Exhaled breath analysis

CHAPTER 8

General discussion and summary

Introduction

The present thesis focuses on the analysis and implementation of exhaled breath volatile organic compound (VOC) analysis in clinical settings of gastrointestinal diseases in terms of disease diagnosis and prognosis. VOC analysis can be performed in various means (e.g. faeces, blood, urine, saliva, bile, breath), although breath is the most prominent due to its patient-friendliness in sampling. It is a well-known and documented fact that dogs can smell cancer [1-4], and in general, animal sniffing studies have shown some fascinating results; animal canine olfactory acuity is over 100.000 times stronger than human acuity [4]. Another example is the case of giant African pouched rats that showed superiority in diagnosing tuberculosis over microscopy [4]. A few years ago, the first-ever human sniffing case was also reported, whereby a British woman could smell Parkinson's [5]. This woman's extraordinary smell helped scientists identify ten molecules that could lead to the first diagnostic test for the condition [5]. Breath has also been investigated since ancient times when clinicians used the smell of breath as a diagnostic tool for various illnesses. For example, the Greek physician Hippocrates of Cos noted the importance of breath smell in diagnosing liver disease, using the term "foetor hepaticus" to describe the characteristic breath odour associated with liver impairment [6]. The aforementioned fascinating results, the high costs for training and housing animals, and the genuine interest in breath research over the centuries led to significant technological developments in sampling, storing, and analysing breath for volatile chemicals. These technological developments spiked even more interest in breath research.

Exhaled breath applications

Exhaled breath VOC analysis holds a lot of potential due to its promising use as a non-invasive, cost-effective, and easy-to-use diagnostic and monitoring tool. Despite all the interest and technological advances, exhaled breath is yet to find diagnostic implementations in the clinics. Many confounding factors can influence exhaled breath, such as lifestyle, environment, medication, smoking, or diet [7]. Exhaled breath also generates enormous and complex datasets that are difficult to handle; for example, how one should analyse their data to separate background noise from biological information [8]. Nevertheless, there are good implementation examples of exhaled breath tests, such as the alcohol consumption [9], C13 isotope labelled substrate [10] monitoring, and the hydrogen [11] breath tests. The alcohol breath test measures how much alcohol there is in the blood. In beverage consumption, ethanol goes to the stomach and the small intestine, and from there, it is absorbed in the blood, carried through the body to the lungs, and then excreted through breath. The C13 isotope test monitors in-vivo metabolic activities. A probe containing a C13 isotope (e.g. C13-labelled methacetin) is administered to a subject, which is then metabolised in the body, and ultimately excreted via the breath in the form of C13O2. The breath excretion of this isotope is used as an indication of the metabolic activity of enzymes in organs such as the liver. An example of such a test is the methacetin breath test (MBT), which monitors postoperative liver metabolism and impairment in subjects undergoing hepatectomy [10]. C13-labelled methacetin is de-alkylated in the liver by the CYP1A2 enzyme, forming paracetamol and C13-formaldehyde, which is then converted to C13O2 and excreted in the breath. The production of C13O2 correlates with general liver function, and it does not say anything regarding the stage of liver impairment. The design of a C13 isotope labelled substrate breath test should also be based on knowledge of a specific metabolic function or malfunction. Established liver metabolic pathways and their associated excreted C13O2 (Chapter 2) could potentially be used to develop other C13 isotope breath tests. Lastly, the hydrogen breath test is fundamentally different from the C13-labelled isotope test because it involves using various substrates such as glucose, lactose, lactulose, and fructose to diagnose small intestine bowel overgrowth (SIBO), or lactose or fructose malabsorption [11]. Such a test measures the amount of hydrogen in breath. Bacteria, especially anaerobic, colonizing the large bowel in healthy and the small bowel in diseased conditions produce hydrogen by fermentation of unabsorbed carbohydrates. Though small amount of hydrogen is produced from limited amounts of unabsorbed carbohydrate reaching the colon, large amounts of hydrogen may be produced if there is malabsorption of carbohydrates in the small intestine, allowing larger amount to reach the colon or if there is excess of bacteria in the small bowel. The hydrogen produced by the bacteria is absorbed through the wall of the small or large intestine or both. The hydrogen-containing blood travels to the lungs where the hydrogen is released and exhaled in the breath. The aforementioned examples indicate that there

is information to be found in exhaled breath. Although it seems as if the right way of analysing it and capturing it consistently in more advanced settings such as these in the clinics remains a challenge.

Exhaled breath data analysis

Breathomics

Measuring the vast amount of volatile chemicals in exhaled breath relates to the overall –omics field (including proteomics, metabolomics, genomics, and transcriptomics), as large and biologically complex datasets similarly characterise it. “Omics” technologies have a broad range of applications, and they are aimed at the detection of genes (genomics), mRNA (transcriptomics), proteins (proteomics), and metabolites (metabolomics) in biological samples in a non-targeted manner. Advances in microarray technology busted genomics and transcriptomics research, whereas advances in mass spectrometry boosted proteomics and metabolomics research. Similarly, breathomics (or volatilomics) are aimed at detecting volatile organic compounds, and they have advanced due to advances in mass spectrometry. “Omics” technologies adopt a holistic view of molecules that make up a cell, tissue, or organism, and they are considered hypothesis-generating since no hypothesis is known and all data are acquired and analysed to define a hypothesis that can be tested. Furthermore, “omics” technologies can be applied to understand better healthy physiological and diseased processes used for screening, diagnosis, prognosis, or understanding disease aetiology. “Omics” are also used in biomarker discovery, and multiple molecules are simultaneously investigated.

Difficulties arise when analysing genomics, transcriptomics, proteomics, and metabolomics data concerning how one should properly collect, handle, and analyse the data; breathomics data are no different. For example, genomics and transcriptomics analysis requires real-time PCR validation regarding microarray changes, whereas proteomics analysis requires complex algorithms to match the data to theoretical databases to enable protein identification and quantification. Metabolomics analysis requires using univariate, multivariate, supervised, or unsupervised statistical methods to look for underlying data patterns and uncover biological information that can be used for further hypothesis-testing. Multiple studies have paved the way for how such complex biological data should be approached based on their type (e.g. genomics or metabolomics) [12-17]. Like metabolomics analysis, breathomics analysis requires statistical methods to uncover biological information. Data preprocessing is crucial when dealing with numerically complex volatilome data (Chapter 3). Pre-processing typically comprises noise and baseline removal, correcting peak shifts due to column ageing, temperature drift or biochemical interaction, and peak picking. Additionally, most of the VOCs in breath samples are not present in all samples. This leads to

another preprocessing step: the retention of compounds present in at least 10% of the samples. Next, normalisation, transformation, and scaling steps should be applied before supervised or unsupervised methods are applied to the breathomics data for further analysis (Chapter 3). The steps above should all be considered standard practice in breathomics analyses, although this is not always the case [8].

Data quality and challenges when performing exhaled breath analysis

Data science has seen tremendous development and implementation in the last decades. Artificial intelligence, machine learning, and deep learning algorithms find implementation in almost every field, such as scientific [8, 18, 19], economic [20, 21], political [22, 23], or geographic [24, 25], to name a few. This is because technology has advanced, and the way of life has become digital and involves large amounts of data. These algorithms have promoted a healthier and improved way of living through, for example, automated cars (e.g. Tesla) or wearables (e.g. smart watches) and have ultimately invaded the most challenging and complex research fields and questions to be answered to date. A prime example of this is the medical research field and its research questions. There have been various successful big data implementations of these algorithms in clinics. They have helped in clinical decision support systems (e.g. the surgical intelligent knife [26, 27]) or medical imagining (e.g. diabetic retinopathy screening [28]). These algorithms can deal with multi-variable and high complexity datasets; however, they do require the data to be of high quality (e.g. no background noise, no non-biological variation present or instrumental artefacts). Assuming that biological information is present in the data, poor data quality is one of the main reasons why these algorithms struggle to solve certain medical challenges and questions.

Breathomics data are characterised by high complexity and multi-variable datasets. Both aspects can be explained by diving into the origin of the exhaled VOCs. VOCs are detected in different body matrices such as breath, faeces, urine, bile, breast milk, and blood, resulting from exogenous or endogenous sources (Chapter 2). Exogenous VOCs (EVOCs) originate from the gut microbiome or the environment. The latter are absorbed through the skin, inhaled, or ingested with food and beverages. Moreover, they might be the result of therapeutic interventions. Biochemical processes in body cells and tissues produce endogenous VOCs, such as in lung and airway tissues or other organ tissues (e.g. liver or kidney); these VOCs can reflect apoptosis, inflammation or oxidative stress [29]. VOCs may arise from body chemical reaction cascades in diseased individuals due to cellular damage; they are released in the bloodstream and spread among the body excretions (Chapter 2). A single breath sample contains thousands of VOCs, leading to multi-variable datasets [30]. These VOCs also interact, causing non-linearities in the data, which translates to even higher data complexity.

Proper handling and preprocessing of breathomics data (Chapter 3) does not necessarily lead to high-quality data or reproducible results. The clinical study design plays a crucial role in both aspects. The clinical study design term refers to the formulation of trials and experiments, as well as observational studies in research involving humans. Several pitfalls and mistakes that occur when one performs a breath VOC analysis lead to low-quality data and non-reproducible results (Chapter 2 and Chapter 4). A vital pitfall that influences data quality and does not allow for reproducible results is the different ways of sampling, storing, or analysing the breath samples are used, which most likely has introduced bias in the data. Therefore, it is paramount to develop a standardisation framework for breath analysis research; currently, attempts toward this are ongoing [31-33]. A common mistake that further hampers result reproducibility is that many studies do not perform any internal or external validation of their findings, or correction of possible confounding factors is also not considered (Chapter 2). Another pitfall is that there is no consensus on what should be regarded as a proper way of handling the data regarding statistical modelling. There is an abundance of available tools to conduct statistical modelling, though it is not always clear what should be chosen or how should one approach their data (Chapter 3).

Another critical challenge in achieving high-quality and trustworthy data is getting "good" control cohorts to compare the diseased groups and determine whether found VOCs are disease-specific or not. It is also challenging to define "healthy" in the context of breath since hidden, underlying issues may be present in each participating individual (Chapter 4). It has recently been reported that 1488 VOCs have been found in the exhaled breath of healthy individuals [34], meaning that it is challenging to say whether identified VOCs are indeed disease-specific are not. A solution to this could be to perform in vitro and animal studies to identify biomarker VOCs that are exclusive, reliably produced, and disease-specific before human studies. This would also require identification of VOC origin, chemical structure, and the possibility of VOCs originating from human disease. A targeted VOC human study could be conducted as soon as these steps are performed.

High-quality breathomics data and reproducible results are hard to generate also because they are prone to batch effects [18]. Batch effects are sources of variation unrelated to the examined samples or inter- or intra-sample class differences. Environmental or methodological differences can cause batch effects during sample collection, chemical analysis, and data handling (Chapter 4). Batch effects are a common problem; they also occur in the other –omics fields. To eliminate batch effects as much as possible, ideally, every sample would have to be measured by the same personnel, at the same location, at the same time, and under the same conditions, and this is not achievable. Batch effects might still occur even if one takes all precautions possible. This is because analytical techniques such as gas chromatography-mass spectrometry or nuclear magnetic resonance have become

highly sophisticated and sensitive, capturing biological and non-biological variations. Scientific literature suggests statistical ways to deal with batch effects in genomics, transcriptomics, proteomics, and metabolomics data [18]; no batch-effect correction techniques have been reported in the literature yet for breathomics data (Chapter 4). The batch effect correction techniques are data-specific (e.g. specifically made for metabolomics data), and therefore, they could not be applied to breathomics data. A way to circumvent batch effects could be using quality controls in regular intervals when running a breath VOC analysis. The use of quality controls is a known practice in metabolomics studies, with demonstrated successful applications [35]. Quality controls can help improve and monitor analysis and data quality, and their use should become standard practice when conducting breathomics analysis (Chapter 4). Monitoring analysis and data quality can lead to high-quality and reproducible data and eventually allow for cross-sectional study comparisons. These, together with a standardised framework and a consensus on analytical and statistical analysis, can help bring exhaled breath to the clinics.

Statistical modelling in exhaled breath analysis

Exhaled breath analysis strongly relies on statistical modelling when multiple VOCs are simultaneously considered, and the development of a successful exhaled breath VOC test would require high model classification and prediction accuracy. Numerous options exist when it comes to building a predictive model. Scientific literature suggests ensemble and linear regression techniques successfully built high-accuracy predictive models [36]. The linear regression techniques are the most well-known and applied in biological data (e.g. Partial Least Squares Regression Analysis [36]). Ensemble techniques are split into three main categories: boosting, bagging, and stacking. The most well-known are AdaBoost, Random Forest, and Gradient Boosting [36], and they include a wide range of successful applications such as flood hazard, earthquake damage, or sleep pattern identification, to name a few [37-44]. Ensemble techniques and mainly Random Forest have only recently gained attention in the field of breath analysis research, and they have started to be applied (Chapter 5). Breath research should continue shifting its interest towards ensemble techniques because they can deal better with multi-variable and complex biological data (e.g. breathomics data) than the linear regression techniques (Chapter 5). The reason is that linear regression techniques assume only linear relations amongst the dataset variables (e.g. VOCs), whereas ensemble techniques assume both linear and nonlinear relations. VOCs in breath samples interact with each other, which means that nonlinear relations are formed.

Applications in computational science have shown that more than one data source can often lead to better classification or prediction results [19]. It is a common belief that the more data, the merrier the result since all these statistical approaches can cope with large volumes of data. Generally, their success ratio improves when more data

are fed into them. Although combining different types of data does not always yield higher model performance, considerations have to be taken into account before any analysis is conducted based on the type of study and the ultimate analysis aim [19]. The idea behind the "the more, the better" principle is that different data sources can generate complementary datasets by capturing different entities (Chapter 6). There is no gold standard for what can be complementary to what; various data sources can be considered complementary depending on the type of analysis and the question at hand each time. Important to be considered before performing data fusion would be a proper data pre-fusion treatment. Variable scaling is required before any data from different sources are concatenated since the magnitude of data coming from various sources is most likely different. From a breath VOC research standpoint, it would make sense to fuse data from different sources. Different sources would mean VOCs produced by, for example, an inflamed organ, which could be released via different routes (e.g. faeces, urine, or breath) through the bloodstream. There are three main ways of data fusion (i.e., low-level, mid-level, and high-level), which have been successfully implemented in biological data; however, there is no available literature on fusion of VOCs either coming from different sources or combined with other types of data (e.g. metabolomics). The breath community should more deeply examine the concept of data fusion. It should also keep in mind that as the complexity and amount of data increase, more advanced and sophisticated fusion methods might be needed (Chapter 6). Advanced fusion ways have been recently proposed, outperforming traditional fusion methods when biological data were used [19, 45].

Breath VOC biomarker discovery also relies on identifying and interpreting VOC that help build good predictive models. VOC identification and interpretation could become a bottleneck when advanced predictive models (e.g. ensemble techniques) are used instead of linear regression techniques due to data complexity. Variable (e.g. VOC) transformation is often needed when advanced predictive techniques are used. If advanced predictive models are used, advanced ways of tracing and visualization of VOCs might be required, too [46]. The pseudo-sample principle has proved to be a successful way of doing so [19, 47] by visualizing VOC importance and behaviour in biological samples (Chapter 6). The pseudo-sample principle is based on a nonlinear plot idea to represent variable importance as a set of artificial samples constructed to evaluate each variable independently. The pseudo-sample principle seems promising and helpful for future investigations, but it can also prove troublesome due to its complexity. Nonetheless, this approach presents a way of dealing with a common problem in biomarker discovery research.

Case study based on acquired knowledge

The present thesis performed a case study that took into account knowledge gained here and tried to use the latest strengths and technological developments in the field to test and validate the theories and points made thus far. Primary Sclerosing Cholangitis was the examined disease; PSC is an orphan liver disease since it roughly affects 60.000 individuals in the western world. Many VOCs coming from breath have been linked to liver impairment (Chapter 2), and the case study used these compounds in a targeted way to see whether they could be used to differentiate between PSC diseased individuals and Inflammatory Bowel Disease diseased individuals with concurring PSC. The choice of IBD was made given the high correlation between IBD/PSC patients with PSC patients. As discussed in Chapter 4, the case study also used quality controls to monitor data and analysis quality for possible batch effects, and it preprocessed the data by following the preprocessing steps mentioned in Chapter 3. Statistical modelling of the PSC breathomics data was conducted by implementing unsupervised and supervised machine learning approaches, as suggested in Chapter 5.

The case study gave good classification results, confirming that the selected VOCs can also potentially be used for PSC detection. The good classification results also confirmed the Chapter 5 statement that ensemble methods work better on complex biological data than linear regression methods. Linear regression was also implemented, but no satisfactory results were obtained. The study results were validated by using a test set, and the found VOCs were tested for the significance of confounding factors such as smoking or diet (as discussed in Chapter 2 in common mistakes and pitfalls in breath research). The case study also aims to use and validate the proposed in Chapter 6 data fusion and variable interpretation approaches by combing the breath VOCs with faecal VOCs. This is still a work in progress; therefore, it is not discussed in the present thesis. Data fusion would be believed to improve the case study classification results based on the theory of “leaky-gut” [48]. This theorem states that an ongoing inflammatory stimulus, which originates from the gut, preserves a bile duct inflammation in PSC patients, leading to molecule excretion in breath samples, faecal samples, or blood samples. This would render breath and faecal VOCs as complementary data.

Standard practices, alternatives, and future perspectives in breath VOC analysis

Breath VOC research has been mainly focused on using Gas Chromatography-Mass Spectrometry in biomarker discovery [41, 45, 49-53]; this is also what was used in the present thesis case study (Chapter 7). Less commonly used yet successful techniques are Proton Transfer Reaction-MS, Selected Ion Flow-Tube-MS, Ion-Molecule

Reaction-MS, Field Asymmetric Ion Mobility Spectrometry, and E-nose (Chapter 2). In GC-MS, a mixture is split into individual substances with heating, and the heated gases are carried through a column with an inert gas (e.g. Helium). As the separated substances emerge from the column opening, they flow into the MS, where they are identified by the mass of the analyte molecule. In PTR-MS, the organic trace gases are ionized by undergoing a proton-transfer reaction with H3O+ ions. The product ions are then mass analysed and detected by a quadrupole mass spectrometer, yielding information about the neutral precursors. The reaction is exothermic and efficient for those compounds with a proton affinity (PA) higher than the proton affinity of water. In SIFT-MS, the selected reagent ion is injected into the flow tube, and excess energy is removed through collisions with the carrier gas. The sample is then introduced at a known flow rate, and the reactive compounds it contains are ionized by the reagent ion to form well-characterized product ions. FAIMS is a technique based on gas phase separations on a millisecond timescale at atmospheric pressures and ambient temperature. It separates ions based on their differential mobility in high and low electric fields, a function of mass, charge, size, and shape. E-nose mimics human olfaction, whose functions are non-separate mechanisms (i.e. the smell or flavour is perceived as a global fingerprint); it consists of a sensor array, pattern reorganization modules, and headspace sampling to generate a signal pattern that is used for characterizing smells. Compared to GC-MS, PTR-MS seems to provide a more complex picture of the compounds, and it can distinguish between different disease severity classes, whereas SIFT-MS provides a higher detection sensitivity for compound concentrations lower than parts per billion and real-time quantification. IMR-MS is more selective and sensitive than GC-MS and does not require any pre-concentration step before analysis compared to other MS-based technologies. FAIMS exceeds other MS-based methods because it can be applied at the point of care since it offers an immediate compound response (as long as the compounds are known); this establishes it as a cost-effective clinical test. Lastly, E-nose provides a rapid profile of detected compounds on a point-of-care base because it can be performed instantaneously in an outpatient care setting, whereas MS-based methods cannot. The disadvantage of the E-nose technology is that the individual compounds are not identifiable compared to MS-based technologies.

It cannot be said whether one of the techniques above is the best for breath VOC analysis since each one has its advantages and disadvantages over the others. More research would be needed on the less commonly used MS-based techniques, and even so, a standardised breath analysis framework based on an MS-based approach might not be what could ultimately lead to breath diagnostic test applications in the clinics. MS-based technologies are generally not portable (micro-GC-MS has been developed [56]) and are expensive, whereas the E-nose technology is inexpensive, portable, and rapid. E-nose does not allow for compound identification; however, a good starting point for bringing breath tests into clinics would be a reliable screening

or monitoring tool, and for such a tool, compound identification does not seem to be a necessity. Breath research has focused on untargeted approaches by blindly looking into breath samples for VOCs. Analysing breath in such a way provides a holistic overview of the breath content, making it difficult to say whether these changes are either specific to a particular disease or more general markers of underlying mechanisms such as inflammation. As noted in Chapter 2, other approaches (e.g. exogenous VOCs; EVOCs) might be more beneficial. Especially in liver breath VOC research, such approaches would make sense due to liver metabolic capacity, and they should be investigated in more depth in the future. These approaches would require exposing or ingesting a cohort to a particular compound concentration (i.e. probe), sampling their breath after exposure or ingestion, and measuring the associated EVOC metabolite in inhaled air to determine liver function. An EVOC analysis enables a tailored, controlled exposure to a compound of interest, providing a better chance to identify disease-specific markers. An EVOC analysis would also be more robust to background VOCs (e.g. environmental VOCs), which are often one of the major confounding factors in the field. However, there are weaknesses to such an approach too. Exposure to or ingesting a specific probe that leads to a particular EVOC product in the breath may require METC approval, patient preparation, and most importantly, it might be a source of a potential allergy (Chapter 2). An EVOC approach would also require an extensive understanding of the probe metabolism, and to achieve this, more in vitro analyses are needed.

Focus on technological developments should also be given; developments such as the ReCIVA sampling apparatus [54] are guaranteed to help advance the breath research field further. However, breath VOC research must first ensure a high-quality laboratory practice by establishing a common and consistent framework before exploring new ways such as the ones mentioned above. It is of paramount importance to have a standardised framework with standard rules of analysis because that way, external data influential factors can be eliminated or significantly reduced (Chapter 2 and Chapter 4).

Final considerations and conclusion

The present thesis aimed to answer whether breath VOC analysis could find diagnostic and prognostic clinical applications. The present dissertation is imperfect and cannot answer this fully; however, it can speculate on the future of the breath field. Breath research has remained stagnant in the last couple of decades regarding clinical applications regarding disease diagnosis and prognosis. In the financial and banking sector, there is the expression of “path to green” when managing risks that the banks are exposed to and how to keep these risks within risk appetite. Risk appetite is the level of risk that an organization is prepared to accept to pursue its objectives before action is deemed necessary to reduce the risk. In breath research,

such a “path to green” would mean successful diagnostic and prognostic clinical day-to-day applications. Research conducted in the present thesis shows a future in exhaled breath research, and such a “path to green” would entail identifying and dealing with the reasons that led to this stagnation. Three main components have led to this stagnation: lack of a standardised framework in terms of clinical design, lack of a consensus in data handling and statistical tool availability and use, and wavering ideologies on whether targeted or untargeted approaches should be considered. The present thesis findings illustrate that exhaled breath could find diagnostic and prognostic clinical applications if these three components are resolved. Scientific literature and the present thesis case study suggest that there is information to be captured in breath, although it cannot be disclosed consistently yet.

In monitoring and screening, breath analysis has already had some successful implementations. Currently, many tests are used in the clinics, such as the methacetin breath test, which monitors postoperative liver metabolism and impairment in subjects undergoing hepatectomy. Available literature suggests that liver research could further benefit from shifting more interest towards breath analysis. The present thesis also demonstrated the potential applicability of breath analysis as a means of diagnosis in liver research since it showed that challenging distinctions (e.g. PSC from IBD patients) could be satisfactorily achieved. Nonetheless, screening/monitoring and diagnostic tests would still require a deep and extensive understanding of compound origin via in-vitro analyses before further implementation in human studies.

Multiple VOC breath analysis strongly relies on statistical modelling. Perhaps, the breath community should more closely join forces with the data science community to see what other ideas could be used in modelling, data fusion, or variable interpretation to help the breath research field flourish. Technological advancements to detect VOCs should also be given attention; however, this should go hand-in-hand with the breath community’s expanding knowledge on compound origin.