There's something in the air : Volatile organic compounds in exhaled breath in pulmonary diseases

There's something in the air

Volatile organic compounds in exhaled breath in pulmonary diseases

Joep JBN van Berkel

ISBN: 978-94-6159-011-4
Production: Datawyse, Universitaire Pers Maastricht

The studies presented in this thesis were performed within NUTRIM School for Nutrition, Toxicology and Metabolism which participates in the graduate School VLAG (Food Technology, Agrobiotechnology, Nutrition and Health Sciences), accredited by the Royal Netherlands Academy of Arts and Sciences.

There's something in the air

Volatile organic compounds in exhaled breath in pulmonary diseases

Proefschrift

Ter verkrijging van de graad van doctor aan de Universiteit Maastricht op gezag van van de Rector Magnificus, Prof mr. G.P.M.F. Mols volgens het besluit van het College van Decanen, in het openbaar te verdedigen op 3 november 2010 om 14:00 uur

door

Joep Jozeph Benjamin Nathan van Berkel
geboren te Rijen op 20 september 1979

UNIVERSITAIRE PERS MAASTRICHT

Promotores:
Prof. dr. F.J. van Schooten
Prof. dr. E.F.M. Wouters

Copromotor:
Dr. J.W. Dallinga

Beoordelingscommissie:
Prof. dr. A. Masclee (Voorzitter)
Prof. dr. A. Bast
Prof. dr. H. Bisgaard (Copenhagen University Hospital, Denmark)
Prof. dr. J.C. de Jongste (Erasmus Universiteit Rotterdam)
Dr. G. Rohde

Contents

Nomenclature

ABPA	Allergic bronchopulmonary aspergillosis
AUC	Area under curve
BAL	Bronchoalveolar lavage
CF	Cystic fibrosis
COPD	Chronic obstructive pulmonary disease
FeNO	Fractional exhaled nitric oxide
FEV_1	Forced expiratory volume in the first second
FID	Flame ionization detector
GC	Gas chromatograph
GC-TOF-MS	Gas chormatograph time-of-flight mass spectrometer
IL-10	Interleukin-10
IL-6	Interleukin-6
MF	Match factor
MRSA	Methicillin-resistant *Staphylococcus aureus*
MS	Mass spectrometer
MSSA	Methicillin-sensitive *Staphylococcus aureus*
ppb	Parts-per-billion
PTR-MS	Proton transfer-reaction mass spectroscopy

PUFA	Polyunsaturated fatty acids
RI	Rentention index
ROS	Reactive oxygen species
RT	Retention time
RV	Residual volume
SIFT-MS	Selected ion flow tube mass spectrometry
SMO	Sequential minimal optimization
SPME	Solid phase micro extraction
SVM	Support vector machine
TNF-α	Tumor necrosis factor alpha
VC	Vital capacity
VOC	Volatile organic compound

CHAPTER 1

General introduction

ANALYSIS OF BREATH

Ancient Greek physicians were aware of the relation between the odor of a subjects' breath and possible diseases associated with it for a long time. They realized it could provide insight into physiological and pathophysiological processes in the body.[1] For example, the sweet acetonic smell of breath might indicate uncontrolled diabetes, a fishy musty reek of breath relates to liver disease and a urine-like smell is associated with kidney failure.[2] Apparently there is something in breath that might enable us to diagnose certain diseases or provide the means to monitor metabolic processes in the body. Due to the great potential of applications in clinical diagnostics and its non-invasive nature the exhaled air analysis has become of increased interest in recent years. During the last 20 to 30 years the technical advances in analytical analysis have been responsible for the recent developmental improvements in diagnostics and partial understanding of metabolic and biological pathways leading to the discovery of new biomarkers in exhaled air able to characterize and identify disease.

Exhaled air is a mixture of nitrogen, oxygen, carbon dioxide, water, inert gases and traces of volatile organic compounds (VOCs).[3] Several hundreds of compounds are detectable with current analytical methods in each sample and several thousand different compounds have been identified sofar demonstrating the vast diversity of compounds available in exhaled air. To date a few compounds found in breath have demonstrated their value as biomarkers; nitric oxide (NO) levels are generally accepted as an indication of inflammation and oxidative stress in the respiratory tract in for instance asthma.[4–6] Brindicci et al. observed slightly elevated alveolar NO and increased levels of exhaled NO during exacerbations in chronic obstructive pulmonary disease (COPD) patients.[7] Additionally carbon monoxide (CO) has been investigated as a biomarker, however, contrasting results are published. Yamaya et al. found a significant relationship between exhaled CO concentrations and certain lung function markers, and exhaled CO appeared to correlate with the eosinophil count in sputum.[8] In contrast others found no correlation of exhaled CO with lung function.[9] The application of CO as a diagnostic marker is also limited because exhaled CO levels are seriously affected by environmental CO, which may fluctuate considerably and is influenced by active and passive smoking making its use as a biomarker at the least questionable.[10] Exhaled hydrogen peroxide (H_2O_2) has also been studied as a potential biomarker in exhaled breath since H_2O_2 levels are thought to reflect the underlying state of oxidative stress in the lungs. Schleiss et al. demonstrated that levels of exhaled H_2O_2 are higher in stable COPD patients compared to young non-diseased non-smoking controls.[11] However, exhaled H_2O_2 measurements are not yet standardized and demonstrate large intra-individual variability. Table 1.1 lists some additional examples of previous studies performed on the analysis of exhaled air related to disease.

This thesis describes the analysis of breath related to inflammatory lung diseases such as COPD and cystic fibrosis (CF). The following paragraphs describe

in short the pathophysiology of these lung diseases and the related benefit of breath analysis since these two diseases will be dealt with in this thesis.

Table 1.1: List of previous research on exhaled air analysis

VOCs	Condition	Reference	Year
Ethane and pentane	Asthma and COPD[12]	Kharitonov	2004
	Cystic Fibrosis[13]	Barker	2006
Methyl alkanes	Lung and breast cancer[14,15]	Phillips	1999, 2003
Acetone	Dextrose metabolism and lipolysis[3]	Miekisch	2004
	Diabetis mellitus and ketonemia[16]	Deng	2004
Isoprene	Cholesterol metabolism[17]	Stone	1993
Sulfur-containing	Liver impairment[18]	Di Francesco	2005
Nitrogen-containing	Uremia, kidney impairment[18]	Di Francesco	2005
Methyl nitrates	Hypoglycemia in children[19]	Novak	2007
VOC profiles	Asthma[20]	Dallinga	2009
	Lung cancer[14,21,22]	Machado	2007
		Phillips	2005
		Dragonieri	2007
	COPD[22]	Van Berkel	2009
	Smokers[23]	Van Berkel	2008
	Tuberculosis[24]	Phillips	2007

Table 1.1: Examples of previously published research regarding VOCs related to disease present in exhaled air.

COPD

COPD is characterized by the progressive development of non-fully reversible airflow limitation as a result of emphysematous destruction this way increasing the resistance of the small airways. The three pathologic conditions include chronic obstructive bronchitis, emphysema and mucus plugging; the relative extent of emphysema and obstructive bronchitis may vary on a patient to patient basis.
In COPD inhaled irritants like cigarette smoke or mining dust trigger an abnormal inflammatory response during which inflammatory cells infiltrate the lungs causing the airways to become thickened and inflamed.[25,26] However, recent research demonstrates that inflammatory cells and mediators generated in the lungs enter the bloodstream and may have systemic effects on other susceptible areas of the body, thus suggesting that inflammation is not confined to the lungs, but exhibits a more systemic profile.
It is now clear that the structural changes leading to COPD are due to the inflammatory response in the lungs.[25,26] Increased numbers of neutrophils[25] and CD8+ T-cells are observed in COPD.[26] Upon arrival of the neutrophils in the alveoli these neutrophils become activated and generate reactive oxygen species (ROS) that on their turn take part in a process called oxidative stress. Inflammatory mediators and activated inflammatory cells released into

the circulation in COPD include tumor necrosis factor alpha (TNF-*alpha*) and interleukin-6 (IL-6).[27,28] Studies show that reduced lung function is associated with elevated systemic inflammatory factors. These factors, increased during exacerbations, likely contribute to the comorbidities observed in patients with COPD.[28] An illustration of comorbidities related to low-grade inflammation is the increased risk of atherosclerosis in patients with inflammatory rheumatic diseases.[29] The increase in systemic inflammation triggers elevations in mediators such as C-reactive protein, which may contribute to the increased risk of cardiovascular disease[30] and may be a marker of impaired health status.[31] Elevated TNF-alpha may contribute to muscle wasting and cachexia, and TNF-alpha and IL-6 are also both associated with atherosclerosis demonstrating the systemic profile of COPD.[28]

During the last few years the mortality of COPD has increased worldwide, even in industrialized countries. Recent predictions from the World Health Organization state that within 15 years this disease will rise in the current ranking of the most common cause of death from the sixth to the third place mainly due to decreased mortality of other diseases like cardiovascular diseases and a marked increase regarding environmental pollution and smoking as number one cause for COPD. Early diagnosis and treatment will be necessary in order to control its high morbidity and mortality and consequently high healthcare cost.[32] Direct measures of inflammation in bronchial biopsy specimens, bronchoalveolar lavage fluid and sputum samples and measurements of pulmonary function are a few of the standard diagnostic tools available regarding COPD. Spirometric testing is one of the oldest clinical tests still in use today. It is a straightforward test that has the patient maximally exhale from total lung capacity. The forced expiratory volume in the first second (FEV_1) and the maximum exhaled volume (vital capacity [VC]) remain the most indicative measurements. In obstructive lung diseases such as COPD, the characteristic changes in spirometry are a reduction in the FEV_1 with respect to the vital capacity (FEV_1/VC ratio). Using this measurement one can diagnose the presence of airway obstructions. This can be used to guide therapies and predict outcomes.[33] Currently, with exception of lung function tests which are the currently applied diagnostic tool for COPD diagnosis, there are no well validated biomarkers or surrogate endpoints that can be used to establish efficacy of novel drugs for COPD.[34] However, the lung function test is not an ideal measure since according to literature it (1) does not provide information regarding disease activity or the underlying pathologic process, (2) cannot separate the various phenotypes of COPD, (3) is not specific for COPD, (4) is relatively unresponsive to known therapies that prolong survival[30] and (5) depends heavily upon the quality of equipment, the patient cooperation, and the skill of the technician performing the test.[33] Spirometry should thus be considered a medical test and not simply a vital sign that can be performed by minimally trained personnel. Additionally since the applied diagnostic tools cause a varying degree of discomfort for the patient, are time-consuming and sometimes require additional information to ascertain a diagnostic outcome, the analysis of exhaled air might prove useful as a new, non-invasive, safe and

fast diagnostic tool regarding the diagnosis of COPD.

Cystic fibrosis

CF is a hereditary disease affecting exocrine glands. It will, among other multisystem failures, manifest itself as a progressive lung disease characterized by recurrent infectious events. Inflammation in the lungs of patients with CF is characterized by persistent and excessive neutrophil infiltration. Degradation of neutrophils in the airways leads to release of substantial amounts of DNA into sputum, thereby increasing viscoelasticity of secretions.[35] Neutrophils also release large quantities of destructive ROS and proteases, including neutrophil elastase. Increased quantities of ROS lead to increased oxidative stress. The increased amount of neutrophil elastase directly damages the airway wall by digesting elastin and other structural proteins, ultimately leading to bronchiectasis. It also cleaves opsonins and receptors necessary for phagocytosis.[35] Neutrophil elastase, as well as neutrophils and macrophages, generate or stimulate production of pro-inflammatory cytokines and chemokines. Airway concentrations of these chemoattractants are dramatically increased in patients with CF.[36–38] In contrast, CF airways appear deficient in the anti-inflammatory cytokine interleukin (IL)-10, which is constitutively produced by bronchial epithelial cells in healthy lungs.[36,37] In healthy subjects pulmonary epithelia are protected from the destructive effects of neutrophil elastase by antiproteases. The balance between proteases and antiproteases suppresses neutrophil elastase activity and prevents damage to respiratory epithelia,[39] disruption of this balance, due to increased amounts of proteases over the amount of antiproteases leads to structural damage. The worsening of symptoms or exacerbations leads to acute changes in pulmonary symptoms related to increased airway secretions. Pulmonary insufficiency has a great impact on the quality of life and is the main reason of morbidity in cystic fibrosis. In case of an exacerbation, often, intravenous antibiotics and hospital admission are necessary and exacerbations have also been associated with diminished lung function in later life.[40] The yearly exacerbation rate increases with more severe pulmonary impairment and is clearly related to survival.[41,42] Therefore CF exacerbations are an important determinant for quality of life of patients and healthcare cost. Despite the importance of these exacerbations, standardization of the definition and detectability of an exacerbation remains arguably diffuse.[43] In the outpatient clinic, diagnosing a CF exacerbation may be difficult. Changes in symptoms and lung function indices occur when an exacerbation is already manifest. Most clinical trials define an exacerbation by means of surrogate measures of exacerbation, such as hospitalization or intravenous antibiotic administration. To date there is no predictive measure regarding the occurrence of an exacerbation by means of an objective chemical, physiologic or histologic marker. Therefore, it is not possible to anticipate and by means of treatment prevent lung damage on the long term. Early and accurate diagnosis of exacerbations will facilitate quality improvement activities aimed at this important aspect of CF care.[44] Exhaled air analysis holds the potential to deliver markers from breath that might en-

able identification of CF and early detection of the exacerbations associated with it as well as monitoring of disease and exacerbations progression.

Oxidative stress

Oxidative stress is a process in which there appears to be an imbalance between the production of reactive oxygen species and the subsequent biological system's ability to detoxify and repair the resulting damage; an imbalance between the oxidants and anti-oxidants.
Inflammatory mediators and activated inflammatory cells released into the circulation in for example COPD are responsible for generation of ROS. These highly reactive molecules lower the reducing environment - naturally occurring in every organisms cell - even more, and are capable of causing extensive cellular damage. It has been shown that oxidative stress as occurring in diseases like atherosclerosis, Parkinson's disease, myocardial Infarction, Alzheimer's disease and COPD causes cells to be extensively damaged. In a process called lipid-peroxidation cell membranes are oxidized by the ROS. This process proceeds by a free radical chain reaction mechanism as depicted in figure 1.1. It most often affects polyunsaturated fatty acids, because they contain multiple double bonds in between which lie methylene -CH_2- groups that possess especially reactive hydrogens. It starts with a process (initiation) whereby a ROS abstracts a hydrogen atom from the lipid and a fatty acid radical is produced. This phase is followed by a process called propagation in which the previously generated fatty acid radical reacts fast with molecular oxygen. This creates a peroxyl-fatty acid that on its turn is also a very highly reactive molecule reacting with another fatty acid, thus generating a chain reaction giving rise to more and more fatty acid radicals. This cycle continues until a reaction between two reactive radicals produces a non radical thus lowering the overall reactivity of the mechanism.
Ethane belongs to the group of VOCs that is formed as a product of lipid peroxidation of cell membranes and was one of the first VOCs demonstrating availability in exhaled air to be studied.[45,46] It was soon demonstrated to be elevated in exhaled air of COPD patients compared to controls.[45] A correlation was demonstrated between levels of ethane and the degree of airway obstruction, smoking habits and FEV_1.[46] However, the analysis of single compounds from exhaled air is hampered by its low sensitivity and specificity. In 1971 Linus Pauling demonstrated the availability of hundreds of different VOCs in exhaled air[47] and many more have been detected since. Philips et al. have recently successfully demonstrated the possibility of using a profile of VOCs in breath as biomarkers of lung cancer and pulmonary tuberculosis.[14,15,48] In 2006, Barker et al. concluded that analysis of VOC profiles is feasible in exhaled air and holds potential for non-invasive diagnosis as demonstrated in young patients suffering from CF.[13] Therefore particular interest has risen in multi-component analysis of VOCs in order to increase the performance of the diagnostic tool as a whole using a biomarker profile approach. A mathematical model based on a profile will be more predictive and robust as compared to the

analysis of a single compound available in exhaled air. An advanced classification model based on a set of VOCs able to discriminate diseased subjects from controls holds great potential regarding clinical application for the assessment of for example airway inflammation, especially since analysis of exhaled breath might provide a fast, non-invasive, cost beneficial and easy to perform diagnostic tool. Additionally the gas matrix is a relatively simple matrix to analyze, far more straight forward compared to blood and its sampling does not require skilled medical staff.[49]

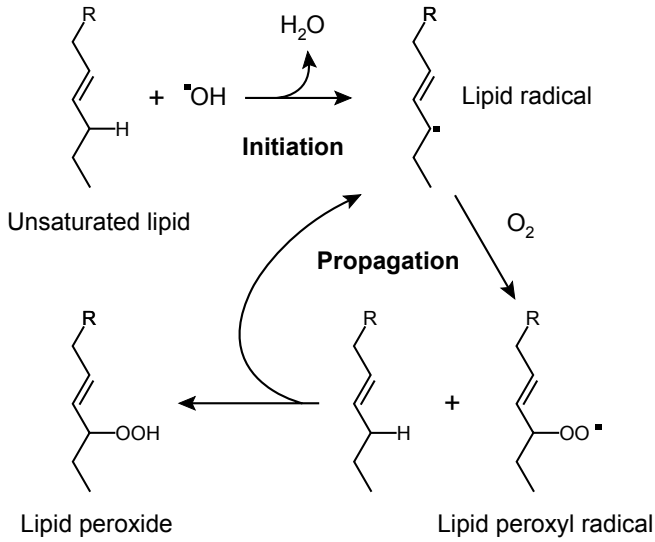

Figure 1.1: Schematic presentation of lipid peroxidation process.

TECHNIQUES

Nowadays the hunt for the best biomarkers is driven by the fast-paced technological advancements in high quality sensors, high throughput analyzes and increases in processing speeds combined with the advances in bio-informatics and biostatistics. Datamining strategies and powerful computers enable researchers to currently find their way through the very large datasets acquired with for example microarray analysis or analysis of exhaled air. Due to the high number of compounds compared to the number of subjects advanced datamining routines are crucial in extracting only the relevant information from the data in order to study the availability of possible biomarkers in the exhaled air.

In order to determine both quantitatively and qualitatively the contents of an exhaled air sample several advanced technologies are available. Research regarding the analysis of exhaled air published recently describe a variety of

these techniques, a few of which will be highlighted briefly followed by a more in depth description of the analysis methodology employed for the studies described in this thesis.

Colorimetric analysis and optical absorbance

Colorimetric analysis is a technique applied to measure certain VOCs in breath. Hydrogen cyanide and acetone have been measured through chemical reactions. In the case of acetone, the VOC reacts with alkaline salicylaldehyde creating a product that absorbs light in a very specific wavelength range centered around 465 nm. By analyzing the degree of absorption the amount of availability of the VOC can be calculated.[50] Especially for acetone the setup has proven to be of high sensitivity. More compounds are currently assessed in a similar fashion. Another optical gas detection system is optical absorption spectroscopy. Recent advances resulting in an increased sensitivity have provided the means to measure trace species in gas samples very accurately with this technology.
A downfall, however, remains the impossibility to measure the entire profile of VOCs in breath. Additionally the high degree of moisture in breath will account for sensitivity issues if the optical absorbance methodology is applied.

Proton transfer reaction mass spectroscopy

Proton transfer-reaction mass spectroscopy (PTR-MS) was first used to determine contents of gas mixtures like breath by Taucher et al.[51] The analysis technique is based on chemical ionization of the target molecules by proton-transfer reactions. During this process the H_3O^+ molecule acts as the primary reactant initiating the reaction. After the chemical ionization the protonated molecules are accelerated and detected with an inline mass spectrometer. H_3O^+ is used since it is most suitable regarding oxygen-containing compounds if a large variety of trace VOCs is to be studied, this due to the fact that almost all of the available oxygen containing VOCs demonstrate proton affinities higher than H_2O, resulting in the occurrence of proton transfer reactions for these compounds. The major downside however remains the fact that it remains a very selective technique and the oxygen-containing compounds only make up a relatively small fraction of the total VOC content of breath
It is nowadays used frequently in research to analyze exhaled air samples. A relation between human breath isoprene levels and cholesterol syntheses was proven by Karl et al. with the use of PTR-MS.[52] Rieder et al. demonstrated PTR-MS proved highly sensitive in measuring exhaled concentrations of certain VOCs and potential tumor markers.[53]
Advantages of this technique are that the samples can be readily analyzed without the need for preconcentration or applying separation processes. Very fast response times (in the order of 100 ms) can be used thus enabling realtime measurements. Additionally the degree of fragmentation of the molecules is

minimized leading to enhanced sensitivity. However, since almost only protonated molecular ions are detected chemical identification of these compounds remains elusive; other techniques need to be applied to identify compounds.

Selected ion flow tube mass spectrometry

Selected ion flow tube mass spectrometry, or SIFT-MS in short, is an analytical technique for the simultaneous real-time quantification of several gases. It is based on chemical ionization of the gases in the gas mixture of interest. In a so-called flow tube, the gases react with a precursor ion, usually, H_3O^+ or NO^+. The product ions produced through this reaction are analyzed with a quadrupole mass spectrometer. Advantages of this technology are the online real-time analysis possibility and the absolute concentration that can be measured down to the parts-per-billion (ppb) levels making it valuable for exhaled air analyzes. Abbott et al. demonstrated the quantification of acetonitrile in exhaled breath with use of Selected ion flow tube mass spectrometry (SIFT-MS).[54]

Chemical sensors

Recent advances in the field of chemical sensors have facilitated a huge progress in the field of these non-selective sensors more familiar under the name 'electronic noses'. Devices that contain a series of non-specific sensors that are able to bind or react with VOCs from gas mixtures like breath. Molecules from gas mixtures generate a so-called profile of the responses of the sensors due to interactions with these trace gases. These responses are then implemented into a training or test set in order to train the device to recognize a certain mixture based on responses from these sensors. The main drawback of this system is that the sensors implemented are -to date- not selective for a single VOC or other trace gas. A large number of VOCs acts on a single sensor demonstrating this technology is still hampered by low sensitivity and specificity regarding classification or prediction of the analyzed gas mixtures. Different sensor principles are applied. Polymer based sensors demonstrate volume changes as they come in contact with VOC from the gas mixture, this way changing the conductance of the polymer. This response can be measured and quantified. In semiconductor based sensors like the quartz microbalance, gas sensors coated with different metalloporphyrins are used. Several research groups now work with electronic noses in the detection of lung diseases and diabetes. Di Natale et al.[55] used quartz microbalance gas sensors showing 100% correct classification of patients with various forms of lung cancer, and 94% classification of controls. Machado et al.[21] used a polymer array sensor based enose and showed promising discrimination between patients with lung cancer and those from other groups, like healthy controls. The authors found 71.4% sensitivity and 91.9% specificity for detecting lung cancer by the electronic nose. Mazzone et al.[56] used a colorimetric sensor array that predicted

the presence of non-small cell lung cancer with a sensitivity of 73.3% and a specificity of 72.4%.

Gas chromatography

The most commonly applied methodology to date to accurately measure trace gases in gas mixtures is based on gas chromatography.[57] This is a two-stage setup consisting of a gas chromatograph (GC) that separates the different compounds in the mixture followed by a standard detector or mass spectrometer (MS).
The GC consists of a capillary column into which the sample is injected. It is then transported through the column by a mobile phase, in most cases an inert gas like helium. Separation of the compounds in the gas mixture is based on differentiated specific interactions of the compounds with the column material and the mobile phase. The lining of the column consists of a polar layer onto which some compounds adhere more than others thus lengthening the passage time of the compound through the column. During the separation procedure the temperature is steadily increased, the higher the temperature the shorter it takes for the compounds to traverse the column. After compounds have been transported through the column they reach a detector, installed to detect and identify the compounds eluting from the column. Different detectors are used of which the flame ionization detection (FID) is the most common one. As the trace gas of interest enters the FID detector hydrogen and air are added to the gas and ignited inside the FID. The organic compounds burning in the flame produce electrons and ions and a large electrical potential set across the flame accelerates these freed ions and electrons towards a collector electrode or plate. The resulting current between the collector plate and the release nozzle is measured with a high-impedance current-meter and fed into an integrator. This measured current corresponds roughly to the proportion of reduced carbon atoms in the flame. Specifically how the reduced carbon atoms are produced is not necessarily understood, but the response of the detector is determined by the number of ionized carbon atoms hitting the detector per unit of time.
Another type of detector is the ion mobility spectrometer. Based on the principle to separate ions according to their mobility as they move through what is called a drift tube. A tube filled with a purified gas (i.e., air of nitrogen) through which the different ions will move at different velocities as they are forced through the tube by means of the applied electric field. The separation of these ions in the drift tube can be optimized by changing the drift length, drift gas, electric field strength, temperature and pressure. The advantages of this type of detector is that it is highly sensitive to certain target compounds and it is relatively inexpensive and portable.
Again another type of detector is based on the time-of-flight (TOF) principle. The basic concept of this TOF-mass spectrometer is identifying compounds by means of the mass-to-charge ratio (m/z). As compounds enter the mass spectrometer they are ionized in an ionization chamber. After ionization the molecular ions and formed fragment ions are accelerated towards a detector

that is set at a known distance from the source. When the charged particle is accelerated into the time-of-flight tube by the acceleration voltage, its potential energy is converted to kinetic energy and will be equal for all equally charged particles. The heavier the particle the longer it takes to complete the path of the time-of-flight tube; the time of flight of the ion varies with the square root of its mass-to-charge ratio. The velocity of the charged particle after acceleration will not change since it moves in a field-free time-of-flight tube. The velocity of the particle can be determined in a time-of-flight tube since the length of the path of flight of the ion is known and the time of the flight of the ion can be measured using a transient digitizer. This kind of setup has proven to be highly sensitive and robust adding to the high degree of reproducibility. Additionally in contrast to the previously described detectors the MS in combination with a GC is capable of producing highly sensitive chromatograms with full mass spectral information of all compounds. Therefore this technique was chosen to be used in the research described in this thesis. Figure 1.2 shows an example of a breathogram (chromatogram of a breath sample).

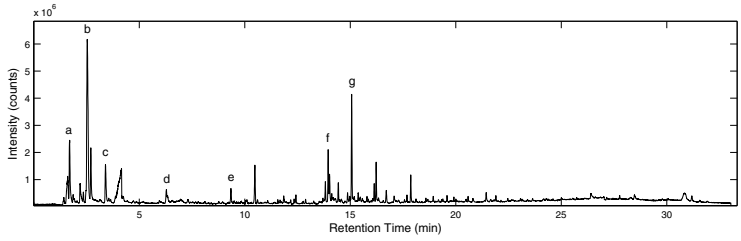

Figure 1.2: Example of a breathogram from a control subject as analyzed by GC-TOF-MS containing a few identified compounds. a) acetone; b) 1,3-pentadiene; c) pentane, 2-methyl; d) benzene; e) toluene; f) 5-hepten-2-one, 6-methyl; g) phenol. The area-under-peak is related to the concentration of the compound.

Data processing

The data generated by the gas chromatograph time-of-flight mass spectrometer (GC-TOF-MS) or any other gas mixture analyzer have to be preprocessed in order to decrease noise, optimize the signal, determine sequences of interest and convert the data into a format suitable to be used by routines to perform the real datamining and isolation of compounds of interest.
Several preprocessing steps have to be undertaken. The data needs to be filtered to minimize the degree of high frequency noise generated by the detector or other instrumental noise introducers. Subsequently baseline corrections have to be performed to improve the very small signals resulting from trace gases in very low concentrations. Peak detection and availability of each compound was determined by analyzing the area under the curve (AUC) for each peak, since

every peak represents a compound and the AUC for every peak is related to the concentration of the compound. Finally every peak was linked to its mass spectrum and retention time. These parameters - retention time and mass spectrum - will be used to identify the compounds and build the database in which all compounds from all subjects will be combined. The resulting dataset will be used in the statistical analyzes. Many different classification algorithms and datamining techniques have been investigated and are explained in more detail in this thesis.

Biomarkers

Introduction of a biomarker or a profile of biomarkers as a new diagnostic tool requires a number of qualities this biomarker should demonstrate as described in the next paragraph. Also both instrumental reproducibility and intra/inter-individual variability have to be mapped and optimized. In case of the analysis of exhaled air this is a difficult task since a large variety of confounding factors will influence the composition of breath within and between individuals.

Breath biomarker qualities

Over the last few decades techniques offering the possibility of a clinical breath test have progressed significantly and a high increase in research and validation on breath tests has been performed in recent years. In order for breath tests to be of clinical relevance this biomarker or profile of biomarkers should exhibit the following characteristics:

- Breath tests should demonstrate a high sensitivity and specificity and breath tests results should be available in a short period of time, resulting in an accurate and fast methodology. It should be a cost effective procedure, with low appliance and operating cost.
- The new biomarker approach should excel in simplicity. This means both in a methodological as in a interpretive way. It should favorably be a non-invasive procedure with a low degree of discomfort for the patient and a high degree of ease-of-use for the medical staff performing the procedure (clinical accessibility). It should, however, also generate an easy to interpret result that leaves very little room for false interpretation. Current research with regards to highly accurate sensor development resulting in quantification of specific VOCs in breath will be of huge benefit to this ease of use if breath biomarker assessment could be performed by handhold breath analyzers like e-nose devices.
- Reproducibility and reliability of the overall applied methodology should be established to a high degree. Instrumental variability should be held at a minimum adding to a high sensitivity. Validation of the biomarker(s)

should be well documented and tested for, for example on separate validation sets, in order to determine the real value of the predictive or discriminatory biomarker outside the test sets.

- In an ideal case the relationship between disease and biomarker(s) should be well defined and linked to disease development, severity and progression. The pathophysiological meaning of the biomarkers should be described and well understood.

Aim and outline of this thesis

This thesis presents an overview of a newly developed breath analysis methodology and the advanced bio-informatics and statistics applied. We chose to use GC-MS in order to measure as many compounds as possible in an exhaled air sample and implemented all detected compounds into a database. In our studies we were not confined to only a one or a few compounds as is performed by many other research groups right now, but we wanted to extract as much information from the exhaled breath samples and use this information to select those compounds of interest that provide insight and information about the health/disease status of the patient or subject analyzed. **Chapter 2** describes our developed methodology of exhaled air analysis and more specifically the validation of this methodology. A closer look will be provided into the sampling procedure, chemical analysis, data handling and preprocessing and finally the advanced statistics applied to isolate the compounds of interest. **Chapter 3** demonstrates the ability of the proposed GC-MS methodology to differentiate bacterial cultures based on VOCs available in bacterial headspace. This is of high value since easy and fast identification of these microbacteria might provide early treatment thus minimizing bacterial-initiated exacerbations in patients suffering from lung diseases. Followed by **chapter 4** presenting data of the first clinical study performed. 50 COPD patients from GOLD classes I to IV and 29 controls were sampled and advanced attribute evaluators were used to isolate those compounds that combined in a classification model provided the highest degree of correct classification. **Chapters 5** focusses on the identification of patients with cystic fibrosis by means of analysis of exhaled air. Several interesting compounds have been identified able to correctly classify a large number of samples based on a model implementing a small number of VOCs. **Chapters 6** provides a detailed analysis of breath with regards to exacerbations in cystic fibrosis patients. VOCs providing insight into the occurrence and phase of an exacerbation were extracted and identified. The study demonstrated that VOCs in exhaled breath are able to indicate CF exacerbations weeks before these adverse events are clinically manifest. Some VOCs demonstrated a trend towards the phase of the exacerbation and might be relevant in monitoring the exacerbation. The thesis will be concluded by a general discussion in **chapter 7** regarding the issues and overall value of the developed methodology.

CHAPTER 2

Development of accurate classification method based on the analysis of volatile organic compounds from human exhaled air

Van Berkel JJBN, Dallinga JW, Möller GM, Godschalk RWL, Moonen EJ, Wouters EFM, Van Schooten FJ
J Chromatogr B Analyt Technol Biomed Life Sci, 861(1):1017, 2008

ABSTRACT

Analysis of exhaled air leads to the development of fast accurate and non-invasive diagnostics. A comprehensive analysis of the entire range of volatile organic compounds (VOCs) in exhaled air samples will enable the identification of VOCs unique for certain patient groups. This study demonstrates proof of principle of our developed method tested on a smoking/non-smoking study population. Thermal desorption and gas chromatography coupled to time-of-flight mass spectrometry were used to analyze exhaled air samples. The VOC profiles obtained from each individual were combined into one final database based on similarity of mass spectra and retention indexes (RI), which offers the possibility for a reliable selection of compounds of interest. As proof of principle we correctly classified all subjects from population of smoking (n=11) and non-smoking (n=11) based on the VOC profiles available in their exhaled air. Support vector machine (SVM) analysis identified 4 VOCs as biomarkers of recent exposure to cigarette smoke: 2,5-dimethyl hexane, dodecane, 2,5-dimethylfuran and 2-methylfuran. This approach contributes to for future development of fast, accurate and non-invasive diagnostics of inflammatory diseases including pulmonary ailments.

INTRODUCTION

Medical diagnostics and monitoring devices are developing at a fast pace greatly improving public health. Non-invasive analytic methods based on the presence of hundreds of volatile organic compounds (VOCs) in exhaled air could further expand the use of diagnostics. Exhaled air is easily obtained from patients which facilitates repeat sampling of the same patient but also larger populations of patients at a lower cost. Many hundreds of VOCs are present in human breath and the opinion is rising these compounds contain valuable information on an individual's disease status.[58] The presence of some of these VOCs in human breath is thought to be due to degradation of polyunsaturated fatty acids by oxidative stress. This process called lipid peroxidation is a chain reaction process in which reactive oxygen species (ROS) remove an allylic hydrogen atom from lipid membrane structures. This gives rise to a conjugated radical that is peroxidized by oxygen and this way prolongs the chain reaction. Among the final stable reaction products of this process are saturated hydrocarbons like ethane and pentane. These hydrocarbons enter the blood stream and due to their low solubility in blood they are excreted into breath within minutes after formation. Therefore, they could potentially be used to monitor the process of oxidative stress in tissues.[3] One of the first exhaled air related studies was performed by Pauling et al.[47] who identified over 200 compounds present in human exhaled air. And indeed some of these compounds have been associated with different pathological conditions. For instance ethane and pentane levels have been linked to oxidative stress and lipid peroxidation[59] and a decrease of exhaled isoprene levels correlated with exacerbations of cystic fibrosis.[60] In 1985, Gordon et al. identified alkanes and mono-methylated alkanes in exhaled air of lung cancer patients,[61] stating the use of the identified compounds as possible biomarkers. In 1999, Phillips et al. selected 22 VOCs to classify subjects with and without lung cancer,[62] and in 2003 modified the VOC pattern by reducing their number to nine.[48] More recently in 2007 Phillips et al. concluded that volatile biomarkers in breath were sensitive and specific for pulmonary tuberculosis.[15] In 2006 Barker et al. proved the feasibility of chemical breath analysis for VOCs as they studied 12 volatile compounds in exhaled air in relation to cystic fibrosis. Only one component demonstrated to be significantly different in CF patients compared to healthy subjects.[13] We developed a more accurate approach of investigating the full range of VOCs in exhaled air and obtained proof of principle by correctly classifying human breath of smokers and non-smokers.

EXPERIMENTAL

Study subjects

A total of 22 subjects, 11 smokers and 11 non-smokers free from chronic lung disease or respiratory tract infection, as confirmed by medical history, were included in this study. Patient characteristics are shown in table 2.1. No

restrictions were applied regarding drugs, alcohol or diet. Subjects were all sampled at one centrally ventilated room at the university. Participation to this study was voluntary. The authors are aware of the small group size, but this study is set up to provide analytical proof of principle of the methodology presented here and will in the future be used on very large subject groups.

Table 2.1: Subject characteristics

	Smoking (n=11)	Non-Smoking (n=11)
Age(years)	54 ± 13	47 ± 11
Packyears	26 ± 19	–
Sex (M/F)	4/7	6/5

Table 2.1: Overview of study population characteristics.

Sample collection and analysis

Exhaled air was collected by exhaling into inert Tedlar bags (5L). Subjects were asked to inhale, hold their breath for 5 seconds and subsequently fully exhale into the Tedlar bag. All Tedlar bags were washed twice with high-grade nitrogen as described by the manufacturer before usage to make sure all contaminants were eliminated. The content of the Tedlar bag was transported under standardized conditions onto desorption tubes; stainless steel two-bed sorption tubes, filled with carbograph 1TD/Carbopack X (Markes International, Llantrisant, Wales, UK). These desorption tubes were placed inside the thermal desorption unit (Marks Unity desorption unit, Marks International Limited, Llantrisant, Wales, UK) and quickly heated to 270 °C in order to release all VOCs and transport the released VOCs onto the GC-capillary. The used desorption unit was highly suitable for repeated, quantitative and reproducible measurements. Ten percent of the sample was injected into the GC, the remaining 90% transported to another adsorption tube for storage and may be used for later reanalysis. Just before the sample enters the GC the sample is trapped by a cold trap at 5 degrees Celsius in order to concentrate the sample. Next VOCs were separated by capillary gas chromatography (column: RTX-5ms, 30 m x 0.25 mm 5% diphenyl, 95% dimethylsiloxane, film thickness 1.0μm, Thermo Electron Trace GC Ultra, Thermo Electron Corporation, Waltham, USA). The temperature of the gas chromatograph was programmed as follows: 40 °C during 5 min., then raised with 10 °C/min until a final maximum temperature 270 °C in the final step this temperature was maintained for 5 min. Time-of-flight mass spectrometry (TOF-MS) (Thermo Electron Tempus Plus time-of-flight mass spectrometer, Thermo Electron Corporation, Waltham, USA) was used to detect and identify components available in the samples. Electron ionization mode was set at 70 eV and the mass range m/z 35-350 was measured. Sample frequency of the mass spectrometer was set to 5 Hz and analysis run time to 33 minutes.

Data-acquisition and data mining

Analysis of the data output files from the GC-TOF-MS was performed in successive steps as described below

Peak detection and corrections

Automated peak detection and baseline correction were performed on the chromatographic raw GC/MS output data files. Baseline correction adjusts the variable background by the following steps: first the background is estimated within multiple shifted windows of width 200 m/z, next the varying baseline is regressed to the window points using a spline approximation, and finally adjusts the background of the input signal. Peak detection consisted of first smoothing the signal. After this step peak locations were assigned. Finally peaks not satisfying specific criteria, like full peak width at half height and maximum base width were eliminated. The Raw GC/MS files contain mass spectra at every MS-scan performed (sample frequency of 5 Hz). The resulting output was saved to a file containing detected peak areas and respective scan numbers. To combine mass spectra and areas belonging to the detected peaks the raw GC/MS files and the peak detection output files were merged through combination of scan numbers. This resulted in a file containing four columns: scan numbers, retention times (RT), peak areas and mass spectra belonging to the detected peaks. Normalization of RTs to retention indices (RI) is necessary to reduce the instrumental variation by adjusting the retention times within each sample run. This was achieved by normalizing RT to the toluene retention time. Next the data were corrected for chromatographic drifting by determining retention indices of 13 widely available compounds (Acetone, 1-propanol, benzene, toluene, furfural, xylene, styrene, heptanal, phenol, D-limonene, decanal, diethylphtalate, diphenylsulfide) in each chromatogram. RI times of these compounds were used in applying corrections in order to line-up the RI indices of all the sample files against one reference sample file. Polynomial functions and interpolation was used to obtain the best fit and correct all RT entries.

Matching peaks based on similarity of mass spectra and retention indices

Subsequently all the corrected files - one for each exhaled air sample - were combined into one large database file by lining up all calculated peak areas of the according compounds based on RI window settings and similarity match factors (MFs) between mass spectra. The MFs between mass spectra were calculated using the best performing routine according to Stein and Scott;[63] the dot-product function that measures the cosine of the angle between spectra represented as vectors. In order to combine the output files of all individuals into one working file suited for statistical analysis, one file was chosen as reference file based on the overall quality of the measurement. Next a second output file was selected. Compounds from this second file were to be combined with

the complementary compounds from the reference file. Combination of these compounds was based on mass spectra similarity with use of the MF-values and the potentially complementary compounds needed to be within a certain RI-range. If no good fit was found the peak was added to the reference file as a new entry. This data combination routine was repeated for every file to be included. Finally the resulting dataset was checked for RI inconsistencies and compounds demonstrating these RI-inconsistencies were removed if necessary. This RI-inconsistencies-check is based on the fact that if the same instrumental procedure is used for the analysis of different samples the RI-order of the detected compounds must be the same.

Quantification of peak areas

After all the corresponding peak areas of the complementary compounds were combined into one large dataset, normalization of the peak-area data was performed in order to be able to compare the different peak areas from different samples. This was necessary because the exhaled air samples contained different unknown absolute volumes of exhaled air, which makes comparison of amounts of compounds impossible. Another reason for normalization is to correct for fluctuations in the response of the mass spectrometer. Different types of global normalization have been evaluated. The most promising rescaling factor used in this study is based on the cumulative area under the detected peaks and implemented into the final database file. Since all chromatograms display rather similar profiles this method of normalization is most robust. Another benefit regarding this area scaling factor is that it does account for the baseline noise present in the raw chromatographic signal. A measure to rule out most of the noise resulted in discarding peaks with RI$<$0.15 and RI$>$2.8. The deleted noise was mostly due to a high degree of column bleeding after RI$>$2.8. Also the very light compounds that elute from the column before RI$<$0.15 usually contain noisy mass spectra in our setup.

Classification model

To determine which compounds added to the database were of interest with regard to the classification of smoking and non-smoking subjects, we used support vector machines (SVM). Several experiments have been performed with different classifiers like random forrest, discriminant analysis, principal component analysis. These experiments demonstrated SVM to outperform all others regarding compound selection. SVM was able to select those compounds that provided the best performance as implemented into a classifier. SVM demonstrate the ability to construct predictive models with large generalization power even in the case of large dimensionality of the data or when the number of observations available for training is low. SVM always seeks a globally optimized solution and avoid over-fitting. This implies a large number of attributes (i.e. compounds) is allowed. This encouraged us to implement this subset selection algorithm into this study. This algorithm will select the most optimal subset

of compounds able to correctly classify our dataset. A variety of subset selection methods was tested, like gain-ratio attribute evaluator. The best subset of compounds was selected using the attribute selection option implemented in Weka:[64] a collection of machine learning algorithms for data mining tasks. Attributes were selected using an SVM attribute evaluator. The attribute evaluator we used evaluated the worth of a subset of attributes by considering the individual predictive ability of each feature along with the redundancy between them. Preferably features will be selected showing high correlations within the class and low inter-correlation. Next the selected attributes were analyzed and ranked with use of SVM using recursive feature selection and removing one attribute at a time. This way attributes were selected using the weight magnitude as ranking criterion. After every run the least efficient attribute was removed. All resulting subsets were analyzed for classification performance with use of support vector classifiers based on John Platt's sequential minimal optimization algorithm and the random forest classification algorithm.[65]

RESULTS

Reproducibility and variability

To validate the newly developed method to extract the discriminating compounds, the instrumental reproducibility and inter- and intra-individual variability were tested as well differences in exhalation patterns.

Instrumental reproducibility

Instrumental reproducibility was determined by analyzing identical exhaled air samples that were obtained by emptying a filled bag over y-shaped connector onto two absorption tubes. The two absorption tubes were subsequently analyzed by GC-TOF-MS. This experiment was repeated 6 times. The instrumental reproducibility was demonstrated by comparing the two complementary chromatograms as demonstrated in figure 2.1. Already from visual inspection of the two chromatograms it can be concluded that the two chromatograms are highly similar, confirming a high degree of instrumental reproducibility. The quantification of the similarity was done by means of calculation of a distance measure (dot-product rule) and is presented in the boxplot of figure 2.2. This distance measure is based on the similarity of the entire raw chromatogram. Distance measure calculation of all complementary files resulted in a distance measure ranging from 0.96 to 0.99. A value of '1' denotes identical samples, the lower the value the lesser the degree of similarity.

Inter- and intra-individual variability in VOC-profiles

Intra-individual and inter-individual variability were also mapped. Intra-individual variability was examined by repeated sampling of exhaled air from 10 non-smoking subjects for 5 consecutive days and comparing the results per subject from day to day. Inter-individual variability was examined by sampling 10

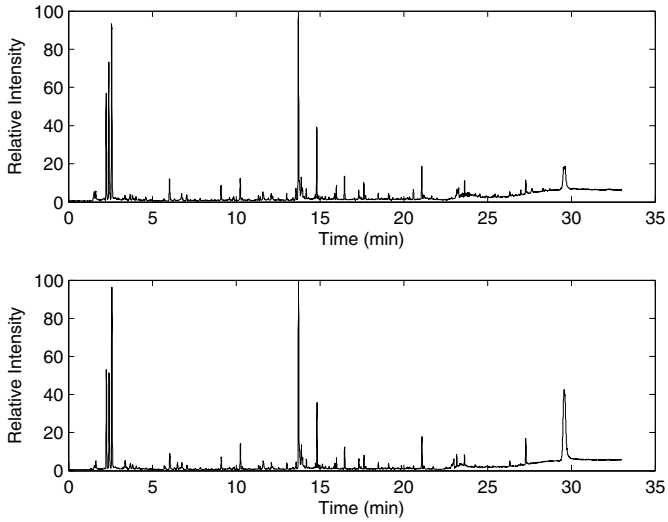

Figure 2.1: Example of two chromatograms demonstrating Instrumental reproducibility. The measured samples contained identical exhaled air samples. Visual inspection confirms high degree of similarity.

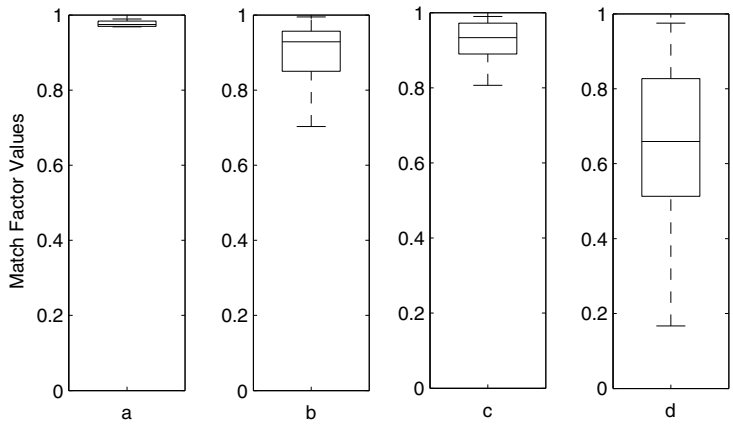

Figure 2.2: Boxplots of match factors that are based on similarity between raw chromatograms. Boxplot representing a) instrumental reproducibility match factors, b) exhalation flow rate depended match factors c) intra-individual variability match factors and d) inter-individual variability match factors. As demonstrated the instrumental reproducibility (a) is by far the smallest, and as expected the inter-individual variability is larger (d) compared to the inter-individual variability (c).

non-smoking subjects and comparing the data from subject to subject. Examples of the resulting chromatograms are shown. In figure 2.3 one subject sampled at two consecutive days is presented and it can be seen that the two chromatograms show a high degree of similarity. Figure 2.4shows chromatograms from two different subjects sampled in the same room at the same time. Shown chromatograms demonstrate that the degree of similarity is less as compared to the chromatograms of figure 2.3. Again the similarity between several chromatograms was quantified using a distance measure as previously mentioned. The results regarding inter-individual and intra-individual variability are shown in figure 2.2. This figure shows boxplots representing a) instrumental reproducibility match factors, b) exhalation flow rate depended match factors c) intra-individual variability match factors and d) inter-individual variability match factors. As expected it can be seen from this figure that the intra-individual variability ranging from 0.80 to 0.99 is far smaller than the inter-individual variability ranging from 0.16 to 0.98; this is consistent with previously performed studies.[66,67]

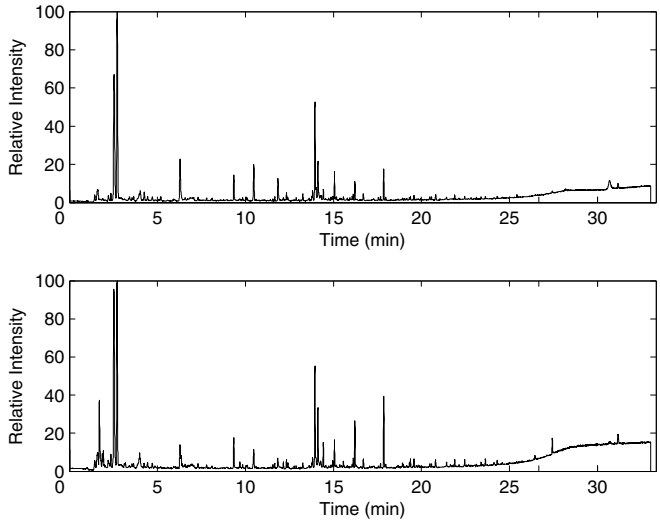

Figure 2.3: Examples of representative chromatograms from a subject sampled at two consecutive days to examine intra-individual variability.

Exhalation characteristics and its impact on VOC-profiles

In order to determine whether standardization of the sampling method of the subjects is necessary an experiment was performed to explore the effect of different exhalation patterns on VOC profiles. To determine whether differences in exhalation air sampling of subjects was a variable in our newly developed methodology 5 non-smoking subjects inflated 2 Tedlar bags as follows: one bag

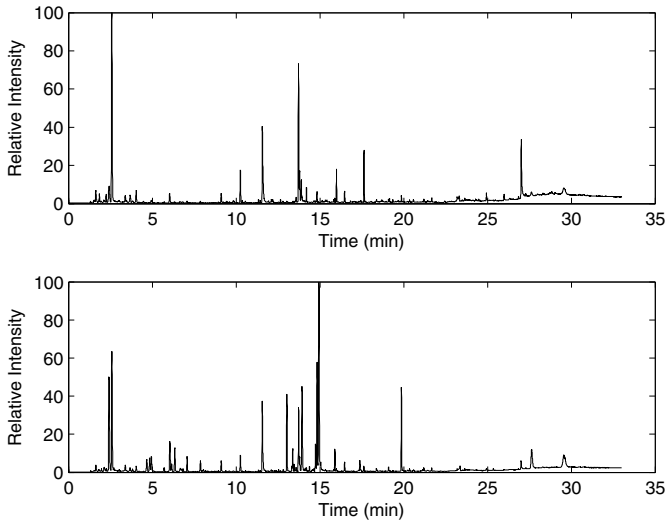

Figure 2.4: Examples of representative chromatograms from two different subjects sampled at the same time to examine inter-individual variability.

was inflated by superficial exhalation and the other one was inflated after deep inspiration, a 5 second breath hold and subsequent total exhalation into the sample bag as suggested by Barker et al.[13] This procedure was repeated five times with approximately 90 minutes intervals in the same centrally ventilated room. In figure 2.5 the resulting chromatograms are shown and as judged already from visual inspection it can be concluded that these complementary chromatograms demonstrate a high degree of similarity suggesting that superficial and deep exhalation are resulting in similar VOC profiles. Mann-Whitney testing showed that only 58 out of the total of 1201 overall detected compounds proved to be statistically different (p=0.05) for the two different exhalation methods. After correcting for multiple testing by applying Bonferroni correction for the alpha-value no compound proved to be significantly affected by the exhalation characteristics. Again quantification of the similarity was done by means of calculation of a distance measure and is presented in the boxplot of figure 2.2b. As can be seen the degree of similarity is comparable to the intra-individual similarity, proving difference in exhalation patterns did not lead to significant difference in VOC profiles within an individual.

Validation of methodology on exhaled air from smokers and non-smokers

To validate the methodology we analyzed the exhaled air from 11 smoking and 11 non-smoking subjects. The subjects exhaled a mean of 381 identified different VOCs. All 22 subjects were combined into one large database. In

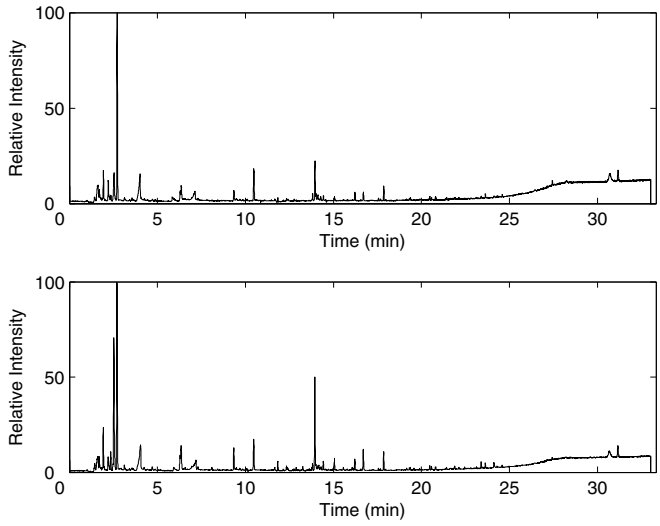

Figure 2.5: Examples of representative chromatograms from a subject inflating a bag superficially (upper graph) and deeply (lower graph).

order to correctly combine the peaks from different subjects an MF-threshold value of 0.85 and an RI-window value of 0.045 were used. This resulted in a database of 22 subjects and 3211 compounds, 467 compounds were present in at least 5 of the 22 subjects. Compounds that were detected in only 2 of the subjects or less were discarded since these compounds do not exert any discriminatory power due to their low occurrence rate and might introduce noise if implemented into the classification model. This value of at least 3 times availability has been introduced by trial and error testing of different threshold values and from similar experiments as mentioned in literature.[68] Applying this threshold criterion resulted in a database consisting of 1095 components. The selection of peaks that discriminate smokers from non-smokers as described in the 'Classification model' section was based on this final database. The most optimal classification model was based on a support vector classifier using just 4 VOCs. This model classifies all subjects correctly regarding their smoking behavior as tested with a 10 times cross validation. Other classification models like random forest, random tree, multilayer perceptrons and Bayesian classifier were also used but the various classifiers tested did not yield improved performances. The same observation was reported by Guyon et al.[69] Since the model based on SVM outperformed other classifiers, this type of classifier was selected. We identified VOCs implemented into the classification model with spectrum recognition using the NIST library in combination with spectrum interpretation by an experienced mass-spectrometrist and identification based on retention times of components. Table 2.2 shows the identified VOCs. Figure 2.6 shows the relative amount of the relevant compounds available in the ex-

Table 2.2: Identified VOCs

Compound Name	Retention Time (min.)	No. times detected in 22 samples
2,5-dimethyl hexane	7.98	11
Dodecane *	17.45	17
2,5-dimethylfuran	7.46	8
2-methylfuran	4.16	7

Table 2.2: Compounds used in the classification model to classify smokers and non-smokers using VOCs in exhaled air. *:confirmed by retention index.

haled air. Bars left from the dotted line represent non-smoking subjects, bars to the right of the dotted line represent the smoking group; the height of the bar represents the normalized integrated peak area of the selected component. As can be seen from figure 2.6 the combined classification power of these four compounds is in most subjects based on availability in smoking subjects versus absence or levels below detection limit of these compounds in exhaled air of non-smoking subjects. As can be seen from figure 2.6 the individual classification power of each compound is not 100%. The SVM implementing and combining data from all 4 compounds is however able to classify all subjects correctly.

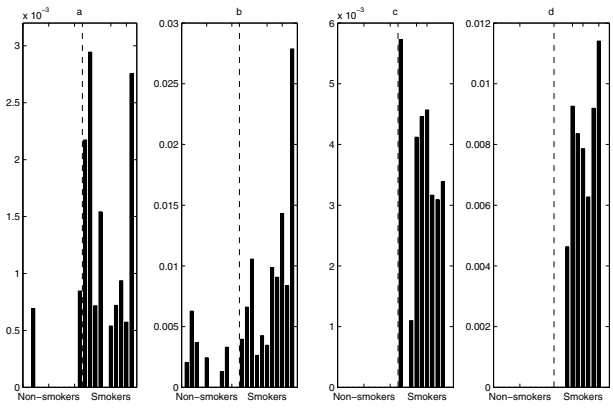

Figure 2.6: Compound availability of identified discriminatory compounds used in an SVM model able to classify all smoking/non-smoking subjects correctly. Left sides of the graphs depict the non-smoking subjects, the right sides depict the smoking subjects. a) 2,5-dimethylhexane b) dodecane c) 2,5-dimethylfuran , d) 2-methylfuran.

DISCUSSION

VOCs in exhaled air are thought to represent several processes in the human body, like metabolism and lipid peroxidation, and therefore have a great potential as non-invasive biomarkers of human health, presence and possibly severity

of disease. In this paper a newly developed method for analyzing and data processing of exhaled air samples has been presented and tested on a small validation set of exhaled air samples of 11 smoking and 11 non-smoking subjects. We are aware that the analysis of exhaled air and more specifically the analysis of VOCs from exhaled air and relating them with disease is not a new approach.[47,62] We want to emphasize that in the present study, a more robust method was developed for sampling and datamining of the acquired data than was published until now. One of the main advantages of our approach is that raw mass spectra are used to find complementary compounds in all subjects, instead of combining compounds based on identity. The match factor as described by Stein et al.[63] is implemented to determine the degree of similarity between measured mass spectra instead of comparison against library mass spectra. We experienced that first identifying the compounds and than finding complementary compounds in the samples based on compound names, introduced more mismatches compared to matching based on the raw mass spectra. We are confident that comparing the retention times and match factors will result in more correctly combined compounds in the respective subjects. The selected subjects exhaled into a Tedlar bag and volatile organic compounds were trapped on desorption tubes and analyzed with use of a gas-chromatograph in line with a time-of-flight mass spectrometer. The resulting data were processed using newly developed routines. To validate this analytical method several reproducibility and variability measurements were performed to assess instrumental variability and both inter- and intra- individual variability. As demonstrated by figure 4 the instrumental variability (a) is very small which confirmed the high reproducibility of our technology. As expected, inter-individual variability is larger than intra-individual variability and both show greater variation than instrumental variability, again confirming the reliability of our methodology. Other studies detecting VOCs in exhaled air mentioned the necessity to correct for chemical background appearing in their samples. In our case, no background corrections have been taken into account. This due to the fact it will not be possible to correct for the complex interdependencies between excretion and uptake of VOCs by easily subtracting the inhaled from the exhaled air.[3] Moreover, background noise will be randomly distributed between subjects' samples and would thus not exert any discriminatory power, nor interfere with the outcome of the analyses. We are aiming with discriminative analysis only to select those compounds that are specific for the disease or condition and should thus principally not depend on background chemicals. The data-analysis design was finally tested on a dataset containing 1095 VOCs from 11 smoking and 11 non-smoking subjects. After classification analysis, a support vector classifier based on only 4 compounds - identified as 2,5-dimethylhexane, dodecane, 2,5-dimethylfuran and 2-methylfuran - was able to correctly classify all subjects based on 10-times cross validation. The authors are aware that simpler statistical approaches like T-statistics or discriminant analysis will perform similar in a small group size as the one used in this study. But since this methodology was designed to be used on large groups with hundreds of subjects a powerful approach like SVM was chosen.

We are aware of the fact that use of an SVM classifier to correctly classify 22 subjects is a bit overpowered, but here we merely provide a proof of principle. The origin of the discriminating compounds in exhaled air remains unclear so far, although these compounds have been identified previously in relation to smoking. In 2002 Gordon et al. already demonstrated 2,5-dimethylfuran to be a promising breath biomarker in detection of active smoking[70] and Sanchez et al. in 2006 identified 2,5-dimethylfuran and 2-methylfuran as strong indicators of smoking status.[66] Although it is well known that active cigarette smoking directly affects the levels of benzene and other VOCs in breath of smokers and previous research demonstrated that concentrations of benzene detected in exhaled air of smokers are always higher than of non-smokers,[67,71] benzene and other important constituents of cigarette smoke have not been included in our most optimized model. The exclusion of for example benzene is because this model represents the best subset of compounds that provides the most optimal classification, also taken the redundancy of the compounds into account. In conclusion, this study demonstrated the functionality of our approach of exhaled air analysis by demonstrating discrimination based on smoking status of subjects. The presented methodology is very accurate and has great power. This design regarding the analysis and identification of discriminatory biomarkers in exhaled air might allow for non-invasive monitoring of inflammation and oxidative stress in the respiratory tract in patients suffering from (inflammatory) lung diseases.

ACKNOWLEDGEMENTS

The authors acknowledge the province of Limburg for the financial support of this research at the University of Maastricht.

CHAPTER 3

Identification of micro organisms based on gas chromatography-mass spectrometry of volatile organic compounds in headspace gases

Van Berkel JJBN, Stobberingh EE, Boumans MLL, Moonen EJ, Wouters EFM, Dallinga JW, Van Schooten FJ
Submitted

ABSTRACT

Background

The elucidation of volatile chemical compounds specifically produced by microorganisms may assist in developing a fast and accurate methodology to determine pulmonary bacterial infections. Identification of microorganisms could be done by determination of predefined compounds in culture headspace. Development of this methodology might ultimately lead to the identification of bacterial species.

Methods

Over 300 bacterial headspace samples from 4 different microorganisms were analyzed by gas chromatography-mass spectrometry to identify relevant VOCs, and compose a profile of VOCs enabling identification of the different microorganisms (*Escherichia coli*, *Pseudomonas aeruginosa*, *Staphylococcus aureus* and *Klebsiella pneumoniae*). Both differently abundant VOCs were determined and classification models based on support vector machines (SVM) were used to allow classification of the samples. Cross validation was applied to validate the results.

Findings

We were able to identify a large number of compounds demonstrating a highly significant difference in availability in bacterial cultures compared to medium and in cultures compared to one another. We have also determined highly significant compounds differing between the four *Escherichia coli* strains and between the two *Staphylococcus aureus* isolates: methicillin-resistant *Staphylococcus aureus* (MRSA) and methicillin-sensitive *Staphylococcus aureus* (MSSA). SVM models were able to classify the microorganisms with very high degrees of sensitivity and specificity based on -on average- 6 VOCs from headspace.

Interpretation

We demonstrated that identification of the studied microorganisms is possible based on a few compounds measured in headspace of these cultures. It provides a fast, non-invasive, cost effective and sensitive technique as a potential diagnostic approach in medical microbiology.

INTRODUCTION

Recent advances in technologies allow researchers to analyze in more detail metabolic processes and the products evolving from these processes in microorganisms. Detection and identification of microorganisms might prove useful in fields as diverse as biotechnology and infection biology. There is an urgent need for rapid diagnostic tests that can at point of care identify the pathogen(s) and the presence of markers of resistance. For serious infections and when resistance is present immediate early therapy is critical. Therefore, the use of rapid diagnostic tests to confirm the presence of a pathogen and/or a resistance biomarker directly from the clinical specimen and subsequent accurate antimicrobial therapy, can potentially improve clinical outcome and facilitate the efficient conduct of clinical trials.[72] In clinical practice, rapid diagnostic tools can avoid useless antibacterial treatment in case of viral infection and can allow selection of appropriate antimicrobial treatment for specific pathogens. Quantitative and qualitative analysis of certain specific volatile organic compounds (VOCs) in headspace might enable mapping of the metabolic processes in microorganisms and enable fast identification of these microorganisms. Probert et al. investigated the composition of VOCs in headspace from stool samples from patients suffering from infectious diarrhea. Different VOC profiles were demonstrated to correlate with different causative organisms.[73]
Micro organisms release these VOCs mainly as metabolic products during growth, as secondary metabolites for protection against antagonists and competitors, or as signalling molecules in cell-to-cell communication. Previous research has already isolated profiles of VOCs from ascomycetous yeast strains and profiles of VOCs have also been determined for the classification of endophytic fungi. Gas chromatography-mass spectrometry has been used for the identification of bacterial VOCs from cultures of cyanobacteria. However no research has been reported sofar on the analysis of the entire profile of VOCs produced or altered by microorganisms.
In the present time VOCs related to microorganisms are once again considered for their potential in identification and monitoring due to the development of efficient sample collection, sensitive analytic technologies for separation and identification and more advanced statistical analyses able to datamine the generated output. Gas chromatography-mass spectrometry (GC-MS) is highly suitable for identifying and characterizing VOCs related to certain microorganisms that at a later stage could be exploited through sensor-based detection like enose.[74] The objective of this study is to analyze all VOCs from various microbial cultures by GC-MS and determine which VOCs or profiles of VOCs might enable rapid and easy identification of the cultures. Here we present a sampling and analysis methodology that is capable of identifying significantly different VOCs between different microbial cultures. Additionally we report the use of classification models based on support vector machine (SVM) classifiers. These VOCs as implemented into an SVM classify these cultures with a high degree of accuracy.

METHODS

Samples

Four different bacterial species were tested: *Escherichia coli (E. coli)* (n=75), *Pseudomonas aeruginosa (P. aeruginosa)* (n=52), *Staphylococcus aureus (S. aureus)* (n=81) and *Klebsiella pneumoniae (K. pneumoniae)* (n=40). In order to assemble the necessary control samples 70 flasks containing only medium underwent the same procedure in order to achieve the highest degree of methodological standardization. Four different *E. coli* strains were used: *E. coli* ATCC 25922 (n=30), the extended spectrum bèta-lactamase producing (ESBL) *E. coli* ATCC 35218 (n=24), *E. coli* 47.4.1.039b (n=10) and *E. coli* 34.4.2.038 (n=11). The *S. aureus* species were subdivided into two strains: methicillin-resistant *Staphylococcus aureus* (MRSA) (n=41), methicillin-sensitive *Staphylococcus aureus* (MSSA) (n=40).

Sample collection and analysis

Bacteria were grown on blood agar plates and incubated overnight at 37°C. Next bacteria were transferred into 4.5 ml sterile Brain Heart Infusion broth (CM0225 Oxoid) and grown for 4 hours under constant agitation at 37°C. Subsequently 0.5 ml of the culture was transferred into 100 ml sterile Mueller-Hinton broth (CM405 Oxoid) in 1L culture flasks. After overnight incubation at 37°C the culture flasks were flushed with high-grade nitrogen connected to the inlet of a custom made flaskcap. Designed in order to transport the contents of the headspace of the cultures, under standardized conditions, onto stainless steel two-bed sorption tubes connected to the outlet connection of the flaskcap. Effort was made to minimize air contamination at opening and connecting the flasks. The custom made flaskcap provided the means to do so. The connected desorption tubes are packed with carbograph 1TD/Carbopack X (Markes International, Llantrisant, Wales, UK) that trap VOCs. These desorption tubes were placed inside a thermal desorption unit (Marks Unity desorption unit, Markes International Limited, Llantrisant, Wales, UK) and subsequently heated to 270°C in order to release all VOCs onto the gas chromatography capillary column (RTX-5ms, 30 m x 0.25 mm 5% diphenyl, 95% dimethylsiloxane capillary, film thickness 1.0 μm). The desorption unit is highly suitable for repeated, quantitative and reproducible measurements. VOCs are separated by GC (ThermoFisher Scientific., Austin, Texas, USA) and subsequently detected by a time-of-flight mass spectrometer (TOF-MS) (Thermo Electron Tempus Plus time-of-flight mass spectrometer, ThermoFisher Scientific, Austin, Texas, USA). The temperature of the gas chromatograph was programmed as follows: 40°C during 5 min., then raised with 10°C/min until the final temperature of 270°C. This temperature was maintained for 5 min. Electron ionization at 70 eV was used combined with a 5Hz scanning rate over a mass range of m/z 35-350 amu. An example of a bacterial headspace chro-

matogram is shown in figure 3.1. Each peak represents a compound and the area underneath the peak is related to the amount the compound was available in.

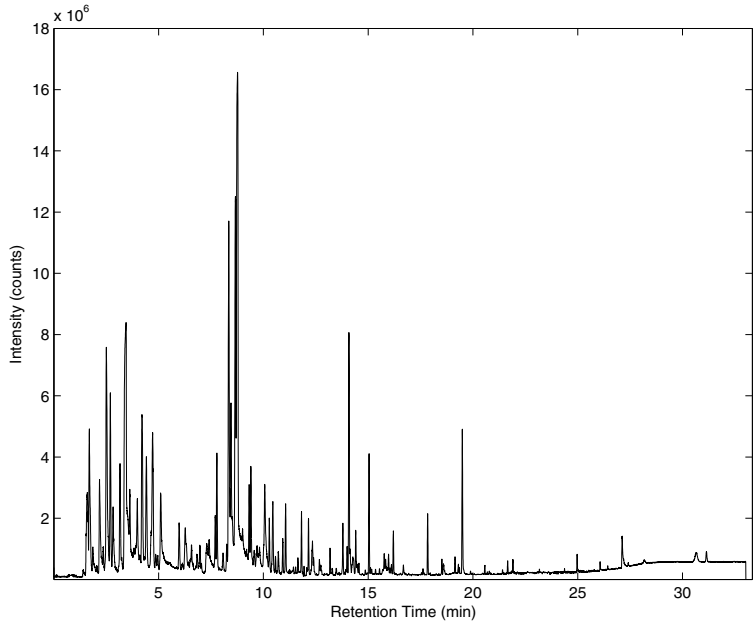

Figure 3.1: Example of a chromatogram produced by a GC-TOF-MS measurement of *E. coli* ATCC 35218 headspace. Each peak represents a compound, compound quantification can be done by calculation of the area under the peak.

Data-acquisition

Analysis of the data output files from the GC-TOF-MS was performed in successive steps as previously described in detail.[22] The output files generated by the GC-MS software were converted to ASCII format and subsequently processed by our in house developed software based on MATLAB®. This included the following steps. After denoising the data and application of baseline corrections on all analysis output files peak detection was performed and the area under the peak was calculated. Normalization of the calculated peak areas was performed using an area scaling factor. This rescaling factor is based on the cumulative area under the detected peaks since this standardization principle proved to be the most robust.[22] To correct for chromatographic drifting the retention times (RT) of all samples were adjusted. This was done by selecting a number of peaks available in all samples and using a linear regression fit. This correction is very effective, easy to perform and eliminates the use of an added internal standard, adding to the straight-forwardness and robustness of

the presented methodology.
The output files were then merged by combining corresponding compounds based on retention time and degree of similarity of the corresponding mass spectra, by determining the match factor values (MFs). The degree of mass spectra similarity using a match factor was based on the similarity index as described by Stein et al.:[63] the dot-product function that measures the cosine of the angle between spectra represented as vectors. These match factors were only determined for compounds within a selectable RT-window. MF-threshold values were determined based on a variety of complementary compounds manually combined. The MFs calculated for these compounds demonstrated to be at least 0.842. Since compounds available in a very low number of samples might obstruct the statistical analysis (since these compounds will be highly effective in discriminating the corresponding samples) a minimum availability threshold of 8% of samples was introduced according to Penn et al.[68]

Component selection

For the selection and determination of interesting compounds two different approaches were used:

- Analyzing the compounds and determining the significantly different compounds based on a t-test combined with a Bonferroni correction as the dataset demonstrated to be normally distributed. These significantly different compounds might hold valuable information and effort was made to select those compounds that hold the potential of a single biomarker for the detection of the bacteria. Analysis of these compounds might also provide insight into the changes in physiology for future research. We are however aware that numerous compounds are still not biologically linked to metabolic or disease pathways.

- Informative compounds fitted into a support vector machines (SVM) classification model were selected. This approach provides information regarding a profile of VOCs combined into an SVM model that holds predictive power towards detection of the bacteria. Previous research already demonstrated classification based on a profile of VOCs is far superior compared to classification based on single compounds.[22]

The SVM approach was chosen for its ability to construct predictive models with large generalization power even in the case of large dimensionality of the data when the number of observations available for training is low,[69] which is obviously the case here. SVM are specifically useful since it seeks a globally optimized solution and avoids over-fitting, so a large number of features or compounds is allowed. The compounds are selected through a number of variable selection criteria. This selection algorithm will select the optimal subset of compounds able to correctly classify the dataset. A variety of subset selection

methods was tested, among which the gain-ratio attribute evaluator. In order to obtain the best subset of compounds the attribute selection option implemented in Weka (Waikato Environment for Knowledge Analysis)[64] was used. Compounds were selected using an SVM attribute evaluator. The attribute evaluator we used evaluated the worth of a subset of compounds by considering the individual predictive ability of each compound along with the redundancy between them. Preferably compounds were selected showing high correlations within the class and low inter-correlation. After every run the least efficient compound was removed. A subset of the highest ranking compounds was implemented into an SVM classifier trained with John Platt's sequential minimal optimization algorithm.[65] The SVM classifiers were validated and performance was tested using 10 times cross-validation in which the entire dataset is split repeatedly into a test set (90% of samples) and a validation set (10% of samples).

RESULTS

In headspace of the 318 samples a total of 5000 compounds were detected by GC-TOF-MS. From the total VOCs measured a number of 912 different compounds were implemented into the final dataset based on the 8% inclusion rule as proposed by Penn,[68] which means that each implemented compound was available in at least 25 samples. This is necessary in order to avoid selection of compounds detected in a small number of samples (or even only one sample) resulting in a large discriminatory power but low predictive value of these compounds with regards to the few number of samples they were detected in. The data were tested for normal distribution before significantly different compounds were determined in different approaches.
The first approach was comparison of samples of headspace from flasks containing only growth medium to headspace from samples containing microorganisms. The second approach studied a cross-comparison of headspace of all individual bacterial cultures, the third approach looked for discriminating compounds between the four different strains of *E. coli* and finally the fourth approach studied discriminating compounds between the two *S. aureus* strains. For each configuration significantly different compounds were selected ($p \leq 0.05$). Tables 3.1-3.10 display the most significantly different compounds, their availability and the relevant p-value.
The total number of significantly different compounds between bacterial cultures and flasks containing only medium was 44. Table 3.1 shows the 10 most significant compounds and their identity as based on the NIST library and spectrum interpretation by an experienced spectrometrist. Presented in figure 3.2 is a heatmap demonstrating the availability of these ten highly significant compounds in all samples as identified in table 3.1 in which compound number is equal to compound number as denoted in the heatmap. The compounds displayed in this heatmap are all normalized to maximum (=1) in order to produce an interpretable graph. As shown in figure 3.2 the discriminatory power of the

most significantly different compounds for bacterial cultures are those with a high availability of these compounds in medium and absence or lower availability in cultures. There are however also several compounds that demonstrate their discriminatory power based on availability of these compounds in cultures and absence or lower availability in medium. This indicates constituents of the growth medium are used for growth of the bacteria and as such consumed by them resulting in disappearance of VOCs.
The total number of significantly different compounds with regards to the second approach determined with a cross-comparison of headspace of all individual bacterial cultures ranged from 23 to 41. Tables 3.2-3.5 show for each culture the 10 most significantly different compounds and their identity based on the NIST library and interpretation of the mass spectra by an experienced spectrometrist. Presented in figure 3.3 is a heatmap demonstrating the availability of the ten most significantly different compounds in all cultures demonstrating p-values ranging from 0 to 0.00011. As shown in figure 3.3 the discriminatory power of these ten most significantly different compounds is based predominantly on availability in the selected bacterial culture versus absence of these compounds in the remaining bacterial cultures or vice versa. This indicates that subsets of VOCs are found that are highly specific for a certain culture.
The third approach analyzed significantly different compounds from the 4 dif-

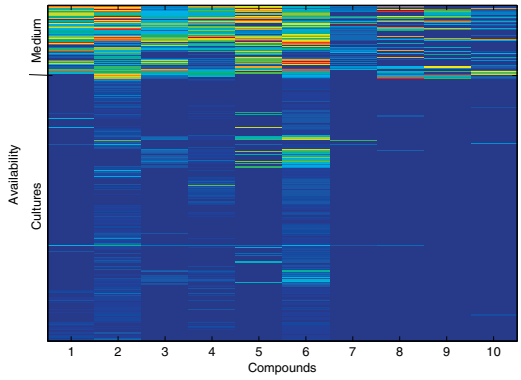

Figure 3.2: Heatmap demonstrating significantly different compounds and their availability. Compounds were selected as comparing headspace from medium samples to all headspace samples containing bacteria. Compound numbers denote identification of compounds as presented in table 3.1.

ferent *E. coli* strains. Depending on which strains were compared the number of significantly different compounds ranged from 3 to 15 compounds. Results are presented in a similar fashion as described above in tables 3.6-3.9 and figure 3.4. The number of significantly different compounds is less compared to the previous studied configurations demonstrating a higher degree of similarity between the 4 strains. *E. coli* 34.4.2.038 only demonstrated three compounds of significant interest, *E. coli* 47.4.1.039b a total of 12 *E. coli* ATCC 25922 a

total of 15 and *E. coli* ATCC 35218 a total of 10. Figure 3.4 demonstrates that the discriminatory power of significantly different compounds in *E. coli* 34.4.2.038 is based both on availability and absence (or lower availability) of these compounds in contrast to the other configurations where the discriminatory power is mainly based on availability in the selected culture versus unavailability in the remaining cultures. The lower number of different compounds regarding the four *E. coli* strains is in accordance with the hypothesis that cultures demonstrating a high degree of similarity do also demonstrate a high degree of similarity in their headspace.
The final approach determined significantly different compounds between the 2 *S. aureus* strains; MRSA and MSSA. Results are presented in table 3.10 and figure 3.5. As can be seen from these data the MSSA strain demonstrates availability of compounds for 8 out of the 10 most significantly different ones. The remaining 2 compounds demonstrate a higher degree of availability in MRSA. Almost all 10 compounds could accomplish full separation of the two strains if applied individually.

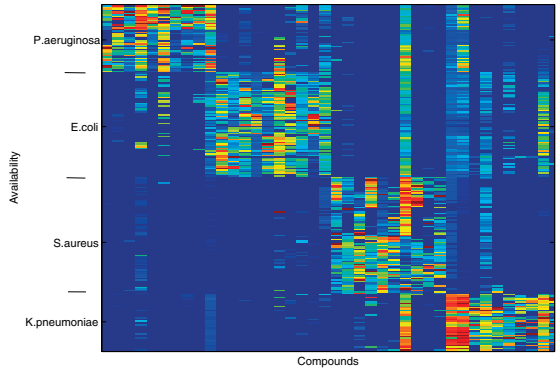

Figure 3.3: Heatmap demonstrating significantly different compounds and their availability. Compounds were selected as comparing cultures to one another. Compound numbers denote identification of compounds as presented in tables 3.2-3.5.

Support vector machine analysis

For all four approaches (medium vs cultures, culture vs remaining cultures, *E. coli* cross-culture analysis, *S. aureus* cross-culture analysis) all 912 compounds were used in the SVM analysis in order to deduce the optimal SVM model. Regarding all approaches an SVM model was build and performance was tested with 10-times cross validation. The SVM models demonstrating optimal classification of discrimination between medium and culture samples based on an increasing number of VOCs implemented is shown in table 3.11. This table shows the correct classification fraction related to the number of VOCs implemented into the model. The optimal model was based on 5 VOCs only misclassifying 2

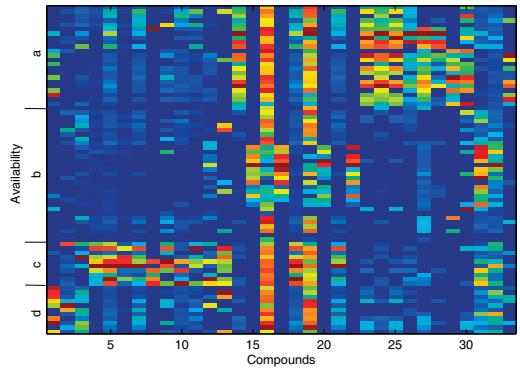

Figure 3.4: Heatmap demonstrating significantly different compounds and their availability. Compounds were selected as comparing *E. coli* cultures to one another. a) *E. coli* ATCC 35218, b) *E. coli* ATCC 25922, c) *E. coli* 47.4.1.039b and d) *E. coli* 34.4.2.038. Compound numbers denote identification of compounds as presented in table 3.6-3.9.

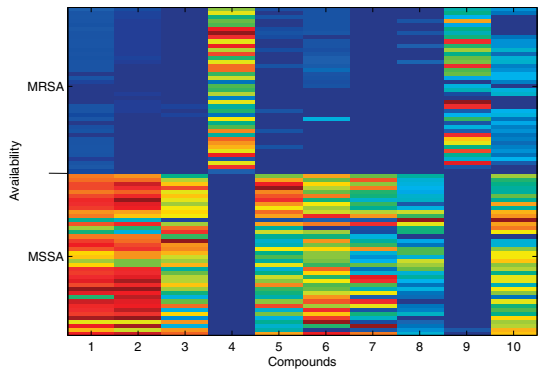

Figure 3.5: Heatmap demonstrating significantly different compounds and their availability. Compounds were selected as comparing *S. aureus* cultures to one another. Compound numbers denote identification of compounds as presented in table 3.10.

samples. Addition of more VOCs to the model did not improve performance. By removing the least valuable VOC of these 5 selected VOCs results in a profile of 4 VOCs; the optimal SVM model based on these 4 VOCs classified 97% of samples correctly, misclassifying 10 samples. Again after removing the least valuable remaining VOC an SVM model based on three VOCs classified correctly 97% of samples, the SVM based on two VOCs correctly classified 97% and the SVM model based on only one VOC correctly classified 94% of samples. By looking at table 3.1 the significantly different compounds between samples of headspace from medium to samples of headspace from all bacterial cultures three compounds have been implemented in the optimal SVM model: 2-methyl-3-hexanol, 2-methyl-butanal and benzaldehyde. The remaining compounds do not demonstrate a high degree of significance but do add valuable information regarding the classification. This due to the fact that selection of the VOCs to be implemented into the SVM model was based on adding information while lowering redundancy of the information therefore not only highly significant VOCs are implemented since these will apparently provide more redundant information.
In case of classification of the cultures against one another, we found 6 VOCs implemented into the optimal SVM model classifying 96% of samples correctly, based on 5 VOCs the model correctly classified 94% of samples. Degree of correct classification went below 90% with an SVM model based on 4 VOCs. The most significantly different VOC comparing *P. aeruginosa* to the remaining cultures proved to be trimethoxy-methane and like demonstrated in table 3.12 this compound is able to correctly classify 82% of *P.aeruginosa* samples. Also 2-methyl-1-dodecanol is implemented into the most optimal performing SVM model while also demonstrating highly significant differences in availability comparing *K. pneumoniae* and *P. aeruginosa* to all remaining cultures as shown in tables 3.4 and 3.5.
The optimal SVM model classifying the strains of the *E. coli* samples correctly did so with a 92% correct classification. Only 4 VOCs were implemented in this model as shown in table 3.13. Two out of four compounds implemented into this best performing SVM model were also highly significantly different in the *E. coli* strains. 2-Methyl-undecanal proved the most significant VOC in both *E. coli* ATCC 25922 and *E. coli* ATCC 35218 as compared to all other strains (tables 3.8 and 3.9) and 1-methyl-1-ethoxy-cyclobutane proved highly significant in comparing *E. coli* ATCC 35218 to all remaining *E. coli* cultures. The two strains of *S. aureus* only one compound was selected enabling 100% correct classification of the 81 samples; 3-methyl-2-butanal. This also demonstrated to be the most significantly different compound as shown in table 3.10. This highly significantly different compound demonstrated a higher degree of availability in all MSSA samples compared to MRSA samples and might be used as an individual biomarker. This compound has not been described in literature as related to MSSA.

DISCUSSION

In the present study headspace was analyzed from cultures from 4 different bacterial species with use of GC-TOF-MS. Extensive tests were previously performed to ensure robustness and reliability of the standardized measurement protocol.[22] Different bacterial cultures were grown in culture flasks and headspace was collected and analyzed by means of GC-MS. Over 300 compounds were detected in each sample and after combining all analysis results (*E. coli* (n=75), *P. aeruginosa* (n=52), *S. aureus* (n=81), *K. pneumoniae* (n=40) and samples containing only medium (n=70)) a total of 912 compounds were implemented into the database. These compounds demonstrated an availability of at least 8% in all samples.[68] We were able to identify a large number of compounds demonstrating a highly significant difference in availability between bacterial cultures compared to medium and compared to one another. We have also determined the highly significant compounds between the four *E. coli* strains. The number of significantly different compounds between the four *E. coli* strains was lower compared to the number of compounds demonstrating significant differences between cultures of different microorganisms thus demonstrating the likeness of the VOC profile of the four *E. coli* strains.
When comparing the samples containing only medium to all bacterial culture samples the compounds with the highest degree of differences between the two groups were those present in headspace from samples containing only medium and absent in headspace from cultures. Most likely these compounds are derived from medium and metabolized by different bacteria. Nonetheless origin of the identified VOCs remains elusive because it shows that bacteria may give rise to changes in the profile of VOCs as metabolic products are degraded or formed as described by Syhre et al. who already monitored bacterial headspace changes.[75] A few identified compounds are already mentioned in literature as VOCs found in headspace of bacterial cultures. For instance organosulfur compounds like dimethylsulfide and methanethiol have been shown to be available in *E. coli* headspace as is confirmed by our analysis. Indole is mentioned as a metabolic byproduct in *E. coli*[76] and proved to be the most significantly different compound in headspace comparing all *E. coli* strains combined to all other bacterial headspaces in our samples.
Our study demonstrates that an SVM model is able to correctly classify the different cultures with a high degree of sensitivity and specificity even differentiating between the four different strains of *E. coli* using a very small number of compounds. Most of the compounds implemented into the model demonstrated a high degree of significant difference in availability and compounds appeared to be mainly long chain hydrocarbons.
One of the innovative steps of our approach compared to other studies in this field is that we use the raw mass spectra to find matching compounds in all subjects instead of searching for matching compounds based on their chemical identification. This results in a far more precise and accurate dataset since chemical identification remains elusive and numerous misidentifications result in attenuating the discriminatory power of the analysis. Introducing

the match factor in order to directly match compounds from different samples based on their mass spectra increases accuracy. This is due to the fact that compounds are compared to one another as measured with the same experimental setup thus eliminating the differences between library mass spectra and measured mass spectra arising from use of different setups. As a result a superior database can be formed. Nonetheless, final identification of the compounds is opportune, though it remains difficult and in a few cases uncertain.
One of the disadvantages of our methodology is the use of GC-MS and the fact that it is not suitable to do temporal measurements. Therefore we made sure all analyzed cultures were sampled at the same time after the addition of the cultures to the medium. We only performed one measurement on each sample. An alternative to this might be the use of proton transfer reaction mass spectrometry (PTR-MS) that supports continuous measurements but this methodology is not able to identify compounds with a high degree of accuracy since PTR-MS cannot differentiate between compounds with the same molecular weight and does not provide identification by means of mass spectra. On top of that fragmentation and clustering of product ions further complicates the qualitative interpretation of the mass spectra and require detailed correlation analyses of the observed signals in PTR-MS.[77] Since we wanted to map and identify the discriminating VOCs from headspace with a high degree of accuracy GC-TOF-MS was used.
We are aware that use of different growth media gives rise to different VOC profiles in bacterial headspace. We chose not to use and test for different media in order to achieve a high degree of standardization. This results in a more homogeneous database of detected VOCs thus increasing the power of our statistical analysis to extract those VOCs that actually discriminate between different microorganisms.
Our main goal was to employ a robust methodology as a whole not influenced by too many confounding factors. We recently demonstrated our chemical analysis, data handling and accurate data mining provide a highly reproducible methodology.[22]
Now that interesting VOCs have been selected and identified new advances in diagnostic tools like the electronic nose[55,78] might be designed to be specifically tuned to detect the aforementioned VOCs in order to provide a fast, non-invasive, cost effective and sensitive diagnostic technique to be used for the detection of described (and likely other) microorganisms. Without these specifically tuned sensors however the presented GC-MS methodology might be of great value since analysis times are fast and results can be obtained within short time additionally it is highly cost-effective while demonstrating a high degree of sensitivity and specificity.

ACKNOWLEDGEMENTS

The authors acknowledge the province of Limburg (The Netherlands) for the financial support.

Table 3.1: Significant VOCs present in headspace from medium compared to all bacterial cultures.

Chemical structure	Medium n=70 Mean±SE	Cultures n=248 Mean±SE
2-ethyl-hexanal	276.77 ± 17.04	32.51 ± 2.75
benzaldehyde	310.85 ± 21.92	18.55 ± 3.29
1-penten-3-ol	32.03 ± 2.61	2.04 ± 0.29
2-pentanal	24.36 ± 1.94	1.81 ± 0.42
5-nonanone	737.67 ± 51.29	162.25 ± 13.57
2-methyl-butanal	12.23 ± 1.48	0.27 ± 0.17
3-cyclohexen-1-ol	7.88 ± 1.18	0.1 ± 0.05
unknown	18.36 ± 2.86	0 ± 0
unknown	7.26 ± 0.96	0.13 ± 0.04
2-methyl-3-hexanol	20.3 ± 3.24	0.24 ± 0.16

Table 3.1: Ten most significantly different compounds present in headspace from medium compared to all bacterial cultures. Values are normalized availability of a compound in the samples.

Table 3.2: Significant VOCs present in headspace from *S. aureus* compared to all other cultures.

Chemical structure	*S. aureus* n=81 Mean±SE	Cultures n=167 Mean±SE
3-methyl-2-butenal	9.88 ± 0.94	1.88 ± 0.3
2-methylamino-1-phenyl-2-propanone	3.73 ± 0.73	0.25 ± 0.07
unknown	0.66 ± 0.1	0.05 ± 0.01
unknown	4.22 ± 0.62	0.56 ± 0.13
4-methyl-cyclopentadecanone	0.35 ± 0.06	0.05 ± 0.01
unknown	8.49 ± 0.35	1.29 ± 0.25
5-methoxy-1-pentene	4.22 ± 0.48	0.53 ± 0.12
dimethyltrisulfide	102.18 ± 7.68	21.77 ± 3.15
2-methyl-butanal	74.27 ± 13.51	7.82 ± 2.3
1-methylcyclopropane-methanol	1.04 ± 0.26	0.03 ± 0.02

Table 3.2: Ten most significantly different compounds present in headspace from *S. aureus* compared to all other cultures. Values are normalized availability of a compound in the samples.

Table 3.3: Significant VOCs present in headspace from *E.coli* compared to all other cultures.

	E. coli n=75	Cultures n=173
Chemical structure	Mean±SE	Mean±SE
indole	196.3 ± 11.43	0.68 ± 0.21
3-pentanol	291.01 ± 18.89	0.03 ± 0.03
methylpropyl-disulfide	8.04 ± 0.55	0.25 ± 0.06
propylester-propanoic acid	23.72 ± 2.07	0.01 ± 0.01
unknown	0.74 ± 0.07	0 ± 0
2-ethyl-furan	19.85 ± 0.92	2.8 ± 0.59
unknown	52.17 ± 4.6	1.82 ± 0.55
nonanal	39.05 ± 2.11	6.1 ± 1.14
ethylester-propanoic acid	20.1 ± 2.12	0.13 ± 0.05
1-ethyl-3-methyl-benzene	28.3 ± 1.92	5.58 ± 0.68

Table 3.3: Ten most significantly different compounds present in headspace from *E.coli* compared to all other cultures. Values are normalized availability of a compound in the samples.

Table 3.4: Significant VOCs present in headspace from *P. aeruginosa* compared to all other cultures.

	P. aeruginosa n=52	Cultures n=196
Chemical structure	Mean±SE	Mean±SE
trimethoxy-methane	22.32 ± 2.31	0.01 ± 0.02
2-methyl-1-dodecanol	46.28 ± 5.09	0.24 ± 0.18
2-heptene	0.62 ± 0.07	0 ± 0
2-methyl-undecanal	84.61 ± 5.97	10.04 ± 2.78
3,3-diethoxy-1-propene	45.58 ± 5.42	0.01 ± 0.02
2,2-dimethyl-pentane	5.95 ± 0.41	0.51 ± 0.23
2-methyl-butanoic acid	4.44 ± 0.56	0 ± 0
2-methylthio-propane	67.03 ± 9.11	0.17 ± 0.17
3-methyl-formate-1-butanol	5.85 ± 0.68	0.21 ± 0.18
1,3-pentadiene	209.58 ± 13.39	39.63 ± 6.91

Table 3.4: Ten most significantly different compounds present in headspace from *P. aeruginosa* compared to all other cultures. Values are normalized availability of a compound in the samples.

Table 3.5: Significant VOCs present in headspace from *K. pneumoniae* compared to all other cultures.

Chemical structure	*K. pneumoniae* n=40 Mean±SE	Cultures n=203 Mean±SE
2-methyl-1-butanol	121.71 ± 21.88	21.12 ± 7.62
2-pentene	234.43 ± 9.43	63.65 ± 5.43
2-methyl-1-dodecanol	0.37 ± 0.05	0 ± 0
unknown	239.26 ± 11.46	65.12 ± 4.27
unknown	0.67 ± 0.09	0.02 ± 0.01
unknown	18.98 ± 1.86	2.56 ± 0.38
1-octen-3-one	3.19 ± 0.51	0.02 ± 0.02
3,5-dimethyl-1-Hexene	0.26 ± 0.05	0 ± 0
1,2 dimethyl-cyclopropane	69.67 ± 3.24	15.75 ± 1.73
1-hepten-3-one	0.63 ± 0.11	0.01 ± 0.01

Table 3.5: Ten most significantly different compounds present in headspace from *K. pneumoniae* compared to all other cultures. Values are normalized availability of a compound in the samples.

Table 3.6: Significant VOCs present in headspace from *E. coli* 34.4.2.038 compared to all other *E.coli* cultures.

Chemical structure	*E. coli* 34.4.2.038 n=11 Mean±SE	*E. coli* n=64 Mean±SE
dimethylsulfide	19.95 ± 4.1	2.48 ± 0.8
unknown	0.34 ± 0.11	0.07 ± 0.02
5-undecene	2 ± 0.26	0.64 ± 0.13

Table 3.6: Three most significantly different compounds present in headspace from *E. coli* 34.4.2.038 compared to all other *E. coli* cultures. Values are normalized availability of a compound in the samples.

Table 3.7: Significant VOCs present in headspace from *E. coli* 47.4.1.039b compared to all other *E. coli* cultures.

Chemical structure	*E. coli* 47.4.1.039b n=10 Mean±SE	*E. coli* n=65 Mean±SE
cyclodecane	26 ± 4.24	0.86 ± 0.17
tridecanone	69.8 ± 7.47	13.47 ± 1.51
2-methyl-1-hexadecanal	7.91 ± 1.82	0.12 ± 0.03
2-heptadecanone	16.74 ± 3.27	4.52 ± 0.6
3-methyl formate-1-butanol	3.18 ± 1.1	0.13 ± 0.13
3,7,11-trimethyl-1-dodecanol	1.03 ± 0.28	0.22 ± 0.04
unknown	1.1 ± 0.34	0.25 ± 0.04
unknown	3.37 ± 0.99	0.64 ± 0.14
unknown	0.77 ± 0.21	0.17 ± 0.04
2-methyl-1-hexene	4.85 ± 0.89	1.01 ± 0.29

Table 3.7: Ten most significantly different compounds present in headspace from *E. coli* 47.4.1.039b compared to all other *E. coli* cultures. Values are normalized availability of a compound in the samples.

Table 3.8: Significant VOCs present in headspace from *E. coli* ATCC 25922 compared to all other *E. coli* cultures.

	E. coli ATCC 25922 n=30	*E. coli* n=45
Chemical structure	Mean±SE	Mean±SE
2-methyl-undecanal	3.17 ± 1.54	75.93 ± 10.32
eicosyl-benzene	1.9 ± 0.42	0 ± 0
2-methyl-1-butanol	213.25 ± 20.43	314.54 ± 7.58
hexadecyl-benzene	0.38 ± 0.09	0 ± 0
tridecanone	5.94 ± 1.05	31.01 ± 3.88
methyl-pyrazine	11.69 ± 1.62	19.84 ± 0.71
1,4-dichloro-benzene	5.45 ± 1.34	0 ± 0
2-pentadecanone	1.86 ± 0.52	9.01 ± 1.14
unknown	1.15 ± 0.29	0 ± 0
2-methyl-dodecanal	0.43 ± 0.21	4.51 ± 0.67

Table 3.8: Ten most significantly different compounds present in headspace from *E. coli* ATCC 25922 compared to all other *E. coli* cultures. Values are normalized availability of a compound in the samples.

Table 3.9: Significant VOCs present in headspace from *E. coli* ATCC 35218 compared to all other E.coli cultures.

	E. coli ATCC 35218 n=24	*E. coli* n=51
Chemical structure	Mean±SE	Mean±SE
2-methyl-undecanal	129.63 ± 10.12	7.86 ± 1.95
unknown	7.12 ± 0.94	0.89 ± 0.22
2-methyl-dodecanal	41.89 ± 6.17	5.39 ± 1.28
1-methyl-1-ethoxycyclobutane	88.07 ± 11.59	23.77 ± 3.29
2-heptanone	15.97 ± 3.55	0.89 ± 0.33
2-octanone	0.3 ± 0.07	0.02 ± 0.02
unknown	1.39 ± 0.21	0.33 ± 0.09
unknown	6.64 ± 1.52	25.37 ± 2.59
ethylester-propanoic acid	6.23 ± 0.69	13.92 ± 1.22
unknown	0.37 ± 0.1	0.07 ± 0.02

Table 3.9: Ten most significantly different compounds present in headspace from *E. coli* ATCC 35218 compared to all other *E. coli* cultures. Values are normalized availability of a compound in the samples.

Table 3.10: Significant VOCs present in headspace from two different *S. aureus* strains.

Chemical structure	MSSA n=40 Mean±SE	MRSA n=41 Mean±SE
3-methyl-2-butanal	1121.81 ± 31.88	133.4 ± 8.97
unknown	281.21 ± 9.7	4.8 ± 2.33
2-pentene	69.67 ± 3.24	0.21 ± 0.97
3-octyn-1-ol	0 ± 0	74.66 ± 5.09
1,2-pentadiene	41.17 ± 2.19	1.65 ± 1.73
3-methylene-heptane	21.17 ± 1.15	2.4 ± 1.7
2-tridecanone	13.65 ± 1.14	0 ± 0
dodecanal	2.79 ± 0.23	0.13 ± 0.64
3-nonynoic acid	0 ± 0	0.49 ± 0.53
unknown	239.26 ± 11.46	88.69 ± 7.52

Table 3.10: Ten most significantly different compounds present in headspace from two different *S. aureus* strains. Values are normalized availability of a compound in the samples.

Table 3.11: SVM performance on medium compared to cultures.

Number of compounds and identification					% Correct
			Medium	Cultures	
5	2-methyl-3-hexanol	Medium	69	1	99
		Cultures	1	247	
4	5-nonylamine	Medium	61	9	97
		Cultures	1	247	
3	2-methyl-butanal	Medium	64	6	97
		Cultures	2	246	
2	benzaldehyde	Medium	63	7	97
		Cultures	3	245	
1	4-amino-heptane	Medium	54	16	94
		Cultures	3	245	

Table 3.11: SVM Performance on medium compared to cultures based on 10-times cross-validation. Providing the confusion matrix and percentage correctly classified instances. Each row of the confusion matrix represents the instances in the predicted class, while each column represents the instances in the actual class.

Table 3.12: SVM performance on four cultures.

Number of compounds and identification		Confusion matrix				% Correct
		P. aeruginosea	*E. coli*	*S. aureus*	*K. pneumoniae*	
6	2-methyl-	48	3	1	0	96
	1-dodecanol	0	75	0	0	
		0	0	77	4	
		0	0	2	38	
5	1-propanol	42	6	4	0	94
		0	75	0	0	
		0	0	77	4	
		0	0	2	38	
4	trimethylene-	42	6	4	0	87
	oxide	0	74	0	1	
		0	0	76	5	
		0	0	16	24	
3	2-methyl-	42	6	4	0	79
	2-butene	0	75	0	0	
		0	0	81	0	
		0	0	40	0	
2	unknown	43	0	9	0	78
		0	69	6	0	
		0	0	81	0	
		0	0	40	0	
1	trimethoxy-	43	9	0	0	48
	methane	0	75	0	0	
		0	81	0	0	
		0	40	0	0	

Table 3.12: SVM Performance on four cultures based on 10-times cross-validation. Providing the confusion matrix and percentage correctly classified instances. Each row of the confusion matrix represents the instances in the predicted class, while each column represents the instances in the actual class.

Table 3.13: SVM performance on four *E.coli* species.

No. of compounds and identification		Confusion matrix				% Correct
		E. coli				
		34.4.2.038	47.4.1.039b	25922	35218	
4	1-decanol	8	0	3	0	92
		0	9	1	0	
		0	0	30	0	
		1	0	1	22	
3	2-methyl-	8	0	3	0	88
	undecanal	0	6	4	0	
		0	0	30	0	
		1	0	1	22	
2	pentacosanone	7	0	4	0	91
		0	8	2	0	
		0	0	30	0	
		0	0	1	23	
1	1-methyl-	7	0	3	1	51
	1-ethoxy-	0	0	10	0	
	cyclobutane	0	0	30	0	
		4	0	19	1	

Table 3.13: SVM Performance on four *E. coli* species based on 10-times cross-validation. Providing the confusion matrix and percentage correctly classified instances. Each row of the confusion matrix represents the instances in the predicted class, while each column represents the instances in the actual class.

CHAPTER 4

A profile of volatile organic compounds in breath discriminates COPD patients from controls

Van Berkel JJBN, Dallinga JW, Möller GM, Godschalk RWL, Moonen EJ, Wouters EFM, Van Schooten FJ
Respir Med, 104(4):55763, 2010

ABSTRACT

Background

Chronic obstructive pulmonary disease (COPD) is an inflammatory condition characterized by oxidative stress and the formation of volatile organic compounds (VOCs) secreted via the lungs. We recently developed a methodological approach able to identify profiles of VOCs in breath unique for patient groups. Here we demonstrate this methodology also identifies COPD patients.

Methods

Fifty COPD patients en 29 controls provided their breath and VOCs were analyzed by gas chromatography-mass spectrometry to identify relevant VOCs. An additional 16 COPD patients en 16 controls were sampled in order to validate the model, and 15 steroid naive COPD patients were sampled to determine whether steroid use affects performance.

Findings

1179 different VOCs were detected, of which thirteen were sufficient to correctly classify all 79 subjects. Six of these 13 VOCs classified 92% of the subjects correctly (sensitivity: 98%, specificity: 88%) and correctly classified 29 of 32 subjects (sensitivity: 100%, specificity: 81%) from the independent validation population. Fourteen out of 15 steroid naive COPD patients were correctly classified thus excluding treatment influences.

Interpretation

This is the first study distinguishing COPD subjects from controls solely based on the presence of VOCs in breath. Analysis of VOCs might be highly relevant for diagnosis of COPD.

INTRODUCTION

In the last few decades the mortality of chronic obstructive pulmonary disease (COPD) has increased worldwide, even in industrialized countries. The main feature of COPD is irreversible airflow limitation as a result of emphysematous destruction, increasing both compliance of the lung and the resistance of the small airways. COPD is a leading cause of mortality and morbidity, and estimates from the World Health Organization state that in 2001 COPD was the fifth leading cause of death in high-income countries and the sixth leading cause of death in low and middle income countries.[32] Early diagnosis and treatment will be necessary in order to control its high morbidity and mortality and consequently high healthcare cost.[79] Spirometry is currently the gold standard for diagnosing and monitoring progression of COPD. However in order to ensure quality, the practice nurse has to be trained on how to perform spirometry and general practitioners on how to evaluate spirograms.[80] Only under these prerequisites, office spirometry can help identify the presence of asthma and COPD if breathing symptoms are present. In the early stage of COPD, breathing symptoms might not be clinically manifest. Especially identification of these early stages of COPD by using VOC analysis in exhaled air might prove clinically relevant. Our proposed methodology indeed detected inflammatory related compounds in exhaled air, in contrast to early detection of COPD with the current gold standard; spirometry.

A relatively new concept, the analysis of exhaled air, might provide more accurate diagnosis and might prove useful as a new non-invasive, safe and fast diagnostic tool regarding the diagnosis of inflammatory lung diseases, including asthma, cystic fibrosis and COPD. Regarding the analysis of exhaled air several biomarkers for several diseases have already been identified. Nitric oxide (NO) levels are generally accepted as an indication of inflammation and oxidative stress in the respiratory tract in for instance asthma.[81] However in COPD the use of NO is limited since exhaled NO levels are not or only marginally elevated in COPD patients.[12] Additionally carbon monoxide (CO) has been investigated as a COPD biomarker, and likewise with NO, contrasting results are published. Yamaya et al. found a significant relationship between exhaled CO concentrations and FEV_1, and exhaled CO appeared to correlate with the eosinophil count in sputum.[8] Others found no correlation of exhaled CO with lung function.[9] The application of CO as a diagnostic marker is also limited because exhaled CO levels are affected by environmental CO, which may fluctuate considerably and is influenced by active and passive smoking making its use as a biomarker for COPD at the least questionable.[10]

Additionally exhaled volatile compounds have been studied regarding their hypothesized function as biomarkers of oxidative stress as for example ethane. Ethane belongs to the group of volatile organic compounds (VOCs) and is demonstrated to be elevated in exhaled air of COPD patients compared to controls. A correlation was also demonstrated between levels of ethane and the degree of airway obstruction, smoking habits and FEV_1.[45] However, the analysis of single compounds from exhaled air is hampered by low sensitivity

and specificity. In 1971 Pauling et al. already demonstrated the availability of hundreds of different VOCs in exhaled air[47] and many more have been identified since. Philips et al. have recently successfully demonstrated the possibility of using a profile of VOCs in breath as biomarkers of lung cancer and pulmonary tuberculosis.[14,15,48] Therefore we are particularly interested in multi-component analysis of VOCs in COPD patients in order to increase performance using a biomarker profile approach. This approach may have great potential regarding clinical application for the assessment of airway inflammation. Especially since analysis of exhaled air provides a fast, non-invasive, cost beneficial and easy to perform diagnostic tool.
The aim of this study was to identify COPD biomarkers from exhaled air able to discriminate patients suffering from COPD from non-diseased controls. We investigated VOCs in exhaled air from COPD patients and controls by means of a recently developed approach and identified profiles that were able to identify the diseased state.[22]

MATERIALS AND METHODS

Study subjects

As a training population a total of 50 COPD patients and 29 non-diseased controls both smoking and non-smoking were recruited at Maastricht University, The Netherlands. Subject characteristics are shown in table 4.1. Diagnosis of COPD was based on examination of pulmonary function according to international guidelines. Subjects were sampled at centrally ventilated treatment rooms located in the hospital. Non-diseased control subjects were staff members. A medical interview confirmed absence of respiratory disease in these subjects.
Another set of 16 COPD patients and 16 non-diseased controls were obtained from the 'center for Integrated Rehabilitation Organ Failure' (CIRO). This population was used as a validation population to validate the results that were obtained in the training population. Additionally, a set of 15 subjects was sampled at the University Hospital Maastricht diagnosed as COPD patients, and exhaled air was obtained before therapeutic treatment with steroids was initiated. All subjects gave their informed consent and the study protocol was approved by the medical ethics committee of Maastricht University.

Sample collection and analysis

Participants were asked to inhale, hold their breath for 5 seconds and subsequently fully expire into resistance free tedlar bags (5L). The content of the bag was transported under standardized conditions onto stainless steel two-bed sorption tubes, filled with carbograph 1TD/Carbopack X (Markes International, Llantrisant, Wales, UK) that trap VOCs. These sorption tubes were placed inside a thermal desorption unit (Marks Unity desorption unit, Markes International Limited, Llantrisant, Wales, UK) and subsequently heated to

Table 4.1: Study population characteristics

Study population	Training set		Validation set		Steroid naive
	Controls (n=29)	COPD (n=50)	Controls (n=16)	COPD (n=16)	COPD (n=15)
Age (years)	50 ± 9	71 ± 8	51 ± 6	63 ± 6	57 ± 8
Sex (M/F)	14/15	38/12	8/8	11/5	8/7
FEV_1 (% predicted)	0 ± 14	50 ± 15	92 ± 7	53 ± 13	57 ± 12
RV (% predicted)	77 ± 19	176 ± 34	72 ± 13	154 ± 29	148 ± 13
Smoking(current/ex/non)	9/7/13	38/6/6	3/3/10	2/14/0	3/9/3
Packyears	18 ± 7	49 ± 12	12 ± 5	35 ± 6	29 ± 26

Table 4.1: Study population characteristics.

270°C in order to release all VOCs onto the gas chromatography capillary column (RTX-5ms, 30m x 0.25mm 5% diphenyl, 95% dimethylsiloxane capillary, film thickness 1.0μm). VOCs are separated by GC (ThermoFisher Scientific., Austin, Texas, USA) and subsequently detected by a time-of-flight mass spectrometer (TOF-MS) (Thermo Electron Tempus Plus time-of-flight mass spectrometer, ThermoFisher Scientific, Austin, Texas, USA). The temperature of the gas chromatograph was programmed as follows: 40 °C during 5 min., then raised with 10 °C/min until the final temperature of 270 °C, this temperature was maintained for 5 min. Electron ionization mode at 70 eV was used with a 5Hz scanning rate over a mass range of m/z 35-350 amu. An example breathogram is shown in Figure 4.1.

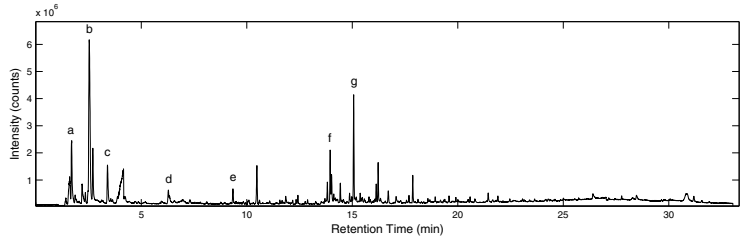

Figure 4.1: Example of a breathogram from a control subject as analyzed by GC-TOF-MS containing a few identified compounds. a) acetone; b) 1,3-pentadiene; c) pentane, 2-methyl; d) benzene; e) toluene; f) 5-hepten-2-one, 6-methyl; g) phenol. The area-under-peak is related to the concentration of the compound.

Data-acquisition

Analysis of the data output files from the GC-TOF-MS was performed in successive steps as previously described in detail.[22] In summary the first step was to perform peak detection and baseline corrections on all analysis output files. Normalization of the calculated peak areas was performed using an area scaling factor. This rescaling factor used is based on the cumulative area under

the detected peaks, since all chromatograms display rather similar profiles this method of normalization is most robust.
Next retention times (RT) of all subjects were corrected for chromatographic drifting and lined up. Applying this correction for retention times is very effective and easy to perform eliminating the use of an added internal standard, adding to the straight-forwardness and easy to perform routine of the presented methodology.
Finally, the output files were merged by combining corresponding compounds based on degree of similarity of the corresponding mass spectra - by determining the match factor values (MFs) - and similarity of RT. The degree of mass spectra similarity was calculated using a match factor based on the similarity index as described by Stein et al.[63] These match factors were only determined for compounds within a selectable RT-window.

Component selection

To determine which compounds in the database were of interest regarding the classification of diseased versus controls, we applied support vector machines (SVM). Support vector machines demonstrate the ability to construct predictive models with large generalization power even in the case of large dimensionality of the data or when the number of observations available for training is low. SVM always seeks a globally optimized solution and avoid over-fitting. This implies that a large number of features (i.e. compounds) is allowed.[69] This encouraged us to implement this subset selection algorithm into this study, since it will select the most optimal subset of compounds able to correctly classify our dataset. A variety of selection methods was tested using the software program 'Weka': a collection of machine learning algorithms for data mining tasks. Compounds were selected using an SVM attribute evaluator. The attribute evaluator we used evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the redundancy between them. Preferably features will be selected showing high correlations within the class and low inter-correlation. Next, the selected compounds were analyzed and ranked with use of SVM using recursive feature selection and removing one attribute at a time. This way attributes were selected using the weight magnitude as ranking criterion. After every run the least efficient attribute was removed. All resulting subsets were analyzed for classification performance with use of support vector classifiers based on John Platt's sequential minimal optimization algorithm.[65]

RESULTS

Discriminating VOCs as subtracted from training dataset

The resulting analysis of exhaled air based on 50 COPD patients and 29 non-diseased subjects demonstrated that the exhaled air is rich in a wide variety of VOCs. A total of 3778 different compounds were found and a mean of 332

different VOCs were detected per subject. The final COPD dataset consisted of 50 COPD patients and 29 non-diseased subjects. Compounds detected in less than 8% of subjects were discarded resulting in a dataset of 79 subjects and 1179 compounds using a match factor-threshold of 0.85 and an RT-window of 12 seconds. Used MF-threshold values were determined based on a variety of complementary compounds manually combined. The MFs calculated for these compounds demonstrated to be at least 0.852. The RT-window value was chosen based on analysis of the maximum RI-range of complementary compounds found in several subject files.
As shown in table 4.2, a classification model was constructed based on 13 VOCs. This model is able to classify both controls and COPD patients for 100% correctly, tested with 10 times cross validation. Classification performance of SVM based on more than 13 VOCs performed equal to the SVM based on this minimal amount of 13 VOCs. These 13 VOCs were identified as: isoprene, C16 hydrocarbon, 4,7-dimethyl-undecane, 2,6-dimethyl-heptane, 4-methyl-octane, hexadecane, 3,7-dimethyl, 1,3,6-octatriene, 2,4,6-trimethyl-decane, hexanal, benzonitrile, octadecane, undecane, terpineol. For one compound it was not possible to identify the compound with certainty only that it consisted of a hydrocarbon containing a 16-carbon chain. Figure 2 shows the component mean area under peak and availability of the compounds.

Table 4.2: VOC classification performance

Chemical structure	RT	Training		Validation		Steroid-naive
	(min.)	sens.	spec.	sens.	spec.	sens.
isoprene *	2.6	0.96	0.41	1	0.56	0.46
C16 hydrocarbon *	23.3	0.96	0.55	1	0.75	0.6
4,7- dimethyl undecane *	19.5	0.96	0.69	1	0.75	0.6
2,6- dimethyl heptane *	10.5	0.98	0.69	1	0.5	0.8
4-methyl octane *	11.3	0.98	0.76	1	0.69	0.87
hexadecane *	23.2	0.98	0.83	1	0.82	0.93
3,7-dimethyl 1,3,6-octatriene	14.5	1	0.83	1	0.56	0.87
2,4,6-trimethyl decane	19.5	1	0.9	0.94	0.69	0.87
hexanal	9.9	1	0.9	0.94	0.69	0.87
benzonitrile	14.1	1	0.93	0.94	0.69	0.93

Table 4.2: Classification performance tested with ten times cross-validation on COPD dataset consisting of 79 subjects (50 COPD patients, 29 controls). A Support vector classifier was used trained by the sequential minimal optimization algorithm (SMO). Correct classification was achieved with use of 13 VOCs from exhaled air. The displayed VOCs are ranked according to their cumulative contribution and performance is evaluated in a cumulative manner. VOC profiles were also tested on an independent validation set consisting of 32 subjects (16 COPD patients, 16 controls), and on steroid naive COPD patients (n=15). The compounds implemented into the optimal classification algorithm are denoted by an asterisk (*). The second column shows the retention time (RT) for every compound.

Evaluation performance of discriminating VOCs in validation datasets

The performance of the generated SVM classifier based on the 13 VOCs as shown in Table 4.2 was evaluated on the validation population, consisting of

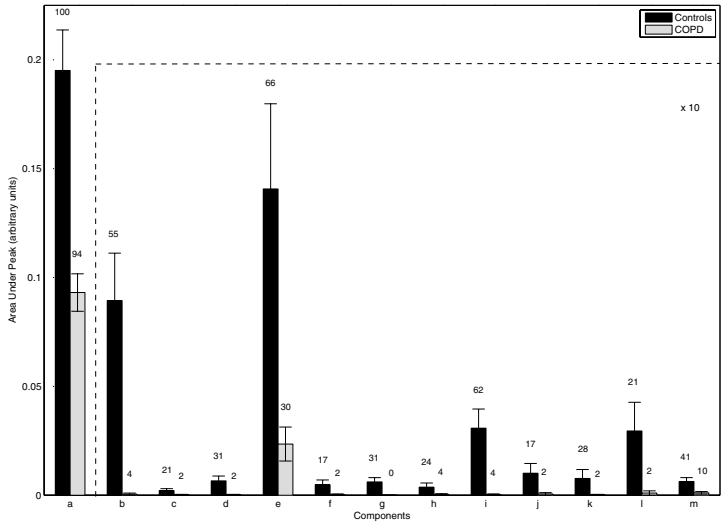

Figure 4.2: Relative amounts of compounds used in the classification algorithm. Each bar represents mean and standard error values for one compound detected in either control subjects or COPD patients. The height of every bar denotes the mean and standard errors are denoted by the errorbars. Grey bars represent patients suffering from COPD, black bars represent control subject data. a) isoprene, b) C16 hydrocarbon, c) 4,7-dimethyl undecane, d) 2,6- dimethyl heptane, e) 4-methyl octane, f) hexadecane, g) 3,7-dimethyl 1,3,6-octatriene, h) 2,4,6-trimethyl decane, i) hexanal, j) benzonitrile, k) octadecane, l) undecane, m) terpineol. The numbers over the errorbars denote the availability; the percentage of subjects in which the compound was detected.

16 COPD patients and 16 controls. The COPD patients and controls used as validation population were sampled at separate locations and another period of time compared to the training population. The results of testing the 13 VOCs on the validation set show that misclassifications occur due to overfitting of the data leading to lower levels of specificities (table 4.2). That means that the classification model based on the 13 VOCs is too specific towards the COPD training dataset and lacks performance when tested on the independent validation dataset. The logical step was to lower the number of VOCs implemented into the classification algorithm thus reducing misclassifications. As shown in table 4.2, a classifier trained on the COPD training set containing 6 VOCs performed best on the validation set. Therefore the preferred classification model is based on 6 VOCs and these are indicated by an asterisk in table 4.2. The support vector classifier build from the COPD training set based on these 6 VOCs demonstrated an optimum of 91% correct classification as tested on the 32 subject validation set. This corresponds to a misclassification of only 3 samples out of 32 demonstrating a specificity of 81% and a sensitivity of 100%. So, all COPD patients were correctly identified and only 3 healthy subjects were incorrectly seen as COPD patients. This illustrates the robustness of our approach that even when sampling is done under different circumstances the discriminating VOCs show similar levels of specificity and sensitivity in both study populations. In order to be sure that the 6 VOCs were not discriminating between controls and COPD patients because of the use of medication by COPD patients, their performance was tested on exhaled air samples obtained from 15 steroid naive COPD patients. Fourteen out of these 15 subjects suffering from COPD not using any medication were correctly classified as COPD by our optimized 6 component classifier demonstrating a sensitivity of 93

DISCUSSION

We have shown in the present paper that measuring multiple VOCs in exhaled air offers an excellent possibility as future diagnostics of COPD patients. In addition to the good performance of the VOC profile in discriminating diseased versus healthy subjects, sampling of exhaled air is noninvasive, safe and does not cause any degree of discomfort to the patients. We present a highly accurate classification model based on 6 VOCs that is not confounded by time of sampling, ambient air, use of medication and former or current tobacco smoking. Using sophisticated bioinformatics tools we extracted 6 VOCs out of nearly 1200 exhaled compounds that combined into an accurate SVM model. This model proved to be highly sensitive and specific to determine whether a person is diseased or healthy. The implemented compounds appeared to be mainly long chain hydrocarbons.

It was our aim to develop a robust diagnostic methodology that can be implemented in clinical daily practices without encountering too many confounding obstacles. In principal the excretion of VOCs can be influenced by many factors of intrinsic nature (gender, age, weight, genetic background) or of exogenous

origin (ambient air, diet or medication). It is practically impossible to control for all these potentially confounding factors. Therefore it was our goal to select those compounds that give information regarding diseased versus healthy status, independently from other endogenous or exogenous factors. We recently published our sampling procedure, chemical analysis, data handling and accurate data mining methodology[22] and extensively showed that the developed methodology is highly reproducible. Since the performance of the 6 discriminating VOCs was similar in both training populations sampled at different locations it is clear ambient air is not a major confounder. Furthermore, we studied whether the use of medication can be considered as a confounder by examining the performance of the 6 VOCs on COPD patients who received no treatment. These persons were sampled immediately before the diagnosis by lung function and the VOC profiles were again highly sensitive to diagnose the COPD patients since they were all but one recognized as diseased. However as shown in table 4.1 the steroid naive validation group consisted of subjects mainly in GOLD classes I and II, while the training group consisted of subjects mainly in GOLD classes II and III. This due to the low availability of steroid naive COPD patients. Nonetheless the SVM model based on 6 VOCs was able to correctly classify 14 out of 15 COPD subjects.
As can be concluded from table 4.1 controls and COPD subjects are not optimally matched regarding age. This difference however does not demonstrate to be of great importance since the degree of overlap between controls and COPD in the training set with regards to age is large but due to a few very old COPD patients included age seems to be poorly matched. According to the results from the validation set and steroid-naive set - where subjects are far better matched for age - we can clearly state that the selected compounds do relate to disease status instead of age or smoking behavior.
In the field of exhaled air analyses, there is an ongoing discussion whether or not background measurements should be taken into account in order to correct for ambient air influences. In the presented work, no background extraction was applied to keep the methodology as easy and straightforward as possible. Moreover, Miekisch et al. mentioned that it will not be possible to correct for the complex interdependencies between excretion and uptake of VOCs by easily subtracting the peak areas obtained from inhaled and exhaled air.[3] Additionally background noise is expected to be randomly distributed between subjects' samples and would thus not exert any discriminatory power, nor interfere with the outcome of the analyses.
Another point of discussion is the difference between sampling alveolar and dead space air. Our sampling approach samples a mixture of alveolar air and dead space air, which is about 150 ml during tidal breathing. Participants were asked to inhale, hold their breath for 5 seconds and subsequently fully expire. Since most of the subjects were able to fully inflate the Tedlar bag in 2 to 4 expirations the contribution of dead space air to the total volume is indeed significant but nonetheless most of the obtained volume originates from alveolar air. From this mixture one can argue that alveolar air is diluted with dead space air which could lead to sensitivity problems. However, in our analysis no

sensitivity problems occurred. Although there are sampling methods available of preventing dead space air sampling, we decided not to use these methods to maintain the ease of sampling. Furthermore, during previous studies we explored the effect of different exhalation patterns on VOC profiles and found no differences in breathograms between superficial exhalation or deep exhalation.[22]

One of the innovative steps of our approach compared to other studies in this field is that we use the raw mass spectra to find the matching compounds in all subjects instead of searching for matching compounds based on their chemical identification. The latter procedure can introduce many mistakes that jeopardize the quality of databases, which attenuates the discriminative power of the analysis (garbage in - garbage out). We introduced the match factor in our routine to determine the degree of similarity between mass spectra. The main advantage is that the mass spectra are compared with each other as measured with the same instrumental setup instead of comparison against mass spectra found in a library. The experience is that most of the time low match factors are found when measured mass spectra are compared to those present in the library. Therefore a far superior database is created when raw mass spectra are compared instead of comparing compounds after improper identification.

The amounts in exhaled air of the compounds that performed best in the classification model are predominantly lower in COPD patients than in controls. An explanation for this observation may be found in the complicated biological equilibrium of formation and removal of VOCs in the human body. The hypothesis is that the inflammation driven oxidative stress is responsible for oxidizing macromolecules, including polyunsaturated fatty acids that are abundantly present in membranes, leading to a series of breakdown products excreted as VOCs. Thus, the relative composition of VOCs in exhaled breath of COPD patients can change as a result of the disease, and this change can be either an increase or a decrease of certain compounds. Indeed we observed a considerable number of compounds that are significantly higher in COPD patients compared to controls but these compounds are not implemented into the optimized classification model due to their limited discriminatory power. On the other hand a decline in certain VOCs may occur since especially the longer chain hydrocarbons are further oxidized into smaller compounds due to enhanced oxidative stress and consequently their amounts are decreased in exhaled air of COPD patients. It appeared that the absence of a number of these long chained VOCs is crucial for the diagnostic ability. Apart from the oxidative stress hypothesis explaining changing VOC composition in COPD patients, an alternative reason may be that lungs are remodeled during COPD resulting in altered gas exchange over the blood lung barrier. Further studies are necessary in clinical settings but also in inflammation models to explain the biochemical origin, the physiological meaning and exhalation kinetics of our selected VOCs. Nevertheless without this mechanistic knowledge the compounds may already be of value to base a diagnostic tool for clinical settings.

Finally, we conclude that analysis of a profile of 6 identified VOCs in exhaled air provides an accurate, non-invasive, easy to perform diagnostic tool in the

diagnosis of COPD patients. Recent developments on real time measurements can bring early-stage detection of COPD into the general practice in the near future.

ACKNOWLEDGEMENTS

The authors acknowledge the province of Limburg (The Netherlands) for the financial support of this research performed at the Maastricht University.

CHAPTER 5

Metabolomics of volatile organic compounds in cystic fibrosis patients and controls

Robroeks CMHHT, Van Berkel JJBN, Dallinga JW, Jöbsis Q, LJ Zimmerman, HJ Hendriks, Wouters EFM, Van der Grienten CP, Van de Kant KD, Van Schooten FJ, Dompeling E
Pediatr Res, 2010

ABSTRACT

In cystic fibrosis (CF), airway inflammation causes an increased production of reactive oxygen species, responsible for degradation of cell membranes. During this process, volatile organic compounds (VOCs) are formed. Measurement of VOCs in exhaled breath of CF patients may be useful for the assessment of airway inflammation. This study investigates whether metabolomics of VOCs could discriminate between CF and controls, and between CF patients with and without *Pseudomonas aeruginosa* (*P. aeruginosa*) colonization. 105 Children (48 CF, 57 controls) were included in this study. After exhaled breath collection, samples were transferred onto tubes containing active carbon in order to adsorb and stabilize VOCs. Samples were analyzed by gas chromatography-time of flight-mass spectrometry to assess VOC profiles. Analysis showed that 1099 VOCs had a prevalence of at least 7%. By using 22 VOCs, a 100% correct identification of CF patients and controls was possible. With 10 VOCs, 92% of the subjects were correctly classified. The reproducibility of VOC measurements with a one-hour interval was very good (matchfactor 0.90±0.038). We conclude that metabolomics of VOCs in exhaled breath was possible in a reproducible way. This new technique was not only able to discriminate between CF patients and controls, but also between CF patients with or without *P. aeruginosa* colonization.

INTRODUCTION

Airway inflammation plays a central role in the pathophysiology of various chronic lung diseases, such as cystic fibrosis (CF), asthma, and chronic obstructive pulmonary disease (COPD). Diagnosing CF is possible at any age, but it is far less simple to diagnose early pulmonary disease in CF patients, particularly in young children. It is known there is a poor correlation between airway inflammatory processes and respiratory symptoms.[82] Airway infection and inflammation can be present before clinical symptoms are evident. Therefore monitoring of airway inflammation and the related oxidative stress in CF may be useful in clinical practice. There has been an increasing interest in non-invasive assessment of airway inflammation and oxidative stress in chronic lung disease. The collection of broncho-alveolar lavage fluid or lung biopsies is invasive, and therefore cannot be applied very easily in children. Non-invasive techniques include measurement of non-volatile inflammatory markers in exhaled breath condensate, and measurement of volatile inflammatory markers in exhaled breath. Fractional exhaled nitric oxide (FeNO) , carbon monoxide (CO), ethane and pentane are the most studied volatile markers,[10,45] of which FeNO is most standardized. In contrast to assessments of pre-selected inflammatory markers, it is possible to assess profiles of volatile organic compounds (VOCs) in exhaled air. An increased production of reactive oxygen species (ROS) is caused by the influx of leukocytes, continuously producing

ROS, leading to an imbalance between oxidants and antioxidants (oxidative stress).[20,83–85] These generated ROS are able to degrade cell membranes in a process called lipid peroxidation. During this process, VOCs are formed as a result of the degradation of polyunsaturated fatty acids. Due to the very low solubility of VOCs in blood, after generation these VOCs are transported through the bloodstream and become available in the exhaled air as soon as the blood containing the VOCs passes the blood-lung barrier. Thus analysis of exhaled air might provide the means to monitor or diagnose inflammatory diseases. One possibility of analysis of VOCs in exhaled air is enabled by using a gas chromatograph-mass spectrometer (GCMS). This technique is highly sensitive and capable of detecting a wide range of VOCs. Other research proved a predictive model employing nine VOCs was sufficiently sensitive and specific to be considered as a screening tool for lung cancer.[48]
The aim of this study was to investigate whether metabolomics of VOCs in exhaled breath was able to discriminate between young subjects with CF and healthy controls, and between subgroups of CF patients (with or without *P.aeruginasa* colonization). It is well known that *P. aeruginosa* colonization of the lung is associated with a less favorable prognosis of CF. The reproducibility of the method, as well as the nature and background of the most discriminating compounds were studied.

Methods and materials

Study subjects

One hundred and five Subjects aged 5-25 years were included in this cross-sectional study: 48 subjects were diagnosed with CF, and 57 control persons. The CF population was recruited from the outpatient clinic of the University Hospital Maastricht, the Netherlands. CF was defined as a combination of typical clinical features and an abnormal sweat test (Chloride >60 mmol/L). Children with CF and/or their parents completed the Shwachman-Kulczycki questionnaire, enabling classification of the CF disease status.[86] The control group consisted of children without any (history of) respiratory problems, as confirmed by the ISAAC questionnaire. Thirty-six (63%) of 57 control children were recruited from primary schools, whereas 21 (37%) subjects were included at the outpatient clinic where enuresis nocturna and constipation were the initial reasons for consultation. At the time of this study, these children were all stable with no signs of infection or any somatic disturbance. We checked the homogeneity of the control group by comparing children recruited from primary schools and outpatient clinic and found no significant differences in VOC patterns between these children (One-way ANOVA tests, p-values¿0.108). Exclusion criteria for both CF and control children were: 1) Patients with mental retardation; 2) Active smokers; 3) (Congenital) heart disease; 4) Technical inability to perform the measurements; 5) Patients with an acute respiratory infection. Informed consent to participate was obtained from the parents of all

children participating in this study. The Ethics Committee of the Maastricht University approved this study. The clinical trial registration number is NCT 00413140. Table 5.1 shows the study population characteristics.

Table 5.1: Study population characteristics

	Cystic Fibrosis n=48	Control n=57
Age (years)	13.0 ± 0.6	$*9.9 \pm 0.4$
Height (cm)	148.3 ± 2.9	141.5 ± 2.3
Weight (kg)	38.6 ± 2.1	34.8 ± 1.5
Sex (M/F)	27/21	29/28
FEV_1 (% predicted)	75.5 ± 3.8	$*101.4 \pm 1.5$
FEV_1/VC (%)	76.7 ± 1.8	$*88.1 \pm 0.9$

Table 5.1: Abbreviations: M, male; F, female; FEV_1, forced expiratory volume in 1 second; VC, vital capacity; Data are given as mean±standard error (SE); * $P<0.05$, cystic fibrosis vs control.

Sample collection and analysis

Subjects were asked to exhale into a resistance free polycarbonate (plastic) bag (Tedlar bag, SKC ltd., Dorset, UK). Severe physical exercise before the test was not permitted. At least three exhalations were necessary to fill the 5-liter bag. Within one hour after collection of the sample, the bag was emptied over a stainless steel two-bed sorption tube, filled with carbograph 1TD/Carbopack X (Markes International, Llantrisant, Wales, UK). Before and after loading, the tubes were airtight capped. The desorption tubes were stored at room temperature until analysis.

Samples were analyzed initially by releasing the volatile compounds trapped on the desorption tubes by means of thermal desorption (Markes International, Llantrisant, Wales, UK). The gaseous mixture of released compounds was then split; 90% of the sample was recollected on a second desorption tube and stored for backup analysis, 10% of the sample was loaded onto a cold trap (5°C, Markes U-T2GPH: general purpose hydrophibic trap, designed for sampling VOC C4-C32), from which it was injected into the capillary column (Restek RTX-5ms, 30m x 0.25 mm) of the gas chromatograph (Trace GC, ThermoFischer Scientific, Austin, Texas, USA) and analyzed by time-of-flight mass spectrometry (Tempus Plus, ThermoFischer Scientific, Austin, Texas, USA). Helium was used as the carrier gas at a flow rate of 1.5 mL/min. Transferline temperature was kept at 250°C. The temperature of the gas chromatograph was programmed as follows: 40°C during 5 min, then raised with 10 °/min until a final maximum temperature of 270°C. In the final step this temperature was maintained for 5 min. Electron ionization at 70 eV was used combined with a 5Hz scanning rate over a mass range of m/z 35-350 amu. An example of a breathogram and the identification of a few peaks is shown in figure 5.1.

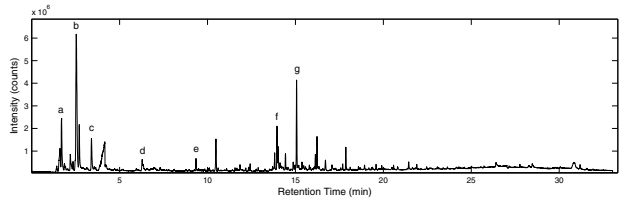

Figure 5.1: Example of a breathogram from a control subject demonstrating a few identified compounds as examples. a) acetone; b) 1,3-pentadiene; c) 2-methyl-pentane; d) benzene; e) toluene; f) 6-methyl-5-hepten-2-one; g) phenol.

Data analysis

The GC-MS chromatograms of the breath samples of the 105 subjects involved in this research were recorded and corrected for retention time differences, using retention indices, and lining up of easily recognizable component peaks. Corrected retention times were based on 6-10 common components with high intensities (e.g. acetone, toluene, phenol and benzene).The parts of the chromatograms that occurred at a retention index ¡0.15 and at a retention index ¿2.8 were removed from the chromatograms, because of unreliable data from these parts, due to noisy mass spectra at the beginning of the chromatograms and column bleeding at the end of the run. Baseline corrections on all analysis output files and peak detection were performed and the area under the peak was calculated. Normalization of the calculated peak areas was performed using an area scaling factor. This rescaling factor is based on the cumulative area under the detected peaks since this standardization principle proved to be the most robust.[22] The intensity of a specific VOC corresponds with the area below that peak. The resulting data files were carefully combined. This was done by means of combining of identical compounds from different samples based on similarity of mass spectra and retention times. This resulted in a final database file containing almost 6000 different chromatographic peaks.
To determine which compounds were of interest regarding the classification of CF patients and controls, we used stepwise discriminant analysis using SPSS (SPSS Inc. Chicago, Illinois, USA). An attribute (VOC) was only included in the discriminant analysis if its prevalence was 7% or more within the study population. Discriminant analysis was assessed with the grouping variables diagnosis (CF n=48, controls n=57), and positive *P. aeruginosa* cultures (yes n=23, no n=17). To exclude bias by different CF treatments (use of an antibiotic, a corticosteroid or DNase) VOC patterns of patients with or without this treatment were compared. The homogeneity of the control group was assessed by comparing VOC patterns between children selected at the primary school versus the subjects recruited at the outpatient clinic. The quantification of the similarity between the VOC profiles was done by means of calculation of a distance measure (dot-product rule) or match factor, as been described in detail by Stein et al.[63] This distance measure is based on the similarity of the entire

raw chromatogram. A value of 1 denotes identical samples, the lower the value the lesser the degree of similarity. Values of 0.8 and over can be characterized as very good agreement.[63]

Selection of informative compounds

All data are expressed as mean±SE. A normal distribution was present for all patient characteristics. Therefore, two-sided independent-sample T-tests were used for the analysis of the subject characteristics. P-values of <0.05 were considered statistically significant. To analyze the VOC profiles, 10 times cross-validation with multiple discriminant analysis was used, to minimize type III errors.[87] Of 10 subsamples, 9 subsamples are used as a training set and the remaining single subsample was used to test the model (validation set). The cross-validation process was repeated 10 times, with each of the 10 subsamples used once as the validation data. The 10 results were averaged to produce a single estimation. Discriminant analysis was applied in order to: 1) investigate differences among groups, 2) determine the most parsimonious way to distinguish among groups, 3) discard variables which are of little interest related to group distinctions, 4) classify cases into groups, and 5) test theory by observing whether cases are classified as predicted. Reproducibility and influence of breathing pattern. The influence of the breathing pattern (unforced breathing pattern versus hyperventilation) on VOC profiles, and the variation in VOCs across the day and between days was analyzed in healthy control children aged 5-15 yrs (mean±SD, 15.5±3 yrs). Sampling of VOCs was applied during an unforced breathing pattern (conform the entire study population), during hyperventilation, after one hour, and after one day.

Results

Subjects with CF were characterized by a light to moderate airway obstruction, air-trapping, and a mild restrictive impairment in comparison with controls. In the CF group, 49% of subjects had positive cultures with *P. aeruginosa* in the past 2 years, and 8 persons (17%) showed a history of allergic bronchopulmonary aspergillosis (ABPA). The meanSE time between sputum collection and breath sample collection was 50±62 days. Patients were mainly treated with antibiotics, antacids, DNase, and corticosteroids (table 5.2).

About 6000 different VOCs were identified in the chromatograms of the entire study population. After selection of attributes with a prevalence of at least 7%, 1099 VOCs were included in the analysis.[68] Variation across the day and between days Distance measure calculation of all complementary files within one hour, resulted in a mean±SD match factor of 0.90±0.038. Similarity of VOC profiles after one day was 0.85±0.096. These data indicate a low within-subject variation in VOC patterns after one hour, and after a day. Influence of breathing patterns Samples were assessed with an unforced breathing pattern and during hyperventilation. Characteristics are shown in Table 5.3.

Table 5.2: Additional subject characteristics (n=48)

Allergy*	
Total IgE (KU/L)	625 ± 263
Active eczema	1(2%)
Allergic rhinitis	2(4%)
Lung Function Indices	
Reversibility †	19%
TLC (% predicted)	95.2 ± 4.8
RV (% predicted)	182.4 ± 18.8
ITGV (% predicted)	114.7 ± 7.3
Treatment	
Oral steroid	4(8%)
DNase	25(52%)
Antibiotic	30(63%)
Antacid	24(50%)
Inhaled steroids	7(15%)
Colonization	
Pseudomonas Aeruginosa	23(48%)
Staphylococcus Aureus	29(60%)
Haemophilus Influenzae	14(29%)
Haemophilus Parainfluenzae	3(6%)
Aspergillus Fumigatus	17(35%)
Candida Albicans	6(13%)

Table 5.2: Abbreviations: ITGV, intra-thoracic gas volume; RV, residual volume; TLC, total lung capacity. Data are given as mean ± SE except were indicated otherwise. * A child was considered allergic when the total IgE level exceeded 20 kU/l and / or the Phadiatop was positive and / or the Radio Allergo Sorbent Test (RAST) had two or more allergens = class 2. †Reversibility is defined as the presence of an increase in FEV_1 of 9% of predicted value or more after inhalation of 400g salbutamol.[50]

Table 5.3: Characteristics of breathing patterns

Attribute	Units	Standard	hyperventilation	P-value
Exhalation pressure	mbar	22.3 ± 11.3	64.2 ± 15.5	<0.001
Sampling duration	seconds	$15:45 \pm 1:28$	$15:35 \pm 1:16$	0.11
number of exhalations		3.5 ± 1.6	3.9 ± 1.3	0.27
Environmental temperature	°C	23.5 ± 0.4	23.7 ± 0.6	0.30
Environmental humidity	%	51.7 ± 9.5	50.5 ± 8.1	0.27

Table 5.3: Data are given as mean±SD.

The match factor (mean±SD) of these breathing patterns was 0.95±0.043 which indicates minimal influence of breathing pattern on VOC profiles. Discriminant analysis: diagnosis of CF. It was possible to assess distinctive VOC profiles, which discriminated between CF children and controls. Based on 10 times cross validation, a 100% correct classification of CF patients and controls was found using 22 attributes or more (figure 5.2). The sensitivity, specificity, and the percentage of correct classification increased with increasing attributes in the analysis. The specificity of the models with 1-26 attributes ranged from 91% -100% with a corresponding sensitivity of 58-100%. Six of these VOCs were classified as hydrocarbons with 4 to 16 carbon atoms. Figure 5.4 shows the mean relative intensity of the 10 most prominent attributes which discriminated children with CF from healthy controls. None of the attributes were significantly different between CF children with of without the use of antibiotics, corticosteroids and DNase use ($p>0.130$, One way ANOVA). In the control group, there was no significant difference in between subjects recruited from the primary school or the children from the outpatient clinic with initial complaints of constipation or enuresis nocturna. As acquisition of *P. aeruginosa* in the sputum is associated with increased morbidity and mortality, we were interested in VOC profiles of CF subjects with and without positive cultures of *P. aeruginosa*. Within the CF group, it was possible to identify patients with or without positive *P. aeruginosa* cultures 100% correctly by means of 14 VOCs in exhaled breath (figure 5.3).

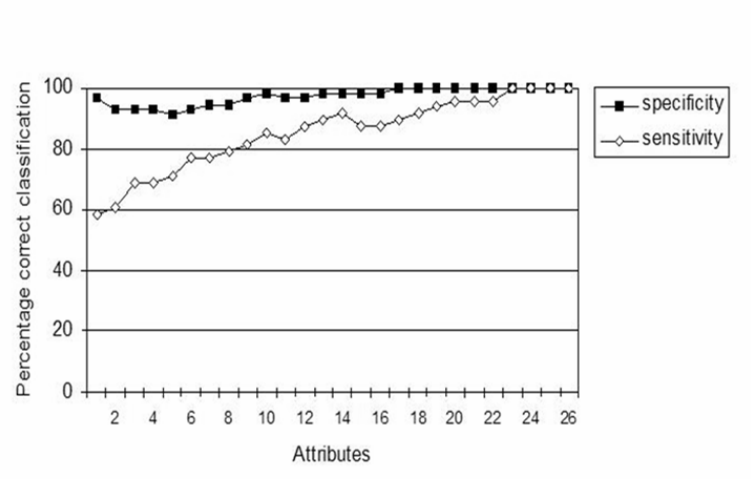

Figure 5.2: The percentage of correct classification of cystic fibrosis with increasing numbers of attributes, using discriminant analysis The discriminant analysis was performed using 10 times cross-validation. An attribute was only included in the analysis if its prevalence was 7% or more within the study population. This figure shows a 100% correct classification of cystic fibrosis patients and controls with use of 22 attributes or more. The sensitivity of the models assessed with 1-26 attributes, ranged from 58-100%. The specificity of these models, ranged from 91-100%.

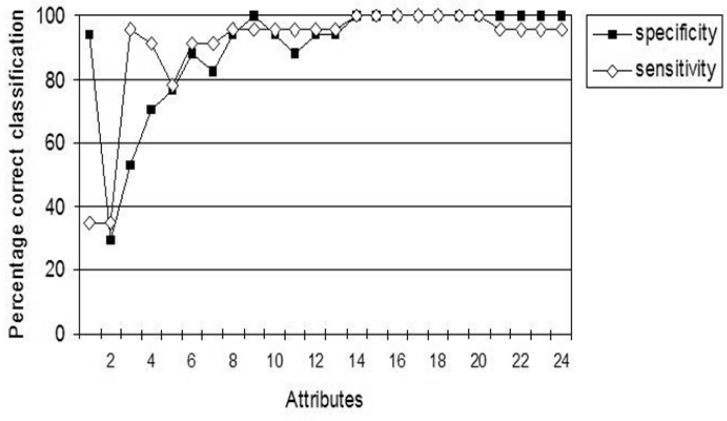

Figure 5.3: The percentage of correct classification of positive or negative *P. aeruginosa* cultures in sputum of cystic fibrosis patients with increasing numbers of attributes using discriminant analysis The discriminant analysis was performed using 10 times cross-validation. This figure shows a 100% correct classification of cystic fibrosis patients with and without positive *P. aeruginosa* cultures in sputum with use of 14 attributes or more. The sensitivity of the models assessed with 1-24 attributes, ranged from 34-100%, the specificity ranged from 29-100%.

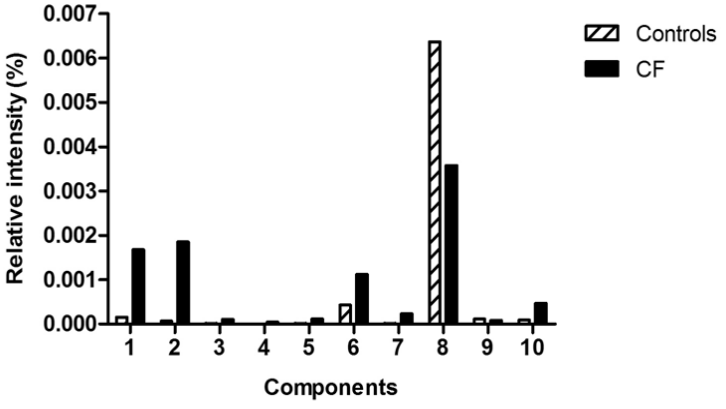

Figure 5.4: Mean relative intensity of the ten most prominent attributes that are used to generate the discriminant function that discriminates children with CF from healthy controls. The numbers represent the order in which the attributes were included in the discriminant analysis.

Discussion

This study showed that metabolomics of VOCs in exhaled breath of young subjects with cystic fibrosis and healthy controls was possible in a reproducible

Table 5.4: Identification of 10 most discriminating VOCs

Identification	Intensities
3,3-dimethylhex-1-ene	$72(41-107)$
2-buten-1-ol	$29(14-71)$
N-methyl-2-methylpropylamine	$8(6-13)$
C8H16 hydrocarbon (2-octene, 3-octene)	$3(1-4)$
Tolualdehyde (o-, m-, or p- isomers)	$35(25-48)$
C16 poly-unsaturated hydrocarbon	$7(17-22)$
C12 saturated hydrocarbon	$11(4-19)$
C13 saturated hydrocarbon	$5(1-9)$
Benzothiazole	$10(4-14)$
Long chain alkylbenzene	$1(0-6)$

Table 5.4: Identification of 10 most discriminating attributes between CF and controls. Compounds have been identified by comparison of the mass spectra to the NIST library and mass spectral interpretation by an experienced mass spectrometrist.

way, resulting in a 100% correct classification rate by using 22 compounds. Moreover, it was possible to discriminate between CF patients with or without positive *P. aeruginosa* cultures, a factor which is clearly associated with CF prognosis. The discrimination between CF and controls was mainly based on C5-C16 hydrocarbons and N-methyl-2-methylpropylamine. This study showed good feasibility of exhaled breath collection in 5 liter Tedlar bags in children of 5 years and over. More than 1000 different VOCs could be measured in exhaled breath of CF patients and controls with a prevalence in the two groups of at least 7%, with good short-term and long-term reproducibility. Breathing patterns did not influence VOC profiles significantly. The match factor is based on a distance measure implementing the dot-product rule was applied to establish the degree of similarity between measured chromatograms.[63]

We acknowledge some limitations in this study: Variance of data is influenced by a large range of patient related and technical factors, therefore there should be some caution in identifying biomarkers. It is of great importance to correct for the selection bias. Ambroise et al. described how selection bias can be assessed and corrected. They recommend 10-fold rather than leave-one-out cross-validation to handle overestimation of the results. Therefore, 10-fold cross-validation was applied in our study.[88] Although 10 times cross-validation is a valid method to prevent overestimation of the predictive value of markers, external validation in a control population of patients and controls is necessary in a second stage in order to further validate the results. The control group in this study was significantly younger that the CF group. This probably did not bias our results as no significant influence of age was found in the discriminant analysis.

As reported previously by van Berkel et al., we did not correct our measurements for chemical background appearing in the samples.[22] This is due to the fact that it will not be possible to correct for the complex interdependencies between excretion and uptake of VOCs by easily subtracting the inhaled from the exhaled air. Moreover, background noise will be randomly distributed between subjects samples and would thus neither exert any discriminatory power,

nor interfere with the outcome of the analyses. We are aiming with discriminative analysis to select only those compounds that are specific for the disease or condition and should thus principally not depend on background chemicals.
In the past few years, there has been increasingly interest in presence of VOCs in exhaled breath of patients with chronic lung diseases. Philips et al. assessed VOCs in exhaled air by means of automated thermal desorption gas chromatography mass spectrometer, and classified patients with and without lung cancer, based on 9-22 VOCs.[15,48] In addition, they studied the potential of VOCs to differentiate between 42 patients with a suspicion of pulmonary tuberculosis and 59 controls, and tuberculosis patients with and without bacterial sputum colonization. These classifications were possible with high sensitivity and specificity.[24] In 2006, Barker et al. studied specific VOCs by means of a customized gas chromatograph in CF patients and controls. They reported a significant lower level of dimethyl sulphide (DMS) in CF compared to controls.[13] However, DMS was not identified as an important discriminating component in the present study. Other methods to measure VOCs are sensor systems like the electronic nose, based on chemical vapor arrays responding to VOCs, and colorimetric sensor arrays. These methods were applied to detect lung cancer.[21,50] In contrast to VOC assessment with GC-TOF-MS, the electronic nose consists of a composite array of 32 organic polymer sensors. This is a limitation as the significance of specific VOCs for specific lung diseases is not established yet. Firstly, it is important to specify the VOCs important for a specific disease, or a specific question (VOCs important for diagnosis may not be the same ones that appear to be important for disease monitoring). In a second stage, when the relevant VOCs are defined, sensor systems like the electronic nose might be very helpful, as analyses can be fast, accurate, easy to perform and inexpensive.
Currently, diagnosis of pulmonary CF exacerbations and monitoring of disease activity are mainly based on clinical features and lung function tests. These parameters reflect changes in the functional abnormalities in the airways by infective and inflammatory processes, instead of the inflammation itself.[81] Therefore, inflammation may be present before clinical parameters change, introducing a time delay between the onset of a pulmonary CF exacerbation and the start of treatment. Persistent inflammation and repeated cycles of infection are present in CF lungs, resulting in progressive lung damage and pulmonary fibrosis.[59] Even in stable patients, chronic airway inflammation is present, as reflected by high airway fluid concentrations of proinflammatory cytokines.[89] Analysis of broncho-alveolar lavage (BAL) fluid has shown a thousand fold increase in the number of neutrophils from the lung of patients with CF compared to controls.
We hypothesize that VOC profiles may be helpful in the early detection of a CF exacerbation even before symptoms occur and lung function deterioration is present and the decision to stop antibiotic treatment in patients recovering from an exacerbation. However, this should be subject of future longitudinal studies. Based on this study, metabolomics of VOCs in exhaled breath discriminate between CF patients and controls. Dallinga et al. and Dragionieri et

al. showed that VOCs could discriminate between, respectively, children and adults with asthma and healthy controls.[20,85] In addition, Fens et al showed good discrimination between asthmatics and patients with chronic pulmonary obstructive disease by means of the electronic nose.[85] Therefore, the hypothesis is that these markers may be of additional value to study on-going processes of airway inflammation and oxidative stress, in addition to lung function and symptoms. As we were interested in the earliest stages of CF lung disease, the CF population of our study had mild to moderately severe pulmonary CF disease. It is not clear whether the results of this study can be generalized to more severe CF. This study showed good feasibility and safety of VOC collection in children. This is an advantage when compared to invasive methods like biopsies, bronchial alveolar lavage (BAL), and induced sputum. The short-term reproducibility of VOCs was excellent. We expect an even higher reproducibility of a typical VOC pattern, because not the absolute values but the values of the spikes relative to each other are important for the reproducibility of a specific pattern. However, this will be subject of future studies. The objective of the present study was to identify VOC profiles of CF disease and control children and to specify which VOCs are important in CF. In our study, products of inhaled medication or their derivatives, such as tobramycin or corticosteroids were not recovered in exhaled breath, and therefore did not contribute to the discrimination. By means of VOC profiles, we were able to differentiate between CF patients with or without positive cultures for *P. aeruginosa*. This result was consistent, although, other microorganisms may be concordantly present in the airways of patients, and a certain time period between collection of sputum samples and the sampling of VOCs was present in some subjects. Future research on VOCs in exhaled breath should not only be focused on clinical questions, but also on methodological issues like the influence of diet and exercise on VOCs. In addition, VOC profiles should be studied in longitudinal studies in order to investigate the additional value of VOC measurements to conventional parameters like symptoms, lung function indices and sputum analyses. Assessment of VOCs in exhaled breath is a new, promising, non-invasive technique, which in addition to conventional parameters can be used to study airway inflammation and oxidative stress in CF patients.

CHAPTER 6

Prediction of exacerbations in cystic fibrosis based on volatile organic compounds in exhaled breath

Van Berkel JJBN, Robroeks CMHHT, Dallinga JW, Jöbsis Q, Moonen EJ, Wouters EFM, Dompeling E, Van Schooten FJ
Submitted

ABSTRACT

Background

Pulmonary exacerbations of cystic fibrosis (CF) are an important cause for the hospitalization of patients, respiratory symptoms, and decreases in lung function. Therefore, prevention of exacerbations in CF is important. Airway infections in CF and changes due to exacerbations are characterized by oxidative stress/airway inflammation and the subsequent changes in the profile of volatile organic VOCs (VOCs) in exhaled breath that are excreted via the lungs. We recently developed a methodological approach to identify profiles of VOCs in exhaled breath that are unique for patient groups. In the present longitudinal study we studied whether changes in VOC profiles can predict pulmonary exacerbations of CF

Methods

A one year longitudinal study was performed in 26 CF patients (age [mean±SE] 15.8±2.1 years). At 2-month intervals, exhaled breath samples were obtained and lung function and symptoms were determined in a standardized way. The VOCs were analyzed by GC-TOF-MS. Advanced statistical unsupervised learning algorithms were applied to find relevant VOCs to predict the occurrence of exacerbations.

Findings

Seventeen CF patients experienced an exacerbation of which 8 experienced a second event during the study. In total, 2667 different VOCs were analyzed. It appeared that 18 VOCs were differentially present in exhaled breath from patients not having an exacerbation versus patients suffering an exacerbation event, 15 VOCs were differentially present in exhaled breath from the comparison of baseline measurements to the first exacerbation measurement at the start of the exacerbation event. Support vector machine classifiers based on these datasets classified 92% respectively 94% of exacerbations correctly.

Interpretation

This study shows that VOCs in exhaled breath are able to indicate CF exacerbations weeks before these adverse events are clinically manifest.

INTRODUCTION

Cystic fibrosis (CF) is a hereditary disease affecting exocrine glands. It will, among other multisystem failures, manifest itself as a progressive lung disease characterized by recurrent infectious events. These exacerbations lead to acute changes in pulmonary symptoms related to increased airway secretions. Pulmonary insufficiency has a great impact on the quality of life and is the main reason of morbidity in CF. In case of an exacerbation intravenous antibiotics and hospital admission are often necessary and exacerbations have also been associated with diminished lung function in later life.[40] The yearly exacerbation rate increases with age and more severe pulmonary impairment and is clearly related to survival.[41,42] To date there is no predictive measure regarding the occurrence of an exacerbation by means of an objective chemical, physiologic or histologic marker. Therefore, it is not possible to anticipate and by means of treatment prevent lung damage in the long term. An early and accurate diagnosis of an exacerbation will facilitate the possibilities to treat these adverse events before they are clinically manifest.[44]

Oxidative stress plays a major role in CF pulmonary exacerbations and before the exacerbations become clinically manifest the increase in oxidative stress might be useful as an indicator of early manifestation of the exacerbations. Literature states that the analysis of breath might prove useful as a diagnostic tool regarding the detection of oxidative stress and might prove useful in the diagnosis and/or early prediction of exacerbations providing clinicians with the possibility to treat these patients with antibiotics in an early stage in order to reduce the severity of the exacerbation or even prevent it.

Regarding the analysis of exhaled breath several biomarkers for diseases have already been identified. Nitric oxide (NO) levels are generally accepted as a marker of inflammation and oxidative stress in the respiratory tract in patients with asthma.[12] Additionally carbon monoxide (CO) has been investigated as a potential biomarker in chronic lung diseases. Yamaya et al. found a significant relationship between exhaled CO concentrations and the forced expiratory volume in one second (FEV_1).[8]

Besides these small and easy to identify VOCs in exhaled breath other volatile organic compounds (VOCs) available in exhaled breath have been studied regarding their hypothesized function as biomarkers.[59] The availability of hundreds of different VOCs in exhaled breath was already demonstrated by Pauling et al. in 1971[47] and many more have been identified since. Philips et al. have recently successfully demonstrated the possibility of using VOC profiles in exhaled breath as biomarkers of lung cancer and pulmonary tuberculosis.[14,15,48]

In previous research we identified subsets of VOCs that proved to correctly classify COPD patients from non-diseased controls[23] and children with asthma or CF from healthy controls.[20,90] Therefore, we are particularly interested in multi-component analysis of VOCs in CF patients suffering an exacerbation in order to increase diagnostic performance using a biomarker profile approach. These previous studies applied a cross sectional setup whereas the study described in this article demonstrates the longitudinal approach to the analysis of

exacerbations in CF patients. This approach may have great potential regarding clinical application for the assessment of exacerbations. Especially since the analysis of breath provides a fast, non-invasive, cost beneficial and easy to perform diagnostic tool.[22] The aim of this study was to identify biomarkers in exhaled breath that are able to 1) predict CF exacerbations reliably and 2) could monitor the course of exacerbations in CF.

METHODS

Study subjects

A total of 26 subjects (age [mean±SE] 15.8±2.1 years) suffering from CF were recruited at the Maastricht University Medical Center, The Netherlands. Subject characteristics are presented in table 6.1. CF disease was defined as the combination of characteristic clinical features (persistent pulmonary symptoms, meconium ileus, failure to thrive, Steatorrhea) and an abnormal sweat test (Chloride >60 mM).[91] CF severity was based on the Shwachman-Kulczycki score (SK-score).[86,92] The following exclusion criteria were applied: diseases that may interfere with the results of the study (e.g. other chronic inflammatory diseases, such as Crohn's disease or rheumatoid arthritis), inability to perform measurements properly or active smoking.

Study design

The design was a one-year prospective controlled longitudinal study. At 2-month intervals, routine clinical visits with assessments of clinical parameters, Shwachman-Kulczycki score, measurement of lung function indices and collection of exhaled breath samples were carried out. In case of an exacerbation, patients were asked to come for four extra visits at day 1, 5, 10 and at the end of the exacerbation. These additional visits were planned at most twice during the study per patient. In order to diagnose a CF exacerbation in a standardized manner, the scoring system of Rosenfeld et al. was applied.[93] The definition of an exacerbation was based on respiratory symptoms, school or work absenteeism, weight loss (over one kg over the preceding month), crackles or rhonchi on auscultation, and a decrease in FEV_1 of more than 20% of personal maximum value. Recognition of an exacerbation in an early stage was important in this study. It was made possible by means of home monitors (AM1, Viasys, Hoechberg, Germany). With the home monitor, (FEV_1), as well as use of antibiotics, overall well-being, and presence of pulmonary symptoms (cough, sputum, dyspnoea) were recorded 3 times a week at fixed times. In case of an exacerbation, patients were called to the hospital for the additional measurements. To monitor the development of an ongoing exacerbation the CF clinical score questionnaire (CFCS) of Kanga et al. was used.[94] All subjects or their parents gave their informed consent and the study protocol was approved

by the medical ethics committee of Maastricht University. The international study number was NCT00404859.

Table 6.1: Subject characteristics (26 CF patients)

Age (yrs)	15.8 ± 2.1
Weight (kg)	37.8 ± 3.1
Height (cm)	143.1 ± 4.1
Sex (M/F)	13/13
Lung Function Indices	
Bronchodilating response in FEV_1 (%)	2.9 ± 1.1
FEV_1 (% predicted)	68.3 ± 4.1
FEV_1/VC (%)	77.9 ± 3.3
FVC (% predicted)	72.3 ± 4.1
MEF50 (% predicted)	51.1 ± 5.9
TLC (% predicted)	106.4 ± 2.6
RV (% predicted)	180.1 ± 13.9
ITGV (% predicted)	121.3 ± 5.2
Treatment	
Antibiotics, prophylactic (inhaled or oral)	7(27%)
DNase	18(69%)
Fat soluble vitamins	23(89%)
Antacids	20(77%)
Inhaled steroids	6(23%)
Long acting 2-agonist	5(19%)
Sputum Cultures	
Positive sputum cultures (yes/no)	22/4
Pseudomonas Aeruginosa	9(43%)
Staphylococcus Aureus	17(77%)
Haemophilus Influenzae	6(29%)
Haemophilus Parainfluenzae	5(24%)
Aspergillus Fumigatus	6(29%)
Candida Albicans	4(19%)

Table 6.1: Subject characteristics: mean±standard error are presented unless stated otherwise. FEV_1: forced expiratory volume in one second, VC: vital capacity, FVC: Forced vital capacity, MEF50: maximal expiratory Flow at 50%, TLC: total lung capacity, RV: residual volume, ITGV: intrathoracic gas volume.

Sample collection and analysis

Subjects were sampled at a centrally ventilated treatment rooms located in the hospital as described previously by van Berkel et al.[22] Subjects were asked to deeply inhale and subsequently exhale into resistance free Tedlar bags (1L). Usually 2 to 3 exhalations were sufficient to inflate the bag. The contents of the bags were transported under standardized conditions onto stainless steel two-bed sorption tubes, filled with carbograph 1TD/Carbopack X (Markes International, Llantrisant, Wales, UK) that trap VOCs. Desorption tubes were stored at room temperature until analysis. For the analysis these desorption tubes were placed inside a thermal desorption unit (Marks Unity desorption unit, Markes International Limited, Llantrisant, Wales, UK) and subsequently heated to 270°C in order to release all VOCs onto a cold trap. This preconcentration step is necessary to increase sensitivity and to provide the means to re-

lease all compounds onto the gas chromatograph (GC) at the same time. VOCs were separated by GC (ThermoFisher Scientific., Austin, Texas, USA). The gas chromatography capillary column RTX-5ms, 30 m x 0.25 mm 5% diphenyl, 95% dimethylsiloxane capillary, film thickness 1.0μm was used for separation. Subsequently detection by a time-of-flight mass spectrometer (TOF-MS) (Thermo Electron Tempus Plus time-of-flight mass spectrometer, ThermoFisher Scientific, Austin, Texas, USA) was applied. The temperature of the gas chromatograph was programmed as follows: 40°C during 5 min., then raised with 10°C/min until the final temperature of 270°C which was maintained for 5 min. Electron ionization at 70eV was used with a 5Hz scanning rate over a mass range of m/z 35-350 amu. An example of a breathogram and the identification of a few peaks is shown in figure 6.1.

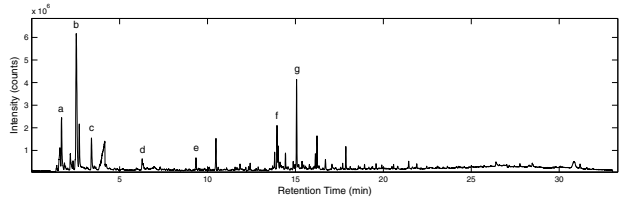

Figure 6.1: Example of a breathogram from a control subject demonstrating a few identified compounds as examples. a) acetone; b) 1,3-pentadiene; c) 2-methyl-pentane; d) benzene; e) toluene; f) 6-methyl-5-hepten-2-one; g) phenol.

Data-acquisition

Analysis of the data output files from the GC-TOF-MS was performed in successive steps as previously described in detail.[22] In summary the first step was to perform peak detection and baseline corrections on all analysis output files. Normalization of the calculated peak areas was performed using an area scaling factor based on the cumulative area under the detected peaks; since all chromatograms display rather similar profiles this method of normalization proved most robust. Retention times (RT) of all samples were corrected for chromatographic drifting. Applying the correction for retention times is very effective and easy to perform. The use of an added internal standard is avoided, since already available compounds that demonstrate availability in a large number of samples are used, adding to the straight-forwardness and robustness of the presented methodology. Finally, the output files were merged by combining corresponding compounds based on retention time and on similarity of the corresponding mass spectra. The degree of mass spectra similarity was calculated using a match factor (MF) based on the similarity index as described by Stein et al.[63] These match factors were only determined for compounds within a selectable RT-window, this value was chosen based on analysis of the maximum RT-range of identical compounds found in several subject files. MF-threshold

values were determined based on a variety of complementary compounds manually combined which resulted in a used MF of 0.859.

Selection of informative compounds

To gain more insight into what VOCs hold predictive value regarding the occurrence of these exacerbations analysis of two settings was performed comparing different phases of the exacerbations.

- Setting 1
  The first explored setting (setting 1) was the comparison of baseline measurements (t_0) to the first exacerbation measurement (e_1) at the start of the exacerbation event. The baseline measurement was defined as the standard sample taken on average one month prior to the start of the exacerbation event and denoted by t_0. The measurements during the exacerbation were chronologically denoted by e_1, e_2, e_3 and e_4. If subjects suffered from a second exacerbation during the time of research these measurements were denoted by e_5, e_6, e_7 and e_8. These phases were treated as e_1, e_2, e_3 and e_4 respectively. The baseline measurement before t_0 is denoted by t_{-1}, the one before that t_{-2} as shown in figure 6.2.

- Setting 2
  The second setting (setting 2) explored was by comparing baseline samples (t_0) of subjects suffering an exacerbation to baseline measurements (t_0) of subjects not suffering an exacerbation. Sampling dates of all incorporated samples in this setting were chosen as close to one another as possible.

Both settings were explored since both settings might provide information with regards to the compounds that contribute to an early detection of exacerbations in CF patients. Setting 1 analyzed the intra-individual differences before and during exacerbations while setting 2 analyzed the inter-individual differences prior to exacerbations.

Component selection

Subsequent component selection and determination of interesting compounds was performed in two ways for both of the above described settings:

- By analyzing the compounds and determining the significantly different compounds based on a t-test, since the dataset demonstrated to be normally distributed. The p-values were corrected for by a Bonferroni correction because of multiple testing. The significantly different compounds could hold valuable information and effort was made to find those compounds that might hold the potential of a single biomarker for the early detection of exacerbations in CF patients. In addition to

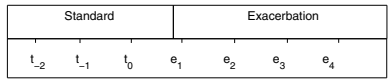

Figure 6.2: Study timeline demonstrating standard and exacerbation measurements. The baseline measurement was defined as the sample 0-2 months (on average 39 days) before the clinical onset of an CF exacerbation, denoted by t_0. Measurements before t_0, were coded t_{-1}, t_{-2},...etc. The time interval between these measurements was 2 months. During the exacerbation, measurements were denoted by e_1, e_2, e_3 and e_4, referring to the additional measurements at day 1, 3, 5 and at the end of the event, respectively. After the exacerbation, standard measurements continued at two months intervals denoted as e_1, e_2,... etc.

.

that, analyses of all compounds individually might provide future insight into the changes in pathophysiology of exacerbations although we are aware that numerous compounds in breath are still not biologically linked to metabolic or disease pathways and studying the pathophysiology of these exacerbations in relation to extracted compounds is currently out of scope.

- By selecting informative compounds fitted into a support vector machines (SVM) classification model that would preferably be able to (1) correctly classify samples of baseline measurements (t_0) compared to the first exacerbation measurement (e_1) and (2) baseline samples from subjects experiencing an exacerbation later on from baseline samples from subjects not suffering from an exacerbation at all. This might provide information regarding a profile of VOCs combined into an SVM model that hold predictive power with regards to early detection of exacerbations. Previous research has concluded that classification based on a profile of VOCs in breath is far superior compared to classification based on single compounds.

The SVM approach was chosen for its ability to construct predictive models with large generalization power even in the case of large dimensionality of the data when the number of observations available for training is low, which is obviously the case here. SVM are specifically useful since they seek a globally optimized solution and avoid over-fitting, so the large number of features or compounds is allowed.[69] The compounds are selected through a number of variable selection criteria. This selection algorithm will select the optimal subset of compounds able to correctly classify the dataset. A variety of subset selection methods was tested, among which the gain-ratio attribute evaluator. In order to obtain the best subset of compounds the attribute selection option implemented in Weka[64] was used. Compounds were selected using an SVM attribute evaluator. The attribute evaluator we used evaluated the worth of a subset of compounds by considering the individual predictive ability of each compound along with the redundancy between them. Preferably compounds were selected

showing high correlations within the class and low inter-correlation. After every run the least efficient compound was removed. A subset of the highest ranking compounds was implemented into an SVM classifier trained with John Platt's sequential minimal optimization algorithm.[65] The SVM classifiers were validated and performance was tested with use of 10 times cross-validation in which the entire dataset is split repeatedly into a test set (90% of samples) and a validation set (10% of samples).

RESULTS

Of 26 included patients, 17 patients experienced one exacerbation and 8 patients had a second exacerbation. Of the patients with one exacerbation, 2 withdrew from the study after the event. These patients were excluded from the analysis.

Significantly different compounds

The analysis of exhaled breath based on 24 subjects demonstrated that the breath contained a wide variety of VOCs. A total of 2667 different compounds were found and a mean of 332 different VOCs were detected per sample. Each subject delivered 7 to 15 samples depending on the occurrence of exacerbations or not. Table 6.2 shows the significantly different compounds in both settings. For each setting the significantly different compounds (p=0.05) were determined by a t-test with a Bonferroni correction. Significantly different compounds regarding setting 1: t_0 compared to e_1 (for subjects experiencing an exacerbation) were obtained using a paired t-test. Regarding setting 2: t_0 from subjects that would experience an exacerbation within 3-8 weeks compared to baseline samples from subjects not suffering an exacerbation were analyzed with a non-paired t-test. Identification of the compounds was performed by comparing the mass spectra to a library and by interpretation of the mass spectra by a experienced mass spectrometrist.
After selection of the significantly different compounds these compounds were tracked in all other samples (t_{-2} to e_4) and their relative intensity was determined in all samples in order to provide more insight in the course of intensities of these compounds prior to and during exacerbations. Analysis of the availability of a selected compound prior to and during the exacerbations ranging from phases t_{-2} to e_4 demonstrated that some of the significantly different compounds do display a correlation towards the phase of the exacerbation and might hold potential as biomakers for monitoring of an exacerbation. As shown in figure 6.3, 7 out of 32 selected compounds showed a trend towards the phase of the exacerbations as tested for with a Friendmann test, namely: eicosane; 1-amino-2-butanol; tetradecane 2,6,10-trimethyl; cyclopentacycloheptene; 5-hexene 5-methyl and heptane - were found using setting 1; 2 out of 7 compounds namely; 1-amino-2-butanol and 2-nonenal - were found using setting 2. One compound (1-amino-2-butanol) was found to be significantly different

in both settings. Other compounds did not show a trend associated with the phase of the exacerbation.

Table 6.2: Identified VOCs

Setting 1	
Compound	P-value
2-methyl-2-penten-1-ol	0.0465
I	0.0044
2-methyl-1-hexadecanol	0.0439
2,3,5,8-tetramethyl-decane	0.0198
II	0.0467
1,4-dione-2,5-cyclohexadiene	0.0387
2-methyl-nonane	0.0099
2,3-dimethyldecane	0.0489
2-chloro-benzenamine	0.0151
III	0.0081
butyl-cyclopropane	0.006
3-buten-2-one	0.0349
heptane	0.0097
2,3,4-trimethyl-pentane	0.0486
IV	0.0498
Setting 2	
Compound	P-value
1-methylthiol-1-propeen	0.0285
5-tridecene	0.0082
2-tetradecen-1-ol	0.0421
V	0.0468
3,5-dimethyl-1-hexene	0.0325
VI	0.0219
VII	0.0415
2-pentene	0.0429
3,6-dimethyl-octane	0.0217
nonadecane	0.0124
diphenylmethane	0.0278
dodecanal	0.04
VIII	0.0201
IX	0.0383
3,3-dimethyl-1-hexene	0.0286
1-propene-1-thiol	0.0238
X	0.025
1,5-pentanediol	0.048

Table 6.2: Significantly different compounds regarding both setting 1 and setting 2. P-value is corrected for with Bonferonni correction. Unidentified VOCs are denoted by I,II,II etc.

Support vector machine classifier

The second approach to determine which combined compounds might be useful for the prediction of an exacerbation and classification of the exacerbation stages was performed by means of building an SVM model. Compounds detected in less than 8% of samples were deleted from the dataset according to Penn et al.[68] in order to avoid inclusion of compounds demonstrating availability in a very low number of samples resulting in a very high discriminatory

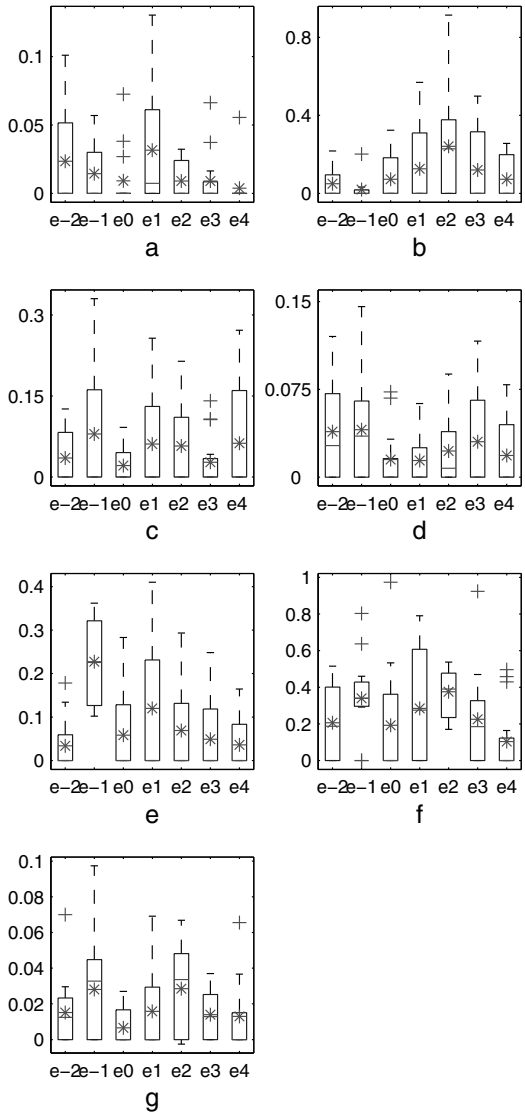

Figure 6.3: Availability of selected significantly different compounds demonstrating a certain trend towards the phase of the exacerbation. The start of the exacerbation is denoted by 'e_1'. a) eicosane; b) 2,6,10-trimethyl-tetradecane; c) cyclopentacycloheptene; d) 1-amino-2-butanol; e) 5-methyl-5-hexene; f) heptane; g)2-nonenal.

power of these compounds with regards to samples they were detected in. This will result in overfitting and has to be avoided. The 912 compounds that were available in at least 8% of samples were added to the SVM analyses. The resulting datasets were selected as described in the previous paragraph (Selection of informative compounds). The identified compounds and their performance when introduced into an SVM model are shown in tables 6.3 and 6.4. It should be emphasized that the best performing compounds implemented in an SVM model do not necessarily exhibit high individual sensitivity or specificity since the performance of the SVM model is based on the cumulative information of all implemented compounds. As demonstrated in table 6.3 the most optimal performing SVM for setting 1 (baseline versus exacerbation within patients) is based on a subset of 6 VOCs classifying 94% of samples correctly (sensitivity 94%, specificity 95%). Adding more VOCs does, to some degree, increase performance, but, as little VOCs as necessary are to be incorporated in order to minimize or avoid the effect of over-fitting. The most optimal SVM model regarding setting 2 (baseline resulting in exacerbation versus baseline not resulting in exacerbation) is based on 8 VOCs. Exacerbations were correctly classified in 92% of samples (sensitivity 100%, specificity 84%) as shown in table 6.3.

Table 6.3: SVM performance

Setting 1 No. of compounds	Specificity	Sensitivity	% Correct
9	0.75	0.89	0.83
8	0.94	0.89	0.91
7	0.94	0.95	0.94
6∗	0.94	0.95	0.94
5	0.81	0.89	0.86
4	0.69	0.95	0.83
3	0.63	1	0.83
2	0.56	0.89	0.74
1	0.44	0.79	0.63
Setting 2 number of compounds	Specificity	Sensitivity	% Correct
11	1	1	1
10	0.84	0.95	0.89
9	0.89	0.95	0.92
8∗	0.84	1	0.92
7	0.79	1	0.89
6	0.79	0.84	0.82
5	0.74	0.89	0.82
4	0.63	0.89	0.76
3	0.47	0.84	0.66
2	0.32	0.95	0.63
1	0.26	0.84	0.55

Table 6.3: Performance of the support vector machines in relation to the number of VOCs implemented into the model. The asterisk denotes the optimal performing SVM for both setting 1 and setting 2. Where in case of analysis of setting 1 a model based on 6 compounds correctly classified 94% of samples, regarding setting 2 a model based on eight VOCs correctly classified 92% of samples.

Table 6.4: Identified VOCs

Setting 1
2-nonenal
7-hexadecenal
5-butyl-4-nonene
3,7-dimethyl-6-nonenal,
2-decanal
3,5-dimethyl-5-hexen-3-ol
Setting 2
7-methoxy-3,7-dimethyl-octanal
2-pentene
2-chloro-hexanal
I
5-diol-cyclohexene-3
1,3-decadiyne
pentane
II

Table 6.4: Identified compounds as implemented into the most optimal SVM classifiers as shown in table 6.3. The two models, regarding both settings, based on these VOCs demonstrated superior performance. Compounds have been identified by comparison of the mass spectra to the NIST library and mass spectral interpretation by an experienced mass spectrometrist. Unidentified VOCs are denoted by I,II,II etc.

DISCUSSION

This paper describes our effort to develop and test a methodology based on the analysis of exhaled breath to obtain information regarding the exacerbations patients suffering from CF experience. The principle is based on identification of VOCs in exhaled breath, a totally non-invasive, easy to perform, cost effective and accurate methodology. Based on the presented data, early recognition of an up-coming exacerbation and consequently early treatment with antibiotics, leading to prevention or suppression of exacerbations in future studies seems feasible.[95]

We identified 33 VOCs that demonstrated a significant difference in availability in breath in the two mentioned settings. Setting 1: t_0 compared to e_1 (for subjects experiencing an exacerbation) using a paired t-test in which 15 VOCs proved to be significantly different. Setting 2: t_0 from subjects that would retrospectively experience an exacerbation within 3-8 weeks compared to baseline samples from a subject not suffering an exacerbation in which 18 VOCs proved to be significantly different. Further analysis of these 33 significantly different VOCs revealed that in 7 VOCs demonstrated a trend towards the phase of the exacerbation. This demonstration might provide the means for detailing an exacerbation or monitoring the phase of the exacerbation. From the VOCs shown in figure 3 however only 5-metyl-5-hexene and 2-nonenal demonstrated a non-significant difference ($p<0.05$) between the t_{-2} and e_4 phases. All other VOCs show a significant difference between these two phases possibly implying that at time point e4 - the end of the clinical manifestation of the exacerbation - these values had not returned to baseline values yet, which makes use of these VOCs as biomarkers for exacerbation progression unlikely. It is however

interesting to notice that all VOCs shown in figure 6.3 do demonstrate a trend towards to phase of the exacerbation with regards to phases t_0 to e_4. It also demonstrates that prediction of an exacerbation in all subjects with acceptable sensitivity and specificity could not be accomplished based on a single VOC. External validation of these results in a larger study population has to be performed.
Our study also demonstrates that SVM classifiers are able to predict an exacerbation with acceptable sensitivity and specificity. To classify subjects suffering an exacerbation from subjects not suffering an exacerbation a model based on VOCs in exhaled breath resulted in 94% correct classification of the samples as validated with 10 times cross validation. Using sophisticated statistical tools we extracted 6 VOCs out of nearly 912 exhaled compounds that combined into an accurate SVM model. This model proved to be highly sensitive and specific regarding the prediction of exacerbations intra-individually. The second SVM model was based on baseline measurement comparing patients suffering an exacerbation 2-4 weeks after the baseline measurement versus patients not suffering an exacerbation. This resulted in an SVM model comprised of 8 VOCs able to classify correctly 92% of exacerbations.
The selected VOCs appeared to be mainly long chain hydrocarbons however, the origin of some of the identified VOCs remains elusive. There are several hypotheses. First, bacterial infections might give rise to changes in the profile of exhaled VOCs as metabolic products are degraded or formed as described by Syhre et al. who monitored bacterial headspace changes.[75] Whether it is possible to detect such small changes in profiles of VOCs produced by these bacteria in human breath - that is if these bacterial VOCs are not altered by human metabolic processes - is unknown. Second, discriminating VOCs might be due to changes in the degree of oxidative stress. Generation of reactive oxygen species (ROS) by activated neutrophils cause degradation of polyunsaturated fatty acids (PUFAs) in a process called lipid peroxidation, giving rise to VOCs like ethane and pentane. These hydrocarbons demonstrate a low solubility in blood and if the bloodstream passes the lung tissue these hydrocarbons are excreted into exhaled breath. In earlier research we have been able to identify VOCs in exhaled breath from smokers possibly related to oxidative stress. These VOCs were already mentioned in literature as smoking behavior related and were not exogenous of.[22,66,96] We also build an SVM model based on a set of 6 VOCs able to discriminate between COPD patients and non-diseased controls.[23]
Especially the VOC 2-nonenal and its hydroxilized form 4-hydroxy-2-nonenal are described in literature as reaction products of lipid peroxidation of omega-7 fatty acids and has already in 2003 been linked to inflammatory processes in COPD patients.[97] Also the chlorinated VOC 2-chloro-hexanal has been described in literature to be generated by a chlorination process in case of a neutrophilic response.[98]
Further studies are necessary in clinical settings but also in inflammation models to explain the biochemical origin, the physiological meaning and exhalation kinetics of our selected VOCs. Analyses of all VOCs individually might provide

future insight into the changes in pathophysiology of exacerbations although we are aware that numerous VOCs in exhaled breath are still not biologically linked to metabolic or disease pathways and studying the pathophysiology of these exacerbations in relation to extracted VOCs is currently out of scope. Nevertheless without this mechanistic knowledge the VOCs may already be of value as a monitoring tool for clinical studies.
In the research field of exhaled breath analyses, there is an ongoing discussion whether or not background measurements should be taken into account in order to correct for ambient air influences. In the presented work, no background extraction was applied to keep the methodology as easy and straightforward as possible. Moreover, Miekisch et al. mentioned that it will not be possible to correct for the complex interdependencies between excretion and uptake of VOCs by easily subtracting the peak areas obtained from inhaled air and breath.[3] In addition, the VOC selection procedure will only select those VOCs that contain information about the disease status and since the subjects have been randomly sampled, all VOCs that are indeed influenced by some confounding factors, like background or time-of-day, will not display information regarding the disease or severity. This random sampling limits the possibility of an exogenous VOC showing up only in a certain measurement (t_{-2} to e_8) in all subjects in order to constitute a highly significant effect.
Another point of discussion is the influence of dead space. No correction for dead space was applied since omitting the start volume of breath might add noise to the samples and reduce sensitivity. In our research we did not correct for dead space by means of sampling only the alveolar air. The applied methodology samples a mixture of alveolar air and dead space air, which is about 90 ml in children of 40 kg. Since most of the subjects were able to fully inflate the 5 L Tedlar bag in 3 to 4 expirations the contribution of dead space to the total volume is small and 90% of the obtained volume originates from alveolar air and in our setup the dilution with dead space air does not lead to sensitivity issues. In addition we wanted to provide a methodology as easy and straightforward as possible. Results regarding the reproducibility of the methodology are provided in previous research.[22]
One of the innovative steps of our approach compared to other studies in this field is that we used the raw mass spectra to find the matching VOCs in all subjects instead of searching for matching VOCs based on their chemical identification. This results in a far more precise and accurate dataset since chemical identification remains elusive and numerous misidentifications result in corrupt datasets. Our solution was to introduce the match factor in order to directly match VOCs from different samples based on their mass spectra. This increases accuracy since VOCs are compared to one another as measured with the same experimental setup thus eliminating the differences between library mass spectra and measured mass spectra arising from use of different setups. This results in a superior database as compared to a database based on identification of VOCs. The final identification of the VOCs as shown in table 6.4 remains difficult and in a few cases questionable, due to GC-resolution limitations.

Our main goal was to develop a robust diagnostic methodology as a whole not influenced by too many confounding factors. In principal the excretion of VOCs can be influenced by many factors of intrinsic nature (gender, age, weight, genetic background) or of exogenous origin (ambient air, diet or medication). It is practically impossible to control for all these potentially confounding factors. Therefore it was our goal to select those VOCs that provide information independently from other endogenous or exogenous confounding factors. We recently concluded that our sampling procedure, chemical analysis, data handling and accurate data mining provide a highly reproducible methodology.[22] Now that interesting VOCs have been selected new advances in diagnostic tools like the electronic nose[55,78,83] could be specifically tuned to detect the aforementioned VOCs in order to provide for a fast, non-invasive cost effective and efficient diagnostic device to detect exacerbations in an early phase. When combined with early (antibiotic) treatment, this method may result in the prevention of exacerbations, an improvement in quality of life and lung function and a reduction in healthcare cost which clearly is a topic for future prospective studies. In further research the pathophysiological meaning of the detected VOCs could be tracked.
Further studies are necessary in clinical settings to explain the biochemical origin, the physiological meaning and exhalation kinetics of selected VOCs. However, without this mechanistic knowledge these VOCs may already be used as a diagnostic tool for exacerbations in future clinical CF studies.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the province of Limburg (The Netherlands), the Jaap Swieringa Foundation, the NCFS and AstraZeneca for their financial support.

CHAPTER 7

General discussion

Introduction

Metabolomics is the systematic study of unique chemical fingerprints that specific cellular processes leave behind.[99] It maps the entire profile of metabolites in a single cell, tissue or (part of the) organism. The concept that individuals might have a 'metabolic profile' that could be reflected in the make up of their biological fluids and gases was introduced by Roger Williams in the late 1940s.[100] Williams used paper chromatography to determine disease-specific metabolic patterns in urine and saliva. Advances in technology finally increased the interest in qualitative analysis of metabolic profiles further in the late 1960s and 1970s. Robinson realized that the patterns of hundreds or thousands of chemical constituents in urine contained useful information as proposed by Linus Pauling in the 1970s. This led to the characterization of metabolic profiles extracted from urine, blood and breath. Although it was not called metabolomics, their first paper devoted to this topic was entitled: 'Quantitative Analysis of Urine Vapor and Breath by Gas-Liquid Partition Chromatography' published in 1971.[47] Demonstrating the metabolomics approach to the analysis of breath is nothing new. For decades it is known that patients suffering from uncontrolled diabetes have an acetonic sweet smell of breath, while patients suffering from liver disease have a fishy smell to their breath. Apparently metabolic and disease pathways give rise to generation of certain volatile molecules that are finally delivered to breath. Identification of discriminating compounds in breath with regards to disease profiles might prove useful regarding disease status and monitoring.

Breath analysis overview

Sampling

Metabolomics on breath has, as described above, been performed since the 1970s. From this time on the applied methodologies regarding breath analysis have evolved tremendously. However to date no consensus has been established regarding one unified and applied operating procedure and different research groups apply different methodologies regarding sampling, sample analysis and data analysis. Different issues regarding the analysis of breath are currently hindering one unified methodology, the most important ones will be explained in the following paragraph where the different designs and issues regarding sampling of exhaled air are discussed. The first one deals with the debate regarding sampling of whole breath fraction compared to the alveolar fraction.
Exhaled air constitutes a mixture of dead-space air and alveolar air. The dead space air is roughly 150 ml and originates from air from the upper airway where no gaseous exchange is facilitated between blood and breath air. This part of the exhaled air demonstrates to a large degree a high significance with the previously inspired air. The alveolar air originates from the lower airways where concentrations of endogenous compounds have proven to be two to three times

as high compared to dead-space air since gaseous exchange between blood and breath air is facilitated in the lower airways. In short there are three ways to sample exhaled air: (a) upper airway collection, this procedure only samples the dead-space air, (b) lower airway collection, this procedure only samples the alveolar air and (c) mixed air collection in which whole breath is sampled; a mixture of dead-space air and alveolar air. Some breath tests like nitric oxide (NO) measurements mainly use the dead-space air since NO is released both into dead-space air as into alveolar air. Therefore to quantify the concentration of NO dead-space air is used. VOCs and alike compounds are however released into breath from the blood in case of systemic-inflammation related generation of VOCs. However in case of inflammatory lung diseases VOCs are also introduced in breath by means of release of VOCs from the target organ, in case of COPD the lung. These systemic inflammatory disease related VOCs require areas where gaseous exchange between blood and air is facilitated and therefore with regards to these compounds the alveolar air is of interest. The dead-space air dilutes the sample with respect to measurements of these VOCs. Different solutions have been described in literature to minimize the effect of dilution of the samples by dead-space air. The most efficient one being a setup in which the alveolar air is sampled by a carbon dioxide (CO_2) controlled valve. The end-tidal CO_2 concentration is used as a marker between dead-space air and alveolar air. A less subtle but more simple solution was proposed by Phillips et al. using a breath-collecting apparatus.[101] This device transports the alveolar air onto a desorption tube using a reservoir filled with every exhalation. After exhalation this reservoir is withdrawn in the reverse order in a flow controlled rate which ensures that the alveolar air is not depleted before the next exhalation, leaving behind the dead-space air[14] in an ideal situation where no diffusion is possible. The sampling procedure used in the research described in this thesis used a mixture of dead-space air and alveolar air since sample collection was based on sampling whole breath by means of inflating a Tedlar bag. No effort was made to reduce the degree of dead-space air in the samples. As mentioned before dead-space air constitutes a mere 150 ml on a total of some 2.5 to 4 liters of exhaled air in adults per exhalation. We have proven that the contribution of dead-space air to the total volume of whole breath does not lead to sensitivity issues in the currently used setup. Sampling whole breath does however add to the simplicity of the methodology both with regards to ease of use for the physician as to the degree of discomfort for the patient. This simplicity is high on the list of qualities a biomarker should exhibit.
Samples are obtained by inflating a Tedlar bag and the samples are subsequently transported onto desorption tubes. With regards to sample collection and storage before analysis it is of great importance to use materials not contaminating the samples. It has been extensively described that certain materials used for tubing, valves, storage and sampling containers will contaminate or alter the composition of the sample. Silicones and certain plastics or rubbers do release certain VOCs that if brought into contact with the sample will introduce noise. Stainless steel or glass containers and teflon fittings and cap-ends will help prevent this contamination. Air tight transportation and capping of sam-

ples is a necessity in order to prevent contamination from room air. We tested application of different materials and selected those that delivered a minimum of noise. Subsequently we tested whether storage time affected the sample composition up to 6 months of storage, and demonstrated that there occurred no (detectable) differences between samples with different storage times.
Due to the low concentrations most VOCs demonstrate in breath, preconcentration of the samples is of importance and can be accomplished in different ways. Cryogenic preconcentration is often used to get hold of the small highly volatile compounds. The exhaled air is cooled with for example liquid nitrogen and the VOCs of interest are trapped. A huge issue with regards to cryogenic preconcentration of breath is that breath contains water vapor and carbon dioxide that will cause ice crystals to form inside the apparatus and interfere with the measurements. Another way of preconcentrating the breath sample is by means of adsorption. This way the sample is transported through an adsorptive material. The VOCs bind with the agent and become trapped. Different adsorptive agents are used with regards to preconcentration of breath of which porous polymers like Tenax or Chromosorb, carbonized molecular sieves and graphitized carbon. These are currently the most widely used. The preconcentration method used in research described in this thesis was based on commercially available graphitized carbon based stainless steel desorption tubes containing a two stage Carbograph packing. This sorbant is widely used for trace level analysis since it exhibits low artifact levels, and enables use on a wide variety of VOCs. A huge benefit of these sorbants is the fact that they exhibit fairly hydrophobic characteristics which makes these sorbants very suitable for sampling under humid conditions, as is the case in breath analysis. Preconcentration can also be done by means of solid phase micro extraction (SPME).[102] This technique is based on a silica fiber coated with a polymer to which the VOCs from the sample adhere. The SPME fiber can be introduced directly into the gas- or high pressure liquid chromatograph. SPME is well suited for breath analysis and research has proven it facilitates detection of VOCs in the nanomolar concentration range. However due to the small sizes of these fibers the number of substances that can be adsorbed is limited. Research described in this thesis used the preconcentration by means of adsorption onto graphitized carbon packed in stainless steel tubes airtight locked with a teflon cap, since these tubes are specifically designed to trap a large degree of present VOCs from breath and are easy to use and store.

Chemical analysis

As the sample of exhaled air is obtained different analysis techniques are useful in determining the content of the samples like mentioned in **Chapter 1**. The methodology used in research described in this thesis is based on gas-chromatography mass-spectrometry. It is the most commonly used technique to analyze trace gases in exhaled air due to its high sensitivity and it enables the user to identify the compounds of interest based on the measured mass spectra.[57] The analysis has proven to be user friendly providing the possibility

of analyzing a large numbers of samples in an automated sequence. **Chapter 2** validated the instrumental reproducibility of this technique and demonstrated it to be of high degree. Both the instrumental reproducibility and inter/intra-individual variability have been mapped. The instrumental reproducibility was tested by analyzing two identical samples and determining the degree of similarity between the two measurements. A global chromatographic comparison match factor was applied presenting scores ranging from 0.96 to 0.99. A value of '1' denotes identical samples, the lower the value the lesser the degree of similarity, thus confirming as expected the intra-individual variability demonstrates rather large variations and even bigger variations are demonstrated by the inter-individual variability caused by the many confounding factors.

Data analysis

Due to advances and innovations in technology from the last few decades researchers in all sorts of disciplines have within their grasp highly sensitive and accurate sensors and measurement technologies facilitating to probe even deeper into space in search of gravitational waves, study vast geological phenomena or map the astonishing biomolecular mechanisms of the human body to name a few. These highly sophisticated technologies however overwhelm researches with ever growing databases and generated machine-outputs and these data structures ask for special models to extract and visualize the interpretable data. This also holds for the analysis of exhaled air. Like already mentioned every sample of breath contains over 300 compounds as measured with the previously described setup and a total of over 4000 compounds have been described demonstrating availability in breath in a large population. Fortunately in case of the exhaled air analysis the field of bio-informatics and bio-statistics hugely aids to interpreting these large datastreams. We improved accuracy of compiling the database by implementing a match factor. This match factor as adapted from Stein et al.[63] determines the degree of similarity between mass spectra in order to find complementary compounds in different samples. In contrast to other studies either manually compiling the dataset or compiling it based on identification of the compounds our methodology proved both more robust and time-efficient. Our self-written routines provide accurate and easy analysis of the data, finetuning all operations according to the needs of the generated data. However great care should be taken in interpreting the results extracted from these huge databases.
In order to extract interpretive data from the large datastructures as generated by the highly sensitive analysis techniques we applied several classification algorithms to isolate the VOCs of interest. The classification algorithm outperforming other classifiers regarding our datastuctures is based on support vector machines or SVMs. SVM analysis is an unsupervised learning algorithm able to perform binary classifications. An SVM is trained to discriminate between the members and the nonmembers of a class. After learning the features of the class, the classifier is able to correctly classify unknown samples as members of

a specific class. A huge benefit regarding SVM analysis remains the fact that it always seeks a globally optimized solution and is designed to avoid over-fitting. This implies a large number of features (or in our case compounds) is allowed. Due to the aforementioned and since SVMs have been shown to perform especially well in multiple areas of biological analyzes, especially functional class prediction from microarray gene expression data and chemometrics, application of the SVMs was the logical step to make.[21]

Biomarker validation and issues

Introduction of a biomarker or a profile of biomarkers as a new diagnostic tool requires a number of qualities this biomarker should exhibit as mentioned in **Chapter 1**.
Over the last few decades techniques offering the possibility of a clinical breath test have progressed significantly and a high increase in research and validation on breath tests have been performed in recent years. In order for breath test to be of clinical relevance this biomarker should exhibit the characteristics as discussed in **Chapter 1**. Besides demonstrating these characteristics there are a few issues regarding the analysis of exhaled air and its clinical relevance.

Validation

One of the issues regarding the analysis of exhaled air are shortcomings related to validation; both methodological validation as validation of obtained results. Like mentioned both instrumental reproducibility as inter/intra-individual variability should be studied. **Chapter 2** of this thesis demonstrated a very high degree of instrumental reproducibility as the same sample was split onto different adsorption tubes and analyzed individually. The same chapter described inter/intra-individual reproducibility and did not show unaccountable results. Another important aspect of validation is the use of validation datasets during data analysis, this in order to test the biomarker performance directly onto an independent dataset. This to account for overfitting occurring for example when a statistical model is too complex but does describe the relation of interest to a high degree. Such a model will generally exhibit poor prediction performance as minor fluctuations in the data will be exaggerated.

Standardization

Standardization remains currently one of the largest issues regarding analysis of breath due to the absence of consensus on almost all different parts of the analysis. Sampling of the subjects is performed with use of different machines or methodologies, sampling both alveolar air and whole breath. Analysis of the samples is performed with different techniques as explained in **Chapter**

1, making data comparison between research groups difficult. Data analysis is left to different software packages employing different algorithms to analyze and interpret the data providing in some cases different results. Study design and validation of the results make comparison of the results difficult for example due to lack of validation subjects to validate the performance of the biomarker in different studies. Therefore for breath tests to gain clinical relevance the different stages of the analysis should be more standardized to facilitate comparison and validation of results. International meetings should address this issue and should initiate the setup of a standard operating procedure or alike.

VOCs in ambient air

One of the major confounding factors is the background signal or room air; these are compounds of exogenous origin. It constitutes a large portion of the exhaled air thus masking or at the least interfering with the endogenous compounds of interest. A debate is going on in the scientific world whether and how to correct for this background air. Several propositions have been made on how to correct for this confounding factor but to date there is non consensus on this subject. In the methodology we developed we do not correct for background or room air since there is no simple and elegant solution on how to correct for it as is well documented by Miekisch et al.[3] Simple subtraction of the room air samples from the background sample is the most popular solution but this correction methodology does not account for the complicated interdependencies between excretion and uptake of VOCs by easily subtracting the peak areas obtained from inhaled and exhaled air.[3] We therefore applied no background correction instead we used a sensitive measurement system and advanced classification models to extract the compounds of interest. Since subjects were randomly sampled at centrally ventilated rooms we tried to randomize the effect of contaminants in background air limiting the discriminatory power of the compounds. Another argument with regards to not employing a background correction is that the overall effect will be minimized since more confounding factors expressing a large effect on breath will not be corrected for. Confounding factors like diet, exercise, smoking, gender and age all present an influence on the composition of breath. These effects can not be easily corrected for if even possible at all. However, noise from ambient air is expected to be randomly distributed between subjects' samples and would thus not exert any discriminatory power, nor interfere with the outcome of the analyses.[23]The huge advantages like the non-invasive nature of the methodology and its ease of use would be minimized if patients were to be sober, smoke free for a certain period of time, rest for a certain period of time before the measurement or inhale through a VOC-filter in order to avoid or minimize the influence of these confounding factors.

Pathophysiological relation

Currently one of the issues regarding the analysis of breath as a new biomarker is the relation of selected VOCs to the biological role these biomarkers play. For only a few of these biomarkers the pathophysiological link has been described but the majority of interesting VOCs with regards to disease remains unknown. In order for a breath biomarker to be of clinical value this pathophysiological link should be determined. Although different scenarios regarding the origin of VOCs from breath samples are possible, it still remains speculative to make any conclusions. Especially since compounds originating in breath can very well be biochemically altered and thus do not relate to disease directly. Further studies are necessary in clinical settings but also in inflammation models to explain the biochemical origin, the physiological meaning and exhalation kinetics of selected VOCs. Nonetheless without this mechanistic knowledge the compounds may already be of value to base a monitoring tool for clinical settings.

Prospectives

COPD heterogeneity

COPD has been long recognized as a heterogeneous disorder or group of disorders - asthma, chronic bronchitis, emphysema, and airflow obstruction - all being important parts of the final disease process.[34] The different components of disease heterogeneity in COPD include different mechanisms in development, presentation and course as explained in the following paragraph detailing the high potential of analysis of exhaled air regarding COPD.
Regarding initiation of the disease smoking is the number one risk factor for the development and progression of COPD. However since less than 25% of smokers develop COPD and more than 15% of COPD mortality occurs in non-smokers other factors are likely of high importance. Smoking cessation is the single most important intervention in COPD management but since the best reported cessation rates are still less than one third,[103] better treatments are needed. Different mechanisms of development have been extensively studied. Snider et al. describe that 1-Antiprotease deficiency is an important cause of COPD in a very small percentage of cases[104] but other undefined genetic factors certainly play important roles in COPD development.[105] The role of infections in both the development and progression of COPD is getting increased attention, especially the role of adenoviral infections in patients with emphysema is widely discussed.[106–108] Other important factors like occupational and environmental exposures to various pollutants also play an significant role in developing COPD.[109] But besides its heterogeneity in development COPD also demonstrates heterogeneity in its presentation. Based on data from NHANES III,[110] a significant proportion of patients with severe airflow limitation (FEV_1 <50% of predicted) may not report symptoms. The symptoms reported most

frequently include wheezing and shortness of breath in some 65% of subjects with FEV_1 values less than 50% of predicted.
COPD has become increasingly recognized as a systemic illness with effects on nutritional status and muscle wasting.[111] Many COPD patients probably have components of chronic bronchitis, asthma, and emphysema occurring simultaneously. Bear in mind some of this overlap might be related to misdiagnosis, however some of it may be a measure of the presence of reversibility. A better definition of the individuals comprising these groups ultimately may help to prepare better interventions and this is the area where exhaled air analysis might play a key role. An indication of disease heterogeneity and reversibility in COPD patients might be obtained by determining exhaled air composition on top of looking at respiratory symptoms, lung function, and activity limitation, thus adding to a new 'gold' standard regarding COPD and subsequently providing opportunities for superior targeted interventions and therapies as currently available.

Exacerbations in CF

Pulmonary exacerbations in patients with CF are associated with poor quality of life, declines in cognitive status, increased healthcare costs, and increased patient mortality.[41,112] Clearly, studies of factors that either predispose or protect CF patients from pulmonary exacerbations have the potential to significantly affect both patient outcomes and healthcare system outcomes. The results of published prospective studies suggest that age, sex, lung function, history of previous exacerbations, and use of inhaled corticosteroids are important factors that are associated with pulmonary exacerbations in CF patients. But needed is a more direct measure of worsening pulmonary symptoms based on easy to obtain biomarker levels. As demonstrated in **Chapter 6** analysis of exhaled air and adequate interpretation of selected VOC levels might provide the means to allow for early identification and closer monitoring of patients with high risk profiles if results will remain true in a larger validation study. Monitoring of high risk patients and early identification of worsening pulmonary symptoms might allow adequate therapeutic interventions with oral antibiotics and/or physiotherapy to prevent severe exacerbations adding to the quality of life in these patients.[113]

Perspectives

Biomarker prediction

Whether novel biomarkers add useful information for disease diagnosis, disease monitoring or risk prediction has been the focus of intense scrutiny in the scientific literature especially regarding cardiovascular diseases but the findings can very well be extrapolated to medical biomarker evaluation in general.[114,115] A

variety of factors are responsible for these conflicting findings: inadequate statistical power, use of older biomarkers, the lack of measures such as calibration and reclassification and disuse of external validation sets have been invoked to explain the poor performance of biomarkers in some studies.[116,117] On the other hand it has been argued that other studies overestimate the relative utility of biomarkers by examining homogeneous or highly selected samples or by using inappropriate endpoints.
The increased levels of scrutiny regarding novel biomarkers is fed by the results presented based on ever increasing statistically significant measures. What may be relevant to clinical care, however, is not whether changes in predicted probabilities are statistically significant but whether application of these identified biomarkers results in reclassification of individuals to new, clinically related risk categories. Breath analysis might provide the means to differentiate on a physiological basis between clinically related disease profiles or sub-stages.
To date several studies regarding the analysis of exhaled air related to disease suggest that breath testing might potentially be valuable in clinical practice because its accuracy proved superior to the reported sensitivity and specificity of current methodologies.[3,24]
The challenge will be to find and validate those biomarkers related to disease able to individually or in combination with existing biomarkers bring about improvements in risk assessment that are not just statistically significant but clinically significant as well. An interesting part of the problem might be solved as discussed and shown in this thesis by implementation of advanced classifiers. Application of the classification algorithms has revealed patterns of breath VOCs that are highly distinctive to constitute a certain VOC profile linked to a disease(status).[15,20,22–24,62,66,90] The superior SVM approach demonstrated to outperform other available unsupervised learning methods for its ability to construct predictive models with large generalization power even in the case of large dimensionality of the data when the number of observations available for training is low,[22,23] which is obviously the case in exhaled air analysis. SVMs are specifically useful since they seek a globally optimized solution and avoid over-fitting, so a large number of features or compounds is allowed.[69]

Conclusion

Breath analysis holds the promise to be of huge interest in clinical practice. It is proven in this thesis that in some disease profiles it is able to perform as a quality biomarker, being a sensitive and non-invasive methodology. However large (prospective) cohort studies are necessary in order to validate the selected biomarkers linked to the different disease profiles. Future research regarding more specific and highly sensitive sensors will provide the means for these biomarkers to be highly cost effective and very simple to use.[83,85] A wide range of sensors is currently under research including metal-oxide sensors and polymer-based sensors, both demonstrating resistance variations related to absorption of specific VOCs and the surface acoustic wave technology demonstrating acoustic variations according to absorption of specific VOCs. It is

however unlikely the currently applied sensors will, even after fine-tuning and incorporation of all necessary adaptations, demonstrate high sensitivities regarding individual VOCs. A more likely-to-work approach is detection of the disease-related VOCs by means of miniaturized GCs in line with highly sensitive en specific sensors (for example MS). Improved sensitivity and specificity and subsequent implementation of these technologies into small 'fool-proof' handheld devices combined with advanced signal processing modules and an easy-to-apply breath collection procedure without the need to correct for all confounding factors will bring breath analysis right into clinical practice.

CHAPTER 8

Summary

Introduction

Technical advances in analytical analysis during the last few decades have been responsible for the recent developmental improvements in diagnostics and partial understanding of metabolic and biological pathways. This could lead to the discovery of new biomarkers able to characterize and identify disease. This thesis aims to describe our efforts regarding design and validation of a diagnostic tool based on the analysis of exhaled air. Exhaled air contains a complex mixture of volatile organic compounds (VOCs), some of which could potentially be applied as biomarkers for lung diseases. To date only single markers in breath are used in clinical diagnostics lacking the degree of specificity and sensitivity in order for the methodology to be taken in clinical practice.
The rapid, accurate and non-invasive diagnosis of respiratory disease represents a challenge to clinicians while the development of new treatments can be confounded by insufficient knowledge of lung disease phenotypes. We have developed a sampling methodology for collecting concentrated samples of exhaled air from patients suffering from COPD and cystic fibrosis against which we employed two-stage thermal desorption gas chromatography-time-of flight mass spectrometry (GC-MS) analysis. Data analysis tools have been developed enabling pipeline analysis of the generated GC-MS sample outputs. Informative VOCs were extracted from the compiled database and implemented into a classifier of which performance was evaluated.

Summary

Chapter 2 describes this developed methodology regarding the analysis of exhaled air. Both the instrumental setup and software are discussed, as well as the statistical analysis. Thermal desorption and gas chromatography coupled to time-of-flight mass spectrometry were used to analyze exhaled air samples. The VOC profiles obtained from each individual were combined into one final database based on similarity of mass spectra and retention indexes, which offers the possibility for a reliable selection of compounds of interest. The developed methodology was validated on a set of smokers and non-smokers. Support vector machine analysis identified 4 VOCs as biomarkers of recent exposure to cigarette smoke. Two of the selected VOCs were already mentioned in literature as possibly smoking behavior related compounds confirming the validity of the setup. After validation of the procedure a large study was setup to identify VOCs in bacterial headspace discriminating between different micro organisms. **Chapter 3** discusses the setup and results from the micro organism headspace analysis study. Different sets of VOCs were found enabling easy and fast identification of the different cultures. Several selected VOCs were previously mentioned in literature regarding bacterial headspace. We were able to identify a large number of compounds demonstrating a highly significant difference in availability in bacterial cultures compared to medium and in cul-

tures compared to one another. We have also determined highly significant compounds differing between the four *Escherichia coli* strains and between the two *Staphylococcus aureus* isolates: methicillin-resistant *Staphylococcus aureus* and methicillin-sensitive *Staphylococcus aureus*. SVM models were able to classify the microorganisms with very high degrees of sensitivity and specificity based on -on average- 6 VOCs from headspace. **Chapter 4** describes the developed methodology and search for biomarkers regarding chronic obstructive pulmonary disease (COPD). Fifty COPD patients en twenty-nine controls were sampled and VOCs were analyzed by gas chromatography-mass spectrometry to identify relevant VOCs. A classifier based on six VOCs correctly classified 92% of the subjects. Additionally several validation data sets were build and performance of the classifier was evaluated. **Chapter 5** investigates whether analysis of VOCs from exhaled air could discriminate between subject suffering from cystic fibrosis and controls, and between cystic fibrosis patients with and without *Pseudomonas aeruginosa* (*P. aeruginosa*) colonization. Statistical analysis identified 10 VOCs that combined into a SVM model correctly classified 92% of samples. This new technique was not only able to discriminate between cystic fibrosis patients and controls, but also between cystic fibrosis patients with and without *P. aeruginosa* colonization. **Chapter 6** emphasizes on application of the developed methodological exhaled air analysis technique on exacerbations in patients with cystic fibrosis. Prevention of exacerbations in cystic fibrosis is important since these pulmonary exacerbations are an important cause for the hospitalization of patients, respiratory symptoms, and decreases in lung function. In the described longitudinal study we studied whether changes in VOC profiles can predict pulmonary exacerbations of cystic fibrosis. Cystic fibrosis patients were sampled at 2-month intervals. Four extra samples where collected in case of an exacerbation. It appeared that 18 VOCs were differentially present in exhaled breath from patients not having an exacerbation versus patients suffering an exacerbation event, 15 VOCs were differentially present in exhaled breath from the comparison of baseline measurements to the first exacerbation measurement at the start of the exacerbation event. The study demonstrated that VOCs in exhaled breath are able to indicate cystic fibrosis exacerbations weeks before these adverse events are clinically manifest. Before these biomarkers can be of any clinical relevance however, the different classifiers need to be validated on external validation sets. **Chapter 7** presents the general discussion pointing out a methodological overview of the presented breath analysis technique and its shortcomings. It discusses the general biomarker properties and projects these onto the exhaled air biomarkers. Additionally several issues regarding the analysis and the biomarkers from exhaled air are described. Finally prospectives - regarding COPD heterogeneity and CF exacerbations - and perspectives are presented in short.

CHAPTER 9

Samenvatting

Introductie

De door technische en wetenschappelijke vooruitgang gedreven optimalisatie van analytische technieken heeft de laatste decennia geleid tot grote verbeteringen in de medische diagnostiek en de beschrijving van metabole pathways. Deze thesis beschrijft de ontwikkeling en validatie van een diagnostische tool gebaseerd op de analyse van componenten aanwezig in de uitademingslucht. Uitademingslucht is een complex mengsel van onder andere stikstof, zuurstof, koolstofdioxide en een groot aantal verschillende vluchtige organische verbindingen (VOCs). De huidige diagnostiek op basis van uitademingslucht is voornamelijk gebaseerd op aan- of afwezigheid van individuele componenten al dan niet aanwezig in de uitademingslucht. De lage sensitiviteit en specificiteit echter van deze huidige methoden vormt een drempel voor grootschalige toepassing van uitademingslucht analyses in de medische diagnostiek. Deze thesis geeft een beschrijving van de ontwikkeling en validatie van een meer nauwkeurige diagnostische tool op basis van uitademingslucht analyses. Er wordt zowel aandacht besteedt aan de manier waarop monsters verzameld worden en de chemische analyse alsmede de data-analyse en selectie van informatieve vluchtige organische verbindingen.

Samenvatting

Hoofdstuk 2 beschrijft de ontwikkelde methodologie van analyse van uitademingslucht. Zowel de instrumentele setup als de software en toegepaste statistiek worden uitgebreid behandeld. De analyse technieken zoals hier toegepast zijn gebaseerd op thermische desorptie en gas chromatografische scheiding van de aanwezige componenten. Uiteindelijke identificatie van de componenten werd uitgevoerd met massa spectrometrie (MS) in de vorm van time-of-flight MS. Na diverse data voorbewerkings stappen, als het verminderen van de ruis en uitvoeren van piekdetectie, zijn de overeenkomstige componenten aanwezig in de verschillende monsters vervolgens op basis van overeenkomst in rententietijd en massaspectrum aan elkaar gekoppeld. Dit resulteert in een database waarin alle in de uitademingslucht aanwezige componenten zijn opgenomen gekoppeld aan de mate van aanwezigheid in ieder monster. De voorgestelde methodologie is vervolgens getoetst op een populatie rokers/niet-rokers waarin informatieve componenten zijn geselecteerd. Vervolgens zijn classificatie algoritmen gebruikt om informatieve componenten te selecteren in de gegenereerde dataset. Het resulterende classificatie model gebaseerd op slecht 4 componenten was is staat om alle monsters correct te classificeren. De geïmplementeerde componenten waren reeds in de literatuur gerelateerd aan rookgedrag. Nadat de methodologie op deze manier is gevalideerd is een studie gestart om de onderscheidende VOCs van bacterie culturen te bepalen. **Hoofdstuk 3** beschrijft deze studie, waarbij de bovenstaande lucht van verschillende bacterie culturen in groei zijn gemeten. Het grote aantal verschillende VOCs aanwezig in die culturen zijn volgens de in hoofdstuk 2 uitgebreid beschreven protocol in een database onderge-

bracht en met de diverse statistische tools geanalyseerd. Zowel supervised als un-supervised learning algoritmen zijn toegepast om de data te verkennen. De verschillende culturen hadden een groot aantal in hoeveelheid significant verschillende componenten in de headspace aanwezig, wat differentiatie met SVM modellen tussen de culturen mogelijk maakte op basis van slechts een beperkt aantal componenten. Zelfs de 4 verschillende *Escherichia coli* stammen waren op basis van componenten aanwezig in de headspace te onderscheiden evenals de twee *Staphylococcus aureus* isolaten: methicilline-resistente *Staphylococcus aureus* en de methicilline-sensitieve *Staphylococcus aureus*. **Hoofdstuk 4** beschrijft toepassing van de ontwikkelde uitademingslucht-analyse methodologie op de inflammatoire luchtwegaandoening COPD. Uitademingslucht van 50 COPD patiënten en 29 niet-COPD personen werd verzameld en geanalyseerd. Een SVM classificatie model gebaseerd op slechts 6 componenten uit de uitademingslucht classificeerde 92 % van de monsters correct. Diverse validatie sets zijn vervolgens samengesteld, bestaande uit steroid-naive COPD patiënten om de invloed van medicatie te bestuderen en een externe validatie set bestaande uit monsters verkregen op andere locaties dan het Academisch Ziekenhuis Maastricht om de invloed van omgevingslucht nader te bestuderen. Classificatie van monsters in deze validatiesets resulteerde eveneens in een correcte classificatie van 92 %. De toepassing van de ontwikkelde methodologie in een klinische setting werd in **Hoofdstuk 5** voortgezet. Hier wordt gekeken naar de mogelijkheid patiënten met cystic fibrosis te onderscheiden van controle personen op basis van VOCs in de uitademingslucht. Statistische analyse resulteerde in een classificatie model gebaseerd op 10 VOCs in staat om 92% van de monsters correct te classificeren. Het bleek eveneens mogelijk om patiënten met en zonder *Pseudomonas aeruginosa* infectie te onderscheiden. In **Hoofdstuk 6** is gekeken naar de mogelijkheid om exacerbaties in patiënten met cystic fibrosis te voorspellen en te monitoren op basis van componenten aanwezig in de uitademingslucht. Het voorspellen en door tijdig handelen voorkomen of verminderen van exacerbaties is van groot belang omdat exacerbaties een belangrijke oorzaak vormen voor afname van longfunctie en kan leiden tot hospitalisatie van patiënten. Patiënten zijn met intervallen van 2 maanden bemonsterd en tijdens een exacerbatie werden 4 extra monsters afgenomen. Statistische analyse wees uit dat 18 componenten in significant verschillende hoeveelheden aanwezig waren in uitademingslucht van patiënten die 3 tot 8 weken na de meting een exacerbatie doormaakten uitgezet tegen patiënten die geen exacerbatie doormaakten. Vijftien componenten waren in significant verschillende hoeveelheden aanwezig in uitademingslucht van patiënten vóór vergeleken met tijdens een exacerbatie. Daarnaast zijn met SVM analyse de meest informatieve componenten geselecteerd welke gecombineerd in en classificatie model 93% van de exacerbatie monsters correct konden classificeren. De studie toont aan dat markers in uitademingslucht het mogelijk maken om exacerbaties bij patiënten met cystic fibrosis weken voor de daadwerkelijk exacerbatie te voorspellen. Er dient echter rekening mee gehouden te worden dat de totale patïentenpopulatie beperkt was en validatie van de resultaten noodzakelijk zal zijn. Bijvoorbeeld door de modellen en componenten te toetsen in een

grotere validatie studie. Ten slotte geeft **Hoofdstuk 7** een beschrijving van de algemene discussie waarbij de nadruk ligt op de gekozen randvoorwaarden van de beschreven methodologie gebaseerd op analyse van uitademingslucht en de huidige tekortkomingen van deze methodologie.

CHAPTER 10

Curriculum vitae

About the author

Joep van Berkel was born September 20th 1979. After completing secondary education in 1996 at the 'Zwijsen College' in Veghel, he went to Eindhoven to study biomedical engineering at the Technical University. After several extensive internships including projects at the Department of Physics (12 months) as at Terumo Cardiovascular Systems in Tustin, LA, California (12 months) he obtained his masters degree in 2005, and started as a PhD-candidate at the Department of Health Risk Analysis and Toxicology at the Maastricht University. During that time he also obtained licenses for working with laboratory animals (Art. 9 according to Dutch law) and among other courses he was trained on advanced datamining and bioinformatics. He was awarded best scientific presentation at the annual meeting of the Dutch society of toxicology in 2009. Since september 2009 he is continuing his research as a postdocteral fellow on the analysis of breath at the department of Health Risk Analysis and Toxicology of Maastricht University. Taking part in extensive collaborations with research groups abroad regarding the analysis of exhaled air.

List of publications

- **Radicals of plasma needle detected with fluorescent probe**
  I.E. Kieft, J.J.B.N. van Berkel, E.R. Kieft, E. Stoffels, Radicals of plasma needle detected with fluorescent probe in *Plasma Processes and Polymers;* Editors: R. d'Agostino, P. Favia, C. Oehr, and M.R. Wertheimer, 295-308, Wiley-VCH, Book Chapter ISBN 3-527-40487-2 (2005)

- **Metabolomics of volatile organic compounds in cystic fibrosis patients and controls.**
  Robroeks CM, van Berkel JJ, Dallinga JW, Jöbsis Q, Zimmermann LJ, Hendriks HJ, Wouters MF, van der Grinten CP, van de Kant KD, van Schooten FJ, Dompeling E.
  Pediatr Res. 2010 Mar 26.

- **Volatile organic compounds in exhaled breath as a diagnostic tool for asthma in children.**
  Dallinga JW, Robroeks CM, van Berkel JJ, Moonen EJ, Godschalk RW, Jöbsis Q, Dompeling E, Wouters EF, van Schooten FJ.
  Clin Exp Allergy. 2010 Jan;40(1):68-76. Epub 2009 Sep 28.

- **Development of accurate classification method based on the analysis of volatile organic compounds from human exhaled air.**
  Van Berkel JJ, Dallinga JW, Möller GM, Godschalk RW, Moonen E, Wouters EF, Van Schooten FJ.
  J Chromatogr B Analyt Technol Biomed Life Sci. 2008 Jan 1;861(1):101-7. Epub 2007 Nov 19.

- **A profile of volatile organic compounds in breath discriminates COPD patients from controls.**

Van Berkel JJ, Dallinga JW, Möller GM, Godschalk RW, Moonen EJ, Wouters EF, Van Schooten FJ.
Respir Med. 2010 Apr;104(4):557-63. Epub 2009 Nov 10.

- **Identification of microorganisms based on gas chromatographic-mass spectrometric analysis of volatile organic compounds in headspace gases.** (submitted)
Van Berkel JJ, Stobberingh EE, Boumans MLL, Moonen EJ, Wouters EFM, Dallinga JW, Van Schooten FJ.

- **Prediction of exacerbations in cystic fibrosis in a one-year longitudinal study based on volatile organic compounds in exhaled breath.** (submitted)
Van Berkel JJ, Robroeks CMHHT, Dallinga JW, Jöbsis Q, Moonen EJ, Wouters EFM, Dompeling E, Van Schooten FJ.

- **Volatile organic compounds in exhaled breath predict exacerbations of childhood asthma in a one year prospective controlled study.** (submitted)
Robroeks CMHHT, van Berkel JJ, Jöbsis Q, van Schooten FJ, Dallinga JW , Wouters EFM, Dompeling E.

Abstracts

- **Design of accurate COPD classification model based on volatile organic compounds from exhaled air.** J. van Berkel, J. Dallinga, G. Mller, R. Godschalk, E. Wouters, F. J. van Schooten (Maastricht, Netherlands), ERS Berlin 2008.

- **Development of accurate classification method based on the analysis of volatile organic compounds from human exhaled air.** J.J.B.N. van Berkel, J.W. Dallinga, R.W.L. Godschalk, E.F.M. Wouters, F.J. van Schooten (Maastricht, The Netherlands), ATS Toronto 2008.

- **Analysis of exhaled breath and its value in prediction of exacerbations in children suffering from CF.** J. van Berkel, C. Robroeks, J. Dallinga, Q. Jobsis, E. Moonen, E. Wouters, E. Dompeling, F. J. van Schooten (Maastricht, Netherlands), ERS Vienna 2009.

- **VOCs in exhaled breath are associated with exacerbations in children with asthma.** J. van Berkel, C. Robroeks, J. Dallinga, Q. Jobsis, E. Moonen, E. Wouters, F. J. van Schooten, E. Dompeling (Maastricht, Netherlands), ERS Vienna 2009.

- **A profile of volatile organic compounds in breath discriminates copd patients from controls.** J.J.B.N. van Berkel, J.W. Dallinga, G.M. Moller, R.W.L. Godschalk, E.J. Moonen, E.F.M. Wouters, F.J. van Schooten (Maastricht, The Netherlands), ATS San Diego 2009.

- **Analysis of volatile organic compounds in exhaled breath as a diagnostic tool for asthma in children.** J.W. Dallinga, M.H.H.T. Robroeks, J.J.B.N. van Berkel, E.J. Moonen, R.W.L. Godschalk, Q. Jobsis, E. Dompeling, E.F.M. Wouters, F.J. van Schooten (Maastricht, The Netherlands), ATS San Diego 2009.
- **Volatile organic compounds in breath completely discriminate COPD patients from controls.** J.J.B.N. van Berkel, J.W. Dallinga, G.M. Moller, R.W.L. Godschalk, E.J. Moonen, E.F.M. Wouters, F.J. van Schooten (Maastricht, The Netherlands), NVT-dagen 2009.

Patents

- **Method for diagnosis of chronic obstructive pulmonary disease by detecting volatile organic compounds in exhaled air.**
  Application number: 08164535.0-1223
  J. van Berkel, J. Dallinga, R. Godschalk, E. Wouters, F. J. van Schooten.
- **Method for the differentiation between Crohn's disease and ulcerative colitis by detecting volatile organic compounds in exhaled air.**
  Application number: 09173487.1-2404
  J. van Berkel, J. Dallinga, F. J. van Schooten.
- **Method for the differentiation between patients with ulcerative colitis and patients in remission by detecting volatile organic compounds in exhaled air.**
  Application number: 09173489.7-2404
  J. van Berkel, J. Dallinga, F. J. van Schooten.
- **Method for diagnosis of inflammatory bowel disease by detecting volatile organic compounds in exhaled air.**
  Application number: 09173486.3-2404
  J. van Berkel, J. Dallinga, F. J. van Schooten.
- **Method for the differentiation between patients with active Crohn's disease and and patients in remission by detecting volatile organic compounds in exhaled air.**
  Application number: 09173488.9-2404
  J. van Berkel, J. Dallinga, F. J. van Schooten.

Bibliography

[1] W. Ma, X. Liu, and J. Pawliszyn. Analysis of human breath with micro extraction techniques and continuous monitoring of carbon dioxide concentration. *Anal Bioanal Chem*, 385(8):1398–408, 2006.

[2] M. Libardoni, P. T. Stevens, J. H. Waite, and R. Sacks. Analysis of human breath samples with a multi-bed sorption trap and comprehensive two-dimensional gas chromatography (gcxgc). *J Chromatogr B Analyt Technol Biomed Life Sci*, 842(1):13–21, 2006.

[3] W. Miekisch, J. K. Schubert, and G. F. Noeldge-Schomburg. Diagnostic potential of breath analysis–focus on volatile organic compounds. *Clin Chim Acta*, 347(1-2):25–39, 2004.

[4] A. Moeller, C. Diefenbacher, A. Lehmann, M. Rochat, J. Brooks-Wildhaber, G. L. Hall, and J. H. Wildhaber. Exhaled nitric oxide distinguishes between subgroups of preschool children with respiratory symptoms. *J Allergy Clin Immunol*, 121(3):705–9, 2008.

[5] R. Katial and L. Stewart. Exhaled nitric oxide: a test for diagnosis and control of asthma? *Curr Allergy Asthma Rep*, 7(6):459–63, 2007.

[6] S. Turner. Exhaled nitric oxide in the diagnosis and management of asthma. *Curr Opin Allergy Clin Immunol*, 8(1):70–6, 2008.

[7] C. Brindicci, K. Ito, O. Resta, N. B. Pride, P. J. Barnes, and S. A. Kharitonov. Exhaled nitric oxide from lung periphery is increased in copd. *Eur Respir J*, 26(1):52–9, 2005.

[8] M. Yamaya, K. Sekizawa, S. Ishizuka, M. Monma, and H. Sasaki. Exhaled carbon monoxide levels during treatment of acute asthma. *Eur Respir J*, 13(4): 757–60, 1999.

[9] P. Montuschi, S. A. Kharitonov, and P. J. Barnes. Exhaled carbon monoxide and nitric oxide in copd. *Chest*, 120(2):496–501, 2001.

[10] I. Horvath, W. MacNee, F. J. Kelly, P. N. Dekhuijzen, M. Phillips, G. Doring, A. M. Choi, M. Yamaya, F. H. Bach, D. Willis, L. E. Donnelly, K. F. Chung, and P. J. Barnes. "haemoxygenase-1 induction and exhaled markers of oxidative

stress in lung diseases", summary of the ers research seminar in budapest, hungary, september, 1999. *Eur Respir J*, 18(2):420–30, 2001.

[11] M. B. Schleiss, O. Holz, M. Behnke, K. Richter, H. Magnussen, and R. A. Jorres. The concentration of hydrogen peroxide in exhaled air depends on expiratory flow rate. *Eur Respir J*, 16(6):1115–8, 2000.

[12] S. A. Kharitonov and P. J. Barnes. Effects of corticosteroids on noninvasive biomarkers of inflammation in asthma and chronic obstructive pulmonary disease. *Proc Am Thorac Soc*, 1(3):191–9, 2004.

[13] M. Barker, M. Hengst, J. Schmid, H. J. Buers, B. Mittermaier, D. Klemp, and R. Koppmann. Volatile organic compounds in the exhaled breath of young patients with cystic fibrosis. *Eur Respir J*, 27(5):929–36, 2006.

[14] M. Phillips, N. Altorki, J. H. Austin, R. B. Cameron, R. N. Cataneo, J. Greenberg, R. Kloss, R. A. Maxfield, M. I. Munawar, H. I. Pass, A. Rashid, W. N. Rom, and P. Schmitt. Prediction of lung cancer using volatile biomarkers in breath. *Cancer Biomark*, 3(2):95–109, 2007.

[15] M. Phillips, R. N. Cataneo, A. R. Cummin, A. J. Gagliardi, K. Gleeson, J. Greenberg, R. A. Maxfield, and W. N. Rom. Detection of lung cancer with volatile markers in the breath. *Chest*, 123(6):2115–23, 2003.

[16] C. Deng, J. Zhang, X. Yu, W. Zhang, and X. Zhang. Determination of acetone in human breath by gas chromatography-mass spectrometry and solid-phase microextraction with on-fiber derivatization. *J Chromatogr B Analyt Technol Biomed Life Sci*, 810(2):269–75, 2004.

[17] B. G. Stone, T. J. Besse, W. C. Duane, C. D. Evans, and E. G. DeMaster. Effect of regulating cholesterol biosynthesis on breath isoprene excretion in men. *Lipids*, 28(8):705–8, 1993.

[18] F. Di Francesco, R. Fuoco, M. G. Trivella, and A. Ceccarini. Breath analysis: trends in techniques and clinical applications. *Microchemical Journal*, 79(1-2): 405–410, 2005.

[19] B. J. Novak, D. R. Blake, S. Meinardi, F. S. Rowland, A. Pontello, D. M. Cooper, and P. R. Galassetti. Exhaled methyl nitrate as a noninvasive marker of hyperglycemia in type 1 diabetes. *Proc Natl Acad Sci U S A*, 104(40):15613–8, 2007.

[20] J. W. Dallinga, C. M. Robroeks, J. J. van Berkel, E. J. Moonen, R. W. Godschalk, Q. Jobsis, E. Dompeling, E. F. Wouters, and F. J. van Schooten. Volatile organic compounds in exhaled breath as a diagnostic tool for asthma in children. *Clin Exp Allergy*, 2009.

[21] R. F. Machado, D. Laskowski, O. Deffenderfer, T. Burch, S. Zheng, P. J. Mazzone, T. Mekhail, C. Jennings, J. K. Stoller, J. Pyle, J. Duncan, R. A. Dweik, and S. C. Erzurum. Detection of lung cancer by sensor array analyses of exhaled breath. *Am J Respir Crit Care Med*, 171(11):1286–91, 2005.

[22] J. J. Van Berkel, J. W. Dallinga, G. M. Moller, R. W. Godschalk, E. J. Moonen, E. F. Wouters, and F. J. Van Schooten. A profile of volatile organic compounds in breath discriminates copd patients from controls. *Respir Med*, 104(4):557–63, 2010.

[23] J. J. Van Berkel, J. W. Dallinga, G. M. Moller, R. W. Godschalk, E. Moonen, E. F. Wouters, and F. J. Van Schooten. Development of accurate classification method based on the analysis of volatile organic compounds from human

exhaled air. *J Chromatogr B Analyt Technol Biomed Life Sci*, 861(1):101–7, 2008.

[24] M Phillips. Breath collection apparatus, 2001.

[25] S. R. Rutgers, D. S. Postma, N. H. ten Hacken, H. F. Kauffman, T. W. van Der Mark, G. H. Koeter, and W. Timens. Ongoing airway inflammation in patients with copd who do not currently smoke. *Thorax*, 55(1):12–8, 2000.

[26] M. Saetta, A. Di Stefano, G. Turato, F. M. Facchini, L. Corbino, C. E. Mapp, P. Maestrelli, A. Ciaccia, and L. M. Fabbri. Cd8+ t-lymphocytes in peripheral airways of smokers with chronic obstructive pulmonary disease. *Am J Respir Crit Care Med*, 157(3 Pt 1):822–6, 1998.

[27] A. G. Agusti, A. Noguera, J. Sauleda, E. Sala, J. Pons, and X. Busquets. Systemic effects of chronic obstructive pulmonary disease. *Eur Respir J*, 21(2): 347–60, 2003.

[28] W. Q. Gan, S. F. Man, A. Senthilselvan, and D. D. Sin. Association between chronic obstructive pulmonary disease and systemic inflammation: a systematic review and a meta-analysis. *Thorax*, 59(7):574–80, 2004.

[29] D. O. Haskard. Accelerated atherosclerosis in inflammatory rheumatic diseases. *Scand J Rheumatol*, 33(5):281–92, 2004.

[30] D. D. Sin and J. Vestbo. Biomarkers in chronic obstructive pulmonary disease. *Proc Am Thorac Soc*, 6(6):543–5, 2009.

[31] R. Garrod, J. Marshall, E. Barley, S. Fredericks, and G. Hagan. The relationship between inflammatory markers and disability in chronic obstructive pulmonary disease (copd). *Prim Care Respir J*, 16(4):236–40, 2007.

[32] D. M. Mannino and A. S. Buist. Global burden of copd: risk factors, prevalence, and future trends. *Lancet*, 370(9589):765–73, 2007.

[33] N. R. Macintyre. Spirometry for the diagnosis and management of chronic obstructive pulmonary disease. *Respir Care*, 54(8):1050–7, 2009.

[34] G. L. Snider. Molecular epidemiology: a key to better understanding of chronic obstructive lung disease. *Monaldi Arch Chest Dis*, 50(1):3–6, 1995.

[35] M. W. Konstan and M. Berger. Current understanding of the inflammatory process in cystic fibrosis: onset and etiology. *Pediatr Pulmonol*, 24(2):137–42; discussion 159–61, 1997.

[36] T. L. Bonfield, M. W. Konstan, and M. Berger. Altered respiratory epithelial cell cytokine production in cystic fibrosis. *J Allergy Clin Immunol*, 104(1):72–8, 1999.

[37] T. L. Bonfield, J. R. Panuska, M. W. Konstan, K. A. Hilliard, J. B. Hilliard, H. Ghnaim, and M. Berger. Inflammatory cytokines in cystic fibrosis lungs. *Am J Respir Crit Care Med*, 152(6 Pt 1):2111–8, 1995.

[38] I. M. Balfour-Lynn and R. Dinwiddie. Role of corticosteroids in cystic fibrosis lung disease. *J R Soc Med*, 89 Suppl 27:8–13, 1996.

[39] P. Birrer, N. G. McElvaney, A. Rudeberg, C. W. Sommer, S. Liechti-Gallati, R. Kraemer, R. Hubbard, and R. G. Crystal. Protease-antiprotease imbalance in the lungs of children with cystic fibrosis. *Am J Respir Crit Care Med*, 150 (1):207–13, 1994.

[40] J. Emerson, M. Rosenfeld, S. McNamara, B. Ramsey, and R. L. Gibson. Pseudomonas aeruginosa and other predictors of mortality and morbidity in young children with cystic fibrosis. *Pediatr Pulmonol*, 34(2):91–100, 2002.

[41] T. G. Liou, F. R. Adler, and B. C. Cahill. Testing lung function decline to time lung transplantation. *Chest*, 128(1):472–3; author reply 473–4, 2005.

[42] O. H. Mayer, A. F. Jawad, J. McDonough, and J. Allen. Lung function in 3-5-year-old children with cystic fibrosis. *Pediatr Pulmonol*, 43(12):1214–23, 2008.

[43] C. Dakin, R. L. Henry, P. Field, and J. Morton. Defining an exacerbation of pulmonary disease in cystic fibrosis. *Pediatr Pulmonol*, 31(6):436–42, 2001.

[44] B. C. Marshall. Pulmonary exacerbations in cystic fibrosis: it's time to be explicit! *Am J Respir Crit Care Med*, 169(7):781–2, 2004.

[45] P. Paredi, S. A. Kharitonov, D. Leak, S. Ward, D. Cramer, and P. J. Barnes. Exhaled ethane, a marker of lipid peroxidation, is elevated in chronic obstructive pulmonary disease. *Am J Respir Crit Care Med*, 162(2 Pt 1):369–73, 2000.

[46] S. Kanoh, H. Kobayashi, and K. Motoyoshi. Exhaled ethane: an in vivo biomarker of lipid peroxidation in interstitial lung diseases. *Chest*, 128(4): 2387–92, 2005.

[47] L. Pauling, A. B. Robinson, R. Teranishi, and P. Cary. Quantitative analysis of urine vapor and breath by gas-liquid partition chromatography. *Proc Natl Acad Sci U S A*, 68(10):2374–6, 1971.

[48] M. Phillips, R. N. Cataneo, R. Condos, G. A. Ring Erickson, J. Greenberg, V. La Bombardi, M. I. Munawar, and O. Tietje. Volatile biomarkers of pulmonary tuberculosis in the breath. *Tuberculosis (Edinb)*, 87(1):44–52, 2007.

[49] J. P. Spinhirne, J. A. Koziel, and N. K. Chirase. Sampling and analysis of volatile organic compounds in bovine breath by solid-phase microextraction and gas chromatography-mass spectrometry. *J Chromatogr A*, 1025(1):63–9, 2004.

[50] P. J. Mazzone, J. Hammel, R. Dweik, J. Na, C. Czich, D. Laskowski, and T. Mekhail. Diagnosis of lung cancer by the analysis of exhaled breath with a colorimetric sensor array. *Thorax*, 62(7):565–8, 2007.

[51] J. Taucher, A. Hansel, A. Jordan, R. Fall, J. H. Futrell, and W. Lindinger. Detection of isoprene in expired air from human subjects using proton-transfer-reaction mass spectrometry. *Rapid Commun Mass Spectrom*, 11(11):1230–4, 1997.

[52] T. Karl, P. Prazeller, D. Mayr, A. Jordan, J. Rieder, R. Fall, and W. Lindinger. Human breath isoprene and its relation to blood cholesterol levels: new measurements and modeling. *J Appl Physiol*, 91(2):762–70, 2001.

[53] J. Rieder, P. Lirk, C. Ebenbichler, G. Gruber, P. Prazeller, W. Lindinger, and A. Amann. Analysis of volatile organic compounds: possible applications in metabolic disorders and cancer screening. *Wien Klin Wochenschr*, 113(5-6): 181–5, 2001.

[54] Sian M. Abbott, James B. Elder, Patrik Spanel, and David Smith. Quantification of acetonitrile in exhaled breath and urinary headspace using selected ion flow tube mass spectrometry. *International Journal of Mass Spectrometry*, 228 (2-3):655–665, 2003.

[55] C. Di Natale, A. Macagnano, E. Martinelli, R. Paolesse, G. D'Arcangelo, C. Roscioni, A. Finazzi-Agro, and A. D'Amico. Lung cancer identification by the analysis of breath by means of an array of non-selective gas sensors. *Biosens Bioelectron*, 18(10):1209–18, 2003.

[56] H. Lord, Y. Yu, A. Segal, and J. Pawliszyn. Breath analysis and monitoring by membrane extraction with sorbent interface. *Anal Chem*, 74(21):5650–7, 2002.

[57] J. E. Szulejko, M. McCulloch, J. Jackson, D. L. McKee, J. C. Walker, and T. Solouki. Evidence for cancer biomarkers in exhaled breath. *Ieee Sensors Journal*, 10(1):185–210, 2010.

[58] W. Cao and Y. Duan. Breath analysis: potential for clinical diagnosis and exposure assessment. *Clin Chem*, 52(5):800–11, 2006.

[59] L. T. McGrath, R. Patrick, P. Mallon, L. Dowey, B. Silke, W. Norwood, and S. Elborn. Breath isoprene during acute respiratory exacerbation in cystic fibrosis. *Eur Respir J*, 16(6):1065–9, 2000.

[60] C. M. Kneepkens, C. Ferreira, G. Lepage, and C. C. Roy. The hydrocarbon breath test in the study of lipid peroxidation: principles and practice. *Clin Invest Med*, 15(2):163–86, 1992.

[61] S. M. Gordon, L. A. Wallace, M. C. Brinkman, P. J. Callahan, and D. V. Kenny. Volatile organic compounds as breath biomarkers for active and passive smoking. *Environ Health Perspect*, 110(7):689–98, 2002.

[62] M. Phillips, K. Gleeson, J. M. Hughes, J. Greenberg, R. N. Cataneo, L. Baker, and W. P. McVay. Volatile organic compounds in breath as markers of lung cancer: a cross-sectional study. *Lancet*, 353(9168):1930–3, 1999.

[63] S. E. Stein and D. R. Scott. Optimization and testing of mass-spectral library search algorithms for compound identification. *Journal of the American Society for Mass Spectrometry*, 5(9):859–866, 1994.

[64] E. Frank, M. Hall, L. Trigg, G. Holmes, and I. H. Witten. Data mining in bioinformatics using weka. *Bioinformatics*, 20(15):2479–81, 2004.

[65] J.C. Platt. *Fast training of support vector machines using sequential minimal optimization*. Advances in kernel methods - support vector learning. MIT press, 1998.

[66] J. M. Sanchez and R. D. Sacks. Development of a multibed sorption trap, comprehensive two-dimensional gas chromatography, and time-of-flight mass spectrometry system for the analysis of volatile organic compounds in human breath. *Anal Chem*, 78(9):3046–54, 2006.

[67] L. Wallace. Environmental exposure to benzene: an update. *Environ Health Perspect*, 104 Suppl 6:1129–36, 1996.

[68] D. J. Penn, E. Oberzaucher, K. Grammer, G. Fischer, H. A. Soini, D. Wiesler, M. V. Novotny, S. J. Dixon, Y. Xu, and R. G. Brereton. Individual and gender fingerprints in human body odour. *J R Soc Interface*, 4(13):331–40, 2007.

[69] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422, 2002.

[70] S. M. Gordon, J. P. Szidon, B. K. Krotoszynski, R. D. Gibbons, and H. J. O'Neill. Volatile organic compounds in exhaled air from patients with lung cancer. *Clin Chem*, 31(8):1278–82, 1985.

[71] L. Wallace, T. Buckley, E. Pellizzari, and S. Gordon. Breath measurements as volatile organic compound biomarkers. *Environ Health Perspect*, 104 Suppl 5: 861–9, 1996.

[72] S. Maddula, L. M. Blank, A. Schmid, and J. I. Baumbach. Detection of volatile metabolites of escherichia coli by multi capillary column coupled ion mobility spectrometry. *Anal Bioanal Chem*, 394(3):791–800, 2009.

[73] C. S. Probert, P. R. Jones, and N. M. Ratcliffe. A novel method for rapidly diagnosing the causes of diarrhoea. *Gut*, 53(1):58–61, 2004.

[74] G. Preti, E. Thaler, C. W. Hanson, M. Troy, J. Eades, and A. Gelperin. Volatile compounds characteristic of sinus-related bacteria and infected sinus mucus: analysis by solid-phase microextraction and gas chromatography-mass spectrometry. *J Chromatogr B Analyt Technol Biomed Life Sci*, 877(22):2011–8, 2009.

[75] M. Syhre and S. T. Chambers. The scent of mycobacterium tuberculosis. *Tuberculosis (Edinb)*, 88(4):317–23, 2008.

[76] W. Hansen, Y. Glupczynski, J. C. Lemper, and E. Yourassowsky. Infective endocarditis due to actinobacillus actinomycetemcomitans. *Acta Clin Belg*, 39 (2):97–102, 1984.

[77] M. Bunge, N. Araghipour, T. Mikoviny, J. Dunkl, R. Schnitzhofer, A. Hansel, F. Schinner, A. Wisthaler, R. Margesin, and T. D. Mark. On-line monitoring of microbial volatile metabolites by proton transfer reaction-mass spectrometry. *Appl Environ Microbiol*, 74(7):2179–86, 2008.

[78] X. Chen, F. Xu, Y. Wang, Y. Pan, D. Lu, P. Wang, K. Ying, E. Chen, and W. Zhang. A study of the volatile organic compounds exhaled by lung cancer cells in vitro for breath diagnosis. *Cancer*, 110(4):835–44, 2007.

[79] M. P. Rutten-van Molken, M. J. Postma, M. A. Joore, M. L. Van Genugten, R. Leidl, and J. C. Jager. Current and future medical costs of asthma and chronic obstructive pulmonary disease in the netherlands. *Respir Med*, 93(11): 779–87, 1999.

[80] E. Derom, C. van Weel, G. Liistro, J. Buffels, T. Schermer, E. Lammers, E. Wouters, and M. Decramer. Primary care spirometry. *Eur Respir J*, 31 (1):197–203, 2008.

[81] S. A. Kharitonov and P. J. Barnes. Exhaled markers of inflammation. *Curr Opin Allergy Clin Immunol*, 1(3):217–24, 2001.

[82] R. L. Gibson, J. L. Burns, and B. W. Ramsey. Pathophysiology and management of pulmonary infections in cystic fibrosis. *Am J Respir Crit Care Med*, 168(8):918–51, 2003.

[83] N. Fens, A. H. Zwinderman, M. P. van der Schee, S. B. de Nijs, E. Dijkers, A. C. Roldaan, D. Cheung, E. H. Bel, and P. J. Sterk. Exhaled breath profiling enables discrimination of chronic obstructive pulmonary disease and asthma. *Am J Respir Crit Care Med*, 180(11):1076–82, 2009.

[84] R. J. Delfino, H. Gong, W. S. Linn, Y. Hu, and E. D. Pellizzari. Respiratory symptoms and peak expiratory flow in children with asthma in relation to volatile organic compounds in exhaled breath and ambient air. *J Expo Anal Environ Epidemiol*, 13(5):348–63, 2003.

[85] S. Dragonieri, J. T. Annema, R. Schot, M. P. van der Schee, A. Spanevello, P. Carratu, O. Resta, K. F. Rabe, and P. J. Sterk. An electronic nose in the discrimination of patients with non-small cell lung cancer and copd. *Lung Cancer*, 64(2):166–70, 2009.

[86] H. Shwachman and L. L. Kulczycki. Long-term study of one hundred five patients with cystic fibrosis; studies made over a five- to fourteen-year period. *AMA J Dis Child*, 96(1):6–15, 1958.

[87] T. L. Edwards, E. Torstensen, S. Dudek, E. R. Martin, and M. D. Ritchie. A cross-validation procedure for general pedigrees and matched odds ratio fitness metric implemented for the multifactor dimensionality reduction pedigree disequilibrium test. *Genet Epidemiol*, 34(2):194–9, 2009.

[88] C. Ambroise and G. J. McLachlan. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci U S A*, 99(10): 6562–6, 2002.

[89] H. Grasemann, I. Ioannidis, R. P. Tomkiewicz, H. de Groot, B. K. Rubin, and F. Ratjen. Nitric oxide metabolites in cystic fibrosis lung disease. *Arch Dis Child*, 78(1):49–53, 1998.

[90] C. M. Robroeks, J. J. van Berkel, J. W. Dallinga, Q. Jobsis, L. J. Zimmermann, H. J. Hendriks, M. F. Wouters, C. P. van der Grinten, K. D. van de Kant, F. J. van Schooten, and E. Dompeling. Metabolomics of volatile organic compounds in cystic fibrosis patients and controls. *Pediatr Res*, 2010.

[91] B. J. Rosenstein and G. R. Cutting. The diagnosis of cystic fibrosis: a consensus statement. cystic fibrosis foundation consensus panel. *J Pediatr*, 132(4):589–95, 1998.

[92] M. C. Oliveira, F. J. Reis, E. A. Oliveira, E. A. Colosimo, A. P. Monteiro, and F. J. Penna. Prognostic factors in cystic fibrosis in a single center in brazil: A survival analysis. *Pediatr Pulmonol*, 34(1):3–10, 2002.

[93] M. Rosenfeld, J. Emerson, J. Williams-Warren, M. Pepe, A. Smith, A. B. Montgomery, and B. Ramsey. Defining a pulmonary exacerbation in cystic fibrosis. *J Pediatr*, 139(3):359–65, 2001.

[94] J. Kanga, R. Kuhn, L. Craigmyle, D. Haverstock, and D. Church. Cystic fibrosis clinical score: a new scoring system to evaluate acute pulmonary exacerbation. *Clin Ther*, 21(8):1343–56, 1999.

[95] C. Johnson, S. M. Butler, M. W. Konstan, W. Morgan, and M. E. Wohl. Factors influencing outcomes in cystic fibrosis: a center-based analysis. *Chest*, 123(1): 20–7, 2003.

[96] D. L. Ashley, M. A. Bonin, B. Hamar, and M. McGeehin. Using the blood concentration of 2,5-dimethylfuran as a marker for smoking. *Int Arch Occup Environ Health*, 68(3):183–7, 1996.

[97] I. Rahman and F. Kelly. Biomarkers in breath condensate: a promising new non-invasive technique in free radical research. *Free Radic Res*, 37(12):1253–66, 2003.

[98] A. K. Thukkani, F. F. Hsu, J. R. Crowley, R. B. Wysolmerski, C. J. Albert, and D. A. Ford. Reactive chlorinating species produced during neutrophil activation target tissue plasmalogens: production of the chemoattractant, 2-chlorohexadecanal. *J Biol Chem*, 277(6):3842–9, 2002.

[99] K. Dettmer, P. A. Aronov, and B. D. Hammock. Mass spectrometry-based metabolomics. *Mass Spectrom Rev*, 26(1):51–78, 2007.

[100] S. C. Gates and C. C. Sweeley. Quantitative metabolic profiling based on gas chromatography. *Clin Chem*, 24(10):1663–73, 1978.

[101] M. Phillips. Method for the collection and assay of volatile organic compounds in breath. *Anal Biochem*, 247(2):272–8, 1997.

[102] Diana Poli, Matteo Goldoni, Massimo Corradi, Olga Acampa, Paolo Carbognani, Eveline Internullo, Angelo Casalini, and Antonio Mutti. Determination of aldehydes in exhaled breath of patients with lung cancer by means of on-fiber-derivatisation spme-gc/ms. *Journal of Chromatography B*, In Press, Corrected Proof, 2010.

[103] S. I. Rennard and D. M. Daughton. Smoking cessation. *Chest*, 117(5 Suppl 2): 360S–4S, 2000.

[104] G. L. Snider. Chronic obstructive pulmonary disease–a continuing challenge. *Am Rev Respir Dis*, 133(5):942–4, 1986.

[105] E. K. Silverman and F. E. Speizer. Risk factors for the development of chronic obstructive pulmonary disease. *Med Clin North Am*, 80(3):501–22, 1996.

[106] J. C. Hogg. Chronic obstructive pulmonary disease: an overview of pathology and pathogenesis. *Novartis Found Symp*, 234:4–19; discussion 19–26, 2001.

[107] M. Kraft, G. H. Cassell, J. E. Henson, H. Watson, J. Williamson, B. P. Marmion, C. A. Gaydos, and R. J. Martin. Detection of mycoplasma pneumoniae in the airways of adults with chronic asthma. *Am J Respir Crit Care Med*, 158(3):998–1001, 1998.

[108] R. G. Hegele, S. Hayashi, J. C. Hogg, and P. D. Pare. Mechanisms of airway narrowing and hyperresponsiveness in viral respiratory tract infections. *Am J Respir Crit Care Med*, 151(5):1659–64; discussion 1664–5, 1995.

[109] M. R. Becklake. Occupational exposures: evidence for a causal association with chronic obstructive pulmonary disease. *Am Rev Respir Dis*, 140(3 Pt 2):S85–91, 1989.

[110] D. M. Mannino, R. C. Gagnon, T. L. Petty, and E. Lydick. Obstructive lung disease and low lung function in adults in the united states: data from the national health and nutrition examination survey, 1988-1994. *Arch Intern Med*, 160(11):1683–9, 2000.

[111] B. Aguilaniu, S. Goldstein-Shapses, A. Pajon, P. Levy, F. Sarrot, X. Leverve, E. Page, and J. Askanazi. Muscle protein degradation in severely malnourished patients with chronic obstructive pulmonary disease subject to short-term total parenteral nutrition. *JPEN J Parenter Enteral Nutr*, 16(3):248–54, 1992.

[112] C. J. Dobbin, D. Bartlett, K. Melehan, R. R. Grunstein, and P. T. Bye. The effect of infective exacerbations on sleep and neurobehavioral function in cystic fibrosis. *Am J Respir Crit Care Med*, 172(1):99–104, 2005.

[113] J. K. Block, K. L. Vandemheen, E. Tullis, D. Fergusson, S. Doucette, D. Haase, Y. Berthiaume, N. Brown, P. Wilcox, P. Bye, S. Bell, M. Noseworthy, L. Pedder, A. Freitag, N. Paterson, and S. D. Aaron. Predictors of pulmonary exacerbations in patients with cystic fibrosis infected with multi-resistant bacteria. *Thorax*, 61(11):969–74, 2006.

[114] B. Zethelius, L. Berglund, J. Sundstrom, E. Ingelsson, S. Basu, A. Larsson, P. Venge, and J. Arnlov. Use of multiple biomarkers to improve the prediction of death from cardiovascular causes. *N Engl J Med*, 358(20):2107–16, 2008.

[115] T. J. Wang, P. Gona, M. G. Larson, D. Levy, E. J. Benjamin, G. H. Tofler, P. F. Jacques, J. B. Meigs, N. Rifai, J. Selhub, S. J. Robins, C. Newton-Cheh, and R. S. Vasan. Multiple biomarkers and the risk of incident hypertension. *Hypertension*, 49(3):432–8, 2007.

[116] P. M. Ridker, J. E. Buring, N. Rifai, and N. R. Cook. Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: the reynolds risk score. *Jama*, 297(6):611–9, 2007.

[117] M. S. Pepe, H. Janes, G. Longton, W. Leisenring, and P. Newcomb. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol*, 159(9):882–90, 2004.

Written in LaTeX