br>[0.0,2.0] | 1.5<br>[1.0,3.0] | **0.022** |
| **ICS (BDPe)** | 3000.0<br>[2000.0,3780.0] | 3000.0<br>[2000.0,3000.0] | 3000.0<br>[2575.0,3755.0] | 3000.0<br>[2000.0,3000.0] | **0.497** |
| **Maintenance OCS** | 31<br>(43.7) | 13<br>(48.1) | 11<br>(32.4) | 24<br>(38.7) | **0.58** |
| **Anti IgE** | 6<br>(8.5) | 3<br>(11.1) | 5<br>(14.7) | 5<br>(8.1) | **0.72** |
| **Anti IL-5** | 11 | 1 | 5 | 7 | **0.432** |

| | | | | | |
|---|---|---|---|---|---|
| | (15.5) | (3.7) | (14.7) | (11.3) | |
| **ACQ6** | 2.3<br>[1.2,3.5] | 2.7<br>[1.8,3.2] | 2.7<br>[1.7,3.2] | 2.2<br>[1.4,3.0] | **0.615** |
| **HADS** | 9.0<br>[6.0,15.0] | 8.0<br>[5.0,13.0] | 13.0<br>[7.5,18.5] | 11.0<br>[6.5,16.5] | **0.267** |
| **SNOT20** | 31.5<br>[23.0,47.8] | 19.0<br>[13.0,34.0] | 33.5<br>[17.5,47.0] | 32.0<br>[18.8,45.0] | **0.128** |
| **FeNO** | 43.5<br>[27.0,71.8] | 23.5<br>[16.2,37.0] | 19.5<br>[12.5,28.8] | 18.0<br>[11.0,28.8] | **<0.001** |
| **PostBD FEV1 (% predicted)** | 67.7<br>[54.8,83.5] | 62.6<br>[54.3,83.7] | 81.2<br>[60.1,92.6] | 87.0<br>[72.3,99.7] | **<0.001** |
| **PostBD FVC (% predicted)** | 91.5<br>[77.3,102.2] | 77.4<br>[68.9,95.2] | 93.0<br>[77.8,101.3] | 95.7<br>[84.8,105.8] | **0.058** |
| **PostBD FEV1/FVC** | 65.0<br>[52.0,70.0] | 64.0<br>[54.0, 77.5] | 68.0<br>[61.8, 75.8] | 74.5<br>[64.5, 81.0] | **<0.001** |
| **PostBD FEF25-75% (% predicted)** | 38.1<br>[27.1,50.1] | 38.3<br>[23.8,82.2] | 54.6<br>[36.0,69.4] | 68.8<br>[39.6,89.7] | **<0.001** |
| **FEV1 Reversibility** | 13.3<br>[4.3,20.8] | 10.9<br>[3.0,24.2] | 4.0<br>[0.1,9.6] | 7.7<br>[4.4,13.8] | **0.016** |
| **Blood Neutrophils** | 5.0 [4.0,6.4] | 6.1 [4.8,7.8] | 6.0 [4.5,7.6] | 4.8 [3.9,6.5] | **0.047** |
| **Blood Eosinophils** | 0.4 [0.2,0.5] | 0.3 [0.2,0.6] | 0.2 [0.1,0.2] | 0.1 [0.1,0.2] | **<0.001** |
| **Serum Total IgE** | 174.8<br>[46.7,411.0] | 70.4<br>[21.6,261.9] | 194.1<br>[56.5,539.3] | 33.0<br>[14.2,136.2] | **<0.001** |
| **Sputum Neutrophils** | 34.8 [19.8,43.2] | 73.7 [70.2,81.6] | 77.7 [68.6,88.7] | 28.3 [17.2,46.6] | **<0.001** |
| **Sputum Eosinophils** | 16.8 [6.4,32.5] | 5.0 [3.2,8.2] | 0.4 [0.2,0.8] | 0.5 [0.1,0.9] | **<0.001** |

### 3.3.3 Assess Currently Available Clinical Biomarkers for Predicting Sputum Inflammatory Phenotypes

Blood cell counts (eosinophils and neutrophils) are commonly available in a respiratory clinic. Their relation to sputum cell counts was explored in the Airway Sampling in Severe Asthma Cohort.

#### 3.3.3.1 Correlations Between Clinical Biomarkers and Sputum Eosinophilia and Neutrophilia

Blood eosinophils and FeNO only share a weak positive correlation (r = 0.231, p<0.001), however both have moderate correlations with sputum eosinophilia (r = 0.563, p<0.001 and r = 0.494, p<0.001 respectively) (Figure 3.5).

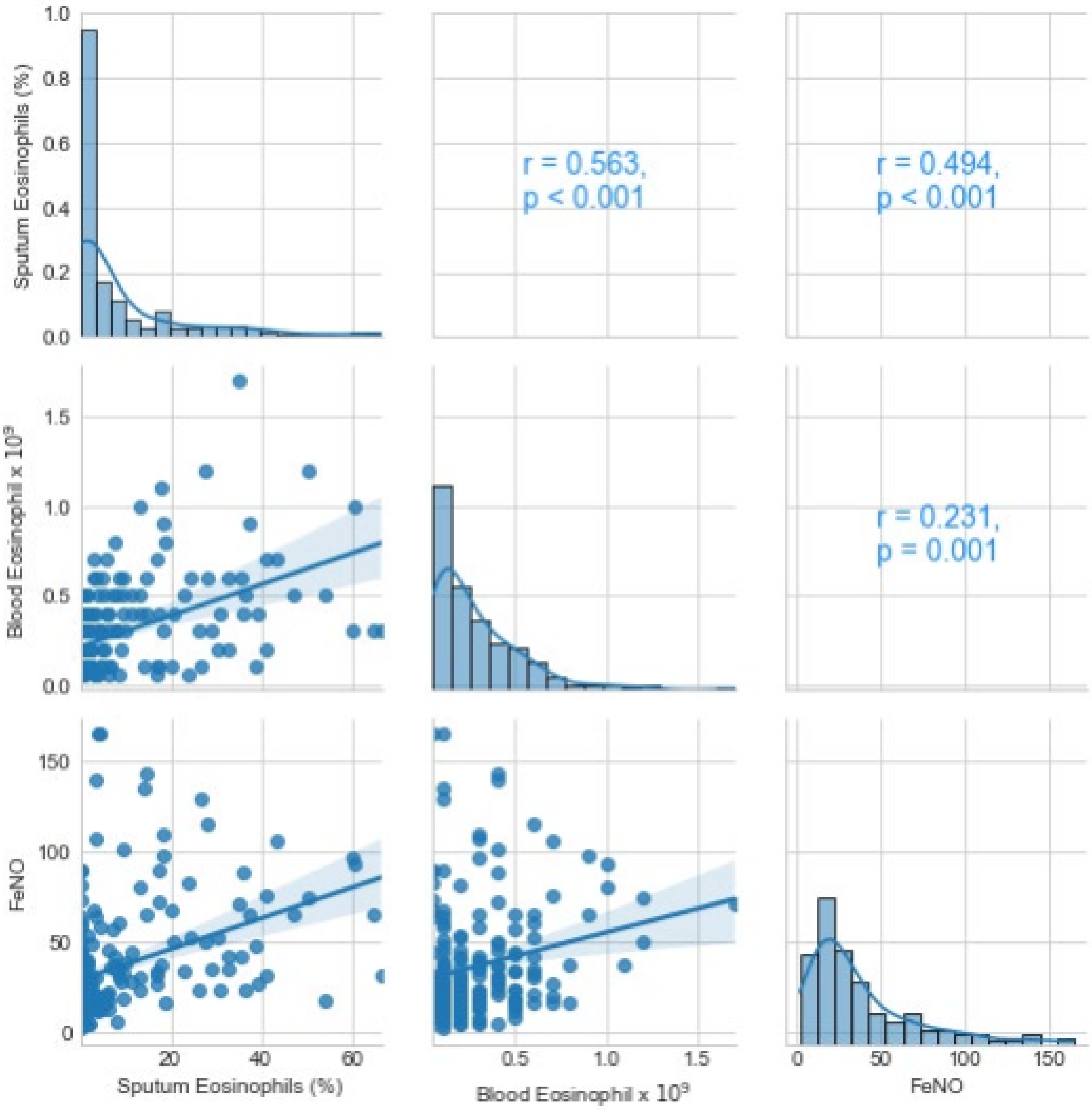


Figure 3.4 Spearman Rank Correlations of Clinical Biomarkers, FeNO (fraction of exhaled nitric oxide in exhaled breath) and Blood Eosinophil counts with Sputum Eosinophilia

There are only weak correlations between blood and sputum neutrophil measures (r = 0.218, p = 0.004). Sputum Neutrophils have a weak negative correlation with FeNO (r = -0.229, p = 0.004) but there is no statistically significant correlation between blood neutrophils and FeNO (Figure 3.6).

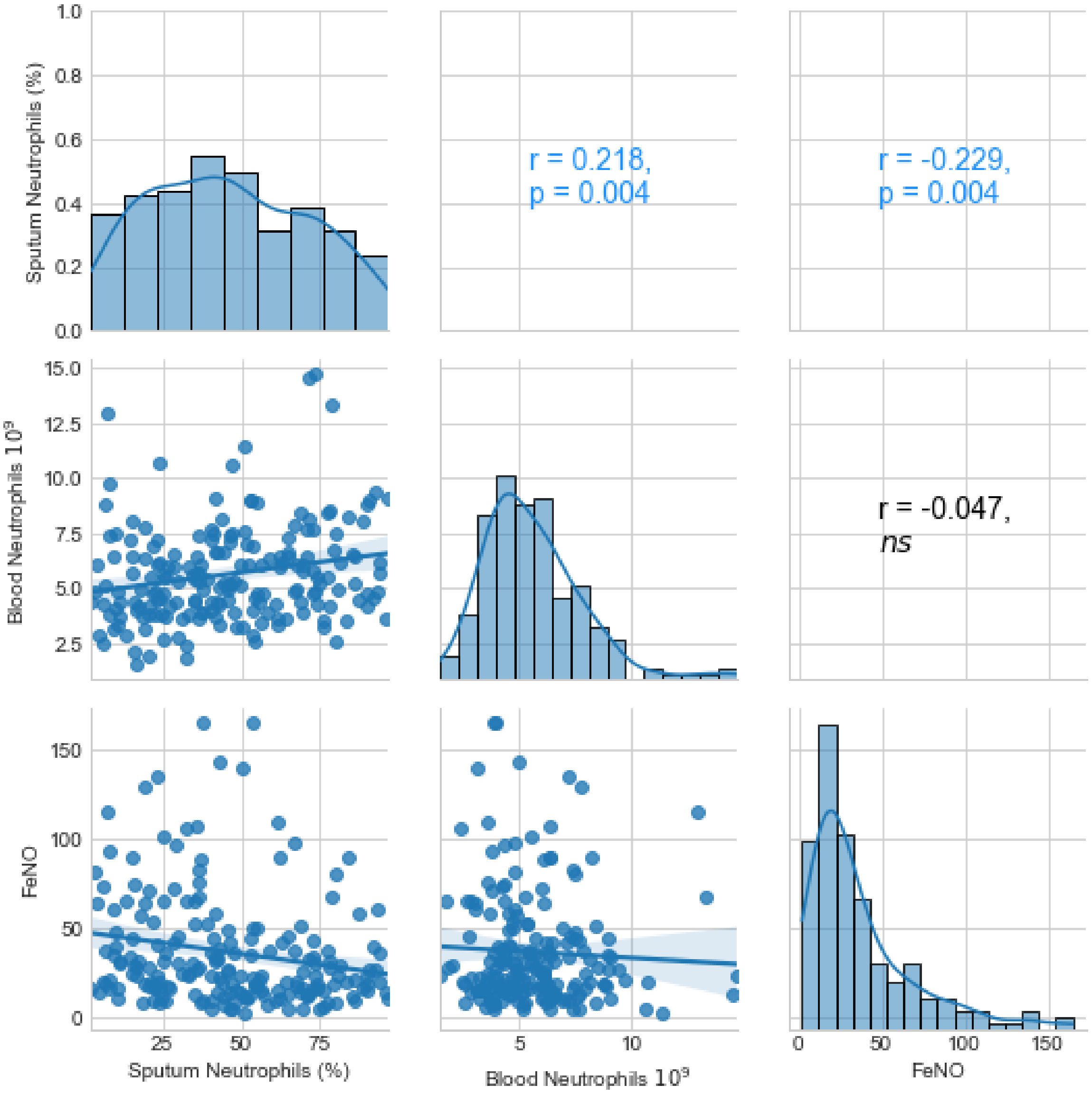


Figure 3.5 Spearman Rank Correlations of Clinical Biomarkers, FeNO (fraction of exhaled nitric oxide in exhaled breath) vs Blood and Sputum Neutrophils

### 3.3.3.2 Predicting Sputum Inflammatory Phenotypes with Clinical Biomarkers

A range of thresholds have been used for blood eosinophil counts and FeNO in clinical trials and in clinical practice, ranging from 150-400 and 20-50 respectively, to predict sputum eosinophil counts of ≥2%.Applying these cut-offs, produce an AUC range from 0.6 and 0.7 (Figure 3.7 and Figure 3.8).

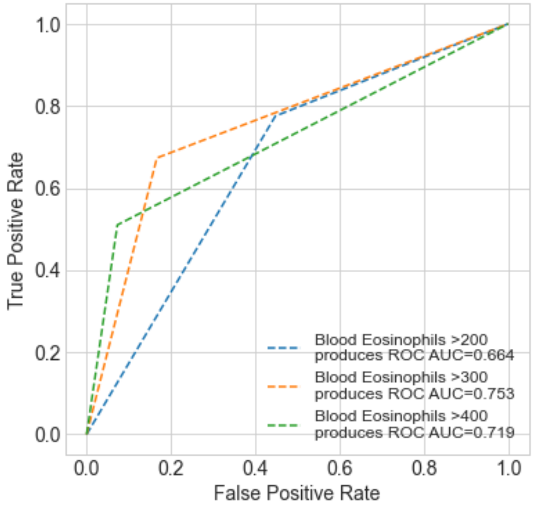


Figure 3.6 ROC Curve for Different Blood Eosinophils Cut-offs for predicting Sputum Eosinophils >2%

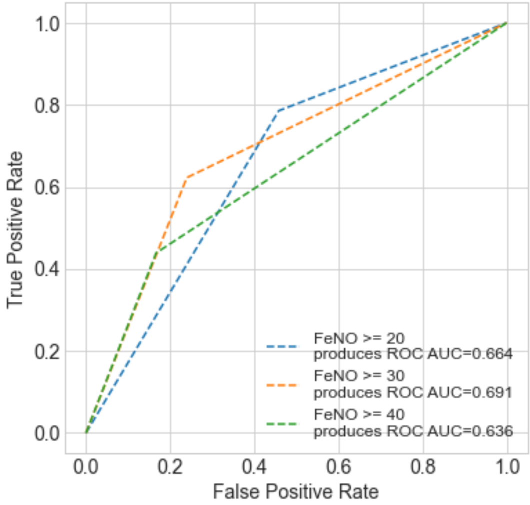


Figure 3.7 ROC Curve for Different FeNO Cut-offs for predicting Sputum Eosinophils >2%

AUC scores are difficult to translate to clinical practice; but the positive predictive value is of clinical salience. The highest cut-offs for blood eosinophils and FeNO produces the highest PPV for predicting sputum eosinophilia >2%: 0.877 and 0.729 respectively (Table 3.5).

Table 3.3 Positive Predictive Value for Different Cut-offs for Clinical T2 Biomarkers in Predicting Sputum Eosinophils >2%

Abbreviations: FeNO, fraction of exhaled nitric oxide in exhaled breath

| Biomarker | PPV |
|---|---|
| **Blood Eosinophils >200** | 0.633 |
| **Blood Eosinophils >300** | 0.795 |
| **Blood Eosinophils >400** | 0.877 |
| **FeNO >20** | 0.631 |
| **FeNO >30** | 0.726 |
| **FeNO >40** | 0.729 |

Using the iterative approach (described in 3.2.2), we can identify the optimal cut-off for these biomarkers (defined as the (lowest cut-off producing the highest f1-score for predicting sputum eosinophils >2%). This allows a check for which thresholds are best for T2 biomarkers and provides a threshold to assess for biomarkers for which there is no recognised threshold. For the clinically used biomarkers, blood eosinophil counts and FeNO, cut-offs towards the lower end of that which is used in clinical practice (200 and 15 respectively). Blood neutrophils and serum Total IgE produce PPVs of 0.508 and 0.546 respectively (Table 3.6).

Table 3.4 Area Under the Curve and Positive Predictive Value for Clinical T2 Biomarkers in Predicting Sputum Eosinophils ≥2% Biomarkers Using F1 Score Optimised Cut-offs

| Variable | Cut-off | AUC | PPV |
|---|---|---|---|
| **Blood Eosinophil** | 200 | 0.664 | 0.633 |
| **Fraction of Exhaled Nitric Oxide (FeNO)** | 15 | 0.639 | 0.587 |
| **Blood Neutrophils** | 2.6 | 0.510 | 0.508 |
| **Serum Total IgE** | 19.1 | 0.573 | 0.546 |

### 3.3.3.3 Effect of OCS

Severe asthma patients are commonly treated with oral prednisolone (Table 3.4) due to their anti-inflammatory effect. The AUC of a blood eosinophil count of ≥300 for predicting sputum eosinophils ≥2% is 0.723 when patients are not on regular oral prednisolone; this drops to 0.630 when patients are on regular oral prednisolone (Figure 3.9). Similarly, the AUC for FeNO ≥20 for predicting sputum eosinophils ≥2% is 0.720 when patients are not on regular oral prednisolone but drops to 0.578 for patients who are (Figure 3.10).

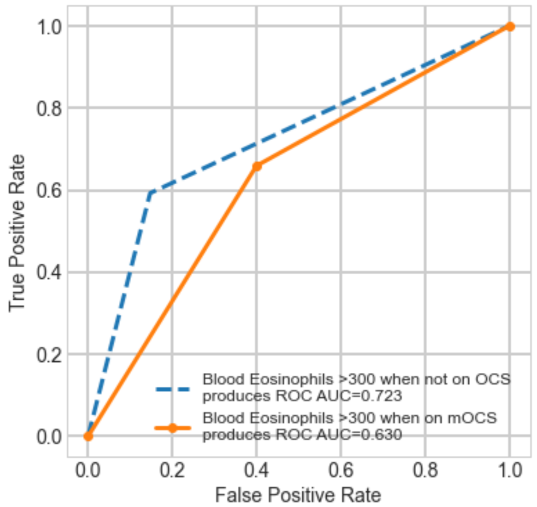


Figure 3.8 Impact of maintenance oral corticosteroids on the Area Under the Curve for Blood Eosinophils ≥300 in Predicting Sputum Eosinophils ≥2%

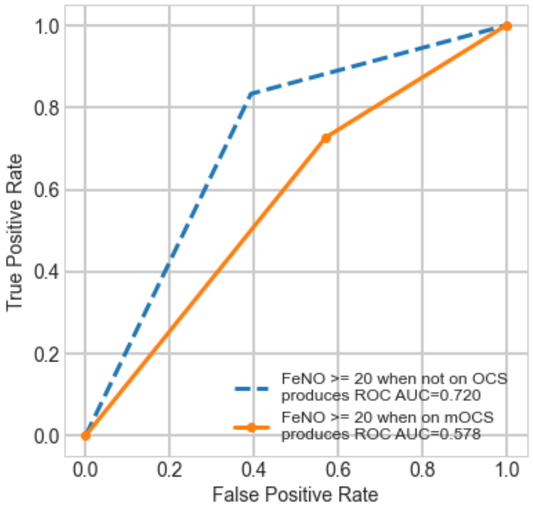


Figure 3.9 Impact of maintenance oral corticosteroids on the Area Under the Curve for Fraction of exhaled Nitric Oxide ≥20 in Predicting Sputum Eosinophils ≥2%

#### 3.3.3.4 Predicting Sputum Neutrophilia with Clinical Biomarkers

The iterative approach (described in 3.2.2) for identifying optimal biomarker cut-offs for predicting sputum eosinophils >2% can be leveraged to predict sputum neutrophilia.

Regardless of the sputum neutrophil cut-off used to define the target phenotype, none of the clinically available biomarkers produced an AUC of greater than 0.55 (Table 3.7, Table 3.8, Table 3.9).

Table 3.5 Area Under the Curve and Positive Predictive Values for for Biomarkers Predicting Sputum Neutrophils >40% Using F1 Score Optimised Cut-offs

Abbreviations: AUC, area under the curver; PPV, positive predictive value; FeNO, fraction of exhaled nitric oxide in breath; IgE, Immunoglobulin E

| Variable | Cut-off | AUC | PPV |
|---|---|---|---|
| **Blood Eosinophil** | 0 | 0.480 | 0.558 |
| **Fraction of Exhaled Nitric Oxide (FeNO)** | 2 | 0.412 | 0.520 |
| **Blood Neutrophils** | 3.1 | 0.536 | 0.593 |
| **Serum Total IgE** | 10 | 0.499 | 0.572 |

Table 3.6 Area Under the Curve and Positive Predictive Values for for Biomarkers Predicting Sputum Neutrophils >61% Using F1 Score Optimised Cut-offs

Abbreviations: AUC, area under the curver; PPV, positive predictive value; FeNO, fraction of exhaled nitric oxide in breath; IgE, Immunoglobulin E

| Variable | Cut-off | AUC | PPV |
|---|---|---|---|
| **Blood Eosinophil** | 0 | 0.565 | 0.358 |
| **Fraction of Exhaled Nitric Oxide (FeNO)** | 5 | 0.469 | 0.293 |
| **Blood Neutrophils** | 4.5 | 0.522 | 0.322 |
| **Serum Total IgE** | 22.1 | 0.533 | 0.336 |

Table 3.7 Area Under the Curve and Positive Predictive Values for Biomarkers Predicting Sputum Neutrophils >76% Using F1 Score Optimised Cut-offs

Abbreviations: AUC, area under the curver; PPV, positive predictive value; FeNO, fraction of exhaled nitric oxide in breath; IgE, Immunoglobulin E

| Variable | Cut-off | AUC | PPV |
|---|---|---|---|
| **Blood Eosinophil** | 0.1 | 0.530 | 0.167 |

| Variable | Cut-off | AUC | PPV |
|---|---|---|---|
| **Fraction of Exhaled Nitric Oxide (FeNO)** | 7 | 0.489 | 0.147 |
| **Blood Neutrophils** | 7.9 | 0.518 | 0.159 |
| **Serum Total IgE** | 27 | 0.525 | 0.164 |

#### 3.3.3.5 Combinations of Biomarkers

Blood eosinophil count thresholds of ≥200, ≥300 and ≥400 produce an AUC of 0.661, 0.748 and 0.719 respectively (Figure 3.7) whilst a FeNO of >20, >30 and >40 produce an AUC of 0.661, 0.693 and 0.647 respectively (Figure 3.8). Using a combination of biomarkers, in an additive or "either-or" manner, produces similar AUC scores (Table 3.10).

Table 3.8 Area Under the Curve for Predicting Sputum Eosinophils ≥2% using a combination of Clinical T2 Biomarkers, Blood Eosinophil count and/or Fraction of Exhaled Nitric Oxide in exhaled breath

| Blood Eosinophil Count | FeNO | AUC using AND conditions | AUC using OR conditions |
|---|---|---|---|
| ≥200 | ≥ 20 | 0.702 | 0.624 |
| ≥200 | ≥30 | 0.683 | 0.671 |
| ≥200 | ≥40 | 0.637 | 0.660 |
| ≥300 | ≥ 20 | 0.754 | 0.660 |
| ≥300 | ≥30 | 0.704 | 0.738 |
| ≥300 | ≥40 | 0.648 | 0.738 |
| ≥400 | ≥ 20 | 0.704 | 0.676 |
| ≥400 | ≥30 | 0.658 | 0.754 |
| ≥400 | ≥40 | 0.622 | 0.734 |

The metric most valuable to clinicians is likely to be the positive predictive value of these biomarkers, rather than the AUC. Blood eosinophil count thresholds of ≥200, ≥300 and ≥400 produce a PPV of 0.633, 0.759, 0.877 respectively whilst a FeNO of >20, >30 and >40 produce a PPV of 0.631, 0.726 and 0.729, respectively (Table 3.5). Combining biomarkers in and additive manner leads to the highest positive predictive value, again, when using the highest available thresholds (blood eosinophils ≥400 and FeNO ≥40): PPV of 0.962 (Table 3.11). Using the T2 biomarkers in an "either or" manner does not lead to an increase in positive predictive value.

Table 3.9 Positive Predictive Value for Sputum Eosinophils ≥2% using a combination of Clinical T2 Biomarkers, Blood Eosinophil count and/or Fraction of Exhaled Nitric Oxide in exhaled breath

| Blood Eosinophil Count | FeNO | PPV using AND conditions | PPV using OR conditions |
|---|---|---|---|
| ≥200 | ≥ 20 | 0.735 | 0.624 |
| ≥200 | ≥30 | 0.810 | 0.671 |
| ≥200 | ≥40 | 0.846 | 0.660 |
| ≥300 | ≥ 20 | 0.889 | 0.660 |
| ≥300 | ≥30 | 0.917 | 0.738 |
| ≥300 | ≥40 | 0.939 | 0.738 |
| ≥400 | ≥ 20 | 0.935 | 0.676 |
| ≥400 | ≥30 | 0.919 | 0.754 |
| ≥400 | ≥40 | 0.962 | 0.734 |

## 3.4 Discussion

### 3.4.1 The Airway Sampling in Severe Asthma Cohort Represents a Severe Asthma Population

The Airway Sampling in Severe Asthma Cohort was established with the aim of characterising a real-world severe asthma patient population. Patients in the cohort were recruited directly from a tertiary NHS service [27], in which complex adult asthma patients receive the same high quality standard of care [26] . All the patients were treated with high dose ICS, with some requiring maintenance oral corticosteroids and/or biologic treatments. Despite this, their asthma was poorly controlled with frequent exacerbations, high T2 biomarkers, poor lung function and poor self-report questionnaire scores. The WATCH cohort has been established as a real-world cohort of difficult to treat asthma patients[238,239] and broad similarities have been observed when comparing to established severe asthma cohorts. As such, The Airway Sampling in Severe Asthma Cohort is broadly similar to the WATCH cohort, except, primarily, with respect to higher objective markers of disease severity: FeNO and blood eosinophil counts, consistent with established severe asthma cohorts [19,240-246]

One striking difference to the established literature is the fact that the Airway Sampling for Severe Asthma cohort demonstrates an almost equal gender split, which contrasts the typically reported 2:1 ratio of females to males. In the WATCH cohort, functional co-morbidities are more commonly observed in female patients [238]. As per the inclusion criteria, these co-morbidities need addressing and optimisation ahead of recruitment, which may have contributed to a sex difference in how quickly new referrals to the asthma clinic were recruited to the study.

Severe asthma cohorts in Southampton tend to be slightly older and treated with higher doses of ICS than other cohorts [238,240] likely reflecting the local demographics and clinical practice respectively. High dose inhaled corticosteroid therapy has long been known to cause systemic effects, specifically adrenal insufficiency [247,248], however the risks have historically been underappreciated, in part due to the lack of an alternative [249]. More recently, with the advent of novel steroid sparing agents, greater focus has been placed on steroid stewardship [250], such that the rates of steroid therapy seen in this cohort will become less frequent. The impact of high dose steroid therapy also has biological implications, including altering the airway inflammatory cell profile [251] and possibly the airway microbiome [252], which may impact the generalisability of the findings from this thesis.

### 3.4.2 Sputum Eosinophils Represent a Robust Phenotype but Existing Biomarkers have Limitations

Consistency between the Airway Sampling in Severe Asthma Cohort and the established severe asthma literature is substantiated by analysis of sputum inflammatory phenotypes. When patients are stratified by sputum eosinophil and neutrophil counts, the eosinophilic phenotype is usually most prevalent [192,236]. This is the case in the Airway Sampling in Severe Asthma Cohort, though the proportions of each phenotype are more evenly distributed, possibly due to the higher doses of prescribed ICS, which are known to suppress sputum eosinophilia [64]. Nevertheless, patients with sputum eosinophilia in the Airway Sampling in Severe Asthma Cohort suffer from the most frequent exacerbations, have a later age of onset, more nasal polyps, higher FeNO, worse post BD spirometry but greater lung function reversibility. This phenotype is entirely consistent with the T2 High phenotype [71].

The importance of identifying frequent exacerbators cannot be understated in asthma. Sputum eosinophilia is a treatable trait [253] but if not identified and/or left untreated, can lead to asthma exacerbations, which can be fatal [254], have significant direct and indirect costs to society [255,256] and lead to accelerated loss of lung function [257]. As discussed extensively in the Introduction, blood eosinophils and FeNO have become established surrogates for airway eosinophilia [65,258], however, it is recognised that these biomarkers are sensitive to steroid treatment [84,259] often leading to underestimation of airway eosinophilia [259]. Considering the high dose ICS treatment of patients in this cohort, it is not surprising that the AUCs for FeNO and blood eosinophils for predicting sputum eosinophilia in the Airway Sampling in Severe Asthma Cohort do not match the AUCs of 0.8 reported in patients with mild asthma (on low doses of steroid therapy) [65]. This is reinforced by the observation that prescription of maintenance oral steroids yields a further drop in AUC performance for these biomarkers. There are numerous reasons why clinical practice forgives this limitation; amongst them are the fact that sensitivity to treatment can be repurposed advantageously, such as in the FeNO Suppression Test [260]. This limitation is nevertheless important to consider when assessing novel biomarkers.

The observation that FeNO and Blood Eosinophils have moderately positive correlations with sputum eosinophilia but only share a weak correlation with each other corroborates the fact that they reflect distinct mechanisms [261]: blood eosinophils reflect IL-5 regulated migration of eosinophils from bone marrow to the circulation [262] and FeNO the increased production of NO due to IL-13 upregulation of inducible NO synthases in airway epithelial and inflammatory cells [263]. This distinction in mechanisms explains why, in addition to independently predicting an increased risk of asthma exacerbation, an elevation in both biomarkers results in an additive effect to exacerbation risk [83,264]. This distinction becomes acute and clinically relevant when we consider biologic therapies. Anti-IL-5 treatments, such as Mepolizumab, lead to a reduction in blood eosinophils but not FeNO [265], whist anti IL-4/13 treatments, such as Dupilumab, lead to a reduction in FeNO but not blood eosinophils [266]. Unsurprisingly, therefore, blood eosinophils are useful predictors for anti-IL-5 treatment response, but FeNO is not [267]. This, in turn, underlines the fact that sputum eosinophilia is heterogenous [198]. As argued, for biomarkers for sputum eosinophilia, this heterogeneity is important to consider when assessing novel biomarkers.

### 3.4.3 T2 Low Phenotypes Remain Poorly Defined

In the Airway Sampling in Severe Asthma Cohort, there are very few distinguishing features for the T2 low asthma phenotypes. Paucigranular patients were more often obese females with high disease burden despite relatively well-preserved lung function; a phenotype that has been described across numerous cluster analyses [16,238,268]. Obesity, however, is commonly seen to be a characteristic of patients with airway neutrophilia [269]: a coalescence only observed in patients with mixed eosinophilic and neutrophilic inflammation in the Airway Sampling in Severe Asthma Cohort and not pure neutrophilic patients. Other commonly cited features of T2 low asthma include older age [270,271] and association with smoking [272], features that are not observed in the Airway Sampling in Severe Asthma Cohort. These characteristics are most commonly identified when clustering on clinical characteristics, which may explain this discrepancy.

Existing clinical biomarkers are able to hint at biological heterogeneity in the eosinophilic but not neutrophilic phenotype which remains poorly understood [273]. In addition to being unable to identify a clear neutrophilic phenotype, we are unable to find useful surrogate markers for airway neutrophilia [81,88]. Just as with sputum eosinophilia, this may suggest that the target is too poorly defined and that the airway neutrophils are themselves a biomarker for another process, possibly the airway microbiome [208,274]

### 3.4.4 Conclusions

The data presented demonstrates that the Airway Sampling in Severe Asthma Cohort represents a real-world severe asthma cohort, maximising the translational potential of the findings in this and subsequent thesis chapters. Describing patients by the inflammatory cell counts in induced sputum samples successfully identifies a phenotype of frequent exacerbators, characterised by airway eosinophilia. However, inflammatory phenotyping is unsuccessful in stratifying patients with airway eosinophilia in a meaningful way. These data corroborate the finding that existing biomarkers for the sputum eosinophilic phenotype are good but have their limitations. Finally, these data demonstrate evidence that sputum eosinophilia is likely heterogenous in itself and warrants further description.

# Chapter 4 Granulocyte Activation Markers in Severe Asthma

## 4.1 Introduction

As described and discussed in 1.4.1 and 3.4, quantitative cytometric analysis of induced sputum samples is a useful and well established method for which to describe the airway inflammation in severe asthma patients [191,193] but it has been proposed that inflammatory phenotypes would be better defined by granulocyte activity rather than granulocyte presence [200]. Alongside a milieu of proinflammatory chemokines and cytokines, eosinophils release four cationic and acid granule proteins: major basic protein, eosinophil cationic protein (ECP), eosinophil-derived neurotoxin (EDN) and eosinophil peroxidase by piecemeal degranulation, exocytosis or cytolysis in response to external stimuli [275]. Neutrophils similarly release granule enzymes, such as Neutrophil Elastase (NE) and Myeloperoxidase (MPO) [276].

Sputum MPO is associated with sputum neutrophilia [47] and sputum ECP and EDN with sputum eosinophilia [277]. However, limiting the definition of an inflammatory phenotypes to sputum cell percentages may underestimate their prevalence, because granulocytes that have degranulated may be missed [278]. Moreover, the proteins are themselves biologically active and produce local (in chronic rhinosinusitis) inflammatory effects. [279].

Compared to healthy controls, Sputum ECP and EDN is elevated in patients with asthma [280] and increases with asthma severity [281]. Whether measurement of these proteins can discriminate between inflammatory phenotypes or inform new inflammatory phenotypes has not been assessed. Similarly, where serum measurement of these proteins increases with asthma severity [282,283], they have been assessed as diagnostic biomarkers but not for phenotyping in severe asthma.

This chapter seeks to describe the granulocyte activation markers in the Airway Sampling in Severe Asthma Cohort. The objectives of this chapter are

- Describe the granulocyte activation markers in the a priori defined sputum inflammatory phenotypes in sputum and serum
- Assess the serum measures of granulocyte activation as biomarkers for sputum inflammatory phenotypes

- Use sputum measures of granulocyte activation to define new phenotypes

## 4.2 Chapter Specific Methods

### 4.2.1 Patient Population

The patients described in this chapter were recruited from the Airway Sampling in Severe Asthma WATCH Sub-Study

### 4.2.2 Statistical Analysis

Statistical analysis was performed using Python scripting language (version 3.8.3) [237]. Clinical characteristics were described using median and 95% confidence intervals with between group comparisons by Mann Whitney U tests for continuous variables and absolute numbers with percentages within each group and Chi Squared tests for categorical variables.

A Jenks optimization method was used to determine the best arrangement of sputum EDN measures into two different classes. In order to prevent the algorithm from simply removing outliers at the tail of the histogram, the algorithm excluded values in the 90th centile and above.

Clustering was performed on the granulocyte degranulation measures only: MPO, neutrophil Elastase, EDN and ECP. These measures were scaled and then used to calculate the Euclidean distance between each subject. Ward's hierarchical clustering was used to stratify patients. Visual inspection of the dendrogram was used to determine the number of clusters.

## 4.3 Results

### 4.3.1 Sputum Granulocyte Activation Markers Across Inflammatory Phenotypes

Eosinophil activation markers (ECP and EDN) in sputum were most elevated in the eosinophilic phenotypes: mixed granular and eosinophilic; sputum MPO was highest in the neutrophilic phenotypes: mixed granular and neutrophilic whilst Sputum Elastase was highest in eosinophilic patients (Table 4.1).

Table 4.1 Granulocyte Activation Markers in Sputum Across Sputum Inflammatory Phenotypes

Continuous variables expressed as median [Q1, Q3] with differences measured by Kruskal Wallis test. Abbreviations: MPO, myeloperoxidase; NE, neutrophil elastase; EDN, eosinophil derived neurotoxin; ECP, eosinophil cationic protein

| | **Paucigranular N = 62)** | **Eosinophilic (n = 71)** | **Neutrophilic (n = 34)** | **Mixed Granular (n = 27)** | **P-Value** |
|---|---|---|---|---|---|
| **Sputum MPO** | 818.8 [414.3,1821.0] | 753.7 [440.2,1239.8] | 2137.0 [1553.5,6176.5] | 1815.0 [1134.0,4081.0] | **<0.001** |
| **Sputum NE** | 6.5 [3.2,9.7] | 8.2 [4.5,13.6] | 4.4 [3.1,6.0] | 4.8 [3.2,8.6] | **0.017** |
| **Sputum EDN** | 113.4 [36.5,295.7] | 917.5 [331.6,1474.0] | 273.8 [159.5,810.3] | 1224.0 [421.0,1729.0] | **<0.001** |
| **Sputum ECP** | 67.1 [17.4,264.9] | 961.0 [353.4,1863.5] | 367.3 [118.2,1412.0] | 1149.0 [269.1,3269.0] | **<0.001** |

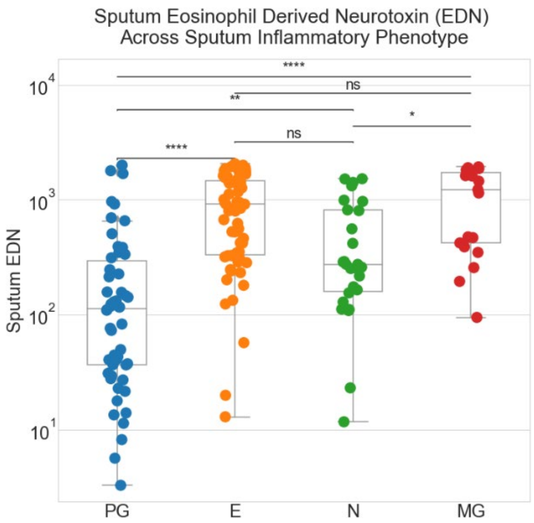


Figure 4.1 Sputum Eosinophil Derived Neurotoxin (EDN) Concentrations Across Sputum Inflammatory Phenotypes

Abbreviations: PG, paucigranular; E, eosinophilic; N, neutrophilic; MG, mixed granular

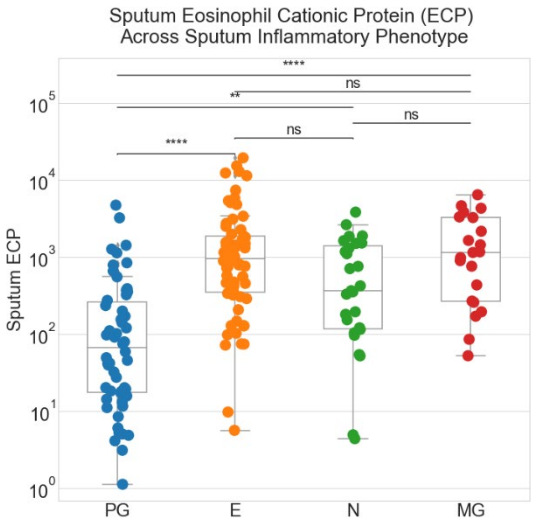


Figure 4.2 Sputum Eosinophil Cationic Protein (ECP) Concentrations Across Sputum Inflammatory Phenotype

Abbreviations: PG, paucigranular; E, eosinophilic; N, neutrophilic; MG, mixed granular

On post hoc pairwise comparisons, there is no difference in sputum ECP or sputum EDN between eosinophilic and mixed granular patients but both are increased compared to paucigranular patients consistent with an "eosinophilic distribution" (Figure 4.1 and Figure 4.2). Sputum ECP and EDN are also increased in neutrophilic patients (though not to the same extent as eosinophilic or mixed granular patients) compared to paucigranular patients (Figure 4.1 and Figure 4.2).

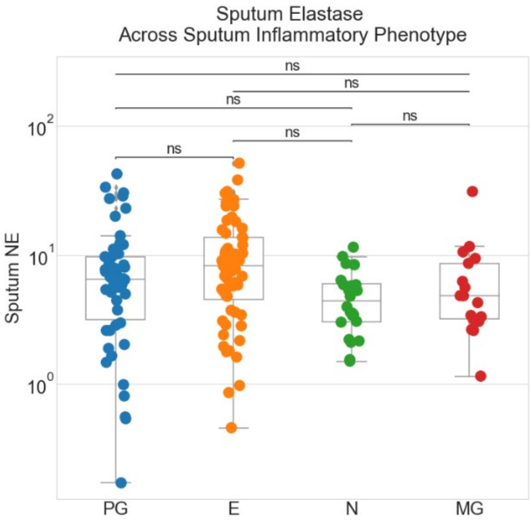


Figure 4.3 Sputum Neutrophil Elastase Concentrations Across Sputum Inflammatory Phenotypes

Abbreviations: PG, paucigranular; E, eosinophilic; N, neutrophilic; MG, mixed granular

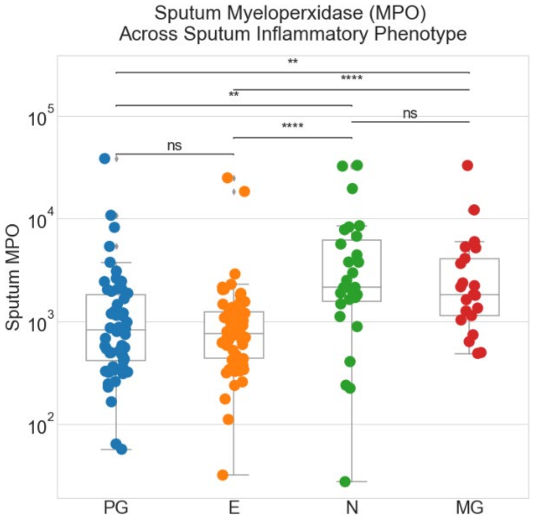


Figure 4.4 Sputum Myeloperoxidase (MPO) Concentrations Across Sputum Inflammatory Phenotypes

Abbreviations: PG, paucigranular; E, eosinophilic; N, neutrophilic; MG, mixed granular

Though sputum Elastase concentrations were statistically significantly different on Kruskal-Wallis testing (Table 4.1), no statistically significant differences were seen on post hoc pairwise comparisons (Figure 4.3).

Sputum MPO concentrations followed a neutrophilic distribution: higher in the Neutrophilic and Mixed Granular groups compared to paucigranular and eosinophilic phenotypes but not statistically significant difference between them (Table 4.1 and Figure 4.4).

### 4.3.2 Serum Granulocyte Activation Markers Across Inflammatory Phenotypes

There were no differences in serum activation markers between the sputum inflammatory phenotypes other than Serum ECP (p= 0.016) with levels higher in paucigranular patients (57.4 (18.0,121.4)), double that of the eosinophilic, mixed granular and neutrophilic phenotypes, which were otherwise similar (25.9 (12.1,58.2), 27.0 (12.5,47.9) and 28.0 (13.3,55.4), respectively) (Table 4.2).

Table 4.2 Granulocyte Activation Markers in Serum Across Sputum Inflammatory Phenotypes

Continuous variables expressed as median [Q1, Q3] with differences measured by Kruskal Wallis test. Abbreviations: MPO, myeloperoxidase; NE, neutrophil elastase; EDN, eosinophil derived neurotoxin; ECP, eosinophil cationic protein

| | Paucigranular<br>N = 62) | Eosinophilic<br>(n = 71) | Neutrophilic<br>(n = 34) | Mixed Granular<br>(n = 27) | P-Value |
|---|---|---|---|---|---|
| Serum MPO | 289.5 [195.4,380.9] | 235.7 [167.9,308.2] | 246.0 [181.3,307.7] | 243.5 [177.0,337.5] | **0.157** |
| Serum NE | 58.3 [38.3,94.6] | 39.2 [28.2,78.8] | 37.9 [33.0,55.1] | 44.6 [18.7,87.4] | **0.132** |
| Serum EDN | 54.9 [36.4,77.8] | 42.0 [33.3,55.6] | 47.8 [32.5,57.8] | 42.2 [35.7,61.2] | **0.217** |
| Serum ECP | 57.4 [18.0,121.4] | 25.9 [12.1,58.2] | 28.0 [13.3,55.4] | 27.0 [12.5,47.9] | **0.016** |

There were no statistically significant differences in serum EDN, serum MPO or serum Elastase across the sputum inflammatory phenotypes (Table 4.2). Serum ECP was increased in paucigranular patients compared to the three other phenotypes, in whom there was no statistically significant difference between each other (Table 4.2).

### 4.3.3 Serum Granulocyte Activation Markers as Biomarkers for Sputum Inflammatory Phenotypes

Sputum EDN shared moderate positive correlations sputum eosinophils (r= 0.667, p<0.001) as well as blood eosinophils and FeNO (0.468, p<0.001 and r = 0.417, p<0.001, respectively). However, there was no correlation between serum EDN and sputum EDN, nor serum EDN with sputum eosinophils, blood eosinophils or FeNO (Figure 4.5).

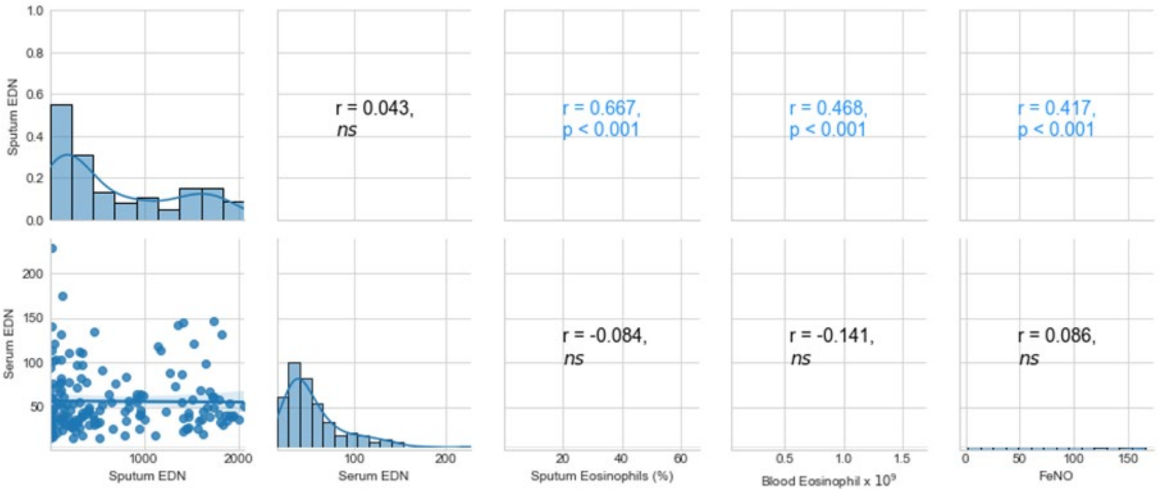


Figure 4.5 Spearman Rank Correlations Between Serum and Sputum Eosinophil Derived Neurotoxin (EDN) with each other, Sputum Eosinophils and Clinical T2 Biomarkers, Blood Eosinophils and Fraction of Exhaled Nitric Oxide (FeNO)

Sputum ECP shared moderate positive correlations sputum eosinophils (r= 0.588, p<0.001) as well as blood eosinophils and FeNO (0.448, p<0.001 and r = 0.407, p<0.001, respectively). Similar to EDN, there was no correlation between serum ECP and sputum ECP. There was no correlation between serum ECP and sputum eosinophils or FeNO, though a weakly negative correlation was seen between serum ECP and blood eosinophils (Figure 4.6).

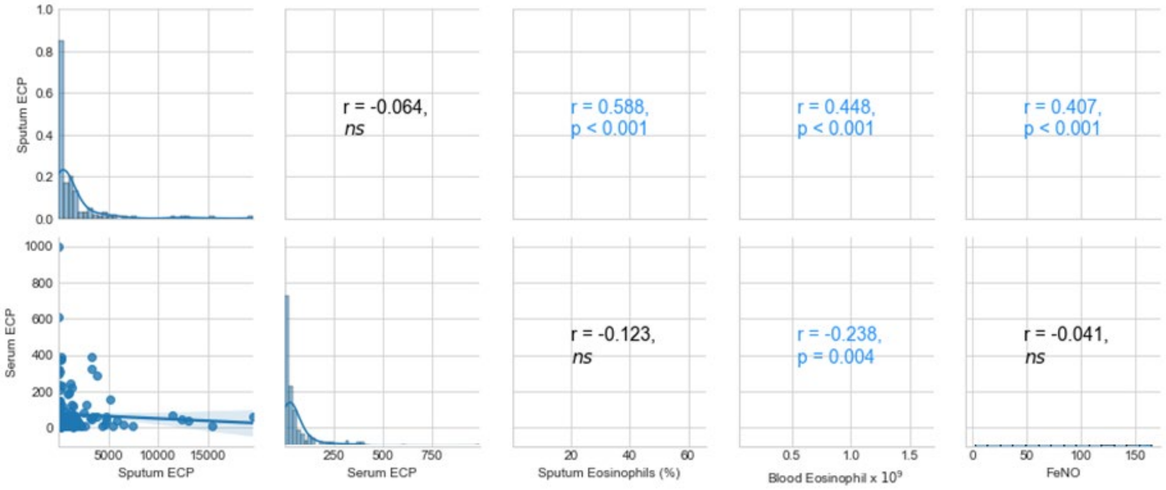


Figure 4.6 Spearman Rank Correlations Between Serum and Sputum Eosinophil Cationic Protein (ECP) with each other, Sputum Eosinophils and Clinical T2 Biomarkers, Blood Eosinophils and Fraction of Exhaled Nitric Oxide (FeNO)

No correlations were seen between serum and sputum MPO and sputum eosinophils, blood eosinophils or FeNO (Figure 4.7). Sputum MPO did share a moderately positive correlation with sputum neutrophils (r = 0.659, p<0.001) but not blood neutrophils. No correlations were observed between serum MPO and either sputum or blood neutrophils (Figure 4.8)

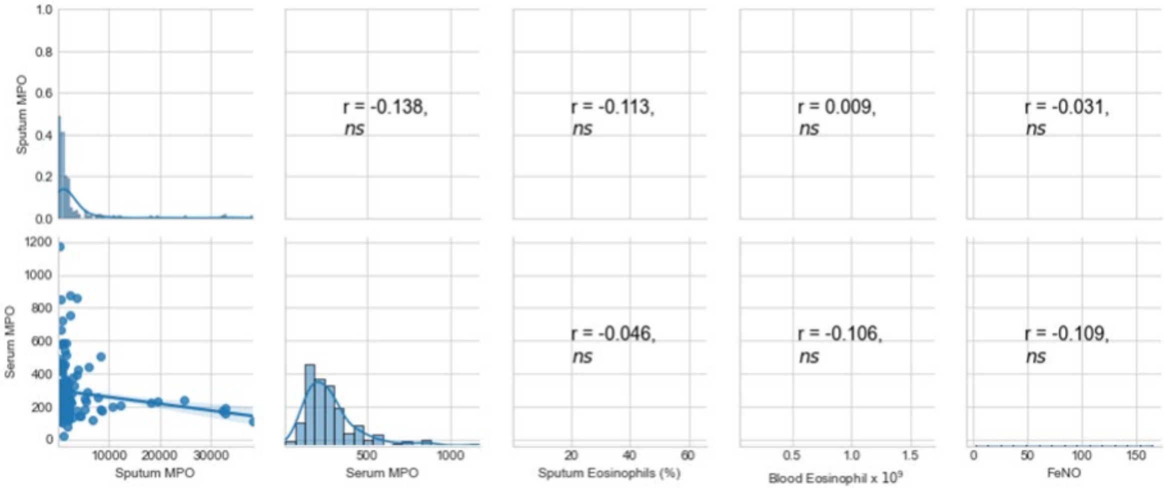


Figure 4.7 Spearman Rank Correlations Between Serum and Sputum Myeloperoxidase (MPO) with each other, Sputum Eosinophils and Clinical T2 Biomarkers, Blood Eosinophils and Fraction of Exhaled Nitric Oxide (FeNO)

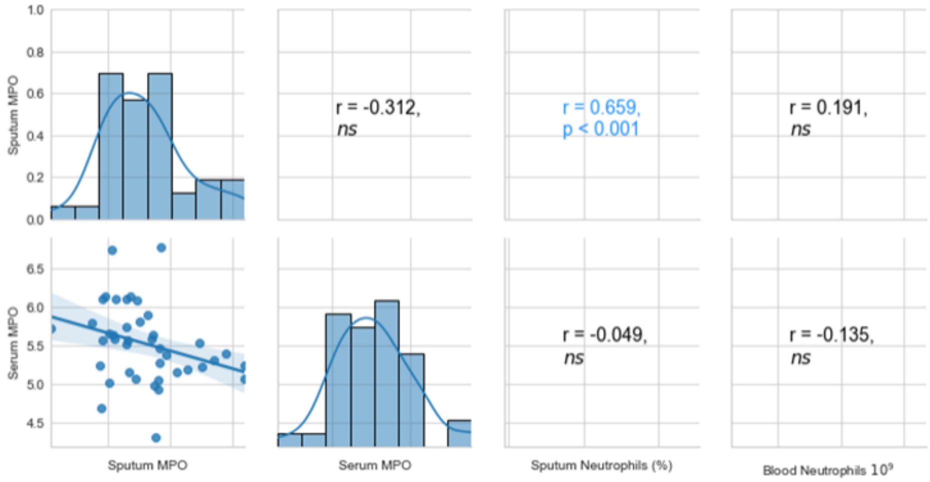


Figure 4.8 Spearman Rank Correlations Between Serum and Sputum Myeloperoxidase (MPO) with each other as well as Sputum and Serum Neutrophils

There were weakly positive correlations between sputum elastase and sputum eosinophils, blood eosinophils and FeNO (r = 0.250, p=0.007, r=0.214, p=0.016, r=0.199, p=0.023, respectively) but nor correlations with sputum or blood neutrophils. There were no correlations between serum elastase and sputum elastase or sputum eosinophils, sputum neutrophils, blood eosinophils, blood neutrophils or FeNO (Figure 4.9 and Figure 4.10).

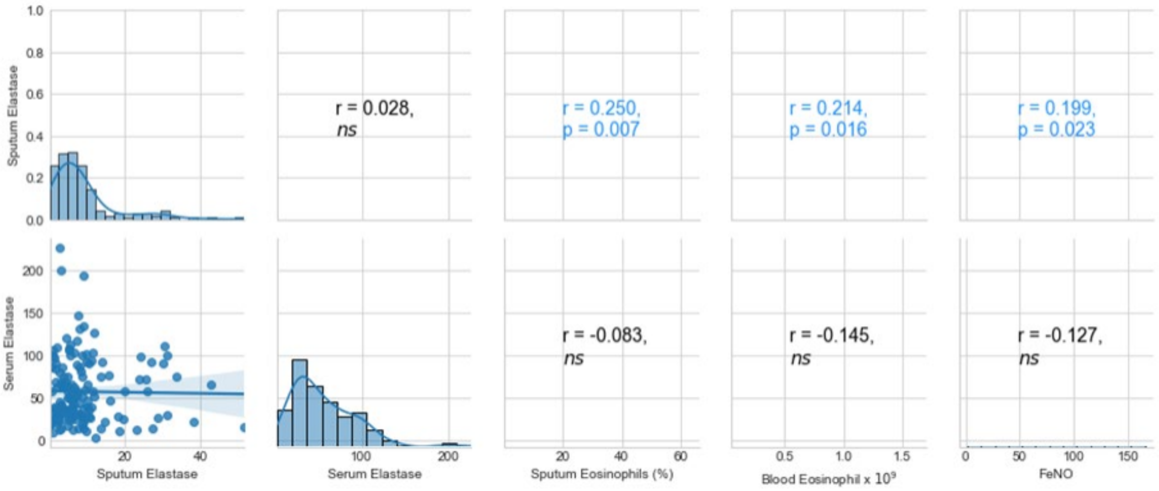


Figure 4.9 Spearman Rank Correlations Between Serum and Sputum Neutrophil Elastase (NE) with each other, Sputum Eosinophils and Clinical T2 Biomarkers, Blood Eosinophils and Fraction of Exhaled Nitric Oxide (FeNO)

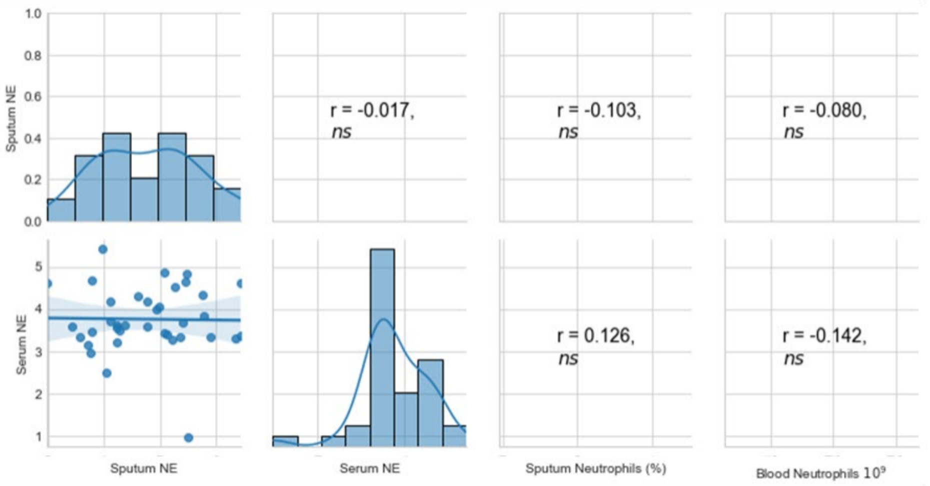


Figure 4.10 Spearman Rank Correlations Between Serum and Sputum Neutrophil Elastase (NE) with each other as well as Sputum and Serum Neutrophils

Applying the data driven approach to identify the optimal cut-off of measures (3.5.2), serum markers of inflammatory cell activation are poor predictors of sputum eosinophils of >2%: none of serum MPO, elastase, EDN or ECP achieved an AUC more 0.52 (Table 4.3). They are similarly limited in predicting sputum neutrophils of >40% (Table 4.4), >61% (Table 4.5) and >76% (Table 4.6).

Table 4.3 Area Under the Curve and Positive Predictive Value for Serum Cell Activation Biomarkers in Predicting Sputum Eosinophils >2%

| Variable | Cut-off | AUC | PPV |
|---|---|---|---|
| **Serum Myeloperoxidase (MPO)** | 74.82 | 0.495 | 0.508 |
| **Serum Elastase** | 8.585 | 0.505 | 0.513 |
| **Serum Eosinophil Derived Neurotoxin (EDN)** | 24.8 | 0.523 | 0.523 |
| **Serum Eosinophil Cationic Protein (ECP)** | 4.581 | 0.505 | 0.513 |

Table 4.4 Area Under the Curve and Positive Predictive Value for Serum Cell Activation Biomarkers in Predicting Sputum Neutrophils >40%

| Variable | Cut-off | AUC | PPV |
|---|---|---|---|
| **Serum Myeloperoxidase (MPO)** | 131.7 | 0.495 | 0.577 |
| **Serum Elastase** | 12.39 | 0.506 | 0.582 |
| **Serum Eosinophil Derived Neurotoxin (EDN)** | 14.18 | 0.480 | 0.568 |
| **Serum Eosinophil Cationic Protein (ECP)** | 4.654 | 0.506 | 0.582 |

Table 4.5 Area Under the Curve and Positive Predictive Value for Serum Cell Activation Biomarkers in Predicting Sputum Neutrophils >61%

| Variable | Cut-off | AUC | PPV |
|---|---|---|---|
| **Serum Myeloperoxidase (MPO)** | 117.3 | 0.492 | 0.317 |
| **Serum Elastase** | 12.39 | 0.504 | 0.323 |
| **Serum Eosinophil Derived Neurotoxin (EDN)** | 18.35 | 0.506 | 0.324 |
| **Serum Eosinophil Cationic Protein (ECP)** | 5.848 | 0.504 | 0.323 |

Table 4.6 Area Under the Curve and Positive Predictive Value for Serum Cell Activation Biomarkers in Predicting Sputum Neutrophils >76%

| Variable | Cut-off | AUC | PPV |
|---|---|---|---|
| **Serum Myeloperoxidase (MPO)** | 158.1 | 0.503 | 0.159 |
| **Serum Elastase** | 36.52 | 0.503 | 0.159 |
| **Serum Eosinophil Derived Neurotoxin (EDN)** | 29.04 | 0.524 | 0.165 |
| **Serum Eosinophil Cationic Protein (ECP)** | 12.27 | 0.503 | 0.159 |

### 4.3.4 Defining Inflammatory Phenotypes Using Granulocyte Activation Markers

Sputum EDN and ECP show a moderate relationship with sputum eosinophils, and sputum MPO with sputum neutrophils. This imperfect overlap indicates that they may be giving slightly different information and the possibility that patients may be discordant e.g. have high sputum eosinophils but low sputum EDN or low sputum eosinophils but high sputum EDN.

#### 4.3.4.1 Defining Eosinophil Activation Phenotype

Visual inspection of scatterplots identifies of sputum eosinophil and sputum EDN (Figure 4.11) and sputum ECP (Figure 4.12) demonstrates potential discordance. In patients with high sputum eosinophils, there is a heterogeneity in sputum EDN and in patients with low sputum ECP, there is heterogeneity in sputum eosinophils.

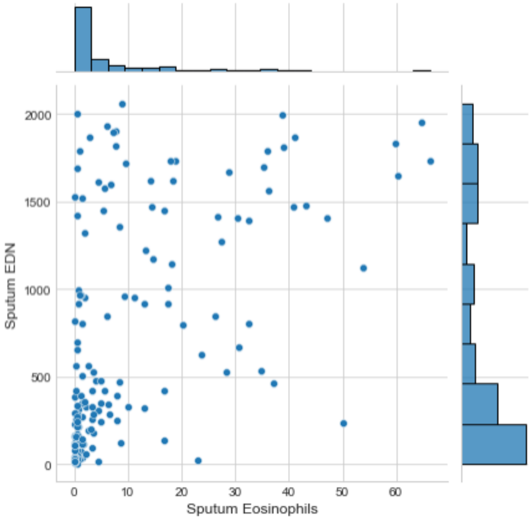


Figure 4.11 Scatterplot with marginal histograms of Sputum Eosinophil Derived Neurotoxin (EDN) and Sputum Eosinophils

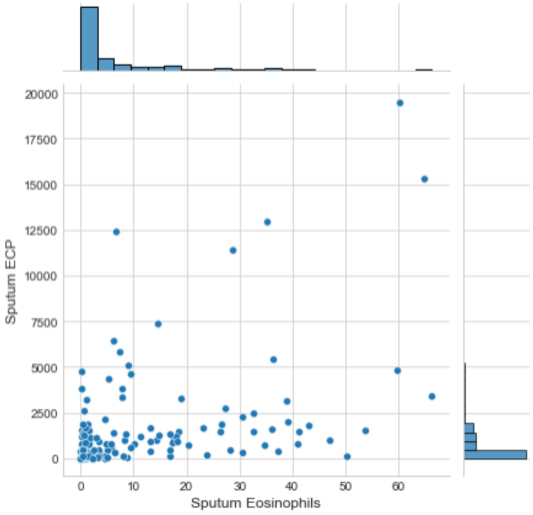


Figure 4.12 Scatterplot with marginal histograms of Sputum Eosinophil Cationic Protein (ECP) and Sputum Eosinophils

The histogram of sputum EDN indicates that this variable may lend itself well to exploration as a dichotomous variable (right-hand y axis of Figure 4.11): there appears to be a bimodal distribution with the first peak close to 0 and a second peak at around 1500 and reasonable numbers associated with the second peak. By contrast, the histogram of sputum ECP (right-hand y axis of Figure 4.12) only indicates a right skewed distribution which would not immediately lend itself to dichotomisation with the same being true for sputum MPO (Figure 4.13and Figure 4.14) A natural break in sputum EDN was identified using a Jenks one dimensional clustering algorithm at 697 (Figure 4.15)

.

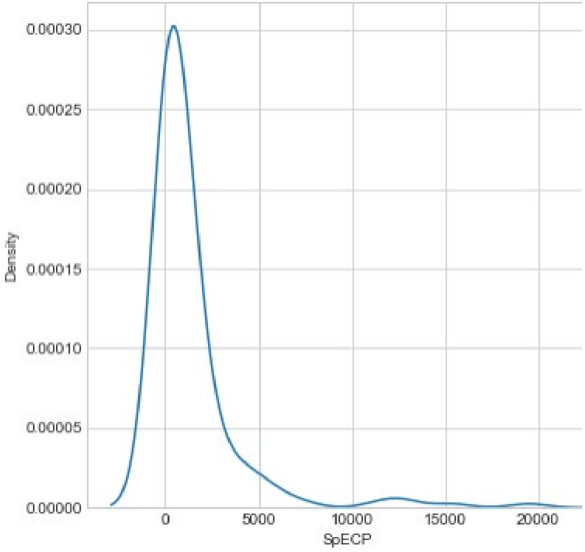


Figure 4.13 Kernel Density Estimate Plot visualizing the distribution of Sputum Eosinophil Cationic Protein (ECP)

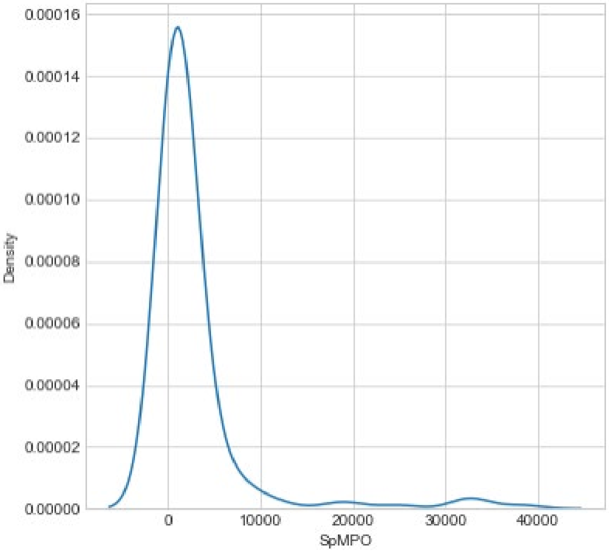


Figure 4.14 Kernel Density Estimate Plot visualizing the distribution of Sputum Myeloperoxidase (MPO)

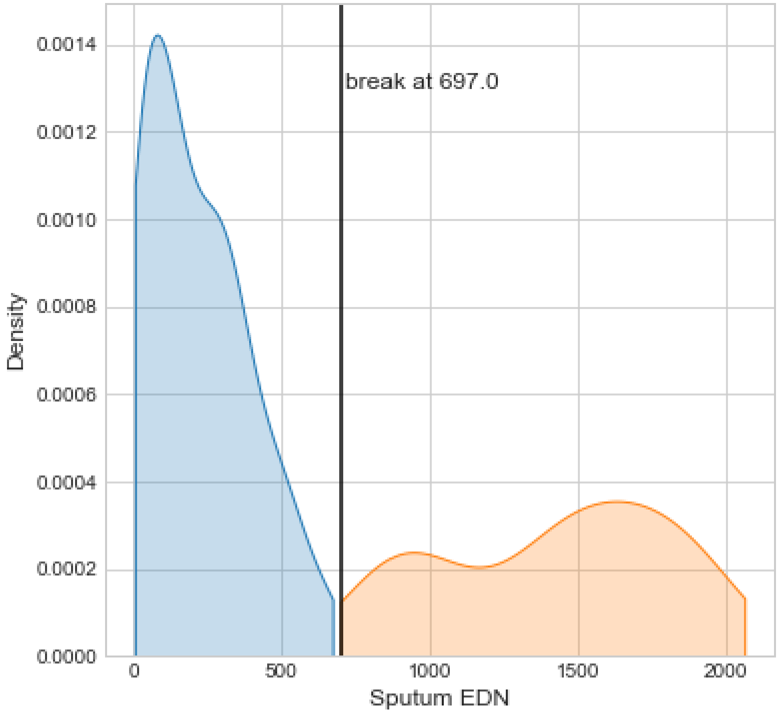


Figure 4.15 Kernel Density Estimation Plot Visualising High and Low Sputum Eosinophil Derived Neurotoxin (EDN) Defined by Identifying a Jenks Natural Break in the Distribution of EDN

Table 4.7 Clinical Characteristics in Patients with Low and High Sputum Eosinophil Derived Neurotoxin (EDN)

Continuous variables expressed as median [Q1, Q3] with differences measured by Mann-Whitney U test. Categorical variables expressed as n (%) with differences measured by chi-square test. Abbreviations: GORD, gastro-oesophageal reflux disease; ICS, inhaled corticosteroid; BDPe, beclomethasone dose equivalent; OCS, oral corticosteroids; IgE, Immunoglobulin E; IL-5, Interleukin 5; ACQ, asthma control questionnaire, HADSTOT, Hospital Anxiety and Depression Total Score; SNOT, SinoNasal Outcome Score; FeNO, fraction of nitric oxide in exhaled breath; post BD, post bronchodilator; FEV1, forced expiratory volume in 1 second; FVC, forced vital capacity; FEF25-75%, forced expiratory flow at 25% to 75% of FVC

| | **Low Sputum EDN (n=131)** | **High Sputum EDN (n=64)** | **P Value** |
|---|---|---|---|
| Female Sex | 74 (56.5) | 29 (45.3) | 0.188 |
| Age | 54.0 [44.0,66.0] | 59.5 [52.2,68.0] | 0.081 |
| BMI | 30.1 [26.1,35.5] | 27.6 [25.6,31.2] | 0.027 |

| | Low Sputum EDN (n=131) | High Sputum EDN (n=64) | P Value |
|---|---|---|---|
| Never Smoker | 81  (61.8) | 43 (67.2) | 0.601 |
| Atopic | 83  (63.4) | 36 (56.2) | 0.424 |
| GORD | 82  (63.6) | 38 (59.4) | 0.684 |
| Nasal Polyps | 28 (21.7) | 20 (31.7) | 0.183 |
| Age of Onset | 15.0 [3.5,37.0] | 28.0 [6.5,49.0] | 0.153 |
| Exacerbations in the Last 12 Months | 1.0 [1.0,3.0] | 3.0 [0.0,5.0] | 0.116 |
| ICS (BDPe) | 3000.0 [2000.0,3200.0] | 3000.0 [2000.0,3125.0] | 0.897 |
| Maintenance OCS | 56 (42.7) | 23 (35.9) | 0.451 |
| Anti IgE | 14 (10.7) | 5 (7.8) | 0.705 |
| Anti IL-5 | 20 (15.3) | 4 (6.2) | 0.117 |
| ACQ6 | 2.3 [1.5,3.1] | 2.3 [1.3,3.4] | 0.904 |
| HADS | 11.0 [6.8,16.0] | 9.0 [6.0,15.0] | 0.236 |
| SNOT20 | 29.0 [16.2,45.0] | 32.0 [23.2,44.8] | 0.423 |
| FeNO | 21.0 [14.0,38.0] | 35.0 [22.5,65.0] | <0.001 |
| PostBD FEV1 (% predicted) | 78.3 [62.3,93.6] | 68.0 [52.9,86.5] | 0.006 |
| PostBD FVC (% predicted) | 91.6 [77.3,102.7] | 91.6 [75.9,101.2] | 0.605 |
| PostBD FEV1/FVC | 70.0 [62.5,80.0] | 62.5 [51.7,69.2] | <0.001 |
| PostBD FEF25-75% (% predicted) | 51.3 [35.8,83.1] | 34.6 [21.7,59.5] | <0.001 |
| FEV1 Reversibility | 8.9 [3.7,16.3] | 10.1 [3.6,17.4] | 0.995 |
| Blood Neutrophils | 5.2 [4.2,7.0] | 5.3 [4.2,6.2] | 0.424 |
| Blood Eosinophils | 0.2 [0.1,0.3] | 0.3 [0.2,0.6] | <0.001 |
| Serum Total IgE | 77.7 [19.6,362.0] | 137.0 [37.4,289.9] | 0.309 |
| Sputum Neutophils | 45.9 [24.9,68.1] | 36.8 [22.3,66.6] | 0.484 |
| Sputum Eosinophils | 0.9 [0.2,3.5] | 13.7 [4.1,31.0] | <0.001 |

High sputum EDN was seen in 32.8% of patients. These patients had a higher FeNO (34.0 (22.5-65.0) vs 21.0 (14.0-38.0), p<0.001), higher blood eosinophil count (0.3 (0.2-0.6) vs 0.2 (0.1-0.3), p<0.001) and sputum eosinophils (13.7 (4.1-31.0) vs 0.9 (0.2-3.5), p<0.001). They also had worse lung function in terms of postBD FEV1 (% predicted) (68.0 (52.0 - 86.5), vs 78.3 (62.3 - 93.6), p =0.006) and postBD FEV1/FVC (62.5 (51.7 - 69.2) vs 70.0 (62.5 - 80.0), p<0.001). However, there were no statistically significant differences between sputum EDN high and low patients in terms T2 clinical characteristics, such as presence of nasal polyps or later age of onset, or of clinical endpoints such as exacerbations or ACQ6 (Table 4.7).

#### 4.3.4.2 Discordance Between Eosinophil Presence and Activation

Patients were categorised in a 2 x 2 matrix of Sputum EDN low and high, using the 697 cut-off described above) and sputum eosinophil low and high, using a 2% cut-off. 83 (42.56%) patients had no evidence of eosinophil presence or of high activity and 50 (25.64%) patients had evidence of eosinophil presence and high activity. 48 (24.62%) of patients had evidence of eosinophil presence but not high activity and 14 (7.18%) of patients had no evidence of eosinophil but high activity (Figure 4.16 ).

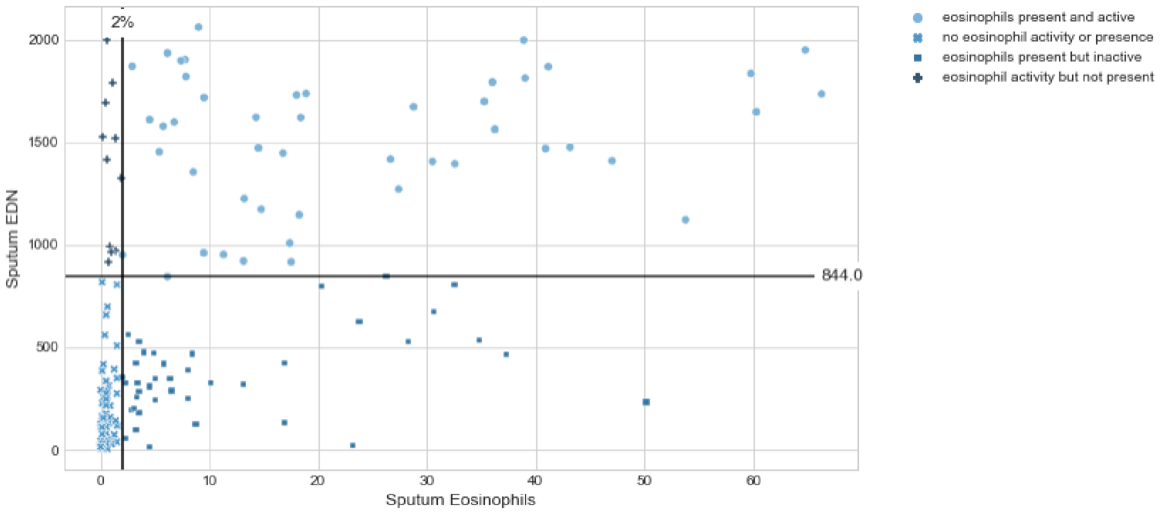


Figure 4.16 Scatterplot of Sputum Eosinophils and Sputum Eosinophil Derived Neurotoxin (EDN).

Horizontal line delineates high and low EDN using the Jenks Natural Break at 697.

Vertical line delineates high and low sputum eosinophils using the 2% cut-off.

The clinical characteristics of patients categorised in this manner is described in Table 4.8. T2 characteristics, such as age of onset and presence of nasal polyps, are statistically significant between these groups, as are objective markers of severity such as FeNO, post BD lung function and exacerbation frequency. The non-eosinophilic (no eosinophils and no evidence of high activity) patients have the youngest age of onset, fewest exacerbations, lowest FeNO and most persevered post BD lung function. Patients with no eosinophils but high EDN appear most similar to the non-eosinophilic group.

Table 4.8 Clinical Characteristics in Patients Stratified by Eosinophil Presence, using Sputum Eosinophil Counts, and Activity, using Eosinophil Derived Neurotoxin (EDN)

Continuous variables expressed as median [Q1, Q3] with differences measured by Kruskal Walis test. Categorical variables expressed as n (%) with differences measured by chi-square test. Abbreviations: GORD, gastro-oesophageal reflux disease; ICS, inhaled corticosteroid; BDPe, beclomethasone dose equivalent; OCS, oral corticosteroids; IgE, Immunoglobulin E; IL-5, Interleukin 5; ACQ, asthma control questionnaire, HADSTOT, Hospital Anxiety and Depression Total Score; SNOT, SinoNasal Outcome Score; FeNO, fraction of nitric oxide in exhaled breath;  post BD, post bronchodilator; FEV1, forced expiratory volume in 1 second; FVC, forced vital capacity; FEF25-75%, forced expiratory flow at 25% to 75% of FVC. Eosinophil; eos, Sputum measure of eosinophil derived neurotoxin; SpEDN,

| | **Sputum Eos Low AND SpEDN Low (n= 83)** | **Sputum Eos Low BUT SpEDN High (n=14)** | **Sputum Eos High BUT SpEDN Low (n=48)** | **Sputum Eos High AND SpEDN High (n=50)** | **P-Value** |
|---|---|---|---|---|---|
| Female Sex | 48 (57.8) | 8 (57.1) | 26 (54.2) | 21 (42.0) | 0.344 |
| Age | 53.0 [42.0,62.5] | 58.0 [49.2,64.5] | 58.5 [46.8,69.2] | 60.0 [53.0,68.0] | 0.057 |
| BMI | 30.4 [26.1,34.5] | 28.6 [26.1,30.9] | 29.6 [26.6,35.6] | 27.4 [25.4,31.1] | 0.18 |
| Never Smoker | 55 (66.3) | 7 (50.0) | 26 (54.2) | 36 (72.0) | 0.33 |
| Atopic | 33  (39.8) | 8  (57.1) | 15  (31.2) | 20 (40.0) | 0.363 |
| GORD | 51  (63.0) | 8  (57.1) | 31  (64.6) | 30 (60.0) | 0.942 |
| Nasal Polyps | 12 (14.8) | 2 (14.3) | 16 (33.3) | 18 (36.7) | 0.013 |
| Age of Onset | 11.0 [3.0,32.2] | 15.0 [3.0,40.0] | 32.0 [13.0,41.0] | 29.5 [12.2,49.8] | 0.02 |
| Exacerbations in the Last 12 Months | 1.0 [1.0,3.0] | 1.0 [0.0,2.0] | 1.5 [0.0,4.0] | 3.0 [1.0,6.0] | 0.021 |
| ICS (BDPe) | 3000.0 [2000.0,3500.0] | 3000.0 [2575.0,3000.0] | 3000.0 [2000.0,3000.0] | 3000.0 [2000.0,3900.0] | 0.97 |
| Maintenance OCS | 32 (38.6) | 3 (21.4) | 24 (50.0) | 20 (40.0) | 0.257 |
| Anti IgE | 9 (10.8) | 1 (7.1) | 5 (10.4) | 4 (8.0) | 0.936 |
| Anti IL-5 | 12 (14.5) | | 8 (16.7) | 4 (8.0) | 0.259 |
| ACQ6 | 2.3 [1.5,3.0] | 2.4 [1.5,3.1] | 2.4 [1.1,3.2] | 2.3 [1.3,3.5] | 0.991 |
| HADS | 11.0 [7.0,17.0] | 9.0 [7.0,10.0] | 10.0 [6.0,14.0] | 8.0 [5.5,15.5] | 0.368 |
| SNOT20 | 32.0 [18.0,45.0] | 23.5 [16.8,42.2] | 23.0 [14.0,42.0] | 33.0 [25.0,45.5] | 0.433 |
| FeNO | 19.0 [11.0,30.0] | 18.5 [12.8,27.8] | 32.0 [18.2,62.5] | 42.0 [27.2,71.0] | <0.001 |
| PostBD FEV1 (% predicted) | 84.9 [67.1,96.4] | 86.4 [74.0,96.5] | 70.5 [60.0,88.7] | 63.9 [48.9,81.2] | <0.001 |
| PostBD FVC (% predicted) | 93.1 [82.6,102.6] | 96.7 [86.4,109.7] | 86.9 [75.1,102.5] | 91.5 [72.8,99.5] | 0.238 |

| | Sputum Eos Low AND SpEDN Low (n= 83) | Sputum Eos Low BUT SpEDN High (n=14) | Sputum Eos High BUT SpEDN Low (n=48) | Sputum Eos High AND SpEDN High (n=50) | P-Value |
|---|---|---|---|---|---|
| PostBD FEV1/FVC | 73.0 [63.5,81.0] | 68.0 [66.0,75.8] | 68.5 [56.8,76.2] | 60.0 [49.5,67.5] | <0.001 |
| PostBD FEF25-75% (% predicted) | 62.1 [38.1,85.0] | 64.8 [36.6,77.5] | 42.3 [31.3,75.6] | 31.5 [21.2,43.2] | <0.001 |
| FEV1 Reversibility | 7.1 [2.9,12.6] | 3.9 [-1.8,14.2] | 15.3 [4.2,26.3] | 11.3 [4.1,17.4] | 0.043 |
| Blood Neutrophils | 5.2 [4.2,7.1] | 4.8 [3.9,6.2] | 5.5 [4.3,6.7] | 5.5 [4.2,6.4] | 0.845 |
| Blood Eosinophils | 0.1 [0.1,0.2] | 0.2 [0.1,0.2] | 0.3 [0.1,0.5] | 0.4 [0.3,0.6] | <0.001 |
| Serum Total IgE | 57.1 [15.8,249.6] | 132.7 [30.2,395.3] | 149.0 [37.3,492.0] | 137.0 [40.4,280.7] | 0.062 |
| Sputum Neutrophils | 45.9 [24.3,66.1] | 68.9 [30.7,84.3] | 45.3 [32.5,70.9] | 36.0 [21.6,55.8] | 0.087 |
| Sputum Eosinophils | 0.4 [0.1,0.8] | 0.8 [0.6,1.3] | 5.0 [3.3,11.9] | 18.1 [9.1,35.8] | <0.001 |

The two groups with high sputum eosinophil presence (i.e. low EDN and high EDN), of almost equal numbers, have the latest age of onset, highest frequency of nasal polyps, highest FeNO and worst post BD lung function. These features are worse in patients with high eosinophils and high EDN activity, in comparison to patients with high eosinophils and low EDN. This is most notable in terms of exacerbation frequency, 1.5 (0.0-4.0), in patients with high sputum eosinophils but low EDN, similar to that of the non-eosinophilic and low EDN patients, 1.0 (1.0-3.0), contrasting eosinophil high and EDN high patients who have an exacerbation frequency in the last 12 months of 3.0 (1.0 – 6.0)(Figure 4.18).

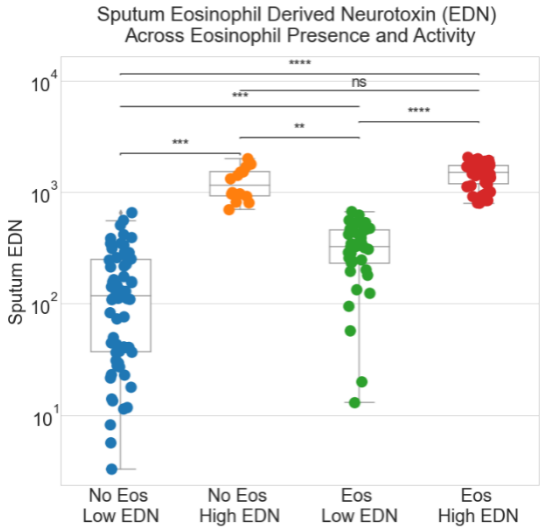


Figure 4.17 Sputum Eosinophil Derived Neurotoxin (EDN) Across Categories of Defined by Eosinophil Presence and Activity.

P values calculated from post-hoc pairwise comparisons corrected for multiple comparisons by Benjamini-Hochberg Procedure.

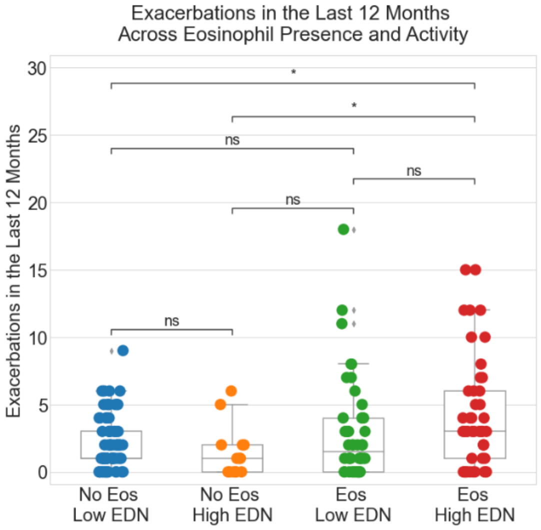


Figure 4.18 Exacerbation Frequency Across Categories of Defined by Eosinophil Presence and Activity.

P values calculated from post-hoc pairwise comparisons corrected for multiple comparisons by Benjamini-Hochberg Procedure.

#### 4.3.4.3 Defining Neutrophil Activation Phenotype

Although, the density plot did not immediately lend itself to dichotomisation, the one-dimensional clustering algorithm applied to EDN was applied to sputum MPO for the purposes of an exploratory analysis. Patients were divided at a break of 844, with 102 (52.31%) patients classified as low neutrophil activity and 93 (47.69%) as high neutrophil activity.

Patients with a high sputum MPO were more often male, older and had higher sputum neutrophils but there were otherwise no other distinguishing clinical characteristics (Table 4.9) There was a trend towards significance in blood neutrophilia being higher in patients with a high sputum MPO.

Table 4.9 Clinical Characteristics in Patients with Low and High Sputum Myeloperoxidase (MPO) Calculated by Jenks Natural Breaks

Continuous variables expressed as median [Q1, Q3] with differences measured by Kruskal Walis test. Categorical variables expressed as n (%) with differences measured by chi-square test. Abbreviations: GORD, gastro-oesophageal reflux disease; ICS, inhaled corticosteroid; BDPe, beclomethasone dose equivalent; OCS, oral corticosteroids; IgE, Immunoglobulin E; IL-5, Interleukin 5; ACQ, asthma control questionnaire, HADSTOT, Hospital Anxiety and Depression Total Score; SNOT, SinoNasal Outcome Score; FeNO, fraction of nitric oxide in exhaled breath; post BD, post bronchodilator; FEV1, forced expiratory volume in 1 second; FVC, forced vital capacity; FEF25-75%, forced expiratory flow at 25% to 75% of FVC.

| | **Low Sputum MPO (n=102)** | **High Sputum MPO (n=93)** | **P Value** |
|---|---|---|---|
| Female Sex | 65 (63.7) | 38 (40.9) | 0.002 |
| Age | 53.0 [42.0,62.0] | 61.0 [52.0,68.0] | <0.001 |
| BMI | 29.4 [26.0,33.5] | 29.1 [25.7,34.6] | 0.523 |
| Never Smoker | 60 (58.8) | 64 (68.8) | 0.127 |
| Atopic | 35 (34.3) | 41 (44.1) | 0.211 |
| GORD | 59 (59.0) | 61 (65.6) | 0.427 |
| Nasal Polyps | 27 (27.0) | 21 (22.8) | 0.617 |
| Age of Onset | 20.0 [5.0,41.0] | 16.0 [4.0,43.0] | 0.905 |
| Exacerbations in the Last 12 Months | 2.0 [1.0,4.0] | 2.0 [0.0,4.0] | 0.412 |
| ICS (BDPe) | 3000.0 [2000.0,3840.0] | 3000.0 [2000.0,3000.0] | 0.429 |

| | Low Sputum MPO (n=102) | High Sputum MPO (n=93) | P Value |
|---|---|---|---|
| Maintenance OCS | 41 (40.2) | 38 (40.9) | 0.959 |
| Anti IgE | 7 (6.9) | 12 (12.9) | 0.238 |
| Anti IL-5 | 16 (15.7) | 8 (8.6) | 0.199 |
| ACQ6 | 2.5 [1.5,3.2] | 2.3 [1.3,3.0] | 0.431 |
| HADS | 11.0 [6.0,16.0] | 10.0 [6.0,15.0] | 0.341 |
| SNOT20 | 30.0 [19.0,45.0] | 29.0 [15.5,45.0] | 0.640 |
| FeNO | 25.0 [15.0,50.8] | 26.0 [17.0,43.0] | 0.835 |
| PostBD FEV1 (% predicted) | 74.9 [62.7,92.7] | 77.4 [56.2,91.0] | 0.457 |
| PostBD FVC (% predicted) | 91.7 [81.7,100.7] | 91.4 [75.3,105.2] | 0.806 |
| PostBD FEV1/FVC | 69.0 [61.0,78.0] | 66.0 [56.0,77.0] | 0.174 |
| PostBD FEF25-75% (% predicted) | 45.4 [31.5,81.3] | 46.1 [28.1,69.7] | 0.394 |
| FEV1 Reversibility | 8.1 [3.2,17.6] | 9.6 [3.8,16.1] | 0.951 |
| Blood Neutrophils | 4.8 [3.9,6.6] | 5.6 [4.5,6.7] | 0.087 |
| Blood Eosinophils | 0.2 [0.1,0.4] | 0.2 [0.1,0.4] | 0.56 |
| Serum Total IgE | 89.7 [29.0,359.9] | 104.9 [21.2,280.7] | 0.921 |
| Sputum Neutrophils | 36.4 [19.0,53.5] | 54.5 [35.8,73.0] | <0.001 |
| Sputum Eosinophils | 2.1 [0.4,14.3] | 1.5 [0.5,7.8] | 0.706 |

The one-dimensional clustering algorithm divided patients into similarly sized groups, indicating that the cut (844) was close to the population median (1020.5). In order to force the dichotomisation of sputum MPO, patients were divided into equal sputum MPO tertiles with a view to comparing high and low sputum MPO patients (Figure 4.19).

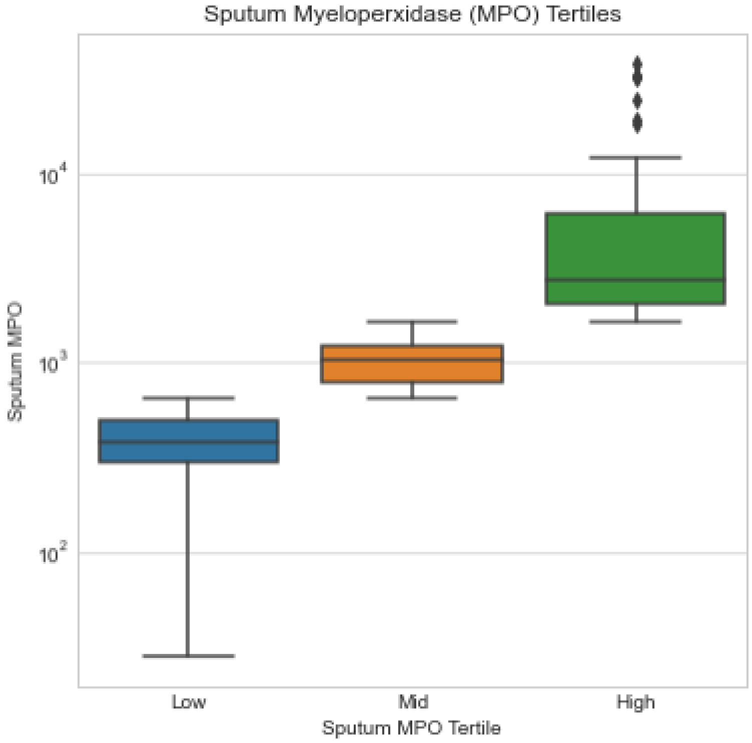


Figure 4.19 Boxplot of Sputum Myeloperoxidase in Patients Stratified into Equal Tertiles

Consistent with the one-dimensional clustering, high sputum MPO (Table 4.9) was associated with older males. They have a high proportion of never smokers and the least bronchodilator reversibility but little else to distinguish them from other patients (Table 4.10).

Table 4.10 Clinical Characteristics in Patients with Low, Mid and High Sputum Myeloperoxidase (MPO), calculated from MPO Tertiles

Continuous variables expressed as median [Q1, Q3] with differences measured by Kruskal Walis test. Categorical variables expressed as n (%) with differences measured by chi-square test. Abbreviations: GORD, gastro-oesophageal reflux disease; ICS, inhaled corticosteroid; BDPe, beclomethasone dose equivalent; OCS, oral corticosteroids; IgE, Immunoglobulin E; IL-5, Interleukin 5; ACQ, asthma control questionnaire, HADSTOT, Hospital Anxiety and Depression Total Score; SNOT, SinoNasal Outcome Score; FeNO, fraction of nitric oxide in exhaled breath; post BD, post bronchodilator; FEV1, forced expiratory volume in 1 second; FVC, forced vital capacity; FEF25-75%, forced expiratory flow at 25% to 75% of FVC.

| | **Lowest Sputum MPO Tertile (n=56)** | **Mid Sputum MPO Tertile (n=56)** | **Highest Sputum MPO Tertile (n=56)** | **P Value** |
|---|---|---|---|---|
| Female Sex | 43 (76.8) | 25 (44.6) | 22 (39.3) | <0.001 |

| | **Lowest Sputum MPO Tertile (n=56)** | **Mid Sputum MPO Tertile (n=56)** | **Highest Sputum MPO Tertile (n=56)** | **P Value** |
|---|---|---|---|---|
| Age | 50.5 [40.8,59.2] | 62.5 [52.8,69.2] | 60.0 [50.0,67.2] | 0.001 |
| BMI | 27.7 [24.8,32.7] | 29.9 [26.8,34.4] | 29.4 [25.7,34.8] | 0.279 |
| Never Smoker | 35 (62.5) | 32 (57.1) | 40 (71.4) | 0.021 |
| Atopic | 38 (67.9) | 32 (57.1) | 33 (58.9) | 0.459 |
| GORD | 28 (51.9) | 38 (67.9) | 38 (67.9) | 0.136 |
| Nasal Polyps | 16 (29.1) | 16 (28.6) | 11 (20.0) | 0.473 |
| Age of Onset | 18.5 [3.0,37.0] | 29.0 [5.0,50.5] | 15.5 [3.0,34.8] | 0.285 |
| Exacerbations in the Last 12 Months | 3.0 [1.0,5.0] | 2.0 [0.0,4.0] | 1.0 [0.0,3.0] | 0.143 |
| ICS (BDPe) | 3000.0 [2000.0,3840.0] | 3000.0 [2490.0,3605.0] | 2960.0 [2000.0,3000.0] | 0.101 |
| Maintenance OCS | 22 (39.3) | 22 (39.3) | 25 (44.6) | 0.801 |
| Anti IgE | 3 (5.4) | 9 (16.1) | 5 (8.9) | 0.160 |
| Anti IL-5 | 9 (16.1) | 5 (8.9) | 6 (10.7) | 0.478 |
| ACQ6 | 2.3 [1.3,3.2] | 2.2 [1.3,3.0] | 2.6 [1.5,3.2] | 0.891 |
| HADS | 11.0 [5.0,16.0] | 10.0 [7.0,16.0] | 9.0 [5.0,16.0] | 0.571 |
| SNOT20 | 32.0 [23.0,43.0] | 26.0 [22.0,45.0] | 29.0 [10.0,46.5] | 0.500 |
| FeNO | 25.0 [14.0,64.5] | 27.0 [17.5,51.0] | 24.5 [15.8,42.0] | 0.776 |
| PostBD FEV1 (% predicted) | 74.5 [63.5,90.5] | 74.3 [58.9,88.1] | 81.9 [55.9,91.8] | 0.824 |
| PostBD FVC (% predicted) | 92.2 [85.5,101.3] | 87.1 [74.5,102.5] | 93.1 [77.0,108.1] | 0.303 |
| PostBD FEV1/FVC | 66.0 [56.0,77.0] | 68.0 [61.8,78.0] | 66.0 [54.0,75.5] | 0.666 |
| PostBD FEF25-75% (% predicted) | 41.9 [29.0,72.0] | 41.4 [32.5,76.1] | 47.8 [26.3,73.3] | 0.873 |
| FEV1 Reversibility | 11.9 [4.7,18.9] | 11.7 [6.5,21.1] | 5.5 [1.1,9.9] | 0.004 |
| Blood Neutrophils | 4.8 [3.8,6.9] | 5.0 [4.4,6.1] | 5.9 [4.6,7.5] | 0.142 |
| Blood Eosinophils | 0.2 [0.1,0.4] | 0.3 [0.1,0.4] | 0.2 [0.1,0.3] | 0.166 |
| Serum Total IgE | 75.6 [22.7,304.8] | 148.9 [39.8,311.7] | 62.5 [17.5,244.6] | 0.337 |
| Sputum Neutrophils | 24.3 [14.2,42.2] | 41.1 [26.9,51.6] | 67.8 [46.4,78.8] | <0.001 |
| Sputum Eosinophils | 1.9 [0.3,26.3] | 5.1 [0.9,15.2] | 1.1 [0.4,5.6] | 0.013 |

## 4.4 Discussion

### 4.4.1 Serum Activity Relates to Airway Activity

Generally we find that serum and sputum EDN and ECP poorly correlate between compartments, contrasting the relationship between blood and sputum eosinophils, which likely represents the migration of eosinophils from bone marrow to the airways via the bloodstream [262]. Rather, we observe that ECP and EDN measures only correlate with eosinophil numbers in the same compartment (blood or sputum). This is consistent with the finding that the predictive power or blood EDN for predicting anti IL-5 response is largely due to its correlation with blood eosinophil counts [284] and that sputum eosinophil counts are not good predictors of Mepolizumab response[267]. This finding is in keeping with the premise that the ECP and EDN should be released in response external stimuli [275], i.e. at the site of its activity rather than in the bloodstream through transit.

Curiously, therefore, we find that paucigranular patients had the highest levels of serum ECP. This suggests that there is some form of eosinophil mediated systemic inflammation. This phenotype has a higher BMI (Table 3.4), which has numerous reported links to eosinophilia: blood eosinophilia has been associated with obesity and metabolic syndrome [285], waist-to-hip ratio with IL-5 [286] and, in a large population study, BMI was found to increase with ECP [282]. Serum ECP likely reflects this systemic inflammatory process rather than airway inflammation.

### 4.4.2 Sputum Eosinophil Measurement Remains Important to Asthma Phenotyping

The EDN dichotomisation identifies a cohort of patients with T2 high features, however this does not appear as clinically distinct as using sputum eosinophils. This may simply reflect a poorly defined cut-off. The one-dimension clustering identified a cluster close to the median value, which may not force separation well enough. Alternatively, the patients in this study were characterised whilst stable, outside of an exacerbation state. It is possible, under these conditions, the sputum eosinophils have yet to degranulate and release the proteins contained within. If so, cell numbers may in fact be more relevant than activation markers, which corroborates the wealth of data surround airway eosinophils and risk of exacerbations [42].

### 4.4.3 Conclusions

To summarise, there is heterogeneity of granulocyte activation in this severe asthma cohort but activation markers in serum are not related to activation markers in sputum. As such serum activation markers have only a limited role in phenotyping severe asthma. Examination of sputum activation markers identify two eosinophilic phenotypes, one of which is characterised by high granulocyte activity; it is likely that they represent distinct underlying mechanisms. The role of sputum neutrophils remains unclear and the T2 low phenotypes remain poorly defined.

# Chapter 5 Airway Microbiome in Severe Asthma

## 5.1 Introduction

In Chapter 3 and Chapter 4, it has proven difficult to define T2 low phenotypes using granulocyte cell counts or granulocyte activation markers. The association, however, between sputum neutrophilia and airway colonisation by potentially pathogenic bacteria [208] suggest that host-microbial interactions are a better way to understand these poorly-characterised asthma endotypes [209]. Over the past decade, a number of studies utilising culture-independent techniques (based on sequencing the variable regions of bacterial 16S ribosomal RNA genes) have described differences in the airway microbiome between asthma and healthy controls [209,287 ,288,289] and differences between inflammatory phenotypes in asthma patients [236,289,290].

Firmicutes, such as *Streptococcus* are associated with airway eosinophilia [289] whilst neutrophilic phenotypes have been associated with an increased abundance of proteobacteria, such as *Haemophilus*, *Moraxella* and *Pseudomonas* [208,236,291]. Increased Proteobacteria with parallel reduction in Bacteroidetes and Fusobacteria commensals are consistently associated with asthma[287,292]. This profile of airway colonisation is associated with increased risk of asthma development (if colonised in early life)[293,294], increased risk of asthma exacerbations [295] as well as reduced lung function and sputum neutrophilia[208,296].

Here, I describe the airway microbiome in the Airway Sampling in Severe Asthma Cohort. The objectives of this chapter are:

- Describe the airway microbiome in the a priori defined sputum inflammatory phenotypes
- Identify and describe a phenotype of patients characterised by excessive *Haemophilus* colonisation
- Assess existing biomarkers for the identification of patients who are colonised with *Haemophilus*

## 5.2 Chapter Specific Methods

### 5.2.1 Patient Population

The patients described in this chapter were recruited from the Airway Sampling in Severe Asthma WATCH Sub-Study.

### 5.2.2 DNA Extraction

DNA was extracted from whole sputum on the Qiacube DNA extraction machine using the DNeasy PowerSoil Pro Kit 250 in two batches. The V3-V4 region of the 16S rRNA gene was amplified from sputum DNA; the resulting DNA amplicon underwent paired end 300bp microbiome sequencing on the Illumina MiSeq platform (Illumina, San Diago, USA), amplifying the V3 and V4 region of the 16S rRNA gene. This work was done by collaborators at the University of Dundee who provided demultiplexed (split by sample) FASTQ Files/

### 5.2.3 Data Pre-Processing

The demultiplexed (split by sample) FASTQ Files were processed using the DADA2 pipeline [297] in R Programming language on the Southampton IRIDIS 5 High Performance Computing Facility. Sequences were clustered into amplicon sequence variants (ASV) table (described in 5.3.1); each Run was processed separately.

#### 5.2.3.1 Primer Removal

Primers were removed from each FASTQ file using scripts adapted from the DADA2 ITS Pipeline Workflow (1.8) (https://benjjneb.github.io/dada2/ITS_workflow.html) due to the variable length of primer sequences.

Table 5.1 Primer Identification in Run1 Before Removal

| | Forward | Complement | Reverse | ReverseComplement |
|---|---|---|---|---|
| **FWD.ForwardReads** | 116 | 0 | 0 | 0 |
| **FWD.ReverseReads** | 0 | 0 | 0 | 0 |
| **REV.ForwardReads** | 0 | 0 | 0 | 0 |
| **REV.ReverseReads** | 56 | 0 | 0 | 0 |

Table 5.2 Primer Identification in Run2 Before Removal

| | Forward | Complement | Reverse | ReverseComplement |
|---|---|---|---|---|
| **FWD.ForwardReads** | 0 | 0 | 0 | 0 |
| **FWD.ReverseReads** | 0 | 0 | 0 | 0 |
| **REV.ForwardReads** | 0 | 0 | 0 | 0 |
| **REV.ReverseReads** | 0 | 0 | 0 | 0 |

### 5.2.3.2 Filter and Trimming

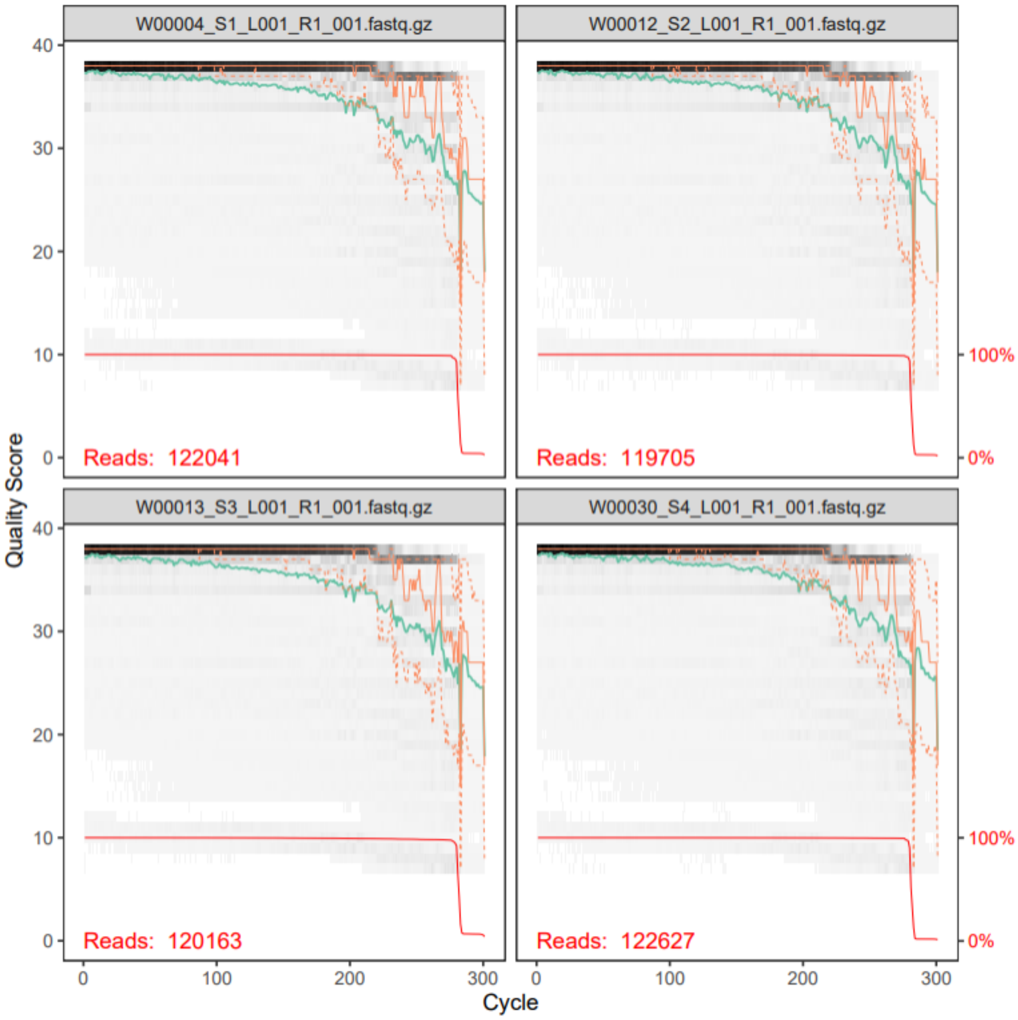


Figure 5.1 Quality Profile for Four Forward Reads in Run 1.The green line represents the mean quality score at each position. The orange lines represent the quartiles of the quality score distribution. The red line shows the scaled proportion of reads that extend to at least that position – flat in these plots as Illumina reads are typically all the same length)

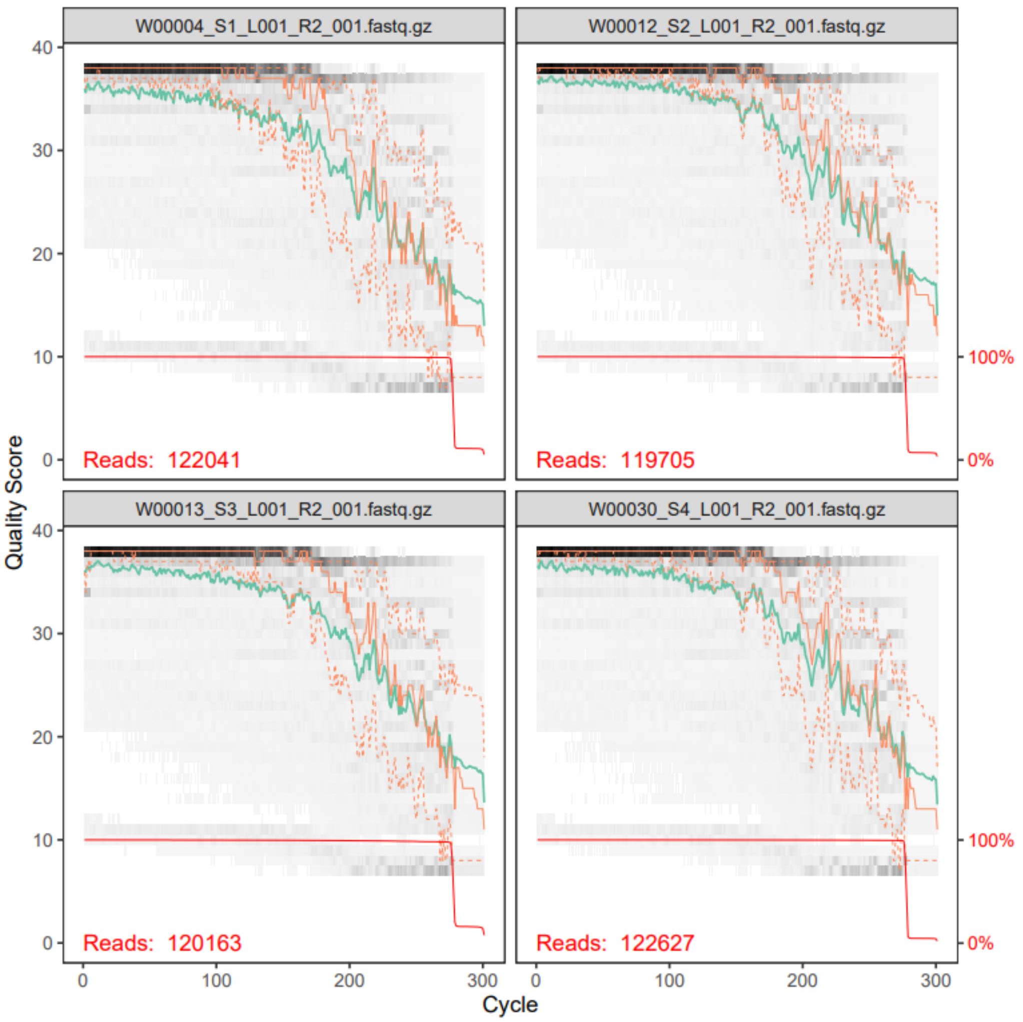


Figure 5.2 Quality Profile for Four Reverse Reads in Run 1.The green line represents the mean quality score at each position. The orange lines represent the quartiles of the quality score distribution. The red line shows the scaled proportion of reads that extend to at least that position – flat in these plots as Illumina reads are typically all the same length)

Samples were trimmed using the median read length of forward and reverse reads on the first FASTQ file in each run, as per the University of Dundee protocol. Samples were trimmed to 281, 277 bases in the forward and reverse reads respectively for both Run 1 and Run2. Visual inspection of Quality Plots generated for each forward and reverse FASTQ (first four samples illustrated) indicates that the Phred score (a measure of base quality i.e. the chance that the base has been correctly labelled) at these thresholds for the first four samples remains above 20. A Phred Score of 20 indicates the likelihood of finding 1 incorrect base call among 100 bases. In other words, the precision of the base call is 99%. Compared to forward reads, the reverse reads are of worse quality, especially at the end, which is common in Illumina sequencing.

### 5.2.3.3 Learning Error Rates

The DADA2 algorithm applies a parametric error model to learn the error rates of every amplicon dataset: This error model quantifies the rate at which an amplicon read is produced from a sample sequence as a function of sequence composition and quality [297]. Each plot (Figure 5.3) displays the error rates for transition from between each base. Each point represents the observed error rates for each consensus quality score with the black line representing the estimated error rates after convergence of the machine-learning algorithm. For each transition, the error frequency declines as the consensus quality score increases; moreover, the estimated error rates (black line) are a good fit for observed error rates (points). These observations give confidence in proceeding with pre-processing.

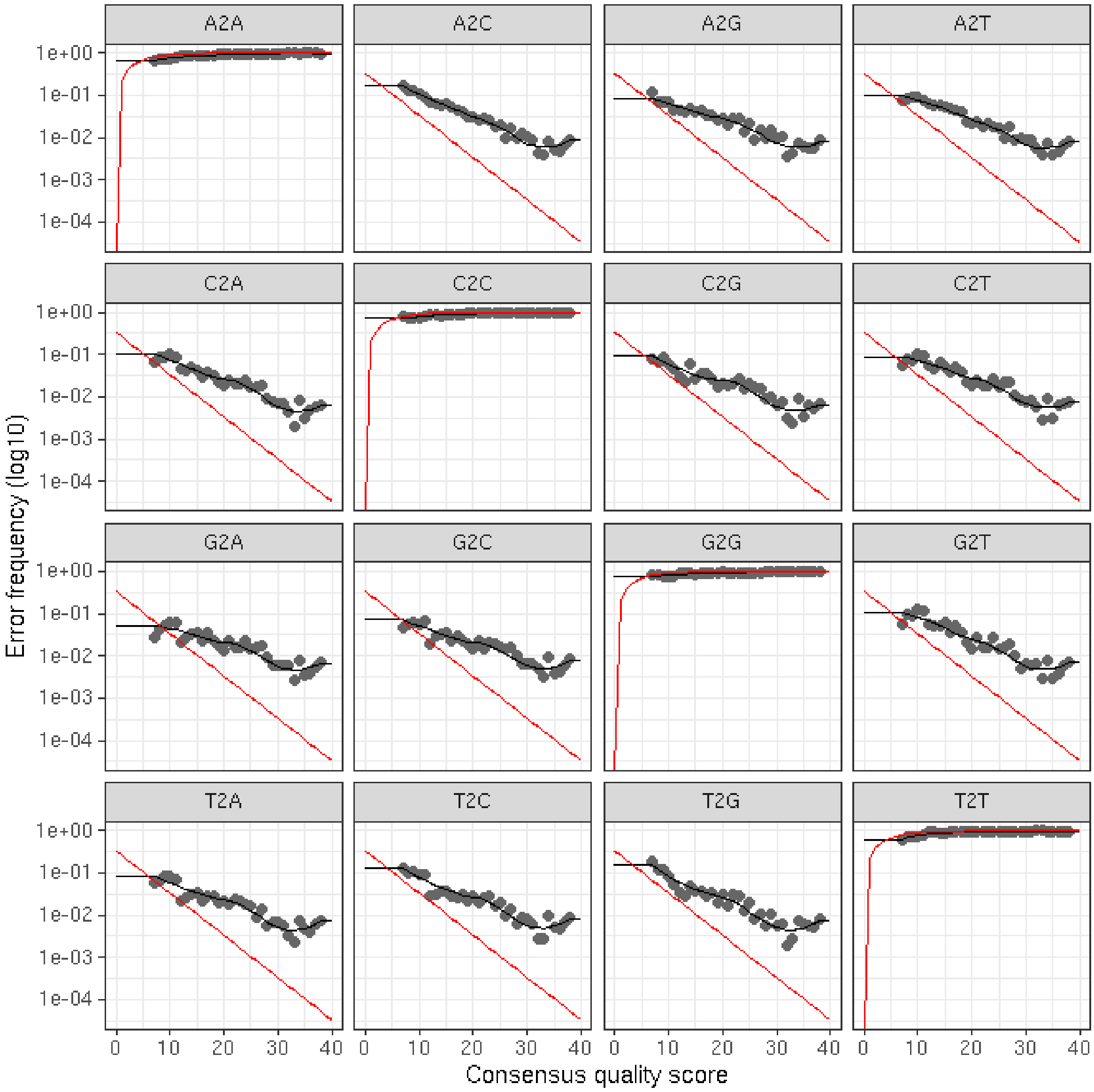


Figure 5.3 Error Rates for Each Possible Transition in the Forward Reads of Run 1 Each point represents the observed error rates for each consensus quality score. The black line

shows the estimated error rates after convergence of the machine-learning algorithm. The red line shows the error rates expected under the nominal definition of the Q-score.

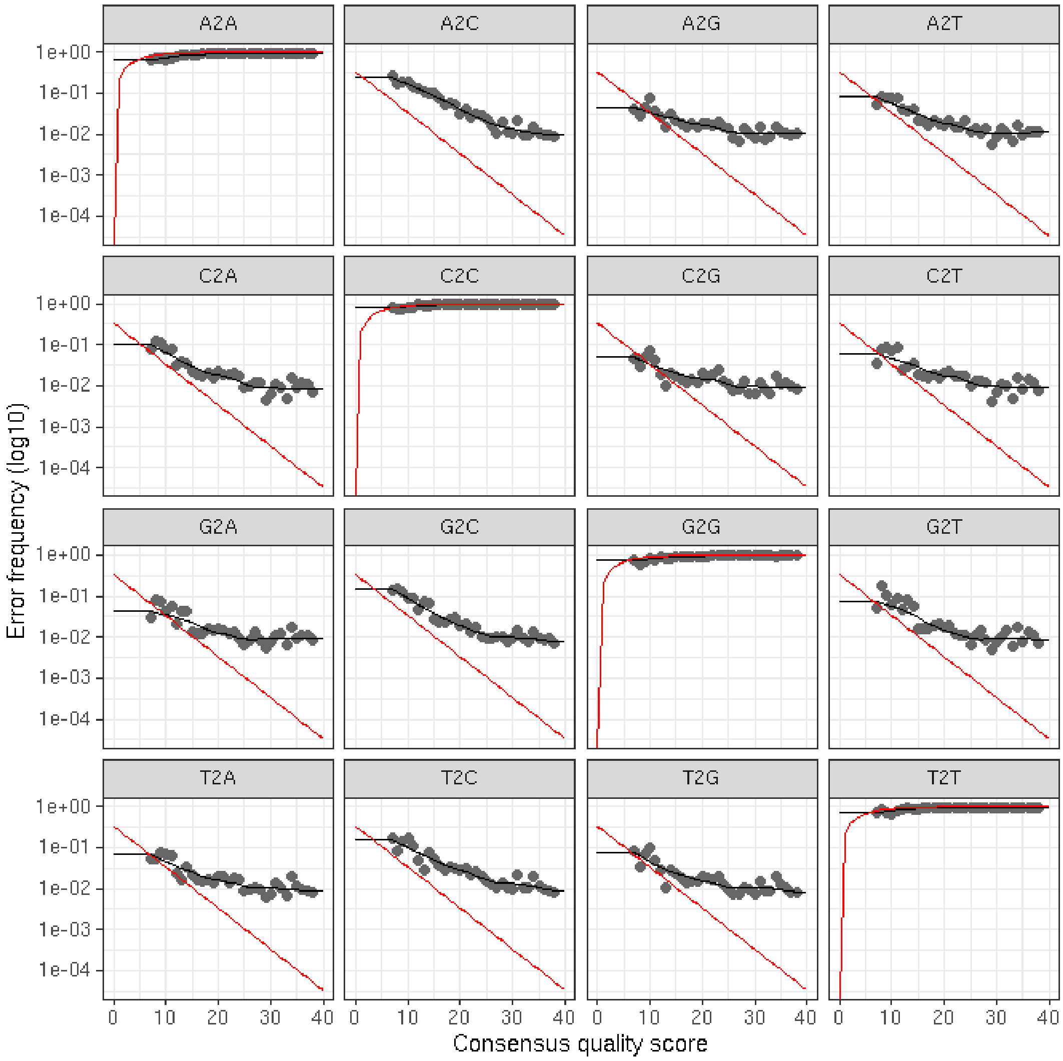


Figure 5.4 Error Rates for Each Possible Transition in the Reverse Reads of Run 1 Each point represents the observed error rates for each consensus quality score. The black line shows the estimated error rates after convergence of the machine-learning algorithm. The red line shows the error rates expected under the nominal definition of the Q-score.

### 5.2.3.4 Sample Inference and Taxonomic Classification

The sample inference algorithm central to the DADA2 toolkit was applied to each forward and reverse read, which were then merged. Chimera sequences (artifact sequences formed by two or more biological sequences incorrectly joined together) were removed. Taxonomic classifications were assigned to each amplicon sequence variant (ASV) from the SILVA_SSU_r138_2019 library using the IDtaxa algorithm [298]. The DECIPHER package [299] was used to maintain alignment quality across the multiple sequences and phangorn package used to construct a phylogenetic tree [300].

#### 5.2.3.5 Phyloseq Object Curation

A Phyloseq [301] object for each Run was constructed from the Amplicon Sequence Variant (ASV table), taxonomy table, clinical metadata and phylogenetic tree. The following contaminant identification step was performed in each Run-specific Phyloseq object independently of each other. Contaminants were identified using the decontam package [302] in R: 236 and 37 ASVs were identified as contaminants in Runs 1 and 2 respectively (Figure 5.5). A further 3 and 1 contaminants in Runs 1 and 2 respectively due to their presence in the negative controls (Figure 5.6).

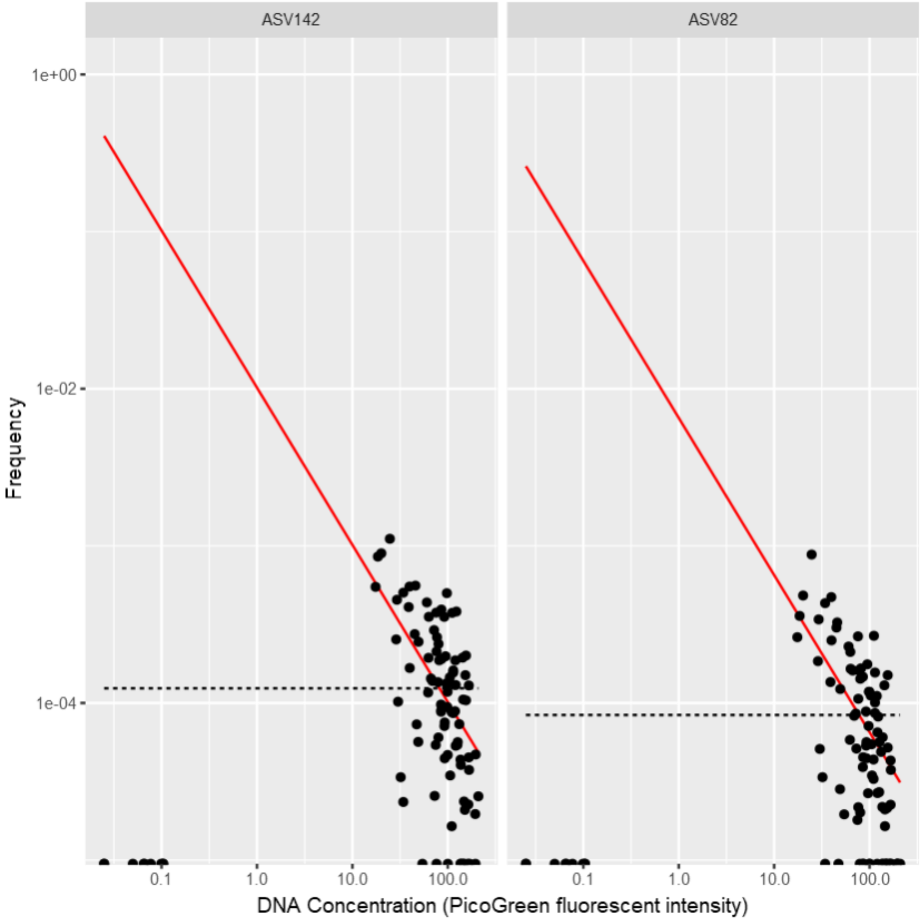


Figure 5.5 Model of Two Contaminant Sequence Features in Run 1 constructed using decontam. The dashed black line represents a model of a noncontaminant sequence feature for which frequency is expected to be independent of the input DNA concentration. The red line shows the model of a contaminant sequence feature, for which frequency is expected to be inversely proportional to input DNA concentration. In this diagram, ASV 142 and ASV82 fit the red line (contaminant model)

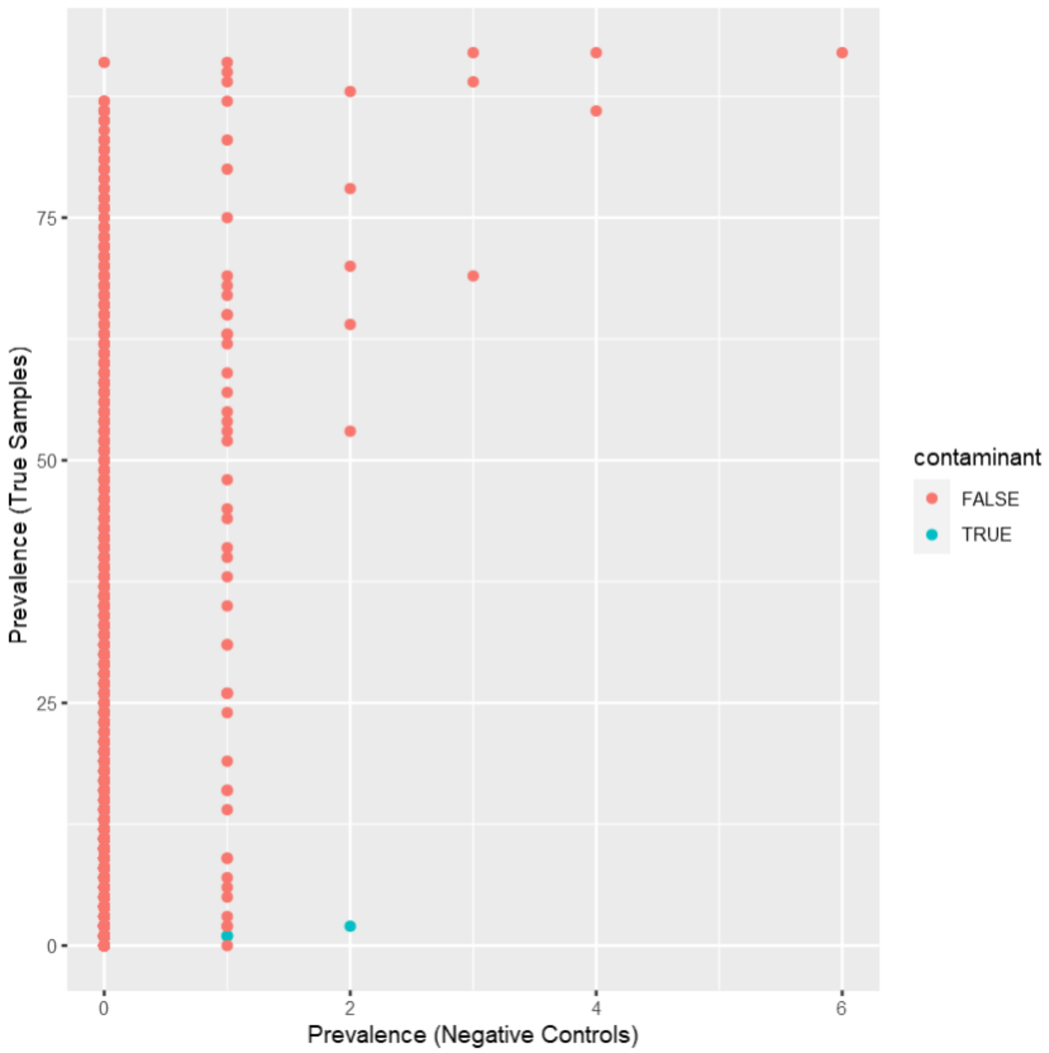


Figure 5.6 Prevalence of Contaminants in Samples and Controls Samples identified as Contaminants (blue) are more prevalent in Negative Controls than in True Samples in Run 1

Following decontamination, the Phyloseq objects were merged, retaining only overlapping taxa. This resulted in the exclusion of 270 taxa, which were determined to have a high risk of representing run specific contaminants.

Taxa were then removed if a phylum could not be assigned, as these taxa are likely to represent artefact sequences that do not exist in nature. Next, the prevalence of each taxa was calculated: this was defined as how many samples that taxon appeared in. The prevalence of each Phylum was calculated by summing the prevalence of each affiliated taxa. Phylum that were lowly prevalent in both runs were filtered and taxa not seen more than 3 times in at least 10% of samples were removed. This step was included to limit the FDR penalty paid when testing lower powered taxa seen in a small number of samples and better meet statistical model assumptions: retaining taxa with a small mean and trivially large coefficient of variation risk skewing downstream statistical analysis. Chloroflexi, Cyanobactera, Deinococcota and Desulfobacterota were lowly abundant in Run 1 and not present at all in Run 2. Taxa from these Phylum were therefore removed.

| Phylum | Sum Prevalence of Taxa Within Phylum (Run1) | Sum Prevalence of Taxa Within Phylum (Run 2) |
|---|---|---|
| **Actinobacteriota** | 5965 | 1115 |
| **Bacteroidota** | 2706 | 984 |
| **Campilobacterota** | 478 | 98 |
| **Chloroflexi** | 6 | 0 |
| **Cyanobacteria** | 8 | 0 |
| **Deinococcota** | 4 | 0 |
| **Desulfobacterota** | 31 | 0 |
| **Firmicutes** | 12867 | 2114 |
| **Fusobacteriota** | 2459 | 472 |
| **Patescibacteria** | 752 | 137 |
| **Proteobacteria** | 2793 | 639 |
| **Spirochaetota** | 581 | 67 |
| **Synergistota** | 201 | 11 |

#### 5.2.3.6 Batch Correction

Combat_seq batch correction was applied to the raw count matrix, using sputum eosinophilia as a proxy for case:control subsetting. Any negative values generated from this process were adjusted to 0: these counts represent “less than one” after rescaling and this adjustment prevents non-sensical values whilst having negligible impact on downstream statistical analysis. Batch correction was assessed by visual inspection of a Non-metric Multi-dimensional Scaling (NMDS) plot using a Bray-Curtis Distance measure of samples.

### 5.2.4 Rarefaction and Relative Abundance

Depending on the downstream analysis, four phyloseq objects were used. The first using absolute abundances of taxa; the second using relative abundances (taxa expressed as a percentage of all taxa in a sample) and the third log transformed absolute abundances.

A common concern in ecological analyses, such as described herein, is that species richness increases with sample size. In order to remove this effect, rarefaction can be applied, but has its limitations [303], in which all samples are limited to the same minimum sample size. This cutoff was determined by plotting rarefaction curves and used to generate a rarefied object.

### 5.2.5 Statistical Analysis

#### 5.2.5.1 Ordination

The microbial data was ordinated using Multidimensional scaling (MDS) on Bray Curtis dissimilarity distances (quantification of the compositional dissimilarity between two samples) of rarefied data. The Bray Curtis distance is preferred in microbial analysis due to its relative tolerance of sparse data. Clustering algorithms were applied to this ordinated data as described in 4.2.

#### 5.2.5.2 Permutational Multivariate Analysis of Variance

Permutational Multivariate Analysis of Variance was used to investigate variance across partitions in distance matrices such as across inflammatory phenotypes. This was conducted using the vegan package in R. Pairwise calculations were conducted using the pairwise_adonis package.

#### 5.2.5.3 Unsupervised Clustering

Clustering was performed on the first principal components to capture 95% of variance in the original data by Partitioning Around Medoids (PAM) and Ward's Hierarchichal clustering on bray Curtis distances. The optimum number of clusters was identified using a variety of indices: average silhouette width, gap statistic and total within-cluster sum of square. Cluster assignment was internally validated by assessing cluster assignment agreement with the Rand Index.

#### 5.2.5.4 Differential Expression

Differential abundance analysis was performed using DESeq2. Sex, ICS and mOCS were included in the model when comparing between clusters. Clusters were compared in a pairwise manner and visualised using a volcano plot using the EnhancedVolcano package.

#### 5.2.5.5 BLAST

Nucleotide sequences from the ASVs of interest were submitted to the BLAST website [304] and referenced against the "rRNA_typestrains/16S_ribosomal_RNA" database. Though taxonomic assignment on short amplicon reads such as by BLAST allows the opportunity for species identification, it has a very high false positive rate.

## 5.3 Results

### 5.3.1 Patient Population

The patients described in this chapter were recruited from the Airway Sampling in Severe Asthma WATCH Sub-Study. Sputum 16S sequencing was available in 119 patients. Patients with sputum microbiome data were not demographically different to those without , however, they did have significantly worse disease, as represented by worse spirometry and sputum eosinophilia; they were also statistically less frequently atopic.

Table 5.3 Clinical Characteristics of Patients with and without 16S Sequencing Data

Continuous variables expressed as median [Q1, Q3] with differences measured by Mann-Whitney U test. Categorical variables expressed as n (%) with differences measured by chi-square test. Abbreviations: ICS, inhaled corticosteroid; BDPe, beclomethasone dose equivalent; FeNO, fraction of nitric oxide in exhaled breath; post BD, post bronchodilator; FEV1, forced expiratory volume in 1 second; FVC, forced vital capacity; FEF25-75%, forced expiratory flow at 25% to 75% of FVC; ACQ, asthma control questionnaire, HADSTOT, Hospital Anxiety and Depression Total Score; SNOT, SinoNasal Outcome Score

| | **16S available (n = 119)** | **16S not available (n = 76)** | **P-Value** |
|---|---|---|---|
| Sex (% Female) | 61 (51.3) | 42 (55.3) | 0.690 |
| Age | 55.0 [44.0,65.0] | 58.0 [47.0,69.0] | 0.258 |
| BMI | 29.4 [25.6,34.8] | 29.1 [25.9,33.4] | 0.983 |
| Smoker (% Never) | 74 (62.2) | 50 (65.8) | 0.447 |
| Atopy | 65 (54.6) | 54 (71.1) | 0.032 |
| Age of Onset | 15.0 [3.0,35.5] | 23.0 [7.0,48.5] | 0.165 |
| Exacerbations in Last 12 months | 2.0 [1.0,4.0] | 1.0 [0.0,3.0] | 0.168 |
| ICS (BDPe) | 3000.0 [2000.0,3500.0] | 3000.0 [2000.0,3000.0] | 0.688 |
| FeNO | 27.0 [17.0,48.0] | 23.0 [14.8,45.0] | 0.391 |
| Blood Eosinophil Count | 0.2 [0.1,0.5] | 0.2 [0.1,0.3] | 0.034 |
| Sputum Eosinophil % | 2.6 [0.5,15.6] | 1.5 [0.2,6.8] | 0.067 |
| Sputum Neutrophil % | 40.5 [23.1,66.8] | 49.8 [34.3,68.0] | 0.141 |
| PostBD FEV1 | 73.3 [54.9,88.0] | 80.0 [65.2,97.5] | 0.007 |
| PostBD FEV1/FVC | 65.0 [54.0,76.5] | 70.0 [64.0,80.0] | 0.003 |

| | 16S available (n = 119) | 16S not available (n = 76) | P-Value |
|---|---|---|---|
| PostBD FEF25-75 %predicted | 39.8 [27.7,69.4] | 54.9 [38.2,89.6] | 0.001 |
| ACQ6 | 2.5 [1.5,3.3] | 2.2 [1.1,3.0] | 0.198 |
| HADSTOT | 10.0 [6.0,16.0] | 10.0 [6.0,15.0] | 0.674 |
| SNOT20 | 32.0 [22.5,46.5] | 27.0 [15.0,45.0] | 0.273 |

### 5.3.2 Data Pre-Processing Results

The average number of raw reads per sample from both runs was >100,000. The majority of reads were lost at the filtering step (Table 5.4 and Table 5.5) and on merging. However, there were, on average, more than 40,000 reads per sample in both Runs.

Table 5.4 Mean Read Counts at Each Pre-processing Step in Run 1

| | Input | Filtering | Denoising Forward | Denoising Reverse | Merging | Chimera Removal |
|---|---|---|---|---|---|---|
| Number of Reads | 127178.19 | 60318.09 | 57236.67 | 54923.46 | 47340.33 | 44713.09 |
| Percentage of Reads Lost | | 52.57 | 5.11 | 8.94 | 21.52 | 5.55 |

Table 5.5 Mean Read Counts at Each Pre-processing Step in Run 2

| | Input | Filtering | Denoising Forward | Denoising Reverse | Merging | Chimera Removal |
|---|---|---|---|---|---|---|
| Number of Reads | 135108.2 | 61223.33 | 56034.13 | 52188.73 | 42683.97 | 41414.33 |
| Percentage of Reads Lost | | 54.69 | 8.48 | 14.76 | 30.28 | 2.97 |

The sequencing depths for samples ranged from 19758 to 59912 in Run 1 and 28,749 to 45110 in Run2. Histograms of sequencing depths from each run (Figure 5.8 and Figure 5.9) demonstrated a normal distribution with no evidence of samples with excessively poor read depth that would require exclusion.

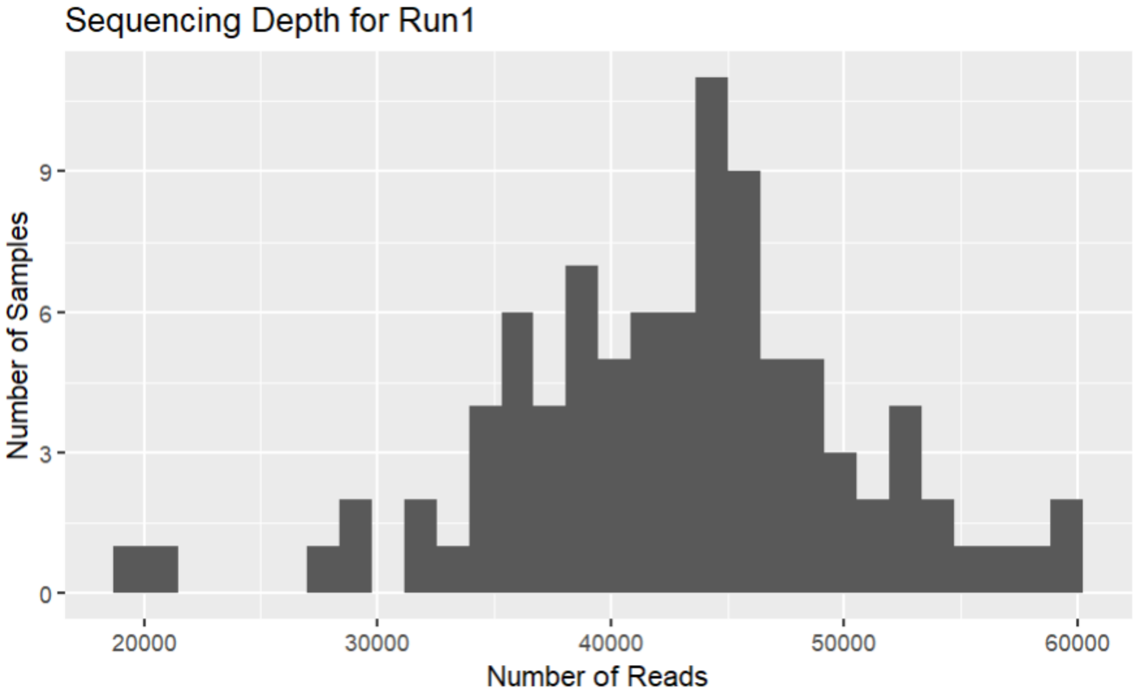


Figure 5.7 Histogram of Sequencing Depth for Samples in Run 1

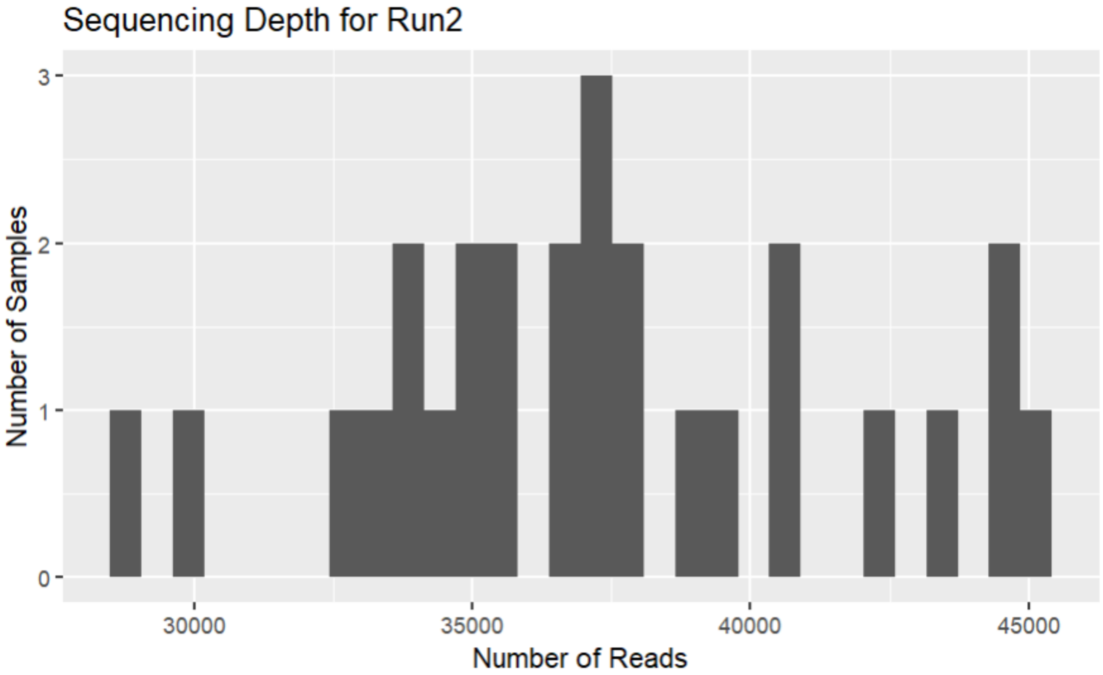


Figure 5.8 Histogram of Sequencing Depth for Samples in Run 2

#### 5.3.2.1 Batch Effect

The clinical characteristics of patients in Run 1 and Run 2 were broadly similar other than Run2 having fewer patients with co-morbid GORD and fewer patients treated with maintenance OCS. There were also trends towards lower BMI and more preserved lung function (Table 5.5).

Nevertheless, there was a similar distribution of inflammatory phenotypes across the two runs: 41.30% and 40.74% eosinophilic in Run 1 and 2 respectively (Table 5.5).

Table 5.6 Clinical Characteristics of patients with 16s rRNA samples in Run1 and Run2 Continuous variables expressed as median [Q1, Q3] with differences measured by Kruskal Walis test. Categorical variables expressed as n (%) with differences measured by chi-square test. Abbreviations: GORD, gastro-oesophageal reflux disease; ICS, inhaled corticosteroid; BDPe, beclomethasone dose equivalent; OCS, oral corticosteroids; IgE, Immunoglobulin E; IL-5, Interleukin 5; ACQ, asthma control questionnaire, HADSTOT, Hospital Anxiety and Depression Total Score; SNOT, SinoNasal Outcome Score; FeNO, fraction of nitric oxide in exhaled breath; post BD, post bronchodilator; FEV1, forced expiratory volume in 1 second; FVC, forced vital capacity; FEF25-75%, forced expiratory flow at 25% to 75% of FVC

| | **Run 1**<br>**(n = 92)** | **Run 2**<br>**(n=27)** | **P Value** |
|---|---|---|---|
| Female Sex | 49 (53.3) | 12 (44.4) | 0.557 |
| Age | 58.0 [48.0,65.0] | 50.0 [39.0,65.5] | 0.259 |
| BMI | 30.2 [26.0,35.5] | 27.0 [25.6,30.3] | 0.075 |
| Never Smoker | 59 (64.1) | 15 (55.6) | 0.716 |
| Atopic | 49 (53.3) | 16 (59.3) | 0.741 |
| GORD | 29 (31.5) | 18 (66.7) | 0.002 |
| Nasal Polyps | 18 (19.6) | 5 (18.5) | 0.89 |
| Age of Onset | 12.5 [0.0,33.0] | 6.0 [0.0,29.0] | 0.513 |
| Exacerbations in the Last 12 Months | 2.0 [0.8,4.2] | 1.0 [0.0,3.0] | 0.18 |
| ICS (BDPe) | 3000.0 [2000.0,3275.0] | 3000.0 [2000.0,3250.0] | 0.473 |
| Maintenance OCS | 38 (41.3) 12 | 4 (14.8) | 0.021 |
| Anti IgE | (13.0) 12 | 2 (7.4) | 0.734 |
| Anti IL-5 | (13.0) 2.5 | 0 | 0.066 |
| ACQ6 | [1.5,3.2] | 2.7 [1.2,3.2] | 0.965 |
| HADS | 10.0 [5.0,16.0] | 9.0 [5.5,15.5] | 0.765 |
| SNOT20 | 23.0 [0.0,39.2] | 24.0 [3.5,40.5] | 0.794 |
| FeNO | 26.5 [17.0,48.2] | 27.0 [14.0,42.0] | 0.587 |
| PostBD FEV1 (% predicted) | 67.8 [53.7,86.1] | 82.9 [70.9,92.8] | 0.028 |
| PostBD FVC (% predicted) | 88.6 [76.9,99.8] | 94.2 [88.7,106.4] | 0.071 |

| | **Run 1**<br>**(n = 92)** | **Run 2**<br>**(n=27)** | **P Value** |
|---|---|---|---|
| PostBD FEV1/FVC | 65.0 [52.0,75.2] | 66.0 [60.0,79.0] | 0.13 |
| PostBD FEF25-75% (% predicted) | 38.2 [26.0,67.3] | 46.5 [32.9,76.1] | 0.09 |
| FEV1 Reversibility | 1.1 [0.0,9.2] | 0.0 [0.0,11.9] | 0.72 |
| Sputum Eosinophils >2% | 49 (53.3) | 14 (51.9) | 1 |
| Sputum Neutrophils >61% | 28 (30.4) | 8 (29.6) | 1 |
| Seurm IgE | 77.5 [26.2,283.0] | 69.0 [18.2,638.5] | 0.739 |
| Serum MPO | 232.1 [169.7,309.2] | 259.7 [167.9,310.1] | 0.859 |
| Serum Elastase | 38.4 [26.4,66.3] | 47.0 [30.0,72.7] | 0.493 |
| Serum EDN | 41.4 [30.1,63.7] | 44.9 [35.7,58.0] | 0.646 |
| Serum ECP | 31.0 [12.0,62.0] | 20.2 [13.8,54.2] | 0.775 |
| Sputum Neutrophils % | 40.4 [22.1,66.6] | 40.5 [24.8,66.6] | 0.487 |
| Sputum Eosinophils % | 2.7 [0.6,15.1] | 2.2 [0.5,14.0] | 0.666 |
| Sputum MPO | 1106.0 [533.6,2093.2] | 873.7 [409.4,3358.0] | 0.889 |
| Sputum Elastase] | 5.9 [3.1,9.5] | 5.6 [3.2,14.1] | 0.894 |
| Sputum EDN | 366.2 [154.0,1468.8] | 796.9 [126.1,1381.0] | 0.889 |
| Sputum ECP | 517.8 [126.4,1715.8] | 844.3 [75.0,1447.5] | 0.944 |

Following batch correction, no separation was identified between the runs on visual inspection of NMDS plot using Bray-Curtis Distance (Figure 5.10).

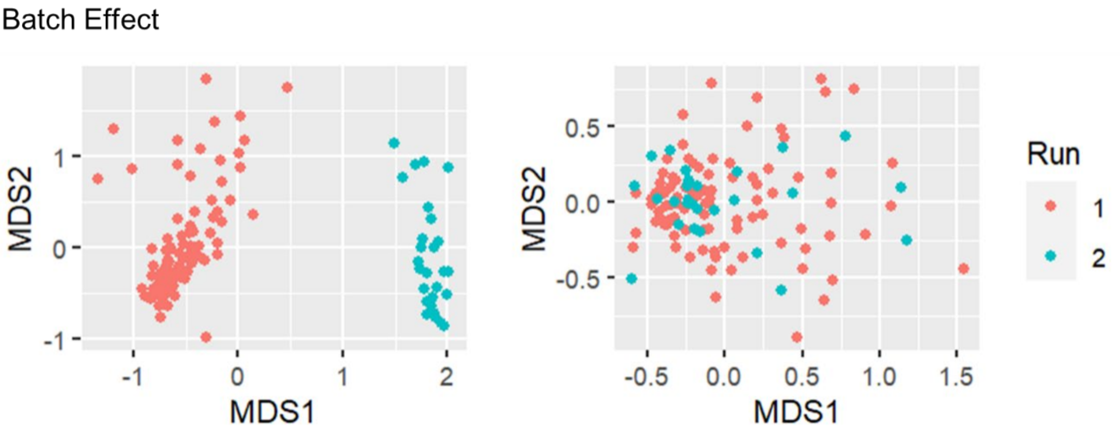


Figure 5.9 Batch Effect Before (left) and After (right) Filtering and CombatSeq

### 5.3.2.2 Rarefaction

On plotting the rarefaction curves (Figure 5.10), plateauing of the number of new species identified occurs at around 10,000, indicating good representation of the microbial community at this size.

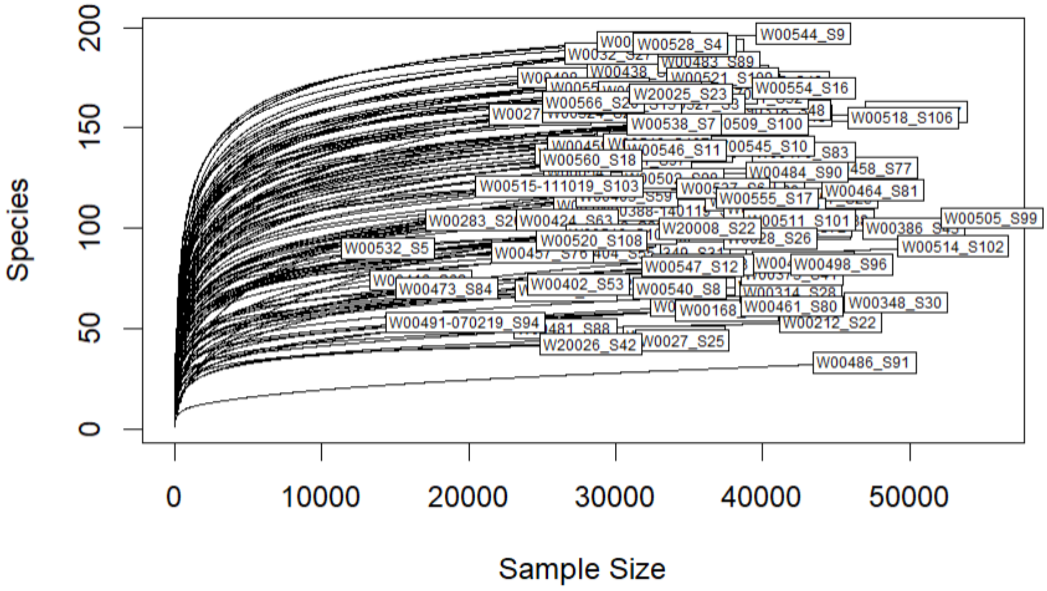


Figure 5.10 Rarefaction Curves for Samples

### 5.3.2.3 Summary of Pre-Processing Events

In summary, an average of 40,000 reads per sample were used to construct an amplicon sequence variant (ASV) table containing 5291 unique taxa. 5059 taxa were removed (Figure 5.11), leaving 232 taxa in 119 samples on which downstream microbial analysis was performed.

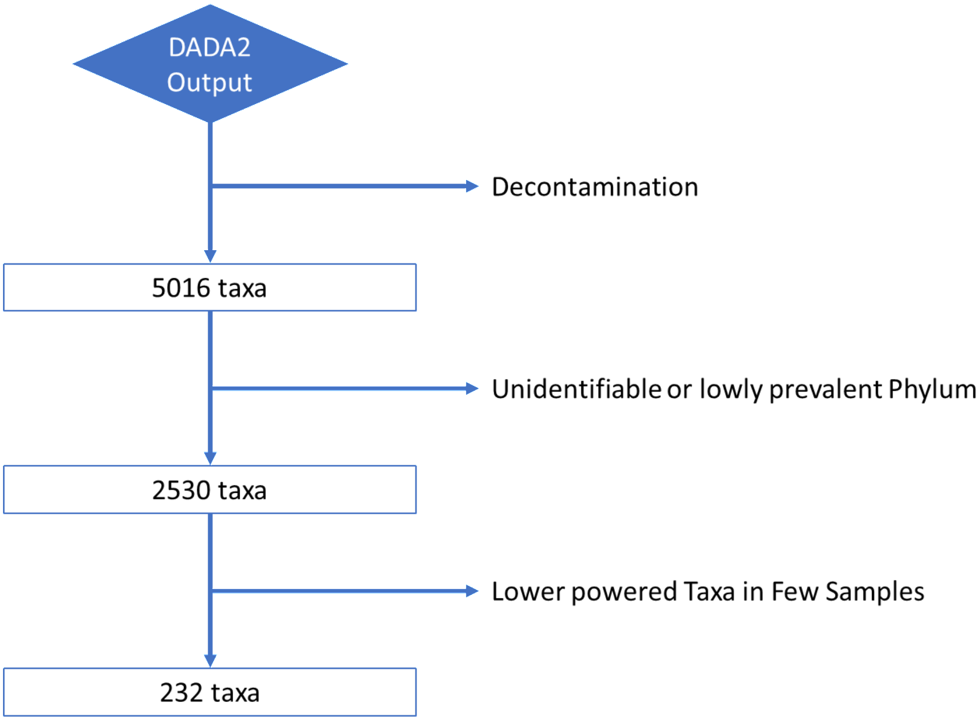

DADA2 Output
Decontamination
5016 taxa
Unidentifiable or lowly prevalent Phylum
2530 taxa
Lower powered Taxa in Few Samples
232 taxa

Figure 5.11 Summary of Taxa Filtering

### 5.3.3 Microbial Description of Cohort

In assessing the relative abundances of ASVs, the most abundant Phylum in the cohort was Firmicutes (Figure 5.12) and most abundant Genus identified in the sputum of this cohort was *Streptococcus*, followed by *Rothia* and *Haemophilus* (Table 5.6). Almost a quarter of bacteria identified belonged to a Genus that had a relative abundance of less that 1%.

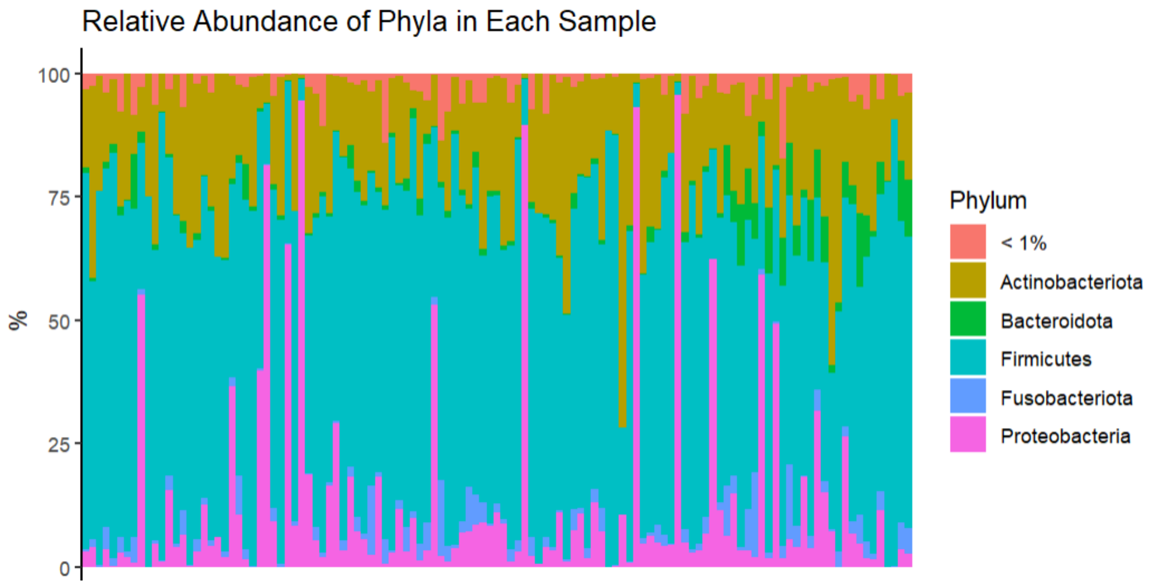


Figure 5.12 Relative Abundance of Phylum per Sample. Each column represents a sputum sample. The distribution of colours in each column reflects the relative abundance of different Phyla in that sample.

Table 5.7 Relative Abundance of Genus in Overall Cohort

| **Genus** | **Relative Abundance (%)** |
|---|---|
| Streptococcus | 36.99 |
| Rothia | 12.62 |
| Haemophilus | 11.81 |
| Veillonella | 8.84 |
| Actinomyces | 4.29 |
| Gemella | 2.88 |
| Other | 22.58 |

### 5.3.4 Microbial Differences Between Inflammatory Phenotypes

Exploratory analysis of rarefied microbial data shows no differences in alpha diversity metrics between the inflammatory phenotypes (Figure 5.14). Accordingly, an NMDS plot using Bray Curtis distances on rarefied data shows no separation between inflammatory phenotypes (Figure 5.13). Permutational Analysis of variance on bray Curtis distance indicates that the inflammatory phenotypes are statistically distinct (Table 5.6), however, on pairwise comparisons, no single phenotype emerges as distinctive (Table 5.7).

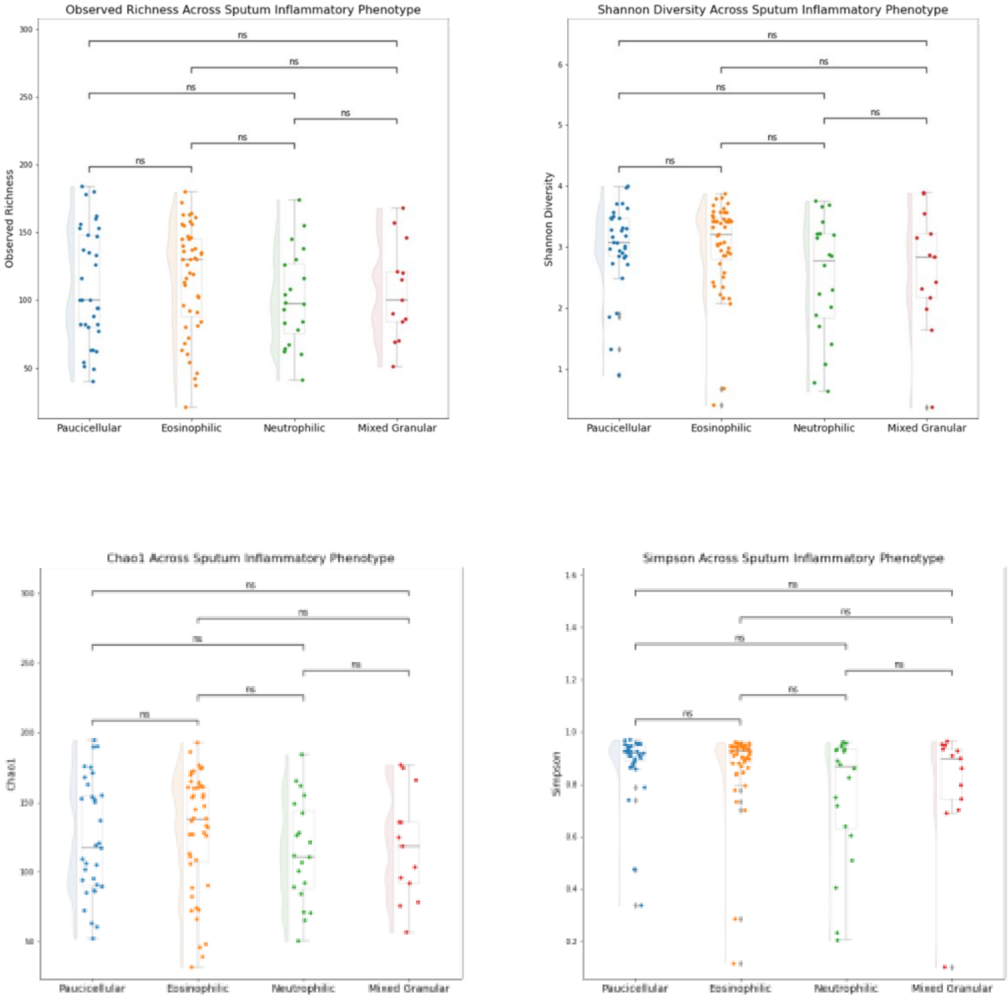


Figure 5.13 Alpha Diversity Measures Between Sputum Inflammatory Phenotypes

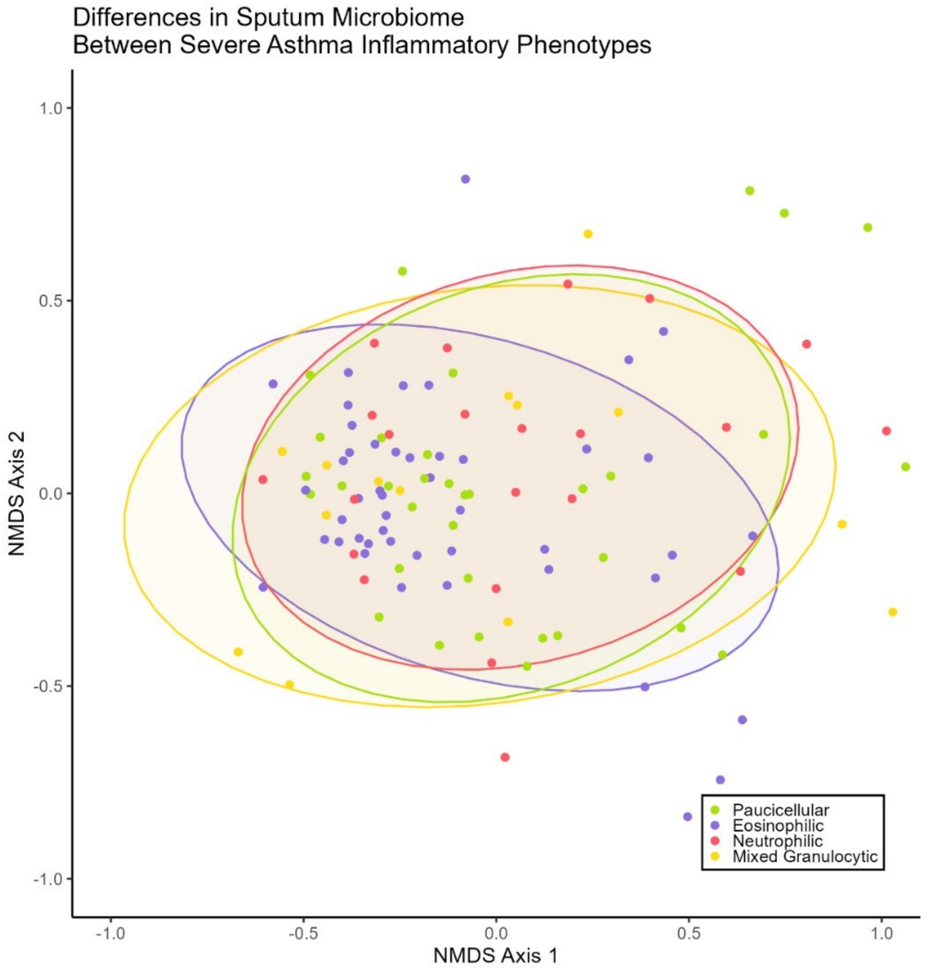


Figure 5.14 NMDS Plot of Differences in Sputum Microbiome Between Inflammatory Phenotypes using Bray Curtis distance.

Table 5.8 Permutational Analysis of variance using distance matrices

| | Df | SumsOfSqs | MeanSqs | F.Model | R2 | Pr(>F) |
|---|---|---|---|---|---|---|
| Sputum Inflammatory Phenotype | 3 | 1.182 | 0.394 | 1.567 | 0.039 | 0.007 |
| Residuals | 115 | 28.912 | 0.251 | NA | 0.960 | NA |
| Total | 118 | 30.094 | NA | NA | 1 | NA |

Table 5.9 Permutational Analysis Pairwise Comparisons

| Pairwise Comparison | corrected p value |
|---|---|
| Eosinophilic vs Mixed | 0.066 |
| Eosinophilic vs Neutrophilic | 0.018 |
| Eosinophilic vs Pauciceulluar | 0.662 |
| Neutrophilic vs Mixed | 0.127 |
| Neutrophilic vs Pauciceulluar | 0.152 |
| Mixed vs Pauciceulluar | 0.012 |

As *Haemophilus* was of particular interest, direct correlations between this genus and airway inflammatory cells were assessed: there was no statistically significant linear correlation between abundance of *Haemophilus* and sputum eosinophils or neutrophils (Figure 5.15).

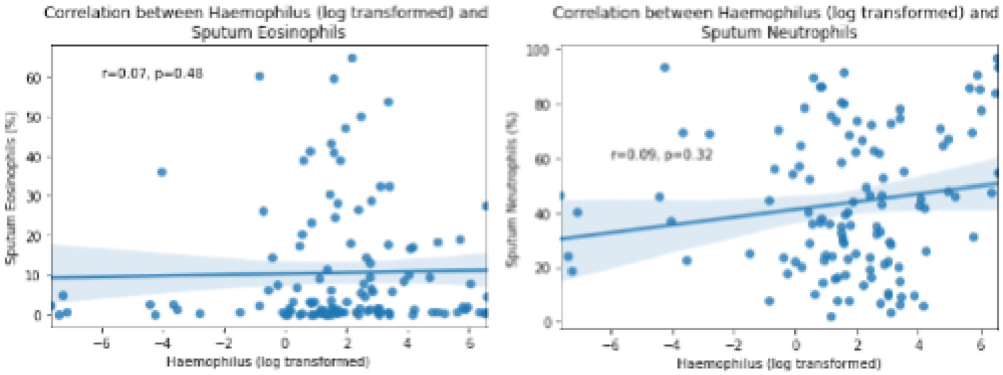


Figure 5.15 Spearman Rank Correlations Between *Haemophilus* and Sputum Inflammatory Cells

### 5.3.5 Clustering Patients on Airway Microbiome

Analyses in this chapter have thus far related bacterial abundance to clinically defined parameters. Clustering was employed to identify microbial data driven arrangements against which clinical parameters could be described.

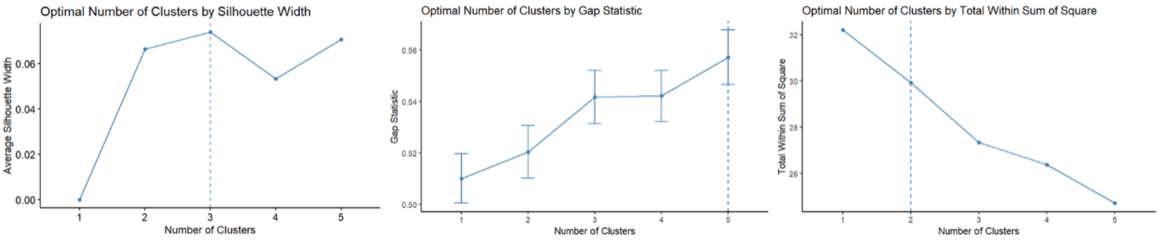


Figure 5.16 Optimum Number of Clusters in Microbiome Data Using Silhouette Width, Gap Statistic and Total Within Sum of Squares

Using the silhouette width, the optimum number of clusters in the microbiome data was determined to be three. This was corroborated by the Gap Statistic and Total Width Sum of Squares which indicated the optimal number to be between 2 and 5.

Hierarchichal clustering and Partition Around the Medioids were employed, as previously described, the rand score of 0.8 suggested strong cluster consensus between Hierarchical and Partition Around the Medioids. The Calinski Harabaz Index for clusters density was 10.34 and 10.21 respectively, indicating that Hierarchical clustering produces marginally more coherent clusters. Further analysis was therefore focussed on 3 clusters produced by Hierarchical clustering.

#### 5.3.5.1 Hierarchical Clusters

Alpha diversity was compared between these three clusters using rarefied abundances. Cluster 3 appears distinct to the other two clusters with low alpha diversity (Figure 5.17). This cluster is dominated by the Proteobacteria Phylum (Figure 5.18) and *Haemophilus* Genus (Figure 5.19). Cluster 1 and 2 are subtly different with reduced diversity in Cluster 2 compared to 1, though the relative abundance barplot does not indicate any major differences between the clusters (Figure 5.19).

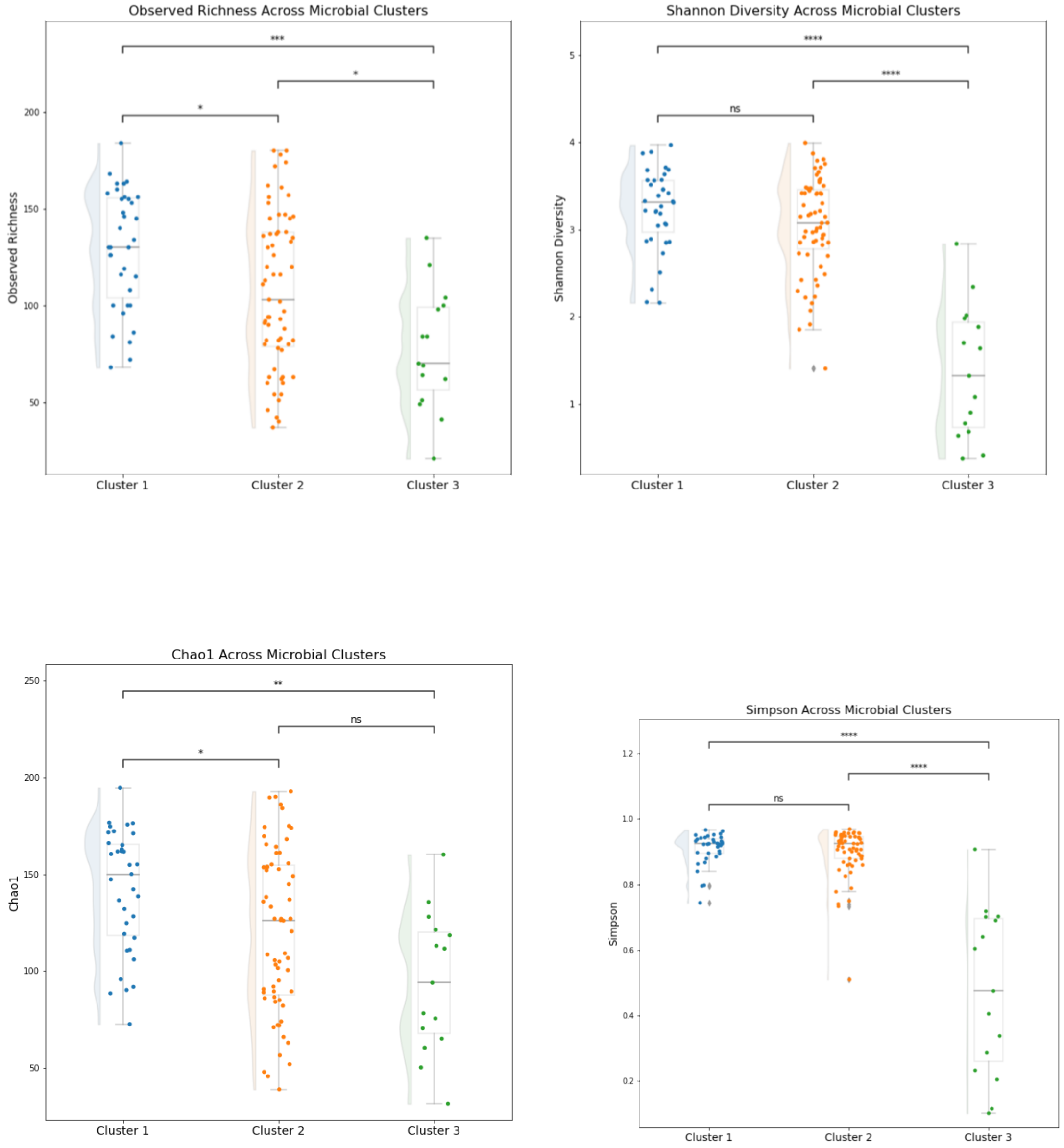


Figure 5.17 Alpha Diversity Measures Between Clusters

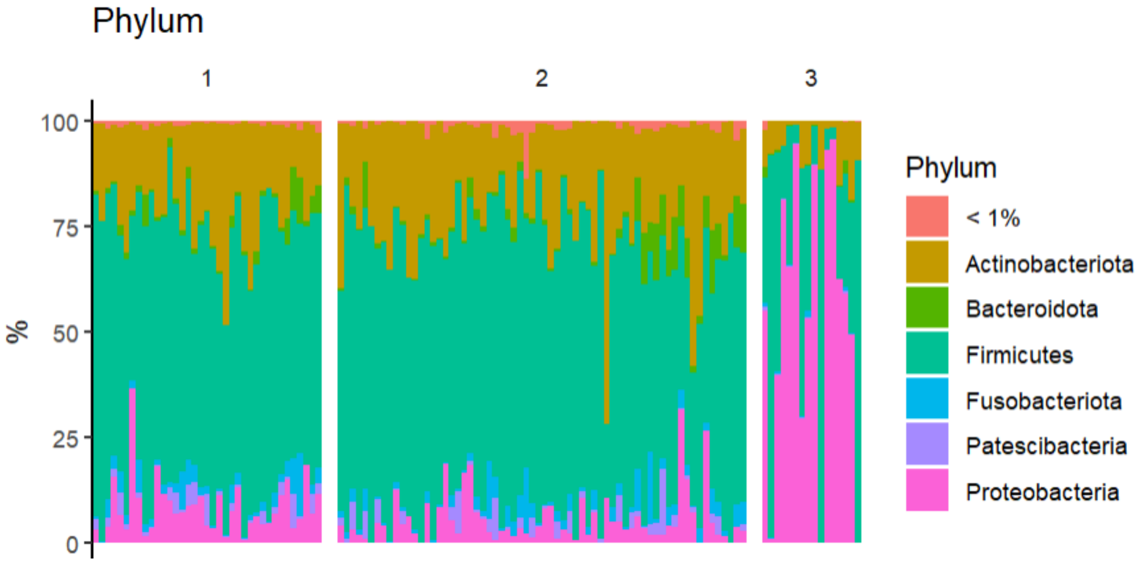


Figure 5.18 Relative Abundance of Phylum Across Clusters.

Each column represents a sample, which have been stratified according to Cluster assignment. The distribution of colours in each column reflects the relative abundance of different Phyla in that sample.

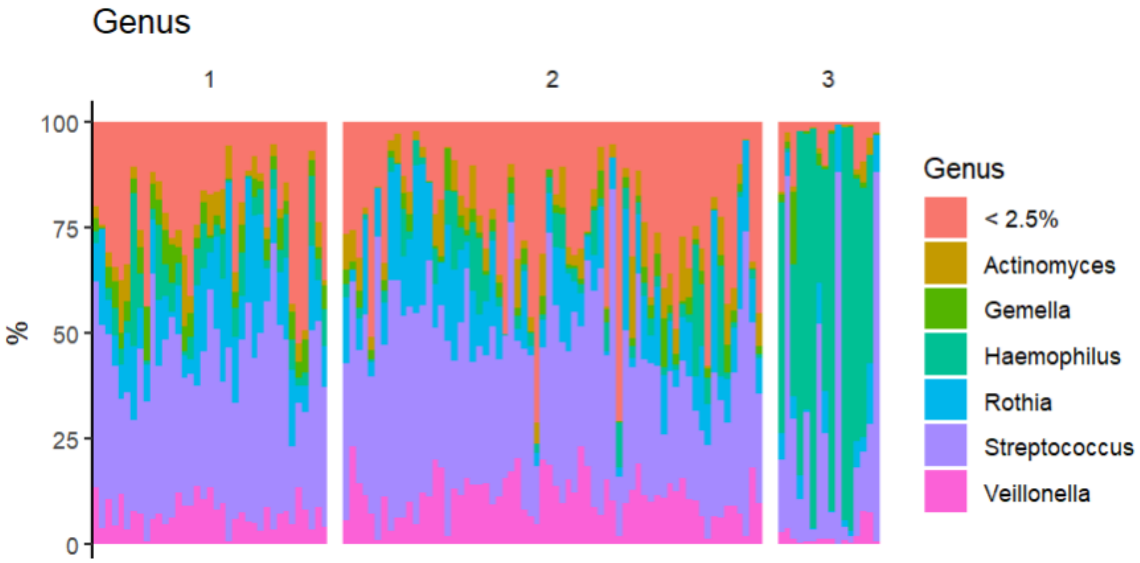


Figure 5.19 Relative Abundance of Genus Across Clusters.

Each column represents a sample, which have been stratified according to Cluster assignment. The distribution of colours in each column reflects the relative abundance of different Genus in that sample.

Further interrogation of the microbial differences was assessed by measuring the differential abundance of taxa between clusters using DESeq2. Relatively few taxa are differentially expressed between Cluster 1 and 2. Cluster 1, which has the highest alpha diversity has increased *Rothia*

(ASV 25) compared to Cluster 2; whilst Cluster 2 had increased *Streptococcus* (ASV2, ASV3), *Lactobacillus* (ASV193) and *Actinomyces* (ASV255).

*Streptococcus* (ASV4) and *Haemophilus* (ASV5) are increased in Cluster 3 compared to both Clusters 1 and 2. Additionally, *Actinomyces* (ASV75) and *Lactobacillus* (ASV193) is increased in Cluster 3 compared to Cluster 1. Compared to Cluster 3, *Rothia* (ASV16) is increased in Cluster 1. Similarly, compared to Cluster 3, *Prevotella* (ASV97), *Rothia* (ASV177), *Streptococcus* (ASV200) is increased in Cluster 2.

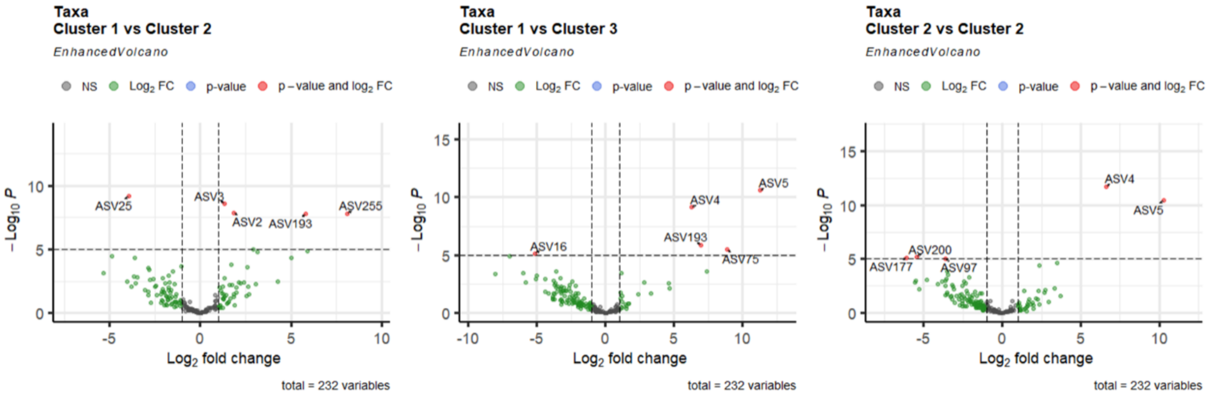


Figure 5.20 Volcano Plot illustrating the differential abundance of taxa

A: Cluster 1 and Cluster 2; B: Cluster 1 and Cluster 3; C: Cluster 2 and Cluster 3.

The differential expression analysis corroborates relative abundance plots indicating *Haemophilus* to be characteristic of Cluster 3. ASV5, identified as *Haemophilus*, was cross-referenced against the rRNA_typestrains/16S_ribosomal_RNA database using BLAST in order to identify it's *Haemophilus* species. The top hit was *Haemophilus influenzae* and *Haemophilus aegypticus* (Table 5.10).

Table 5.10 Top hits for species identification of ASV5 using BLAST

| Scientific Name | Total Score |
|---|---|
| **Haemophilus influenzae** | 773 |
| **Haemophilus aegyptius ATCC 11116** | 773 |
| **Haemophilus aegyptius** | 750 |
| **Haemophilus haemolyticus** | 739 |
| **Haemophilus seminalis** | 739 |

There are few clinical characteristics that differ between the three clusters (Table 5.11). Though it does not reach statistical significance, Cluster 1 has the highest sputum eosinophils. These patients also have the highest proportion of nasal polyposis, are accordingly on higher levels of T2 directed therapy (maintenance oral steroids and anti IgE therapy) and there is a trend towards higher FeNO levels. Cluster 3 is characterised by higher sputum neutrophils and there is a trend towards poorer lung function. However, there are no other discerning clinical characteristics by which to otherwise identify them. Notably, however, Cluster 3 does not appear to on higher levels of T2 supressing therapy, such as Cluster 1. Cluster 2 appears, clinically, to have pauci-cellular airway inflammation.

Table 5.11 Clinical Characteristics Between Clusters

| | **Cluster 1 (n=37)** | **Cluster 2 (n=66)** | **Cluster 3 (n=16)** | **P-Value** |
|---|---|---|---|---|
| Run (% Run 1) | 31 (83.8) | 48 (72.7) | 13 (81.2) | 0.403 |
| Sex (% Female) | 20 (54.1) | 33 (50.0) | 5 (31.2) | 0.298 |
| Age | 53.0 [40.0,61.0] | 56.5 [49.0,67.8] | 56.0 [45.8,65.2] | 0.296 |
| BMI | 29.1 [26.1,34.7] | 29.6 [26.5,34.5] | 29.4 [25.1,35.2] | 0.965 |
| Never Smoker | 28 (75.7) | 37 (56.1) | 9 (56.2) | 0.289 |
| Atopic | 22 (59.5) | 34 (51.5) | 9 (56.2) | 0.732 |
| GORD | 21 (56.8) | 42 (63.6) | 9 (56.2) | 0.737 |
| Nasal Polyps | 25 (67.6) | 54 (81.8) | 15 (93.8) | 0.02 |
| Age of Onset | 4.0 [0.0,28.0] | 12.0 [0.0,34.2] | 16.5 [9.0,33.2] | 0.227 |
| Exacerbations in the Last 12 Months | 1.0 [0.0,5.0] | 2.0 [0.0,3.0] | 2.0 [0.0,6.0] | 0.862 |
| ICS (BDPe) | 3000.0 [2000.0,3840.0] | 3000.0 [2000.0,3000.0] | 3000.0 [2000.0,3125.0] | 0.685 |
| Maintenance OCS | 26 (70.3) | 23 (34.8) | 7 (43.8) | 0.002 |
| Anti IgE | 9 (24.3) | 5 (7.6) | 0 | 0.012 |
| Anti IL-5 | 5 (13.5) | 5 (7.6) | 2 (12.5) | 0.594 |
| ACQ6 | 2.5 [1.2,3.3] | 2.5 [1.7,3.0] | 2.5 [1.5,3.7] | 0.822 |
| HADS | 8.0 [3.0,14.0] | 10.0 [6.0,14.0] | 16.5 [6.8,24.0] | 0.145 |
| SNOT20 | 14.0 [0.0,33.0] | 25.0 [4.0,45.0] | 2.5 [0.0,32.5] | 0.13 |
| FeNO | 35.0 [20.0,57.0] | 25.5 [14.0,47.5] | 22.5 [15.8,39.2] | 0.13 |
| PostBD FEV1 (% predicted) | 62.1 [54.8,88.0] | 76.2 [63.8,89.1] | 56.3 [38.8,83.4] | 0.105 |
| PostBD FVC (% predicted) | 89.9 [72.2,97.5] | 93.1 [83.1,102.0] | 76.8 [60.2,95.5] | 0.06 |
| PostBD FEV1/FVC | 64.0 [54.0,76.0] | 66.0 [60.0,76.8] | 59.5 [45.0,76.0] | 0.443 |
| PostBD FEF25-75% (% predicted) | 32.9 [27.9,69.7] | 42.9 [29.7,73.8] | 39.0 [14.9,54.4] | 0.26 |

| | Cluster 1 (n=37) | Cluster 2 (n=66) | Cluster 3 (n=16) | P-Value |
|---|---|---|---|---|
| FEV1 Reversibility | 0.0 [0.0,9.2] | 2.1 [0.0,12.8] | 0.0 [0.0,1.2] | 0.244 |
| Blood Neutrophils | 4.9 [4.4,6.4] | 5.2 [4.3,6.4] | 6.2 [5.2,7.5] | 0.204 |
| Blood Eosinophils | 0.3 [0.1,0.5] | 0.2 [0.1,0.4] | 0.2 [0.2,0.4] | 0.432 |
| Serum Total IgE | 132.1 [42.0,349.4] | 72.5 [24.7,294.4] | 40.8 [11.7,324.3] | 0.461 |
| Sputum Neutrophils (%) | 28.4 [18.0,58.6] | 38.6 [24.2,56.8] | 73.4 [54.6,87.0] | <0.001 |
| Sputum Eosinophils (%) | 6.1 [0.6,26.6] | 1.6 [0.5,10.0] | 1.6 [0.6,5.7] | 0.181 |

There are no differences in serum activation markers between the three microbial clusters (Table 5.12). However, sputum MPO is markedly increased in Cluster 3, as is sputum ECP; sputum Elastase and sputum EDN are not increased in Cluster 3 (Table 5.13). Sputum ECP is highest in Cluster 3 and lowest in Cluster 2.

Table 5.12 Serum Activation Markers Between Microbial Clusters

| | Cluster 1 (n=37) | Cluster 2 (n=66) | Cluster 3 (n=16) | P-Value |
|---|---|---|---|---|
| Serum MPO | 247.7 [166.8,313.9] | 234.8 [173.6,311.0] | 212.8 [165.1,250.2] | 0.683 |
| Serum Elastase | 46.5 [28.2,65.2] | 38.9 [28.4,72.6] | 39.0 [25.0,87.0] | 0.957 |
| Serum EDN | 40.9 [29.5,61.9] | 42.1 [34.7,61.0] | 43.2 [29.8,58.8] | 0.872 |
| Serum ECP | 31.6 [11.6,62.0] | 24.3 [13.6,61.5] | 22.2 [10.3,68.8] | 0.883 |

Table 5.13 Sputum Activation Markers Between Microbial Clusters

| | Cluster 1 (n=37) | Cluster 2 (n=66) | Cluster 3 (n=16) | P-Value |
|---|---|---|---|---|
| Sputum MPO | 729.4 [466.4,2029.0] | 879.8 [472.0,1914.8] | 2728.0 [2078.2,8273.2] | <0.001 |
| Sputum Elastase | 5.9 [3.6,10.2] | 7.0 [2.7,10.3] | 4.8 [3.1,5.5] | 0.174 |
| Sputum EDN | 801.5 [194.8,1577.0] | 336.9 [112.9,1267.2] | 958.3 [254.2,1465.0] | 0.206 |
| Sputum ECP | 792.0 [171.0,1881.0] | 378.2 [96.9,1286.5] | 1380.0 [1021.7,2286.5] | 0.027 |

### 5.3.6 Biomarkers for the *Haemophilus* Cluster

Confirming the observations from Table 5.9, T2 biomarkers, Blood Eosinophils and FeNO, poorly predict the *Haemophilus* Cluster of patients. Recycling the code from Chapter 4 Granulocyte

Activation Markers in Severe Asthma, none of the currently available biomarkers, or the serum activation markers are able to identify patients with *Haemophilus* (Table 5.14).

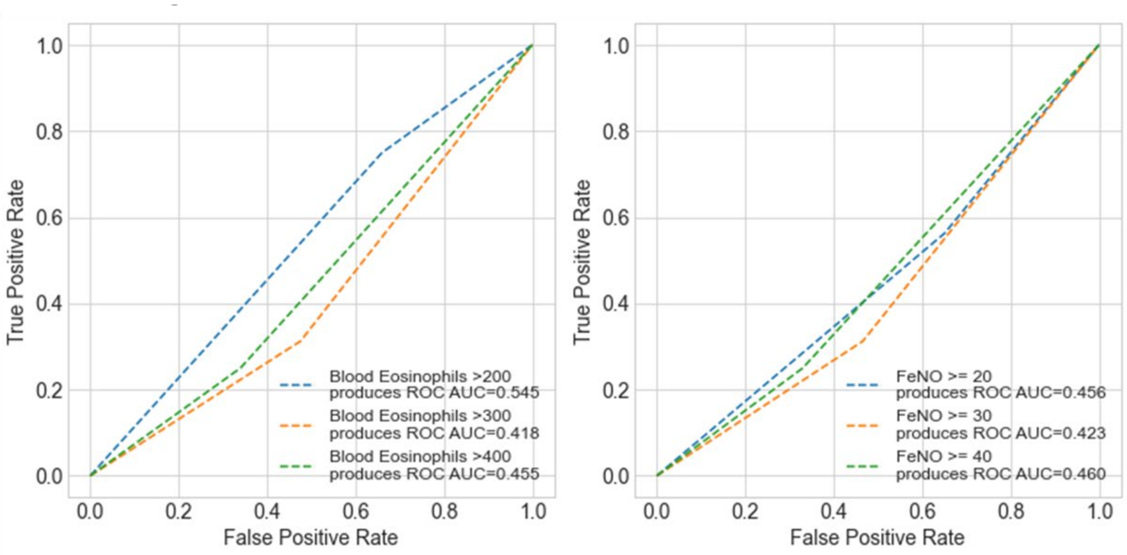


Table 5.14 Biomarkers for Predicting the *Haemophilus* Cluster

| Variable | Cutoff | AUC | PPV |
|---|---|---|---|
| Blood Neutrophils | 5.20 | 0.63 | 0.19 |
| Blood Eosinophil | 0.70 | 0.59 | 0.33 |
| Serum Total IgE | 505.00 | 0.56 | 0.22 |
| Serum Myeloperxidase (MPO) | 117.30 | 0.54 | 0.14 |
| Serum Elastase | 80.59 | 0.57 | 0.22 |
| Serum Eosinophil Derived Neurotoxin (EDN) | 0.00 | 0.50 | 0.13 |
| Serum Eosinophil Cationic Protein (ECP) | 4.54 | 0.51 | 0.14 |
| Fraction of Exhaled Nitric Oxide (FeNO) | 5.00 | 0.51 | 0.14 |

## 5.4 Discussion

### 5.4.1 Importance of *Haemophilus*

Clustering patients on their airway microbiome identifies a unique cluster of patients with loss of diversity due to an increased relative abundance of *Haemophilus*. Consistent with other asthma cohorts, these patients were associated with an increase in sputum neutrophilia[287,289,305]. The relationship between airway microbiome and asthma inflammatory phenotypes is undoubtedly complex and bi-directional[306] but the observation that this neutrophilia is associated with increased MPO indicates an innate immune response. MPO plays an important role in the microbicidal activity of phagocytes [307] through a variety of mechanisms, including NETosis [308].

Ideally, as the *Haemophilus* genus demonstrates a high level of genomic diversity[309], sequencing to species depth would confirm that the causative organism is *Haemophilus influenzae*. Taxonomic assignment on short amplicon reads (such as by using BLAST) has a high false positive rate but corroborates a growing body of evidence from modern culture-independent techniques for microbial profiling of the lower airways in the importance of H. influenzae[208,236,274,310]. Identification of this cohort of patients would be clinically salient as treatment with Azithromycin (macrolide antibiotic) is predicted by baseline colonisation with *Haemophilus influenzae*[311] and can reduce its bacterial load[312], thus representing a treatable trait.

Of course, increases in *Haemophilus* is associated with loss of bacterial diversity and it may be this that is the mechanistically salient feature [313]. Cluster 3 in this analysis was characterised by loss *Rothia* as much as increase in *Haemophilus*. *Rothia* species have been identified as having anti-inflammatory effects[314] and it could be hypothesised that the airway neutrophilia seen in this cluster might be more directly related to that than *Haemophilus*.

Cross-sectional correlations, such as in this study design, are ill suited to exploring this further, but longitudinal profiling of the airway microbiome in asthma is challenging. Broadly, the microbiome appears relatively stable over 18 months[310], unless treated with antibiotics[312]. In COPD, airway colonisation with *Haemophilus* has been demonstrated to vary from days to years[315-317] and, in asthma the expression of factors that promote neutrophilia appear to shift over time[318].

### 5.4.2 Co-morbid Bronchiectasis

One of the limitations of this study is that patients in this study were not radiologically screened with high resolution computed tomography (HRCT) for bronchiectasis, which is an airways disease characterised by chronic cough, expectoration and increased susceptibility to infection[319]. Bronchiectasis commonly occurs co-morbidly with severe asthma[320], possibly as a reflection of a partial immunodeficiency derived from chronic corticosteroid therapy[321]. Though the patients in this study were screened clinically, it is plausible that some of the differences between phenotypes (and clusters) is driven by unrecognised bronchiectasis.

It is widely recognised that airways diseases commonly overlap, and fixation on clinical labels should move towards focus on treatable traits[322]. Bronchiectasis shares many of the characteristics described by the *Haemophilus* cluster[323-325] and so it is plausible that they might also benefit from similar therapies, such as antibiotic (rather than exclusively steroid) therapy [326].

### 5.4.3 Airway Microbiome Provides Granularity to Sputum Inflammatory Phenotypes

Historically studies have described the microbial characteristics across clinically defined phenotypes, identifying airway neutrophilia to associate with increased *Haemophilus* and *Moraxella* taxa[208,236]. Such a clear relationship does not emerge from this data, which appears slightly more nuanced: neutrophilic airways do not always have increased abundance of *Haemophilus* but airways with increased abundance of *Haemophilus* are always neutrophilic, consistent with recent cluster analyses[310,327].

This would suggest that there are heterogenous mechanisms underlying sputum neutrophilia. Corticosteroids are known to improve the survival of neutrophils[205] and, at least in some patients, withdrawal of inhaled corticosteroids is associated with decrease in sputum neutrophils[251]. A variety of other factors have been associated with sputum neutrophilia, including smoking/pollution and obesity/insulin resistance[328]. These heterogenous mechanism likely underlie the reason for the failure of therapeutics targeting neutrophil recruitment have not been proven successful[273].

### 5.4.4 Lack of Biomarkers

Ultimately, identification of target patients is essential for the implementation of targeted therapies, however, our findings demonstrate that there are currently no good biomarkers available to do this. As described in previous chapters, there are no good biomarkers for sputum neutrophilia (Chapter 3.3.3.4) and even then, sputum neutrophilia is a poor proxy for the *Haemophilus* cluster. Rather than predicting sputum neutrophilia, it is critical to develop a well validated biomarker for this *Haemophilus* cluster, which could now be considered a treatable trait.

# Chapter 6 Exhaled VOCs in Severe Asthma

## 6.1 Introduction

The preceding chapters (Chapter 3, 4, 5) describe the heterogeneity of severe asthma, the complexity of airway biology and limitations of existing clinical biomarkers. As detailed in Chapter 1, exhaled breath biomarkers are and attractive solution to the challenge of identifying biomarkers for severe asthma due to their direct contact with the organ of interest and ease of collection. This enthusiasm is bolstered by the report from a number of asthma studies demonstrating sensitivity of exhaled VOCs to airway biology[136].

One of the major limitations to breathomics is the lack of standardisation, which extends to data analysis. In contrast to microbial analysis, for example, a consensus framework exists for almost all aspects basic analysis (as detailed in Chapter 5). Though there has been no consensus approach for assessing sensitivity to airway inflammation, the majority of breathomics studies pair feature reduction techniques with a supervised machine learning model for classification. In the absence of guidelines/consensus on approaches, an analytical pipeline needs to be developed and validated.

Feature reduction is commonly performed in breathomics, through Principal Component Analysis[136]. Broadly, linear dimensionality reduction techniques transform the features into a new set of lower dimension features, whilst attempting to retain the variance of the data. Such an approach assumes that all of the original features are of equal value (or does not assume that any feature is more/less important than another). Less frequently, studies have used feature selection techniques [197], in which features are eliminated according to a particular criterion. One of the functions of feature selection is to optimise the signal to noise ratio. The combination of study characteristics (many features, few observations) risks model overfitting [329] and false discovery [330] but feature selection by a relevant criterion need not be restricted to machine learning.

Noisy VOCs are difficult to identify in cross-sectional analyses but might be easier to identify in longitudinal studies. VOCs that do not change with a stressor (e.g. during an asthma exacerbation) are likely not be important to it. Such a study design would be relevant to asthma but challenging to deliver. The opposing hypothesis, though less elegant, may nonetheless be valuable: VOCs that change (excessively) during a clinically stable state are likely not to be relevant to the that physiology. It is recognised that a number of exhaled VOCs are unrelated to airway biology but related to transient factors such as diet [223], exercise [224]; such VOCs would be transient during the clinically stable state. Transient VOCs, might be identifiable on repeat sampling and could inform feature selection for cross-sectional analyses.

Prior to attempting to predict the novel *Haemophilus* phenotype described in Chapter 5, the eosinophilic phenotype will be used to benchmark the machine learning classifier. This is the the most suitable phenotype for this purpose as, despite the limitations described in Chapter 3, it remains relatively robust and has the advantage of being used in multiple other studies described in the literature.

The objectives of this chapter are:

- Describe the VOCs present in the exhaled breath of severe asthma patients
- Describe the nature of variance in repeat breath samples and describe the short-term repeatability of VOCs
- Describe the relationship between VOCs and sputum eosinophilia
- Define a framework for training VOC machine learning classifier models on breath samples from the cross-sectional arm and testing on the first breath samples from the repeatability arm
- Evaluate the performances of VOC models for predicting the sputum eosinophilic phenotype
- Evaluate the performances of VOC models that incorporates feature selection using measures of short-term repeatability for predicting the sputum eosinophilic phenotype

## 6.2 Chapter Specific Methods

### 6.2.1 Breath Sampling in Severe Asthma WATCH Sub-Study

This study was designed in order to pair breath samples to sputum samples collected in the Airway Sampling of Severe Asthma Patients sub-study and clinical characterisation of the WATCH

Study (Chapter 2.3). As sputum induction can cause airway irritation [195], breath samples were collected either before sputum induction or at least 48 hours after sputum induction, in order to not be confounded by the airway irritation due to sputum induction. There was a maximum of seven days between sputum induction and starting breath sampling.

The Breath Sampling of Severe Asthma Patients study had two arms: “cross sectional” and “repeatability”.

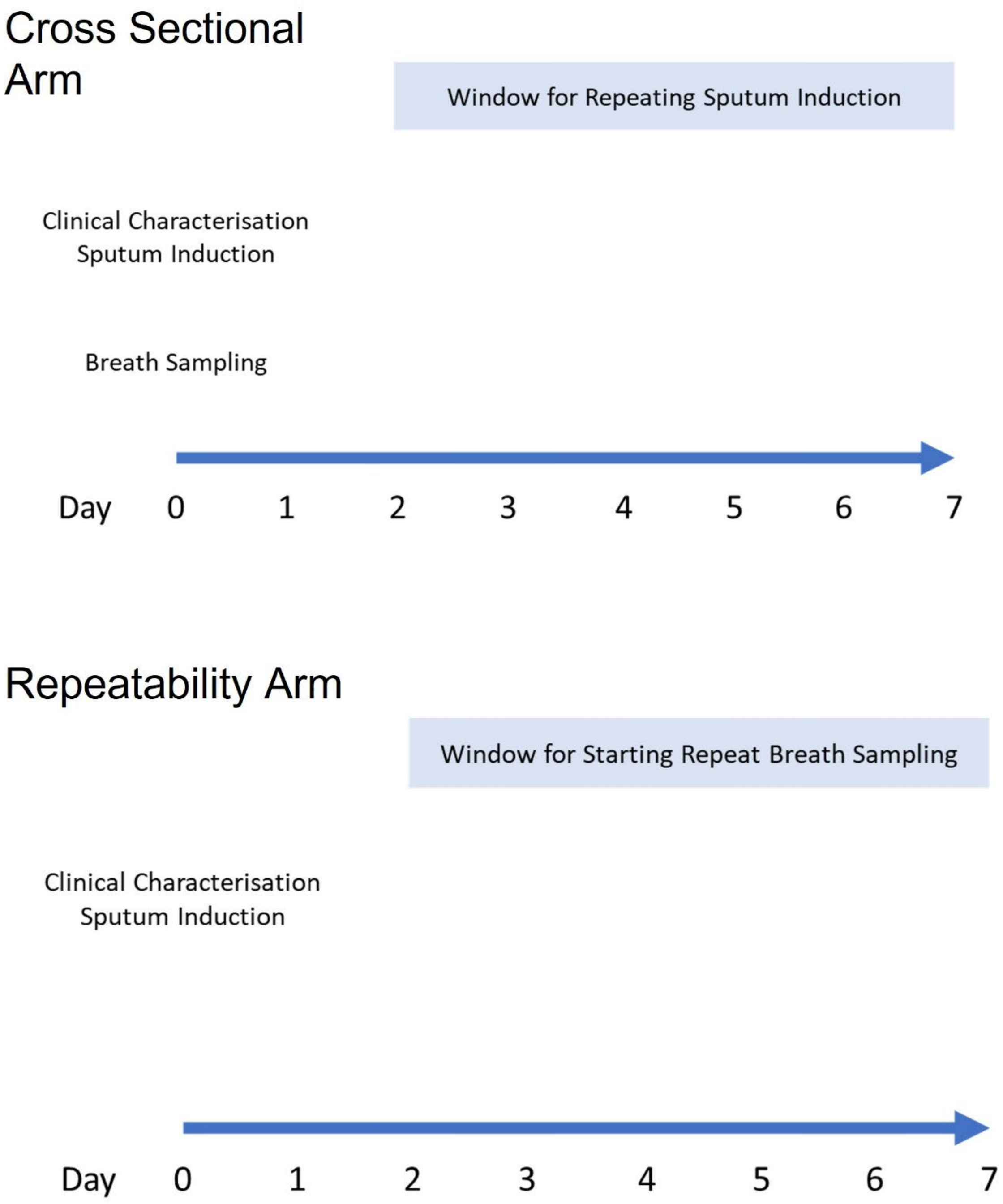

Cross Sectional Arm
Window for Repeating Sputum Induction
Clinical Characterisation
Sputum Induction
Breath Sampling
Day 0 1 2 3 4 5 6 7
Repeatability Arm
Window for Starting Repeat Breath Sampling
Clinical Characterisation
Sputum Induction
Day 0 1 2 3 4 5 6 7

Figure 6.1 Schema for Breath Sampling in each arm

In the cross-sectional arm (Figure 6.1), the full characterisation schedule was performed on the same morning, starting with breath sampling and ending with sputum induction. In some cases, in order to obtain a viable sputum sample, it was necessary to repeat sputum induction on a second date. If so, this was done within 7 days of breath sampling. Subjects were excluded from the analysis if a viable sputum sample was not obtainable.

In the repeatability arm (Figure 6.1), breath sampling was dissociated from the full characterisation schedule. Once a viable sputum sample was obtained, subjects would return to provide breath samples on five consecutive days, starting within 7 days of sputum induction; these breath samples were collected at the same time of day on each occasion.

The first breath sample from the repeatability arm of the study will be collected a maximum of seven days apart from sputum induction, which is true of breath samples from the cross-sectional arm of the repeatability study.

### 6.2.2 Patient Recruitment

Recruitment to the breath sampling sub-study were identical to the airway sampling sub-study. There were no restrictions to the cross-sectional or repeatability arms of the breath sampling sub-study.

### 6.2.3 Breath Sampling

#### 6.2.3.1 Breath Collection

All breath samples were collected within the same room. Breath samples were collected using the ReCIVA Breath Sampler (Owlstone Medical Ltd.). Exhaled breath was collected onto a Breath Biopsy Cartridge, which consists of four Tenax TA/Carbograph 5TD sorbent tubes (Markes International). The ReCIVA Breath Sampler monitored subjects' tidal breathing pattern in real time, using $CO_2$ concentration and pressure sensors. Dynamically determined gates using the real-time pressure levels triggered the sampling pumps to collect breath. Each pump pulled pressure-gated exhaled breath through two sorbent tubes, with 1473 ml being collected on each tube. Each pair of tubes was later combined to give a single sample for TD-GC-MS analysis.

#### 6.2.3.2 Breath Analysis

Samples analysed by Owlstone Ltd: first, they were dry purged to remove excess water and desorbed using a TD100-xr thermal desorption autosampler (Markes International) and transferred onto a Quadrex 007-624 column (30 m x 0.32 mm x 3.00 µm) using splitless injection. Chromatographic separation was achieved via a programmed method (40-250˚C in 84.5 min at 3.0 mL/min) on a 7890B gas chromatography (GC) oven (Agilent Technologies) and mass spectral data acquired using an electron impact ionization time-of-flight (TOF) BenchTOF high definition mass spectrometer (MS) (Markes International). Each sample consisted of two sorbent tubes, both of which were desorbed into the Thermal Desorber cold trap for a single analysis. A cleaning method was run in between each sample to prevent carry-over.

A quality control sample (sorbent tube spiked with a known mixture of chemicals) was run in between every four subject breath samples to monitor the stability of instrumentation. A blank tube was run every four samples and after every quality control sample to monitor background. A set of four samples, quality control samples, and blank tubes are denoted as an “analytical sequence.”

#### 6.2.3.3 Breath Data Pre-Processing

Retention time shifts due to column events were corrected using retention time of QC compounds in QC samples. For each QC sample, a piece-wise linear function was constructed by comparing QC compound retention times in the sample to the compound-specific medians across all QC samples. This piece-wise linear function was then applied to the retention time axis of breath samples that were analysed immediately after the QC sample. Small deviations in peak area, introduced by retention time alignment, was corrected using the scaling factors derived from the piece-wise linear functions.

Untargeted feature extraction was performed for samples that passed all curation checks. TD-GC-MS chromatograms were converted into molecular feature (MF) lists for statistical analysis. Whenever a feature was below the limit of detection (LOD), the baseline for that feature was integrated instead to give a minimum value. If a feature could not be reliably quantified due to issues not associated with LOD (e.g. interference from neighbouring peaks), no baseline integration was performed, and the feature was marked as non-LOD missing.

Features were excluded from downstream analysis if they were not present in at least 80% of samples of each inflammatory phenotype. Each feature was assigned a tentative ID by comparison to the National Institute of Standards and Technology (NIST) mass spectra standard reference database (2017). A tentative ID was assigned if the match score was > 85%.

### 6.2.4 Statistical Analysis

Statistical analysis was performed using Python scripting language (version 3.8.3) [237]. Clinical characteristics were described using median and 95% confidence intervals with between group comparisons by Mann Whitney U tests for continuous variables and absolute numbers with percentages within each group and Chi Squared tests for categorical variables.

Unsupervised clustering was performed as described in 5.2.5.3.

### 6.2.5 Data Transformation

For the exploratory analysis of the breath data, in order to maximise sample number, samples from Visit 1 of the repeatability arm were combined with samples from the cross-sectional arm of the study. Pre-processing steps were performed in all samples together.

For all other analyses, pre-processing was performed in the samples in the cross-sectional study and these parameters were applied independently to data from each Visit in the repeatability arm. The repeatability analysis compares VOCs across visits and the machine learning prediction models used Visit 1 as the test set and the cross-sectional data used as the training set. Pre-processing in the above manner prevents data leakage (e.g., data from the test set contributing to pre-processing of the training set).

Data was log transformed in order to try to achieve a gaussian distribution and then scaled such that each feature was given a range between 0 and 1

### 6.2.6 Batch Effect

Principal Component Analysis (PCA) (sklearn.decomposition.PCA) was used to explore the breath data. Breath features were reduced to principal components, where each principal component attempts to capture the maximum variance in data. The first two principal components were visualised in a scatterplot in order to investigate the impact of technical measures: proportion of target volume collected, breath sampling duration, GCMS platform, storage duration and the resolution between similar standards measured prior to each sample.

Batch correction was performed using pycombat.

### 6.2.7 Descriptive Analysis of Cross-Sectional Data

Each molecular feature was assigned a structural category using the contextual expertise of Dr Paul Afolabi and Dr Grielof Koster.

The absolute abundances of VOCs was assessed using transformed and batch corrected data without scaling.

A correlation matrix for all VOCs was constructed using Pearson's correlation coefficients and visualised as a graph using Cytoscape [331]. Each node represents a molecular feature, which was labelled with its tentative ID. Each line between the nodes (edge) represents a positive correlation coefficient with a p value <0.05; the thickness of the edge reflects the strength of correlation (thicker line represents a stronger correlation). The clusterMaker2 [332] plug in was used, using a GLay network partitioning algorithm, to identify clusters within the graph

Relationships between the abundance of VOCs and log transformed Sputum Inflammatory cell counts were performed using Pearson's correlations

For differential expression analyses, data was exported to R for analysis using the limma package. limma uses an empirical Bayes method to moderate the standard errors of the estimated log-fold changes. This results in more stable inference and improved power [333]. Sex and $FEV_1$/FVC ratio was accounted for in the differential expression model. Volcano plots were constructed using the EnhancedVolcano library, which visualise the logfold change and p value for any VOCs found to be differentially abundant.

### 6.2.8 Principal Component Analysis

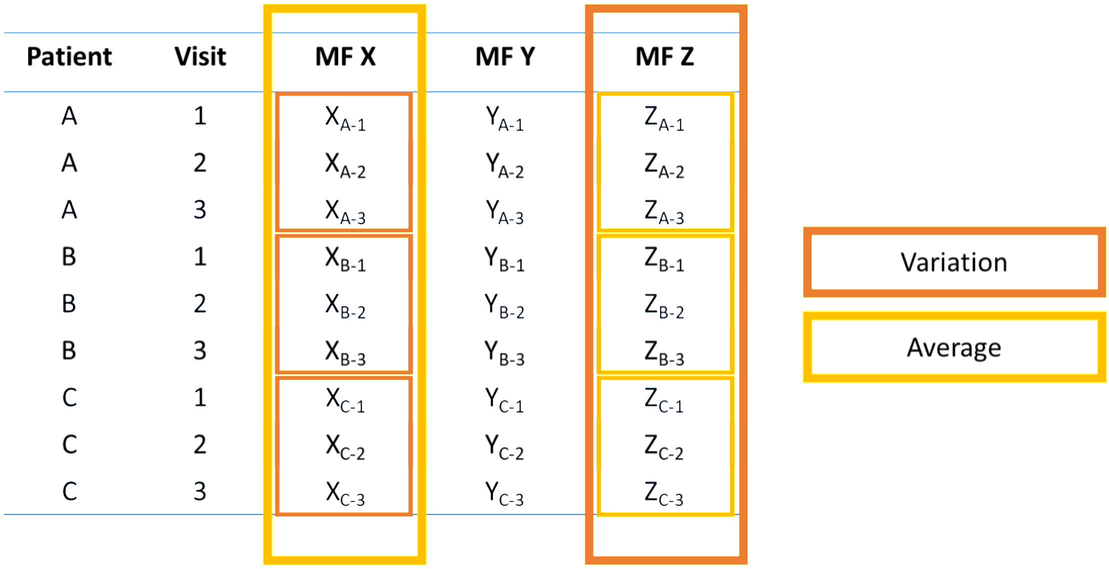
Principal Component Analysis was used for exploratory data analysis. A bar chart was used to visualise the cumulative percentage of variance captured by successive principal components. The first ten principal components were correlated (using Pearson's Correlation coefficient) to a restricted set of clinical characteristics and visualised as a heatmap: Non-significant correlations were coloured grey/white, positive and negative correlations with p-value < 0.05 in red and blue, respectively.

PCA plots were constructed from the first two principal components, where each point represents a breath sample. Ellipses were constructed and positioned using the mean and coefficient of variation of the principal components within the grouping of interest as a representation of a 95% confidence interval. Hierarchichal Clustering of Repeat Sample Visits

A supervised machine learning approach was applied to the data in the Repeat data. Ward's hierarchical clustering on Euclidean distances between the VOC features was used to identify 14 clusters (to match the 14 subjects).

### 6.2.9 Within Subject Variability

“Between Visit” and “Between Subject” variation was calculated for each molecular feature using the coefficient of variation (scipy.stats.variation). “Between Visit” variability (or Within Subject) attempts to capture the variation of each molecular feature in the same individual but measured at different visits. A coefficient of variation was calculated across Visits 1-5 for each patient. From those fourteen CVs, the average (median) value was extracted) (illustrated by MF X in Figure). “Between Subject” variability (or Within Visit) attempts to capture the variation of each molecular feature across individuals. For each patient, the average (median) value for each molecular feature across Visits 1-5 was calculated and a coefficient of variation taken from these (illustrated by MF Z in Figure 6.2).



| Patient | Visit | MF X | MF Y | MF Z |
|---|---|---|---|---|
| A | 1 | $X_{A-1}$ | $Y_{A-1}$ | $Z_{A-1}$ |
| A | 2 | $X_{A-2}$ | $Y_{A-2}$ | $Z_{A-2}$ |
| A | 3 | $X_{A-3}$ | $Y_{A-3}$ | $Z_{A-3}$ |
| B | 1 | $X_{B-1}$ | $Y_{B-1}$ | $Z_{B-1}$ |
| B | 2 | $X_{B-2}$ | $Y_{B-2}$ | $Z_{B-2}$ |
| B | 3 | $X_{B-3}$ | $Y_{B-3}$ | $Z_{B-3}$ |
| C | 1 | $X_{C-1}$ | $Y_{C-1}$ | $Z_{C-1}$ |
| C | 2 | $X_{C-2}$ | $Y_{C-2}$ | $Z_{C-2}$ |
| C | 3 | $X_{C-3}$ | $Y_{C-3}$ | $Z_{C-3}$ |

Figure 6.2 Illustration of How Within and Between Variability in VOCs was Calculated

Any VOC with a mean between-subject variability of ≥30% was considered potentially discriminatory. Any VOC with a mean within-subject variability of ≥30% was considered inconsistent. These criteria were used to categorise VOCs into four categories: “Conserved”: low variability within subjects and between subjects, “Erratic”: high variability within subjects and between subjects, “Potential biomarkers”: low variability within subjects but high variability between subjects, and “Noisy”: high variability within subjects but low variability between subjects.

### 6.2.10 Supervised Machine Learning for Classification

#### 6.2.10.1 Training and Test Sets

Breath samples collected from the cross-sectional arm of the study were used as the training set (n=60) and the first breath sample from the repeat arm of the study were used as the test set (n=14). As described in 1.2.2, pre-processing was conducted in such a way that there was no data leakage from the test set to the training set.

#### 6.2.10.2 Prediction Outcome

Patients were phenotyped according to their sputum inflammatory cell counts: eosinophilia was defined as sputum eosinophils ≥2% and non-eosinophilia as sputum eosinophils <2%.

#### 6.2.10.3 Model Construction

XGBoost (eXtreme Gradient Boosting) is an implementation of gradient boosted decision trees [334]. Tree based ensemble algorithms are well suited to biological data due to their tolerance of non-gaussian data distributions, multi-collinearity of features and outliers. We chose to apply XGBoost to our data due to its predictive accuracy and recent successful application of random forests (parallel ensemble of decision trees) to asthma breathomics data [197]. Boosting describes an ensemble technique in which predictions from new models are sequentially combined to improve the overall performance of the model. Gradient boosting specifically describes the use of a gradient descent algorithm to minimize loss when adding new models.

Each model was tuned to optimise for cross entropy loss using a Bayesian Optimisation algorithm [335], which builds a probability model to search over the most promising model hyperparameters (the number of tress, maximum tree depth, L1 regularization term on weights, L2 regularization term on weights, the minimum sum of weights of all observations required in a child, the minimum loss reduction required to make a split) for the objective function, within a threefold cross validation.

#### 6.2.10.4 Feature Selection

Feature selection was performed using the feature importance tool within the eXtreme Gradient Boosting algorithm. Feature importance was defined using gain, which is a measure of how much the classification metric improves following branching using that feature. The optimal set of features was calculated by splitting the training set such that models could be trained on different subsets of features. The minimum number of features producing the best accuracy were selected to be assessed in the test set.

#### 6.2.10.5 Model Evaluation

The predictive performance of each developed model was evaluated on the test set using prediction measures of discrimination (area under the receiver operating curve, AUC). Where an AUC of >0.6 was achieved, sensitivity, specificity, positive and negative predictive values (PPV and NPV were calculated

## 6.3 Results

### 6.3.1 Patient Population

Recruitment to the Breath Sampling of Severe Asthma study commenced in November 2018, at which time, 81 patients had provided a viable sputum sample to the Airway Sampling of Severe Asthma Study. During the period in which patients were recruited to the Airway Sampling and Breath Sampling studies, in parallel, 113 patients provided a viable sputum sample. A paired breath sample was collected in 74 patients (65.5%). Instances of sputum samples collected without paired breath samples were due, exclusively, to lack of availability of breathomics consumables (Figure 6.3)

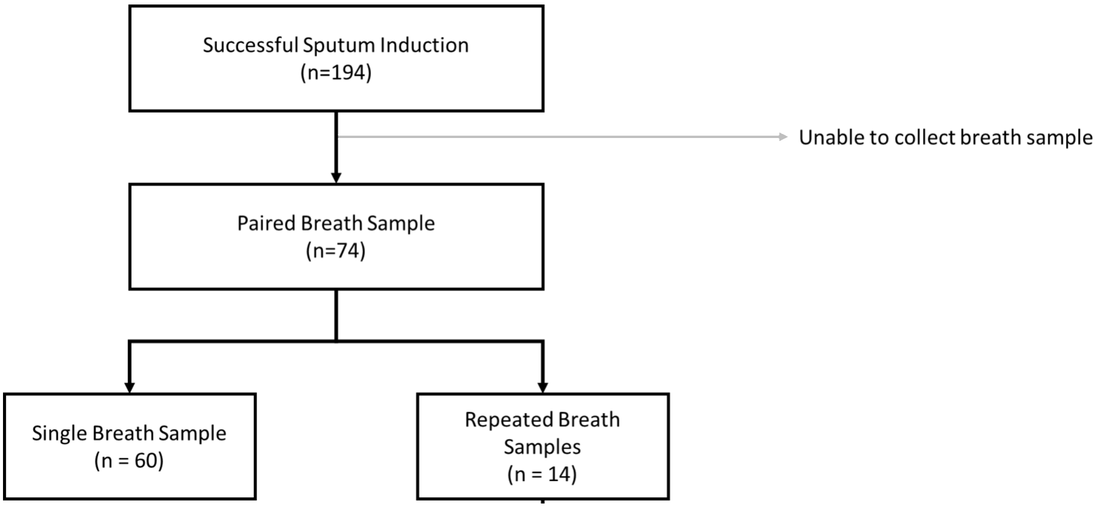

Successful Sputum Induction (n=194)
Unable to collect breath sample
Paired Breath Sample (n=74)
Single Breath Sample (n = 60)
Repeated Breath Samples (n = 14)

Figure 6.3 Consort Diagram of Patients with Paired Sputum and Breath Samples

#### 6.3.1.1 Patients Providing a Breath Sample vs Patients that did not

The only statistically significant difference between those patients providing a breath sample and those that did not as an increase in prescribed ICS in (3000 (2762-3900) vs 3000 (2000-3000) respectively, p=0.047) (Table 6.1).

Table 6.1 Comparison of Patients with and without Breath Samples

Continuous variables expressed as median [Q1, Q3] with differences measured by Mann-Whitney U test. Categorical variables expressed as n (%) with differences measured by chi-square test. Abbreviations: ICS, inhaled corticosteroid; BDPe, beclomethasone dose equivalent; FeNO, fraction of nitric oxide in exhaled breath; post BD, post bronchodilator; FEV1, forced expiratory volume in 1 second; FVC, forced vital capacity; FEF25-75%, forced expiratory flow at 25% to 75% of FVC; ACQ, asthma control questionnaire, HADSTOT, Hospital Anxiety and Depression Total Score; SNOT, SinoNasal Outcome Score

| | **No Breath (n = 121)** | **Breath Sample (n= 74)** | **P-Value** |
|---|---|---|---|
| Sex (% Female) | 70 (58.3) | 33 (44.6) | 0.086 |
| Age | 58.0 [43.0,67.0] | 56.0 [49.0,64.8] | 0.774 |
| BMI | 30.2 [26.4,34.8] | 28.1 [25.2,32.6] | 0.071 |
| Smoker (% Never) | 71 (59.2) | 53 (71.6) | 0.168 |
| Atopy | 74 (61.7) | 45 (60.8) | 0.974 |
| Age of Onset | 20.0 [7.0,40.0] | 14.5 [3.0,43.5] | 0.548 |

| | **No Breath (n = 121)** | **Breath Sample (n= 74)** | **P-Value** |
|---|---|---|---|
| Exacerbations in Last 12 months | 2.0 [1.0,4.0] | 1.0 [0.0,3.0] | 0.242 |
| ICS (BDPe) | 3000.0 [2000.0,3000.0] | 3000.0 [2762.0,3900.0] | **0.047** |
| FeNO | 23.0 [15.0,45.0] | 28.0 [17.0,46.0] | 0.25 |
| Blood Eosinophil Count | 0.2 [0.1,0.4] | 0.3 [0.1,0.4] | 0.079 |
| Sputum Eosinophil (%) | 1.5 [0.5,8.4] | 3.1 [0.5,13.2] | 0.286 |
| Sputum Neutrophil (%) | 43.2 [24.9,62.4] | 45.9 [24.0,73.6] | 0.399 |
| PostBD FEV1 | 75.9 [61.6,92.2] | 77.5 [57.4,90.9] | 0.566 |
| PostBD FEV1/FVC | 68.5 [61.0,78.0] | 66.0 [56.0,77.0] | 0.304 |
| PostBD FEF25-75 %predicted | 46.2 [31.5,79.6] | 41.7 [28.4,71.8] | 0.501 |
| ACQ6 | 2.2 [1.3,3.0] | 2.7 [1.5,3.3] | 0.126 |
| HADSTOT | 10.0 [6.0,15.2] | 10.0 [6.0,16.0] | 0.88 |
| SNOT20 | 28.0 [17.2,45.2] | 32.0 [21.0,45.0] | 0.407 |

#### 6.3.1.2 Patients Providing Repeat Breath Samples vs Patients Providing a Single Breath Sample

Patients took part, mutually exclusively, to the Cross Sectional and Repeat Sampling arms of the study (Figure 6.1). There were no statistically significant differences between patients providing a single breath sample and those that provided repeat breath samples (Table 6.2).

Table 6.2 Comparison of Patients Providing a Single and Repeat Breath Sample

Continuous variables expressed as median [Q1, Q3] with differences measured by Mann-Whitney U test. Categorical variables expressed as n (%) with differences measured by chi-square test. Abbreviations: ICS, inhaled corticosteroid; BDPe, beclomethasone dose equivalent; FeNO, fraction of nitric oxide in exhaled breath; post BD, post bronchodilator; FEV1, forced expiratory volume in 1 second; FVC, forced vital capacity; FEF25-75%, forced expiratory flow at 25% to 75% of FVC; ACQ, asthma control questionnaire, HADSTOT, Hospital Anxiety and Depression Total Score; SNOT, SinoNasal Outcome Score

| | **Single Breath (n = 60)** | **Repeat Breath (n= 14)** | **P-Value** |
|---|---|---|---|
| Sex (% Female) | 28 (46.7) | 5 (35.7) | 0.657 |
| Age | 57.5 [49.0,65.5] | 54.0 [51.2,60.2] | 0.46 |
| BMI | 28.7 [26.2,33.5] | 25.0 [23.3,31.5] | 0.082 |
| Smoker (% Never) | 42 (70.0) | 11 (78.6) | 0.703 |
| Atopy | 39 (65.0) | 6 (42.9) | 0.221 |
| Age of Onset | 16.0 [3.0,49.2] | 14.0 [5.2,30.2] | 0.68 |
| Exacerbations in Last 12 months | 2.0 [0.0,3.0] | 1.0 [0.0,3.0] | 0.456 |
| ICS (BDPe) | 3000.0 [2900.0,3860.0] | 2920.0 [2529.0,3900.0] | 0.451 |
| FeNO | 27.0 [15.5,43.5] | 38.5 [29.8,50.8] | 0.145 |
| Blood Eosinophil Count | 0.3 [0.1,0.4] | 0.2 [0.1,0.4] | 0.894 |
| PostBD FEV1 | 76.7 [59.7,89.5] | 81.5 [45.6,92.9] | 0.751 |
| PostBD FEV1/FVC | 66.0 [56.0,77.2] | 66.0 [54.5,74.5] | 0.907 |
| PostBD FEF25-75 %predicted | 40.6 [28.6,72.9] | 55.0 [25.2,68.6] | 0.989 |
| ACQ6 | 2.7 [1.5,3.3] | 2.5 [1.6,3.5] | 0.809 |
| HADSTOT | 10.0 [6.0,16.0] | 11.0 [6.0,14.0] | 0.827 |
| SNOT20 | 30.5 [20.2,45.0] | 32.0 [30.0,59.0] | 0.226 |

### 6.3.2 Pre-Processing

#### 6.3.2.1 Data Transformation

Broadly, most features demonstrate a right skew distribution, which was successfully transformed by log transformation into a gaussian-like distribution (Figure 6.4)

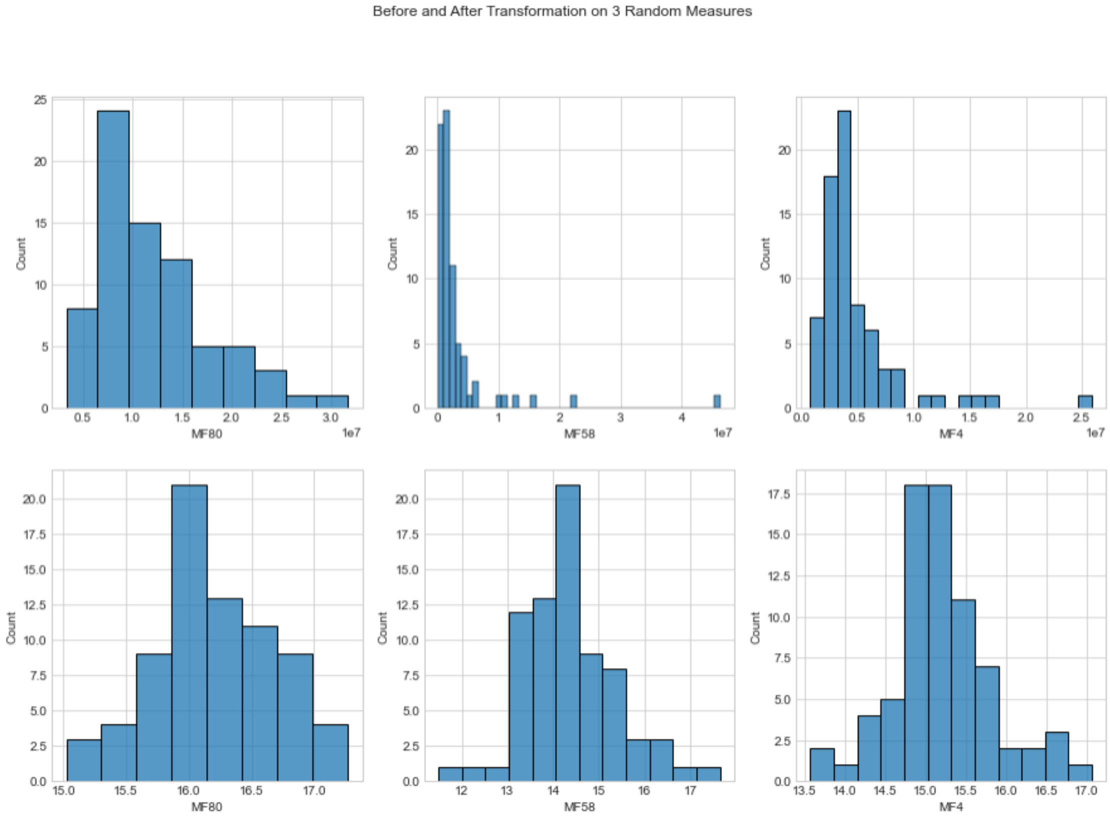


Figure 6.4 Distributions of three random VOCs before and after log transformation

### 6.3.2.2 Batch Effect

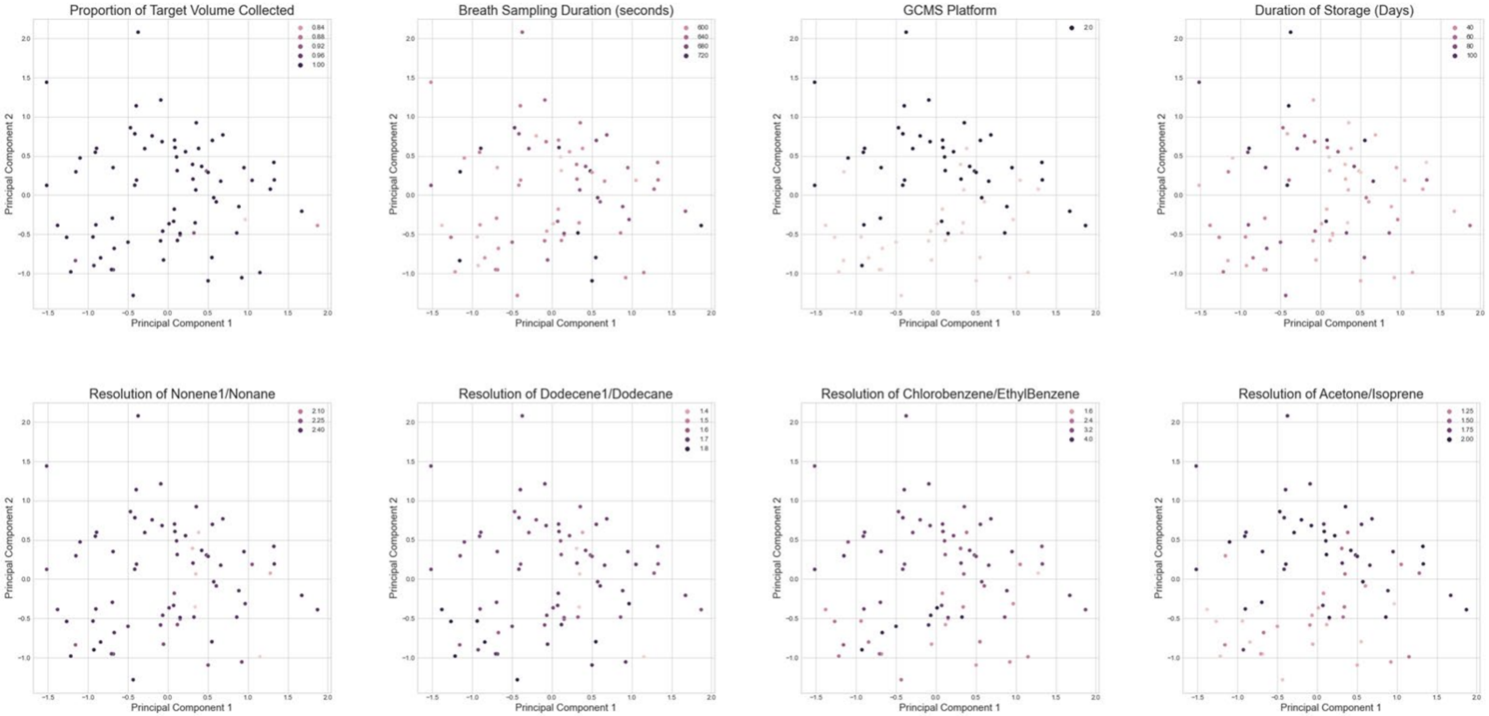


Figure 6.5 PCA plot of Samples Across Technical Breath Parameters

There was minimal separation in samples across target volume collected, breath sampling duration and storage duration (Figure 6.5). However, there was distinct separation on the PCA plots across the GCMS platform used, which was carried over to the resolution measures This is confirmed using boxplots comparing the measurements of the same VOCs across the GCMS platform on which it was analysed (Figure 6.6). This batch effect was resolved by pycombat (Figure 6.7)

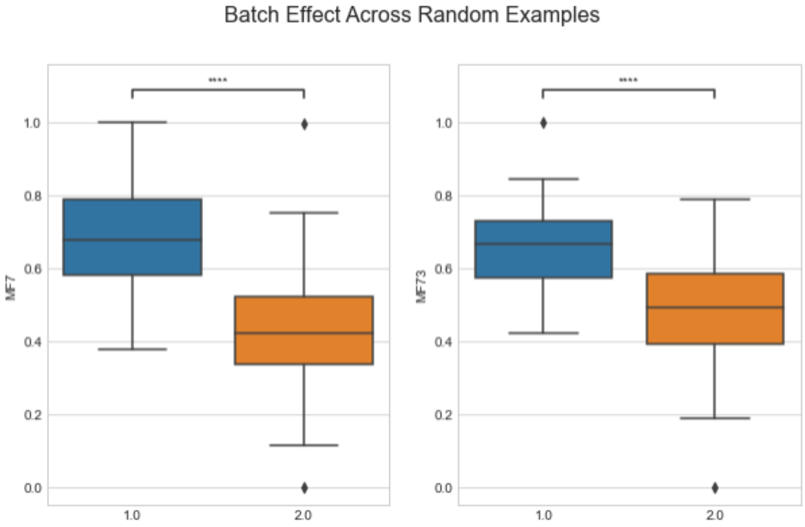


Figure 6.6 Boxplot Demonstrating Batch Effect seen Across Instrumentation for two Random Molecular Features

Blue represents one instrument and orange another instrument. Abundances have been transformed and scaled

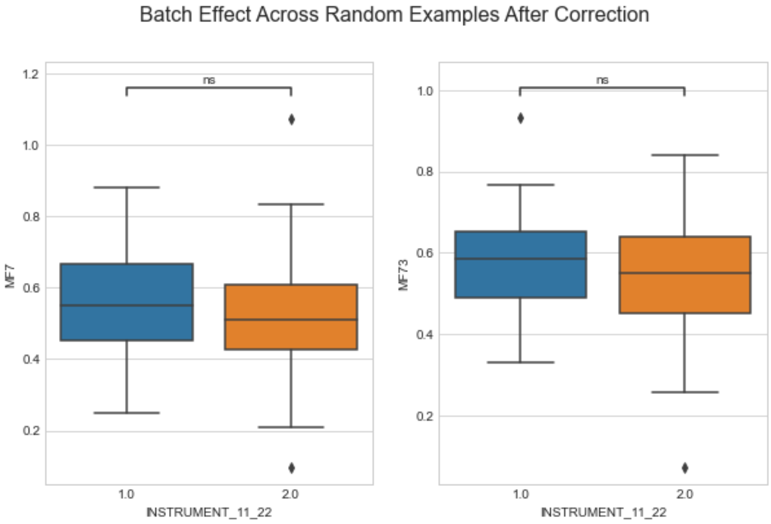


Figure 6.7 Boxplot Demonstrating Resolution of Batch Effect seen Across Instrumentation for two Random Molecular Features

Blue represents one instrument and orange another instrument. Abundances have been transformed and scaled

### 6.3.3 Exploratory Analysis of Cross-Sectional Data

Cross sectional data (n=74) was compiled from samples in the cross-sectional arm (n=60) and visit 1 from the repeated sampling arm (n=14) of the study

#### 6.3.3.1 Abundance

Over a third (42.7%) of the VOCs identified in the exhaled breath samples could be categorised as alkanes or terpenoids. In terms of absolute concentrations, aromatic esters (Furan-2-methyl and Furan-3-methyl) were most abundant, appearing in exhaled breath at a median concentration of 4,708,940 [range 2,354,471.37 - 7,063,410.03] parts per billion (ppb) (Figure 6.5), 10,000 times the concentration of the next identified category, halogenated hydrocarbons, 443.60 [443.60 - 443.60] ppb (Table 6.3).

Table 6.3 Abundance and Frequency of VOCs in exhaled breath belonging to structural categories defined by Dr Afolabi and Dr Koster.

| Category of Volatile Organic Compound | Count | Median Abundance (ppb) and [range] |
|---|---|---|
| Alkane | 21 | 3.17 [2.18 - 4.96] |
| Terpenoid | 17 | 6.05 [4.20 - 8.50] |
| Aldehyde | 7 | 10.10 [4.26 - 165.39] |
| Aromatic hydrocarbon | 7 | 1.68 [1.52 - 2.20] |
| Ketone | 7 | 6.02 [4.86 - 36.63] |
| Halogentaed hydrocarbon | 6 | 10.95 [9.74 - 20.93] |
| Organic sulfide | 5 | 62.41 [20.91 - 111.88] |
| Alcohol | 4 | 13.24 [10.78 - 34.62] |
| Aromatic ester | 2 | 4,708,940.70 [2,354,471.37 - 7,063,410.03] |
| Fluronated benzene | 2 | 8.54 [7.31 - 9.76] |
| Halogentaed aromatic hydrocarbon | 2 | 9.70 [6.07 - 13.34] |
| Cyano compound | 1 | 73.56 [73.56 - 73.56] |
| Di-ether | 1 | 9.45 [9.45 - 9.45] |
| Ester | 1 | 24.93 [24.93 - 24.93] |
| Halogenated aromatic hydrocarbon | 1 | 8.99 [8.99 - 8.99] |
| Halogenated hydrocarbon | 1 | 443.60 [443.60 - 443.60] |
| Monoterpene | 1 | 6.86 [6.86 - 6.86] |
| Organoselenium compound | 1 | 227.60 [227.60 - 227.60] |
| Organosilicon compound | 1 | 143.61 [143.61 - 143.61] |
| Unknown | 1 | 11,380.38 [11,380.38 - 11,380.38] |

#### 6.3.3.2 Network Analysis

Co-abundance of VOCs in exhaled breath (Figure 6.8) was visualised using a graph network of correlations. Clusters within the graph broadly map to the chemical structure categories defined by Dr Afolabi and Dr Koster: monoterpenes (light blue), alkanes (yellow), which includes aromatic hydrocarbons that also cluster closely to one another, aldehydes (green) and aromatic esters (purple). Based upon this network, the single unidentified compound is likely to be an aldehyde.

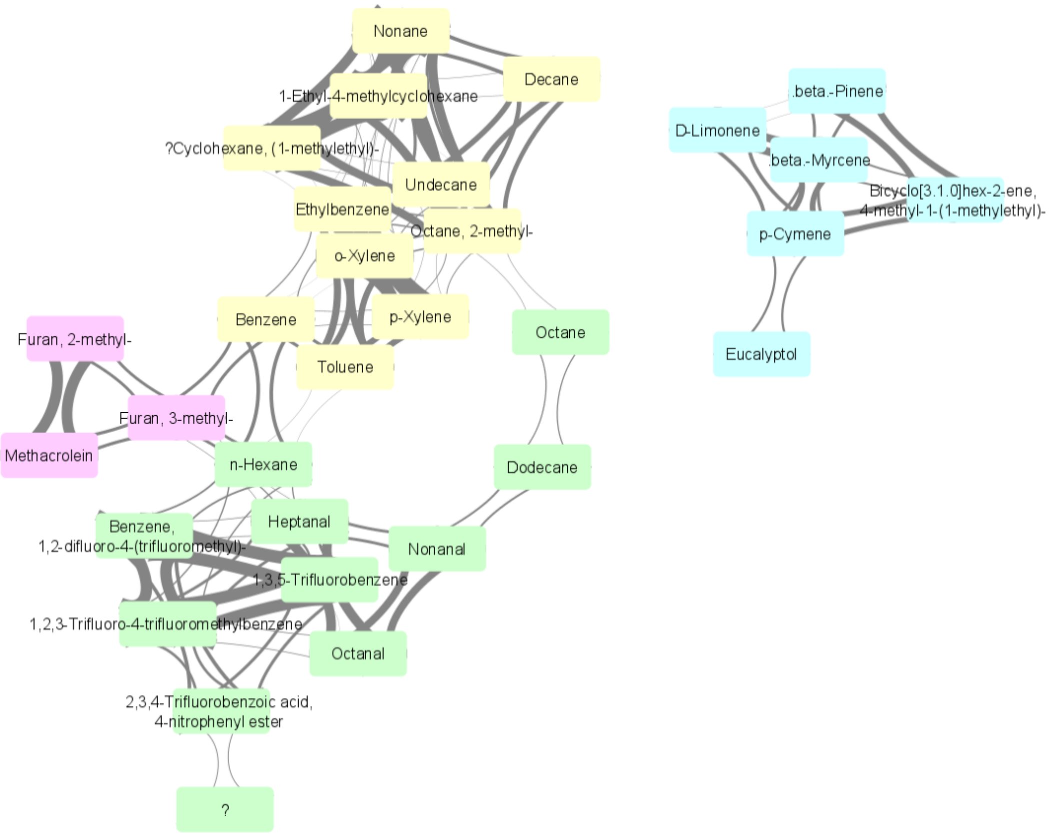


Figure 6.8 Network Representation of a Correlation Matrix of Exhaled VOCs

Each node represents a molecular feature and each line the correlation between those molecular features; the thickness of the line represents the strength of the correlation. The colour of the node represents cluster assignment

### 6.3.3.3 Principal Component Analysis

Feature reduction by Principal Component Analysis demonstrates that 66.58% of variance of variance in the original feature set was explained by the first ten principal components; 28.41% was explained by the first two (Figure 6.9).

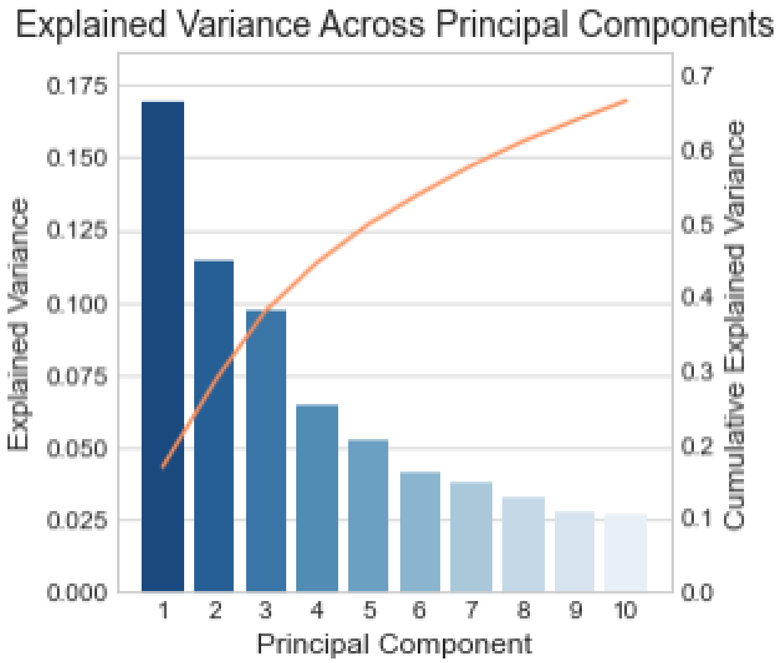
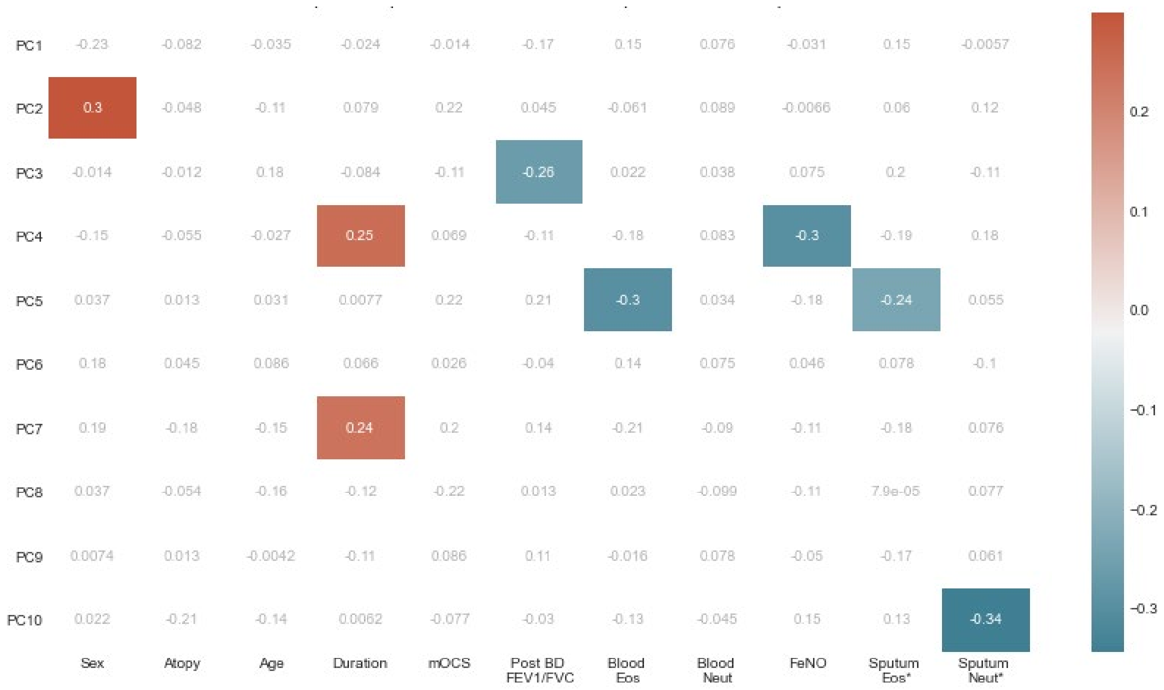


Figure 6.9 Variance Explained by First 10 Principal Components of the Cross Sectional Data

PC1 captures the majority of variance in the VOC dataset, however it does not appear to correlate to any salient clinical characteristics. Sex correlates with second principal component. Subsequent principal components relate to oral corticosteroid use, duration of asthma disease; PC4 and PC5 correlate with objective markers of T2 inflammation (Figure 6.10).



| | Sex | Atopy | Age | Duration | mOCS | Post BD FEV1/FVC | Blood Eos | Blood Neut | FeNO | Sputum Eos* | Sputum Neut* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| PC1 | -0.23 | -0.082 | -0.035 | -0.024 | -0.014 | -0.17 | 0.15 | 0.076 | -0.031 | 0.15 | -0.0057 |
| PC2 | 0.3 | -0.048 | -0.11 | 0.079 | 0.22 | 0.045 | -0.061 | 0.089 | -0.0066 | 0.06 | 0.12 |
| PC3 | -0.014 | -0.012 | 0.18 | -0.084 | -0.11 | -0.26 | 0.022 | 0.038 | 0.075 | 0.2 | -0.11 |
| PC4 | -0.15 | -0.055 | -0.027 | 0.25 | 0.069 | -0.11 | -0.18 | 0.083 | -0.3 | -0.19 | 0.18 |
| PC5 | 0.037 | 0.013 | 0.031 | 0.0077 | 0.22 | 0.21 | -0.3 | 0.034 | -0.18 | -0.24 | 0.055 |
| PC6 | 0.18 | 0.045 | 0.086 | 0.066 | 0.026 | -0.04 | 0.14 | 0.075 | 0.046 | 0.078 | -0.1 |
| PC7 | 0.19 | -0.18 | -0.15 | 0.24 | 0.2 | 0.14 | -0.21 | -0.09 | -0.11 | -0.18 | 0.076 |
| PC8 | 0.037 | -0.054 | -0.16 | -0.12 | -0.22 | 0.013 | 0.023 | -0.099 | -0.11 | 7.9e-05 | 0.077 |
| PC9 | 0.0074 | 0.013 | -0.0042 | -0.11 | 0.086 | 0.11 | -0.016 | 0.078 | -0.05 | -0.17 | 0.061 |
| PC10 | 0.022 | -0.21 | -0.14 | 0.0062 | -0.077 | -0.03 | -0.13 | -0.045 | 0.15 | 0.13 | -0.34 |

Figure 6.10 Heatmap of correlation between clinical characteristics and the first 10 principal components from the Cross Sectional Data.

Non-significant correlations in grey/white, positive and negative correlations with p-value < 0.05 in red and blue, respectively. * Sputum Eosinophils and Sputum Neutrophil Percentages log transformed. Volatile organic compounds (VOCs), principal component (PC), maintenance oral corticosteroid treatment dose (mOCS), post bronchodilator ratio between forced expiratory volume in 1 s and forced vital capacity (Post BD FEV1/FVC), eosinophils (Eos), neutrophils (Neut), Fraction of exhaled Nitric Oxide (FeNO).

### 6.3.3.4 Unsupervised Clustering on VOCs

Silhouette width indicates the optimal number of clusters in this dataset is 2, corroborated by the gap statistic and total within sum of squares (Figure 6.11). As per previous analysis, Hierarchical clustering and Partition Around the Medioids were explored as clustering techniques. There was minimal cluster consensus however (rand index 0.53). The Calinski Harabaz Index, a measure of cluster tightness, was similar between the two clustering techniques (9.02 and 10.84, respectively) indicating similar clustering performance.

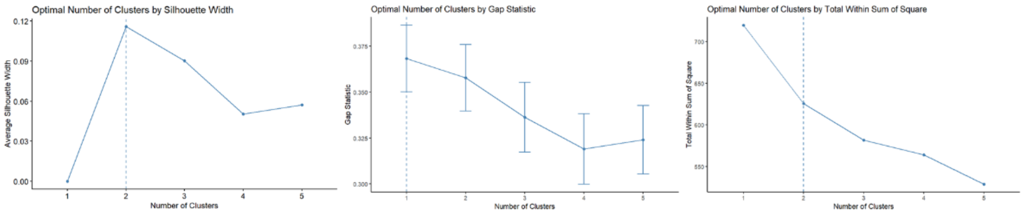


Figure 6.11 Measures of Optimal Number of Clusters in VOC Data.

A. Silhouette Width, B Gap Statistic, C Total Within Sum of Squares

Comparison between the PAM defined clusters (which had the higher Calinski Harabaz Index) indicated that there were few clinical and physiological differences between the two similarly sized clusters (Table 6.4). There were significant differences in spirometry; post bronchodilator spirometry was more impaired in Cluster 2 than Cluster 1 (p=0.043). This difference was accompanied by trends towards loss of diversity and increased exacerbation frequency in Cluster 2.

Table 6.4 Clinical Characteristics Across VOC Clusters Continuous variables expressed as median [Q1, Q3] with differences measured by Mann-Whitney U test. Categorical variables expressed as n (%) with differences measured by chi-square test. Abbreviations: GORD, gastro-oesophageal reflux disease; ICS, inhaled corticosteroid; BDPe, beclomethasone dose equivalent; OCS, oral corticosteroids; IgE, Immunoglobulin E; IL-5, Interleukin 5; ACQ, asthma control questionnaire, HADSTOT, Hospital Anxiety and Depression Total Score; SNOT, SinoNasal Outcome Score; FeNO, fraction of nitric oxide in exhaled breath; post BD, post bronchodilator; FEV1, forced expiratory volume in 1 second; FVC, forced vital capacity; FEF25-75%, forced expiratory flow at 25% to 75% of FVC;

| | Cluster 1<br>(n=44) | Cluster 2<br>(n=30) | P-Value |
|---|---|---|---|
| Female Sex | 21 (47.7) | 12 (40.0) | 0.676 |
| Age | 55.0 [49.0,62.2] | 58.0 [50.5,66.5] | 0.7 |
| BMI | 27.8 [25.8,31.0] | 29.0 [23.9,36.7] | 0.745 |
| Never Smoker | 33 (75.0) | 20 (66.7) | 0.735 |
| Atopic | 30 (68.2) | 15 (50.0) | 0.183 |
| GORD | 21 (47.7) | 14 (46.7) | 0.883 |

| | Cluster 1<br>(n=44) | Cluster 2<br>(n=30) | P-Value |
|---|---|---|---|
| Nasal Polyps | 14 (32.6) | 4 (13.8) | 0.127 |
| Age of Onset | 15.0 [3.5,38.5] | 14.0 [3.0,49.5] | 0.956 |
| Exacerbations in the Last 12 Months | 1.0 [0.0,3.0] | 3.0 [1.0,4.0] | 0.062 |
| mOCS | 18 (40.9) | 14 (46.7) | 0.801 |
| ICS (BDPe) | 3000.0 [2814.0,3840.0] | 3000.0 [2575.0,3980.0] | 0.969 |
| Anti IgE | 5 (11.4) | 4 (13.3) | 1 |
| Anti IL-5 | 2 (4.5) | 5 (16.7) | 0.112 |
| ACQ6 | 2.6 [1.4,3.3] | 2.7 [2.1,3.3] | 0.464 |
| HADS | 10.0 [6.0,13.0] | 11.5 [7.0,16.8] | 0.193 |
| SNOT20 | 29.5 [20.8,44.2] | 35.0 [25.5,48.5] | 0.32 |
| FeNO | 29.0 [17.5,47.5] | 28.0 [17.8,45.8] | 0.982 |
| PostBD FEV1 (% predicted) | 79.2 [61.8,93.6] | 64.5 [48.9,88.2] | 0.043 |
| PostBD FVC (% predicted) | 92.7 [81.2,101.2] | 89.7 [72.3,104.8] | 0.324 |
| PostBD FEV1/FVC | 66.5 [56.8,79.2] | 63.0 [52.5,70.0] | 0.211 |
| PostBD FEF25-75% (% predicted) | 51.7 [32.6,85.0] | 39.4 [19.1,60.9] | 0.025 |
| FEV1 Reversibility | 10.9 [3.4,20.5] | 12.7 [8.7,17.1] | 0.709 |
| Blood Neutrophils | 5.2 [4.2,6.3] | 5.4 [4.7,7.0] | 0.481 |
| Blood Eosinophils | 0.3 [0.1,0.4] | 0.2 [0.1,0.5] | 0.907 |
| Serum Total IgE | 159.6 [31.0,366.4] | 131.8 [24.8,279.6] | 0.817 |
| Sputum Neutophils | 49.9 [24.0,76.3] | 45.6 [22.6,70.8] | 0.886 |
| Sputum Eosinophils | 2.7 [0.4,11.8] | 3.7 [0.5,15.8] | 0.72 |
| Sputum MPO | 1134.0 [489.4,2100.5] | 979.6 [603.4,1844.5] | 0.984 |
| Sputum NE | 6.1 [3.4,11.7] | 8.6 [2.9,11.8] | 0.932 |
| Sputum EDN | 656.4 [149.1,1133.0] | 474.2 [184.1,1471.5] | 0.527 |
| Sputum ECP | 555.2 [92.0,1543.0] | 916.0 [144.6,1747.5] | 0.43 |
| Observed α Diversity | 137.0 [103.0,147.5] | 100.0 [72.0,146.0] | 0.158 |
| Chao α Diversity | 153.5 [119.9,167.8] | 113.1 [89.0,161.0] | 0.118 |
| Shannon α Diversity | 3.3 [2.9,3.6] | 3.1 [2.5,3.5] | 0.276 |
| Simpson α Diversity | 0.9 [0.9,1.0] | 0.9 [0.8,0.9] | 0.2 |

### 6.3.4 Exploratory Analysis of Repeat Data Samples

Repeat data (n=70) was compiled from samples from 14 patients in the repeated sampling arm (n=14) of the study who each provided 5 samples.

#### 6.3.4.1 Principal Component Analysis

Feature reduction by Principal Component Analysis demonstrates that 76.89% of variance of variance in the original feature set was explained by the first ten principal components; 32.75% was explained by the first two (Figure 6.12).

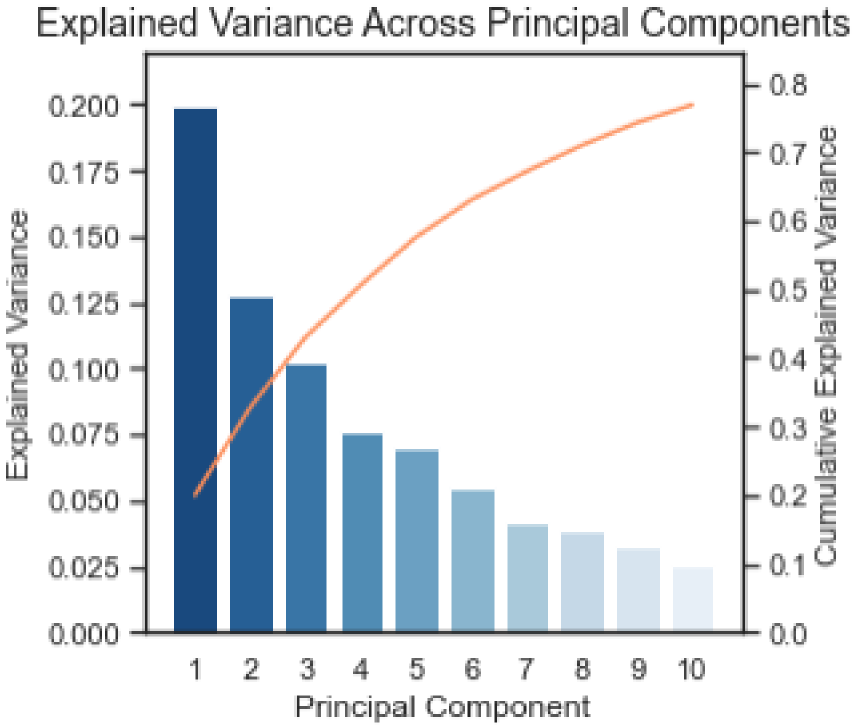


Figure 6.12 Variance Explained by first 10 Principal Components

Subject ID was correlated with the first three principal components but visit number (i.e. day of the week) did not correlate with any of the first ten principal components (Figure 6.13). Contrasting the cross-sectional PCA, the first principal component also correlates with typical T2 characteristics, such as atopy, FeNO, sputum eosinophils and obstructed lung function obstruction. This is simply a confounder correlating with Subject ID (50% of these patients had sputum eosinophilia).

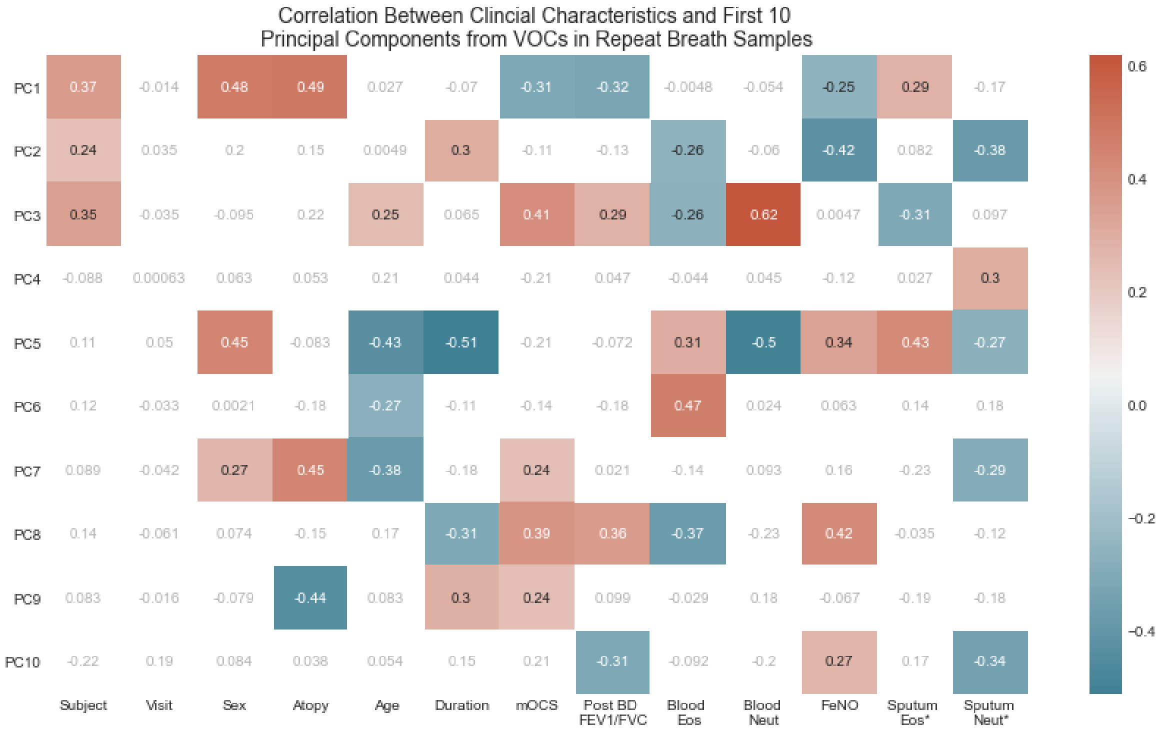


Figure 6.13 Heatmap of correlation between clinical characteristics and the first 10 principal components.

Non-significant correlations in grey/white, positive and negative correlations with p-value < 0.05 in red and blue, respectively. * Sputum Eosinophils and Sputum Neutrophil Percentages log transformed. Volatile organic compounds (VOCs), principal component (PC), maintenance oral corticosteroid treatment dose (mOCS), post bronchodilator ratio between forced expiratory volume in 1 s and forced vital capacity (Post BD FEV1/FVC), eosinophils (Eos), neutrophils (Neut), Fraction of exhaled Nitric Oxide (FeNO).

A PCA plot of breath samples (visualising the first two principal components) corroborates the findings from the heatmap (Figure 6.10) that breath samples from the same patient are closely related. but do show some within-subject variability, as illustrated by the ellipses (Figure 6.14). The size of each ellipsis (representing an individual subject) relative to the spread of all breath samples illustrates that within-subject variability is a fraction of the variability seen across all breath samples. The ellipses are closely connected and, in most cases, overlap, indicating that breath samples from different individuals share some characteristics. When constructing ellipses to represent the day of the Visit, we see no separation (Figure 6.15).

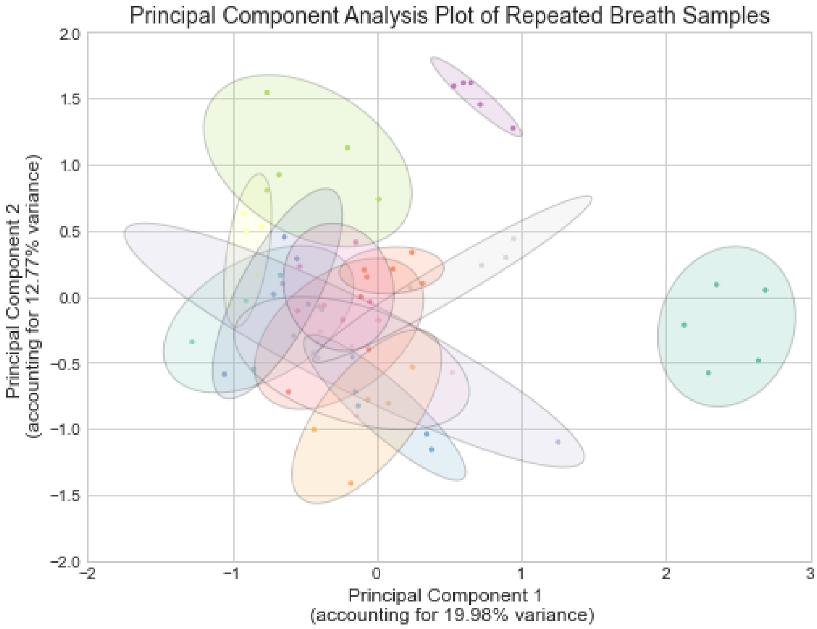


Figure 6.14 PCA plot of all 70 breath samples with ellipses representing subject identifiers (n = 14)

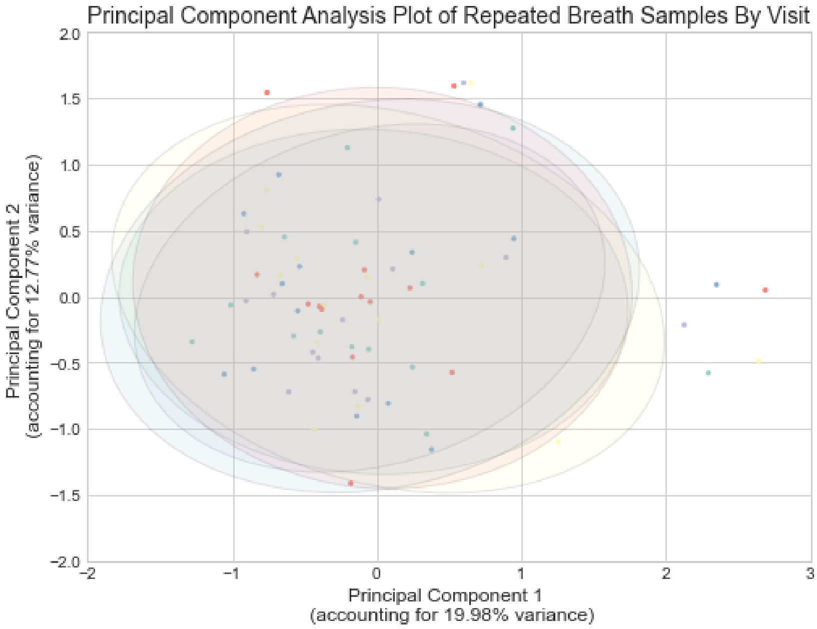


Figure 6.15 PCA plot of all 70 breath samples (5 samples from 14 subjects) with ellipses representing the day of the week on which the sample was collected

### 6.3.4.2 Clustering

When hierarchical clustering of the 70 breath samples was applied to identify 14 clusters, 11 clusters contained all five breath samples from the same patient; the remaining three cluster contained a combination of breath samples from different patient (Figure 6.16)

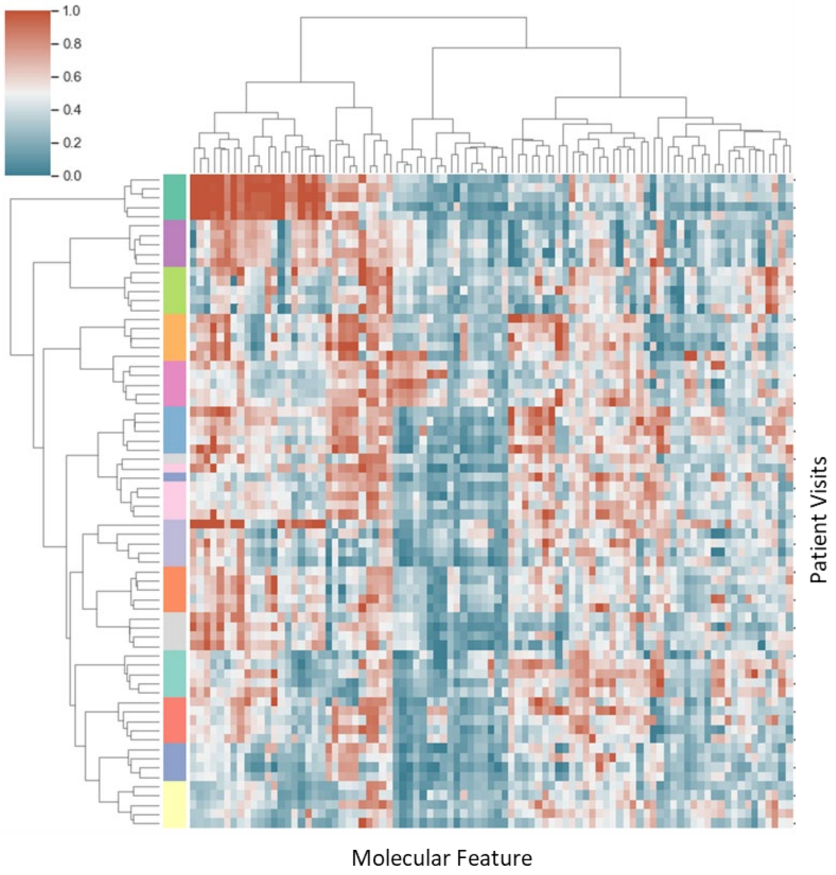


Figure 6.16 Heatmap of VOC Abundance in 80 Breath Samples From 14 Patients.

Columns represent VOCs, rows represent patient visits. Rows ordered according to hierarchical clustering of molecular features using Ward's method on Euclidean distances. Colour bar on y axis indicates subject ID (i.e. blocks of colours indicate that visits from the same patient have clustered together).

### 6.3.5 Within Subject Variability

Cross-sectional analysis allows for identification of VOCs that vary between subjects, the data from the repeat sampling visit allows for identification of VOCs that vary within subjects.

The majority of VOCs (62, 69.66%) had a mean within-subject variation of <30% (Figure 6.17). Of these, 14 VOCs (15.73% of total) were found to be "Conserved", that is, they showed low variability within subjects and between subjects. 30.35% of VOCs (n = 27) were found to be "Erratic", that is, they showed high variability within subjects and between subjects. The remaining 53.93% of VOCs (n = 48) showed low variability within subjects but high variability between subjects ("Potential Biomarker").

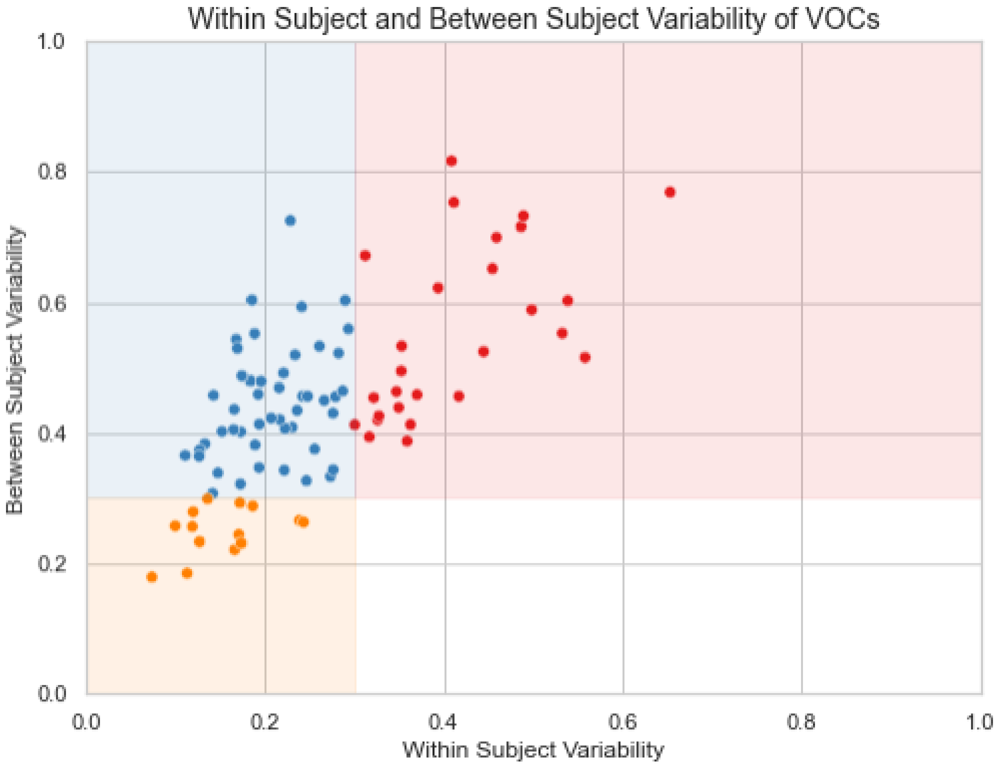


Figure 6.17 Scatterplot of within-subject variability and between-subject variability for each VOC. Cutoffs at 0.3 for both Within and Between Subject Variability

The VOC types described in Table 6.1 had broadly similar relative frequencies of "Erratic", "Conserved" and "Potential Biomarker" assignments. Of note, 52.4% of alkanes were categorised as "potential biomarkers" but, in parallel, 38.1% were categorised as "erratic". The majority of aldehydes (57.1%) were erratic.

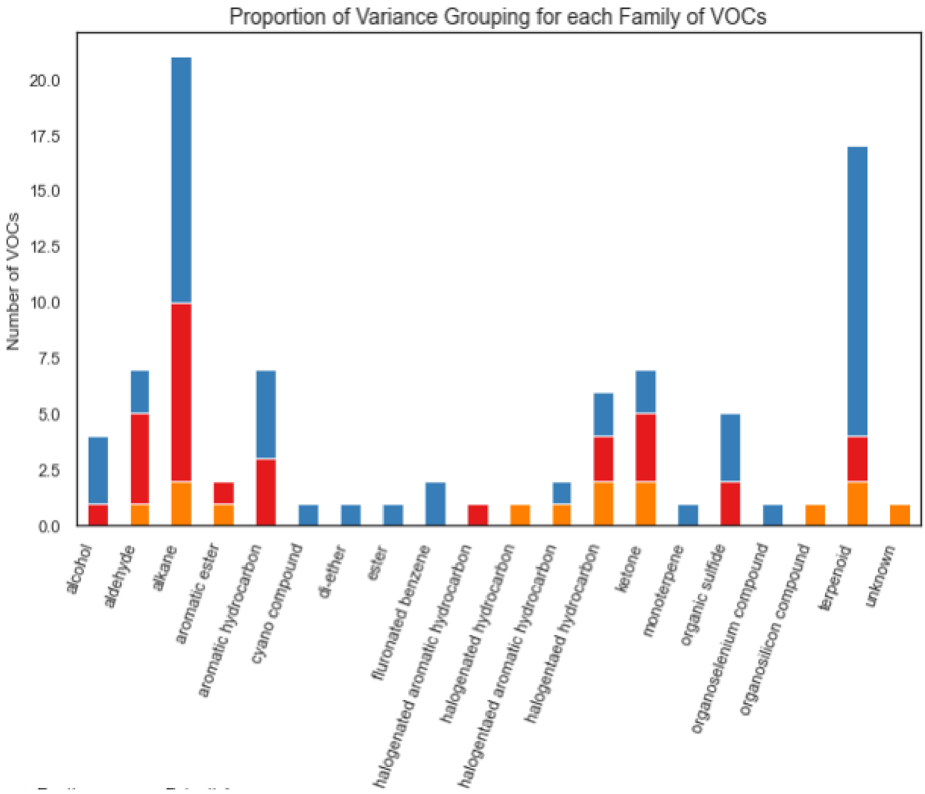


Figure 6.18 Proportion of repeatability assignments in each VOC category

### 6.3.6 Relationship With Granulocyte Counts

PCA analysis identifies some relationship between VOCs and granulocyte count and VOC abundance but this is weak (Figure 1.9). Prior to training machine learning predictors, a number of approaches were used to further understand this.

#### 6.3.6.1 Pearson's Correlation

Weak correlations are observed between VOCs and sputum granulocyte counts. The strongest correlation with sputum eosinophil was negative: 1,4 dioxane (Table 6.5). There was only one significant correlation with sputum neutrophils: Undecane, 3-methyl (Table 6.6).

| | Pearson's Correlation | P Value |
|---|---|---|
| 1,4-Dioxane | -0.330 | 0.004 |
| 2-Butanone | -0.320 | 0.006 |
| Pentane, 3-methyl- | 0.251 | 0.031 |

| | Pearson’s Correlation | P Value |
|---|---|---|
| Phenol | 0.249 | 0.033 |
| 1-Propene, 1-(methylthio)-, (E)- | 0.245 | 0.035 |
| n-Hexane | -0.236 | 0.043 |
| Bicyclo[3.1.0]hex-2-ene, 4-methylene-1-(1-methylethyl)- | 0.235 | 0.044 |
| Sulfide, allyl methyl | 0.233 | 0.046 |

Table 6.5 Significant Correlations Between log transformed Sputum Eosinophils and VOC Abundance

| | Pearson’s Correlation | P Value |
|---|---|---|
| Undecane, 3-methyl | -0.341 | 0.003 |

Table 6.6 Significant Correlations Between log transformed Sputum Neutrophils and VOC Abundance

#### 6.3.6.2 Differential Abundance

Differential abundances of VOCs between sputum eosinophilic and sputum neutrophilic patients, when adjusted for sex and spirometry, identified only minor log fold changes with p values only reaching significance without correcting for false discovery. 1-Hexanol, 2-ethyl- was decreased in eosinophilic patients, whilst Benzene and Thiophene, 3-methyl- were increased (Figure 6.19). Heptanal, .beta.-Pinene and identified (suspected aldehyde (Figure 6.8)) MF were increased in neutrophilic patients (Figure 6.19).

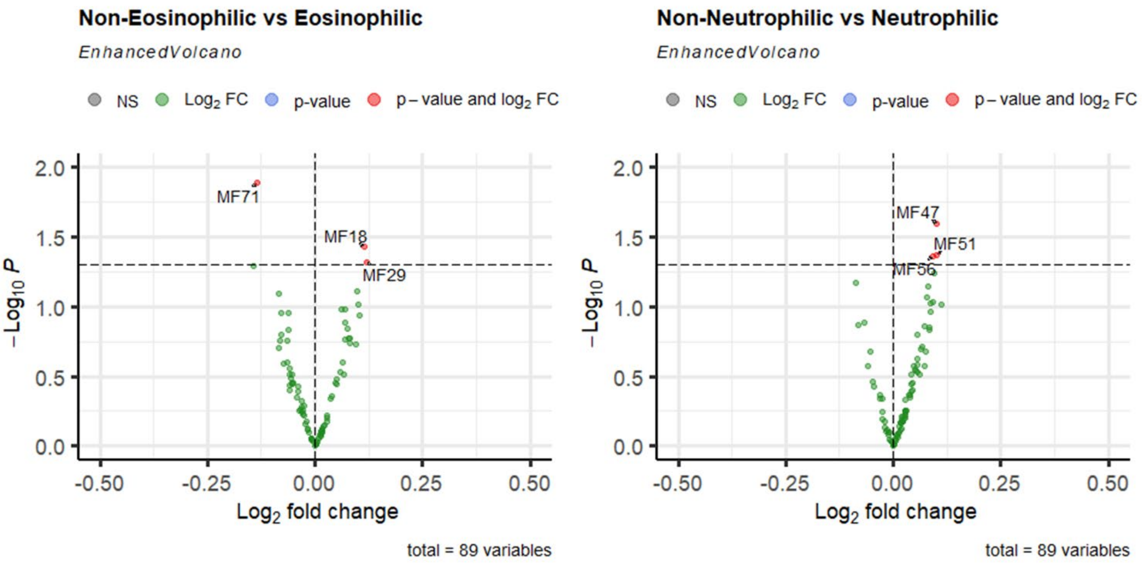


Figure 6.19 Volcano Plots of Differential Abundance of VOCs in inflammatory Phenotypes corrected for Sex and FEV1.

A = Non-eosinophilic vs Eosinophilic. B = Non-neutrophilic vs Neutrophilic. MF71 = 1-Hexanol, 2-ethyl-, MF18 = Benzene, MF29 = Thiophene, 3-methyl-, MF 47 = Heptanal, MF51 = No NIST identification, MF56 = .beta.-Pinene

### 6.3.6.3 Principal Component Analysis

A PCA plot of breath samples in the cross-sectional data (visualising the first two principal components) corroborates shows no separation between sputum eosinophilic and non-eosinophilic patients (Figure 6.20) nor between sputum neutrophilic and non-neutrophilic patients (Figure 6.21).

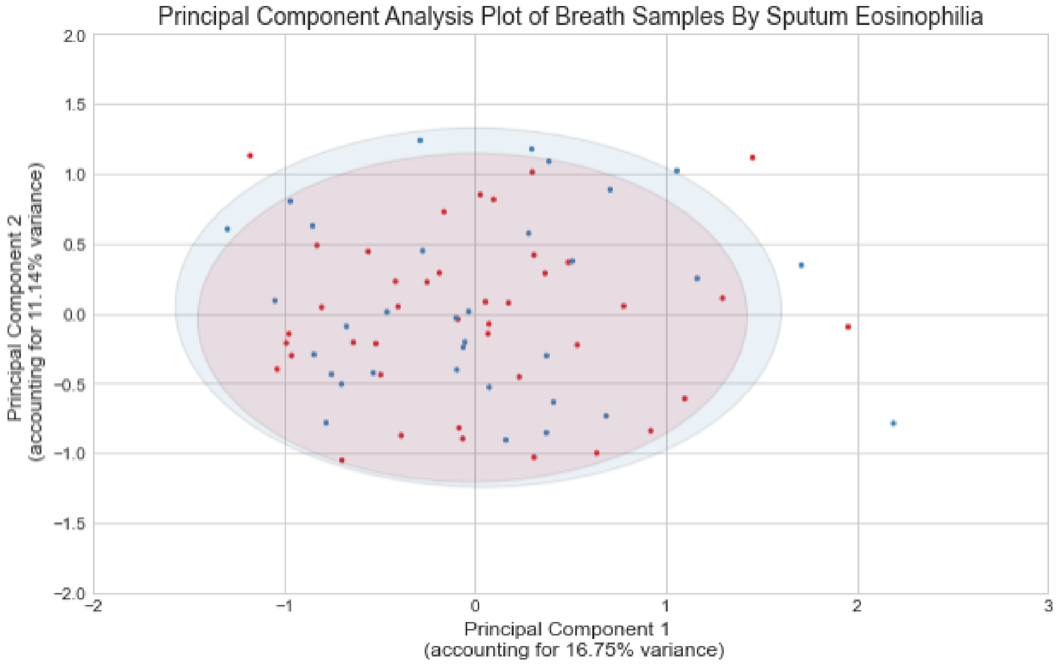


Figure 6.20 PCA Plot of Breath Samples Separated by Sputum Eosinophils >2%.

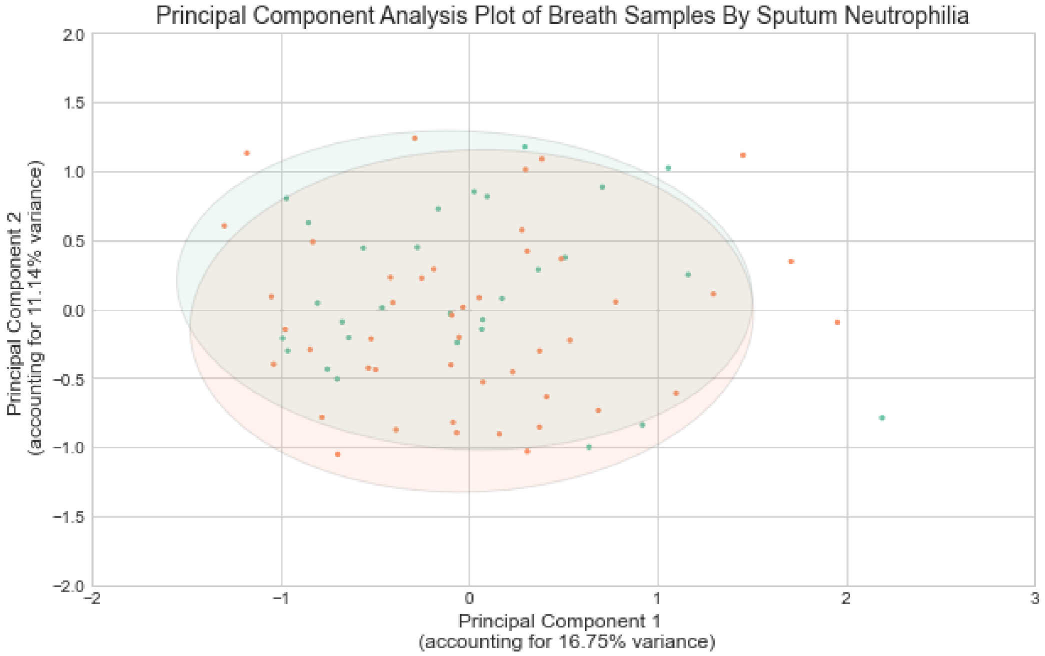


Figure 6.21 PCA Plot of Breath Samples Separated by Sputum Neutrophils >61%.

### 6.3.7 Machine Learning Classifier

An XGBoost classifier was trained on all 89 VOCs in the training set and assessed on the test set. A model using default settings achieved an AUC for predicting sputum eosinophilia (≥2%) of 0.357 (Figure 6.22); following hyper-parameter optimisation, this improved to 0.429 (Figure 6.23).

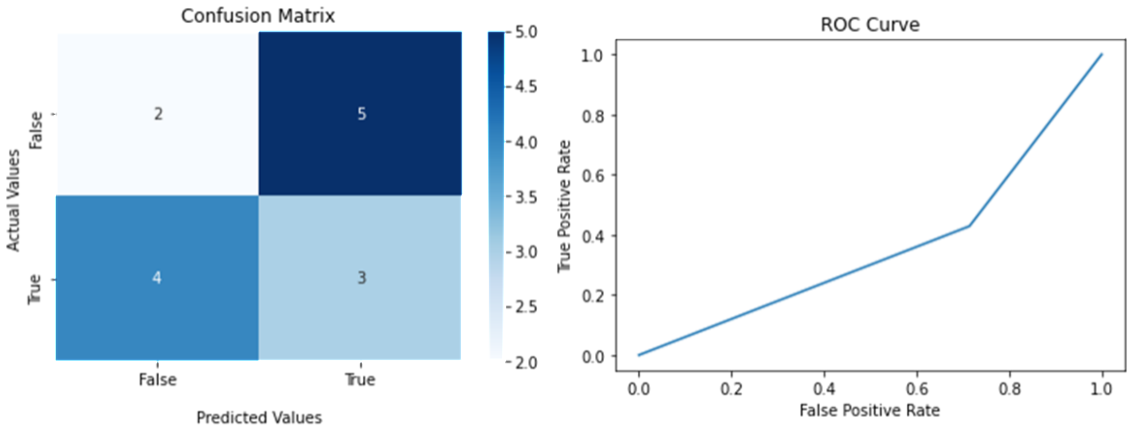


Figure 6.22 Confusion Matrix and ROC Curve for XGB Classifier Trained on All VOCs

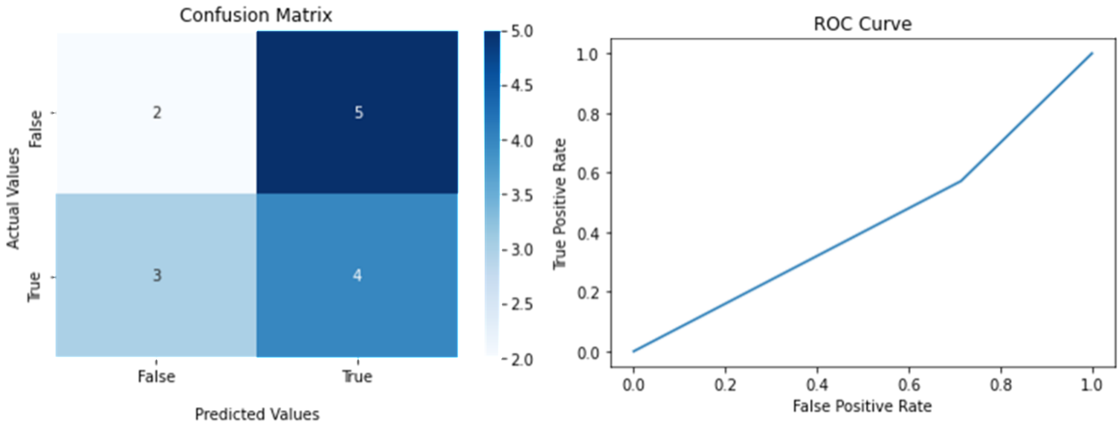


Figure 6.23 Confusion Matrix and ROC Curve for Hyperparameter Optimised XGB Classifier Trained on All VOCs

From the 89 features, the most important features (determined by 'gain') were selected (Figure 6.24). A classifier trained on these features achieved an AUC for predicting sputum eosinophilia (≥2%) of 0.5 (Figure 6.25); following hyper-parameter optimisation, this dropped to 0.429 (Figure 6.26).

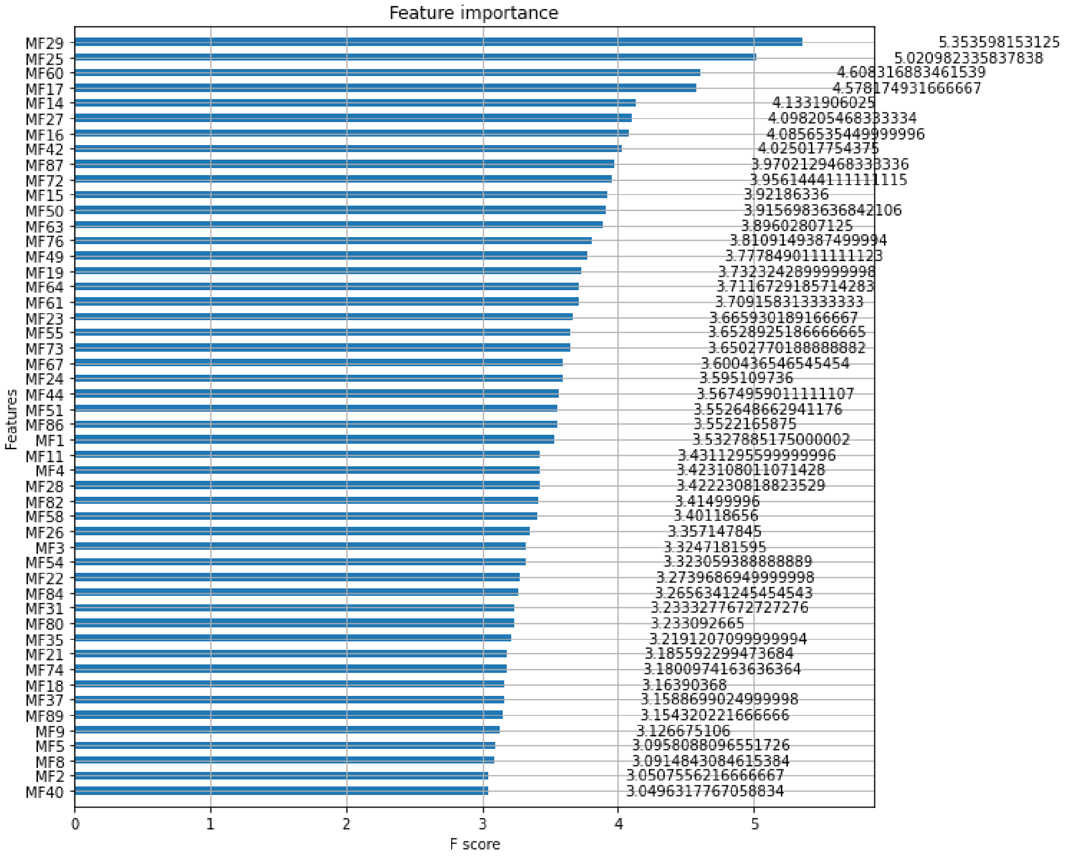


Figure 6.24 Feature Importance for All VOCs in the XGB Classifier

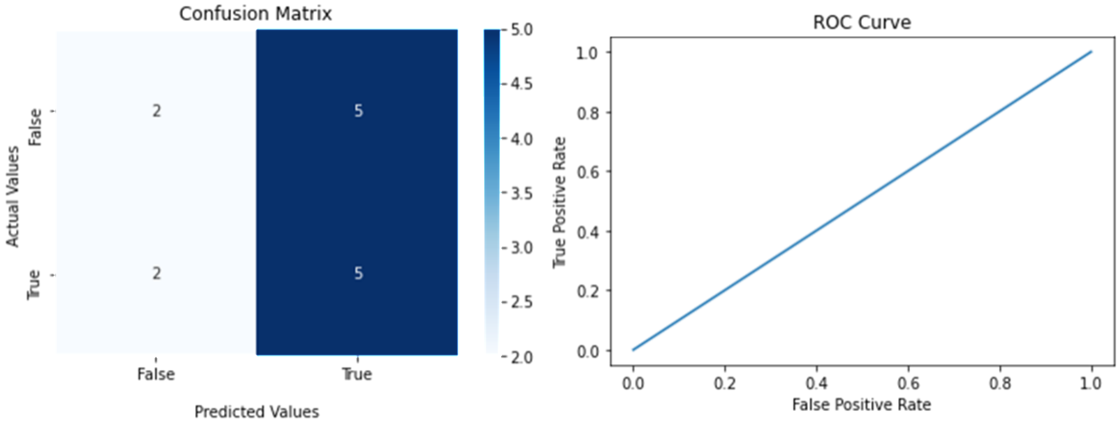


Figure 6.25 Confusion Matrix and ROC Curve for XGB Classifier Trained on Important VOCs Identified from All VOCs

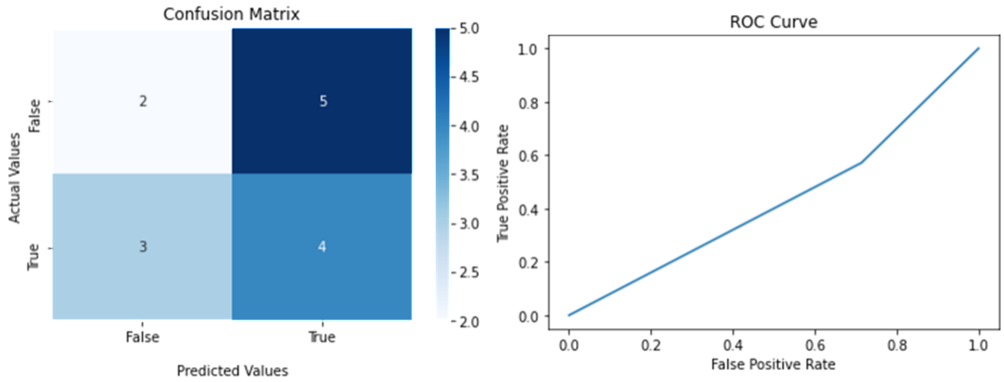


Figure 6.26 Confusion Matrix and ROC Curve for Hyperparameter Optimised XGB Classifier Trained on Important VOCs Identified from All VOCs

In summary, although there is evidence on PCA that VOCs are sensitive to sputum eosinophilia, this appears to be weak. Consequently, in this analysis, machine learning classifiers are unable to predict the sputum eosinophilic phenotype in an independent test cohort.

### 6.3.8 Restricting to Stable Features

Using the within-subject variation findings described in 1.3.5, the clustering approach described in 1.3.4.2 was repeated on the 70 breath samples from the 14 patients, excluding erratic VOCs. When using all VOCs, 11 clusters were perfect (cluster composed of all samples from a single patient, (Figure 6.16)); when restricting to non-erratic VOCs, this improves to 12 (Figure 6.27).

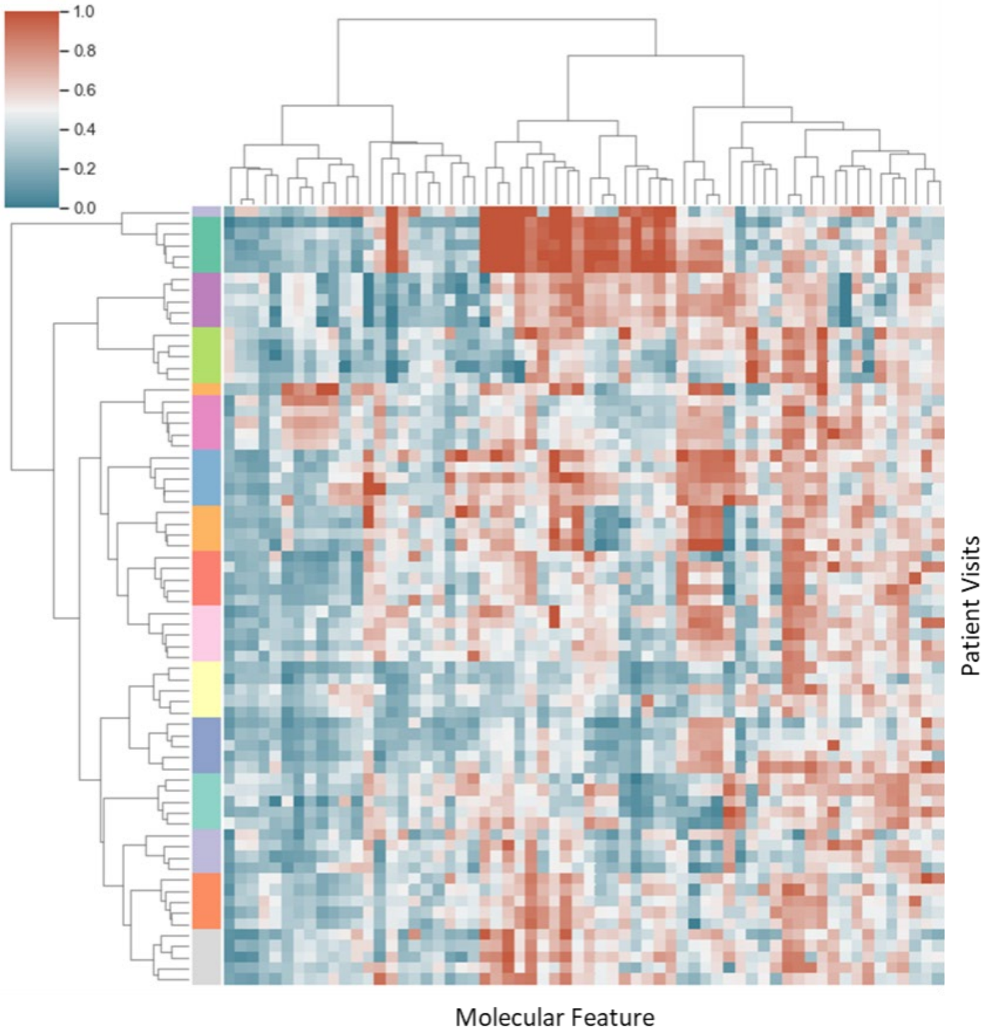


Figure 6.27 Heatmap of VOC Abundance in 80 Breath Samples From 14 Patients restricted to 62 features that were not erratic.

Columns represent VOCs, rows represent patient visits. Rows ordered according to hierarchical clustering of molecular features using Ward's method on Euclidean distances. Colour bar on y axis indicates subject ID (i.e. blocks of colours indicate that visits from the same patient have clustered together).

#### 6.3.8.1 Machine Learning Classifier

An XGBoost classifier was trained on the non-erratic 62 VOCs in the training set and assessed on the test set. A model using default settings achieved an AUC for predicting sputum eosinophilia

(>2%) of 0.286 (Figure 6.28); following hyper-parameter optimisation, this dropped to 0.214 (Figure 6.29).

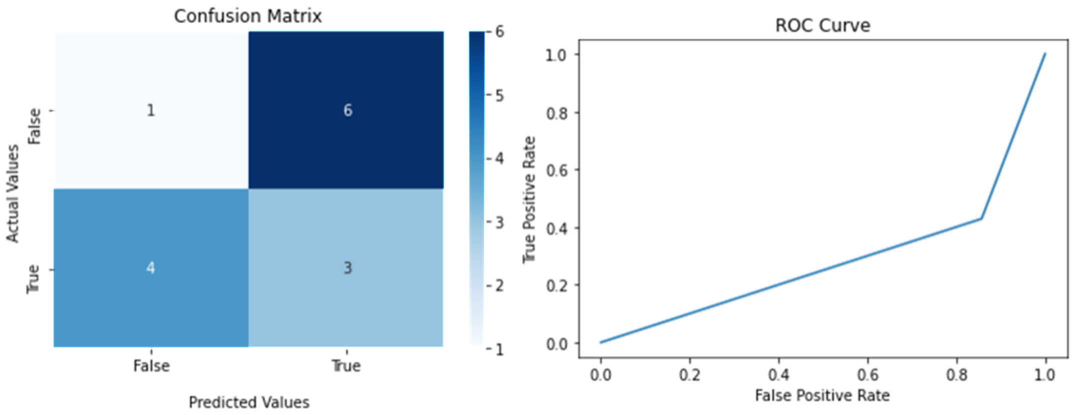


Figure 6.28 Confusion Matrix and ROC Curve for XGB Classifier Trained on All VOCs

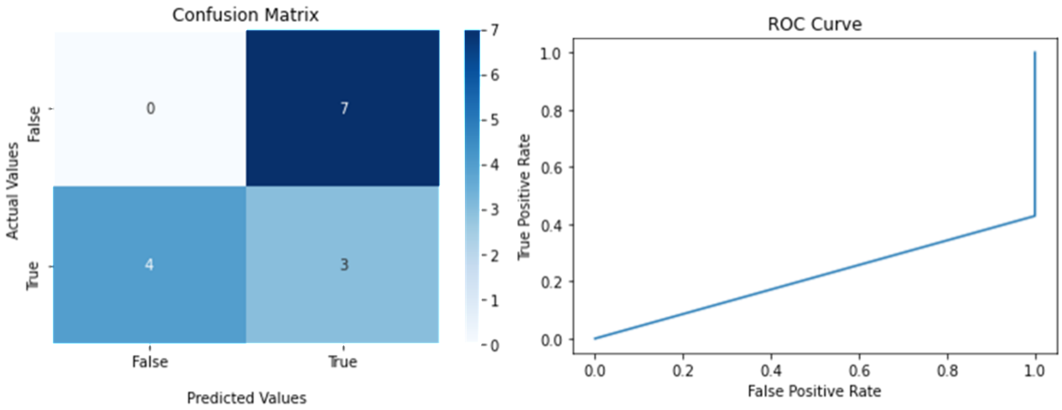


Figure 6.29 Confusion Matrix and ROC Curve for Hyperparameter Optimised XGB Classifier Trained on All VOCs

From the 62 features, the most important features (determined by 'gain') were selected (Figure 6.30). A classifier trained on these features achieved an AUC for predicting sputum eosinophilia (>2%) of 0.571 (Figure 6.31); following hyper-parameter optimisation, this improved to 0.643 (Figure 6.32).

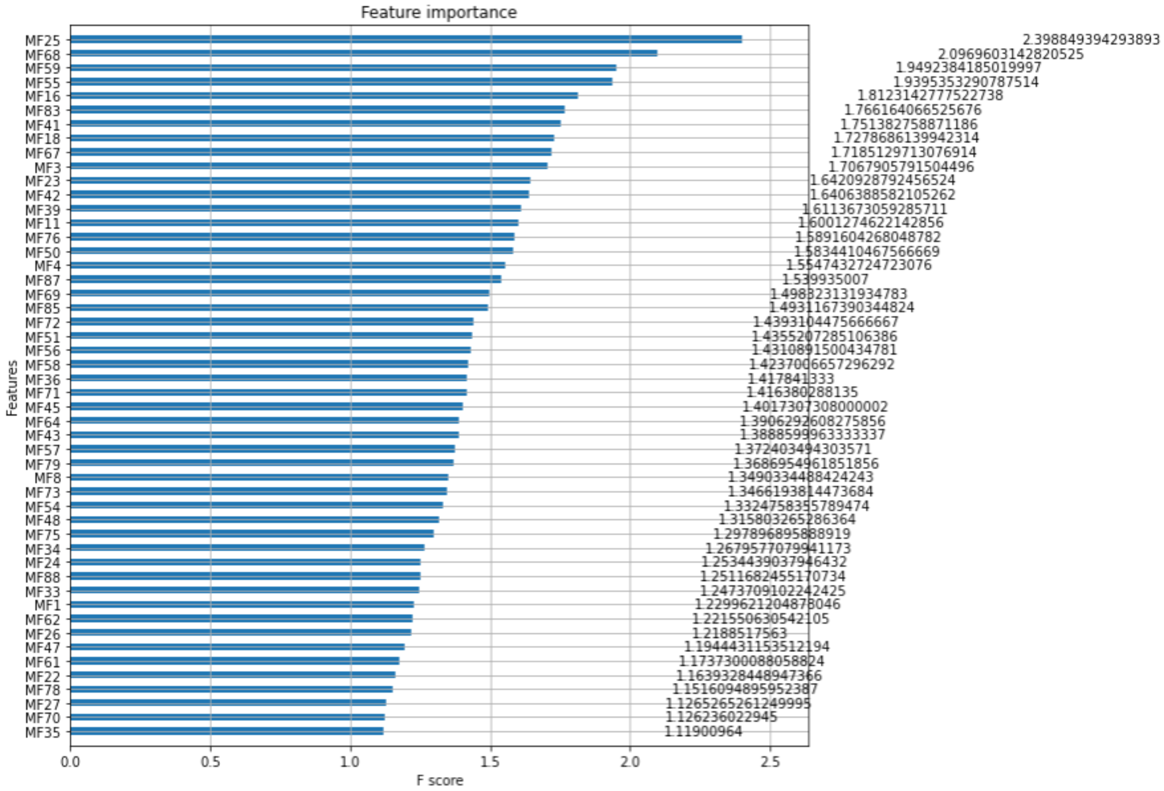


Figure 6.30 Feature Importance for All VOCs in the XGB Classifier

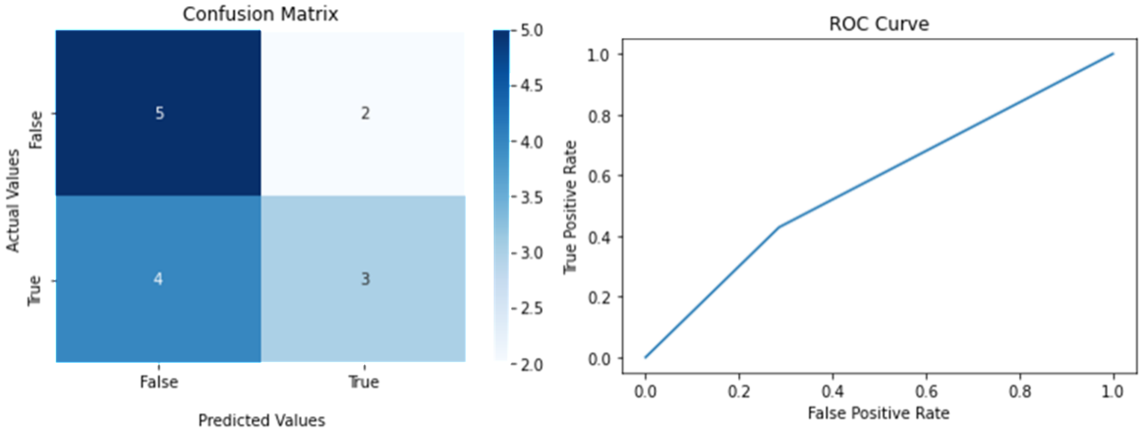


Figure 6.31 Confusion Matrix and ROC Curve for XGB Classifier Trained on Important VOCs Identified from All VOCs

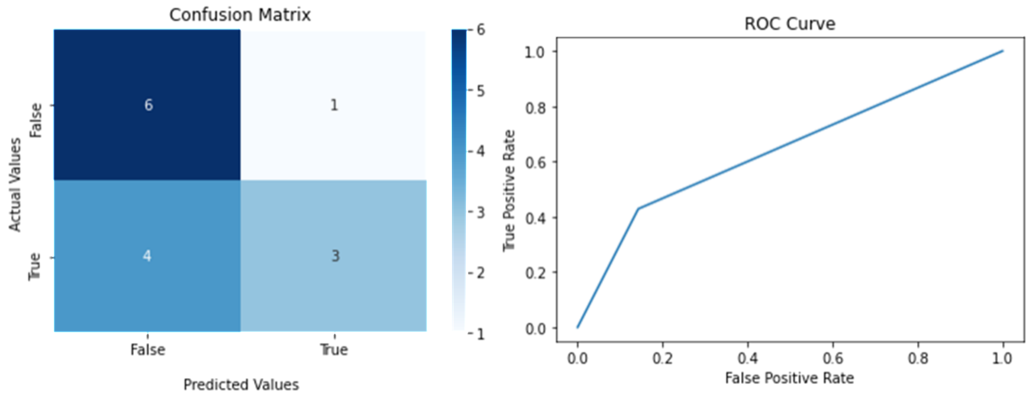


Figure 6.32 Confusion Matrix and ROC Curve for Hyperparameter Optimised XGB Classifier Trained on Important VOCs Identified from All VOCs

The final model had an accuracy of 0.64, sensitivity 0.43, specificity 0.86, f1 score of 0.57, PPV 0.86 and NPV 0.60. The most important VOCs to contribute to the most successful model (Hyperparameter Optimised XGB Classifier Trained on Important VOCs Identified from All VOCs) include butane, 2-methyl and pentane (Table 6.7).

| VOC | Feature Importance |
|---|---|
| **2-Pentanone** | 0.288 |
| **4-Heptanone** | 0.264 |
| **Octane, 1-chloro-** | 0.240 |
| **2,3,4-Trifluorobenzoic acid, 4-nitrophenyl ester** | 0.208 |

Table 6.7 Feature Importance of VOCs used in the Hyperparameter Optimised XGB Classifier Trained on Important VOCs Identified from All VOCs

## 6.4 Discussion

The VOCs identified in this cohort are consistent with those reported by other studies[336] and otherwise broadly consistent with the established literature, which commonly describes aldehydes, aromatic hydrocarbons and ketones[337]. The breath samples in our cohort were dominated, in absolute terms by aromatic esters. These compounds have been associated with asthma and furan based cyclic compounds have been identified as possible biomarkers of asthma [338]. This would explain it’s high abundance in this cohort, though this interpretation is limited by a lack of healthy control group.

### 6.4.1 Cautious Interpretation of Co-abundance analysis

The co-abundance analysis provides a very basic quality control check for the study. The graph illustrated is rich with nodes and edges, demonstrating that there is a structure to the exhaled breath profile across all the patients. This is best illustrated by considering an opposing finding: if no statistically significant correlations were observed, this might indicate that exhaled VOC profiles are totally inconsistent from patient to patient. This would undermine any attempt at searching this matrix for a biomarker.

More detailed interpretation of the co-abundance analysis should be done cautiously. Simply, this demonstrates that the abundance of VOCs which have been identified as structurally similar are correlated. One explanation might relate to the fact that putative identifications were made against the NIST library but not confirmed against pure chemical standards: VOC identifications were made if there was >85% match and so structurally similar compounds could be mistaken for one another.

If the putative identifications are accurate then the analysis indicates there might be value in describing exhaled breath in less granular terms i.e., by VOC class rather than VOC. The Proteobacteria:Firmicutes ratio is a summary statistic of the airway microbiome [327] that shows potential utility, despite only capturing Phylum level data (as opposed to Genus or Species levels of identification that are possible by 16S or metagenomic approaches respectively). This would, for all intents and purposes, be a form of feature reduction. To take the microbial comparison further, future analyses of breath volatile might also be explore summary measures such as diversity.

### 6.4.2 Repeat Analysis Identifies Consistency in Breath VOC Profiles

Breath sampling by the ReCIVA device has been successfully used in acutely breathless patients [339] indicating it to be easy to perform. Nevertheless, to our knowledge, this study is the first to report that repeated breath sampling (more than twice) is feasible in a severe asthma cohort. This is not surprising[260] but affirms one of the oft quoted advantages of breathomics for airways disease.

As with the co-abundance analysis of cross-sectional samples, PCA analysis of the repeated samples serve as a very basic quality control check for the study. Modern dimension reduction techniques, like t-SNE and UMAP, have become popular, due to their use of non-linear techniques in order to maximise local structure separation. PCA was preferred in this analysis because it maintains global structure; we are able to visualise that intra-patient variability of breath samples is less then inter-variability of breath samples. We also observe that some patients are more closely related to one another than others indicating that there is similarity and heterogeneity in exhaled breath profiles.

### 6.4.3 Heterogeneity in Within Subject Variance

Breathprints from e-nose shows some between day repeatability [137 340]. However, these were from just two samples taken, on average, 14 days part. To our knowledge, this is the first report of individual VOCs measured by GCMS to be reported from multiple visits. The within subject variability of VOCs is highly heterogenous (Figure 6.22). As discussed in Chapter 6.1, a number of factors can contribute to this variability. It is possible that the heterogeneity may also reflect analytical sensitivity or indeed true biological variation but would regardless lead to excessive noise that would drown out signals of interest.

Hierarchical clustering of repeat breath samples demonstrates that restricting to non-erratic VOCs does lead to detrimental signal loss: similarities in breath samples from the same patient was identified in the reduced feature set. Moreover, the predictive performance of supervised machine learning models was improved, when compared to a full feature set. It should be noted that a mean cutoff of 30% coefficient of variation is very lenient and would not be acceptable in other fields. The purpose of this analysis was to reduce noise rather than eliminate it altogether. VOCs are inherently volatile and stringent cutoffs would have been prohibitive for biomarker discovery. Ultimately, it is unlikely that any clinically translatable VOC biomarker will based on a single VOC, rather it is likely to be a panel of VOCs. As such it is not the variability of individual VOCs that is important but the variability of that signature. Furthermore, once a signature is identified, the methodology can be adapted to improve the repeatability of that signature, as in the case of exhaled nitric oxide (FeNO) [80,341].

### 6.4.4 Relationship with Airway Inflammation

A variety of approaches were used to relate the VOC profile to airway inflammation but none were conclusive. Only weak correlations and differential expression were seen in relation to sputum eosinophilia. PCA indicated that T2 inflammation was correlated with the fourth and fifth PC (Figure 6.8) but this was, as with the other analyses, only weak. These findings resonate with previous reports that airway inflammation is rarely captured in the first principal component [342] VOCs important to the prediction of sputum eosinophilia were only identified through aggressive feature selection. This is consistent with the PC loading analysis, which demonstrates airway inflammation to only relate to a small amount of variation seen in exhaled breath.

The VOC most important to airway eosinophilia was 2-pentanone, a ketone. This compound has not previously, to our knowledge, been associated with airway eosinophilia, continuing the disappointing trend of failure to externally validate [343]. At least some of this is likely due to methodological heterogeneity but there is supportive evidence for our finding. Measures of 2-pentanone are higher in exhaled breath than parallel ambient air samples, indicating that that it is produced endogenously[344]. Ketones, more generally, are thought to be the product of fatty acid degradation[345] and have been found to be produced by human bronchial epithelial cells[346], indicating that they may originate from the airways (though not necessarily exclusively). These findings suggest that 2 pentanone directly relates to airway biology. Moreover, 2-pentanone differentiates COPD from healthy controls [347]. Though not strictly the same compound, 2-hexanone, a six carbon ketone, is structurally similar to 2 pentanone and has been identified in eosinophilic asthma patients [348].

### 6.4.5 Limitations

Ultimately, particularly when applying machine learning tools, this analysis is limited by a small sample size. Nevertheless, all patients were well characterised and representative of a population in which inflammatory phenotyping is clinically relevant. It would have been ideal for the analysis of between day variability for feature selection to have been performed on a dataset independent to the test set, however, due to the stringent pre-processing steps taken, the degree of data leakage is likely to minimal.

The classification performance of classifier is not as strong as that observed in other studies [144 145], in part, due to our commitment to a separate test cohort (as opposed to internal validation strategies) despite limited numbers. Of course, an AUC of 0.7 is not dissimilar to that of blood eosinophils and FeNO (3.5.2 Predicting Sputum Inflammatory Phenotypes with Clinical Biomarkers) and some of this limitation may relate to our hypothesis that the target variable is of too little resolution.

### 6.4.6 Conclusions

In conclusion, this analysis demonstrates that there is a clear structure to the exhaled breath profile. This matrix is incredibly complex but appears to show intra-individual consistency. The similarities in findings from this analysis to that described in the literature gives confidence that this dataset and analytical pipeline can be used for novel exploratory analysis

# Chapter 7 Exhaled VOCs as Novel Biomarkers

## 7.1 Introduction

In the preceding chapter (Chapter 6), a machine learning framework was established that was able to predict the sputum eosinophilic phenotype in the severe asthma cohort of patients characterised in this thesis. Chapters 3 and 4 demonstrated the limitations of inflammatory phenotypes, particularly for patients with T2 low disease, which was poorly defined and lacking in biomarkers. In Chapter 5, a T2 low phenotype was described, characterised by airway colonisation with *Haemophilus*, but unidentifiable using existing biomarkers.

Bacteria are known to produce a variety metabolites, including those with a low molecular mass (<300 Da), high vapor pressure and low boiling point (VOCs)[349]. This is corroborated by clinical experience: pseudomonal wound infections have a characteristic odour, distinct from wound infections due to other organisms [350,351]. These VOCs have been catalogued [213,349] and investigated in exhaled breath, identifying the potential for breathomics to be used as biomarker respiratory infections [352-354].

In this final results chapter, the breath analysis pipeline developed in Chapter 6 is used to predict the *Haemophilus* cluster defined in Chapter 5. *Haemophilus* dominant patients are of particular interest as *Haemophilus* is associated with corticosteroid resistance [208,296] but may be amenable to macrolide therapy [311,312]. Analysis will be restricted to non-erratic VOCs, as per the analysis in Chapter 5

The objectives of this chapter are

- Review Exploratory Analysis of Breathomics using non-erratic VOCs
- Exploratory Analysis of Breathomics data in relation to microbial data
- Evaluate the performances of VOC models for predicting the *Haemophilus* phenotype

## 7.2 Chapter Specific Methods

### 7.2.1 Patient Recruitment

All patients with sputum 16S samples and breath samples were included in this analysis

### 7.2.2 Supervised Machine Learning for Classification

Models were constructed as described in 6.2.10. However, when there are too few examples of the minority class in the training set, the model is unable to effectively learn how to make boundaries, moreover, it can lead to skewed evaluation metrics further impeding model training. Consequently, oversampling by SMOTE (Synthetic Minority Oversampling Technique) was used [355]. SMOTE was applied with k_neighbours of 2.

## 7.3 Results

### 7.3.1 Patient Population

Of the 74 patients in the Breath Sampling in Severe Asthma Study, only 52 patients had a paired *Haemophilus* cluster assignment from sputum 16S analysis (Figure 7.1).

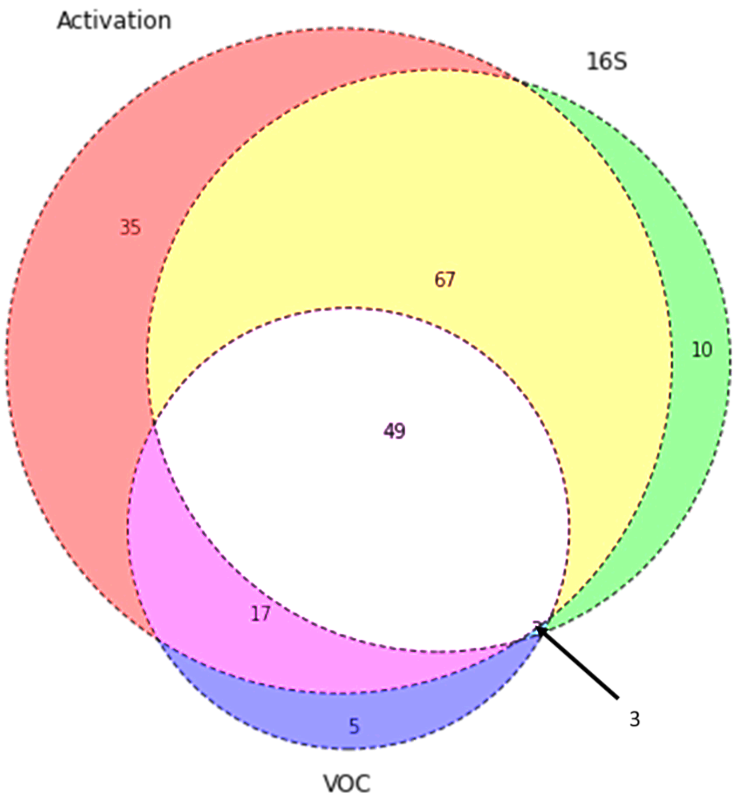


Figure 7.1 Venn Diagram of Patients and Overlapping Samples.

Abbreviations: Granulocyte Activation (Activation), 16S Microbial Sequencing (16S), Volatile Organic Compounds (VOC).

Comparing patients with and without paired 16S and breath data, patients with paired data had statistically significantly elevated blood eosinophil count and worse small airways disease (as measured by FEF25-75) (Table 1.1). Other parameters of disease severity (airway inflammatory cell counts, exacerbation frequency and self-report scores were the same, as were demographic and clinical characteristics.

Table 7.1 Clinical Characteristics of Patients with Paired Sputum 16S and Breath Data Available vs Patients that did not

Continuous variables expressed as median [Q1, Q3] with differences measured by Mann-Whitney U test. Categorical variables expressed as n (%) with differences measured by chi-square test. Abbreviations: ICS, inhaled corticosteroid; BDPe, beclomethasone dose equivalent; FeNO, fraction of nitric oxide in exhaled breath;

post BD, post bronchodilator; FEV1, forced expiratory volume in 1 second; FVC, forced vital capacity; FEF25-75%, forced expiratory flow at 25% to 75% of FVC; ACQ, asthma control questionnaire, HADSTOT, Hospital Anxiety and Depression Total Score; SNOT, SinoNasal Outcome Score

| | **Paired 16S and Breath not available (n = 143)** | **Paired 16S and Breath available (n = 52)** | **P-Value** |
|---|---|---|---|
| Sex (% Female) | 82 (57.3) | 21 (40.4) | 0.053 |
| Age | 58.0 [44.5,67.0] | 55.0 [49.0,62.2] | 0.690 |
| BMI | 29.8 [25.9,34.8] | 28.5 [25.6,32.6] | 0.130 |
| Smoker (% Never) | 49 (34.3) | 12 (23.1) | 0.252 |
| Atopy | 92 (64.3) | 27 (51.9) | 0.160 |
| Age of Onset | 21.0 [7.0,42.0] | 12.0 [3.0,33.5] | 0.102 |
| Exacerbations in Last 12 months | 2.0 [1.0,4.0] | 2.0 [0.0,3.0] | 0.534 |
| ICS (BDPe) | 3000.0 [2000.0,3000.0] | 3000.0 [2814.0,3840.0] | 0.236 |
| FeNO | 23.0 [15.0,45.0] | 31.0 [19.0,47.5] | 0.131 |
| Blood Eosinophil Count | 0.2 [0.1,0.4] | 0.3 [0.1,0.5] | 0.044 |
| Sputum Eosinophil (%) | 1.5 [0.5,8.6] | 3.3 [0.5,18.8] | 0.231 |
| Sputum Neutrophil (%) | 44.1 [25.6,67.1] | 42.3 [22.7,70.0] | 0.682 |
| PostBD FEV1 | 78.0 [62.1,93.4] | 72.1 [54.8,88.0] | 0.111 |
| PostBD FEV1/FVC | 69.0 [61.5,78.0] | 64.0 [54.8,74.5] | 0.034 |
| PostBD FEF25-75 %predicted | 46.7 [31.7,81.3] | 38.6 [27.8,69.3] | 0.066 |
| ACQ6 | 2.3 [1.3,3.0] | 2.7 [1.5,3.3] | 0.149 |
| HADSTOT | 10.0 [6.0,15.0] | 10.0 [5.0,16.0] | 0.720 |
| SNOT20 | 29.5 [17.2,45.0] | 30.0 [23.0,43.8] | 0.758 |

### 7.3.2 Exploratory Data Analysis through PCA

Exploratory analysis in Chapter 5 was performed primarily through PCA, agnostic to the within-subject variability of VOCs. PCA is highly sensitive to noise and so the analyses were repeated, restricting to non-erratic VOCs. Variance in the new PCs shows a similar pattern to previously (Chapter 6.3.3.3) with Sex being the major determinant of variation, followed by airway obstruction. Airway inflammation appears to be captured by PC7 (Figure 1.2).

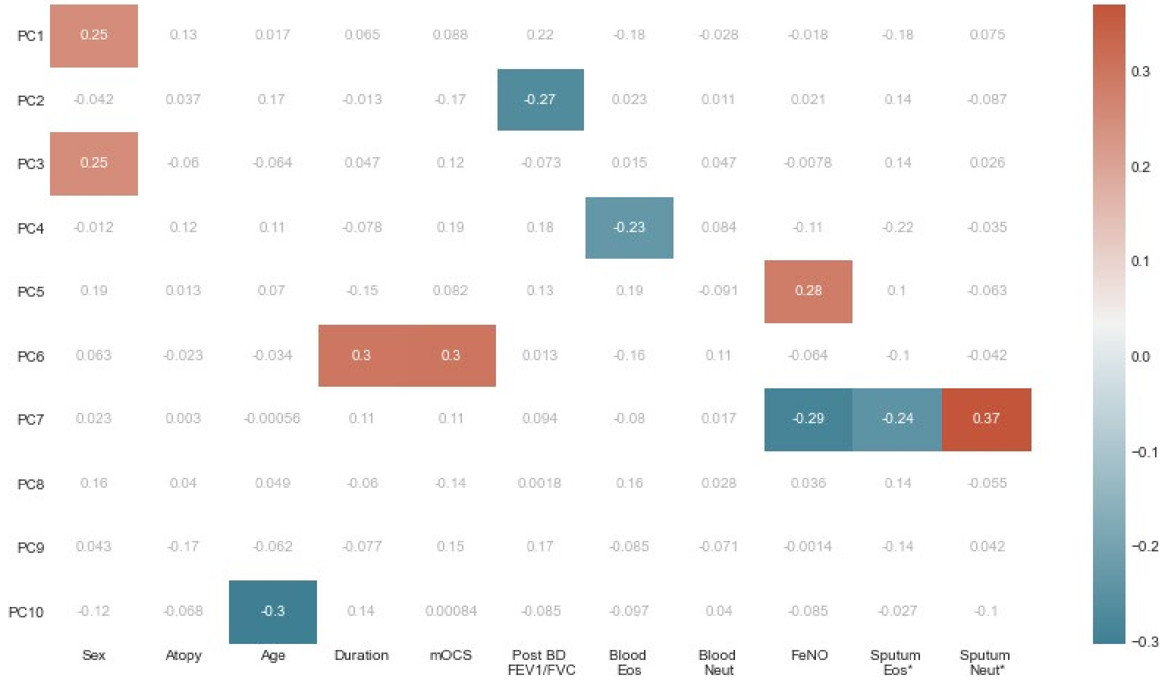
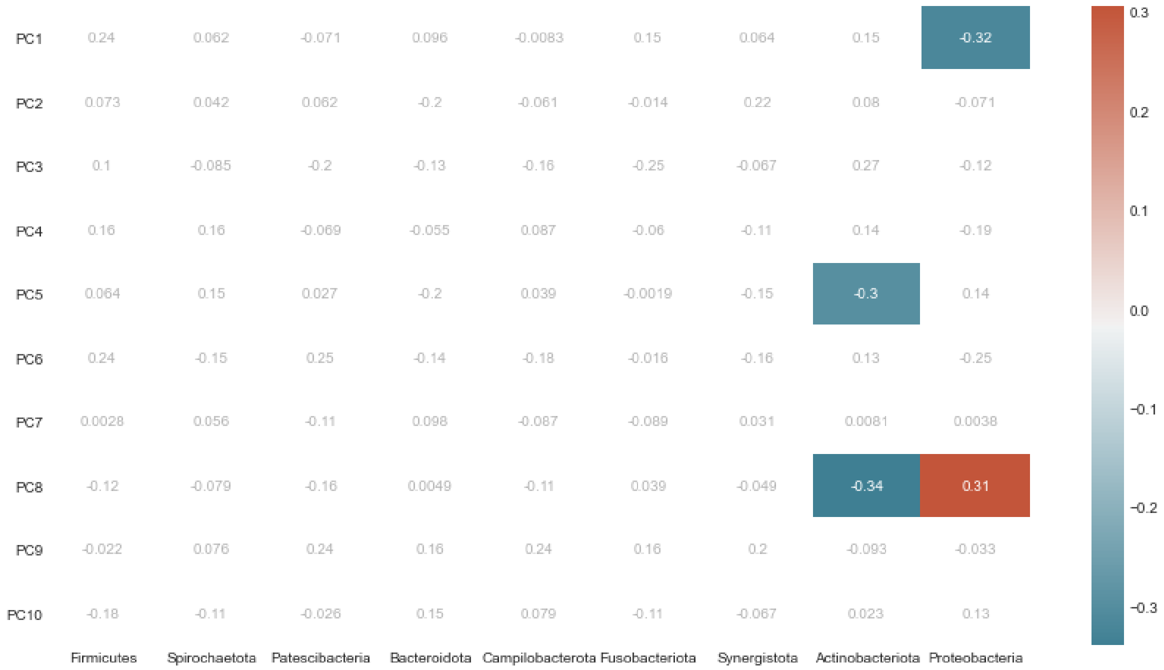


Figure 7.2 Heatmap of correlation between clinical characteristics and the first 10 principal components from the Cross Sectional Data, restricting to non-erratic VOCs.

Non-significant correlations in grey/white, positive and negative correlations with p-value < 0.05 in red and blue, respectively. * Sputum Eosinophils and Sputum Neutrophil Percentages log transformed. Volatile organic compounds (VOCs), principal component (PC), maintenance oral corticosteroid treatment dose (mOCS), post bronchodilator ratio between forced expiratory volume in 1 s and forced vital capacity (Post BD FEV1/FVC), eosinophils (Eos), neutrophils (Neut), Fraction of exhaled Nitric Oxide (FeNO).

#### 7.3.2.1 Microbial Variables

PC1, which captures the majority of variance in exhaled VOC data correlates to the relative abundance of Proteobacteria in the airways (Figure 1.3). PC1 also correlated with sex (Figure 1.2); there was no difference in proteobacteria between sexes: median (IQR) for male 4.80% (3.10-16.74), female 6.23% (3.35-8.27), Mann Whitney U statistic = 1766.0, p=0.495).



| | Firmicutes | Spirochaetota | Patescibacteria | Bacteroidota | Campilobacterota | Fusobacteriota | Synergistota | Actinobacteriota | Proteobacteria |
|---|---|---|---|---|---|---|---|---|---|
| PC1 | 0.24 | 0.062 | -0.071 | 0.096 | -0.0083 | 0.15 | 0.064 | 0.15 | -0.32 |
| PC2 | 0.073 | 0.042 | 0.062 | -0.2 | -0.061 | -0.014 | 0.22 | 0.08 | -0.071 |
| PC3 | 0.1 | -0.085 | -0.2 | -0.13 | -0.16 | -0.25 | -0.067 | 0.27 | -0.12 |
| PC4 | 0.16 | 0.16 | -0.069 | -0.055 | 0.087 | -0.06 | -0.11 | 0.14 | -0.19 |
| PC5 | 0.064 | 0.15 | 0.027 | -0.2 | 0.039 | -0.0019 | -0.15 | -0.3 | 0.14 |
| PC6 | 0.24 | -0.15 | 0.25 | -0.14 | -0.18 | -0.016 | -0.16 | 0.13 | -0.25 |
| PC7 | 0.0028 | 0.056 | -0.11 | 0.098 | -0.087 | -0.089 | 0.031 | 0.0081 | 0.0038 |
| PC8 | -0.12 | -0.079 | -0.16 | 0.0049 | -0.11 | 0.039 | -0.049 | -0.34 | 0.31 |
| PC9 | -0.022 | 0.076 | 0.24 | 0.16 | 0.24 | 0.16 | 0.2 | -0.093 | -0.033 |
| PC10 | -0.18 | -0.11 | -0.026 | 0.15 | 0.079 | -0.11 | -0.067 | 0.023 | 0.13 |

Figure 7.3 Heatmap of correlation between Relative abundance of airway phylum and the first 10 principal components from the Cross Sectional Data, restricting to non-erratic VOCs.

Non-significant correlations in grey/white, positive and negative correlations with p-value < 0.05 in red and blue, respectively. * Sputum Eosinophils and Sputum Neutrophil Percentages log transformed. Volatile organic compounds (VOCs), principal component (PC), maintenance oral corticosteroid treatment dose (mOCS), post bronchodilator ratio between forced expiratory volume in 1 s and forced vital capacity (Post BD FEV1/FVC), eosinophils (Eos), neutrophils (Neut), Fraction of exhaled Nitric Oxide (FeNO).

PCA Plot of exhaled VOCs shows some evidence of microbial cluster separation, specifically that of the *Haemophilus* cluster, though this is minimal (Figure 1.2)

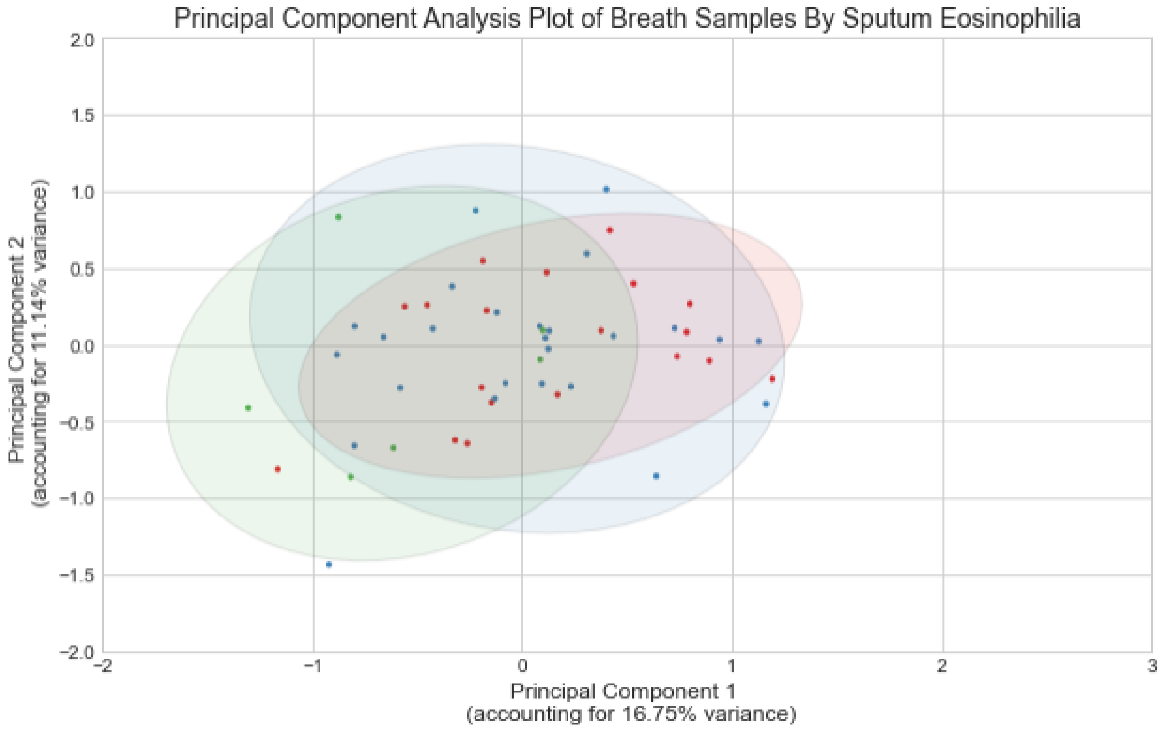


Figure 7.4 PCA plot of all 70 breath samples with ellipses representing Microbial Clusters.

Cluster 1 (Red), Cluster 2 (Blue), Cluster 3 (Green)

### 7.3.3 Predicting Microbial Cluster using VOCs

A model trained on non-erratic VOCs achieved an AUC of 0.571 for predicting patients in the *Haemophilus* clusters. This AUC improved to 0.857 when further feature selecting by feature importance (gain). This model had an accuracy of 0.86, sensitivity 0.71, specificity 1.0, f1 score of 0.83, PPV 1.0 and NPV 0.77. The most important VOC to contribute to this model were Decane, 3-methyl nonane and 1,3,5-Trifluorobenzene.

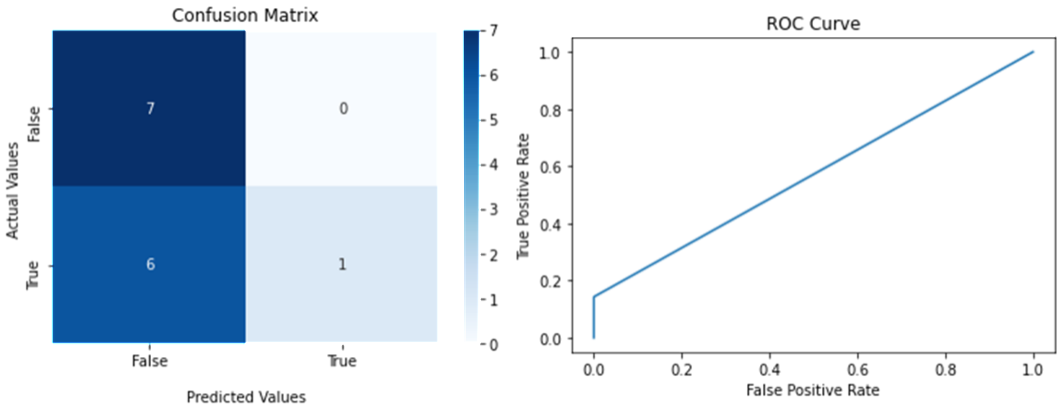


Figure 7.5 Confusion Matrix and ROC Curve for an XGB Classifier Trained on non-Erratic VOCs for predicting Patients in the *Haemophilus* Cluster

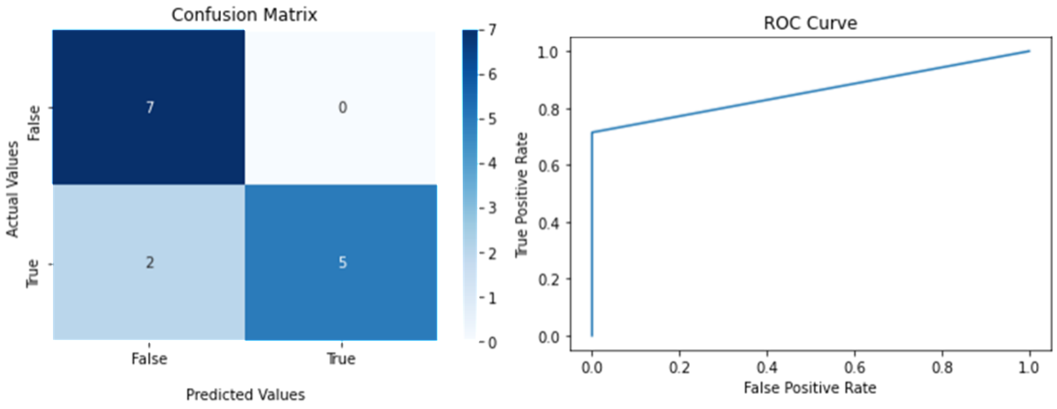


Figure 7.6 Confusion Matrix and ROC Curve for an XGB Classifier Trained on Important Features Determined from non-Erratic VOCs for predicting Patients in the *Haemophilus* Cluster.

## 7.4 Discussion

### 7.4.1 Breath Volatile Sensitive to Airway Microbiome

The model for predicting the *Haemophilus* cluster has a very high classification metrics. In this severe asthma population, this model has better classification performance than even FeNO and blood eosinophils for predicting sputum eosinophilia, as described in Chapter 3 (though they have utility beyond that single application [260,267]). The model for predicting *Haemophilus* outperforms the model for predicting sputum eosinophilia, demonstrating that the full potential for breathomics may be understood when paired with molecularly defined phenotypes.

PCA analysis of non-erratic VOCs indicate that they are sensitive to the relative abundance of sputum Proteobacteria, corroborated by the separation seen on the PCA plot of the *Haemophilus* dominant cluster and its accurate prediction through breath classifier. That PC1 should be correlated with the relative abundance of proteobacteria, albeit via a weak correlation, is surprising, particularly as there is no parallel relation with sputum neutrophilia. One explanation may be that the correlation between exhaled breath (PC1) and sputum proteobacteria may be due to shared correlation with an unobserved variable. The oropharyngeal microbiome is understood to be a major determinant of the lung microbiome [356,357] and, recognised to cause halitosis, thus a major determinant of the exhaled breath profile [358].

Decane was the major determinant of the machine learning model that accurately predicted *Haemophilus* colonisation in the test cohort. This has recently been identified in COPD as relating to viral exacerbations through a combination of in vitro and in vivo experiments [359]. The study made a clear distinction, however, that decane was not associated with *Haemophilus* i*nfluenzae*. Indeed it was not identified from the headspace of pure cultures of *Haemophilus* [360]. In addition methodological heterogeneity, there might be biological reasons for this dissociation, based on subtle physiological differences in asthma and COPD. Moreover, breath samples in this study were taking during a stable state: *Haemophilus* was not involved in an acute exacerbation event.

Without doubt, the observation that Decane is associated with *Haemophilus* requires further validation. Modern breathomics studies pair in vitro work with clinical sampling [354,359], and would complement the findings from this study.

### 7.4.2 Conclusions

Despite the aforementioned limitations, these findings demonstrate that exhaled VOCs may have clinical utility for phenotyping severe asthma patients and identifying a group of patients amenable to macrolide therapy [311]. The accurate identification of these patients could be important in avoiding unnecessary steroid therapy [361] and responsible antibiotic prescribing [362]. An extensive portfolio of future research is required to validate these findings and translate them to a clinical setting.

# Chapter 8 Discussion

## 8.1 Summary of Thesis Findings

### 8.1.1 Deep Phenotyping of a New Severe Asthma Cohort

The aim of this thesis was to evaluate exhaled VOCs as a biomarker for T2 low mechanisms of Severe Asthma. This was achieved by recruiting and clinically characterising a new cohort of severe asthma patients (Chapter 3) who were deeply characterised based upon their airway biology through measurements of sputum inflammatory cell activation (Chapter 4) and microbial composition (Chapter 5). Stratifying patients traditionally across sputum inflammatory cell counts confirmed that our cohort was consistent with existing cohorts. The analysis from this thesis reiterated how poorly T2 low asthma was characterised and the limitations of existing clinical biomarkers for predicting airway inflammation.

### 8.1.2 Description of a *Haemophilus*-MPO Cluster of Severe Asthma Patients

As described in Chapter 5, an association between airways disease and the airway microbiome has long been recognised. The findings from this thesis are consistent with modern studies describing a small haemophilus dominant cluster in severe asthma [310,327]. These patients are also unique by their abnormal airway inflammation, characterised by sputum neutrophilia and sputum MPO. It is increasingly appreciated that colonisation with haemophilus could represent a new treatable trait for severe asthma [363] but there is a desperate lack of accurate biomarkers [364] by which to identify these patients.

### 8.1.3 VOC Feature Selection by Within Patient Variance

Breathomics is still in its infancy and the breath samples collected and analysed as part of this thesis (Chapter 6) represents, at the time of writing, membership to a relatively exclusive collection of severe asthma cohorts [153,348]. Uniquely, however, this thesis describes the short-term repeatability of exhaled VOC measurements in severe asthma patients. This understanding is crucial in understanding the analytical validity of potential breathomics based biomarkers.

This repeated sampling was also used to develop and describe a novel feature reduction step: VOCs are, by definition, volatile and the logistical and financial costs of recruiting and characterising enough patients to overcome the noise this introduces are prohibitive, particularly whilst breathomics remains in its infancy. The analysis in Chapter 6 and Chapter 7 illustrate that within-patient variance for feature selection improves classification by exhaled VOCs

### 8.1.4 Breathomics as a Biomarker of Airway Colonisation with Haemophilus in Severe Asthma

Bringing together the aforementioned novel findings, the analysis described in this thesis (Chapter 7) identifies a potential role for exhaled VOC as a biomarker for patients colonised with Haemophilus and potentially amenable to Macrolide antibiotic therapy. The breathomics findings undoubtedly require validation but could represent a new precision medicine tool for severe asthma.

## 8.2 Limitations

There are a number of limitations to the analyses described in this thesis, which are discussed, where relevant, in the results chapters. Broader limitations are discussed herein.

### 8.2.1 Study Design and Sample Attrition

Though the Breath Sampling in Severe Asthma Study was designed with a priori intention of investigating short term repeatability and discovery of predictors of airway inflammation, it remains, for all intents and purposes a sub-study. The advantages of combining characterisation efforts (logistical and financial) were balanced against the ceiling effect that a cross-sectional study provides, primarily being limited to correlation alone.

Much like subject attrition, which is well recognised in clinical trials [365], the reasons for sample attrition are heterogenous.. When trying to overlap samples, the accumulation of losses can have multiplicative effect: in Chapter 7, just 52 patients were eligible for analysis, which severely underpowers any statistical analysis.

A major reason for sample loss in the Breath Sampling in Severe Asthma Study was the limited shelf life of breath sampling consumables. The study duration was extended, at cost, to improve numbers but was curtailed by the Covid-19 pandemic. Whilst a relatively high attrition rate was anticipated for sputum induction, it was grossly underestimated for breath sampling. This underestimation is somewhat representative of the perception of breathomics in academia: the clinical utility of breath sampling hides the complexity of breath sampling study design.

Identifying solutions to improve the number of observations is not straightforward. Autonomy over breath sampling consumables (e.g., reconditioning the TD tubes on which breath samples are captured) and in house GC-MS analysis of breath samples was not possible at the start of the study: neither the capability, protocols nor expertise had been established locally. External collaboration and the logistical issues associated with this was therefore unavoidable. Extending the study duration was not possible due to the pandemic but would have been prohibited by cost. If, hypothetically, additional investment had been possible then, with retrospect, a multi-centre study design might have been given stronger consideration. Breath samples are shown to vary between sites [157] and so this is not without (logistical) costs [366].

Strong consideration was given to resorting to cross-validation (i.e., internally validating VOC models) rather than keeping the repeat breath samples as a distinct test cohort. Cross validation, specifically leave one out cross validation, is commonly applied to breathomics datasets [136] but severely limits generalisability to the “real world” and is a major contributor the general lack of replication of findings [343]. Maintaining an a priori test cohort achieves a higher level of internal validation [367] than cross validation but was, in this case, tempered by a reliance on imputation [368].

True validation would require the replication of findings from a different group. The unique study design for this study – restricting to severe asthma and pairing with sputum 16S sequencing – limits these prospects. Thus, the analysis presented in this thesis represents a common trap for this emerging technology: prioritisation of innovation over standardisation [338].

### 8.2.2 Limitations of Machine Learning

One of the major and recurring tools used in this thesis was clustering by classical methods of hierarchical and partitioning around medioids. These approaches aim to group patients with similar biology together and separate patients with distinct biology [217]. One of the risks of clustering is that patients can be artificially forced apart rather than represented as existing on a gradient, which speaks to the concern that machine learning algorithms will find clusters even when none exist (or are at least not biologically/clinically meaningful) [369]. This effect is confounded by complex datasets [370] (Garbage In, Garbage Out).

Consideration had been given to applying topological data analysis approaches [371] to the data, which have been successfully applied to biological data in severe asthma [198]. The advantage of TDA would have been to provide a two-dimensional representation that retains the essential features of the original high-dimensional data set, which does not force patients apart. The main obstruction to this approach was the limited numbers and noise of the breathomics data.

Concerns about artificial dichotomisation is mitigated by a constant appreciation that unsupervised approaches are hypothesis generating rather than hypothesis testing. Moreover, these approaches are useful tools by which to digest otherwise incomprehensibly highly dimensional information [372]. The utility and limitations of dimension reduction can be illustrated by the widespread of adoption of scoring systems in clinical practice: e.g. the early warning scoring system, which condenses various parameters of physiological function into a single digit number [373].

VOC biomarker identification in this analysis was done by supervised machine learning classifiers, which, like clustering, have their limitations. One of the limitations of tree based ensemble models is that feature importance does not (without using an explainer) indicate direction: it is, for example, unclear whether Decane is increased or decreased with Haemophilus abundance in Chapter 7. Consequently, black boxes in clinical research are treated with deep suspicion [374,375]. An additional problem with this lack of transparency is that it prohibits power calculations for future studies. Though, there are tools to mitigate some of these issues (SHapley Additive exPlanations (SHAP) [376] for explaining models and approaches to sample size prediction algorithms [377]), the optimal solution is not to rely solely on machine learning approaches. .

One of the reasons machine learning approaches have become central to systems biology approach to understanding disease is their ability to handle multi-dimensional data [97,98]. However, supervised machine learning tools are designed to predict rather than describe biology. Ranking by feature importance in a tree based model (such as XGBoost) reflects importance to the prediction model rather than importance to the target variable. For example, as described by the Network Representation of a Correlation Matrix of Exhaled VOCs (Figure 6.9), the abundance of Decane correlates with other straight chain alkanes such as nonane and undecane. These VOCs might be equally relevant to Haemophilus abundance but due to collinearity, would not add further information to the prediction model beyond that provided by Decane and would therefore be assigned low feature importance. This risks under-reporting of salient associations and may hinder efforts at external validation.

### 8.2.3 Normal Breath Profile

One of the major limitations of this study was that analysis and pre-processing was performed by external collaborators. A normal exhaled breath profile has not been defined making it difficult to quality check the data produced. This is compound by the fact that the published literature rarely reports a full list of VOCs identified in their exhaled breath sample and their average abundances. Sense checking by identifying VOCs present in this study and the reported literature (Chapter 6) is prone to confirmation bias. A number of unusual compounds were also identified: sevoflurane, an anaesthetic agent, as well as chlorinated and fluorinated compounds. These compounds may represent misidentifications (Chapter 1.3) or contaminants but the inability to determine this is challenging.

The within subject variance feature reduction step described in this study is an attempt to mitigate this issue. One concern with removing erratic VOCs is that they may be biologically salient. This argument can be reduced to a debate on balancing type 1 and type 2 error: including erratic VOCs will undoubtedly lead to false positives whilst excluding them will likely lead to false negatives. Due to the concerns described above, reducing type 1 error was prioritised.

## 8.3 Follow up Studies for the Breathomics Findings in this Study

Reporting in this thesis suffers from many of the traps that the published literature falls into. Chiefly, molecular features were identified by comparing against the NIST library (Chapter 6.2.3.3) but true identification requires comparing molecular features to chemical standards. The putative identification of Decane should be verified before further validation work.

Next, as referenced throughout this discussion, biomarker prediction through classification alone is insufficient. In order to follow up the results from this analysis, a number of complimentary studies could be explored. Firstly, headspace experiments of culturing haemophilus or airway epithelial cells infected with haemophilus would give mechanistic insight into the origins of Decane. Headspace experiment methodology has been described [348,378] and successfully used to complement clinical samples [359]. Of course, replicating the association between exhaled Decane and airway Haemophilus in another cohort of asthma patients is critical. Longitudinal studies describing the long term repeatability of Decane in persistently haemophilus colonised patients and attenuation of the Decane signal in those in which haemophilus has been reduced/eliminated would strengthen the correlative findings.

If the Decane signal was validated through the measures described above, it is highly unlikely that GCMS would be used in clinical practice; rather, breathomics is likely to be implemented in the form of an electronic nose. One of the advantages of electronic noses is that it can be used “point of care. The importance of this feature is likely overstated: clinical practice rarely uses investigations with immediately available results, particularly in non-emergency situations (e.g. cross-sectional imaging from the time of requesting to reporting, even in outpatient lung cancer services, can take days-weeks). In this context, characterising patients in an outpatient setting, collecting a breath sample and sending it away for analysis, would probably be acceptable. It is far more likely that cost will be the prohibitive factor. The additional advantage of electronic noses (online testing) is the avoidance of heterogeneity and variance introduced by breath sampling and storage [163]. It follows therefore that there would need to be further method development of an electronic nose and standard operating procedure sensitive to Decane. This would then need to be assessed again in a large-scale population.

The application of an exhaled Decane biomarker proposed by this thesis is to identify patients with airway overabundance of Haemophilus. These patients are likely to respond to Azithromycin therapy [311]. Further studies would have to demonstrate that a Decane electronic nose was superior to standard of care in informing Azithromycin prescription: possible endpoints for such a study might include response rates to Azithromycin prescription, side-effect burden, anti-microbial resistance. As Azithromycin is readily available and relatively cheap, a robust health economic rationale would be required to justify the cost of a Decane electronic nose.

## 8.4 Future of Breathomics Research in Airways Disease

### 8.4.1 Standardisation

No treatise on breathomics would be complete without a call for standardisation, a call which has been repeated many times [182] but appears unheard. Reluctance of the field to standardise likely comes down to the rejection of what is current available to standardise to. Breath Sampling, Sample Analysis and Data Analysis [157] are useful headings to describe the main pillars of breathomics. Optimisation of Sample Analysis and Data Analysis has largely been achieved or can be borrowed from other fields (petrochemical [379] and other high-throughput omics analyses [380], respectively).

Breath Sampling is the likely, therefore, the greatest challenge. The findings from this analysis indicate that, despite collecting multiple breaths through a highly engineered sampler, there was great day to day variability in the exhaled VOCs measured. Further engineering of the sampler or restrictions on breath sampling (e.g on foods/drinks) risk making breathomics more expensive and more difficult. The alternative would be a simpler breath sampling system, which might be easier to adopt in larger numbers, but potentially sacrifice analytical validity. Rather than standardisation, innovation in this aspect is critical.

Where standardisation is required is in study design and reporting. A number of well conducet large scale breathomics studies have now been performed in asthma [151,197] but, despite this, there is little in the way of shared or shareable data. Adherence to the TRIPOD recommendations on reporting multivariable prediction models [180] and STARD guidelines on reporting of diagnostics accuracy studies [181] is still desperately lacking[136] and is perhaps the first priority.

### 8.4.2 Innovation

One of the strengths of this study was the focus on repeated breath samples, which remains an under-utilised quality of breathomics in the existing literature. Challenge tests are common in medicine, including the bronchodilator reversibility test for airways disease; repeated breath sampling would facilitate a similar model for breathomics. In hepatology[381] and lung cancer (NCT05510674) radiolabelled VOC probes are being explored as functional measures. These rely on a priori knowledge of the specific mechanisms and applications to asthma may not be immediately obvious.

Another strength of this study is the emphasis on novel endpoints. Assessment of biomarkers is limited by the gold standard against which they are measured [199]. In this thesis, breathomics identified a biomarker describing a novel phenotype defined by 16S rRNA profiling of the airway microbiome. Molecular endpoints are challenging due to the sample attrition described in Chapter 8.2.1. and already evident in the literature[153] . Moreover, strategies for combining multiple omics platforms are poorly defined [382]. Successful study endpoints, at least in the imminent future are likely to be clinically rather than molecularly driven. These might include longitudinal parameters (e.g. lung function decline), early diagnostics or [151].

## 8.5 Conclusions

This thesis sought to understand whether exhaled VOCs could be used as a biomarker for airway biology in severe asthma. Models for predicting airway inflammatory phenotypes and microbial phenotypes were successfully constructed and found to have good predictive performance. In Chapter 1, a rendition of the EGAPP qualities by which to judge a biomarker [106] was described: is the test result true, meaningful or useful [107]. Sputum eosinophilia and airway *Haemophilus* do not describe the whole of airway biology but, representing, treatable traits they are undoubtedly meaningful. Though the demand (usefulness) for novel biomarkers for these traits is less convincing, ultimately, it is the question of whether the results are true that remains the greatest challenge. As discussed, a host of further experimentation and validation is required in concert with evolution of the field more generally.

Such a cautious and conditional affirmative is unsurprising given the relative immaturity of breathomics. More mature molecular technologies than breathomics are yet to translate to the asthma clinic. Breathomics easily captures the imagination and, small studies such as this, which can advocate for the continued adoption of breathomics into large projects, such as U-BIOPRED [153], MRC-EMBER [110] and most recently in IMI-3TR (3tr-imi.eu) are critical for the field to develop.

As discussed, a combination of standardisation in reporting but innovation in practice is required right now. A number of centres from across Europe have started to produce robust research [348,354,359,383], such that we are likely to be entering a golden age of breathomics research.

# List of References

1. Bergmann KC. Asthma. *Chem Immunol Allergy* 2014;100:69-80.

2. Cohen SG. Asthma among the famous. Henry Hyde Salter (1823-1871), British physician. *Allergy Asthma Proc* 1997;18(4):256-8.

3. Asthma GIf. *Global Strategy for Asthma Management and Prevention* https://ginasthma.org/wp-content/uploads/2020/06/GINA-2020-report_20_06_04-1-wms.pdf (accessed 10/10/20).

4. Network GA. *The global asthma report*. http://www.globalasthmareport.org/Global%20Asthma%20Report%202018.pdf. (accessed 11/09/20).

5. UK A. *Asthma facts and statistics* (accessed 10/10/20).

6. Fleming DM, Crombie DL. Prevalence of asthma and hay fever in England and Wales. *Br Med J (Clin Res Ed)* 1987;294(6567):279-83.

7. Aberg N, Hesselmar B, Aberg B, et al. Increase of asthma, allergic rhinitis and eczema in Swedish schoolchildren between 1979 and 1991. *Clin Exp Allergy* 1995;25(9):815-9.

8. Ereshefsky M. Defining 'health' and 'disease'. *Stud Hist Philos Biol Biomed Sci* 2009;40(3):221-7.

9. Boyd KM. Disease, illness, sickness, health, healing and wholeness: exploring some elusive concepts. *Med Humanit* 2000;26(1):9-17.

10. Scully JL. What is a disease? *EMBO Rep* 2004;5(7):650-3.

11. Lewens T, McMillan J. Defining disease. *Lancet* 2004;363(9409):664.

12. Gross NJ. What is this thing called love? --or, defining asthma. *Am Rev Respir Dis* 1980;121(2):203-4.

13. Castro-Giner F, Kauffmann F, de Cid R, et al. Gene-environment interactions in asthma. *Occup Environ Med* 2006;63(11):776-86, 61.

14. Holgate ST. Genetic and environmental interaction in allergy and asthma. *J Allergy Clin Immunol* 1999;104(6):1139-46.

15. Spycher BD, Silverman M, Kuehni CE. Phenotypes of childhood asthma: are they real? *Clin Exp Allergy* 2010;40(8):1130-41.

16. Haldar P, Pavord ID, Shaw DE, et al. Cluster analysis and clinical asthma phenotypes. *Am J Respir Crit Care Med* 2008;178(3):218-24.

17. Haselkorn T, Fish JE, Zeiger RS, et al. Consistently very poorly controlled asthma, as defined by the impairment domain of the Expert Panel Report 3 guidelines, increases risk for future severe asthma exacerbations in The Epidemiology and Natural History of Asthma: Outcomes and Treatment Regimens (TENOR) study. *J Allergy Clin Immunol* 2009;124(5):895-902 e1-4.

18. Sastre J, Fabbri LM, Price D, et al. Insights, attitudes, and perceptions about asthma and its treatment: a multinational survey of patients from Europe and Canada. *World Allergy Organ J* 2016;9:13.

19. Shaw DE, Sousa AR, Fowler SJ, et al. Clinical and inflammatory characteristics of the European U-BIOPRED adult severe asthma cohort. *Eur Respir J* 2015;46(5):1308-21.

20. Uchmanowicz B, Panaszek B, Uchmanowicz I, et al. Clinical factors affecting quality of life of patients with asthma. *Patient Prefer Adherence* 2016;10:579-89.

21. Chen H, Gould MK, Blanc PD, et al. Asthma control, severity, and quality of life: quantifying the effect of uncontrolled disease. *J Allergy Clin Immunol* 2007;120(2):396-402.

22. Sullivan PW, Ghushchyan VH, Slejko JF, et al. The burden of adult asthma in the United States: evidence from the Medical Expenditure Panel Survey. *J Allergy Clin Immunol* 2011;127(2):363-69 e1-3.

23. Physicians RCo. *Why Asthma Still Kills: the National Review of Asthma Deaths (NRAD) Confidential Enquiry Report*. www.rcplondon.ac.uk/sites/default/files/why-asthma-still-kills-full-report.pdf (accessed 11/10/2020).

24. Smith DH, Malone DC, Lawson KA, et al. A national estimate of the economic costs of asthma. *Am J Respir Crit Care Med* 1997;156(3 Pt 1):787-93.

25. Barnes PJ, Jonsson B, Klim JB. The costs of asthma. *Eur Respir J* 1996;9(4):636-42.

26. (adult) ENSRS. *Severe Asthma* https://www.england.nhs.uk/wp-content/uploads/2017/04/specialised-respiratory-services-adult-severe-asthma.pdf. (accessed 21/05/2019).

27. Azim A, Mistry H, Freeman A, et al. Protocol for the Wessex AsThma CoHort of difficult asthma (WATCH): a pragmatic real-life longitudinal study of difficult asthma in the clinic. *BMC Pulm Med* 2019;19(1):99.

28. Chung KF, Wenzel SE, Brozek JL, et al. International ERS/ATS guidelines on definition, evaluation and treatment of severe asthma. *Eur Respir J* 2014;43(2):343-73.

29. Hekking PP, Wener RR, Amelink M, et al. The prevalence of severe refractory asthma. *J Allergy Clin Immunol* 2015;135(4):896-902.

30. Wagner PD. The physiological basis of pulmonary gas exchange: implications for clinical interpretation of arterial blood gases. *Eur Respir J* 2015;45(1):227-43.

31. Lloyd CM, Hawrylowicz CM. Regulatory T cells in asthma. *Immunity* 2009;31(3):438-49.

32. Takahama Y. Journey through the thymus: stromal guides for T-cell development and selection. *Nat Rev Immunol* 2006;6(2):127-35.

33. Ricci M, Matucci A, Rossi O. T cells, cytokines, IgE and allergic airways inflammation. *J Investig Allergol Clin Immunol* 1994;4(5):214-20.

34. Mosmann TR, Coffman RL. TH1 and TH2 cells: different patterns of lymphokine secretion lead to different functional properties. *Annu Rev Immunol* 1989;7:145-73.

35. Holgate ST. A look at the pathogenesis of asthma: the need for a change in direction. *Discov Med* 2010;9(48):439-47.

36. Kline JN, Hunninghake GW. T-lymphocyte dysregulation in asthma. *Proc Soc Exp Biol Med* 1994;207(3):243-53.

37. Bousquet J, Chanez P, Lacoste JY, et al. Eosinophilic inflammation in asthma. *N Engl J Med* 1990;323(15):1033-9.

38. Robinson DS, Hamid Q, Ying S, et al. Predominant TH2-like bronchoalveolar T-lymphocyte population in atopic asthma. *N Engl J Med* 1992;326(5):298-304.

39. Choy DF, Modrek B, Abbas AR, et al. Gene expression patterns of Th2 inflammation and intercellular communication in asthmatic airways. *J Immunol* 2011;186(3):1861-9.

40. Bentley AM, Menz G, Storz C, et al. Identification of T lymphocytes, macrophages, and activated eosinophils in the bronchial mucosa in intrinsic asthma. Relationship to symptoms and bronchial responsiveness. *Am Rev Respir Dis* 1992;146(2):500-6.

41. Kuruvilla ME, Lee FE, Lee GB. Understanding Asthma Phenotypes, Endotypes, and Mechanisms of Disease. *Clin Rev Allergy Immunol* 2019;56(2):219-33.

42. Green RH, Brightling CE, McKenna S, et al. Asthma exacerbations and sputum eosinophil counts: a randomised controlled trial. *Lancet* 2002;360(9347):1715-21.

43. Wenzel SE, Schwartz LB, Langmack EL, et al. Evidence that severe asthma can be divided pathologically into two inflammatory subtypes with distinct physiologic and clinical characteristics. *Am J Respir Crit Care Med* 1999;160(3):1001-8.

44. Wenzel SE. Asthma phenotypes: the evolution from clinical to molecular approaches. *Nat Med* 2012;18(5):716-25.

45. Woodruff PG, Modrek B, Choy DF, et al. T-helper type 2-driven inflammation defines major subphenotypes of asthma. *Am J Respir Crit Care Med* 2009;180(5):388-95.

46. Moore WC, Meyers DA, Wenzel SE, et al. Identification of asthma phenotypes using cluster analysis in the Severe Asthma Research Program. *Am J Respir Crit Care Med* 2010;181(4):315-23.

47. Hinks TSC, Brown T, Lau LCK, et al. Multidimensional endotyping in patients with severe asthma reveals inflammatory heterogeneity in matrix metalloproteinases and chitinase 3-like protein 1. *J Allergy Clin Immunol* 2016;138(1):61-75.

48. Newby C, Heaney LG, Menzies-Gow A, et al. Statistical cluster analysis of the British Thoracic Society Severe refractory Asthma Registry: clinical outcomes and phenotype stability. *PLoS One* 2014;9(7):e102987.

49. Kuo CS, Pavlidis S, Loza M, et al. A Transcriptome-driven Analysis of Epithelial Brushings and Bronchial Biopsies to Define Asthma Phenotypes in U-BIOPRED. *Am J Respir Crit Care Med* 2017;195(4):443-55.

50. Brusselle GG, Maes T, Bracke KR. Eosinophils in the spotlight: Eosinophilic airway inflammation in nonallergic asthma. *Nat Med* 2013;19(8):977-9.

51. McKenzie ANJ, Spits H, Eberl G. Innate lymphoid cells in inflammation and immunity. *Immunity* 2014;41(3):366-74.

52. Lambrecht BN, Hammad H. The immunology of asthma. *Nat Immunol* 2015;16(1):45-56.

53. Lotvall J, Akdis CA, Bacharier LB, et al. Asthma endotypes: a new approach to classification of disease entities within the asthma syndrome. *J Allergy Clin Immunol* 2011;127(2):355-60.

54. Pavord ID, Beasley R, Agusti A, et al. After asthma: redefining airways diseases. *Lancet* 2018;391(10118):350-400.

55. Siddiqui S, Denlinger LC, Fowler SJ, et al. Unmet Needs in Severe Asthma Subtyping and Precision Medicine Trials. Bridging Clinical and Patient Perspectives. *Am J Respir Crit Care Med* 2019;199(7):823-29.

56. Muraro A, Lemanske RF, Jr., Hellings PW, et al. Precision medicine in patients with allergic diseases: Airway diseases and atopic dermatitis-PRACTALL document of the European Academy of Allergy and Clinical Immunology and the American Academy of Allergy, Asthma & Immunology. *J Allergy Clin Immunol* 2016;137(5):1347-58.

57. Biomarkers Definitions Working G. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin Pharmacol Ther* 2001;69(3):89-95.

58. Bakhtiar R. Biomarkers in drug discovery and development. *J Pharmacol Toxicol Methods* 2008;57(2):85-91.

59. Selleck MJ, Senthil M, Wall NR. Making Meaningful Clinical Use of Biomarkers. *Biomark Insights* 2017;12:1177271917715236.

60. Reddel HK, Taylor DR, Bateman ED, et al. An official American Thoracic Society/European Respiratory Society statement: asthma control and exacerbations: standardizing endpoints for clinical asthma trials and clinical practice. *Am J Respir Crit Care Med* 2009;180(1):59-99.

61. Green RH, Brightling CE, Woltmann G, et al. Analysis of induced sputum in adults with asthma: identification of subgroup with isolated sputum neutrophilia and poor response to inhaled corticosteroids. *Thorax* 2002;57(10):875-9.

62. Pavord ID, Brightling CE, Woltmann G, et al. Non-eosinophilic corticosteroid unresponsive asthma. *Lancet* 1999;353(9171):2213-4.

63. Fahy JV, Boushey HA. Effect of low-dose beclomethasone dipropionate on asthma control and airway inflammation. *Eur Respir J* 1998;11(6):1240-7.

64. van Rensen EL, Straathof KC, Veselic-Charvat MA, et al. Effect of inhaled steroids on airway hyperresponsiveness, sputum eosinophils, and exhaled nitric oxide levels in patients with asthma. *Thorax* 1999;54(5):403-8.

65. Wagener AH, de Nijs SB, Lutter R, et al. External validation of blood eosinophils, FE(NO) and serum periostin as surrogates for sputum eosinophils in asthma. *Thorax* 2015;70(2):115-20.

66. Zhang XY, Simpson JL, Powell H, et al. Full blood count parameters for the detection of asthma inflammatory phenotypes. *Clin Exp Allergy* 2014;44(9):1137-45.

67. Price DB, Rigazio A, Campbell JD, et al. Blood eosinophil count and prospective annual asthma disease burden: a UK cohort study. *Lancet Respir Med* 2015;3(11):849-58.

68. Hancox RJ, Pavord ID, Sears MR. Associations between blood eosinophils and decline in lung function among adults with and without asthma. *Eur Respir J* 2018;51(4).

69. Tran TN, Khatry DB, Ke X, et al. High blood eosinophil count is associated with more frequent asthma attacks in asthma patients. *Ann Allergy Asthma Immunol* 2014;113(1):19-24.

70. Vedel-Krogh S, Fallgaard Nielsen S, Lange P, et al. Association of Blood Eosinophil and Blood Neutrophil Counts with Asthma Exacerbations in the Copenhagen General Population Study. *Clin Chem* 2017;63(4):823-32.

71. Buhl R, Humbert M, Bjermer L, et al. Severe eosinophilic asthma: a roadmap to consensus. *Eur Respir J* 2017;49(5).

72. Lloyd CM, Snelgrove RJ. Type 2 immunity: Expanding our view. *Sci Immunol* 2018;3(25).

73. Ortega HG, Liu MC, Pavord ID, et al. Mepolizumab treatment in patients with severe eosinophilic asthma. *N Engl J Med* 2014;371(13):1198-207.

74. Chupp GL, Bradford ES, Albers FC, et al. Efficacy of mepolizumab add-on therapy on health-related quality of life and markers of asthma control in severe eosinophilic asthma (MUSCA): a randomised, double-blind, placebo-controlled, parallel-group, multicentre, phase 3b trial. *Lancet Respir Med* 2017;5(5):390-400.

75. FitzGerald JM, Bleecker ER, Nair P, et al. Benralizumab, an anti-interleukin-5 receptor alpha monoclonal antibody, as add-on treatment for patients with severe, uncontrolled, eosinophilic asthma (CALIMA): a randomised, double-blind, placebo-controlled phase 3 trial. *Lancet* 2016;388(10056):2128-41.

76. Bleecker ER, FitzGerald JM, Chanez P, et al. Efficacy and safety of benralizumab for patients with severe asthma uncontrolled with high-dosage inhaled corticosteroids and long-acting beta2-agonists (SIROCCO): a randomised, multicentre, placebo-controlled phase 3 trial. *Lancet* 2016;388(10056):2115-27.

77. Kroes JA, Zielhuis SW, van Roon EN, et al. Prediction of response to biological treatment with monoclonal antibodies in severe asthma. *Biochem Pharmacol* 2020;179:113978.

78. Stuehr DJ. Mammalian nitric oxide synthases. *Biochim Biophys Acta* 1999;1411(2-3):217-30.

79. Gustafsson LE, Leone AM, Persson MG, et al. Endogenous nitric oxide is present in the exhaled air of rabbits, guinea pigs and humans. *Biochem Biophys Res Commun* 1991;181(2):852-7.

80. American Thoracic S, European Respiratory S. ATS/ERS recommendations for standardized procedures for the online and offline measurement of exhaled lower respiratory nitric oxide and nasal nitric oxide, 2005. *Am J Respir Crit Care Med* 2005;171(8):912-30.

81. Korevaar DA, Westerhof GA, Wang J, et al. Diagnostic accuracy of minimally invasive markers for detection of airway eosinophilia in asthma: a systematic review and meta-analysis. *Lancet Respir Med* 2015;3(4):290-300.

82. Dupont LJ, Rochette F, Demedts MG, et al. Exhaled nitric oxide correlates with airway hyperresponsiveness in steroid-naive patients with mild asthma. *Am J Respir Crit Care Med* 1998;157(3 Pt 1):894-8.

83. Malinovschi A, Fonseca JA, Jacinto T, et al. Exhaled nitric oxide levels and blood eosinophil counts independently associate with wheeze and asthma events in National Health and Nutrition Examination Survey subjects. *J Allergy Clin Immunol* 2013;132(4):821-7 e1-5.

84. Silkoff PE, McClean P, Spino M, et al. Dose-response relationship and reproducibility of the fall in exhaled nitric oxide after inhaled beclomethasone dipropionate therapy in asthma patients. *Chest* 2001;119(5):1322-8.

85. McNicholl DM, Stevenson M, McGarvey LP, et al. The utility of fractional exhaled nitric oxide suppression in the identification of nonadherence in difficult asthma. *Am J Respir Crit Care Med* 2012;186(11):1102-8.

86. Tiotiu A. Biomarkers in asthma: state of the art. *Asthma Res Pract* 2018;4:10.

87. Chung KF. Personalised medicine in asthma: time for action: Number 1 in the Series "Personalised medicine in respiratory diseases" Edited by Renaud Louis and Nicolas Roche. *Eur Respir Rev* 2017;26(145).

88. Schleich F, Demarche S, Louis R. Biomarkers in the Management of Difficult Asthma. *Curr Top Med Chem* 2016;16(14):1561-73.

89. Fahy JV. Eosinophilic and neutrophilic inflammation in asthma: insights from clinical studies. *Proc Am Thorac Soc* 2009;6(3):256-9.

90. Moore WC, Hastie AT, Li X, et al. Sputum neutrophil counts are associated with more severe asthma phenotypes using cluster analysis. *J Allergy Clin Immunol* 2014;133(6):1557-63 e5.

91. Ayers D, Day PJ. Systems Medicine: The Application of Systems Biology Approaches for Modern Medical Research and Drug Development. *Mol Biol Int* 2015;2015:698169.

92. Eberhardt M, Lai X, Tomar N, et al. Third-Kind Encounters in Biomedicine: Immunology Meets Mathematics and Informatics to Become Quantitative and Predictive. *Methods Mol Biol* 2016;1386:135-79.

93. Friboulet A, Thomas D. Systems Biology-an interdisciplinary approach. *Biosens Bioelectron* 2005;20(12):2404-7.

94. Kirschner MW. The meaning of systems biology. *Cell* 2005;121(4):503-04.

95. Breitling R. What is systems biology? *Front Physiol* 2010;1:9.

96. Kitano H. Systems biology: a brief overview. *Science* 2002;295(5560):1662-4.

97. Molinelli EJ, Korkut A, Wang W, et al. Perturbation biology: inferring signaling networks in cellular systems. *PLoS Comput Biol* 2013;9(12):e1003290.

98. Ge H, Walhout AJ, Vidal M. Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet* 2003;19(10):551-60.

99. Tang HHF, Sly PD, Holt PG, et al. Systems biology and big data in asthma and allergy: recent discoveries and emerging challenges. *Eur Respir J* 2020;55(1).

100. Dunn WB, Broadhurst DI, Atherton HJ, et al. Systems level studies of mammalian metabolomes: the roles of mass spectrometry and nuclear magnetic resonance spectroscopy. *Chem Soc Rev* 2011;40(1):387-426.

101. Jung J, Kim SH, Lee HS, et al. Serum metabolomics reveals pathways and biomarkers associated with asthma pathogenesis. *Clin Exp Allergy* 2013;43(4):425-33.

102. Loureiro CC, Oliveira AS, Santos M, et al. Urinary metabolomic profiling of asthmatics can be related to clinical characteristics. *Allergy* 2016;71(9):1362-5.

103. Kolmert J, Gomez C, Balgoma D, et al. Urinary Leukotriene E4 and Prostaglandin D2 Metabolites Increase in Adult and Childhood Severe Asthma Characterized by Type-2 Inflammation. *Am J Respir Crit Care Med* 2020.

104. Reinke SN, Gallart-Ayala H, Gomez C, et al. Metabolomics analysis identifies different metabotypes of asthma severity. *Eur Respir J* 2017;49(3).

105. Patti GJ, Yanes O, Siuzdak G. Innovation: Metabolomics: the apogee of the omics trilogy. *Nat Rev Mol Cell Biol* 2012;13(4):263-9.

106. Teutsch SM, Bradley LA, Palomaki GE, et al. The Evaluation of Genomic Applications in Practice and Prevention (EGAPP) Initiative: methods of the EGAPP Working Group. *Genet Med* 2009;11(1):3-14.

107. Bossuyt PM. Clinical validity: defining biomarker performance. *Scand J Clin Lab Invest Suppl* 2010;242:46-52.

108. Fowler SJ. Breath analysis for label-free characterisation of airways disease. *Eur Respir J* 2018;51(1).

109. Fitzpatrick AM. Biomarkers of asthma and allergic airway diseases. *Ann Allergy Asthma Immunol* 2015;115(5):335-40.

110. Ibrahim W, Wilde M, Cordell R, et al. Assessment of breath volatile organic compounds in acute cardiorespiratory breathlessness: a protocol describing a prospective real-world observational study. *BMJ Open* 2019;9(3):e025486.

111. Gisbert JP, Pajares JM. Review article: 13C-urea breath test in the diagnosis of Helicobacter pylori infection -- a critical review. *Aliment Pharmacol Ther* 2004;20(10):1001-17.

112. Kharitonov SA, Yates D, Robbins RA, et al. Increased nitric oxide in exhaled air of asthmatic patients. *Lancet* 1994;343(8890):133-5.

113. Nassar BS, Schmidt GA. Capnography During Critical Illness. *Chest* 2016;149(2):576-85.

114. Phillips M, Gleeson K, Hughes JM, et al. Volatile organic compounds in breath as markers of lung cancer: a cross-sectional study. *Lancet* 1999;353(9168):1930-3.

115. Phillips M. Breath tests in medicine. *Sci Am* 1992;267(1):74-9.

116. Pauling L, Robinson AB, Teranishi R, et al. Quantitative analysis of urine vapor and breath by gas-liquid partition chromatography. *Proc Natl Acad Sci U S A* 1971;68(10):2374-6.

117. Bujak R, Struck-Lewicka W, Markuszewski MJ, et al. Metabolomics for laboratory diagnostics. *J Pharm Biomed Anal* 2015;113:108-20.

118. Haick H, Broza YY, Mochalski P, et al. Assessment, origin, and implementation of breath volatile cancer markers. *Chem Soc Rev* 2014;43(5):1423-49.

119. van de Kant KD, van der Sande LJ, Jobsis Q, et al. Clinical use of exhaled volatile organic compounds in pulmonary diseases: a systematic review. *Respir Res* 2012;13:117.

120. van der Schee MP, Paff T, Brinkman P, et al. Breathomics in lung disease. *Chest* 2015;147(1):224-31.

121. Wilson AD. Advances in electronic-nose technologies for the detection of volatile biomarker metabolites in the human breath. *Metabolites* 2015;5(1):140-63.

122. Tisch U, Haick H. Chemical sensors for breath gas analysis: the latest developments at the Breath Analysis Summit 2013. *J Breath Res* 2014;8(2):027103.

123. de Vries R, Brinkman P, van der Schee MP, et al. Integration of electronic nose technology with spirometry: validation of a new approach for exhaled breath analysis. *Journal of Breath Research* 2015;9(4):10.

124. de Vries R, Dagelet YWF, Spoor P, et al. Clinical and inflammatory phenotyping by breathomics in chronic airway diseases irrespective of the diagnostic label. *Eur Respir J* 2018;51(1).

125. Mondello L, Tranchida PQ, Dugo P, et al. Comprehensive two-dimensional gas chromatography-mass spectrometry: a review. *Mass Spectrom Rev* 2008;27(2):101-24.

126. Smith D, Spanel P, Herbig J, et al. Mass spectrometry for real-time quantitative breath analysis. *J Breath Res* 2014;8(2):027101.

127. Mochalski P, Wiesenhofer H, Allers M, et al. Monitoring of selected skin- and breath-borne volatile organic compounds emitted from the human body using gas chromatography ion mobility spectrometry (GC-IMS). *J Chromatogr B Analyt Technol Biomed Life Sci* 2018;1076:29-34.

128. Bayrakli I. Breath analysis using external cavity diode lasers: a review. *J Biomed Opt* 2017;22(4):40901.

129. Schwoebel H, Schubert R, Sklorz M, et al. Phase-resolved real-time breath analysis during exercise by means of smart processing of PTR-MS data. *Anal Bioanal Chem* 2011;401(7):2079-91.

130. Cope KA, Watson MT, Foster WM, et al. Effects of ventilation on the collection of exhaled breath in humans. *J Appl Physiol (1985)* 2004;96(4):1371-9.

131. Dragonieri S, Schot R, Mertens BJ, et al. An electronic nose in the discrimination of patients with asthma and controls. *J Allergy Clin Immunol* 2007;120(4):856-62.

132. Fens N, van der Schee MP, Brinkman P, et al. Exhaled breath analysis by electronic nose in airways disease. Established issues and key questions. *Clin Exp Allergy* 2013;43(7):705-15.

133. Xia J, Broadhurst DI, Wilson M, et al. Translational biomarker discovery in clinical metabolomics: an introductory tutorial. *Metabolomics* 2013;9(2):280-99.

134. Neerincx AH, Vijverberg SJH, Bos LDJ, et al. Breathomics from exhaled volatile organic compounds in pediatric asthma. *Pediatr Pulmonol* 2017;52(12):1616-27.

135. Rufo JC, Madureira J, Fernandes EO, et al. Volatile organic compounds in asthma diagnosis: a systematic review and meta-analysis. *Allergy* 2016;71(2):175-88.

136. Azim A, Barber C, Dennison P, et al. Exhaled volatile organic compounds in adult asthma: a systematic review. *Eur Respir J* 2019;54(3).

137. Fens N, Zwinderman AH, van der Schee MP, et al. Exhaled breath profiling enables discrimination of chronic obstructive pulmonary disease and asthma. *Am J Respir Crit Care Med* 2009;180(11):1076-82.

138. Montuschi P, Santonico M, Mondino C, et al. Diagnostic performance of an electronic nose, fractional exhaled nitric oxide, and lung function testing in asthma. *Chest* 2010;137(4):790-6.

139. Fens N, Roldaan AC, van der Schee MP, et al. External validation of exhaled breath profiling using an electronic nose in the discrimination of asthma with fixed airways obstruction and chronic obstructive pulmonary disease. *Clin Exp Allergy* 2011;41(10):1371-8.

140. Timms C, Thomas PS, Yates DH. Detection of gastro-oesophageal reflux disease (GORD) in patients with obstructive lung disease using exhaled breath profiling. *J Breath Res* 2012;6(1):016003.

141. Dragonieri S, Quaranta VN, Carratu P, et al. Exhaled breath profiling by electronic nose enabled discrimination of allergic rhinitis and extrinsic asthma. *Biomarkers* 2018.

142. Ibrahim B, Basanta M, Cadden P, et al. Non-invasive phenotyping using exhaled volatile organic compounds in asthma. *Thorax* 2011;66(9):804-9.

143. Meyer N, Dallinga JW, Nuss SJ, et al. Defining adult asthma endotypes by clinical features and patterns of volatile organic compounds in exhaled air. *Respiratory Research* 2014;15:9.

144. Plaza V, Crespo A, Giner J, et al. Inflammatory Asthma Phenotype Discrimination Using an Electronic Nose Breath Analyzer. *J Investig Allergol Clin Immunol* 2015;25(6):431-7.

145. Fens N, van der Sluijs KF, van de Pol MA, et al. Electronic nose identifies bronchoalveolar lavage fluid eosinophils in asthma. *Am J Respir Crit Care Med* 2015;191(9):1086-8.

146. Brinkman P, Wagener AH, Hekking PP, et al. Identification and prospective stability of electronic nose (eNose)-derived inflammatory phenotypes in patients with severe asthma. *J Allergy Clin Immunol* 2018.

147. Schleich FN, Zanella D, Stefanuto PH, et al. Exhaled Volatile Organic Compounds are Able to Discriminate between Neutrophilic and Eosinophilic Asthma. *Am J Respir Crit Care Med* 2019.

148. van der Schee MP, Palmay R, Cowan JO, et al. Predicting steroid responsiveness in patients with asthma using exhaled breath profiling. *Clin Exp Allergy* 2013;43(11):1217-25.

149. Paredi P, Kharitonov SA, Barnes PJ. Elevation of exhaled ethane concentration in asthma. *Am J Respir Crit Care Med* 2000;162(4 Pt 1):1450-4.

150. Bruce C, Chan HP, Mueller L, et al. Effect of hydrofluoroalkane-ethanol inhalers on estimated alcohol levels in asthmatic subjects. *Respirology* 2009;14(1):112-6.

151. Brinkman P, Ahmed WM, Gomez C, et al. Exhaled volatile organic compounds as markers for medication use in asthma. *Eur Respir J* 2020;55(2).

152. Olopade CO, Zakkar M, Swedler WI, et al. Exhaled pentane levels in acute asthma. *Chest* 1997;111(4):862-5.

153. Brinkman P, Wagener AH, Hekking PP, et al. Identification and prospective stability of electronic nose (eNose)-derived inflammatory phenotypes in patients with severe asthma. *J Allergy Clin Immunol* 2019;143(5):1811-20 e7.

154. Lazar Z, Fens N, van der Maten J, et al. Electronic nose breathprints are independent of acute changes in airway caliber in asthma. *Sensors (Basel)* 2010;10(10):9127-38.

155. van der Schee MP, Fens N, Brinkman P, et al. Effect of transportation and storage using sorbent tubes of exhaled breath samples on diagnostic accuracy of electronic nose analysis. *J Breath Res* 2013;7(1):016002.

156. Larstad MA, Toren K, Bake B, et al. Determination of ethane, pentane and isoprene in exhaled air--effects of breath-holding, flow rate and purified air. *Acta Physiol (Oxf)* 2007;189(1):87-98.

157. Ahmed WM, Brinkman P, Weda H, et al. Methodological considerations for large-scale breath analysis studies: lessons from the U-BIOPRED severe asthma project. *J Breath Res* 2018;13(1):016001.

158. Miekisch W, Schubert JK, Noeldge-Schomburg GF. Diagnostic potential of breath analysis--focus on volatile organic compounds. *Clin Chim Acta* 2004;347(1-2):25-39.

159. Chen M-L, Chen S-H, Guo B-R, et al. Relationship between environmental exposure to toluene, xylene and ethylbenzene and the expired breath concentrations for gasoline service workers. *Journal of Environmental Monitoring* 2002;4(4):562-66.

160. Cao W, Duan Y. Breath analysis: potential for clinical diagnosis and exposure assessment. *Clin Chem* 2006;52(5):800-11.

161. Paredi P, Kharitonov SA, Leak D, et al. Exhaled ethane, a marker of lipid peroxidation, is elevated in chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 2000;162(2 Pt 1):369-73.

162. Scholpp J, Schubert JK, Miekisch W, et al. Breath markers and soluble lipid peroxidation markers in critically ill patients. *Clin Chem Lab Med* 2002;40(6):587-94.

163. Lawal O, Ahmed WM, Nijsen TME, et al. Exhaled breath analysis: a review of 'breath-taking' methods for off-line analysis. *Metabolomics* 2017;13(10):110.

164. Silkoff PE, McClean PA, Slutsky AS, et al. Marked flow-dependence of exhaled nitric oxide using a new technique to exclude nasal nitric oxide. *American Journal of Respiratory and Critical Care Medicine* 1997;155(1):260-67.

165. Miekisch W, Kischkel S, Sawacki A, et al. Impact of sampling procedures on the results of breath analysis. *J Breath Res* 2008;2(2):026007.

166. Basanta M, Koimtzis T, Singh D, et al. An adaptive breath sampler for use with human subjects with an impaired respiratory function. *Analyst* 2007;132(2):153-63.

167. Basanta M, Jarvis RM, Xu Y, et al. Non-invasive metabolomic analysis of breath using differential mobility spectrometry in patients with chronic obstructive pulmonary disease and healthy smokers. *Analyst* 2010;135(2):315-20.

168. Phillips M. Method for the collection and assay of volatile organic compounds in breath. *Anal Biochem* 1997;247(2):272-8.

169. Garcia-Morin M, Lopez-Sanguos C, Vazquez P, et al. Lactate Dehydrogenase: A Marker of the Severity of Vaso-Occlusive Crisis in Children with Sickle Cell Disease Presenting at the Emergency Department. *Hemoglobin* 2016;40(6):388-91.

170. Beauchamp J, Herbig J, Gutmann R, et al. On the use of Tedlar(R) bags for breath-gas sampling and analysis. *J Breath Res* 2008;2(4):046001.

171. Stein VB, Narang RS, Wilson L, et al. A simple, reliable method for the determination of chlorinated volatile organics in human breath and air using glass sampling tubes. *J Anal Toxicol* 1996;20(3):145-50.

172. Miekisch W, Schubert JK. From highly sophisticated analytical techniques to life-saving diagnostics: Technical developments in breath analysis. *Trends in Analytical Chemistry* 2006;25(7):665-73.

173. Dettmer K, Engewald W. Adsorbent materials commonly used in air analysis for adsorptive enrichment and thermal desorption of volatile organic compounds. *Anal Bioanal Chem* 2002;373(6):490-500.

174. Tangerman A, Meuwese-Arends MT, van Tongeren JH. New methods for the release of volatile sulfur compounds from human serum: its determination by Tenax trapping and gas chromatography and its application in liver diseases. *J Lab Clin Med* 1985;106(2):175-82.

175. Beale DJ, Jones OA, Karpe AV, et al. A Review of Analytical Techniques and Their Application in Disease Diagnosis in Breathomics and Salivaomics Research. *Int J Mol Sci* 2016;18(1).

176. Lourenco C, Turner C. Breath analysis in disease diagnosis: methodological considerations and applications. *Metabolites* 2014;4(2):465-98.

177. Levitt MD, Ellis C, Furne J. Influence of method of alveolar air collection on results of breath tests. *Dig Dis Sci* 1998;43(9):1938-45.

178. Mardis ER. The $1,000 genome, the $100,000 analysis? *Genome Med* 2010;2(11):84.

179. Leopold JH, Bos LDJ, Sterk PJ, et al. Comparison of classification methods in breath analysis by electronic nose. *Journal of Breath Research* 2015;9(4):12.

180. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015;162(1):W1-73.

181. Cohen JF, Korevaar DA, Altman DG, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open* 2016;6(11):e012799.

182. Horvath I, Barnes PJ, Loukides S, et al. A European Respiratory Society technical standard: exhaled biomarkers in lung disease. *Eur Respir J* 2017;49(4).

183. Dunn WB, Wilson ID, Nicholls AW, et al. The importance of experimental design and QC samples in large-scale and MS-driven untargeted metabolomic studies of humans. *Bioanalysis* 2012;4(18):2249-64.

184. de Lacy Costello B, Amann A, Al-Kateb H, et al. A review of the volatiles from the healthy human body. *J Breath Res* 2014;8(1):014001.

185. Lemfack MC, Gohlke BO, Toguem SMT, et al. mVOC 2.0: a database of microbial volatiles. *Nucleic Acids Res* 2018;46(D1):D1261-D65.

186. van Oort PM, Nijsen T, Weda H, et al. BreathDx - molecular analysis of exhaled breath as a diagnostic test for ventilator-associated pneumonia: protocol for a European multicentre observational study. *BMC Pulm Med* 2017;17(1):1.

187. Broadhurst DI, Kell DB. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics* 2006;2(4):171-96.

188. Pavord I, Bahmer T, Braido F, et al. Severe T2-high asthma in the biologics era: European experts' opinion. *Eur Respir Rev* 2019;28(152).

189. Chung KF. New treatments for severe treatment-resistant asthma: targeting the right patient. *Lancet Respir Med* 2013;1(8):639-52.

190. Freedman DH. *Why Scientific Studies Are So Often Wrong: The Streetlight Effect*. https://www.discovermagazine.com/the-sciences/why-scientific-studies-are-so-often-wrong-the-streetlight-effect (accessed 28/10/2020).

191. Simpson JL, Scott R, Boyle MJ, et al. Inflammatory subtypes in asthma: assessment and identification using induced sputum. *Respirology* 2006;11(1):54-61.

192. Schleich FN, Manise M, Sele J, et al. Distribution of sputum cellular phenotype in a large asthma cohort: predicting factors for eosinophilic vs neutrophilic inflammation. *BMC Pulm Med* 2013;13:11.

193. Hargreave FE, Nair P. The definition and diagnosis of asthma. *Clin Exp Allergy* 2009;39(11):1652-8.

194. Pizzichini E, Pizzichini MM, Efthimiadis A, et al. Indices of airway inflammation in induced sputum: reproducibility and validity of cell and fluid-phase measurements. *Am J Respir Crit Care Med* 1996;154(2 Pt 1):308-17.

195. Djukanovic R, Sterk PJ, Fahy JV, et al. Standardised methodology of sputum induction and processing. *Eur Respir J Suppl* 2002;37:1s-2s.

196. Pavord ID, Shaw DE, Gibson PG, et al. Inflammometry to assess airway diseases. *Lancet* 2008;372(9643):1017-9.

197. Schleich FN, Zanella D, Stefanuto PH, et al. Exhaled Volatile Organic Compounds Are Able to Discriminate between Neutrophilic and Eosinophilic Asthma. *Am J Respir Crit Care Med* 2019;200(4):444-53.

198. Schofield JPR, Burg D, Nicholas B, et al. Stratification of asthma phenotypes by airway proteomic signatures. *J Allergy Clin Immunol* 2019;144(1):70-82.

199. Waikar SS, Betensky RA, Emerson SC, et al. Imperfect gold standards for biomarker evaluation. *Clin Trials* 2013;10(5):696-700.

200. Kim CK, Callaway Z, Kim DW, et al. Eosinophil degranulation is more important than eosinophilia in identifying asthma in chronic cough. *J Asthma* 2011;48(10):994-1000.

201. Carr TF, Berdnikovs S, Simon HU, et al. Eosinophilic bioactivities in severe asthma. *World Allergy Organ J* 2016;9:21.

202. Kyriakopoulos C, Gogali A, Bartziokas K, et al. Identification and treatment of T2-low asthma in the era of biologics. *ERJ Open Res* 2021;7(2).

203. Medrek SK, Parulekar AD, Hanania NA. Predictive Biomarkers for Asthma Therapy. *Curr Allergy Asthma Rep* 2017;17(10):69.

204. Hastie AT, Moore WC, Meyers DA, et al. Analyses of asthma severity phenotypes and inflammatory proteins in subjects stratified by sputum granulocytes. *J Allergy Clin Immunol* 2010;125(5):1028-36 e13.

205. Saffar AS, Ashdown H, Gounni AS. The molecular mechanisms of glucocorticoids-mediated neutrophil survival. *Curr Drug Targets* 2011;12(4):556-62.

206. O'Byrne PM, Metev H, Puu M, et al. Efficacy and safety of a CXCR2 antagonist, AZD5069, in patients with uncontrolled persistent asthma: a randomised, double-blind, placebo-controlled trial. *Lancet Respir Med* 2016;4(10):797-806.

207. Nair P, Aziz-Ur-Rehman A, Radford K. Therapeutic implications of 'neutrophilic asthma'. *Curr Opin Pulm Med* 2015;21(1):33-8.

208. Green BJ, Wiriyachaiporn S, Grainge C, et al. Potentially pathogenic airway bacteria and neutrophilic inflammation in treatment resistant severe asthma. *PLoS One* 2014;9(6):e100645.

209. Huang YJ, Nariya S, Harris JM, et al. The airway microbiome in patients with severe asthma: Associations with disease features and severity. *J Allergy Clin Immunol* 2015;136(4):874-84.

210. Lavelle A, Sokol H. Gut microbiota-derived metabolites as key actors in inflammatory bowel disease. *Nat Rev Gastroenterol Hepatol* 2020;17(4):223-37.

211. Schulz S, Dickschat JS. Bacterial volatiles: the smell of small organisms. *Nat Prod Rep* 2007;24(4):814-42.

212. Thorn RM, Reynolds DM, Greenman J. Multivariate analysis of bacterial volatile compound profiles for discrimination between selected species and strains in vitro. *J Microbiol Methods* 2011;84(2):258-64.

213. Bos LD, Sterk PJ, Schultz MJ. Volatile metabolites of pathogens: a systematic review. *PLoS Pathog* 2013;9(5):e1003311.

214. Wu W, Bleecker E, Moore W, et al. Unsupervised phenotyping of Severe Asthma Research Program participants using expanded lung data. *J Allergy Clin Immunol* 2014;133(5):1280-8.

215. Kuo CS, Pavlidis S, Loza M, et al. T-helper cell type 2 (Th2) and non-Th2 molecular phenotypes of asthma using sputum transcriptomics in U-BIOPRED. *Eur Respir J* 2017;49(2).

216. Deliu M, Sperrin M, Belgrave D, et al. Identification of Asthma Subtypes Using Clustering Methodologies. *Pulm Ther* 2016;2:19-41.

217. Bourdin A, Chanez P. Clustering in asthma: why, how and for how long? *Eur Respir J* 2013;41(6):1247-8.

218. Meyer N, Dallinga JW, Nuss SJ, et al. Defining adult asthma endotypes by clinical features and patterns of volatile organic compounds in exhaled air. *Respir Res* 2014;15:136.

219. Johnson CH, Ivanisevic J, Siuzdak G. Metabolomics: beyond biomarkers and towards mechanisms. *Nat Rev Mol Cell Biol* 2016;17(7):451-9.

220. Johnson CH, Patterson AD, Idle JR, et al. Xenobiotic metabolomics: major impact on the metabolome. *Annu Rev Pharmacol Toxicol* 2012;52:37-56.

221. Phillips M, Cataneo RN, Greenberg J, et al. Effect of age on the breath methylated alkane contour, a display of apparent new markers of oxidative stress. *J Lab Clin Med* 2000;136(3):243-9.

222. Crane MA, Levy-Carrick NC, Crowley L, et al. The response to September 11: a disaster case study. *Ann Glob Health* 2014;80(4):320-31.

223. Bikov A, Paschalaki K, Logan-Sinclair R, et al. Standardised exhaled breath collection for the measurement of exhaled volatile organic compounds by proton transfer reaction mass spectrometry. *BMC Pulm Med* 2013;13:43.

224. Bikov A, Lazar Z, Schandl K, et al. Exercise changes volatiles in exhaled breath assessed by an electronic nose. *Acta Physiol Hung* 2011;98(3):321-28.

225. Cheng ZJ, Warwick G, Yates DH, et al. An electronic nose in the discrimination of breath from smokers and non-smokers: a model for toxin exposure. *J Breath Res* 2009;3(3):036003.

226. Gaugg MT, Engler A, Nussbaumer-Ochsner Y, et al. Metabolic effects of inhaled salbutamol determined by exhaled breath analysis. *Journal of Breath Research* 2017;11(4):046004.

227. Amann A, Mochalski P, Ruzsanyi V, et al. Assessment of the exhalation kinetics of volatile cancer biomarkers based on their physicochemical properties. *J Breath Res* 2014;8(1):016003.

228. Wild CP. Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiol Biomarkers Prev* 2005;14(8):1847-50.

229. Pleil JD, Sheldon LS. Adapting concepts from systems biology to develop systems exposure event networks for exposure science research. *Biomarkers* 2011;16(2):99-105.

230. Miller MR, Hankinson J, Brusasco V, et al. Standardisation of spirometry. *Eur Respir J* 2005;26(2):319-38.

231. ten Brinke A, de Lange C, Zwinderman AH, et al. Sputum induction in severe asthma by a standardized protocol: predictors of excessive bronchoconstriction. *Am J Respir Crit Care Med* 2001;164(5):749-53.

232. Bafadhel M, McCormick M, Saha S, et al. Profiling of sputum inflammatory mediators in asthma and chronic obstructive pulmonary disease. *Respiration* 2012;83(1):36-44.

233. Loza MJ, Djukanovic R, Chung KF, et al. Validated and longitudinally stable asthma phenotypes based on cluster analysis of the ADEPT study. *Respir Res* 2016;17(1):165.

234. Lefaudeux D, De Meulder B, Loza MJ, et al. U-BIOPRED clinical adult asthma clusters linked to a subset of sputum omics. *J Allergy Clin Immunol* 2017;139(6):1797-807.

235. Society BT. BTS_SIGN Guideline for the management of asthma 2019. 2019.

236. Taylor SL, Leong LEX, Choo JM, et al. Inflammatory phenotypes in patients with severe asthma are associated with distinct airway microbiology. *J Allergy Clin Immunol* 2018;141(1):94-103 e15.

237. Van Rossum GaD, Fred L. *Python 3 Reference Manual*: CreateSpace; 2009.

238. Azim A, Freeman A, Lavenu A, et al. New Perspectives on Difficult Asthma; Sex and Age of Asthma-Onset Based Phenotypes. *J Allergy Clin Immunol Pract* 2020;8(10):3396-406 e4.

239. Akuthota P, Busse WW. How Sex and Age of Asthma Onset Influence Difficult Asthma Heterogeneity. *J Allergy Clin Immunol Pract* 2020;8(10):3407-08.

240. Brown T, Jones T, Gove K, et al. Randomised controlled trials in severe asthma: selection by phenotype or stereotype. *Eur Respir J* 2018;52(6).

241. Moore WC, Bleecker ER, Curran-Everett D, et al. Characterization of the severe asthma phenotype by the National Heart, Lung, and Blood Institute's Severe Asthma Research Program. *J Allergy Clin Immunol* 2007;119(2):405-13.

242. Schleich F, Brusselle G, Louis R, et al. Heterogeneity of phenotypes in severe asthmatics. The Belgian Severe Asthma Registry (BSAR). *Respir Med* 2014;108(12):1723-32.

243. Kupczyk M, Haque S, Sterk PJ, et al. Detection of exacerbations in asthma based on electronic diary data: results from the 1-year prospective BIOAIR study. *Thorax* 2013;68(7):611-8.

244. The ENFUMOSA cross-sectional European multicentre study of the clinical phenotype of chronic severe asthma. European Network for Understanding Mechanisms of Severe Asthma. *Eur Respir J* 2003;22(3):470-7.

245. Heaney LG, Busby J, Hanratty CE, et al. Composite type-2 biomarker strategy versus a symptom-risk-based algorithm to adjust corticosteroid dose in patients with severe asthma: a multicentre, single-blind, parallel group, randomised controlled trial. *Lancet Respir Med* 2021;9(1):57-68.

246. Jackson DJ, Busby J, Pfeffer PE, et al. Characterisation of patients with severe asthma in the UK Severe Asthma Registry in the biologic era. *Thorax* 2021;76(3):220-27.

247. Law CM, Marchant JL, Honour JW, et al. Nocturnal adrenal suppression in asthmatic children taking inhaled beclomethasone dipropionate. *Lancet* 1986;1(8487):942-4.

248. Patel L, Wales JK, Kibirige MS, et al. Symptomatic adrenal insufficiency during inhaled corticosteroid treatment. *Arch Dis Child* 2001;85(4):330-4.

249. Russell G. Very high dose inhaled corticosteroids: panacea or poison? *Arch Dis Child* 2006;91(10):802-4.

250. Menzies-Gow A, Gurnell M, Heaney LG, et al. Oral corticosteroid elimination via a personalised reduction algorithm in adults with severe, eosinophilic asthma treated with benralizumab (PONENTE): a multicentre, open-label, single-arm study. *Lancet Respir Med* 2022;10(1):47-58.

251. Cowan DC, Cowan JO, Palmay R, et al. Effects of steroid therapy on inflammatory cell subtypes in asthma. *Thorax* 2010;65(5):384-90.

252. Martin MJ, Zain NMM, Hearson G, et al. The airways microbiome of individuals with asthma treated with high and low doses of inhaled corticosteroids. *PLoS One* 2020;15(12):e0244681.

253. McDonald VM, Fingleton J, Agusti A, et al. Treatable traits: a new paradigm for 21st century management of chronic airway diseases: Treatable Traits Down Under International Workshop report. *Eur Respir J* 2019;53(5).

254. Martin MJ, Beasley R, Harrison TW. Towards a personalised treatment approach for asthma attacks. *Thorax* 2020;75(12):1119-29.

255. Mukherjee M, Stoddart A, Gupta RP, et al. The epidemiology, healthcare and societal burden and costs of asthma in the UK and its member nations: analyses of standalone and linked national databases. *BMC Med* 2016;14(1):113.

256. Bahadori K, Doyle-Waters MM, Marra C, et al. Economic burden of asthma: a systematic review. *BMC Pulm Med* 2009;9:24.

257. O'Byrne PM, Pedersen S, Lamm CJ, et al. Severe exacerbations and decline in lung function in asthma. *Am J Respir Crit Care Med* 2009;179(1):19-24.

258. Westerhof GA, Korevaar DA, Amelink M, et al. Biomarkers to identify sputum eosinophilia in different adult asthma phenotypes. *Eur Respir J* 2015;46(3):688-96.

259. Fowler SJ, Tavernier G, Niven R. High blood eosinophil counts predict sputum eosinophilia in patients with severe asthma. *J Allergy Clin Immunol* 2015;135(3):822-4 e2.

260. Heaney LG, Busby J, Bradding P, et al. Remotely Monitored Therapy and Nitric Oxide Suppression Identifies Nonadherence in Severe Asthma. *Am J Respir Crit Care Med* 2019;199(4):454-64.

261. Couillard S, Pavord ID, Heaney LG, et al. Sub-stratification of type-2 high airway disease for therapeutic decision-making: A 'bomb' (blood eosinophils) meets 'magnet' (FeNO) framework. *Respirology* 2022;27(8):573-77.

262. George L, Brightling CE. Eosinophilic airway inflammation: role in asthma and chronic obstructive pulmonary disease. *Ther Adv Chronic Dis* 2016;7(1):34-51.

263. Chibana K, Trudeau JB, Mustovich AT, et al. IL-13 induced increases in nitrite levels are primarily driven by increases in inducible nitric oxide synthase as compared with effects on arginases in human primary bronchial epithelial cells. *Clin Exp Allergy* 2008;38(6):936-46.

264. Couillard S, Laugerud A, Jabeen M, et al. Derivation of a prototype asthma attack risk scale centred on blood eosinophils and exhaled nitric oxide. *Thorax* 2022;77(2):199-202.

265. Pavord ID, Korn S, Howarth P, et al. Mepolizumab for severe eosinophilic asthma (DREAM): a multicentre, double-blind, placebo-controlled trial. *Lancet* 2012;380(9842):651-9.

266. Castro M, Corren J, Pavord ID, et al. Dupilumab Efficacy and Safety in Moderate-to-Severe Uncontrolled Asthma. *N Engl J Med* 2018;378(26):2486-96.

267. Yancey SW, Keene ON, Albers FC, et al. Biomarkers for severe eosinophilic asthma. *J Allergy Clin Immunol* 2017;140(6):1509-18.

268. Moore WC, Fitzpatrick AM, Li X, et al. Clinical heterogeneity in the severe asthma research program. *Ann Am Thorac Soc* 2013;10 Suppl:S118-24.

269. Telenga ED, Tideman SW, Kerstjens HA, et al. Obesity in asthma: more neutrophilic inflammation as a possible explanation for a reduced treatment response. *Allergy* 2012;67(8):1060-8.

270. Thomas RA, Green RH, Brightling CE, et al. The influence of age on induced sputum differential cell counts in normal subjects. *Chest* 2004;126(6):1811-4.

271. Ducharme ME, Prince P, Hassan N, et al. Expiratory flows and airway inflammation in elderly asthmatic patients. *Respir Med* 2011;105(9):1284-9.

272. Chalmers GW, MacLeod KJ, Thomson L, et al. Smoking and airway inflammation in patients with mild asthma. *Chest* 2001;120(6):1917-22.

273. Sze E, Bhalla A, Nair P. Mechanisms and therapeutic strategies for non-T2 asthma. *Allergy* 2020;75(2):311-25.

274. Azim A, Green B, Lau L, et al. Peripheral airways type 2 inflammation, neutrophilia and microbial dysbiosis in severe asthma. *Allergy* 2021;76(7):2070-78.

275. Sastre B, Rodrigo-Munoz JM, Garcia-Sanchez DA, et al. Eosinophils: Old Players in a New Game. *J Investig Allergol Clin Immunol* 2018;28(5):289-304.

276. Sheshachalam A, Srivastava N, Mitchell T, et al. Granule protein processing and regulated secretion in neutrophils. *Front Immunol* 2014;5:448.

277. Ronchi MC, Piragino C, Rosi E, et al. Do sputum eosinophils and ECP relate to the severity of asthma? *Eur Respir J* 1997;10(8):1809-13.

278. Kjarsgaard M, Adatia A, Bhalla A, et al. Underestimation of airway luminal eosinophilia by quantitative sputum cytometry. *Allergy Asthma Clin Immunol* 2021;17(1):63.

279. Tsuda T, Maeda Y, Nishide M, et al. Eosinophil-derived neurotoxin enhances airway remodeling in eosinophilic chronic rhinosinusitis and correlates with disease severity. *Int Immunol* 2019;31(1):33-40.

280. Pizzichini E, Pizzichini MM, Efthimiadis A, et al. Measuring airway inflammation in asthma: eosinophils and eosinophilic cationic protein in induced sputum compared with peripheral blood. *J Allergy Clin Immunol* 1997;99(4):539-44.

281. Bartoli ML, Bacci E, Carnevali S, et al. Quality evaluation of samples obtained by spontaneous or induced sputum: comparison between two methods of processing and relationship with clinical and functional findings. *J Asthma* 2002;39(6):479-86.

282. Granger V, Zerimech F, Arab J, et al. Blood eosinophil cationic protein and eosinophil-derived neurotoxin are associated with different asthma expression and evolution in adults. *Thorax* 2021.

283. Lee Y, Lee JH, Yang EM, et al. Serum Levels of Eosinophil-Derived Neurotoxin: A Biomarker for Asthma Severity in Adult Asthmatics. *Allergy Asthma Immunol Res* 2019;11(3):394-405.

284. Howarth P, Quirce S, Papi A, et al. Eosinophil-derived neurotoxin and clinical outcomes with mepolizumab in severe eosinophilic asthma. *Allergy* 2020;75(8):2085-88.

285. Hartl S, Breyer MK, Burghuber OC, et al. Blood eosinophil count in the general population: typical values and potential confounders. *Eur Respir J* 2020;55(5).

286. El-Wakkad A, Hassan Nel M, Sibaii H, et al. Proinflammatory, anti-inflammatory cytokines and adiponkines in students with central obesity. *Cytokine* 2013;61(2):682-7.

287. Hilty M, Burke C, Pedro H, et al. Disordered microbial communities in asthmatic airways. *PLoS One* 2010;5(1):e8578.

288. Marri PR, Stern DA, Wright AL, et al. Asthma-associated differences in microbial composition of induced sputum. *J Allergy Clin Immunol* 2013;131(2):346-52 e1-3.

289. Zhang Q, Cox M, Liang Z, et al. Airway Microbiota in Severe Asthma and Relationship to Asthma Severity and Phenotypes. *PLoS One* 2016;11(4):e0152724.

290. Pang Z, Wang G, Gibson P, et al. Airway Microbiome in Different Inflammatory Phenotypes of Asthma: A Cross-Sectional Study in Northeast China. *Int J Med Sci* 2019;16(3):477-85.

291. Simpson JL, Daly J, Baines KJ, et al. Airway dysbiosis: Haemophilus influenzae and Tropheryma in poorly controlled asthma. *Eur Respir J* 2016;47(3):792-800.

292. Huang YJ, Nelson CE, Brodie EL, et al. Airway microbiota and bronchial hyperresponsiveness in patients with suboptimally controlled asthma. *J Allergy Clin Immunol* 2011;127(2):372-81 e1-3.

293. Stiemsma LT, Turvey SE. Asthma and the microbiome: defining the critical window in early life. *Allergy Asthma Clin Immunol* 2017;13:3.

294. Teo SM, Mok D, Pham K, et al. The infant nasopharyngeal microbiome impacts severity of lower respiratory infection and risk of asthma development. *Cell Host Microbe* 2015;17(5):704-15.

295. Bisgaard H, Hermansen MN, Bonnelykke K, et al. Association of bacteria and viruses with wheezy episodes in young children: prospective birth cohort study. *BMJ* 2010;341:c4978.

296. Goleva E, Jackson LP, Harris JK, et al. The effects of airway microbiome on corticosteroid responsiveness in asthma. *Am J Respir Crit Care Med* 2013;188(10):1193-201.

297. Callahan BJ, McMurdie PJ, Rosen MJ, et al. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods* 2016;13(7):581-3.

298. Murali A, Bhargava A, Wright ES. IDTAXA: a novel approach for accurate taxonomic classification of microbiome sequences. *Microbiome* 2018;6(1):140.

299. Wright ES. Using DECIPHER v2.0 to Analyze Big Biological Sequence Data in R. *The R Journal* 2016;8(1):352-59.

300. Schliep KP. phangorn: phylogenetic analysis in R. *Bioinformatics* 2011;27(4):592-3.

301. McMurdie PJ, Holmes S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* 2013;8(4):e61217.

302. Davis NM, Proctor DM, Holmes SP, et al. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome* 2018;6(1):226.

303. McMurdie PJ, Holmes S. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput Biol* 2014;10(4):e1003531.

304. Medicine NNLo. *Basic Local Alignment Search Tool*. https://blast.ncbi.nlm.nih.gov/Blast.cgi (accessed 27/11/2022).

305. Durack J, Lynch SV, Nariya S, et al. Features of the bronchial bacterial microbiome associated with atopy, asthma, and responsiveness to inhaled corticosteroid treatment. *J Allergy Clin Immunol* 2017;140(1):63-75.

306. Rogers GB, Hoffman LR, Carroll MP, et al. Interpreting infective microbiota: the importance of an ecological perspective. *Trends Microbiol* 2013;21(6):271-6.

307. Klebanoff SJ. Myeloperoxidase: friend and foe. *J Leukoc Biol* 2005;77(5):598-625.

308. Parker H, Albrett AM, Kettle AJ, et al. Myeloperoxidase associated with neutrophil extracellular traps is active and mediates bacterial killing in the presence of hydrogen peroxide. *J Leukoc Biol* 2012;91(3):369-76.

309. Power PM, Bentley SD, Parkhill J, et al. Investigations into genome diversity of Haemophilus influenzae using whole genome sequencing of clinical isolates and laboratory transformants. *BMC Microbiol* 2012;12:273.

310. Abdel-Aziz MI, Brinkman P, Vijverberg SJH, et al. Sputum microbiome profiles identify severe asthma phenotypes of relative stability at 12 to 18 months. *J Allergy Clin Immunol* 2021;147(1):123-34.

311. Taylor SL, Ivey KL, Gibson PG, et al. Airway abundance of Haemophilus influenzae predicts response to azithromycin in adults with persistent uncontrolled asthma. *Eur Respir J* 2020;56(4).

312. Taylor SL, Leong LEX, Mobegi FM, et al. Long-Term Azithromycin Reduces Haemophilus influenzae and Increases Antibiotic Resistance in Severe Asthma. *Am J Respir Crit Care Med* 2019;200(3):309-17.

313. Lynch SV. The Lung Microbiome and Airway Disease. *Ann Am Thorac Soc* 2016;13 Suppl 2(Suppl 5):S462-S65.

314. Rigauts C, Aizawa J, Taylor SL, et al. R othia mucilaginosa is an anti-inflammatory bacterium in the respiratory tract of patients with chronic lung disease. *Eur Respir J* 2022;59(5).

315. Murphy TF, Brauer AL, Schiffmacher AT, et al. Persistent colonization by Haemophilus influenzae in chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 2004;170(3):266-72.

316. Gallo MC, Kirkham C, Eng S, et al. Changes in IgA Protease Expression Are Conferred by Changes in Genomes during Persistent Infection by Nontypeable Haemophilus influenzae in Chronic Obstructive Pulmonary Disease. *Infect Immun* 2018;86(8).

317. Wilkinson TMA, Aris E, Bourne SC, et al. Drivers of year-to-year variation in exacerbation frequency of COPD: analysis of the AERIS cohort. *ERJ Open Res* 2019;5(1).

318. Kermani NZ, Pavlidis S, Xie J, et al. Instability of sputum molecular phenotypes in U-BIOPRED severe asthma. *Eur Respir J* 2021;57(2).

319. McShane PJ, Naureckas ET, Tino G, et al. Non-cystic fibrosis bronchiectasis. *Am J Respir Crit Care Med* 2013;188(6):647-56.

320. Porsbjerg C, Menzies-Gow A. Co-morbidities in severe asthma: Clinical impact and management. *Respirology* 2017;22(4):651-61.

321. Lujan M, Gallardo X, Amengual MJ, et al. Prevalence of bronchiectasis in asthma according to oral steroid requirement: influence of immunoglobulin levels. *Biomed Res Int* 2013;2013:109219.

322. Polverino E, Dimakou K, Hurst J, et al. The overlap between bronchiectasis and chronic airway diseases: state of the art and future directions. *Eur Respir J* 2018;52(3).

323. Richardson H, Dicker AJ, Barclay H, et al. The microbiome in bronchiectasis. *Eur Respir Rev* 2019;28(153).

324. Gaga M, Bentley AM, Humbert M, et al. Increases in CD4+ T lymphocytes, macrophages, neutrophils and interleukin 8 positive cells in the airways of patients with bronchiectasis. *Thorax* 1998;53(8):685-91.

325. Dente FL, Bilotta M, Bartoli ML, et al. Neutrophilic Bronchial Inflammation Correlates with Clinical and Functional Findings in Patients with Noncystic Fibrosis Bronchiectasis. *Mediators Inflamm* 2015;2015:642503.

326. Wong C, Jayaram L, Karalus N, et al. Azithromycin for prevention of exacerbations in non-cystic fibrosis bronchiectasis (EMBRACE): a randomised, double-blind, placebo-controlled trial. *Lancet* 2012;380(9842):660-7.

327. Diver S, Richardson M, Haldar K, et al. Sputum microbiomic clustering in asthma and chronic obstructive pulmonary disease reveals a Haemophilus-predominant subgroup. *Allergy* 2020;75(4):808-17.

328. Crisford H, Sapey E, Rogers GB, et al. Neutrophils in asthma: the good, the bad and the bacteria. *Thorax* 2021.

329. Friedman JH. On Bias, Variance, 0/1—Loss, and the Curse-of-Dimensionality. *Data Mining and Knowledge Discovery* 1997;1(1):55-77.

330. Johnstone IM, Titterington DM. Statistical challenges of high-dimensional data. *Philos Trans A Math Phys Eng Sci* 2009;367(1906):4237-53.

331. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;13(11):2498-504.

332. Morris JH, Apeltsin L, Newman AM, et al. clusterMaker: a multi-algorithm clustering plugin for Cytoscape. *BMC Bioinformatics* 2011;12:436.

333. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015;43(7):e47.

334. Chen T, Guestrin C. XGBoost *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; 2016.

335. Bayesian Optimization: Open source constrained global optimization tool for Python [program], 2014.

336. Statheropoulos M, Agapiou A, Georgiadou A. Analysis of expired air of fasting male monks at Mount Athos. *J Chromatogr B Analyt Technol Biomed Life Sci* 2006;832(2):274-9.

337. Fenske JD, Paulson SE. Human breath emissions of VOCs. *J Air Waste Manag Assoc* 1999;49(5):594-8.

338. Peel AM, Wilkinson M, Sinha A, et al. Volatile organic compounds associated with diagnosis and disease characteristics in asthma - A systematic review. *Respir Med* 2020;169:105984.

339. Holden KA, Ibrahim W, Salman D, et al. Use of the ReCIVA device in breath sampling of patients with acute breathlessness: a feasibility study. *ERJ Open Res* 2020;6(4).

340. de Vries R, Brinkman P, van der Schee MP, et al. Integration of electronic nose technology with spirometry: validation of a new approach for exhaled breath analysis. *J Breath Res* 2015;9(4):046001.

341. Alving K, Malinovschi A. Basic aspects of exhaled nitric oxide. In: Horvath I, DeJongste JC (eds.) *EXHALED BIOMARKERS*; 2010 p1-31.

342. Bos LD, Sterk PJ, Fowler SJ. Breathomics in the setting of asthma and chronic obstructive pulmonary disease. *J Allergy Clin Immunol* 2016;138(4):970-76.

343. Holz O, Waschki B, Watz H, et al. Breath volatile organic compounds and inflammatory markers in adult asthma patients: negative results from the ALLIANCE cohort. *Eur Respir J* 2021;57(2).

344. Filipiak W, Ruzsanyi V, Mochalski P, et al. Dependence of exhaled breath composition on exogenous factors, smoking habits and exposure to air pollutants. *J Breath Res* 2012;6(3):036008.

345. van den Velde S, Quirynen M, van Hee P, et al. Halitosis associated volatiles in breath of healthy subjects. *J Chromatogr B Analyt Technol Biomed Life Sci* 2007;853(1-2):54-61.

346. Filipiak W, Sponring A, Filipiak A, et al. TD-GC-MS analysis of volatile metabolites of human lung cancer and normal cells in vitro. *Cancer Epidemiol Biomarkers Prev* 2010;19(1):182-95.

347. Allers M, Langejuergen J, Gaida A, et al. Measurement of exhaled volatile organic compounds from patients with chronic obstructive pulmonary disease (COPD) using closed gas loop GC-IMS and GC-APCI-MS. *J Breath Res* 2016;10(2):026004.

348. Schleich FN, Dallinga JW, Henket M, et al. Volatile organic compounds discriminate between eosinophilic and neutrophilic inflammation in vitro. *J Breath Res* 2016;10(1):016006.

349. Lemfack MC, Nickel J, Dunkel M, et al. mVOC: a database of microbial volatiles. *Nucleic Acids Res* 2014;42(Database issue):D744-8.

350. Mutluoglu M, Uzun G. Pseudomonas infection in a postoperative foot wound. *CMAJ* 2011;183(8):E499.

351. Barankin B, Levy J. Dermacase. Can you identify this condition? Pseudomonas aeruginosa infection. *Can Fam Physician* 2012;58(10):1103-4.

352. Filipiak W, Beer R, Sponring A, et al. Breath analysis for in vivo detection of pathogens related to ventilator-associated pneumonia in intensive care patients: a prospective pilot study. *J Breath Res* 2015;9(1):016004.

353. Gao J, Zou Y, Wang Y, et al. Breath analysis for noninvasively differentiating Acinetobacter baumannii ventilator-associated pneumonia from its respiratory tract colonization of ventilated patients. *J Breath Res* 2016;10(2):027102.

354. Ahmed WM, Fenn D, White IR, et al. Microbial volatiles as diagnostic biomarkers of bacterial lung infection in mechanically ventilated patients. *Clin Infect Dis* 2022.

355. Chawla NV. SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research* 2002;16:321-57.

356. Mammen MJ, Scannapieco FA, Sethi S. Oral-lung microbiome interactions in lung diseases. *Periodontol 2000* 2020;83(1):234-41.

357. Dong J, Li W, Wang Q, et al. Relationships Between Oral Microecosystem and Respiratory Diseases. *Front Mol Biosci* 2021;8:718222.

358. Porter SR, Scully C. Oral malodour (halitosis). *BMJ* 2006;333(7569):632-5.

359. Kamal F, Kumar S, Edwards MR, et al. Virus-induced Volatile Organic Compounds are Detectable in Exhaled Breath During Pulmonary Infection. *Am J Respir Crit Care Med* 2021.

360. Filipiak W, Sponring A, Baur MM, et al. Characterization of volatile metabolites taken up by or released from Streptococcus pneumoniae and Haemophilus influenzae by using GC-MS. *Microbiology (Reading)* 2012;158(Pt 12):3044-53.

361. Bleecker ER, Al-Ahmad M, Bjermer L, et al. Systemic corticosteroids in asthma: A call to action from World Allergy Organization and Respiratory Effectiveness Group. *World Allergy Organ J* 2022;15(12):100726.

362. Siddiqi A, Sethi S. Optimizing antibiotic selection in treating COPD exacerbations. *Int J Chron Obstruct Pulmon Dis* 2008;3(1):31-44.

363. Brown MA, Jabeen M, Bharj G, et al. Non-typeable Haemophilus influenzae airways infection: the next treatable trait in asthma? *Eur Respir Rev* 2022;31(165).

364. Ghebre MA, Pang PH, Diver S, et al. Biological exacerbation clusters demonstrate asthma and chronic obstructive pulmonary disease overlap with distinct mediator and microbiome profiles. *J Allergy Clin Immunol* 2018;141(6):2027-36 e12.

365. Flick SN. Managing attrition in clinical research. *Clinical Psychology Review* 1988;8:499-515.

366. Wilkinson M, White I, Hamshere K, et al. The peppermint breath test: a benchmarking protocol for breath sampling and analysis using GC-MS. *J Breath Res* 2021;15(2).

367. Akobeng AK. Assessing the validity of clinical trials. *J Pediatr Gastroenterol Nutr* 2008;47(3):277-82.

368. Lin W-C, Tsai C-F. Missing value imputation: a review and analysis of the literature (2006–2017). *Artificial Intelligence Review* 2020;53(2):1487-509.

369. Prosperi MC, Sahiner UM, Belgrave D, et al. Challenges in identifying asthma subgroups using unsupervised statistical learning techniques. *Am J Respir Crit Care Med* 2013;188(11):1303-12.

370. Dhindsa K, Bhandari M, Sonnadara RR. What's holding up the big data revolution in healthcare? *BMJ* 2018;363:k5357.

371. Singh G, Mémoli F, Carlsson GE. Topological methods for the analysis of high dimensional data sets and 3d object recognition. *PBG@ Eurographics* 2007;2.

372. Fontanella S, Cucco A, Custovic A. Machine learning in asthma research: moving toward a more integrated approach. *Expert Rev Respir Med* 2021;15(5):609-21.

373. Gao H, McDonnell A, Harrison DA, et al. Systematic review and evaluation of physiological track and trigger warning systems for identifying at-risk patients on the ward. *Intensive Care Med* 2007;33(4):667-79.

374. Kemper J, Kolkman D. Transparent to whom? No algorithmic accountability without a critical audience. *Information, Communication & Society* 2019;22(14):2081-96.

375. Price WN. Big data and black-box medical algorithms. *Sci Transl Med* 2018;10(471).

376. Lundberg SM, Erion G, Chen H, et al. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat Mach Intell* 2020;2(1):56-67.

377. Figueroa RL, Zeng-Treitler Q, Kandula S, et al. Predicting sample size required for classification performance. *BMC Med Inform Decis Mak* 2012;12:8.

378. Ahmed WM, Geranios P, White IR, et al. Development of an adaptable headspace sampling method for metabolic profiling of the fungal volatome. *Analyst* 2018;143(17):4155-62.

379. Demeestere K, Dewulf J, De Witte B, et al. Sample preparation for the analysis of volatile organic compounds in air and water matrices. *J Chromatogr A* 2007;1153(1-2):130-44.

380. Kaur P, Singh A, Chana I. Computational Techniques and Tools for Omics Data Analysis: State-of-the-Art, Challenges, and Future Directions. *Archives of Computational Methods in Engineering* 2021;28(7):4595-631.

381. Berthold HK, Giesen TA, Gouni-Berthold I. The stable isotope ketoisocaproic acid breath test as a measure of hepatic decarboxylation capacity: a quantitative analysis in normal subjects after oral and intravenous administration. *Liver Int* 2009;29(9):1356-64.

382. Li CX, Wheelock CE, Skold CM, et al. Integration of multi-omics datasets enables molecular classification of COPD. *Eur Respir J* 2018;51(5).

383. Ibrahim W, Wilde MJ, Cordell RL, et al. Visualization of exhaled breath metabolites reveals distinct diagnostic signatures for acute cardiorespiratory breathlessness. *Sci Transl Med* 2022;14(671):eabl5849.