Detection of Staphylococcus aureus and Aspergillus fumigatus infections in exhaled breath of cystic fibrosis patients

E.M. Golbach

12 February 2021

Detection of *Staphylococcus aureus* and *Aspergillus fumigatus* infections in exhaled breath of cystic fibrosis patients

Chairman	**Dr.ir. R. Hagmeijer** University of Twente Faculty of Engineering Technology, Faculty of Engineering Fluid dynamics
Technical Supervisor	**Dr.ir. F.H.C. de Jongh** University of Twente Faculty of Engineering Technology, Faculty of Engineering Fluid dynamics
Medical Supervisor	**Dr. J. Altenburg, MD** Amsterdam UMC location AMC Department of Respiratory Medicine
Daily Supervisor	**Msc. A. Lammers** Amsterdam UMC location AMC Department of Respiratory Medicine
Technical Supervisor	**Dr.ir. P. Brinkman** Amsterdam UMC location AMC Department of Respiratory Medicine
Professional Behavior Supervisor	**Drs. B.J.C.C. Hessink-Sweep** University of Twente Faculty of Science and Technology
External member	**Prof.dr. J.G.E. Gardeniers** University of Twente Faculty of Science and Technology, Mesoscale Chemical Systems

Do what you can,
with what you have,
where you are.

Theodore Roosevelt

Erik Golbach
Master Thesis
12 February 2021

University of Twente, Technical Medicine, Faculty of Science and Technology
Amsterdam UMC location AMC, Department of Respiratory Medicine

Abstract

Introduction - Cystic fibrosis (CF) is the most life-threatening monogenic disease in western populations caused by mutation in the CF transmembrane conductance regulator (CFTR). The survival age of new-born CF patients in 2016 was 47.7 years. The disease causes damage to all organs consisting of epithelial cell membranes, mostly affecting the lungs. CF is characterized by thickened mucus in the lungs that can obstruct the airways and hinder the removal of pathogens, causing bacterial infections. The most common bacterial infections in CF patients are *Pseudomonas aeruginosa* (*P. aeruginosa*) and *Staphylococcus aureus* (*S. aureus*), both associated with greater risk of exacerbations, hospitalisations, and decline of lung function. A fast and sensitive screening method for pathogens in the lungs is needed to successfully treat the infections at an early stage and prevent further harm. A possible solution might be the detection of volatile organic compounds (VOCs) in exhaled breath. This study aimed to identify candidate VOCs from literature linked to *S. aureus* and *A. fumigatus* in exhaled breath using the method of gas chromatography-mass spectrometry (GC-MS). The second objective was to assess if these VOCs can be used to distinguish *S. aureus* and *A. fumigatus* and *P. aeruginosa* in exhaled breath.

Methods - To identify candidate VOCs that can possibly distinguish *S. aureus* and *A. fumigatus* in exhaled breath a systematic review of previous literature is conducted. These candidate VOCs are then examined in the *in-vivo* GC-MS breath data, originating from the BioMerieux study, a longitudinal study focused on exhaled breath of CF patients. The VOCs present in the breath samples were then compared to microbiology results (sputum and cough swaps) after which the targeted VOCs are analysed with multiple analyses. This study used univariate (Mann-Whitney U test), multivariate (PLS-DA) and topological (mapper) methods to analyse and assess the performance of breath data to predict pathogen presence.

Results - In a total of 16 articles, 10 candidate VOCs coupled to the presence of *S. aureus* were found, and 6 articles resulted in one compound to link to *A. fumigatus*. Complete data (GC-MS and microbiology) was available from 54 patients, 29 adults and 25 children. Benzaldehyde displayed a significant difference ($p < 0.05$) between the *S. aureus* positive versus negative sputum samples as well as the *S. aureus* chronically infected versus not-chronically infected groups, with a AUROC of respectively 0.61 and 0.62 these groups could not be distinguished. Using PLS-DA and mapper did not deliver group distinguishing results. In the search for a discriminating compound for *A. fumigatus* α-pinene displayed a significant difference ($p < 0.05$) between positive versus negative sputum samples, with a AUROC of 0.62. Applying multivariate analysis strategies, PLS-DA and mapper, did not increase the discriminatory power between groups. Using all the candidate VOCs linked to *S. aureus* and *A. fumigatus* combined with previous research to compounds for *P. aeruginosa* a difference can be seen between groups of samples infected with *S. aureus*, *A. fumigatus*, and uninfected samples and groups that next to these samples also contain multiple infections.

Discussion - Combinations of mass and retention time were significantly different between colonised and not colonised samples for both *S. aureus* and *A. fumigatus*. However, these specific data columns are not confirmed by other mass and retention time combinations linked to these same candidate VOCs making it is less plausible that these VOCs as discriminating. This study was the first to validate previous literature found in a systematic review using a targeted analysis with in vivo GC-MS data for *S. aureus* or *A. fumigatus*. Previous work of Kos et al. performed a similar protocol, using different compound confirmation in their breath samples and statistical methods. For future research, the mapper method could be of use to validate a suspected relationship between compounds and the presence of the pathogens. However it is difficult to quantify such relationships as mapper relies on visual analysis. Using exhaled breath is a non-invasive, low-cost, and time-efficient way of checking patients. However, more research should be done before it could be used in practice. When the performance of these methods will increase, exhaled breath has the right benefits to improve the possibilities in home monitoring.

Conclusion - This study found several components to be associated with *S. aureus* and linked one component with *A. fumigatus* in literature, but was not able to extract significant candidate VOCs to discriminate positive and negative samples for both *S. aureus* and *A. fumigatus* by using GC-MS data.

Acknowledgements

This thesis is the final product of the research that I conducted during a 10 month internship in the department of respiratory medicine in the Amsterdam UMC location AMC in order to obtain my master of science degree at the University of Twente. During this internship I developed myself as a healthcare professional, I learned a lot both clinically and scientifically. While improving these important skills I also learned about myself, working habits, and my future ambitions, making this past 10 months an experience that I am very grateful for. I would like to take a moment to express my gratitude to the people who helped me during this period.

First of all, I would like to thank Marije Lammers for her enthusiasm and pleasant balance in working together while supervising. You were always available for help or opinion if needed. Likewise, I would like to thank Paul Brinkman for teaching me with patience about the technical background and for always coming up with the explanatory metaphors. Together with the other members you made me feel part of the group from the start, I had a great time. Thereby, I would like to thank Frans de Jongh, for the useful insights during the fun (digital) update sessions, which turned out a lot more creative and cultural than I imagined. Additionally, I would like to thank Josje Altenburg, Niels Rutjes and Suzanne Terheggen-Lagro, for giving me the chance to develop myself in the clinical setting of the outpatient clinic and trusting me with their ’Luchtbrug’ platform. Bregje Hessink-Sweep I would like to thank for her contribution to my learning process during two years of internships, learning me to look for new perspectives. Finally, I would like to thank all my friends and family for being supportive towards my ambitions and for being patient during not only this graduation but my whole study time. You made studying worth the effort and the whole period a lot of fun.

I hope you enjoy reading this thesis as much as I enjoyed the internship and my final months as a student. Many thanks again to all of you.

Erik Golbach

Utrecht, February 2021

Abbrevations

AMC	Amsterdam UMC, location Amsterdam Medisch Centrum
AMDIS	Software for GC-MS data interpretation (Automated Mass Spectral Deconvolution and Identification System)
A. fumigatus	Fungus *Aspergillus fumigatus*
AUC	Area under the curve
CAS number	A unique number to identify chemical elements, components, polymers and alloys. CAS is short Chemical Abstracts Service, a division of the American Chemical Society
CF	Cystic Fibrosis
CFTR	Cystic Fibrosis Transmembrane conductance Regulator (both the gene and protein are called this abbreviation)
DT	Desorption tubes
eNose	electronic Nose
GC-MS	Gas Chromatography – Mass Spectrometry
GC-TOF-MS	Gas Chromatography - time of flight - mass spectrometry
HSSE	Headspace sorptive extraction
HS-SPME	Headspace – solid phase microextraction
KEGG	Kyoto Encyclopaedia of Genes and Genomes; a collection of databases dealing with genomes, biological pathways, diseases, drugs, and chemical substances
M/Z-ratio	Mass to charge ratio
NIST	National Institute of Standards and Technology
P. aeruginosa	Bacteria *Pseudomonas aeruginosa*
PLS-DA	partial least squares – discriminant analysis
ROC	Receiver operating characteristic
S. Aureus	Bacteria *Staphylooccus aureus*
SBSE	Stir bar sorptive extraction
SESI-MS	Secondary electrospray ionisation – Mass spectrometry
SIFT-MS	Selected ion flow tube – Mass spectrometry
TDA	Topological data analysis
VOC	Volatile organic component

Contents

List of Figures

List of Tables

1 Introduction **1**

1.1 Background 1

1.2 Clinical Problem 2

1.3 Pathogen detection in breath 3

1.4 Goals and hypothesis 6

2 Methods **7**

2.1 Study Population and data 7

2.2 Part I: Systematic review 9

2.3 Part II: Targeted analysis 10

3 Results **12**

3.1 Part I: Systematic review 12

3.2 Part II: Targeted analysis 15

3.2.1 Characteristics 15

3.2.2 Detection of *S. aureus* 15

3.2.3 Detection of *A. fumigatus* 20

3.2.4 Co-infection analysis using mapper 22

4 Discussion **26**

5 Future recommendations **30**

6 Conclusion **31**

7 References **32**

A Articles systematic review *S. aureus* **38**

B Articles systematic review *A. fumigatus* **39**

List of Figures

1 CF pathogenesis cascade . . . 2
2 Age specific bacteria prevalence in CF patients . . . 3
3 VOC detection methods . . . 4
4 Schematical outline of the study . . . 7
5 Schematical outline of the data selection . . . 9
6 Example of mapper . . . 11
7 Literature review focused on *S. aureus* . . . 12
8 Literature review focused on *A. fumigatus* . . . 14
9 ROC-curves for GC-MS data column X85.858 . . . 17
10 PLS DA executed focusing on *S. aureus* . . . 18
11 Mapper plot X85.858 and *S. aureus* positive versus negative samples . . . 18
12 Mapper plot X85.858 and *S. aureus* chronically versus not chronically infected samples . . . 19
13 ROC-curve for GC-MS data column X53.817 . . . 21
14 PLS DA executed focusing on *A. fumigatus* . . . 22
15 Mapper plot X53.817 and *A. fumigatus* positive versus negative samples . . . 22
16 Mapper analysis of co-infections . . . 23
17 Co-infection analysis for *S. aureus* . . . 24
18 Co-infection analysis for *A. fumigatus* . . . 24
19 Co-infection analysis for *P. aeruginosa* . . . 25
20 Mass spectrum of Benzaldehyde . . . 28
21 Mass spectrum of α-pinene . . . 28

List of Tables

1 Compounds found in the literature linked to *S. aureus* . . . 13
2 Compounds found in the literature linked to *A. fumigatus* . . . 14
3 Patient characteristics . . . 15
4 Sample characteristics . . . 15
5 Mann-Whitney U test results targeted GC-MS columns for *S. aureus* . . . 16
6 Mann-Whitney U test results targeted GC-MS columns for *A. fumigatus* . . . 20

1 Introduction

1.1 Background

Cystic fibrosis (CF) is the most life-threatening monogenic disease in western populations with a median survival age of new-born CF patients of 47.7 years in 2016 [1]. The diagnosis CF is established in more than 70,000 individuals worldwide, of which around 50,000 patients are living in Europe [2]. A mutation in the CF transmembrane conductance regulator (CFTR) gene causes the disease, which is characterised by thick and viscous mucus secretions. These affected secretions are present in the liver, pancreas [3], intestines [4] and most notably the lungs, making CF a multi-organ disease [5,6]. This CFTR gene encodes for cAMP-regulated chloride (Cl^-) and bicarbonate ($HCO3^-$) channels expressed in the membrane of epithelial cells [7]. If these channels are malfunctioning, the amount of water in- and outside the epithelial cells is out of balance, which results in thickened mucoid secretion. The CFTR gene mutations can be divided into six different classes, all with their specific characteristic effects on the CFTR protein functioning. The CFTR protein is either not synthesised (class 1), not processed (class 2), not regulated (class 3), not conducted (class 4), partly produced or processed (class 5) or there is a deregulation of these CFTR channels (class 6) [6]. The most common mutation in CF patients is of the class 2 mutation, more specific the ΔF508 mutation. This mutation causes absence of phenylalanine resulting in a deficiency of the CFTR protein and is accounted for a total of 86.4% of all the CF cases in 2016 [1].

The European Cystic Fibrosis Society (ECFS) states that the neonatal screening for CF done via a heel prick is preferable [8]. In the Netherlands, this advice is carried out on national level since 2011 [9]. This diagnosis is confirmed by a sweat Cl^- test. The damaged CFTR channels cannot provide sufficient migration of Cl^- ions, resulting in more salted sweat on the skin. According to the guidelines from the cystic fibrosis foundation, an amount of $\geq$ 60mmol/L is the limit to definitively diagnose CF with the sweat test [10]. The symptoms which raise the suspicion of CF, mainly focus on the respiratory and gastrointestinal tract [7].

Of all organs consisting of epithelial cell membranes, and thus the CFTR protein, the pathological thickened mucus mostly affects the lungs. CF patients suffer from extensive coughing, lower lung function, and have a higher risk of exacerbations. They are limited in their physical activity and daily routine by their condition. Looking at cell level, the genetic defects start the so-called cystic fibrosis pathogenesis cascade in the lungs (Figure 1), which ultimately leads to respiratory disease [11]. The first step is the lack of CFTR protein leading to deficient chloride ion transport. This malfunctioning regulation leads to less water in the secretion of airway surface liquid, caused by the increased water retention by the ion rich secretion cells. The lower amount of water results in thickened mucus, which can obstruct the airway and eventually cause bacterial infection, inflammation of the tissue and ultimately lung damage [7]. The main cause of the infection is the inability of patients with CF to clear their lungs by coughing up the pathogens.

There are five types of pathogens; viruses, bacteria, fungi, protozoa and worms [12]. A number of these pathogens are present in most people's airways, and not hurtful for healthy persons. The pathophysiological changes in CF patients' lungs result in their inability to clear their lungs properly of pathogens. Some regular pathogens cause a lot of problems in this group of patients. The most common cause of problems in the lungs of CF patients is bacterial infection, specifically *Pseudomonas aeruginosa (P. aeruginosa)* and *Staphylococcus aureus (S. aureus)* [6]. As displayed in figure 2, infection caused by *S. aureus* is more prevalent in younger patients, with even a percentage of more than 60% (with a peak of above 80% in teen years). Bacterial infection by *P. aeruginosa* becomes more prevalent around 30 years, with a peak of more than 70% in the 35-44 years group.

Of these pathogens in CF, only *S. aureus* can be pathogenic in immunocompetent patients. The other bacteria such as *P. aeruginosa* are considered opportunistic pathogens, meaning that the microorganisms are non-harming for healthy hosts and only become a problem in already damaged environments [5]. Both *P. aeruginosa* and *S. aureus* are reported to be associated with a declined lung function [13], increased hospitalisation and lower survival rate [14,15]. Next to

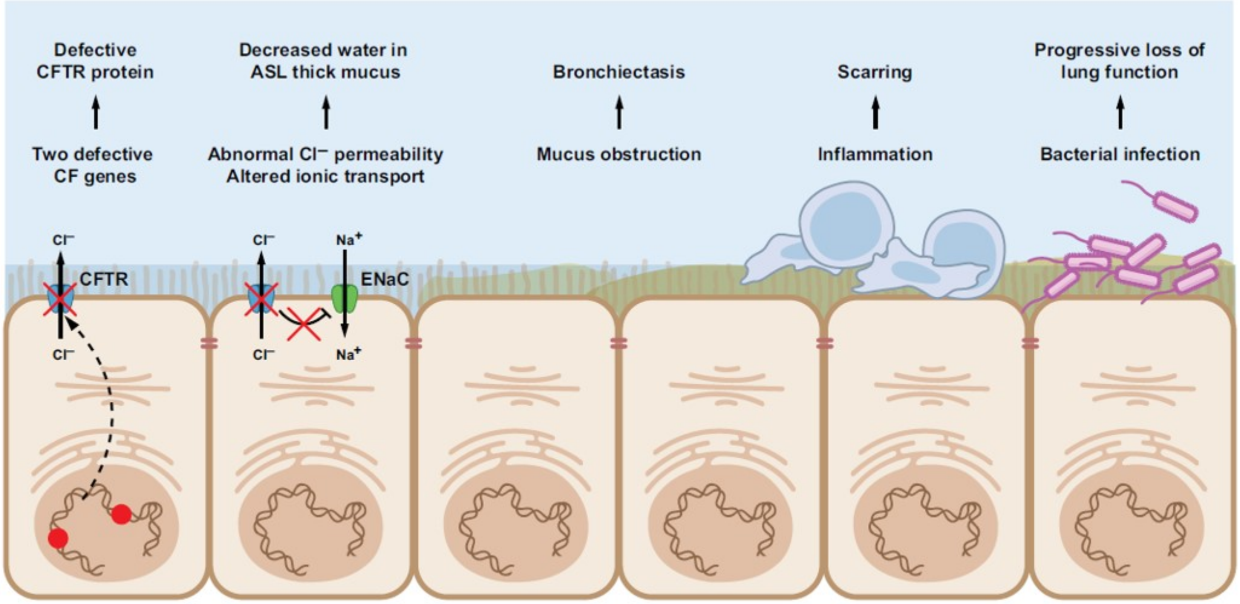


Figure 1: CF pathogenesis cascade. The lack of CFTR (left part) ultimately leading to inflammation and infection (right part)[7]

these bacteria, another opportunistic pathogen known to culture the lungs of CF patients is the fungus *Aspergillus fumigatus* (A. *fumigatus*). *A. fumigatus* colonisation is associated with greater risk of exacerbations, hospitalisations and lung function decline[16].

1.2 Clinical Problem

To date, the presence of pathogens in the lungs are determined through culturing of sputum or throat swabs. With a sensitivity of 78% and specificity of 100% in the case of *P. aeruginosa*, and a sensitivity of 100% and specificity of 63% for *S. aureus* (throat swaps showing even fewer discriminating results for both pathogens) the method lacks accuracy[15]. For CF patients, especially children, coughing up sputum is hard and sometimes even impossible. Moreover, routine culturing takes several days. A more quick and still reliable detection could be useful to predict the pathogens in CF patients.

The importance of early detection is underlined by better long-term outcomes for patients when using earlier treatment. Antibiotic treatment in early *P. aeruginosa* infection has a high eradication of the infection[5]. Due to this antibiotic treatment against the *P. aeruginosa* bacteria the lungs will remain eradicated for a more extended period[17]. Treating the pathogen in a later stadium gives a higher probability of chronic infection, worsening long-term outcomes for the patient. An added difficulty of late treatment of (meanwhile chronic) *P. aeruginosa* is that this pathogen can construct biofilms. The biofilm consists of an extracellular polymeric substance matrix. This so-called mucoid phenotype of *P. aeruginosa* provides a more stable environment for the bacteria, in which it can defend itself against host defences and antibiotic-therapy[18] increasing the chance on morbidity and mortality[19]. Just like *P. aeruginosa*, *S. aureus* in CF patients is also treated with a variety of antibiotics[17]. Treating these pathogens effectively results in improved lung function and higher life expectancy in CF patients[20]. Treatment for patients with established *A. fumigatus* culturing by antifungal therapy has improved the clinical condition while these patients do not respond to antibacterial therapy[21]. Overall, CF patients' clinical outcome benefit from early treatment against *P. Aeruginosa,*

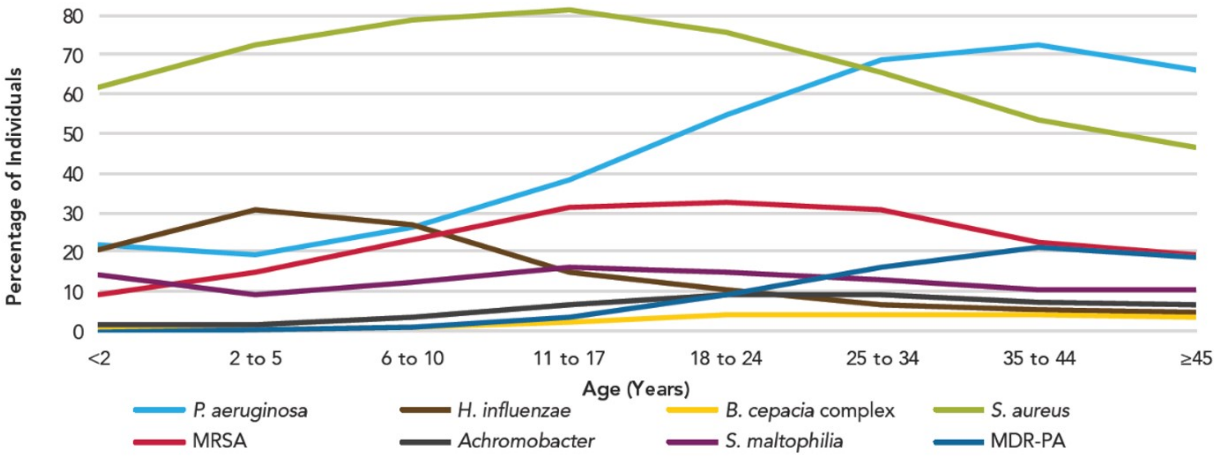


Figure 2: Age-specific prevalence of airway infections caused by bacteria in CF patients in 2016. Showing the high proportional of patients infected with SA at a younger age and PA at a later age. [1]

S. aureus and *A. fumigatus.*

To know when and how to treat the pathogenic infections at an early stage and prevent exacerbations, a fast and sensitive screening method for pathogens in the lungs is needed and could be of great help in treating infections in CF patients.

1.3 Pathogen detection in breath

A possible solution might be the detection of volatile organic compounds (VOCs) in exhaled breath. This method is non-invasive and, unlike sputum collection, almost always feasible for CF patients. VOCs are either produced in the body by all sorts of metabolic processes (endogenous VOCs) or obtained via inhaled air (exogenous VOCs). A mix of these VOCs can be measured in the exhaled breath. The endogenous VOCs in the exhaled breath are partly produced by the lung tissue, making them potential indicators (also called: candidate biomarkers for pulmonary diseases) [22]. Some main methods to analyse these VOCs are gas chromatography (GC) in combination with mass spectrometry (MS) and pattern-based techniques which resemble the mammalian olfactory system (electronic nose). Both techniques are displayed in figure 3 [23].

Electronic nose

A promising, time-efficient, and cost-efficient technique for breath analysis is the electronic nose (eNose) technology. The eNose technology detects a VOC mixture and projects it into a sensor response pattern. The eNose consists of cross-reactive sensors, enabling multiple different VOCs to interact with one sensor, and the other way around, multiple sensors interacting with the same organic compound. This constructed pattern can be compared to previously obtained mixtures of a population to retrieve information about the breath's current underlying pathology. The eNose technology has proven to be of potential use in different kinds of applications [24], especially in a clinical setting [25]. A downside of eNose measurements is the black-box principle, as no individual VOCs can be identified. This makes it challenging or even impossible to understand the chemical processes behind the exhaled breath pattern and which factors play a role in the potential distinction between groups.

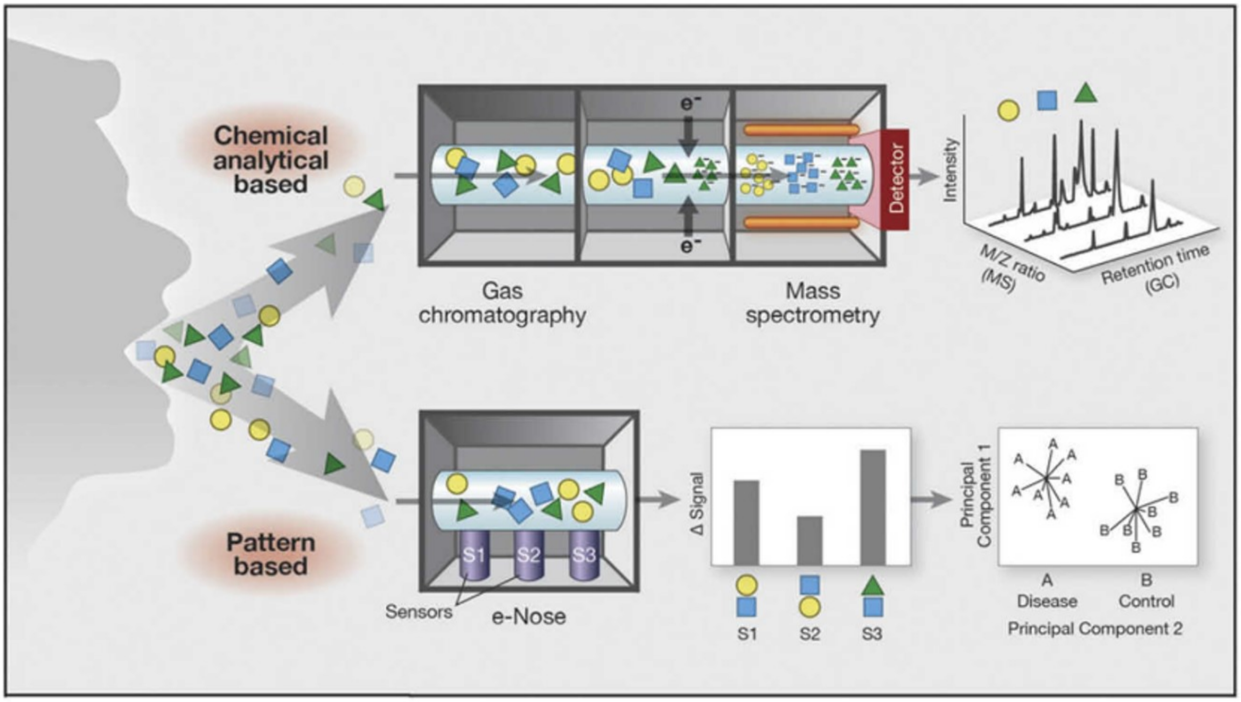


Figure 3: VOC detection methods GC-MS (upper part) and eNose (lower part). The first part of the GC-MS is gas chromatography, and it can be seen in the figure that the chemicals are separated in this part, after which they are ionised by the mass-spectrometry. The retention time is given for every m/z ratio, which can be coupled to volatile compounds. With the eNose method the different sensors of the device all collect values of a breath sample and this is combined in a PCA result, giving a discriminating pattern instead of insights about differentiating volatile compounds [23]

Gas Chromatography-Mass Spectrometry (GC-MS)

Another technique for breath analysis is the GC-MS method, which – in contrast to the pattern-based eNose - identifies chemical compounds based on the retention time (GC) and their ratio of mass to charge (m/z) ratio (MS) [26]. This technique can distinguish individual VOCs, which can be compared manually with an earlier obtained reference library. GC-MS is a broadly used method for metabolic analysis, especially for identification and quantification of individual chemical compounds [26].

The first part of the GC-MS, the gas chromatograph, is used to separate the different VOCs in an exhaled breath sample over time. The first part of the GC-MS, the gas chromatograph, is used to separate the different VOCs in an exhaled breath sample over time, as can be seen in the upper part of figure 3. The separation is realised by sending the breath sample through a long heated column with a mobile phase (i.e. carrier gas – often helium) and a stationary phase (coating on the wall). The interaction of the VOC to the stationary phase determines the mobility of the VOC in the column. A component that has a short time of interaction will elute relatively quickly out of the column [27].

The mass spectrometry (MS) makes it possible to identify the by GC separated compounds from the sample, as can be seen in the right upper part in figure 3. The first step is the ionisation of the mixture. As a result, the VOCs present in the sample will fracture and obtain charge. Next, the ions are accelerated by electric plates and deflected by a magnet. The higher the mass, the lower the deflection of the ion is. These flying ions eventually result in a mass spectrum, which gives information about the mass of the different building blocks out of which this specific VOC is built. GC-MS is beneficial for identifying specific compounds and gives precise results. Disadvantages are that this method is expensive, labour-intensive, and takes time[28].

A quick, easy and accurate method to identify present pathogens and hereby prevent upcoming bacterial infections could be of great added value in CF-specialised healthcare. Using VOCs seems to be a promising procedure that is worth investigating.

1.4 Goals and hypothesis

The goals of this study are:

I. To evaluate which VOCs have been associated with *S. aureus* and *A. fumigatus* in previous literature

II. To determine the diagnostic potential of these compounds for detection of *S. aureus* and *A. fumigatus* in exhaled breath of CF patients

III. To determine if associated VOCs can make a distinction between patients with *S. aureus*, *A. fumigatus,* and *P. aeruginosa*

The primary outcome of this study is to evaluate if GC-MS data linked to individual VOCs can be used to detect and predict *S. aureus* and *A. fumigatus* colonisation in exhaled breath. As a secondary outcome, we want to know if these methods can be used to distinguish *S. aureus*, *A. fumigatus* and *P. aeruginosa*.

Previous work in our group resulted in promising results as it concerns the detection of *P. aeruginosa* through exhaled VOCs[29]. Therefore, we hypothesize that exhaled VOCs measured by GC-MS can be used to detect and identify *S. aureus* and *A. fumigatus* in exhaled breath and that these VOCs can discriminate patients colonised by *S. aureus, A. fumigatus* and *P. aeruginosa.* For this study, we will make use of data that originates from the BioMerieux study, an observational 1-year follow-up study on exacerbations among CF patients able to perform breath measurements. Study visits were performed at Amsterdam UMC (AMC and VUmc) and took place between 2013 and 2015.

2 Methods

A targeted analysis was conducted to identify compounds that can distinguish *S. aureus* and *A. fumigatus* in exhaled breath. More precisely, first, potential VOC candidates are sought after which their predictive value is checked by statistic tests instead of looking for statistical findings in a big dataset. In order to execute this targeted analysis, it was first needed to identify candidate VOCs to look for in the GC-MS data. A systematic review of the literature provided this information. The targeted analysis steps are the same for *S. aureus* and *A. fumigatus* but conducted separately for both pathogens. In figure 4, the outline of the research is shown schematically.

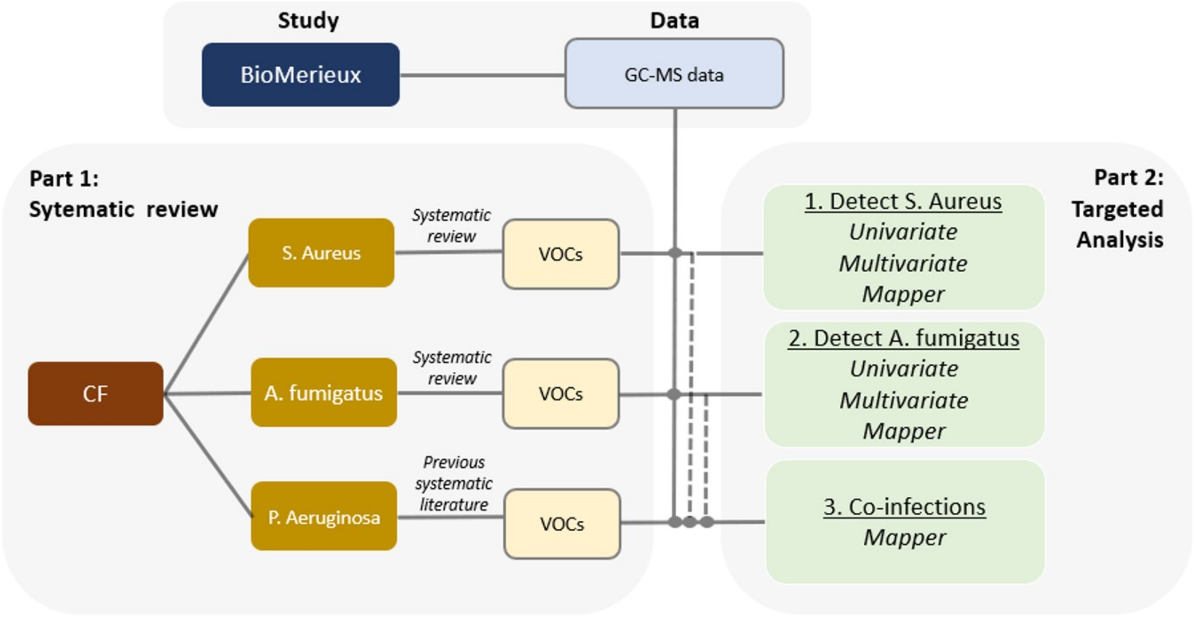


Figure 4: Schematical outline of the study. In the first part a systematic review is performed; focusing on patients with established cystic fibrosis a literature search is performed for *S. aureus*, *A. fumigatus* and *P. aeruginosa* to determine candidate VOCs. These candidate VOCs were then checked in the GC-MS data of the BioMerieux dataset. VOCs that were both present in the dataset and in more than 2 articles resulting from the literature search aimed on *S. aureus* and *A. fumigatus*, are used to perform univariate and multivariate data analysis. With all columns combined a topological data analysis for coinfections is performed.

2.1 Study Population and data

The data used in this study all originates from the 'BioMerieux' study. The 'BioMerieux' study is a longitudinal observational study which investigated the exhaled VOCs in CF patients (adult and paediatric) from the Amsterdam UMC (Amsterdam, the Netherlands) using GC-MS analysis and clinical data. Patients were followed for one year, in which exhaled breath and sputum or cough samples were collected during their 3-monthly regular outpatient visits with maximally five visits per patient. CF patients can have extra visits in case of exacerbation. Patients were included based on mutations in one or both CFTR alleles, cystic fibrosis-related clinical symptoms and chloride sweat test. They had to be stable for at least six weeks (no exacerbations) and able to perform lung function tests. Patients younger than 18 years old were considered children and patients of at least 18 years old were considered adults. Children were included from the point that they were able to perform the breath experiments. Exclusion criteria were mental retardation, CF-related diabetes, smoking, waiting for lung transplantation or participation in other studies or inability to perform requested manoeuvres or measurements.

Data

Sputum samples

Sputum samples were taken of all patients, for children and patients unable to expectorate sputum a cough swab was obtained. The sputum samples and cough swabs were part of standard CF clinical care and used to test for infections with pathogens. Samples were considered positive for a pathogen *(S. aureus*, *A. fumigatus* or *P. aeruginosa*) when the test returned positive for this specific pathogen in the concerned sputum sample for this visit, if the label was not positive the label for this sample was considered negative. Patients were considered chronically colonised if 50% of all the samples in the last year before the study returned positive. If they were not chronically colonised, the patient was considered not-chronically colonised.

Breath samples

Breath samples were collected using a nose clip to prevent breathing through the nose. First, a wash-out period took place where patients breathed through a mouthpiece connected to a VOC filter (A1-filter, North, N7500-1U) for 5 minutes continuously, so that the external air is completely filtered from the VOCs present in the environment. This was followed by an inspiratory capacity manoeuvre, after which the inhaled air was exhaled into a nalophan bag. A total of 500mL of exhaled air was sampled in desorption tubes (TD tubes) filled with Tenax (GR60/80 Interscience, Breda, The Netherlands) using a peristaltic pump at a flow pace of 250mL/min for 2 minutes, the sampling tubes were stored in a refrigerator.

GC-MS analysis

Using a thermal desorption unit (TD100, Markes, Cincinnati, Ohio, USA), the TD tubes were heated to 280°C for 15 minutes with a flow of 30 ml/min which enabled release of the captured VOCs. After cold trapping compounds at 10°C, they were quickly heated to 300°C for one minute. Compounds were injected through a transfer line at a speed of 1.2mL/min and a temperature of 180°C, using an Intercap 5MS/Sil GC column (30m", 0.25mm ID, 1μm film thickness, 1,4-bis(dimethylsiloxy)phenylene dimethyl polysiloxane, Restek, Breda, The Netherlands). The isothermal temperature remained 40°C for 5 minutes, after which it was increased in steps of 10°C/min to 280°C. Molecules were ionised using electron ionisation at 70 eV. A quadrupole mass-spectrometer (GC-MS-GP2010, Shimadzu, Den Bosch, The Netherlands) was used to detect fragment ions at a scan range of 37-300 Da.

GC-MS data

The results of the GC-MS analysis are stored in an Excel-database, whereby every row represents an analysed breath sample. Each column shows the intensity for a m/z-value (ion) that was detected at a specific retention time. The columns are named by the mass and retention time. Column 'X45.127' for example is the data column that indicates the compounds' mass of 45 molecular mass and retention time of 127 seconds. These columns are the columns referred to when linking data columns to a volatile organic compound.

2.2 Part I: Systematic review

The first step of this study was a systematic review, in which VOCs linked to the pathogens (*S. aureus* and *A. fumigatus*) in previous literature were searched; the VOCs had to be linked to a pathogen at least in two separate articles. The literature search resulted in two sets of compounds, for which all linked GC-MS data columns were selected. A more detailed description of the systematic literature search and following steps are described below.

Systematic literature review

A systematic search was conducted using the terms "volatile organic compounds [MESH] AND s. aureus [MESH]" and terms "volatile organic compounds [MESH] AND a. fumigatus [MESH]" in PubMed. Articles were selected based on the title, abstract or tables containing volatile compounds produced by the pathogens. No publication date cut-off was used. At least two researchers performed selection steps. If there was no agreement between researchers whether to include or exclude an article, the article was discussed until an agreement was found. Next, the complete text of the included articles was read. If there were no specific VOCs reported specific for *S. aureus* or *A. fumigatus*, the articles were excluded.

Compounds

Subsequently, chemical names of volatile compounds, structural formulas and detection methods were listed. CAS registration numbers, i.e., unique numerical identifying numbers found using the chemical standards reference database (version: 2008) of the National Institute of Standards and Technology (NIST), were assigned to each compound to prevent the same compound's occurrence in the list twice.

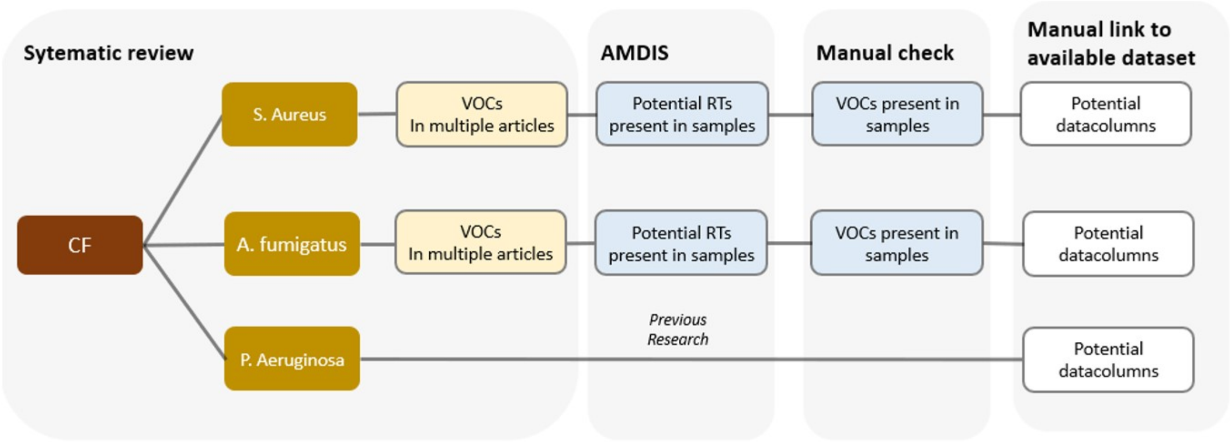


Figure 5: Schematical outline of the selection of data columns in the BioMerieux dataset. From left to right the results of literature research make it possible to distil the potential data by first making use of AMDIS software, after which a manual check leads to definitive data columns to include in the targeted analysis.

Compound analysis of GC-MS data from BioMerieux

The available GC-MS data consists of data columns with retention times in combination with a mass-charge-ratio. A schematic outline of the procedure from potential candidate compounds to GC-MS data is shown in figure 5. To link the compounds found in the literature to the available GC-MS data, first the retention times linked to the VOCs of interest needed to be found. This was done by using an automated batch job in AMDIS software, that identifies the candidate compounds in the samples of BioMerieux data and the coupled retention times. The batch job analysis compared the known mass spectra to the information stored in the breath samples of the BioMerieux samples. To ensure the accuracy of the batch job results, the list of compounds in the samples were checked manually thereafter. The manual check was done by looking at the mass spectral information in the results of the batch job and checking if corresponding retention times were realistic for that compound. If these steps all had a positive outcome, the compounds were confirmed as prevalent in the breath data. Retention times linked to these prevalent compounds were now manually linked to mass-to-charge-ratios based on their characteristic mass-spectra peaks. For each candidate compound was checked which relevant columns were present in the dataset. These data columns were selected for part 2 of the study, the targeted analysis.

2.3 Part II: Targeted analysis

This study's outcomes can be divided into three subsections of analysis, all with their own statistical tools.

1. The detection of *S. aureus* using **univariate**, **multivariate** and **topological data analysis**
   (a) Positive samples vs. negative samples
   (b) Positive and chronic infected samples vs. negative samples
2. The detection of *A. fumigatus* using **univariate**, **multivariate** and **topological data analysis**
   (a) Positive samples vs. negative samples
3. The distinction between *S. aureus, A. fumigatus,* and *P. aeruginosa* using **topological data analysis**

The clinical characteristics and GC-MS data were analysed using R studio (RStudio Inc., Boston, United States of America; Version 1.2.1335) in combination with R-libraries (mixomics, ROC, tidyverse, ggplot2, TDAmapper, devtools, forcats, reshape2, dbstats, tableone, plotROC, Rcolorbrewer, rms, dplyr, cowplot, sva, limma, pROC, lattice, lme4, vegan, locfit, igraph)

Univariate analysis

Univariate analysis was performed on each separate column linked to the compounds resulting from the systematic review. This was done by performing a Mann–Whitney U test. The Mann-Whitney U test is suitable to check if two independent groups are comparable as a group but shifted in their values. A p-value of under 0.05 was considered to indicate a significant difference between groups.

Multivariate analysis

Partial Least Squares Discriminant Analysis (PLS-DA) was performed on all columns linked to all found compounds for the multivariate analysis. PLS-DA is used to find the largest covariance between two labelled groups, i.e. can one distinguish between two given groups based on the measured variables. The labels, in this case, were based on the sputum culture and cough swaps. Due to the targeted character of PLS-DA, there is risk for overfitting, therefore results should be interpreted with caution. Preferably external validation is performed after this type of modelling.

Topological data analysis (TDA)

Another analysis that was included is TDA. This method is based on the topology of a dataset, it connects datapoints in a point cloud by using information about their geometric shape. TDA is useful for complicated large multidimensional data sets and has already proven to be of use in the field of (micro)biology, for example in genomic data[30].

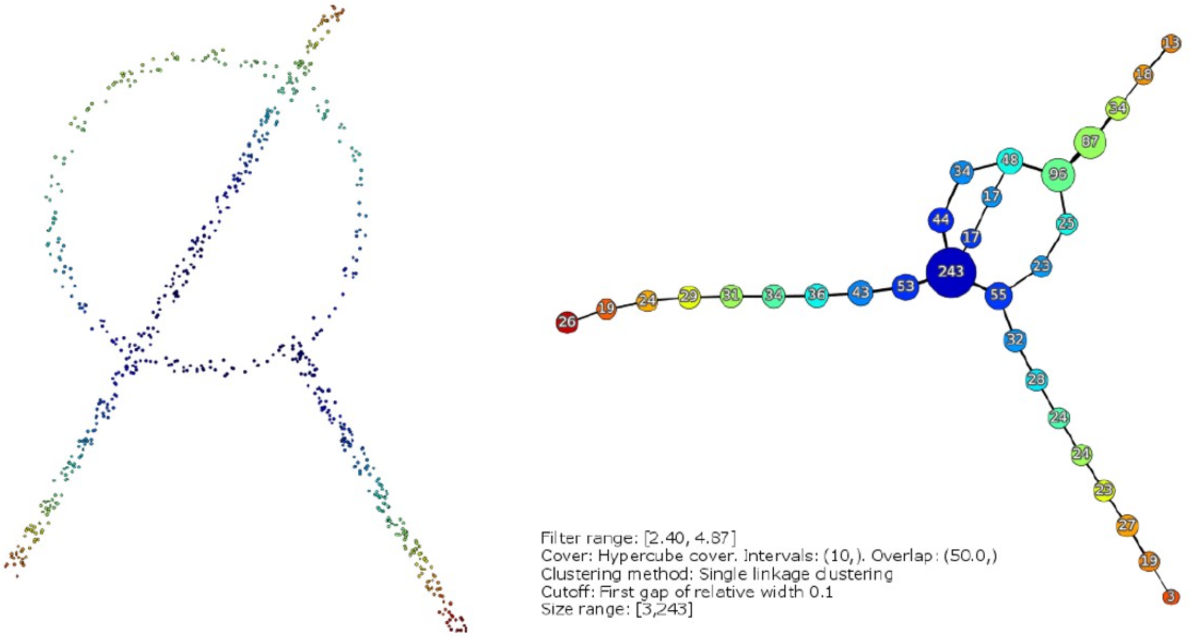
To combine all numerical values of a data point, the Euclidean distance is used to create proximity between the data points. The Euclidean distance gives a distance (d) between data points (p and q) with given coordinates of infinite dimensions (n), see equation 1.

$$d(\mathbf{p},\mathbf{q}) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + (p_3 - q_3)^2 + (p_4 - q_4)^2} \tag{1}$$

When the Euclidean distance calculates distances of all these data points, a point cloud is formed, which can be used to look into shape information.[30]

Mapper

In this study, the topological analysis method used is called mapper[31]. This method is a popular way to visualise the structure of a complicated high dimensional dataset. Specific for the mapper function is the filter that is applied after calculating the distance. The filter used in this analysis is a kernel density estimator (KDE). This filter estimates the probability density in the datapoints without assuming a gaussian distribution. Other groups of filters can be divided in focusing on eccentricity or Laplacian operators[31]. The mapper function can be best explained by using a visualisation. In figure 5 an example is shown of the mechanism used in mapper, in this example executed in python[32]. The algorithm simplifies the point cloud on the left into a more structured map of dots on the right side. Note that this example is in two dimensions, making it easier to grasp the concept of what is happening. In practice, datasets have more dimensions, resulting in the fact that the constructed map cannot be compared to the point cloud.



Filter range: [2.40, 4.87]
Cover: Hypercube cover. Intervals: (10,). Overlap: (50.0,)
Clustering method: Single linkage dustering
Cutoff: First gap of relative width 0.1
Size range: [3,243]

Figure 6: Visualisation of an example of the mapper algorithm. A point cloud is displayed on the left, on the right the same data is shown but this time by making use of the mapper algorithm. This example is from a paper of Mullner et al, has nothing to do with this study and functions merely as an example[32]

3 Results

3.1 Part I: Systematic review

S. aureus

Systematic literature review

Thirty-four articles were listed as a result of the search query in PubMed. Next, the articles were selected based on title, abstract or tables containing volatile compounds produced by *S. aureus*. The literature review resulted in the inclusion of sixteen articles. A flow chart of literature search process is displayed in figure 7. A complete list of articles is included in appendix A.

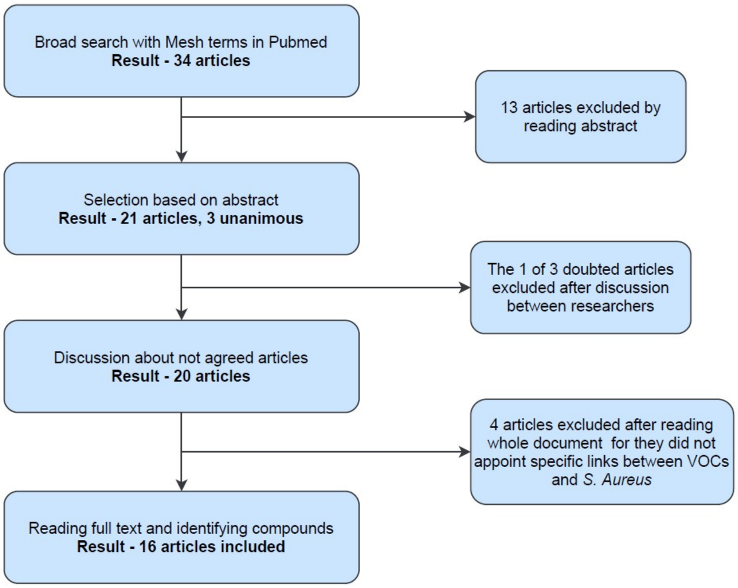

Broad search with Mesh terms in Pubmed
Result - 34 articles
13 articles excluded by reading abstract
Selection based on abstract
Result - 21 articles, 3 unanimous
The 1 of 3 doubted articles excluded after discussion between researchers
Discussion about not agreed articles
Result - 20 articles
4 articles excluded after reading whole document for they did not appoint specific links between VOCs and S. Aureus
Reading full text and identifying compounds
Result - 16 articles included

Figure 7: Flow chart of the systematic literature review focused on *S. aureus*

Compounds

A total of 145 compounds, of which 129 unique CAS numbers, were identified in all literature articles. Twenty-seven of these compounds were designated in at least two articles, and thus used in further steps to look for linked retention times. These compounds and their number of designations in articles are listed in table 1.

Compound analysis of GC-MS data

Out of 27 compounds found in multiple articles, 10 compounds were identified in the BioMerieux patients' breath samples using AMDIS batch job analysis and manual checking. Out of 10 identified compounds, 8 were possible to link to a total of 27 m/z-columns of the GC-MS database. These 27 columns were extracted from the database and analysed using R Studio.

Table 1: Compounds found in the literature linked to *S. aureus*

Name [1]	CAS [2]	M/z [3]	Number of mentions [4]
Acetaldehyde	75-07-0	44	3
ethanol	64-17-5	46	4
formic acid	64-18-6	46	2
Aceton	67-64-1	58	3
acetoin	513-86-0	58	2
acetic acid	64-19-7	60	3
2-butanone	78-93-3	72	2
2-methyl-1-propanol	78-83-1	74	3
3-methyl-butanal	590-86-3	86	5
2,3-butanedione	431-03-8	86	3
2-methyl-butanal	96-17-3	86	2
Pentanal	110-62-3	86	2
3-methyl-1-butanol	123-51-3	88	5
1-methyl-propylhydrazine	4986-49-6	88	2
butanoic acid	107-92-6	88	2
ethyl acetate	141-78-6	88	2
dimethyldisulfide	624-92-0	94	2
3-methyl-butanoic acid	503-74-2	102	5
Benzaldehyde	100-52-7	103	4
ethyl butanoate	105-54-4	116	2
n-butyl acetate	123-86-4	116	2
1H-indole	120-72-9	117	2
isopentyl acetate	123-92-2	130	3
ethyl isovalerate	108-64-5	130	2
2-methyl napthalene	91-57-6	142	2
Undecan-2-one	112-12-9	170	2
1,3,5,7-tetraazatricyclo[3.3.1.1]decane	60168-84-5	262	2

Compounds are displayed with their most-used chemical name [1], their CAS registration number [2], their mass-to-charge ratio [3] and the amount of mentions in articles as a result of the systematic review of literature [4]

A. fumigatus

Systematic literature review

Ten articles were listed as a result of the search query in PubMed. Next, the articles were selected based on title, abstract or tables containing volatile compounds produced by *A. fumigatus*. No publication date cut-off was used. The result of the literature review is inclusion of 8 articles, the process of the review is displayed in figure 8. The complete list of articles is included in appendix B.

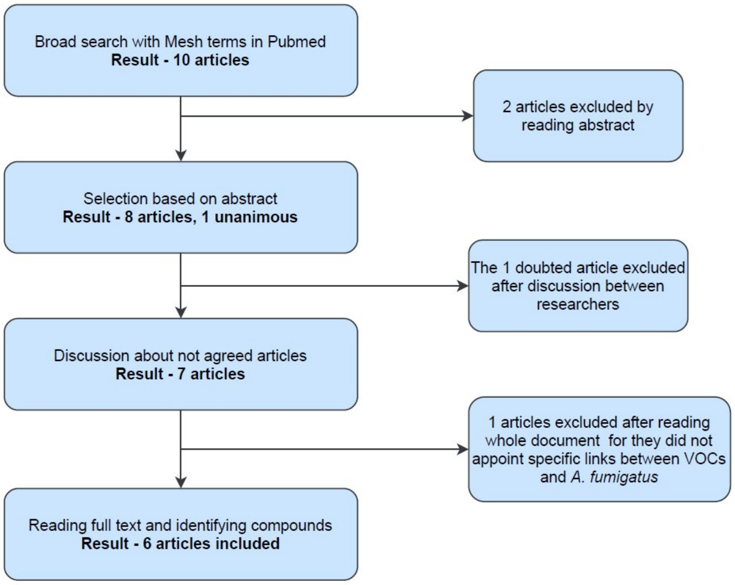

Broad search with Mesh terms in Pubmed
Result - 10 articles
2 articles excluded by reading abstract
Selection based on abstract
Result - 8 articles, 1 unanimous
The 1 doubted article excluded after discussion between researchers
Discussion about not agreed articles
Result - 7 articles
1 articles excluded after reading whole document for they did not appoint specific links between VOCs and A. fumigatus
Reading full text and identifying compounds
Result - 6 articles included

Figure 8: Flow chart of the systematic literature review focused on *A. fumigatus*

Compounds

A total of 54 compounds, of which 33 unique CAS numbers, were identified in the literature search focusing on *A. fumigatus*. Five of these compounds were designated in at least two articles and were used to find linked retention times. These compounds and their number of designations in articles is listed are table 2.

Table 2: Compounds found in the literature linked to *A. fumigatus*

Name [1]	CAS [2]	M/z [3]	Number of mentions [4]
α-pinene	67762-73-6	93	2
camphene	79-92-5	93	2
α-trans-bergamotene	13474-59-4	119	2
β-trans-bergamotene	15438-94-5	119	2
2-pentylfuran	3777-69-3	81	2

Compounds are displayed with their most-used chemical name [1], their CAS registration number [2], their mass-to-charge ratio [3] and the amount of mentions in articles as a result of the systematic review of literature [4]

Compound analysis of GC-MS data

Out of 5 noted compounds mentioned in multiple articles, 1 compound was identified in the BioMerieux patients' breath samples using AMDIS analysis and a manual check. From this one compound it was possible to link to a total of 20 m/z-columns of the GC-MS database. These 20 columns were extracted from the database and analysed using R Studio.

3.2 Part II: Targeted analysis

3.2.1 Characteristics

In this study, the GC-MS data and sputum culture data of 54 patients, of whom 29 adults and 25 children, is used with a total of 204 visits along with 204 samples from both breath and sputum. Patient characteristics are described in table 3, the colonisation of all samples is noted in table 4. Both positive sputum samples as well as samples of chronic infected patients by *S. aureus* are more prevalent compared to *P. aeruginosa* and *A. fumigatus*, especially in children. In the adult group the colonisation tends to shift in presence from *S. aureus* to *P. aeruginosa*.

Table 3: Patient characteristics

	Total (n = 54)	**Adults (n = 29)**	**Children (n = 25**
Gender - Male [N (%)]	30 (55.6)	17 (59)	13 (52)
Age [Mean (SD)]	20.9 (14.0)	30.1 (13.0)	10.1 (3.1)
Chronic culture Positive [N (%)]			
Pseudomonas aeruginosa	13 (24.1)	9 (31.0)	4 (16)
Aspergillus fumigatus	6 (11.1)	5 (17.2)	1 (4)
Staphylococcus aureus	27 (50.0)	13 (44.8)	14 (56)
Number of visits [Mean (SD)]	3.8 (0.9)	3.9 (1.0)	4.1 (0.8)

Data is presented numerical (N (%)) or as mean (N (SD)); SD = standard deviation

Table 4: Sample characteristics

	Total (n = 54)	**Adults (n = 29)**	**Childs (n = 25)**
Samples [N]	204	110	94
Visit culture Positive [N (%)]			
Pseudomonas aeruginosa	51 (25.4)	35 (31.8)	16 (17.0)
Aspergillus fumigatus	35 (17.2)	22 (20)	13 (13.8)
Staphylococcus aureus	98 (48.0)	46 (41.8)	52 (55.3)

Data is presented numerical (N (%)) or as mean (N (SD)); SD = standard deviation

3.2.2 Detection of *S. aureus*

Univariate analysis

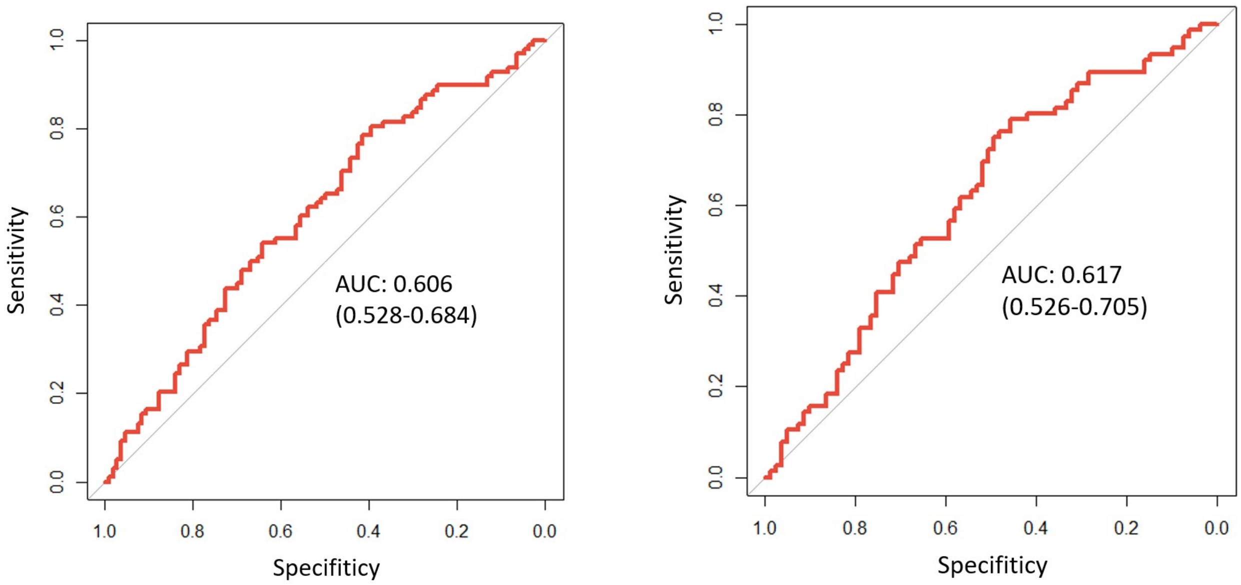
For all 27 columns linked to the eight compounds, the Mann-Whitney-U test's resulting scores - for both the positive versus negative samples and chronically versus not chronically infected patients' sputum samples - are given in table 5. A p-value of under 0.05 is considered to display a significant difference between the two groups. In column 'X85.858', a significant difference is present for positive versus negative samples and chronically versus not chronically infected patients (*S. aureus*) with p-values both under 0.05. No other column in this analysis showed statistically significant results.

Table 5: Mann-Whitney U test results targeted GC-MS columns for *S. aureus*

Compound	**Columns**	**Positive vs. negative**	**Chronic vs. Not chronic**
		P - values	**P - values**
ethanol	x45.127	0.90	0.55
	x46.127	0.97	0.57
	x47.127	0.95	0.68
formic acid	x48.127	0.42	0.86
aceton	x43.141	0.76	0.49
	x44.140	0.64	0.39
	x57.140	0.53	0.35
	x58.140	0.75	0.45
	x59.141	0.47	0.23
2-butanone	x43.221	0.26	0.16
	x44.230	0.20	0.06
2,3-butanedione	x56.213	0.89	0.93
	x86.222	0.84	0.59
butanoic acid, 3-methyl	x87.327	1.00	0.84
	x88.326	0.79	0.90
Benzaldehyde	x77.863	0.07	0.53
	x78.865	0.11	0.43
	x85.858	<0.05	<0.05
	x105.860	0.46	0.71
	x106.863	0.15	0.48
	x107.863	0.09	0.30
1-butanol, 3-methyl-	x70.1007	0.77	0.88
	x112.1008	0.31	0.40
	x70.1127	0.73	0.48
	x97.1127	0.74	0.46
	x100.1127	0.69	0.36
	x112.1127	0.87	0.61

Results are displayed with the most-used chemical name of the compound [1], the data column which gives information about the m/z ratio and retention time [2] and the p-values for statistical tests of both groups [3,4] . Breath samples are considered positive when the result of sputum sampling was positive for *S. Aureus* and negative when its result was negative. The sample was considered chronic if the patient linked to the sample is considered chronically infected with *S. aureus* and not chronic if this was not the case.

ROC analysis showed that positive versus negative breath samples and chronic versus not chronic samples could be discriminated from each other with an AUC of respectively 0.61 (CI 0.53-0.68) and 0.62 (CI 0.53-0.71) (figure 9).



(a) ROC Curve for predicting *S. aureus* positive samples versus negative samples using X85.858

(b) ROC Curve for predicting *S. aureus* chronically versus not-chronically infected patients using X85.858

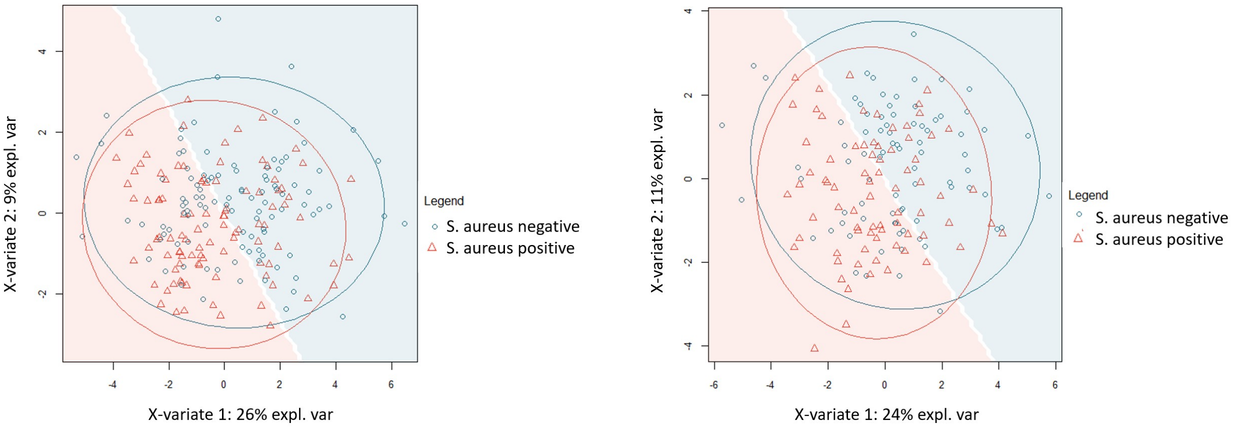
Figure 9: ROC-curves for GC-MS data column X85.858 predicting *S. aureus*

Multivariate analysis

PLS-DA was executed to assess the predictive values for the set of targeted columns. There was no significant differentiation between the two groups (again *S. aureus* positive and negative and *S. aureus* chronic and *S. aureus* not chronic) when using discriminating variates on the X and Y-axis (see figure 10). Respectively, 24% and 26% are explained by the X1-variate and 9% and 11% by the X2-variate, displayed on the Y-axis. No discriminating clusters of samples can be determined for the colonisation of *S. Aureus* (left side of figure 10) or chronic infections with *S. aureus* groups (right side of figure 10). To compare the predictive performance of the multivariate analysis the AUROC of the first two components was calculated for both models, resulting in of 0.61 and 0.65 respectively for the positive versus negative sample and the chronic versus not chronic analysis of 0.63 and 0.67.

Mapper

With the GC-MS data (27 data columns) linked to candidate VOCs for *S. aureus* a mapper network is constructed. In this network, the samples are all structured in a way that corresponding samples are either in the same vertex (dots containing samples) or in vertices positioned close to each. The relations between the vertices gives information about the coherence of these samples. The direction or shape of the structure is not of additional value. The occurrence of multiple splits is a result of the coherence of the vertices. It is important to note that one sample can be included in multiple vertices if the dimensions are similar in multiple ways. The layer of interest of this network can be visualised to understand how the different columns of the data are represented in the network, this can be varied to look at multiple aspects.

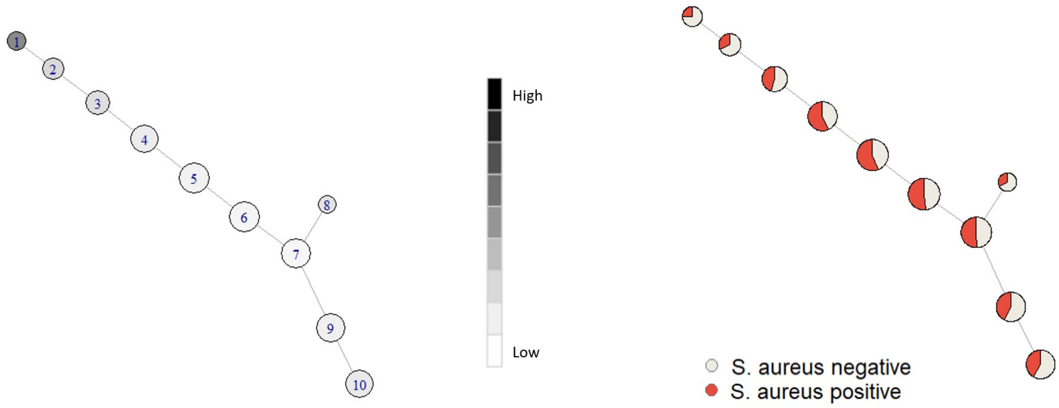


(a) Positive samples versus negative samples

(b) Chronically versus not-chronically infected samples

Figure 10: PLS DA executed with all targeted GC-MS data columns focusing on *S. aureus*

Figure 15 includes the network in two different layers in two different plots, where mean values of column X85.858 (the most significant differing GC-MS data column for *S. aureus*) is used as a colour scale for the vertices on the left and the amount of *S. aureus* in the vertices is displayed on the right. In the right plot, a more considerable amount of *S. aureus* is observed in the middle of the string. The value of column X85.858 seems to decrease continuously over the string's length from vertex 1 until 10. One could say that the values of both networks are inversely related until vertex 7, where the amount of *S. aureus* and the mean value of the vertices 9 and 10 both decreases.

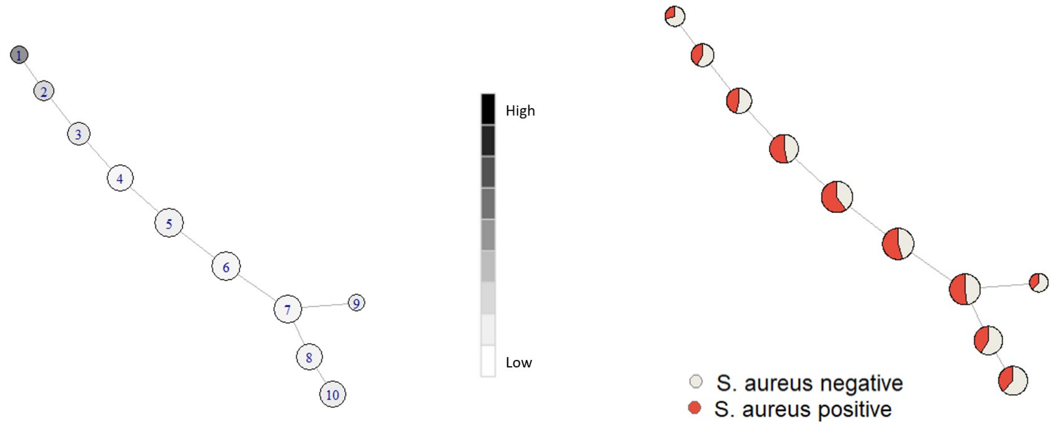


(a) Coloured by GC-MS data column X85.858

(b) Coloured by the ratio of *S. aureus* in samples

Figure 11: (TDA) Mapper plot used to visualise the value of GC-MS data column X85.858 versus the ratio of *S. aureus* in samples. 10 vertices are formed based on how the samples correlate, laying in a string with only one extra split on vertex 7. The mean values of column X85.858 plotted on the vertices decrease steadily from vertex 1 until 10. The highest amount of *S. aureus* is found in vertices 4, 5 and 6.

Similar to the network in figure 15 based on *S. aureus* positive versus negative samples, in figure 12 a network is plotted based on the *S. aureus* chronic versus the not chronic samples. In figure 12, the layers are coloured with the same method in the left and right network as in figure 15 and the same structures are formed, except for some minor changes caused by the different properties of the dataset between positive vs. negative and chronically infected vs. not chronically infected sputum samples.



(a) Coloured by GC-MS data column X85.858

(b) Coloured by the ratio of *S. aureus* in samples

Figure 12: (TDA) Mapper plot used to visualise the value of GC-MS data column X85.858 next to the ratio of *S. aureus* chronic samples. 10 vertices are formed based on the coherence of samples, laying in a string with only one extra split on vertex 7. The mean values of column X85.858 plotted on the vertices decrease steadily from vertex 1 until 10. The highest amount of S. aureus is found in vertices 4, 5 and 6.

3.2.3 Detection of *A. fumigatus*

Univariate analysis

The resulting scores of the Mann-Whitney-U test for positive against negative samples are given in table 6 of all columns matched to α-pinene a p-value of under 0.05 is considered to display a significant difference between the two groups – positive versus negative sputum samples. In column 'X53.817', a significant difference is present with a p-value of < 0.05. No other column showed a significant difference between groups.

Table 6: Mann-Whitney U test results targeted GC-MS columns for *A. fumigatus*

Compound [1]	**Columns [2]**	**P − values [3]**
α-pinene	X51.817	0.40
	X52.817	0.22
	X53.817	<0.05
	X63.817	0.14
	X65.817	0.46
	X77.817	0.26
	X78.817	0.16
	X79.817	0.31
	X80.817	0.28
	X91.817	0.31
	X92.817	0.31
	X94.817	0.24
	X103.817	0.46
	X105.817	0.65
	X106.817	0.30
	X107.817	0.40
	X119.817	0.28
	X121.817	0.36
	X136.817	0.33
	X137.817	0.30

Results are displayed with the most-used chemical name of the compound [1], the data column which gives information about the m/z ratio and retention time [2] and the p-values for statistical test between the positive versus negative samples[3] . Breath samples are considered positive when the result of sputum sampling was positive for *A. fumigatus* and negative when its result was negative.

To visualise the predictive value of this column '53.817', the ROC curve is constructed with an AUC of 0.62, see figure 13.

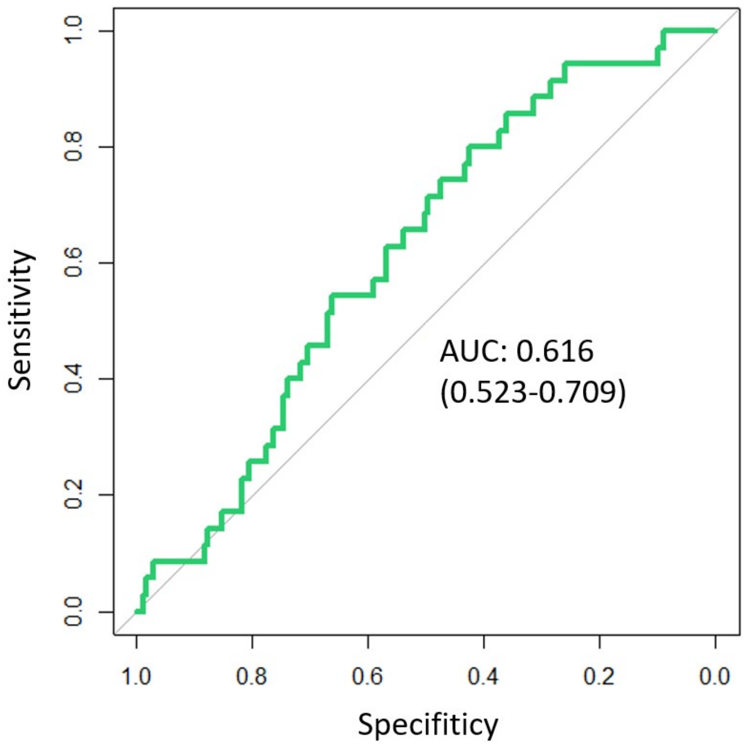


Figure 13: ROC-curve for GC-MS data column X53.817 predicting *A. fumigatus*

Multivariate analysis

As for *S. aureus* the PLS-DA was also executed for all GC-MS data columns linked to α-pinene and in this manner indirectly to *A. fumigatus*. The two groups differ on the most discriminating variates on the X and Y-axis, 85% explained by the X1-variate and 2% by the X2-variate on the Y-axis, see figure 14. No explicit groups or clusters can be distinguished by plotting over the (most discriminating) X-variate and Y-variate. To check the predictive value of the multivariate analysis the AUROC of the first two components was calculated, resulting in of 0.57 and 0.63 respectively.

Mapper

With all 21 GC-MS data columns targeted on *A. fumigatus* a mapper network is constructed to check the coherence between linked columns and *A. fumigatus*. Just like *S. aureus* detection, the layers of interest of this network can be varied. In figure 15, the same network is plotted twice, where the values of column X53.817 is used as a colour scale on the left and the amount of *A. fumigatus* is displayed on the right. In the right plot, a more considerable amount of *A. fumigatus* is observed in the left part and median part of the string, while the right particle is displaying no *A. fumigatus* in the vertices. The values of column X53.817 seem to be have the same pattern for the 'clean' samples in the right lower part.

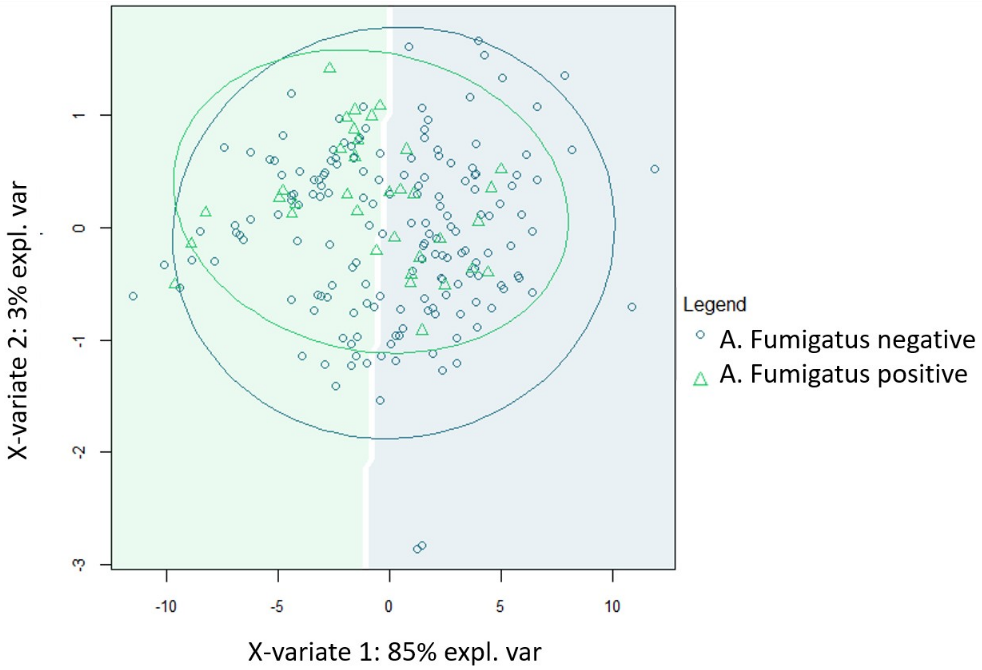
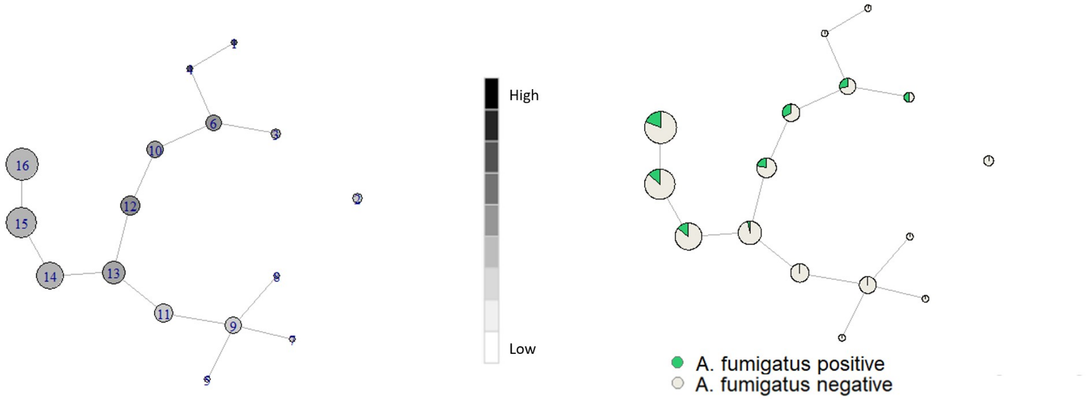


Figure 14: PLS DA executed with all targeted columns for positive versus negative samples



(a) Coloured by GC-MS data column X53.817

(b) Coloured by the ratio of *A. fumigatus* in samples

Figure 15: (TDA) Mapper used to visualise the value of GC-MS data column X53.817 (left) versus the ratio of *A. fumigatus* in samples (right). 16 vertices are formed based the GC-MS information, forming a network with extra splits on vertex 6, 13 and 9. The mean values of column X53.817 is the highest in the upper part of the network, becomes lower in the middle part and has the lowest values in the lower part. The amount of *A. fumigatus* plotted on the vertices are similarly the highest in the upper and left section and the lowest in the lower section.

3.2.4 Co-infection analysis using mapper

With the mapper algorithm, a network is constructed with all GC-MS data linked to *S. aureus* and *A. fumigatus* and *P. aeruginosa.* The GC-MS data columns linked to *S. aureus* and *A. fumigatus* resulted from the targeted analysis of this study. The columns linked to compounds to look for *P. aeruginosa* are chosen accordingly to the results of Kos et al[29]. In figure 16 the constructed network is visualised. Based on the GC-MS data a distinction is made between the breath samples included in the up left vertices, which contain mainly *S. aureus, P. aeruginosa,* and 'clean' samples in specific amounts for each vertex, see figure 16. The rest of the string also contains mixed samples.

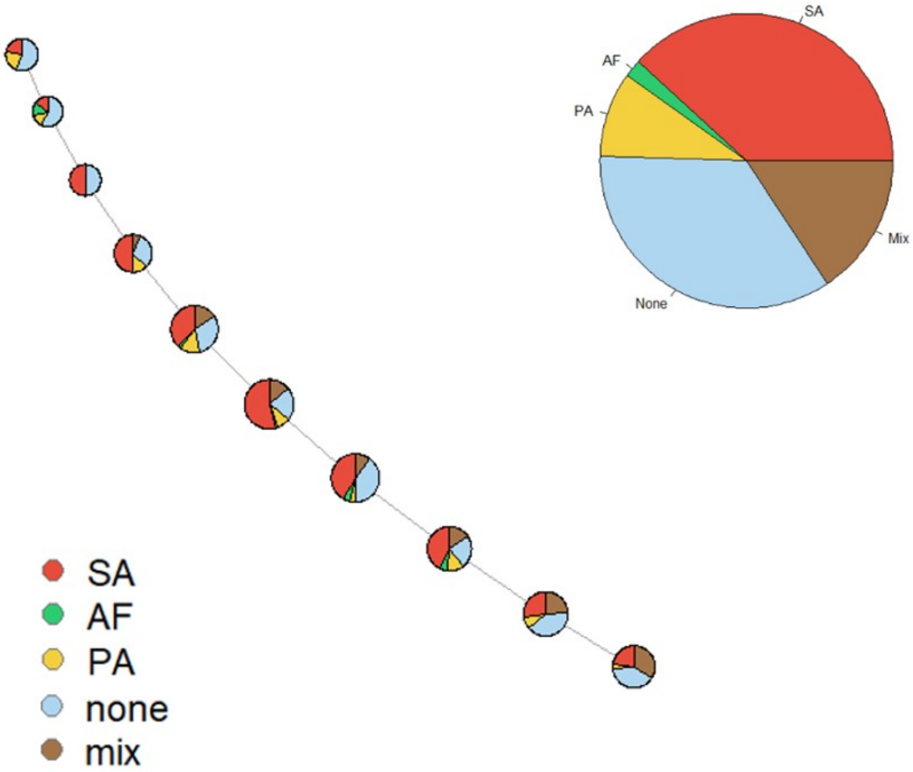


Figure 16: Mapper used to visualise the distribution of all samples included in the vertices with total population distribution shown in a pie chart upright. All samples either only positive for *S. aureus* (SA), *A. fumigatus* (AF), *P. aeruginosa* (PA), none of these three pathogens or multiple pathogens (mix). The amount of samples in the vertices range from 7 to 56.

To get insights on specific pathogens and their role in the total network, we focus on their linked columns - i.e. X85.858 for *S. aureus*, X53.817 for *A. fumigatus* and X73.285 for *P. aeruginosa* - and the coherence of these plots to the presence of the pathogen. The potential correlation of a compound to the pathogen's presence can be confirmed qualitatively by matching patterns in both plots.

Comparing amount of *S. aureus* coupled to the mean value of X85.858 (see figure 17) shows a clear outlier in the mean value of X85.858 in node 2. Next to the higher value in node 2 also the nodes 1, 3, and 4 display higher values just like a small difference on the other end in node 9 and 10. In the left plot it is visible that in the middle of the string the ratio of S. aureus is the highest, becoming smaller towards both ends.

In figure 18, the same data is showed for the distribution of *A. fumigatus* on the left and the values of column X53.817 on the right. There is a outlier visible in the mean value of X53.817 in node 2, being higher than all the other nodes and thus coloured darker. Other patterns are not visible.

In figure 19 the same plots are constructed for the distribution of *P. aeruginosa* on the left and the values of column

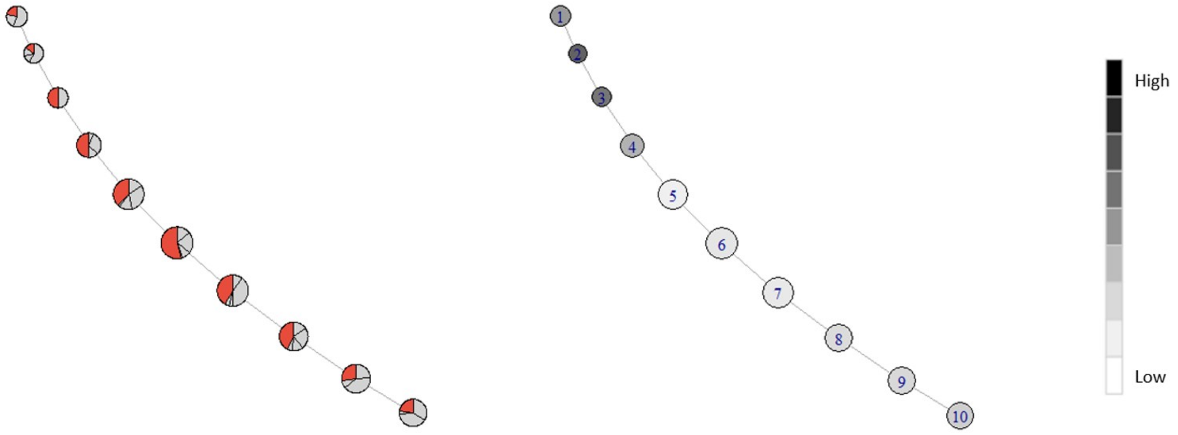


Figure 17: Mapper network to visualise to the distribution of *S. aureus* in the vertices (left) in comparison the pattern of GC-MS data column X85.858 (right). The piecharts give information about the ratio of *S. aureus* (red) and the clean samples. The high (black) and low amounts (white) in the right network are the values of the data column.

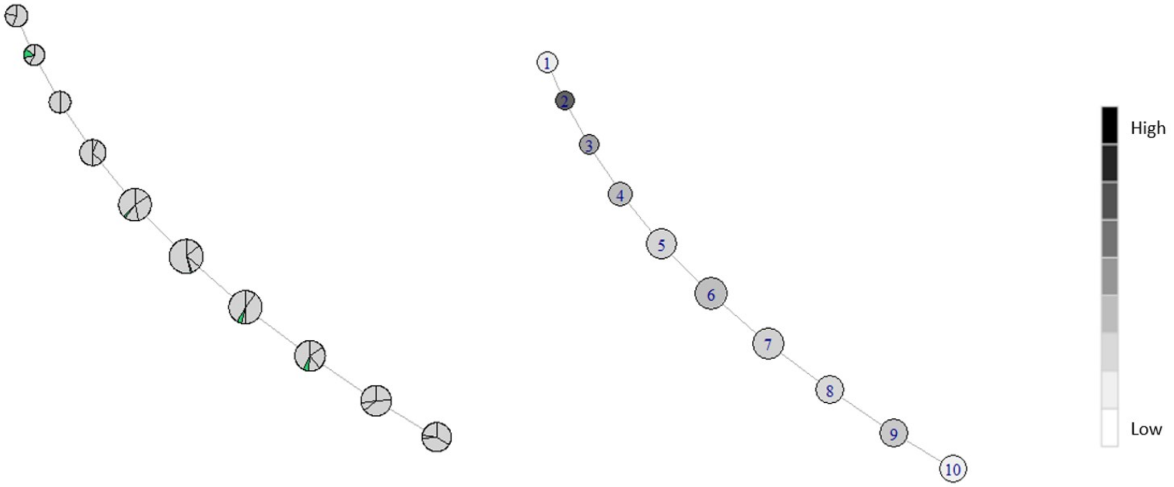


Figure 18: Mapper network to visualise the pattern of GC-MS data column X53.817 (right) in comparison to the distribution of *A. fumigatus* in the vertices (left).The piecharts give information about the ratio of *A. fumigatus* (green) and the clean samples. The high (black) and low amounts (white) in the right network are the values of the data column

X73.258 on the right. There are higher values visible in the mean value of X73.258 in node 1, 4, 7, and 9. In the plotted groups of *P. aeruginosa*, no real pattern can be appointed, except for a increasing amount of *P. aeruginosa* in the nodes 1, 4-6, and 9-10.

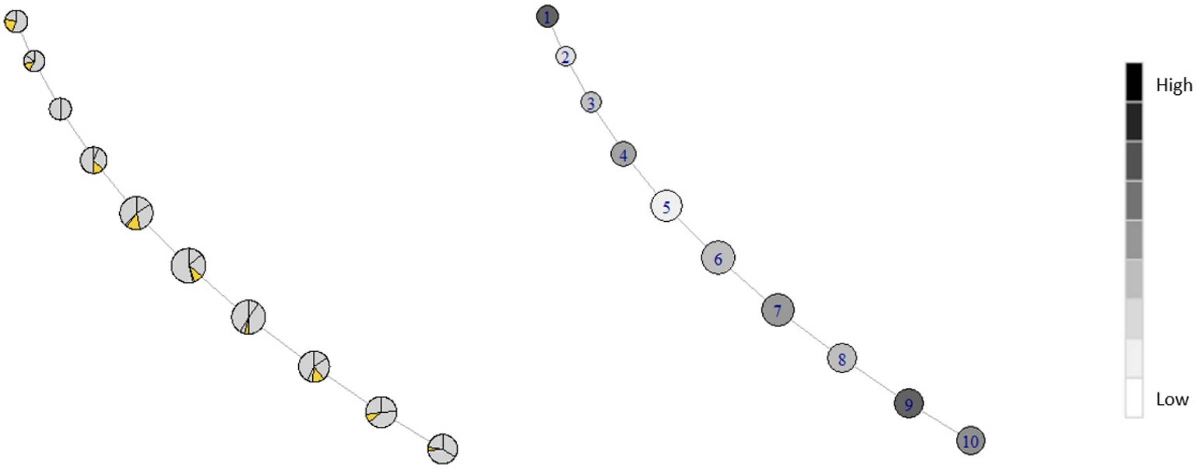


Figure 19: Mapper network to visualise the pattern of GC-MS data column X73.258 (right) in comparison to the distribution of *P. aeruginosa* in the vertices (left). The piecharts give information about the ratio of *P. aeruginosa* (yellow) and the clean samples. The high (black) and low amounts (white) in the right network are the values of the data column.

4 Discussion

This study has found several components to be associated with *S. aureus* and one with *A. fumigatus* in literature. There were some data columns (combination of mass and retention times) significantly different between colonised and not colonised samples, however not consequently pointing in the same direction as other columns linked to these same compounds. Meaning that no single component with a high predictive value in univariate and multivariate analyses was appointed in this study's breath samples, for *S. aureus* and *A. fumigatus*. This study could therefore not indicate candidate VOCs in exhaled breath to distinct *S. aureus* or *A. Fumigatus*. It did, however, show a number of VOCs linked to *S. aureus* and *A. fumigatus* published in multiple studies. Next, it did manage to show the usefulness of the mapper algorithm when using GC-MS data, by visualising properties of a population colonised with *S. aureus, A. fumigatus* or *P. aeruginosa*. These results should be followed by new studies focusing on more pathogens, gaining more knowledge about using VOCs in exhaled breath as predictors for lung infections in CF patients.

This study was the first to validate previous literature that was found in a systematic review using a targeted analysis with in vivo GC-MS data for *S. aureus* or *A. fumigatus*. The (unpublished) work of Kos. et al has performed a similar study though focusing on *P. aeruginosa*[29]. Differentiating from our study is the step of compound confirmation in the breath samples, instead of performing this check manually we used a batch job analysis in AMDIS and checked the results by hand. The difference in presence of benzaldehyde in the breath sample analysis of both groups (positive vs. negative and chronically infected vs. not chronically infected) in this study is according to the findings of four articles in our systematic review. Filipiak et al, Berrou et al., and Boots et al. published findings that benzaldehyde was consumed by *S. aureus* in their studies leading to a lower amount of benzaldehyde[33–35]. Karami et al. mentioned a higher value of benzaldehyde in cultures with *S. aureus*[36], meaning that the bacteria produced the benzaldehyde. Our results show a lower amount of the column X85.858 in the group where *S. aureus* was present, pointing in the direction of consumption by the bacteria. However, all other columns that are linked to benzaldehyde (X77.863, X78.865, X105.860, X106.863 and X107.863) do not result in significant changes. Significantly different amounts of data columns linked to α-pinene between the positive and negative breath samples, is in line with the findings in two articles in our systematic review. Ahmed et al. and Heddergott et al. both showed that α-pinene was increased when *A. fumigatus* was present in the samples[37,38].

Using topological data analysis for multidimensional datasets is not a new approach, though other studies that used the mapper algorithm focused on different datasets. Using this method researchers found predictive factors of wage[39], gene expression on profiling breast cancer [Lum] and clustered player performance in the National Basketball Association (NBA)[40]. The high dimensional datasets that are used is similar in these divergent studies. It is used as validation method for clustering methods before[41], but never as an independent approach for GC-MS data. To the best of the authors' knowledge, this study was the first to use a mapper algorithm's topological method to assess GC-MS data.

This study has several strengths. The first being its reproducibility, due to a very clear protocol and transparent steps this research is easy to repeat, making it suitable for future research to use this as reference for further steps. Secondly, is the verification of previous research in *in-vivo* clinical data. This study combines previous knowledge and tests it in clinical data, adding more context to the previous literature by underlining or disproving . In this field of interest a lot of studies are conducted to look for the new candidate VOCs that point to pathogens. These studies are all executed separately, making it look like a competition. Furthermore, the study methods with regards to the analysis are complete, using multiple angles of approach. Working with multiple statistical analyses, quantitative outcome measures and a standardised analysis protocol means that the conclusion of this study is not ambiguous. The findings can be trusted and there is a minimal chance of getting the wrong predictors with this method. Finally, the standardised method when analysing GC-MS data of the breath samples. The factor of manual work is reduced making use of a batch job option in the software of AMDIS. Hereby, directly making chance of subjective decisions or errors smaller. The fact that these batch results are checked manually makes it more standardised but at the same time still robust.

This study had several limitations. First, the use of the standardised systematic search. By using such a predefined search protocol, articles with valuable data could be missed. A balance is needed between time investment to check all articles and looking for the best search terminology. The broader the search, the more time is needed to check the articles. This systematic search's reproducibility means that everybody can rerun the search and add more information if needed. Moreover, this study entirely focused on checking and validating old findings in a new*in-vivo* dataset. A direct result is that no new findings could be pointed out by using this new data. The execution of GC-MS analysis in this research has some potential downsides. The batch job's check is done by hand and only by one researcher. This manual check of compounds is subjected to intra-observer variability. Thereby, the AMDIS software works with relative intensities of the GC-MS data. By not being able to see the absolute values of these peaks it becomes harder to check the presence of compounds. Finally, data about the colonisation of pathogens in this study is used as a categorical value, either present or absent. It would be nice to come up with a way of quantifying the pathogen in the sample, this gives more subtlety to the analysis.

When interpreting GC-MS data one should keep in mind that this technique relies on specifics in the measuring setup. Factors that could influence a GC-MS analysis results are the length and material of the column, the choice of carrier gas, or the material in which the breath samples are stored. All these factors can influence the results of the MS analysis.

In this study the information is gathered by a systematic review in the literature, not making it explicit which method could or could not be used. A wide range of different techniques is present in the compounds list, GC-time-of-flight MS (GC-TOF-MS), headspace sorptive extraction (HSSE), stir bar sorptive extraction (SBSE), and solid phase microextraction (SPME), secondary electrospray ionisation MS (SESI-MS) or selected ion flow tube MS (SIFT-MS) are all encountered with their own specifications.

Another aspect to keep in mind is that GC-MS data is caught with a breath manoeuvre on a particular moment in time. The exhaled breath mixture is often based on an end-to-end process displayed in the exhaled breath from lungs but gives no information about the steps in between. These processes can take place on another location, while the end products will be visible in the breath. It is thus hard to reason which processes are behind some of the compounds.

One should note that presence of a VOC in a pathway or process does not necessarily mean that this presence can be measured. The amounts and concentrations are sometimes of such small numbers that the sensors cannot measure them or cannot discriminate them from samples in which the VOCs are not present.

All data columns except for X85.858 are not significant, making it doubtful that benzaldehyde is a compound that plays a significant role in this column. The mass spectrum of benzaldehyde is shown in figure 21, the highest peaks – often the most specific for their compound – are 77, 105 and 106. The columns close to these peaks show not sufficient differences between *S. aureus* positive and negative samples, making it very hard to justify the claim that benzaldehyde can be appointed a predictor for *S. aureus.*

If a difference between columns linked to benzaldehyde was present in a higher amount when *S. aureus* is present, one should still be critical about this finding. It could be the case that the toluene degradation or aminobenzoate degradation play a role in these differences[43].

Given that all columns except for X53.817 are not significant gives uncertainty that α -pinene is a compound that could be used as a predictor for *A. fumigatus.* The columns with a higher relative intensity (see figure 21) should be more specific for α-pinene when present in the sample. Most of these columns are not differentiating, pointing to

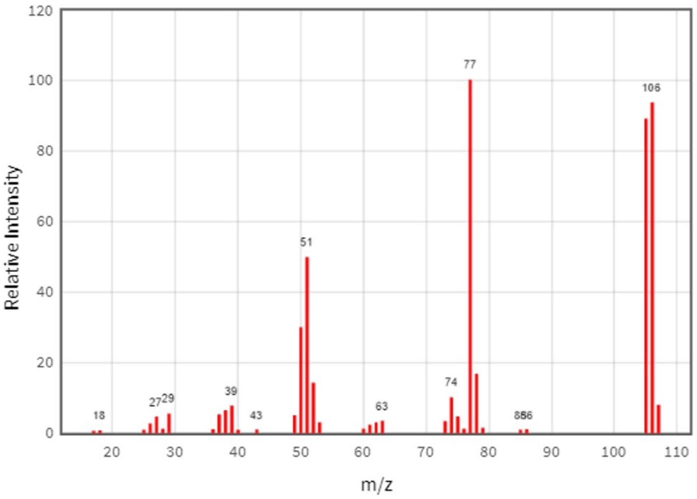


Figure 20: Mass spectrum of Benzaldehyde[42]

the conclusion that using α-pinene as a trustworthy predictor for *A. fumigatus* should be approached with a critical attitude.

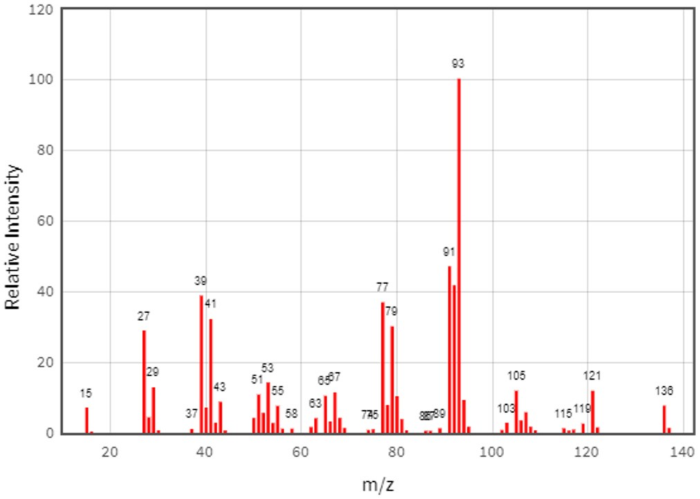


Figure 21: Mass spectrum of α-pinene[42]

If more columns corresponding to α-pinene were differentiating in the presence of *A. fumigatus*, one should still be critical about this finding, it could be due to limonene and pinene degradation or due to the biosynthesis of secondary metabolites[43].

PLS-DA is a reliable method to look for a possible distinction in a dataset between two groups, with a risk of overfitting because it is tailored to the concerning dataset. If this method does not give a usable outcome, there is a high chance that there is not enough information in the data to predict the distribution of samples. PLS-DA is a method that looks for – and displays - the highest co variance in the specific dataset and can distinguish between the groups. If the plot suggests no separation, there is most likely no separation possible with this kind of mathematical method(s). Looking at the separation in the results of the PLS-DA for both the pathogens, one can conclude that this set of columns cannot predict in which group a sample is in.

For the results of both *S. aureus* and *A. fumigatus* a relationship can be seen between the pathogen presence in different groups and the corresponding column values for linked compounds. The analysis in this form is to indicate a relationship and not quantify it, this is because it relies merely on visualising and not counting or calculating certain outcome measures. The mapper algorithm's functioning in this study did not provide a separation between the pathogen groups. There are however, parts of the algorithm that could contribute to this field of GC-MS research.

Clinical relevance

While looking into exhaled breath analysis, this study found no concrete VOC that could be pinpointed directly to detect *S. aureus, A. fumigatus* or discriminate between the pathogens. For further research focused on improving the clinical relevance it should be noted that using pathogenic information in exhaled breath can be an essential tool to predict pathogen-induced exacerbations. However, to act on it, the type of the causing pathogen needs to be known, including the reactivity to antibiotics or other treatment of that specific pathogen. The current method to reach knowledge about these characteristics is culturing the sputum. No matter how good the results of breath analysis will become, in-house sputum culturing would still be needed when personalising treatment for patients colonised by pathogens. An ideal case could be to monitor the patient at home until presence of a possible pathogen is indicated and more analysis is needed. When pathogen-specific sensors detect such a pathogen, the patient could then go to the hospital to analyse more specifically which pathogen colonised the lungs and how it reacts on different antibiotics. Patients who do not cough up sputum at all have an even higher improvement of their personalised care, being able to track their lungs' colonization in the future, where this was not possible before. This could mean a major step forward in healthcare for these patients.

5 Future recommendations

The systematic review in combination with targeted analysis executed in this study did not pinpoint a specific biomarker or VOC to predict or prove a presence of either *S. aureus* or *A. fumigatus*. However, a targeted analysis is an exquisite method to pinpoint candidate VOCs and still be able to know the reasoning behind the possible separation. A couple of recommendations can be made for further research in this field.

For example adding an earlier analysis step to the process. If one is interested in using the targeted approach, it is first useful to know if there is information in the columns of the concerning dataset. When PLS-DA is executed on all columns and there seems to be information in the columns, a targeted analysis only added value if a substantiated list of interesting compounds – and thus data columns – can be found. If beforehand it is shown that there is no information in the data, further analysis in this direction could prove to be not very useful.

The use of mapper in this study was exploratory. I would recommend the mapper tool to be used for future research as a tool to look at the data before starting a quantitative analysis. For example by colouring a network to look for overlapping patterns with regards to pathogen presence. Computing this kind of exploratory analysis is time consuming. I would instead use the mapper method to check if data is distributed with randomisation. For example, if samples differentiate or cluster in the network in terms of gender or age. In multi centre studies it is possible to check if the samples from all centres are evenly distributed or very clustered, and in the same way for longitudinal data or even to check for differences between executing researchers. It can also be used as a validation of an already suspected relationship, in this way the specific column needing validation is already known and can be used as colour map. The computing the network should be done with as much (relevant) information as possible, as long as it is numerical. The number of dimensions used to build the network only adds more information. It does not affect the results of the colouring. A specific recommendation for the colouring of categorical data is to include a weighting factor as additional layer, for example using it as transparency value of the concerning vertex. In this way one can combine the information about different groups in the vertex with the amount of variety. It is useful to know if the 80% majority belongs to one variable and the other 20% consists of 4 other variables or just one. Another way – and in my opinion the one with the most potential - to use the mapper algorithm as a qualitative predictor. If a network of labelled samples is calculated, one can add a new sample. The location or the vertex of this new sample in the already known network could potentially give information about the properties of this sample. For example, about colonisation, based solely on the GC-MS data.

Further research to candidate VOCs of pathogens can improve the care for CF patients. Looking at more pathogens in future research will only provide the clinicians with a more complete idea about the pathogens and in this way, contribute to the health of the patients.

Development of these technical solutions makes the risk of unexpected exacerbations smaller. Thereby, using this solution in the home setting could make more information available for the health professionals when using video consults. It gives more efficiency and freedom for both the patient and the doctor in the outpatient clinic if the patients' situation remains stable. Using home monitoring based on the exhaled breath can be a non-invasive, low-cost, and time-efficient way of checking this patients' vital lung values, without letting them travel to the hospital for each check-up. I think this kind of combination of digital consults and home monitoring devices will play a significant role in the bright future of personalised medicine.

6 Conclusion

This study found several components to be associated with *S. aureus* and linked one component to *A. fumigatus* according to literature. Testing these components as candidate VOCs resulted in the conclusion that it was not able discriminate positive and negative samples for both *S. aureus* and *A. fumigatus.* There were some promising significant GC-MS data columns linked to compounds. However, other columns related to the same compounds did not show the same predictive value. This study did thus not succeed in appointing VOCs to the studies' pathogens. Nonetheless, the combination of a systematic review with a targeted analysis is a suitable method to look for VOCs and could be used in further research. The use of the mapper algorithm can be a useful tool to look into GC-MS data and should be considered as an exploratory method when continuing research in the exhaled breath field.

7 References

[1] Michael Barley, Joe Mcnally, Bruce Marshall, Al Faro, Alexander Elbert, Aliza Fink, Ase Sewall, Deena Loeffler, Kristofer Petren, Thomas O'Neil, and Samar Rizvi. Annual Data Report 2016 Cystic Fibrosis Foundation Patient Registry. *Cystic Fibrosis Foundation Patient Registry*, pages 1–92, 2016.

[2] A Zolin, EF McKone, and J van Rens. ECFS Patient Registry Annual Data Report 2013. page 127, 2018.

[3] Michael Wilschanski and Peter R. Durie. Patterns of GI disease in adulthood associated with mutations in the CFTR gene. *Gut*, 56(8):1153–1163, 2007. ISSN 00175749. doi: 10.1136/gut.2004.062786.

[4] Li Li and Shawn Somerset. Digestive system dysfunction in cystic fibrosis: Challenges for nutrition therapy. *Digestive and Liver Disease*, 46(10):865–874, 2014. ISSN 18783562. doi: 10.1016/j.dld.2014.06.011.

[5] Ronald L. Gibson, Jane L. Burns, and Bonnie W. Ramsey. Pathophysiology and Management of Pulmonary Infections in Cystic Fibrosis. *American Journal of Respiratory and Critical Care Medicine*, 168(8):918–951, 2003. ISSN 1073449X. doi: 10.1164/rccm.200304-505SO.

[6] Felix Ratjen and Gerd Döring. Cystic fibrosis. *Adherence and Self-Management in Pediatric Populations*, 196: 107–132, 2003. doi: 10.1038/nrdp.2015.10.Cystic.

[7] M. D. Amaral. Novel personalized therapies for cystic fibrosis: treating the basic defect in all patients. *Journal of Internal Medicine*, 277(2):155–166, 2 2015. ISSN 09546820. doi: 10.1111/joim.12314.

[8] Carlo Castellani, Alistair J.A. Duff, Scott C. Bell, Harry G.M. Heijerman, Anne Munck, Felix Ratjen, Isabelle Sermet-Gaudelus, Kevin W. Southern, Jurg Barben, Patrick A. Flume, Pavla Hodková, Nataliya Kashirskaya, Maya N. Kirszenbaum, Sue Madge, Helen Oxley, Barry Plant, Sarah Jane Schwarzenberg, Alan R. Smyth, Giovanni Taccetti, Thomas O.F. Wagner, Susan P. Wolfe, and Pavel Drevinek. ECFS best practice guidelines: the 2018 revision. *Journal of Cystic Fibrosis*, 17(2):153–178, 2018. ISSN 18735010. doi: 10.1016/j.jcf.2018.02.006. URL https://doi.org/10.1016/j.jcf.2018.02.006.

[9] Cystic fibrosis toegevoegd aan hielprikscreening | RIVM. URL https://www.rivm.nl/nieuws/cystic-fibrosis-toegevoegd-aan-hielprikscreening.

[10] Philip M. Farrell, Terry B. White, Clement L. Ren, Sarah E. Hempstead, Frank Accurso, Nico Derichs, Michelle Howenstine, Susanna A. McColley, Michael Rock, Margaret Rosenfeld, Isabelle Sermet-Gaudelus, Kevin W. Southern, Bruce C. Marshall, and Patrick R. Sosnay. Diagnosis of Cystic Fibrosis: Consensus Guidelines from the Cystic Fibrosis Foundation. *Journal of Pediatrics*, 181:S4–S15.e1, 2017. ISSN 10976833. doi: 10.1016/j.jpeds.2016.09.064. URL http://dx.doi.org/10.1016/j.jpeds.2016.09.064.

[11] Margarida D. Amaral and Karl Kunzelmann. Molecular targeting of CFTR as a therapeutic approach to cystic fibrosis. *Trends in Pharmacological Sciences*, 28(7):334–341, 2007. ISSN 01656147. doi: 10.1016/j.tips.2007.05.004.

[12] Jr Charles A Janeway, Paul Travers, Mark Walport, and Mark J Shlomchik. Immunobiology: The Immune System in Health and Disease. chapter Infectious. Garland Science, 5th editio edition, 2001. URL https://www.ncbi.nlm.nih.gov/books/NBK27114/.

[13] Heather G Ahlgren, Andrea Benedetti, Jennifer S Landry, Joanie Bernier, Elias Matouk, Danuta Radzioch, Larry C Lands, Simon Rousseau, and Dao Nguyen. Clinical outcomes associated with Staphylococcus aureus and Pseudomonas aeruginosa airway infections in adult cystic fibrosis patients. pages 1–6, 2015. doi: 10.1186/s12890-015-0062-7.

[14] Gillian M Nixon, B Mbc, David S Armstrong, and Rosemary Carzino. C linical outcome after early Pseudomonas aeruginosa infection in cystic fibrosis. 2001. doi: 10.1067/mpd.2001.112897.

[15] Darius Seidler, Mary Griffin, Amanda Nymon, Katja Koeppen, and Alix Ashare. Throat swabs and sputum culture as predictors of P. aeruginosa or S. aureus lung colonization in adult cystic fibrosis patients. *PLoS ONE*, 11(10):8–15, 2016. ISSN 19326203. doi: 10.1371/journal.pone.0164232.

[16] Pierre-Régis Burgel, André Paugam, Dominique Hubert, and Clémence Martin. Infection and Drug Resistance Dovepress Aspergillus fumigatus in the cystic fibrosis lung: pros and cons of azole therapy. 2016. doi: 10.2147/IDR.S63621. URL http://dx.doi.org/10.2147/IDR.S63621.

[17] Gerd Döring, Patrick Flume, Harry Heijerman, and J. Stuart Elborn. Treatment of lung infection in patients with cystic fibrosis: Current and future strategies. *Journal of Cystic Fibrosis*, 11(6):461–479, 2012. ISSN 15691993. doi: 10.1016/j.jcf.2012.10.004. URL http://dx.doi.org/10.1016/j.jcf.2012.10.004.

[18] Nicholas M. Maurice, Brahmchetna Bedi, and Ruxana T. Sadikot. Pseudomonas aeruginosa biofilms: Host response and clinical implications in lung infections. *American Journal of Respiratory Cell and Molecular Biology*, 58(4):428–439, 2018. ISSN 15354989. doi: 10.1165/rcmb.2017-0321TR.

[19] Bobbi Pritt, Linda O'Brien, and Washington Winn. Mucoid Pseudomonas in cystic fibrosis. *American Journal of Clinical Pathology*, 128(1):32–34, 2007. ISSN 00029173. doi: 10.1309/KJRPC7DD5TR9NTDM.

[20] Edith T. Zemanick and Lucas R. Hoffman. Cystic Fibrosis: Microbiology and Host Response. *Pediatric Clinics of North America*, 63(4):617–636, 8 2016. ISSN 15578240. doi: 10.1016/j.pcl.2016.04.003.

[21] David Shoseyov, Keith G. Brownlee, Steven P. Conway, and Eitan Kerem. Aspergillus bronchitis in cystic fibrosis. *Chest*, 130(1):222–226, 7 2006. ISSN 00123692. doi: 10.1378/chest.130.1.222. URL https://linkinghub.elsevier.com/retrieve/pii/S0012369215509768.

[22] Agnes W. Boots, Lieuwe D. Bos, Marc P. van der Schee, Frederik Jan van Schooten, and Peter J. Sterk. Exhaled Molecular Fingerprinting in Diagnosis and Monitoring: Validating Volatile Promises, 10 2015. ISSN 1471499X. URL https://pubmed.ncbi.nlm.nih.gov/26432020/.

[23] Marc Philippe Van Der Schee, Tamara Paff, Paul Brinkman, Willem Marinus Christiaan Van Aalderen, Eric Gerardus Haarman, and Peter Jan Sterk. Breathomics in lung disease. *Chest*, 147(1):224–231, 2015. ISSN 19313543. doi: 10.1378/chest.14-0781. URL http://dx.doi.org/10.1378/chest.14-0781.

[24] Diclehan Karakaya, Oguzhan Ulucan, and Mehmet Turkan. Electronic Nose and Its Applications: A Survey, 4 2020. ISSN 17518520. URL https://link.springer.com/article/10.1007/s11633-019-1212-9.

[25] Mariana Valente Farraia, João Cavaleiro Rufo, Inês Paciência, Francisca Mendes, Luís Delgado, and André Moreira. The electronic nose technology in clinical diagnosis. *Porto Biomedical Journal*, 4(4):e42, 7 2019. ISSN 2444-8664. doi: 10.1097/j.pbj.0000000000000042. URL /pmc/articles/PMC6924976/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC6924976/.

[26] David J. Beale, Farhana R. Pinu, Konstantinos A. Kouremenos, Mahesha M. Poojary, Vinod K. Narayana, Berin A. Boughton, Komal Kanojia, Saravanan Dayalan, Oliver A.H. Jones, and Daniel A. Dias. Review of recent developments in GC–MS approaches to metabolomics-based research, 11 2018. ISSN 15733890. URL https://link.springer.com/article/10.1007/s11306-018-1449-2.

[27] O. David Sparkman, Zelda Penton, and Fulton G. Kitson. *Gas Chromatography and Mass Spectrometry: A Practical Guide*. Elsevier Inc., 2011. ISBN 9780123736284. doi: 10.1016/C2009-0-17039-3.

[28] Frank Röck, Nicolae Barsan, and Udo Weimar. Electronic nose: Current status and future trends. *Chemical Reviews*, 108(2):705–725, 2008. ISSN 00092665. doi: 10.1021/cr068121q.

[29] R. Kos, P. Brinkman, A.H. Neerinx, T. Paff, M.G. Gerritsen, A. Lammers, A.D. Kraneveld, H.G.M. Heijerman, H.M. Janssens, J.C. Davies, C.J. Majoor, E.J. Weersink, P.J. Sterk, E.G. Haarman, L.D. Bos, and A.H Maitland-van der zee. Targeted analysis of volatile organic compounds for detection of Pseudomonas aeruginosa in cystic fibrosis patients by exhaled breath analysis. *Submitted (not published)*, 2020.

[30] Elizabeth Munch. A User's Guide to Topological Data Analysis. *Journal of Learning Analytics*, 4(2), 7 2017. ISSN 1929-7750. doi: 10.18608/jla.2017.42.6. URL `https://learning-analytics.info/index.php/JLA/article/view/5196`.

[31] Gurjeet Singh, Facundo Mémoli, and Gunnar Carlsson. Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition. Technical report, 2007.

[32] Müllner and Babu. Welcome to the Python Mapper documentation! — Python Mapper documentation. URL `http://danifold.net/mapper/`.

[33] Wojciech Filipiak, Andreas Sponring, Maria Magdalena Baur, Anna Filipiak, Clemens Ager, Helmut Wiesenhofer, Markus Nagl, Jakob Troppmair, and Anton Amann. Molecular analysis of volatile metabolites released specifically by staphylococcus aureus and pseudomonas aeruginosa. *BMC Microbiology*, 12(1):1, 2012. ISSN 14712180. doi: 10.1186/1471-2180-12-113. URL `BMCMicrobiology`.

[34] Kevin Berrou, Catherine Dunyach-Remy, Jean Philippe Lavigne, Benoit Roig, and Axelle Cadiere. Comparison of stir bar sorptive extraction and solid phase microextraction of volatile and semi-volatile metabolite profile of staphylococcus aureus. *Molecules*, 25(1), 2020. ISSN 14203049. doi: 10.3390/molecules25010055.

[35] A W Boots, A Smolinska, J J B N van Berkel, R R R Fijten, E E Stobberingh, M L L Boumans, E J Moonen, E F M Wouters, J W Dallinga, and F J Van Schooten. Identification of microorganisms based on headspace analysis of volatile organic compounds by gas chromatography–mass spectrometry. *Journal of Breath Research*, 8 (2):027106, 4 2014. ISSN 1752-7155. doi: 10.1088/1752-7155/8/2/027106.

[36] Najmeh Karami, Hassan Rezadoost, Fateme Mirzajani, Abdollah Karimi, Alireza Ghassempour, Atousa Aliahmadi, and Fatemeh Fallah. Resistant/susceptible classification of respiratory tract pathogenic bacteria based on volatile organic compounds profiling. *Cellular and Molecular Biology*, 2018. ISSN 01455680.

[37] Waqar M. Ahmed, Pavlos Geranios, Iain R. White, Oluwasola Lawal, Tamara M. Nijsen, Michael J. Bromley, Royston Goodacre, Nick D. Read, and Stephen J. Fowler. Development of an adaptable headspace sampling method for metabolic profiling of the fungal volatome. *Analyst*, 2018. ISSN 13645528. doi: 10.1039/c8an00841h.

[38] C. Heddergott, J. P. Latgé, and A. M. Calvo. The volatome of Aspergillus fumigatus. *Eukaryotic Cell*, 13(8): 1014–1025, 2014. ISSN 15359778. doi: 10.1128/EC.00074-14.

[39] Rami Kraft. Illustrations of data analysis using the mapper algorithm and persistant homology. 2016. URL `www.kth.se/sci%0Ahttp://www.diva-portal.org/smash/record.jsf?pid=diva2%3A900997&dswid=-1942`.

[40] P. Y. Lum, G. Singh, A. Lehman, T. Ishkanov, M. Vejdemo-Johansson, M. Alagappan, J. Carlsson, and G. Carlsson. Extracting insights from the shape of complex data using topology. *Scientific Reports*, 3(February), 2013. ISSN 20452322. doi: 10.1038/srep01236.

[41] Paul Brinkman, Ariane H. Wagener, Pieter Paul Hekking, Aruna T. Bansal, Anke Hilse Maitland-van der Zee, Yuanyue Wang, Hans Weda, Hugo H. Knobel, Teunis J. Vink, Nicholas J. Rattray, Arnaldo D'Amico, Giorgio Pennazza, Marco Santonico, Diane Lefaudeux, Bertrand De Meulder, Charles Auffray, Per S. Bakke, Massimo Caruso, Pascal Chanez, Kian F. Chung, Julie Corfield, Sven Erik Dahlén, Ratko Djukanovic, Thomas Geiser, Ildiko Horvath, Nobert Krug, Jacek Musial, Kai Sun, John H. Riley, Dominic E. Shaw, Thomas Sandström, Ana R. Sousa, Paolo Montuschi, Stephen J. Fowler, and Peter J. Sterk. Identification and prospective stability of electronic nose (eNose)–derived inflammatory phenotypes in patients with severe asthma. *Journal of Allergy and Clinical Immunology*, 143(5):1811–1820, 2019. ISSN 10976825. doi: 10.1016/j.jaci.2018.10.058.

[42] Stephen Stein. NIST MS Search, 2008. URL https://chemdata.nist.gov/mass-spc/ms-search/.

[43] KEGG PATHWAY Database. URL https://www.genome.jp/kegg/pathway.html.

[44] J. Chen, J. N. Tang, K. L. Hu, Y. Y. Zhao, and C. Tang. The production characteristics of volatile organic compounds and their relation to growth status of Staphylococcus aureus in milk environment. *Journal of Dairy Science*, 101(6):4983–4991, 2018. ISSN 15253198. doi: 10.3168/jds.2017-13629.

[45] Juan Chen, Junni Tang, Hui Shi, Cheng Tang, and Rong Zhang. Characteristics of volatile organic compounds produced from five pathogenic bacteria by headspace-solid phase micro-extraction/gas chromatography-mass spectrometry. *Journal of Basic Microbiology*, 57(3):228–237, 3 2017. ISSN 15214028. doi: 10.1002/jobm.201600505. URL https://pubmed.ncbi.nlm.nih.gov/27874211/.

[46] Kevin Berrou, Catherine Dunyach-Remy, Jean Philippe Lavigne, Benoit Roig, and Axelle Cadiere. Multiple stir bar sorptive extraction combined with gas chromatography-mass spectrometry analysis for a tentative identification of bacterial volatile and/or semi-volatile metabolites. *Talanta*, 195(July 2018):245–250, 2019. ISSN 00399140. doi: 10.1016/j.talanta.2018.11.042.

[47] Carrie L. Jenkins and Heather D. Bean. Influence of media on the differentiation of Staphylococcus spp. By volatile compounds. *Journal of Breath Research*, 14(1):16007, 2020. ISSN 17527163. doi: 10.1088/1752-7163/ab3e9d. URL http://dx.doi.org/10.1088/1752-7163/ab3e9d.

[48] Mavra Nasir, Heather D. Bean, Agnieszka Smolinska, Christiaan A. Rees, Edith T. Zemanick, and Jane E. Hill. Volatile molecules from bronchoalveolar lavage fluid can 'rule-in' Pseudomonas aeruginosa and 'rule-out' Staphylococcus aureus infections in cystic fibrosis patients. *Scientific Reports*, 8(1):1–11, 2018. ISSN 20452322. doi: 10.1038/s41598-017-18491-8. URL http://dx.doi.org/10.1038/s41598-017-18491-8.

[49] Mohammed Ashrafi, Lilyann Novak-Frazer, Matthew Bates, Mohamed Baguneid, Teresa Alonso-Rasgado, Guoqing Xia, Riina Rautemaa-Richardson, and Ardeshir Bayat. Validation of biofilm formation on human skin wound models and demonstration of clinically translatable bacteria-specific volatile signatures. *Scientific Reports*, 8(1): 1–16, 2018. ISSN 20452322. doi: 10.1038/s41598-018-27504-z.

[50] Yu Wang, Sijing Liu, Qikang Pu, Yongxin Li, Xixi Wang, Yang Jiang, Danni Yang, Yi Yang, Jinling Yang, and Chengjun Sun. Rapid identification of Staphylococcus aureus, Vibrio parahaemolyticus and Shigella sonnei in foods by solid phase microextraction coupled with gas chromatography–mass spectrometry. *Food Chemistry*, 262 (April 2017):7–13, 2018. ISSN 18737072. doi: 10.1016/j.foodchem.2018.04.088.

[51] Flavio A. Franchina, Giorgia Purcaro, Alison Burklund, Marco Beccaria, and Jane E. Hill. Evaluation of different adsorbent materials for the untargeted and targeted bacterial VOC analysis using GC×GC-MS. *Analytica Chimica Acta*, 1066:146–153, 2019. ISSN 18734324. doi: 10.1016/j.aca.2019.03.027.

[52] Haorong Li and Jiangjiang Zhu. Differentiating Antibiotic-Resistant Staphylococcus aureus Using Secondary Electrospray Ionization Tandem Mass Spectrometry. *Analytical Chemistry*, 90(20):12108–12115, 2018. ISSN 15206882. doi: 10.1021/acs.analchem.8b03029.

[53] Mohammed Ashrafi, Lilyann Novak-Frazer, Julie Morris, Mohamed Baguneid, Riina Rautemaa-Richardson, and Ardeshir Bayat. Electrical stimulation disrupts biofilms in a human wound model and reveals the potential for monitoring treatment response with volatile biomarkers. *Wound Repair and Regeneration*, 27(1):5–18, 2019. ISSN 1524475X. doi: 10.1111/wrr.12679.

[54] A. H. Neerincx, B. P. Geurts, J. Van Loon, V. Tiemes, J. J. Jansen, F. J.M. Harren, L. A.J. Kluijtmans, P. J.F.M. Merkus, S. M. Cristescu, L. M.C. Buydens, and R. A. Wevers. Detection of Staphylococcus aureus in cystic fibrosis patients using breath VOC profiles. *Journal of Breath Research*, 10(4):46014, 2016. ISSN 17527163. doi: 10.1088/1752-7155/10/4/046014. URL http://dx.doi.org/10.1088/1752-7155/10/4/046014.

[55] Carolin Drees, Wolfgang Vautz, Sascha Liedtke, Christopher Rosin, Kirsten Althoff, Martin Lippmann, Stefan Zimmermann, Tobias J. Legler, Duygu Yildiz, Thorsten Perl, and Nils Kunze-Szikszay. GC-IMS headspace analyses allow early recognition of bacterial growth and rapid pathogen differentiation in standard blood cultures. *Applied Microbiology and Biotechnology*, 103(21-22):9091–9101, 2019. ISSN 14320614. doi: 10.1007/s00253-019-10181-x.

[56] Heather D Bean, Jaime Jimézes-Diaz, Jiangjiang Zhu, and Jane E. Hill. Breathprints of model murine bacterial lung infections are linked with immune response. *Physiology & behavior*, 176(1), 2016. doi: 10.1117/12.2549369. Hyperspectral.

[57] Kseniya Dryahina, Kristýna Sovová, Alexandr Nemec, and Patrik Španěl. Differentiation of pulmonary bacterial pathogens in cystic fibrosis by volatile metabolites emitted by their in vitro cultures: Pseudomonas aeruginosa, Staphylococcus aureus, Stenotrophomonas maltophilia and the Burkholderia cepacia complex. *Journal of Breath Research*, 10(3), 2016. ISSN 17527163. doi: 10.1088/1752-7155/10/3/037102.

[58] Jiangjiang Zhu, Heather D. Bean, Matthew J. Wargo, Laurie W. Leclair, and Jane E. Hill. Detecting bacterial lung infections: In vivo evaluation of in vitro volatile fingerprints. *Journal of Breath Research*, 7(1), 2013. doi: 10.1088/1752-7155/7/1/016003.Detecting.

[59] Malina K. Storer, Kim Hibbard-Melles, Brett Davis, and Jenny Scotter. Detection of volatile compounds produced by microbial growth in urine by selected ion flow tube mass spectrometry (SIFT-MS). *Journal of Microbiological Methods*, 87(1):111–113, 10 2011. ISSN 01677012. doi: 10.1016/j.mimet.2011.06.012.

[60] Chihiro Osaki, Kyoshiro Yamaguchi, Shinji Urakawa, Yukihiko Nakashima, Kazutoshi Sugita, Masaki Nagaishi, Shinji Mitsuiki, Takuya Kuraoka, Yukiko Ogawa, and Hiroshi Sato. The bacteriological properties of bacillus strain TM-I-3 and analysis of the volatile antifungal compounds emitted by this bacteria. *Biocontrol Science*, 24 (3):129–136, 2019. ISSN 18840205. doi: 10.4265/bio.24.129.

[61] Layla J. Barkal, Clare L. Procknow, Yasmín R. Álvarez-Garciá, Mengyao Niu, José A. Jiménez-Torres, Rebecca A. Brockman-Schneider, James E. Gern, Loren C. Denlinger, Ashleigh B. Theberge, Nancy P. Keller, Erwin Berthier, and David J. Beebe. Microbial volatile communication in human organotypic lung models. *Nature Communications*, 8(1), 2017. ISSN 20411723. doi: 10.1038/s41467-017-01985-4. URL http://dx.doi.org/10.1038/s41467-017-01985-4.

[62] Christiaan Rees, Pierre-Hugues Stefanuto, and et al Beattie, Sarah. Sniffing out the hypoxia volatile metabolic signature of Aspergillus fumigatus. *Physiology & behavior*, 176(1):139–148, 2018. doi: 10.1016/j.physbeh.2017.03.040.

[63] Benoit Briard, Christoph Heddergott, and Jean Paul Latgé. Volatile compounds emitted by pseudomonas aeruginosa stimulate growth of the fungal pathogen aspergillus fumigatus. *mBio*, 7(2), 3 2016. ISSN 21507511. doi: 10.1128/mBio.00219-16. URL http://mbio.asm.org/.

[64] A. H. Neerincx, B. P. Geurts, M. F.J. Habets, J. A. Booij, J. Van Loon, J. J. Jansen, L. M.C. Buydens, J. Van Ingen, J. W. Mouton, F. J.M. Harren, R. A. Wevers, P. J.F.M. Merkus, S. M. Cristescu, and L. A.J. Kluijtmans. Identification of Pseudomonas aeruginosa and Aspergillus fumigatus mono- and co-cultures based on volatile biomarker combinations. *Journal of Breath Research*, 2016. ISSN 17527163. doi: 10.1088/1752-7155/10/1/016002.

[65] Neus Planas Pont, Catherine A. Kendall, and Naresh Magan. Analysis of volatile fingerprints for monitoring anti-fungal efficacy against the primary and opportunistic pathogen Aspergillus fumigatus. *Mycopathologia*, 173 (2-3):93–101, 2012. ISSN 15730832. doi: 10.1007/s11046-011-9490-y.

[66] Thorsten Perl, Melanie Jünger, Wolfgang Vautz, Jürgen Nolte, Martin Kuhns, Margarete Borg-von Zepelin, and Michael Quintel. Detection of characteristic metabolites of Aspergillus fumigatus and Candida species using ion mobility spectrometry - metabolic profiling by volatile organic compounds. *Mycoses*, 54(6):828–837, 2011. ISSN 09337407. doi: 10.1111/j.1439-0507.2011.02037.x.

[67] Stephen T. Chambers, Shrawan Bhandari, Amy Scott-Thomas, and Mona Syhre. Novel diagnostics: Progress toward a breath test for invasive Aspergillus fumigatus. *Medical Mycology*, 49(SUPPL. 1):54–61, 2011. ISSN 13693786. doi: 10.3109/13693786.2010.508187.

A Articles systematic review *S. aureus*

Title	**citation**
The production characteristics of volatile organic compounds and their relation to growth status of Staphylococcus aureus in milk environment	[44]
Characteristics of volatile organic compounds produced from five pathogenic bacteria by headspace-solid phase micro-extraction/gas chromatography-mass spectrometry	[45]
Molecular analysis of volatile metabolites released specifically by Staphylococcus aureus and Pseudomonas aeruginosa	[33]
Comparison of Stir Bar Sorptive Extraction and Solid Phase Microextraction of Volatile and Semi-Volatile Metabolite Profile of Staphylococcus Aureus	[46]
Influence of media on the differentiation of Staphylococcus spp. by volatile compounds	[47]
Volatile molecules from bronchoalveolar lavage fluid can 'rule-in' Pseudomonas aeruginosa and 'rule-out' Staphylococcus aureus infections in cystic fibrosis patients	[48]
Validation of biofilm formation on human skin wound models and demonstration of clinically translatable bacteria-specific volatile signatures	[49]
Resistant/susceptible classification of respiratory tract pathogenic bacteria based on volatile organic compounds profiling	[36]
Rapid identification of Staphylococcus aureus, Vibrio parahaemolyticus and Shigella sonnei in foods by solid phase microextraction coupled with gas chromatography-mass spectrometry	[50]
Evaluation of different adsorbent materials for the untargeted and targeted bacterial VOC analysis using GCxGC-MS	[51]
Differentiating Antibiotic-Resistant Staphylococcus aureus Using Secondary Electrospray Ionization Tandem Mass Spectrometry	[52]
Multiple stir bar sorptive extraction combined with gas chromatography-mass spectrometry analysis for a tentative identification of bacterial volatile and/or semi-volatile metabolites	[46]
Identification of microorganisms based on headspace analysis of volatile organic compounds by gas chromatography-mass spectrometry	[35]
Electrical stimulation disrupts biofilms in a human wound model and reveals the potential for monitoring treatment response with volatile biomarkers	[53]
Detection of Staphylococcus aureus in cystic fibrosis patients using breath VOC profiles	[54]
GC-IMS headspace analyses allow early recognition of bacterial growth and rapid pathogen differentiation in standard blood cultures	[55]
Breathprints of model murine bacterial lung infections are linked with immune response	[56]
Differentiation of pulmonary bacterial pathogens in cystic fibrosis by volatile metabolites emitted by their in vitro cultures: Pseudomonas aeruginosa, Staphylococcus aureus, Stenotrophomonas maltophilia and the Burkholderia cepacia complex	[57]
Detecting bacterial lung infections: in vivo evaluation of in vitro volatile fingerprints	[58]
Detection of volatile compounds produced by microbial growth in urine by selected ion flow tube mass spectrometry (SIFT-MS)	[59]

Articles marked in yellow turned out not useful for literature review and were excluded, see figure 7 on page 12.

B Articles systematic review *A. fumigatus*

Title	citation
The Bacteriological Properties of Bacillus Strain TM-I-3 and Analysis of the Volatile Antifungal Compounds Emitted by this Bacteria.	[60]
Development of an adaptable headspace sampling method for metabolic profiling of the fungal volatome.	[37]
Microbial volatile communication in human organotypic lung models.	[61]
Sniffing out the hypoxia volatile metabolic signature of Aspergillus fumigatus.	[62]
Volatile Compounds Emitted by Pseudomonas aeruginosa Stimulate Growth of the Fungal Pathogen Aspergillus fumigatus.	[63]
Identification of Pseudomonas aeruginosa and Aspergillus fumigatus mono- and co-cultures based on volatile biomarker combinations.	[64]
The volatome of Aspergillus fumigatus.	[38]
Analysis of volatile fingerprints for monitoring antifungal efficacy against the primary and opportunistic pathogen Aspergillus fumigatus.	[65]
Detection of characteristic metabolites of Aspergillus fumigatus and Candida species using ion mobility spectrometry-metabolic profiling by volatile organic compounds.	[66]
Novel diagnostics: progress toward a breath test for invasive Aspergillus fumigatus.	[67]

Articles marked in yellow turned out not useful for literature review and were excluded, see figure 8 on page 14.