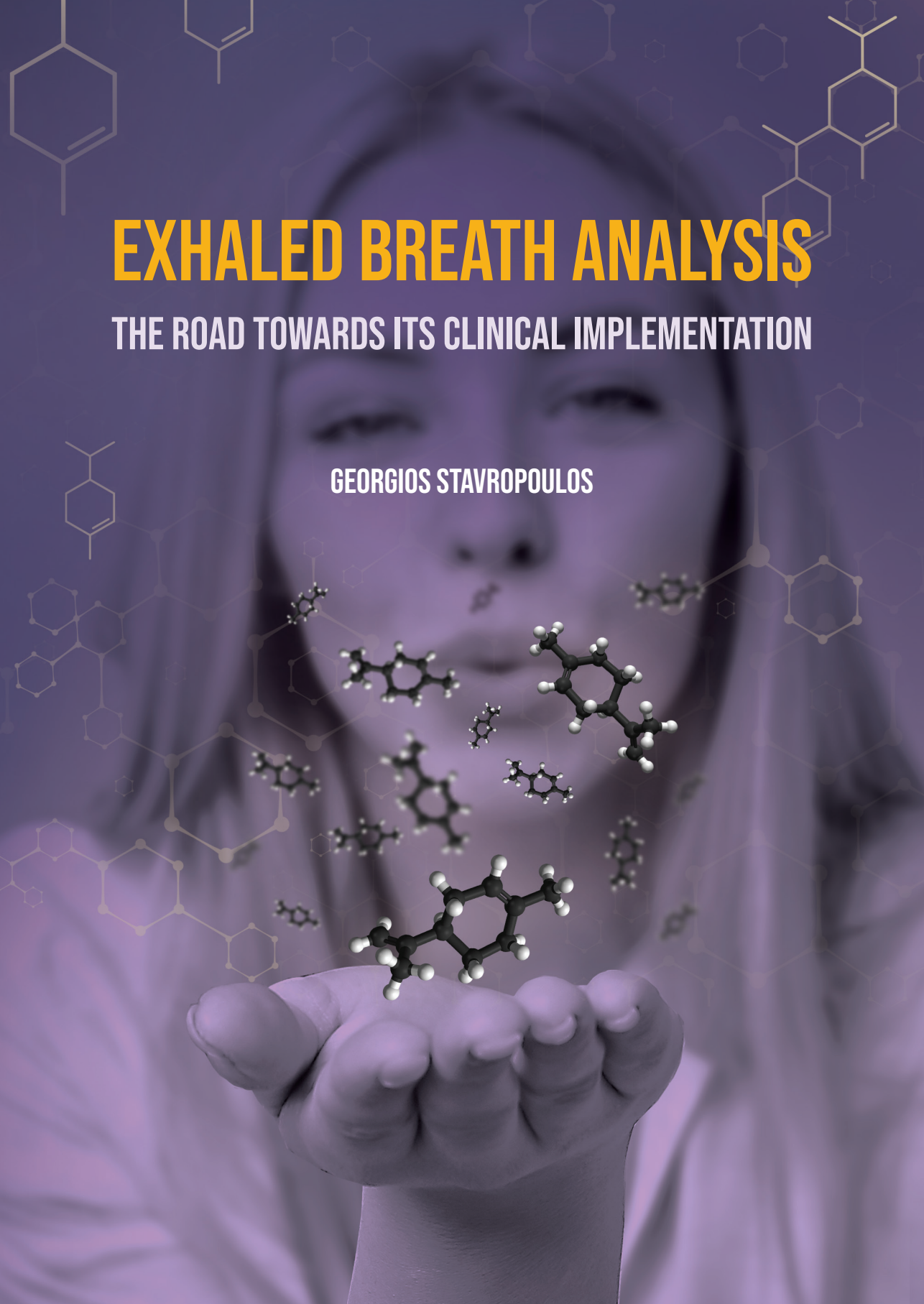
Exhaled breath analysis

EXHALED BREATH ANALYSIS

THE ROAD TOWARDS ITS CLINICAL IMPLEMENTATION

GEORGIOS STAVROPOULOS



Exhaled breath analysis:
the road towards its clinical implementation

Georgios Stavropoulos

The work described in the present thesis was conducted at the Department of Pharmacology and Toxicology at the School of Nutrition and Translational Research in Metabolism (NUTRIM) at the Maastricht University, Maastricht, The Netherlands.

The present work was supported by the Nederlandse Wetenschap Organisatie (NWO) Talent Program as part of the Talent Scheme funding instrument VENI.

ISBN: 978-94-6458-827-9
Cover design and Layout: Publiss | www.publiss.nl
Printed by: Ridderprint | www.ridderprint.nl

Exhaled breath analysis:
the road towards its clinical implementation

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Maastricht, op gezag van de Rector Magnificus, Prof. dr. N. D. Bouvy volgens het besluit van het College van Decanen, in het openbaar te verdedigen op donderdag, 25 mei 2023 om 16.00 uur.

door

Georgios Stavropoulos
Geboren op 1 mei 1991 te Patras

Promoter
Prof. dr. Frederik-Jan van Schooten

Co-promoter
Dr. Agnieszka Smolinska

Examination committee
Prof. dr. Nicole Bouvy (Maastricht University, chairman)
Prof. dr. Christina Davis (University of California)
Dr. Ger Koek (Maastricht University)
Dr. Edoardo Saccenti (Wageningen University and Research)

This research received financial support from the Nederlandse Wetenschap Organisatie (NOW) under the funding instrument VENI.

Table of Contents

CHAPTER 1

General introduction

Volatile organic compound analysis

Volatile organic compound (VOC) analysis has gained a lot of attention the last few decades due to its promising application as a non-invasive, patient-friendly, inexpensive, and easy to use diagnostic or monitoring tool [1]. The first report on VOC analysis dates back in the early seventies [2, 3]; although, technological advances in the decades to follow pioneered modern VOC analysis. VOCs are carbon-containing low molecular weight (i.e. < 1 kDa) compounds that exhibit high-vapour pressure at room temperature, and they can be clustered into several chemical classes such as alcohols, esters, hydrocarbons, aldehydes, to name a few. VOC concentration changes or appearance are seen as a response to human health issues (e.g. inflammation), and therefore, VOCs can be potentially used as biomarkers of human health. To date, thousands of VOCs have been identified and have been linked to several ailments such as liver impairment diseases [4], gastrointestinal tract diseases [5, 6], or pulmonary diseases [7-9]. However, the origin of the VOCs in breath can vary. They can originate from endogenous sources, and then, they are released into the human bloodstream; eventually, they are emitted in the air through a variety of bodily excretion sources. At the same time, VOCs from exogenous sources can be taken up and excreted again. The most common source of emitted VOCs, and the most widely studied, is exhaled breath; other VOC emission sources include faeces, blood, urine, skin, saliva, sweat, and the bile [1, 4].

Endogenous VOCs

VOCs are characterised as endogenous when they originate within the body; they can be produced biochemically by bodily cells and tissues or commensal and pathogenic microorganisms residing in the body. Endogenous VOCs are also considered VOCs that are already present in the body, but due to metabolic processes, their concentrations may change. Generally, a compound is considered endogenous when its concentration in a subject/patient is higher than in ambient air. It is hypothesised that these VOCs are by-products of normal cellular processes; these processes are altered when a disease occurs, and thus, they lead to the production of these VOCs, which are then diffused, due to their high volatility, from their point of origination to the bloodstream [10]. The point of origination for most of the endogenous VOCs that have been reported in the literature is largely unknown because there are many contributing factors to the VOC production. For instance, it has been hypothesized that hydrocarbons and aldehydes have resulted from the oxidative stress and lipid peroxidation processes [11]. However, there have also been instances where the point of origination has been pin-pointed. Such instances are the dimethyl-sulphide, which is known to be a late-stage liver failure indication [4], and acetone, which is produced in sepsis [12] or in uncontrolled diabetes [1]. It should be noted that microbiome is a big contributor to the endogenous VOC production [1]; however, there is no consensus whether VOCs originating from the microbiome should be purely considered as

endogenous due to the foreign nature of some of the pathogens that reside in it. Endogenous VOCs are also known as the human volatilome.

Exogenous VOCs

VOCs are characterised as exogenous when they originate from external factors such as the environment, lifestyle, diet, or therapeutic interventions. For example, environmental and lifestyle-related compounds can be absorbed by the skin or inhaled, whereas dietary or therapeutic compounds are ingested. On the one hand, the point of origination for most of the environmental and lifestyle-related VOCs can be found either in nature-related sources (e.g. terpenoid is emitted by plants to protect them against the ozone) or pollution-related sources (e.g. methane is emitted as a result of CO2) [13]. On the other hand, the point of origination for most of the dietary compounds is food and beverages. Other examples of known exogenous VOCs are the acetonitrile, which is found in the exhaled breath of smokers [14], and ethanol, which is found in the exhaled breath of alcohol drinkers. Exogenous VOCs are also known as the human exposome.

Exhaled breath: composition, sampling, and analysis

In the literature, VOC analysis has been almost exclusively performed on exhaled breath data [4], and this is because exhaled breath meets most of the criteria of the ideal diagnostic/monitoring tool: non-invasive, fast and cheap to perform, patient-friendly, it can be performed at the point-of-care, and it can be applied to every age group. It is also believed to, generally, be equally disease-information potent alongside faeces when it comes to VOCs contained in their respective samples [15]; although, sampling stools does not meet the patient-friendly criterion, thus, making exhaled breath the preferred VOC analysis means.

Exhaled breath consists of a mixture of gases: nitrogen, oxygen, carbon dioxide, noble gases, and VOCs that are present in concentrations ranging from nmol/L to pmol/L [16]. An exhaled breath sample can also be distinguished into the following fractions: the dead-space air, the alveolar air, and a mixture of dead-space and alveolar air. Different fractions are needed for different research questions or purposes; however, in a VOC analysis, the most informative fraction is the alveolar air (i.e. air in the alveoli) because this is where the pulmonary exchange of gases between air and blood happens [15]. Dead-space air fraction consists of CO2, and it is also exhaled first in an exhalation; therefore, proper sampling of the alveolar air requires monitoring of the CO2 exhalation levels. VOCs are already present in very small concentrations in the alveolar air, and sampling the alveolar air mixed with the dead-space air would further dilute the VOC concentrations, thus, making it even more difficult to detect them and properly quantify them.

Sampling of exhaled breath can be achieved with a variety of ways, with the most commonly used being: Tedlar bag, ReCIVA breath-sampler, and eNose sensor [1]. Tedlar (polymer) bag is the simplest and cheapest to use, especially in a large-scaling sampling of exhaled breath; however, it samples both the dead-space air and the alveolar air, which is not preferred in a VOC analysis, and it might contaminate the sample as well. ReCIVA [17] deals with this problem since it was especially designed to sample the alveolar air by monitoring the CO2 exhalation levels; it is also contamination-proof since it filters the air that the subject inhales before sampling. Although, the use of ReCIVA comes at a high cost since it is a lot more expensive to acquire the equipment, and it is also not easily portable. As far as the eNose sensor is concerned, it is considerably less expensive than ReCIVA, and it is portable, which is suitable for clinical use; however, they do not permit for individual VOC identification, whereas Tedlar bag and ReCIVA do when connected to analytical instrumentation such as, for instance, gas chromatography-mass spectrometry (GC-MS), and this is why they are less suitable for research purposes.

As far as analysis of the samples coming from these three sampling ways is concerned, Tedlar bag and ReCIVA samples, first, require use of adsorption tubes to store the samples. This is advantageous because the VOC traces are already present in very low concentrations in the samples, and these tubes extract the volatiles from the samples to pre-concentrate/enrich them onto the sorbent material. The most widely technique to analyse these tubes is gas chromatography-mass spectrometry (GC-MS); others are used too, such as proton transfer reaction-mass spectrometry (PTR-MS), ion mobility spectrometry (IMS), and selected-ion flow tube-mass spectrometry (SIFT-MS) [1]. This is why (i.e. the use of MS approaches) Tedlar bag and ReCIVA samples permit for VOC identification. The eNose sensor samples do not require any further analysis steps; immediately after sampling, a VOC profile is given. As a final note, MS approaches require large machines and a lot of equipment, they are expensive, and they require experienced personnel too.

The design of the clinical trial is one of the most critical components to discover reliable and reproducible biomarkers for specific medical conditions. The selection of subjects and patients to be enrolled is vital to the success of identifying breath biomarkers specific to a particular disease. Breath collection and measurements should be standardized to eliminate the possibility of errors in different centres around the world. Several attempts are made internationally to come up with guidelines to tackle this problem within both the International Association of Breath Research [18, 19] and the European Respiratory Society [20]. Once diagnostic biomarkers are selected in initial studies, these markers should then be thoroughly validated in independent retrospective studies in patients from different centres to demonstrate their robustness in identifying diseased individuals (susceptibility and specificity).

Volatilomics

Volatilome research is defined as quantitative and qualitative ways to find changes in VOCs present in bodily excretion means, and it aims to discover patterns of VOCs that are linked to deviant metabolic processes (e.g. inflammation) that take place in the human body [21]. The continued development of the analytical platforms (e.g. GC-MS) that are used for VOC analysis has resulted in large and complex biological datasets, which require extensive and advanced data processing and preprocessing. The biggest challenge in volatilome analysis is to separate biological signal/variance from noise or redundant information. It is for a fact that these analytical platforms introduce non-biological signal in the data due to their high sensitivity, and that most of the VOCs present in the samples contain redundant information [21]. Importantly, these non-biological signals are referred to as batch effects [22], and they are addressed in chapter 4. Machine learning algorithms have been developed to cope with such issues, and statistical modelling should lie at the core of a proper volatilome analysis.

Multivariate statistics

In statistical modelling, multivariate statistics is defined as the modelling process where more than one variables/parameters are examined or taken into account simultaneously to build a model, whereas univariate statistics is defined as the modelling process where only one variable is examined each time to build a model. A univariate statistics approach is the so-called student's t-test distribution test, which is used to determine whether two means of two different populations are statistically significant [23]. An example of a multivariate approach is the so-called principal component analysis (PCA) [24], where multiple variables are linearly combined to generate new variables (i.e. principal components) that capture most of the variation (i.e. information) present in the data. Multivariate statistics is the way to go in VOC analysis since volatilomics datasets consist of hundreds, if not thousands, of variables; put differently, there are hundreds of VOCs present in each sample. Multivariate statistics are divided into two main categories: unsupervised and supervised multivariate approaches. Unsupervised approaches are those that do not use any apriori information of the dataset, whereas supervised approaches do use apriori information of the dataset. For instance, such information can be class-related information. Furthermore, both unsupervised and supervised approaches can be used either for exploratory, predictive, or classificatory purposes. PCA is one of the most known unsupervised exploratory algorithms, whereas random forest [25] is one of the most known supervised predictive and classificatory algorithms. PCA, random forest, as well as other multivariate statistical approaches are discussed later in the present thesis.

Data preprocessing

Data preprocessing is, perhaps, the most important part of the VOC data analysis because, as highlighted already at the beginning of this section, the biggest challenge in volatilome analysis is to separate biological signal from noise. Noise can be introduced in the data either from instrumental artefacts or they can be caused by the personnel that conducted the analysis. Other reasons such as environmental contamination are also prone to introducing noise in the data [21]. A proper VOC data preprocessing strategy typically consists of the following steps: baseline removal, correcting for peak shifts, and peak picking. An extensive documentation of all the steps that a proper strategy consists of as well as several algorithms that have been developed to deal with these issues are addressed later in this thesis in chapter 3.

Batch effects

As stated already at the beginning of this section, batch effects are non-biological signals introduced in the data. Most of the times, proper data preprocessing prevents batch effects from happening; oftentimes, however, these non-biological signals are not possible to be prevented. Inevitable batch effects have been known to other fields such as transcriptomics or genomics, and specific algorithms have been developed to cope with them [22]. Volatilomics field is also prone to batch effects even after a proper data preprocessing strategy has been followed. No algorithms have been specifically designed to correct for these effects in volatilomics data, and it seems that even if such an algorithm were to be developed, it would come at the cost of losing important information due to the complexity of the volatilomics data. It has been proved that the existing algorithms do not properly work in volatilomics data, therefore, an additional preprocessing step maybe needed to account for these batch effects in VOC data—implementation of quality controls. This idea and detailed documentation on batch effects in volatilomics data is discussed in chapter 4.

Ensemble techniques

Ensemble techniques represent a particular category of multivariate supervised approaches that have become an integral part of statistical modelling, and especially of statistical prediction and classification modelling due to their high performance on complex datasets [26]. Ensemble techniques can be divided into three main categories: boosting, bagging (or aggregative bootstrapping), and stacking. Random forest is, perhaps, the most renowned one and it belongs to the bagging category. It is worth digging and properly understanding these techniques and how to validate them when working in the volatilomics field because they have proved to outperform other supervised techniques (e.g. partial least square analysis [27, 28]) that are very popular in various fields (e.g. metabolomics, proteomics) [26]. Ensemble methods are addressed in depth in chapter 5.

Variable interpretation

The use of advanced machine learning algorithms is not always enough in itself to provide proper, if at all, VOC identification or sample classification due to the complexity of the volatilomics data. Oftentimes, data transformation is needed, and the biggest challenge to be dealt with in such a case is to trace back the original variables/VOCs that led to the eventual study outcome [29, 30]. For instance, such transformation could be a kernel transformation of variables into samples [31]. The pseudo-sample principle that was initially proposed by Krooshof et al. [32] and further explored by Smolinska et al. [31] has proved to help trace back the original variables and demonstrate their behaviour in the dataset samples as well as their importance in the model in volatilomics data. It does so by providing a so-called trajectory plot and a bar plot [30]. An in-depth documentation on how the pseudo-sample principle is applied is discussed in chapter 6.

Data fusion

Data fusion is another integral part of statistical modelling that has gained a lot of attention in life sciences because analysis of biological data, such as the volatilomics data, might require multiple platform datasets to be combined to express the samples fully. The principle behind data fusion lies in the idea that different data platforms (e.g. GC-MS, nuclear magnetic resonance) detect different biological entities, which when combined can provide a comprehensive profiling of the research question in hand [30]. There are three main data fusion categories: low-level, mid-level, and high-level fusion approaches. In 2012, Smolinska et al. [31] introduced an advanced kernel fusion approach; they stacked kernels of their different data platforms to build their final model. In 2021, Stavropoulos et al. [30] stacked random forest weighted proximities of their different data platforms (volatilomics and metabolomics data) to build their final model. It is believed that data fusion would be of interest in VOC analysis, and it should be explored more in the volatilomics world. Therefore, data fusion as a whole as well as these two specific aforementioned cases are thoroughly discussed in chapter 7.

Thesis outline

The general aim of the present thesis is two-fold. First, the thesis aims to introduce VOC analysis by reviewing available literature on VOC analysis studies, what has been the main focus in all these VOC analyses thus far, and where the field, perhaps, should be headed to eventually make VOC analysis applicable in the clinics. Second, it aims to extensively discuss several technical aspects that should be considered and implemented before a proper VOC analysis is performed. In the end, it also presents a VOC case study that has considered and implemented all the aspects addressed here, and it concludes with a general discussion and summary.

The first aim of the thesis is addressed in chapter 2, where an extensive review of VOC studies in the liver is accomplished. More specifically, all the available literature on VOCs in liver diseases has been addressed, compounds of interest have been identified and summarized, limitations and flaws of the studies that were conducted are pointed out, and future suggestions are given based on the review findings.

The second aim of the thesis is addressed over the chapters 3 to 7. In particular, chapter 3 discusses in-depth why and how volatilomics data preprocessing should be performed, what steps need to be followed, and what algorithms have been developed for these purposes. Chapter 4 addresses one major problem that occurs if proper data preprocessing is not achieved, which is the batch effects. Although, batch effects may still appear even if proper data preprocessing has been done because of various reasons. Why existing algorithms fail to correct for batch effects in volatilomics data and what further preprocessing steps can be done to prevent or at least diminish even further inevitable batch effects are also covered in chapter 4. Chapter 5 presents ensemble techniques, advanced statistical modelling algorithms that can be applied to complex biological datasets, such as the volatilomics datasets, for predictive or classification purposes. Chapter 5 also describes how they can be properly run, optimized, and validated. Nevertheless, the application of these algorithms in itself is not always enough because of the complexity of the volatilomics data; therefore, data transformations might be needed to achieve proper classification or prediction. Chapter 6 emphasizes on the fact that, sometimes, data from one data platform are not always enough to cover the topic of interest fully. This is why data fusion should be considered. The major data fusion approaches are discussed and compared to another, more sophisticated fusion approach that is proposed here too. Chapter 7 presents a VOC case study, where exhaled breath VOC data from primary sclerosing cholangitis and inflammatory bowel disease patients were used to build a classification model and to find which VOCs are important in classifying the patients.

Finally, chapter 8 critically discusses what was achieved in the present work, whether the results of the present work met the thesis aims, and it provides suggestions as to where the field should be headed and future work that should be done to follow up on what was addressed here. Chapter 8 finishes with a concluding summary.

References

1. Davis, C., J. Pleil, and J. Beauchamp, Breathborne biomarkers and the human volatilome. 2020.
2. Chen, S., L. Zieve, and V. Mahadevan, Mercaptans and dimethyl sulfide in the breath of patients with cirrhosis of the liver: Effect of feeding methionine. Translational Research, 1970. 75(4): p. 628-635.
3. Pauling, L., et al., Quantitative analysis of urine vapor and breath by gas-liquid partition chromatography. Proceedings of the National Academy of Sciences, 1971. 68(10): p. 2374-2376.
4. Stavropoulos, G., et al., Liver Impairment—The Potential Application of Volatile Organic Compounds in Hepatology. Metabolites, 2021. 11(9): p. 618.
5. Bodelier, A.G., et al., Volatile Organic Compounds in Exhaled Air as Novel Marker for Disease Activity in Crohn's Disease: A Metabolomic Approach. Inflamm Bowel Dis, 2015. 21(8): p. 1776-85.
6. Smolinska, A., et al., Volatile metabolites in breath strongly correlate with gut microbiome in CD patients. Analytica chimica acta, 2018. 1025: p. 1-11.
7. Fijten, R.R.R., et al., The necessity of external validation in exhaled breath research: a case study of sarcoidosis. J Breath Res, 2017. 12(1): p. 016004.
8. Smolinska, A., et al., Profiling of volatile organic compounds in exhaled breath as a strategy to find early predictive signatures of asthma in children. PloS one, 2014. 9(4): p. e95668.
9. Schnabel, R., et al., Analysis of volatile organic compounds in exhaled breath to diagnose ventilator-associated pneumonia. Sci Rep, 2015. 5: p. 17179.
10. Van Der Schee, M.P., et al., Breathomics in lung disease. Chest, 2015. 147(1): p. 224-231.
11. Kneepkens, C.F., G. Lepage, and C.C. Roy, The potential of the hydrocarbon breath test as a measure of lipid peroxidation. Free Radical Biology and Medicine, 1994. 17(2): p. 127-160.
12. Vary, T.C., et al., A biochemical basis for depressed ketogenesis in sepsis. The Journal of trauma, 1986. 26(5): p. 419-425.
13. Kansal, A., Sources and reactivity of NMHCs and VOCs in the atmosphere: A review. Journal of hazardous materials, 2009. 166(1): p. 17-26.
14. Pauwels, C.G., et al., Smoking regular and low-nicotine cigarettes results in comparable levels of volatile organic compounds in blood and exhaled breath. Journal of Breath Research, 2020. 15(1): p. 016010.
15. Amann, A., et al., The human volatilome: volatile organic compounds (VOCs) in exhaled breath, skin emanations, urine, feces and saliva. Journal of breath research, 2014. 8(3): p. 034001.
16. Miekisch, W., J.K. Schubert, and G.F. Noeldge-Schomburg, Diagnostic potential of breath analysis—focus on volatile organic compounds. Clinica chimica acta, 2004. 347(1-2): p. 25-39.
17. Doran, S.L., A. Romano, and G.B. Hanna, Optimisation of sampling parameters for standardised exhaled breath sampling. Journal of breath research, 2017. 12(1): p. 016007.
18. Herbig, J. and J. Beauchamp, Towards standardization in the analysis of breath gas volatiles. Journal of breath research, 2014. 8(3): p. 037101.
19. Gaude, E., et al., Targeted breath analysis: exogenous volatile organic compounds (EVOC) as metabolic pathway-specific probes. Journal of breath research, 2019. 13(3): p. 032001.
20. Horváth, I., et al., A European Respiratory Society technical standard: exhaled biomarkers in lung disease. European Respiratory Journal, 2017. 49(4): p. 1600965.
21. Stavropoulos, G., et al., Preprocessing and analysis of volatilome data, in Breathborne Biomarkers and the Human Volatilome. 2020, Elsevier. p. 633-647.
22. Stavropoulos, G., et al., Implementation of quality controls is essential to prevent batch effects in breathomics data and allow for cross-study comparisons. Journal of Breath Research, 2020. 14(2): p. 026012.
23. Upton, G. and I. Cook, A Dictionary of Statistics 2 rev. 2008.
24. Bro, R. and A.K. Smilde, Principal component analysis. Anal. Methods, 2014. 6(9): p. 2812-2831.
25. Breiman, L., Random Forest. Machine Learning, 2001. 45: p. 5-32.
26. Stavropoulos, G., et al., Random Forest and Ensemble Methods, in Comprehensive Chemometrics: Chemical and Biochemical Data Analysis. 2020, Elsevier BV. p. 661-672.

27. Barker, M. and W. Rayens, Partial least squares for discrimination. Journal of Chemometrics: A Journal of the Chemometrics Society, 2003. 17(3): p. 166-173.
28. Wold, S., et al., The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. SIAM Journal on Scientific and Statistical Computing, 1984. 5(3): p. 735-743.
29. Blanchet, L., et al., Constructing bi-plots for Random Forest: tutorial. Analytica Chimica Acta, 2020.
30. Stavropoulos, G., et al., Advanced data fusion: Random forest proximities and pseudo-sample principle towards increased prediction accuracy and variable interpretation. Analytica Chimica Acta, 2021: p. 339001.
31. Smolinska, A., et al., Interpretation and visualization of non-linear data fusion in kernel space: study on metabolomic characterization of progression of multiple sclerosis. PLoS One, 2012. 7(6).
32. Krooshof, P.W., et al., Visualization and recovery of the (bio) chemical interesting variables in data analysis with support vector machine classification. Analytical Chemistry, 2010. 82(16): p. 7000-7007.

CHAPTER 2

Liver impairment—the potential application of volatile organic compounds in Hepatology

Georgios Stavropoulos, Kim van Munster, Giuseppe Ferrandino, Marius Sauca, Cyriel Ponsioen, Frederik-Jan van Schooten, Agnieszka Smolinska

Metabolites (ISSN 2218-1989),
doi: 10.3390/metabo11090618

Abstract

Background: Liver diseases are currently diagnosed through liver biopsy. Its invasiveness, costs, and relatively low diagnostic accuracy require new techniques to be sought. Analysis of volatile organic compounds (VOCs) in human bio-matrices has received a lot of attention. It is known that a musty odour characterises liver impairment, resulting in the elucidation of volatile chemicals in breath, and other body fluids such as urine and stool, that may serve as biomarkers of a disease.

Aims: To review all the studies found in the literature regarding VOCs in liver diseases, and to summarise all the identified compounds that could be used as diagnostic or prognostic biomarkers.

Methods: The literature search was conducted on ScienceDirect and PubMed, and each eligible publication was qualitatively assessed by two independent evaluators using the SANRA critical appraisal tool.

Results: 58 publications were found, 28 were kept for inclusion—23 were about VOCs in the breath, one in the bile, three in urine, and one in faeces. Each publication was graded from zero to 10.

Conclusion: A graphical summary of the metabolic pathways showcasing the known liver disease-related VOCs and suggestions on how VOC analysis on liver impairment could be applied in clinical practice are given.

Keywords: VOCs, liver diseases, breath, faeces, bile, urine, non-invasive

Introduction

Fetor hepaticus, a musty breath aroma, has been among the most prominent liver insufficiency signs available to clinicians, and it was in the seventies when Chen et al. [1] identified the first responsible compounds. The authors reported that several mercaptans and aliphatic acids (i.e., predominantly acetic and propionic acid) were elevated in the exhaled breath of individuals with liver cirrhosis [2]. However, it was not until the nineties that Tangerman et al. [3] pinpointed dimethyl-sulphide as the primary source of fetor hepaticus. These studies [1-3] were the first liver-related volatile organic compound (VOC) analyses in the breath and paved the way for further research in the field. Many pathophysiological conditions in the liver alter various hepatic metabolic pathways, modifying the abundance of specific exhaled VOCs. Derivatives of cell membrane peroxidation can generate different VOCs as a result of oxidative stress in hepatic inflammation. Metabolic pathway alterations can lead to increased amounts of several compounds such as sulphur derivatives through the incomplete transamination of sulphur-containing amino acids [1] or ammonia through the altered urea cycle [4]. Elevated ketones can result from a combination of impaired hepatic gluconeogenesis, increased insulin resistance, and glycogen exhaustion [5], whereas exhaled acetic and propionic acid increase due to reduced hepatic clearance of short-chain fatty acids from the gut microbiome as a result of increased sinusoidal pressure and portosystemic shunts [1]. Many liver diseases that ensue in the sequence of hepatitis, fibrosis, cirrhosis, and end-stage liver failure still pose diagnostic and monitoring challenges; non-alcoholic fatty liver disease (NAFLD), non-alcoholic steatohepatitis (NASH), autoimmune hepatitis (AH), chronic cholestatic diseases including primary sclerosing cholangitis (PSC) and primary biliary cirrhosis are such examples. All these conditions require an invasive liver biopsy for diagnosis, which frequently does not confirm but rather suggest a specific diagnosis. Metabolically, the liver is the main active organ; therefore, VOC analysis in the breath and other body fluids or faeces could hold great noninvasive, patient-friendly potential for diagnostic purposes and for gauging functional reserve of liver impairment.

Liver pathophysiology and liver function tests

A wide variety of viral, immune-mediated, cholestatic, and toxic conditions may cause chronic liver tissue inflammation. In response to this, the liver accumulates extracellular matrix components, leading to fibrous tissue and scarring [6, 7]. In prolonged and severe liver damage, fibrosis might turn into cirrhosis and end-stage liver disease. Substantial liver damage leads to impaired liver function, causing health issues such as disturbed coagulation and hepatic encephalopathy. Moreover, increased hepatic flow resistance leads to portal hypertension that causes hemodynamic insufficiency, which subsequently leads to ascites, varices, and several other critical conditions [8]. Finally, liver cirrhosis is a premalignant condition with an increased risk for hepatocellular carcinoma [9]. Diagnosis and monitoring of liver disease progression

are essential to establish an optimal treatment strategy and evaluate therapeutic effects [10]. However, only a handful of biomarkers demonstrate sufficient specificity and sensitivity to develop a reliable diagnosis and monitoring of chronic liver injury. For example, anti-mitochondrial antibodies are used to diagnose primary biliary cholangitis, whereas polymerase chain reaction is used for viral hepatitis. Although, both examples fail to tell something about the severity of liver injury. Liver biopsy is considered the reference method for diagnosis and evaluation of liver impairment; although its invasiveness and cost make it less suitable for frequent sampling. Additionally, in some liver diseases such as cholestatic liver diseases, liver fibrosis is patchy and not homogenous, which decreases the representability, and thus, accuracy of the biopsy.

In the past few decades, several noninvasive biomarkers have entered the liver research field, some of which have already been used in clinical trials, and the most widely used are the enhanced liver fibrosis score (ELF) [11], the FibroTest [12], and the Pro-C3 [13]. All these biomarkers measure molecules involved in fibrogenesis or fibrinolysis; however, they are influenced by confounding factors (e.g. fibrous tissue elsewhere), leading to suboptimal sensitivity and specificity [14]. Moreover, liver fibrosis can be detected through imaging techniques such as ultrasound elastography, which measures liver stiffness (liver fibrosis has been associated with liver stiffness) and is currently widely used in clinical trials and daily clinical practices. Other imaging techniques include magnetic resonance imaging (MRI), computed tomography (CT), or magnetic resonance elastography. However, other pathophysiological processes that increase liver stiffness, such as cholestasis, decrease elastography reliability in its capability to measure fibrosis [14]. Concerning the liver functional reserve, which is vital to determine the moment patients qualify for liver transplantation, the end-stage liver disease model (MELD) is widely applied [15]. This model uses serum bilirubin, the international normalised ratio (INR) for prothrombin time (i.e. a measure of clotting factors), and serum creatinine; these parameters combined to constitute a model as a proxy for the liver function that predicts mortality within 90 days. Mortality and disease severity should be considered; however, the combination of such parameters makes the model dependent on a kidney function read-out, which is not an optimal solution either [16]. Despite the different invasive and noninvasive methods to assess liver diseases, more than 50% of the cases are detected at advanced stages when decompensation episodes occur [8, 17]. As a result, the need for new, reliable, and effective biomarkers in the context of liver function or disease diagnosis for example, remains.

Breath tests are already used in clinical setups; an example is identifying *Helicobacter pylori* infection via the C13 urea breath test [18]. Here, labelled C13 urea is administered to patients, and then their exhaled breath is collected, where the isotope-labelled carbon dioxide is measured. Other C13 breath tests, such as the C13 aminopyrine breath test, have also been used to examine liver diseases [19, 20]; however, C13

implementations are outside the scope of the present review since they are not based on VOC analysis. The current review focuses on endogenously formed compounds that have been connected with liver impairment, among which are nitrogen derivates [4], ketones [21], alkanes [21], sulphur derivates [1], and alcohols [22].

VOC analysis

In human research, VOCs arise from different body matrices such as breath, faeces, urine, bile, breast milk, and blood, resulting from exogenous or endogenous sources [23-25]. Exogenous VOCs 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 [26]. A compound is considered endogenous when its concentration in a subject/patient sample is higher than in ambient air [27, 28]. Endogenous VOCs are produced biochemically by body cells and tissues, such as lung and airway tissues, or from other organ tissues (e.g. liver or kidney) [29]; these VOCs are a reflection of the biochemical reactions such as apoptosis, inflammation or oxidative stress [30-32]. These VOCs arise from body chemical reaction cascades in diseased individuals due to cellular damage [33]; they are released in the bloodstream and spread among the body excretions. In particular, liver diseases alter VOC abundances in the blood [34, 35], leading to different amounts of VOCs present in body excretions.

Many studies explored different approaches to quantifiably detect VOCs in liver disease patients [22, 34-36]. The vast majority of these studies examined breath as the means of discovering discriminatory VOCs, whereas only a handful of studies used body excretions other than breath [24, 37, 38]. Thus far, examining liver diseases via VOC analysis has mainly focused on cirrhosis and NAFLD, and currently, no VOC detection test has been implemented in the clinics yet, despite the diagnostic potential of VOC analysis, in general [39-41]. This review aims to discuss the available VOCs literature on liver diseases examined through, primarily, breath, and secondarily, through faeces, urine, and the bile. Finally, conclusions on possible causes for the lack of clinical VOC tests for liver diseases are drawn, and possible future directions are suggested.

Materials & methods

Literature search

The scientific literature search focused on liver disease diagnosis, prognosis, and monitoring via VOCs in the breath or faeces. For breath related VOCs, PubMed and ScienceDirect were interrogated with the following search terms:

(((((liver disease) OR "Liver Diseases"[Mesh]) OR ((Diagnosis/Broad[filter]) AND ("Liver Diseases"[Mesh]))))) AND ((volatile organic compounds) OR "Volatile Organic Compounds"[Mesh])) AND ((breath analysis) OR "Breath Tests" [Mesh]).

The search terms for faeces were:
(((((“Liver Diseases"[Mesh]) OR liver disease) OR ((Diagnosis/Broad[filter]) AND ("Liver Diseases"[Mesh]))))) AND ((volatile organic compounds) OR "Volatile Organic Compounds"[Mesh])) AND (((fecal analysis) OR faecal analysis) OR "Feces" [Mesh]).

Replacing the word “Diagnosis” with “Prognosis” or “Monitoring” yielded the same results for both biological matrices. Additional studies cited by the initially identified research papers were also included and discussed in this review. These additional studies examined liver diseases related to VOCs in the breath and faeces and other body fluids such as urine, blood, and bile. The number of the latter was minimal; therefore, it was decided to discuss these as well. Only articles published in English, reporting original research in humans, and focused on different VOC patterns between healthy and diseased liver subjects were included. Engineering or technical studies were excluded since they fall outside the scope of this review. Finally, no year of publication criterion was imposed as an exclusion criterion. An overview of the literature search and the exact numbers of the publications found and used herein can be seen in the Results section in Figure 1.

Quality assessment

Two independent evaluators assessed the eligible studies using the Scale for the Assessment of Narrative Review Articles (SANRA) [42]. SANRA is a brief critical appraisal tool used to assess the quality of narrative reviews and research articles, and it consists of a six-question questionnaire. Each question is evaluated on a scale from zero to two (i.e., 0, 1, and 2), resulting in a maximum cumulative score of 12 for the paper at hand. However, in the present review, question number three (“Description of literature search”) was excluded from the evaluation of the papers because it is not applicable for scientific research papers. The whole SANRA questionnaire can be found elsewhere [42]. As a result, the SANRA assessment score was on a scale from zero to 10. Papers with a maximum aggregate score of five (i.e. (0-5]) were considered as low-quality, those with a total score from five to seven (i.e. (5-7]) were regarded as a medium-quality, and those with an aggregate score from seven to ten (i.e. (7-10]) were considered as high-quality. However, the SANRA quality assessment tool was deemed not strict enough when the assessment was finalised (i.e. almost all the papers were scored with eight or more; the scores are illustrated in the Results section, in Table 1). This is because the questions are made to assess general scientific guidelines; thus, five additional assessment questions were included in the overall assessment. The two assessors construed these questions following the present review purposes;

these questions can be seen in the Supporting information in Table S1. The new questions were also graded on a scale from zero to two (the same as the SANRA questions), and the new scores (i.e. from the five SANRA questions and the added five summed up) are also illustrated in Table 1.

Results

The literature search performed in both PubMed and ScienceDirect resulted in 58 hits in total, of which one was not accessible, 16 were either engineering or technical, and 13 were reviews. Thus, the final number of papers to be discussed here was 28. From these 28 articles, 23 found VOCs in the breath, one in the bile, three in urine, and one in faeces.

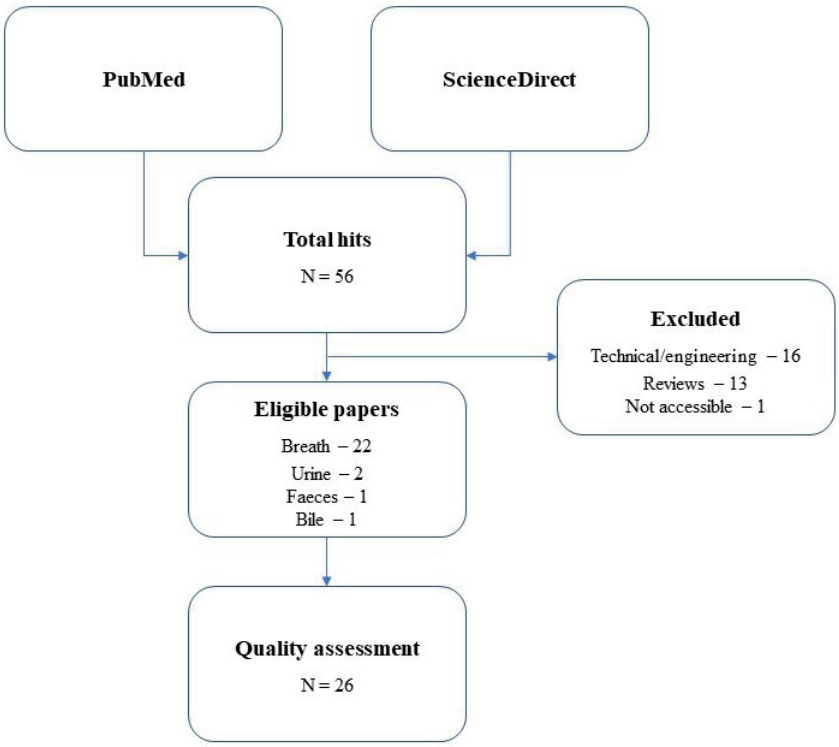


FIGURE 1: SCHEMATIC REPRESENTATION OF THE LITERATURE SEARCH PERFORMED IN THE PRESENT REVIEW. THE TOTAL NUMBER OF PAPERS FOUND IS 58, AND THE NUMBER OF PUBLICATIONS ELIGIBLE TO BE REVIEWED IS 28.

Figure 1 represents a scheme of the literature search that was performed here. Table 1 shows the average over the two independent evaluators' scores per publication and the publication categorisation into low, medium, and high-quality paper.

TABLE 1: EVALUATION OF THE PAPERS THAT WERE INCLUDED IN THE PRESENT REVIEW. BOTH SCORE COLUMNS (I.E. SANRA SCORES AND SANRA & ADDED QUESTIONS SCORES) ARE AVERAGED OVER THE TWO INDEPENDENT EVALUATORS. THE QUALITY OF THE PAPERS IS CHARACTERISED AS LOW (I.E. [0-5]), MEDIUM (I.E. (5-7]), OR HIGH (I.E. (7-10]).

Publication	Means of analysis	SANRA scores (averaged)	SANRA & added questions scores (averaged)	Quality
Friedman et al. 1994 [1]	Breath	6.5	6.25	Medium
Hiroshi et al. 1978 [2]	Breath	7	5	Low
Letteron et al. 1993 [3]	Breath	9	6.5	Medium
Van den Velde et al. 2008 [4]	Breath	9.5	9.25	High
Dadamio et al. 2012 [5]	Breath	10	8.25	High
Pijls et al. 2016 [6]	Breath	10	8	High
Morisco et al. 2013 [7]	Breath	9	8.25	High
Del Rio et al. 2015 [8]	Breath	9	8	High
Eng et al. 2015 [9]	Breath	9.5	7.25	High
Alkhouri et al. 2015 [10]	Breath	10	7.25	High
De Vincentis et al. 2016 [11]	Breath	9	5.75	Medium
Khalid et al. 2013 [12]	Breath	9	6.75	Medium
O'Hara et al. 2016 [13]	Breath	10	8.5	High
Arasaradnam et al. 2015 [14]	Breath	9	5.5	Medium
Solga et al. 2006 [15]	Breath	9	6.75	Medium
Verdam et al. 2013 [16]	Breath	9	6.25	Medium
Alkhouri et al. 2013 [17]	Breath	9.5	6.75	Medium
Millonig et al. 2010 [18]	Breath	7.5	7.5	High
Hanouneh et al. 2014 [19]	Breath	9	7.75	High
Qin et al. 2010 [20]	Breath	7.5	6	Medium
Sinha et al. 2019 [21]	Breath	10	7	Medium
Ferrandino et al. 2020 [22]	Breath	10	7	Medium
Miller-Atkins et al. [23]	Breath	10	8.75	High
Raman et al. 2013 [24]	Faeces	9	6.75	Medium
Navaneethan et al. 2015 [25]	Bile	9	6.75	Medium
Navaneethan et al. 2015 [26]	Urine	9	6.75	Medium
Arasaradnam et al. 2012 [27]	Urine	8.5	6	Medium
Bannaga et al. 2021 [28]	Urine	9.5	7	Medium

Table 2 summarises all the compounds that were found as significant in more than one of the examined research papers analysed in the present review. Table 2 also describes what is believed to be the biological origin of each of the present compounds.

TABLE 2: A SUMMARY OF THE COMPOUNDS THAT WERE FOUND AS SIGNIFICANT IN MORE THAN ONE OF THE EXAMINED RESEARCH PAPERS IN THE PRESENT REVIEW. WHAT IS BELIEVED TO BE THE BIOLOGICAL ORIGIN OF EACH COMPOUND IS DESCRIBED HERE TOO.

Compound	Number of times	Biological origin
Dimethyl-sulphide	11	Incomplete metabolism of sulphur-containing amino acids in the transamination pathway – Cytochrome C oxidase deficiency
Limonene	7	Limonene is not produced in the human body – metabolised by the P450 enzymes CYP2C9 and CYP2C19 – accumulates in the fat of patients
Acetone	7	Due to hepatic insulin resistance that leads to an increase in triglycerides, free fatty acids and ketones
Ethanol	7	Due to increased shunting volumes through portocaval shunts
Isoprene	6	A by-product of cholesterol biosynthesis – the intestinal microbiota may generate isoprene too
Acetaldehyde	6	Oxidation product in ethanol metabolism – CYP2E1 is induced
2-Pentanone	5	Due to hepatic insulin resistance – inhibition of CYP2E1
Carbon-disulphide	4	The oxidative metabolism of carbon disulphide – also due to incomplete metabolism of sulphur-containing essential systems
2-Butanone	4	Due to hepatic insulin resistance, formed during lipolysis – inhibition of CYP2E1
Benzene	4	Environmental pollutant
Pentane	3	Lipid peroxidation – a by-product of the cytochrome P450 metabolism
Hydrogen-sulphide	3	Incomplete metabolism of sulphur-containing amino acids in the transamination pathway – cytochrome C oxidase deficiency (less stable than dimethyl-sulphide)
Ethane	3	Lipid peroxidation of polyunsaturated fatty acids – a by-product of the cytochrome P450 metabolism
Trimethyl-amine (TMA)	3	The intestinal microflora degrades dietary phosphatidylcholine to form trimethylamine – trimethylamine is metabolised by the hepatic flavin monooxygenase family of enzymes
2-Nonene	3	It is yet to be discovered – it has been linked to oxidative stress
2-Propanol	2	It is yet to be discovered – it is speculated to be related to inflammatory processes and/or lipid peroxidation
Indole	2	Derived from the catabolism tryptophan
Dimethyl-selenide	2	Excretion product of the essential micronutrient selenium
Methanol	2	Metabolised mainly by alcohol dehydrogenase – pectin degradation – an imbalance of microflora composition in cirrhotic patients

Compound	Number of times	Biological origin
2-Octanone	2	Due to hepatic insulin resistance, formed during lipolysis – inhibition of CYP2E1
Octane	2	Metabolised by the cytochrome P450 enzymes
Alpha-pinene	2	Metabolised by the cytochrome P450 enzymes
Tridecane	2	It is yet to be discovered – it is speculated that it is related to inflammatory processes and/or lipid peroxidation
Styrene	2	Exogenous sources such as industrial materials – it is oxidised by cytochrome P450

Discussion

Differentiation among general cirrhotic CLD, non-cirrhotic CLD, and healthy individuals

Pauling et al. pioneered breath testing with their unprecedented study published in 1971 [43]. Since then, the 500+ discovered VOCs provided insights into the human body metabolic processes. Lipid peroxidation has been associated with alkanes such as pentane and ethane, whereas cholesterol metabolism has been linked to isoprene and other unsaturated compounds [28, 29, 44, 45]. Dextrose metabolism has been correlated with ketones such as acetone, while the sulphur-containing compounds dimethyl-sulphide, methyl-mercaptans, and ethyl-mercaptans, have been associated with renal failure or liver disease and deemed the cause of fetor hepaticus of cirrhotic patients [28, 29, 44, 45]. Initial studies mainly focused on finding biomarkers related to liver cirrhosis. Hiroshi et al. [46], Tangerman et al. [47], and Friedman et al. [48] paved the way for modern liver breath analysis by comparing cirrhotic patients to healthy controls aiming to identify compounds that differ between the two cohorts by exploiting advances of the gas chromatography-mass spectrometry (GC-MS) technology. All three studies found significantly higher levels of dimethyl-sulphide in the breath of cirrhotic patients. However, Friedman et al. [48] also reported that hydrogen-sulphide was substantially higher in patients with less severe forms of cirrhosis than healthy controls. More interestingly, they also found elevated levels of limonene in half of the cirrhotic patients. The additionally reported compounds in the [48] study might have resulted from the fact that the GC detector used was different than the one used in the [46, 47] studies.

Van den Velde et al. [34] and Dadamio et al. [49] also analysed liver cirrhosis patients' and healthy controls' breath to identify VOCs related to liver cirrhosis by using GC-MS. Van den Velde et al. found that acetone, dimethyl-sulphide, 2-butanone, and 2-pentanone were elevated, while indole and dimethyl-selenide were reduced in the

patients compared to controls. The discriminative model based on these compounds showed a sensitivity and specificity of 100% and 70%, respectively. Dadamio et al. found more than 20 compounds elevated in the breath of cirrhotic patients. The resulting classification models provided an overall average sensitivity and specificity of 83% and 100%, respectively. Morisco et al. [22] also stratified cirrhotic patients and healthy volunteers to evaluate the capability of breath testing in distinguishing among different levels of disease severity in addition to liver cirrhosis diagnosis employing proton transfer reaction-MS (PTR-MS). Twelve compounds (i.e. heptadienol, methanol, 2-butanone, 3-pentone, 2-octanone, C8-ketone, 2-nonanone, C9-ketone, monoterpene, p-cymene, sulphoxide compounds, an S-compound, an NS-compound, and an N-compound) had significantly higher concentrations, except for the S-compound, which had significantly lower concentration, in liver cirrhosis patients compared to controls. Morisco et al. [22] further stratified their patients into two groups (i.e. mild cases and severe cases) to assess the different VOC concentrations according to disease severity. They found that five VOCs (i.e. heptadienol, C8-ketone, monoterpene (tentatively identified as limonene), 2-butanone, and an NS-compound) had higher concentrations in the severe cases, while the S-compound and the N-compound had lower concentrations in the severe cases. Limonene had the highest diagnostic performance with a sensitivity and specificity of 83% and 86%, respectively. Mild cases were discriminated from controls with a sensitivity and specificity of 83% and 86%, respectively, and with a sensitivity and specificity of 100% from the severe cases. Interestingly, the monoterpene, tentatively identified as limonene, had the highest diagnostic performance again with a sensitivity and specificity of 100% when discriminating mild from severe cases. In general, the [22] study found different compounds than the [34, 49] studies (Table 3); however, the chemical classes of the discovered VOCs were the same (i.e. sulphur compounds and ketones). PTR-MS seems to provide a more complex picture of the breath compounds in liver cirrhosis patients and it seems to be able to distinguish between different disease severity classes, which may explain the identification of different compounds in the [22] study. Noteworthy, the [34, 49] studies did not enforce a fasting state for their volunteers, whilst the [22] study did, and fasting could explain the appearance of ketone bodies in the breath.

In 2015, Del Rio et al. [50] also compared cirrhotic patients against healthy cohorts and aimed to identify breath biomarkers of liver diseases by employing PTR-MS. Cirrhotic patients who had undergone a liver transplant were compared to their pre-transplant samples, effectively becoming their controls and allowing liver metabolism-related compounds isolation. It was found that methanol, 2-butanone, carbon disulphide, 2-pentanone, and limonene presented significantly higher concentrations in liver cirrhosis patients than in controls (Table 3). Limonene levels were monitored in post-liver transplant patients, and they were steadily decreasing in the following days. Results generated by this study design support Del Rio et al. claim that all previous studies were only hypothesis-generating since there was a lack of follow-up

to confirm the found biomarkers. These findings also highlight limonene potential as a liver function biomarker in liver transplant patients by monitoring its wash-out [50]. It should be noted, however, that post-liver transplantation, other factors could have influenced the limonene levels such as reduced food intake in the first days after the operation.

Pijls et al. [51] stratified CLD patients with or without cirrhosis and aimed to identify a VOC profile to separate the classes using GC-MS. They identified 11 VOCs (i.e. dimethyl-sulphide, terpene (limonene), 2-methybutanal, propanoic acid, octane, terpenoid, 3-carene, 1-hexadecanol, an unknown compound, as well as a branched C16H34) that discriminate between non-cirrhotic CLD and cirrhotic CLD patients with an accuracy of 84.1% (Table 3).

De Vincentis et al. [52] also compared cirrhotic against non-cirrhotic patients and healthy controls using the emerging e-nose technology, which provides rapid breath-prints (BPs). This technique offers a VOC profile on a point-of-care base because it can be performed instantaneously in an outpatient care setting. De Vincentis et al. identified BPs that discriminate different liver disease severity stages among liver cirrhosis patients with a sensitivity and specificity of 87.5% and 64.7%, respectively. Differences among patients with infectious and non-infectious liver diseases were also achieved with a sensitivity and specificity of 29% and 88%, respectively (Table 3). It is worth mentioning that in a follow-up study, De Vincentis et al. [53] showed that e-nose could significantly identify cirrhotic patients with a high risk of hospitalisation and mortality, thus, making it a substantial alternative to the Child-Pugh and MELD scores in clinical practices, which are considered as the reference method. Successful e-nose discriminatory capabilities have been reported already [54, 55].

In 2015, Eng et al. [56] conducted the first reported paediatric study to differentiate cirrhotic children from healthy children by using the newly developed selected ion flow-tube-MS (SIFT-MS). They identified 1-decene, 1-heptene, 1-octene, and 3-methyl-hexane as significantly higher in cirrhotic children than in controls. These VOCs were also increased in children with advanced liver fibrosis compared to children suffering from no to mild fibrosis. Additionally, 1-nonene, (E)-2-nonene, and dimethyl-sulphide were lower in cirrhotic children than controls and inversely proportional to the degree of liver fibrosis. This finding is unexpected and contradicts previous studies conducted in adults [22, 34, 49, 50], where dimethyl-sulphide was elevated in adult liver disease patients. However, this inconsistency may be explained by differences in hepatic metabolism between children and adults [57]. Eng et al. also generated a predictive model by combining five VOCs (i.e. 1-octene, triethyl-amine, ethane, E2-nonene, and 1-decene) that showed prediction accuracy of cirrhosis with an AUC of 0.97 (Table 3).

Origin of the VOCs reported in general cirrhotic CLD against healthy individuals

The most significant compounds, and the ones that the aforementioned literature (section 4.1) seems to be more certain about their origin, are limonene and dimethyl-sulphide. Limonene is suggested to originate from foods and drinks. Limonene is broken down in the liver by CYP2C19 and CYP2C9 enzymes into other compounds such as perillyl alcohol, trans-isopiperitenol, and trans-carveol [58]. In liver impairment, the CYP2C19 and CYP2C9 enzymes are proportionally reduced and thus, leading to increased limonene levels in the body [22, 48, 50]. Increased dimethyl-sulphide, along with other sulphur-containing compounds, points toward incomplete metabolism of sulphur-containing amino acids in the transamination pathway due to liver impairment. As far as other groups of compounds are concerned, the aforementioned literature also discusses possible metabolic pathways that might be involved in their origin, and they can be summarised as follows. It is suggested that free fatty acids, triglycerides, and ketones such as 2-butanone, 2-pentanone, and acetone may increase due to hepatic insulin resistance [22, 34], which favours lipolysis and free fatty acid beta-oxidation. As for reduced indole and phenol levels, they may have resulted from the impaired ability of the liver to degrade aromatic amino acids such as tryptophan [22, 34], whereas the reduced dimethyl-selenide is explained by lower levels of this micronutrient observed in the blood of patients with cirrhosis [59]. Increased levels of hydrocarbons, such as ethane and pentane, were attributed to the impaired conversion of saturated hydrocarbons into alcohols due to deficient cytochrome P450 activity [34, 49]. Cirrhotic liver inability to metabolise methanol by efficiently using alcohol dehydrogenase [50] or an imbalance in the bacterial flora composition [22] explain the increased methanol levels in liver disease patients, which alters the colon fermentation processes. Finally, high levels of other alkanes such as 3-methyl-trexane, 1-decene, 1-heptene, and 1-octene are thought to be related to oxidative stress [56]. Figure 2 illustrates these suggested pathways.

TABLE 3: SUMMARY OF THE PAPERS THAT EXAMINED CIRRHOSIS/CLD PATIENTS AGAINST HEALTHY COHORTS. THE ARROWS SHOW THE VOC ABUNDANCE IN THE CLD GROUP COMPARED TO THE HEALTHY GROUP IN THE STUDY DESIGN.

Author/Year	Study design	Analytical method	VOCs identified as significant	Discriminatory performance
Friedman et al.1994	24 cirrhotic CLD vs 24 healthy	GC-MS	Hydrogen-sulphide ↑ Limonene ↑	Not reported
Van den Velde et al. 2008	52 cirrhotic CLD vs 50 healthy	GC-MS	Acetone ↑ Dimethyl-sulphide ↑ 2-butanone ↑ 2-pentanone ↑ Indole ↓ Dimethyl-selenide ↓	100% sensitivity 70% specificity
Dadamio et al. 2012	35 cirrhotic CLD vs 49 healthy	GC-MS	Dimethyl-sulphide ↑ Acetone ↑ 2-butanone ↑ 2-pentanone ↑ Indole ↓ Phenol ↓ Dimethyl-selenide ↓ Isoprene ↑ Ethane ↑ Pentane ↑	83% sensitivity 100% specificity
Morisco et al. 2013	12 cirrhotic CLD vs 14 healthy	PTR-MS	Heptadienol ↑ Methanol ↑ 2-butanone ↑ 3-pentone ↑ 2-octanone ↑ 2-nonanone ↑ Monoterpene ↑ P-cymene ↑	83% sensitivity 86% specificity
Del Rio et al. 2015	31 cirrhotic CLD vs 30 healthy	PTR-MS	Methanol ↑ 2-butanone ↑ Carbon-sulphide ↑ 2-pentanone ↑ Limonene ↑	97% sensitivity 70% specificity
Pijls et al. 2016	34 cirrhotic CLD vs 87 non-cirrhotic CLD	GC-MS	Dimethyl-sulphide ↑ Terpene (limonene) ↑ 2-methyl-butanal ↓ Propanoic acid ↑ Octane ↑ Terpenoid ↑ 3-carene ↑ 1-hexadecanol ↓ C16H34 ↓	83% sensitivity 87% specificity
De Vincentis et al. 2016	65 cirrhotic CLD vs 39 non-cirrhotic CLD	E-nose	Not available	86.2% sensitivity 98.2% specificity
Eng et al. 2015	49 cirrhotic CLD children vs 55 healthy children	SIFT-MS	1-decene ↑ 1-heptene ↑ 1-octene ↑ 3-methyl-hexane ↑ 1-nonene ↓ (E)-2-nonene ↓ Dimethyl-sulphide ↓	0.97 AUC

Differentiation among specific cirrhotic CLD, non-cirrhotic CLD, and pre-cirrhotic CLD

VOCs in advanced versus mild fibrosis patients

In 2013, Alkhouri et al. [60] assessed the utility of breath VOC measurements to diagnose advanced fibrosis in CLD patients by employing SIFT-MS. They found reduced acetone, benzene, carbon disulphide, isoprene, pentane, and ethane in the breath of patients with advanced fibrosis compared to those with minimal fibrosis (Table 4). Isoprene had the highest AUC for advanced fibrosis (i.e. AUC = 0.855), and 75% of the patients were correctly classified as advanced fibrosis using certain cut-off levels for isoprene.

VOCs in cirrhotic patients with hepatic encephalopathy or hepatocellular cancer

Hepatic encephalopathy (HE) was investigated by Khalid et al. [61]. They sampled alcoholic cirrhotic patients, of which some had HE and some others did not have HE, along with a few non-alcoholic cirrhotic patients, harmful drinkers, and healthy volunteers; ultimately, they aimed to differentiate cirrhotic HE patients from cirrhotic patients without HE or harmful drinkers by using GC-MS. They reported that methyl-vinyl ketone and, likely, isothiocyanato-cyclohexane contributed to the group separation of alcoholic cirrhotic patients with HE and without HE. The model yielded a 90% sensitivity and 87% specificity. Undecane and an unknown compound contributed to the separation of alcoholic and non-alcoholic cirrhotic patients without HE, and the model yielded 78% sensitivity and 69% specificity. 1-methyl-4-(1-methyl-ethenyl)-benzene (p-cymenene) and two unknown compounds contributed to the group separation of alcoholic cirrhotic patients and harmful drinkers without cirrhosis, and the model yielded 88% sensitivity and 85% specificity. Octanal, a compound tentatively identified as 2,6-dimethyl-7-octen-2-ol, and an unknown compound contributed to distinguishing harmful drinkers from healthy volunteers, and the model yielded 71% sensitivity and 93% specificity. Methyl-vinyl ketone and an unknown compound allowed for the discrimination of non-alcoholic cirrhotic patients from healthy controls, and the model yielded 92% sensitivity and 100% specificity. Finally, heptane, 1-methyl-2-(1-methyl-ethyl)-benzene, phellandrene, and 2-methyl-hexane contributed to discriminating the alcoholic cirrhotic group from the healthy volunteers, and the model yielded 97% sensitivity and 93% specificity.

In 2016, O'Hara et al. [45], a follow-up of the [50] study, stratified the population of cirrhosis patients based on the presence of HE and investigated variations in limonene, methanol, and 2-pentane by using PTR-MS measurements. They found that limonene was higher in the breath of patients with HE and was the only compound able to discriminate from non-HE patients. In contrast, 2-penatanone could not discriminate against cirrhotic patients stratified by the presence/absence of HE complication. However, they did not provide sensitivity and specificity results.

Qin et al. [62] compared healthy volunteers, cirrhotic patients without hepatocellular cancer (HCC), and non-cirrhotic patients with HCC to find breath biomarkers that could be used to diagnose HCC patients—they ran a GC-MS/solid-phase micro-extraction analysis (SPME). 3-hydroxy-2-butanone, styrene, and decane appeared the most promising breath biomarkers for HCC patients. 3-hydroxy-2-butanone was the only one that was significantly different among all three groups, and it could discriminate between healthy volunteers and HCC groups with a sensitivity and specificity of 83.3% and 91.7%, respectively. In contrast, the diagnostic accuracy between HCC and cirrhosis groups was lower, with a sensitivity and specificity of 70% and 70.4%, respectively (Table 4). Styrene was not significantly different between the healthy volunteers and HCC groups, while decane was not significantly different between the cirrhosis and HCC groups. These compounds were significantly higher in HCC patients than in healthy volunteers, which suggests that these VOCs result from cancer metabolism, and thus, they may serve as breath biomarkers of HCC. The [45] study also examined VOCs in HCC patients; however, its results are different from those in [62]. The former study only found that HCC patients had significantly lower limonene levels than patients without HCC. These differences might be because the [45] study used PTR-MS instead of GC-MS/SPME that the [62] study used.

Ferrandino et al. [63] followed up on the limonene-related hypothesis and by sampling cirrhotic patients, cirrhotic patients with HCC, and healthy controls, they focused on comparing the exhaled limonene levels of their groups and see how they relate with each other by performing a GS-MS analysis. They reported that limonene concentration was significantly higher in cirrhotic and cirrhotic patients with HCC when compared to healthy individuals. However, no significant differences in limonene levels were found between the two diseased groups. They also reported that limonene levels correlate with serum bilirubin but not with alanine transferase. Consequently, Ferrandino et al. confirmed that breath limonene levels do not change among patients with HCC over underlying cirrhosis from patients with matching cirrhosis severity.

In 2020, another broader scale HCC study was reported by Miller-Atkins et al. [64]. They sampled healthy volunteers, cirrhotic without HCC, non-cirrhotic with HCC, pulmonary hypertension (PA), and colorectal cancer liver disease (CRLD) patients, and they examined specific VOCs reported in the literature to see whether they could achieve separation of their classes and which VOCs are more or less abundant in which group. They ran a SIFT-MS analysis, and they published that pairwise disease comparisons demonstrated that most of the VOCs were present in significantly different relative abundances. Each pairwise disease comparison had several compounds as significant; therefore, only the most significant metabolite associations for each disease are mentioned here. Comparing HCC against healthy volunteers revealed that (E)-2-nonene, ethane, and benzene increased in HCC patients, whereas hydrogen sulphide decreased. Comparing cirrhotic against healthy controls showed that trimethyl-amine and propanol

significantly increased in cirrhotic individuals. Furthermore, (E)-2-nonene, acetaldehyde, and ethane significantly increased in PA individuals than healthy volunteers, whereas hydrogen sulphide decreased in that pairwise disease comparison. When CRLD patients were compared against healthy controls, (E)-2-nonene, acetaldehyde, and triethyl-amine significantly increased in CRLD individuals, whereas hydrogen sulphide, acetone, and trimethyl-amine decreased. Lastly, Miller-Atkins et al. found that acetone, acetaldehyde, and dimethyl-sulphide were increased in cirrhotic without HCC patients than in non-cirrhotic with HCC patients, while ethanol was increased in the non-cirrhotic HCC patients than the cirrhotic without HCC patients. The authors' classification results can be seen in Table 4.

Arasaradnam et al. [65] investigated breath VOCs in non-cirrhotic HE patients compared to healthy individuals; however, they used the e-nose technology. They found that the resulting BP could distinguish the two groups with a sensitivity and specificity of 88% and 68%, respectively. The BP could also differentiate between overt and covert HE, however, with a moderate sensitivity and specificity of 79% and 50% (Table 4). E-nose technology does not quantify individual compounds that form the BP; nevertheless, this might not be a considerable bottleneck depending on the application.

VOCs in non-alcoholic fatty liver disease versus non-alcoholic steatohepatitis patients

Breath analysis has also been implemented to examine obesity-related liver diseases. Solga et al. [5] compared NAFLD patients, of which some had NASH, to explore the diagnostic capability of breath biomarkers against a standard blood serum test; they performed a GC analysis. Acetone concentrations in breath were found to be significantly increased in patients with severe steatosis (grade 2 or 3), steatohepatitis, and NASH compared to patients with mild forms of steatosis, or steatohepatitis, and NASH. Breath ethanol was also positively associated with hepatic steatosis severity, as it was higher in the breath of patients with severe steatosis (grade 2 and 3).

In 2013, Verdam et al. [66] investigated NASH. They sampled NASH and non-NASH patients, and they aimed to separate the classes—they performed a GC-MS analysis. They reported that NASH and non-NASH patients could be discriminated by using three compounds: N-tridecane, 3-methyl-butanotrile, and 1-proponol with a sensitivity and specificity of 90% and 69%, respectively [66] (Table 4). Their results, however, are very different from the research conducted in the [5] study. The lack of control and validation in the [5] study might have been a reason for this difference.

Alkhouri et al. [67] examined the usage of exhaled breath analysis as a diagnostic tool in children. They aimed to separate obese children with NAFLD from obese children without NAFLD by performing a SIFT-MS breath analysis. They discovered that various

VOCs (i.e. isoprene, acetone, trimethylamine, acetaldehyde, and pentane) could distinguish NAFLD children from those without NAFLD with an AUC of 0.71 (Table 4). The [67] study findings, though, might be questionable since NAFLD was not diagnosed by liver biopsy; it was diagnosed by assessing the presence of fatty infiltration.

VOCs in alcoholic and non-alcoholic fatty liver disease patients versus cirrhotic patients

Millonig et al. [36] demonstrated the usage of exhaled breath VOCs for differentiating among non-cirrhotic alcoholic fatty liver disease (AFLD), non-cirrhotic NAFLD, cirrhotic patients, and healthy cohorts. They aimed to separate these groups of patients by using ion-molecule reaction-MS (IMR-MS) analysis. Millonig et al. reported that 19 compounds showed significantly different exhalation patterns (not compound identification was achieved per class) among the different liver disease types. The most promising compound was acetaldehyde, which was significantly higher in NAFLD and AFLD when compared to healthy controls and cirrhotic patients, and ethanol, which was only increased in cirrhotic patients and not in patients with NAFLD, AFLD, or healthy controls (Table 4).

In 2020, Sinha et al. [68] were the latest to investigate the ability to diagnose NAFLD using exhaled breath. They found that styrene, acetone, isoprene, terpinene, dimethyl-sulphide, acetophenone, and limonene significantly differed among cirrhotic and non-cirrhotic NAFLD patients. More specifically, isoprene, acetophenone, and terpinene were significantly lower in non-cirrhotic NAFLD patients than healthy controls; terpinene had the highest predictive capability, achieving an AUC value of 0.84. Styrene, isoprene, acetophenonene, and terpinene were significantly lower in cirrhotic NAFLD patients than healthy controls, whereas dimethyl-sulphide and limonene were significantly higher in cirrhotic NAFLD patients than in healthy controls—limonene and dimethyl-sulphide combined yielded the highest predictive capability with an AUC value of 0.98. Furthermore, dimethyl-sulphide and limonene were significantly higher in cirrhotic NAFLD patients than non-cirrhotic NAFLD; combined, they achieved an AUC of 0.91 (Table 4).

Letteron et al. [69] conducted a large scale study in which they stratified various liver disease patients. They sampled non-alcoholic liver disease patients categorised into acute hepatitis, chronic hepatitis, viral cirrhosis patients, polyadenomatosis of the liver patients, non-alcoholic HCC, liver metastasis, sclerosing cholangitis, biliary cirrhosis, extrahepatic bile duct obstruction patients, alcohol abusers, as well as healthy individuals. They measured the exhaled ethane levels by using a GC-flame ionisation detector (FID). Their results showed that alcohol abusers had significantly higher ethane levels than other non-alcoholic groups.

VOCs in alcoholic hepatitis patients versus cirrhotic patients

Hanouneh et al. [21] published a study where they investigated alcoholic hepatitis (AH). More specifically, they gathered two groups that consisted of AH patients with liver cirrhosis, patients with acute decompensation (AD) with aetiologies other than alcohol and liver cirrhosis, and a healthy cohort. They aimed to find concentrations of VOCs that correlate with AH diagnosis and the severity of liver disease in AH patients—patient samples were analysed utilising SIFT-MS. Six compounds were identified to be significantly higher in the exhaled breath of liver disease patients compared to controls: acetaldehyde, 2-propanol, ethanol, acetone, pentane, and trimethyl-amine (TMA). Moreover, four compounds (i.e. acetaldehyde, acetone, TMA, and pentane) stood out in patients with cirrhotic AH compared to patients with AD. Finally, Hanouneh et al. also demonstrated that cirrhotic AH patients have a distinct breath VOC pattern characterised by high levels of acetone, pentane, and TMA when compared against patients with liver disease of aetiologies other than alcohol. Their model created using these three compounds gave an excellent diagnostic accuracy for AH with a 97% sensitivity and a 72% specificity (Table 4).

Origin of the VOCs reported in cirrhotic, non-cirrhotic, and pre-cirrhotic stage individuals

The key compounds and their metabolic pathways discussed in the aforementioned literature (sections 4.2.1 – 4.2.5) can be summarised as follows. Increased isoprene levels were found in AFLD and advanced fibrosis stage patients [36, 60, 67], and it is suggested that they are the result of impairment in the cholesterol biosynthesis pathway or that they might be the result of disturbed colon flora. However, other literature suggests that subjects should be at rest before testing because isoprene absence/deficiency maybe the result of exercise and that generally, it should not be attributed to pathophysiological effects onto mevalonate/cholesterol pathways [70, 71]. Increased acetone levels were found in stage 1 or 2 fibrosis patients, as well as NAFLD and AH patients [5, 21, 67]; acetone is believed to be associated with lipolysis and carbohydrate metabolism, where increased expression of the CYP450 enzyme would result in fatty acid beta-oxidation, which then would lead to excess of acetyl-CoA. Another possible explanation could be that reduced NADH levels (Nicotinamide Adenine Dinucleotide) in hepatocellular mitochondria could decrease d-3-hydroxybutyrate and dehydrogenase activity, which also would increase acetone levels. Alkanes such as pentane, heptane, 2-methyl-hexane, and ethane that were found in NAFLD, HE, AH, and alcohol abusers were linked to lipid peroxidation of polyunsaturated fatty acids due to oxidative stress [21, 61, 67, 69]; terpinene, found in NAFLD individuals, was also linked to oxidative stress [68]. Furthermore, isothiocyanato-cyclohexane was characterised as a common environmental pollutant and its increase in HE patients was attributed to impaired liver catabolism, whereas increased 1-methyl-4-(1-methylethenyl)-benzene levels again in HE patients may

have originated from an enhanced aromatase activity due to extensive alcohol abuse that could have been responsible for changes in metabolism. HE patients were also characterised by increased octanal, and a compound tentatively identified as 2, 6-dimethyl-7-octen-2-ol levels, which might have resulted from the P450 induction and catabolism of fatty acids [61]. Compounds such as limonene, dimethyl-sulphide, as well as ketones that were also found in the section 4.1 studies, were given the same possible origin explanations as those discussed in the section 4.1.1. Higher ethanol levels observed in cirrhotic patients are probably caused by increased shunting volumes through portocaval shunts in the liver, preventing the metabolism of endogenous ethanol [36], whereas diminished acetaldehyde levels that were observed in NAFLD, AFLD, and cirrhotic patients were explained by diminished ethanol oxidation [36]. Interestingly, acetaldehyde levels were increased in NAFLD children; however, they were also attributed to the fact that acetaldehyde is a product of liver ethanol metabolism [67]. Finally, TMA either derives from an impaired liver damaged capacity to transform TMA to TMAO (i.e. physiological oxidation of TMA), or it derives from the degradation of dietary phosphatidylcholine and dietary free choline by the intestinal microflora [21, 67]. Figure 2 visualizes all these suggested pathways.

TABLE 4: SUMMARY OF THE PAPERS THAT EXAMINED CIRRHOTIC, NON-CIRRHOTIC AND VARIOUS PRE-CIRRHOTIC STAGE OCCASION PATIENTS AGAINST EACH OTHER. THE ARROWS SHOW (IF APPLICABLE) WHETHER A VOC LEVEL INCREASED OR DECREASED IN THE FIRST GROUP COMPARED TO THE SECOND GROUP IN THE STUDY DESIGN.

Author/Year	Study design	Analytical method	VOCs identified as significant	Discriminatory performance
Alkhouri et al. 2015	20 advanced fibrosis vs 41 mild fibrosis	SIFT-MS	Acetone ↓ Benzene ↓ Carbon disulphide ↓ Isoprene ↓ Pentane ↓ Ethane ↓	0.85 AUC *(Isoprene model)*
Khalid et al. 2013	11 alcoholic cirrhotic with HE vs 23 alcoholic cirrhotic without HE	GC-MS	Methyl-vinyl ketone ↓ Isothiocyanato-cyclohexane ↑	90% sensitivity 87% specificity
	34 alcoholic cirrhotic vs 13 non-alcoholic cirrhotic		Undecane ↑ Unknown ↓	78.3% sensitivity 69.2% specificity
	34 alcoholic cirrhotic vs 7 harmful drinkers		1-methyl-4-(1-methyl-ethenyl)-benzene ↑ Unknown ↓ Unknown ↓	88% sensitivity 85% specificity
	7 harmful drinkers vs 15 healthy		Octanal 2,6-dimethyl-7-octen-2-ol Unknown	71% sensitivity 93% specificity
	13 non-alcoholic cirrhotic vs 15 healthy		Methyl-vinyl ketone 1-methyl-2-(1-methyl-ethyl)-benzene (o-cymene) Unknown	92% sensitivity 100% specificity
	34 alcoholic cirrhotic vs 15 healthy		Heptane 1-methyl-2-(1-methyl-ethyl)-benzene Phellandrene 2-methyl-hexane	97% sensitivity 93% specificity
O'Hara et al. 2016	11 cirrhotic HE vs 11 cirrhotic without HE vs 7 history of HE vs 30 healthy	PTR-MS	Limonene ↑	Not reported
	10 without HCC vs 21 HCC vs 30 healthy		Limonene ↑	Not reported
Qin et al. 2010	30 HCC vs 36 healthy	GC-MS-SPME	3-hydroxy-2-butanone ↑ Styrene ↑ Decane ↑	83.3% sensitivity 91.7% specificity
	30 HCC vs 27 cirrhotic without HCC		3-hydroxy-2-butanone ↑ Styrene ↑	70% sensitivity 70.4% specificity
Ferrandino et al. 2020	32 cirrhotic without HCC vs 12 cirrhotic with HCC vs 40 healthy controls	GC-MS	Limonene ↑	73% sensitivity 77% specificity

Author/Year	Study design	Analytical method	VOCs identified as significant	Discriminatory performance
Miller-Atkins et al. 2020	112 non-cirrhotic HCC vs 54 healthy	SIFT-MS	(E)-2-nonene ↑ Ethane ↑ Benzene ↑ Hydrogen sulphide ↓	*Healthy vs all the rest* 76% sensitivity 97% specificity
†only the three most significant metabolite associations for each disease comparison are shown in the column of significant compounds	30 cirrhotic without HCC vs 54 healthy		Trimethyl-amine ↓ Propanol ↓	*Cirrhotic vs all the rest* 40% sensitivity 96% specificity
	49 PH vs 54 healthy		(E)-2-nonene ↑ Acetaldehyde ↑ Ethane ↑ Hydrogen sulphide ↓	*HCC vs all the rest* 73% sensitivity 71% specificity
	51 CRLM vs 54 healthy		(E)-2-nonene ↑ Acetaldehyde ↑ Triethyl-amine ↑ Acetone ↓	*CRLM vs all the rest* 51% sensitivity 94% specificity
	112 non-cirrhotic HCC vs 30 cirrhotic		Acetone ↓ Acetaldehyde ↓ Dimethyl-sulphide ↓ Ethanol ↑	*PH vs all the rest* 58% sensitivity 93% specificity
Arasaradnam et al. 2016	22 non-cirrhotic HE vs 20 healthy	E-nose	Not available	88% sensitivity 68% specificity
	13 covert non-cirrhotic HE vs 9 overt non-cirrhotic HE		Not available	79% sensitivity 50% specificity
Solga et al. 2008	16 moderate to severe steatosis vs 11 less steatosis	GC	Ethanol ↑ Acetone ↑	Not reported
	24 NASH vs 24 without NASH		Acetone ↑	Not reported
Verdam et al. 2013	39 NASH vs 26 without NASH	GC-MS	n-tridecane ↑ 3-methyl-butanonitrile ↑ 1-propanol ↑	90% sensitivity 69% specificity
Alkhouri et al. 2013	37 obese NAFLD vs 23 obese without NAFLD	SIFT-MS	Isoprene ↑ Acetone ↑ Trimethylamine ↑ Acetaldehyde ↑ Pentane ↑	0.76 AUC

Author/Year	Study design	Analytical method	VOCs identified as significant	Discriminatory performance
Millonig et al. 2010	37 cirrhotic vs 35 healthy	IMR-MS	Ethanol ↑	0.88 AUC
	91 liver diseased vs healthy		Acetaldehyde ↑ Ethanol ↑ Isoprene ↑	0.94 AUC
	34 NAFLD vs healthy controls		Acetaldehyde ↑	0.96 AUC
	20 AFLD vs 35 healthy		Acetaldehyde ↑ Isoprene ↑	0.97 AUC
	20 AFLD vs 34 NAFLD		Isoprene ↑	0.95 AUC
Letteron et al. 1993	89 alcohol abusers vs 52 liver diseased vs 42 healthy	GC-FID	Ethane ↑	Not reported
Hanouneh et al. 2014	80 liver diseased vs 43 healthy	SIFT-MS	2-propanol ↑ Acetaldehyde ↑ Acetone ↑ Ethanol ↑ Pentane ↑ Trimethylamine ↑	Not reported
	40 cirrhotic AH vs 40 cirrhotic AD		Acetaldehyde ↑ Acetone ↑ Pentane ↑ Trimethylamine ↑	97% sensitivity 72% specificity *(Acetone-pentane-trimethylamine)*

Liver diseases examined by VOC measured in faeces, bile and urine

VOCs in faeces

Raman et al. [72] sampled obese NAFLD presence patients and healthy controls to analyse and compare VOCs patterns in the headspace of faecal matter by running a GC-MS analysis. They found a core group of ester VOCs that was more abundant in obese NAFLD patients than healthy controls (normal liver and lean). This suggests that obese NAFLD patients have altered microbiome composition. Using binary data, they found 12 compounds that were significantly less common and 18 compounds that were more common in the faecal headspace of NAFLD patients than in healthy controls. Ester compounds composed most of the identified VOCs (i.e. aliphatic esters of ethanoic, butanoic, propanoic, and pentanoic acids). Most of these compounds were short-chain aliphatic alcohols and carboxylic acids derivatives. The origin of volatile esters coming from the gut microbiota [72] is still elusive. However, it is believed that bacterial enzymes such as esterases could catalyse reactions by using organic acids and alcohols; thus, leading to the formation of ester VOCs such as

those found in their study [72]. Ethanol was seen as a ubiquitous compound since it was present in both NAFLD and healthy individuals; nevertheless, these findings do not allow conclusions to be drawn as they are only qualitative findings. Many confounding factors were present as the researchers did not account for different diets, environment, or smoking. The study population did not include non-NAFLD obese patients; therefore, it is unknown if VOC characteristics are due to NALFD or obesity. The VOCs detected in the [72] study in the faecal headspace (esters of ethanoic, butanoic, propanoic, and pentanoic acids) belonged to the same classes as the compounds found by papers analysing breath (2-butanone, 2-pentanone, ethane). This suggests that breath VOCs could be derivatives of VOCs created by gastrointestinal bacteria, as argued in [72].

VOCs in bile

In 2015, Navaneethan et al. [38] published a pilot study in patients with primary sclerosis cholangitis (PSC), a risk factor for cholangiocarcinoma (CCA). Bile samples from the endoscopic bile repository were selected for analysis, of which some were PSC only patients, and some were PSC with CCA patients. Their objective was to identify potential VOCs in the bile headspace to discriminate CCA progression in PSC patients. They ran a SIFT-MS analysis, and they reported the following significantly different compounds: ethanol, acetonitrile, acrylonitrile, 3-methyl-trexane, benzene, carbon disulphide, acetaldehyde, dimethyl-sulphide, and 2-propanol. Combining 3-methyl-hexane, acrylonitrile, and benzene, they built a predictive model to diagnose PSC patients with CCA with a sensitivity and specificity of 90.5% and 72.7%, respectively. Benzene, an environmental pollutant originating from tobacco smoke and vehicle exhaust [38], was found alongside acrylonitrile and acetonitrile to be significantly less abundant in patients with CCA than PSC only patients. Also, dimethyl-sulphide, carbon disulphide, and mercaptopurines, which are products of incomplete metabolism in the liver of sulphur-containing amino acids [38], are less prominent in PSC patients with CCA. However, it should be noted that all of the compounds found in the [72] study, except for acetonitrile and acrylonitrile, have also been associated with liver disease by multiple papers analysing breath VOCs [34, 47, 50, 56, 67]. The [38] study illustrates that bile VOC analysis has potential for clinical applications. However, bile collection requires invasive procedures, and thus, it may not be the best path towards alternative VOC diagnosis of liver disease.

VOCs in urine

Navaneethan et al. [73] published another pilot study conducted on urinary samples consisting of patients with CCA, patients with pancreatic cancer, and patients with benign biliary strictures (PSC, chronic pancreatitis, and papillary steatosis). They aimed to diagnose biliary strictures in urinary VOCs by running a SIFT-MS analysis.

They found that ethane levels were significantly higher in PSC strictures compared to CCA patients. They also found that 2-propanol and carbon disulphide levels were lower in malignant strictures, which is in line with their previous study in the bile [38]. They generated a model using ethane and octane, which predicted CCA and malignancy with sensitivity and specificity of 80% and 100%, respectively.

Arasaradnam et al. [74] published a proof-of-principle study also focused on urinal VOC analysis. The patients recruited were NASH cirrhotic (NASH-C), NASH non-cirrhotic, and NAFLD; healthy controls (normal liver) were recruited, too. Their objective was to determine whether different stages of NAFLD and NASH had specific urinary VOC patterns and to pursue this, they ran a field asymmetric ion mobility spectrometry (FAIMS) analysis. The [74] study revealed that a urinary VOC breath-print could discriminate between all liver disease patients and healthy controls with low sensitivity of 58% and high specificity of 93%, and an AUC of 0.73. Arasaradnam et al. argued that these results suggest that different liver disease conditions create other chemicals [74]. The analysis also showed that urinary VOCs could distinguish between NASH and NAFLD with a sensitivity and specificity of 73% and 79%, respectively. Their urinary VOC patterns also distinguished well NASH-C and NASH without cirrhosis [74]. Their study suggests that urinary VOCs could be a potential noninvasive diagnostic tool for diagnosing NAFLD and the different NASH stages.

Finally, Bannaga et al. [75] published another pilot urinal VOC analysis examining HCC. They sampled HCC and non-HCC patients, and they tried to find biomarkers to separate the two classes—the non-HCC cases consisted of healthy and various NAFLD stage individuals, including those with or without fibrosis. They ran a GC-IMS analysis to separate their classes and a GC-MS analysis to identify HCC-related biomarkers. More specifically, the GC-IMS data separated the HCC patients from the fibrotic patients with an AUC of 0.97 (sensitivity 43% and specificity 95%), the HCC patients from the non-fibrotic patients with an AUC of 0.62 (sensitivity 60% and specificity 74%), and the fibrotic from the non-fibrotic patients with an AUC of 0.63 (sensitivity 29% and specificity 90%). Five compounds were identified as significantly different between the HCC and non-HCC patients (i.e. 4-Methyl-2,4-bis(p-hydroxyphenyl)pent-1-ene (2TMS derivative), 2-butanone, 2-hexanone, 1-ethyl-2-methyl-benzene, and 3-butene-1,2-diol,1-(2-furanyl)-) from the GC-MS dataset. All compounds but 2-butanone were significantly lower in HCC patients. Bicyclo[4.1.0]heptane, 3,7,7-trimethyl-, [1S-(1a,3ß,6a)]- and sulpiride were also significantly lower in HCC patients than in fibrotic patients. Bannaga et al. neither verified nor quantified their compounds; however, they gave plausible explanations as to why they may have found these compounds based on existing literature. For instance, they stated that 2-butanone has been reported in breath-related VOCs in liver diseases (this is in agreement with [22, 34, 49, 50]), 1-ethyl-2-methyl-benzene has been identified as a blood biomarker of HCC, whereas 3-butene-1,2-diol,1-(2-furanyl)- has been associated with lung cancer [75].

Summary

Figure 2 summarises the VOCs reported in the reviewed studies related to chronic liver diseases and their proposed metabolic pathways.

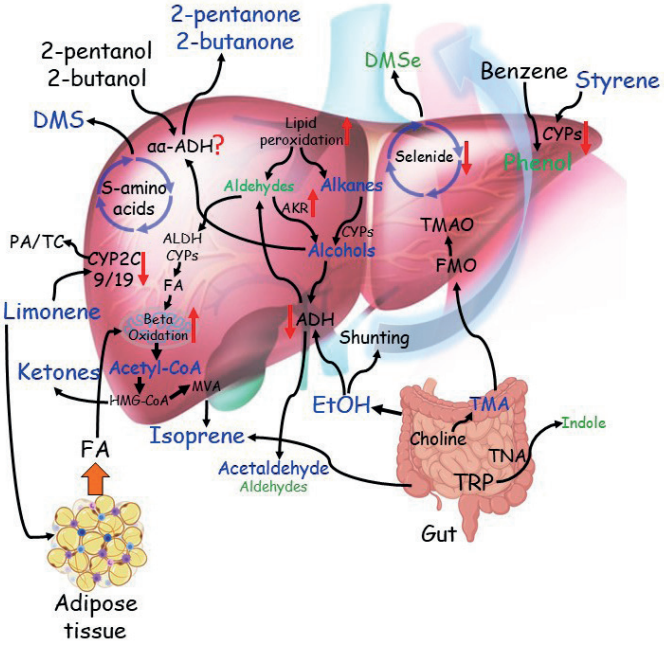


FIGURE 2: THE COMPLEX NETWORK OF ESTABLISHED AND PROPOSED METABOLIC PATHWAYS FROM WHICH VOCS STEM AND THEIR ALTERATIONS IN CHRONIC LIVER DISEASES. COMPOUNDS FOUND ELEVATED IN THE BREATH OF PATIENTS WITH CIRRHOSIS ARE INDICATED IN BLUE, THOSE DOWNREGULATED IN GREEN. RED ARROWS INDICATE CHANGES IN THE METABOLIC PATHWAYS. FROM THE BOTTOM LEFT: INSULIN RESISTANCE INCREASES FATTY ACID (FA) SHUTTLING FROM THE ADIPOSE TISSUE TO THE LIVER. THE RESULTING EXCESS OF ACETYL-COA IS METABOLISED IN THE MEVALONATE PATHWAY (MVA) TO KETONES AND ISOPRENE, THE LATTER ALSO GENERATES FROM GUT MICROBIOTA. DIETARY LIMONENE IS CONVERTED TO PERILLYL ALCOHOL (PA) AND TRANS-CARVEOL (TC) MAINLY BY CYP2C9 AND CYP2C19. PA AND TC ARE MORE SOLUBLE IN THE AQUEOUS ENVIRONMENT AND CAN BE EXCRETED IN THE URINES. IN THE CIRRHOTIC LIVER, REDUCED ACTIVITY OF CYP ENZYMES LEADS TO THE ACCUMULATION OF LIMONENE IN THE ADIPOSE TISSUE AND INCREASES ITS PERMANENCE IN THE BODY, RESULTING IN ELEVATED LEVELS IN THE BREATH. INCOMPLETE METABOLISM OF SULPHUR-CONTAINING AMINO ACIDS IN THE TRANSAMINATION PATHWAY, COUPLED WITH CYTOCHROME C OXIDASE DEFICIENCY IN THE CIRRHOTIC LIVER, LEAD TO ELEVATED LEVELS OF DIMETHYL-SULPHIDE (DMS) IN THE BREATH OF PATIENTS WITH CIRRHOSIS. DIETARY 2-BUTANOL, A FLAVOURING AGENT, AND A COMPOUND CONTAINED IN FRUIT IS CONVERTED TO 2-BUTANONE BY AA-ADH. A SIMILAR PATHWAY MAY ALSO INVOLVE 2-PENTANOL, A SIMILAR COMPOUND. BOTH 2-BUTANONE AND 2-PENTANONE HAVE BEEN FOUND ELEVATED IN THE BREATH OF PATIENTS WITH CIRRHOSIS. LIPID PEROXIDATION, A PROCESS TRIGGERED BY INCREASED INFLAMMATION OF THE CIRRHOTIC LIVER, HAS BEEN PROPOSED TO GENERATE ALKANES, SUCH AS OCTANE, PENTANE AND ETHANE, AND MEDIUM, LONG-CHAIN ALDEHYDES. THESE ALKANES HAVE BEEN FOUND ELEVATED, WHILE

DETECTED ALDEHYDES ARE REDUCED. BOTH CLASSES OF COMPOUNDS CAN BE CONVERTED TO CORRESPONDING ALCOHOLS BY RESPECTIVELY CYPS OR ALDO-KETO REDUCTASES (AKR). MEDIUM-CHAIN PRIMARY ALCOHOLS CAN BE FURTHER METABOLISED BY ALCOHOL DEHYDROGENASES (ADH) BACK TO ALDEHYDES, WHICH CAN BE CONVERTED TO CORRESPONDING FATTY ACIDS AND FEED BETA-OXIDATION. SECONDARY ALCOHOLS SUCH AS 2-BUTANOL AND 2-BUTANONE MAY ALSO BE GENERATED AND CONTRIBUTE TO INCREASING THE CORRESPONDING KETONES. ETHANOL (ETOH), WHICH ORIGINATES FROM THE DIET, SUGAR FERMENTATION FROM GUT MICROBIOTA, AND OXIDATION OF ETHANE, INCREASES IN THE BREATH OF PATIENTS WITH CIRRHOSIS BECAUSE OF SHUNTING AND DOWNREGULATION OF THE MAIN METABOLISING PATHWAY. HOWEVER, ACETALDEHYDE, THE MAIN BIO-PRODUCT OF ETOH METABOLISM, HAS ALSO BEEN ELEVATED DUE TO DOWNREGULATION OF THE DOWNSTREAM ENZYME ALDEHYDE DEHYDROGENASE (ALDH). DIMETHYL SELENIDE (DMSE) IS ONE OF THE EXCRETION PRODUCTS OF SELENIDE METABOLISM. SELENIDE BLOOD LEVELS WERE REDUCED IN PATIENTS WITH CIRRHOSIS, TO AN EXTENT RELATED TO DISEASE SEVERITY. THEREFORE, REDUCED DMSE IN BREATH MAY RESULT FROM A LACK OF SUBSTRATE AND IMPAIRED SELENIDE METABOLIC PATHWAY. BENZENE IS A POLLUTANT GENERATED MAINLY BY PETROL PRODUCTS AND READILY ADSORBED BY THE BODY BY INHALATION. BENZENE IS OXIDISED TO PHENOL BY THE CYP SYSTEM. REDUCED CYP ACTIVITY IN CIRRHOSIS MAY EXPLAIN REDUCED BREATH LEVELS OF PHENOL. EXPOSURE TO STYRENE TAKES PLACE MAINLY BY ADSORPTION OF VAPOURS THROUGH THE LUNGS. ITS REDUCED OXIDATION BY THE CYP SYSTEM EXPLAINS ITS INCREASE IN THE BREATH OF PATIENTS WITH CIRRHOSIS. TRIMETHYLAMINE (TMA) IS DERIVED FROM THE DIET BY MICROBIAL DEGRADATION OF PRECURSORS SUCH AS CHOLINE. TMA IS READILY ABSORBED AND METABOLISED BY FLAVIN-CONTAINING MONOOXYGENASES (FMO) IN TRIMETHYLAMINE N-OXIDE (TMAO) FOR URINE EXCRETION. REDUCED FMO ACTIVITY IN CIRRHOSIS MAY LEAD TO INCREASED TMA IN THE BREATH. INDOLE IS A CATABOLIC PRODUCT OF TRYPTOPHANE (TRP) METABOLISM BY TRYPTOPHANASE (TNA) ACTIVITY OF GUT MICROBIOTA, WHICH ALTERATIONS IN CIRRHOSIS MAY LEAD TO REDUCED INDOLE EXHALATION IN THE BREATH.

VOC analysis might greatly benefit liver disease diagnosis and prognosis; however, it is apparent from the literature findings that implementation of the VOC analysis in clinical liver practices is not ready yet for routine applications since much more research is needed. All conducted studies are either proof-of-concept studies or of a small sample size. Furthermore, many of the presented here studies, did not perform any internal or external validation of their findings. The correction of possible confounding factors was also not considered, and this might have affected their results. Nevertheless, some key concept can be kept from the present review that may point towards the eventual implementation of the VOC analysis in clinical liver practices. Several VOCs have been found in several studies, and as indicated in Figure 2, they have a solid biological explanation. All the compounds reported here are endogenous compounds except for limonene, which is an exogenous compound. This is probably the most striking observation of the present review because it illustrates the possibilities of a different study approach: exogenous VOC exposure. More specifically, one could expose a cohort at a particular limonene concentration with ingestion, sample their breath or maybe urine after exposure, and measure the difference between the inhaled and exhaled limonene concentration to determine liver function. The same principle could be applied to any other exogenous VOC metabolised by the liver. An exogenous VOC analysis enables for a tailored, controlled exposure to a compound of interest, thus providing a better chance in identifying disease-specific markers. Moreover, an exogenous VOC 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. It should be noted, however, that there are weaknesses of such an approach too. An exposure to a specific VOC may require patient preparation, but most importantly, it might be source of a potential allergy. Nonetheless, this approach could potentially help with liver disease diagnosis and prognosis since the exhaled concentration could indicate the level of liver impairment. The authors believe that this could push VOC analysis a step forward towards its clinical implementation in the liver research domain and other clinical settings.

Acknowledgments

Declaration of competing interest: The authors declare that they have no competing financial interests or personal relationships that could appear to influence the work presented in this paper.

Declaration of funding interest: The present study was supported by the VENI grant, Netherlands organization for scientific research (NWO) no. 016 VENI 178.064.

Authorship

Guarantor of the article: Dr Agnieszka Smolinska.

Specific author contributions: GS conducted the literature search; GS and MS performed the quality assessment of the eligible publications; GS wrote the manuscript with assistance from MS and KvM; GF provided the figure 2; GF, AS, FJvS, and CP revised the manuscript for intellectual content. All authors approved the final version of the manuscript.

References

1. Chen, S., L. Zieve, and V. Mahadevan, Mercaptans and dimethyl sulfide in the breath of patients with cirrhosis of the liver: Effect of feeding methionine. Translational Research, 1970. 75(4): p. 628-635.
2. Chen, S., V. Mahadevan, and L. Zieve, Volatile fatty acids in the breath of patients with cirrhosis of the liver. The Journal of laboratory and clinical medicine, 1970. 75(4): p. 622-627.
3. Tangerman, A., M. Meuwese-Arends, and J.M. Jansen, Cause and composition of foetor hepaticus. The Lancet, 1994. 343(8895): p. 483.
4. Shimamoto, C., I. Hirata, and K. Katsu, Breath and blood ammonia in liver cirrhosis. Hepato-gastroenterology, 2000. 47(32): p. 443-445.
5. Solga, S., et al., Breath biomarkers and non-alcoholic fatty liver disease: preliminary observations. Biomarkers, 2006. 11(2): p. 174-183.
6. Bataller, R. and D.A. Brenner, Liver fibrosis. The Journal of clinical investigation, 2005. 115(2): p. 209-218.
7. Kisseleva, T. and D.A. Brenner, Mechanisms of fibrogenesis. Experimental biology and medicine, 2008. 233(2): p. 109-122.
8. Tsochatzis, E.A., J. Bosch, and A.K. Burroughs, Liver cirrhosis. The Lancet, 2014. 383(9930): p. 1749-1761.
9. Simonetti, R.G., et al., Hepatocellular carcinoma. Digestive diseases and sciences, 1991. 36(7): p. 962-972.
10. De Franchis, R., Expanding consensus in portal hypertension: Report of the Baveno VI Consensus Workshop: Stratifying risk and individualizing care for portal hypertension. Journal of hepatology, 2015. 63(3): p. 743-752.
11. Lichtinghagen, R., et al., The Enhanced Liver Fibrosis (ELF) score: normal values, influence factors and proposed cut-off values. Journal of hepatology, 2013. 59(2): p. 236-242.
12. Le Calvez, S., et al., The predictive value of Fibrotest vs. APRI for the diagnosis of fibrosis in chronic hepatitis C. Hepatology, 2004. 39(3): p. 862-863.
13. Boyle, M., et al., Performance of the PRO-C3 collagen neo-epitope biomarker in non-alcoholic fatty liver disease. JHEP Reports, 2019. 1(3): p. 188-198.
14. Wai, J.W., C. Fu, and V.W.-S. Wong, Confounding factors of non-invasive tests for nonalcoholic fatty liver disease. Journal of Gastroenterology, 2020.
15. Kamath, P., R. Wiesner, and M. Malinchoc, A model to predict survival in patients with end-stage liver disease. Hepatology (Baltim Md) 33: 464–470. 2001.
16. Tonon, M., et al., Natural history of acute kidney disease in patients with cirrhosis. Journal of Hepatology, 2021. 74(3): p. 578-583.
17. Holzhütter, H.-G., et al., A novel variant of the 13 C-methacetin liver function breath test that eliminates the confounding effect of individual differences in systemic CO 2 kinetics. Archives of toxicology, 2020. 94(2): p. 401-415.
18. Logan, R., Urea breath tests in the management of Helicobacter pylori infection. Gut, 1998. 43(Suppl 1): p. S47.
19. Verlinden, W., et al., Non-Alcoholic Steatohepatitis Decreases Microsomal Liver Function in the Absence of Fibrosis. Biomedicines, 2020. 8(12): p. 546.
20. Petta, S., et al., Aminopyrine breath test predicts liver-related events and death in HCV-related cirrhosis on SVR after DAA therapy. Liver International, 2020. 40(3): p. 530-538.
21. Hanouneh, I.A., et al., The breathprints in patients with liver disease identify novel breath biomarkers in alcoholic hepatitis. Clinical Gastroenterology and Hepatology, 2014. 12(3): p. 516-523.
22. Morisco, F., et al., Rapid “Breath-Print” of Liver Cirrhosis by Proton Transfer Reaction Time-of-Flight Mass Spectrometry. A Pilot Study. PLoS One, 2013. 8(4).
23. Pleil, J.D., M.A. Stiegel, and T.H. Risby, Clinical breath analysis: discriminating between human endogenous compounds and exogenous (environmental) chemical confounders. Journal of Breath Research, 2013. 7(1): p. 017107.
24. Amann, A., et al., The human volatilome: volatile organic compounds (VOCs) in exhaled breath, skin emanations, urine, feces and saliva. Journal of breath research, 2014. 8(3): p. 034001.
25. Khalid, T., P. Richardson, and C.S. Probert, The liver breath! Breath volatile organic compounds for the diagnosis of liver disease. Clinical Gastroenterology and Hepatology, 2014. 12(3): p. 524-526.

26. Kwak, J. and G. Preti, Volatile disease biomarkers in breath: a critique. Current pharmaceutical biotechnology, 2011. 12(7): p. 1067-1074.
27. Volatile Organic Compounds (VOC) as non-invasive biomarkers for a range of diseases. 2020 [cited 2020; Available from: https://www.owlstonemedical.com/science-technology/.
28. Calenic, B. and A. Amann, Detection of volatile malodorous compounds in breath: current analytical techniques and implications in human disease. Bioanalysis, 2014. 6(3): p. 357-376.
29. Miekisch, W., J.K. Schubert, and G.F. Noeldge-Schomburg, Diagnostic potential of breath analysis—focus on volatile organic compounds. Clinica chimica acta, 2004. 347(1-2): p. 25-39.
30. Janfaza, S., et al., Digging deeper into volatile organic compounds associated with cancer. Biology Methods and Protocols, 2019. 4(1): p. bpz014.
31. Sies, H., What is oxidative stress?, in Oxidative stress and vascular disease. 2000, Springer. p. 1-8.
32. Kwon, J.-W., et al., Exposure to volatile organic compounds and airway inflammation. Environmental Health, 2018. 17(1): p. 65.
33. Hakim, M., et al., Volatile organic compounds of lung cancer and possible biochemical pathways. Chemical reviews, 2012. 112(11): p. 5949-5966.
34. Van den Velde, S., et al., GC–MS analysis of breath odor compounds in liver patients. Journal of Chromatography B, 2008. 875(2): p. 344-348.
35. Lembo, V., et al., Online analysis of breath by proton transfer reaction time of flight mass spectrometry in cirrhotic patients. 2013.
36. Millonig, G., et al., Non-invasive diagnosis of liver diseases by breath analysis using an optimized ion–molecule reaction-mass spectrometry approach: a pilot study. Biomarkers, 2010. 15(4): p. 297-306.
37. Garner, C.E., et al., A pilot study of faecal volatile organic compounds in faeces from cholera patients in Bangladesh to determine their utility in disease diagnosis. Transactions of the Royal Society of Tropical Medicine and Hygiene, 2009. 103(11): p. 1171-1173.
38. Navaneethan, U., et al., Volatile organic compounds in bile for early diagnosis of cholangiocarcinoma in patients with primary sclerosing cholangitis: a pilot study. Gastrointestinal endoscopy, 2015. 81(4): p. 943-949. e1.
39. Phillips, M., Breath tests in medicine. Scientific American, 1992. 267(1): p. 74-79.
40. Risby, T.H. and S. Solga, Current status of clinical breath analysis. Applied Physics B, 2006. 85(2-3): p. 421-426.
41. Pereira, J., et al., Breath analysis as a potential and non-invasive frontier in disease diagnosis: an overview. Metabolites, 2015. 5(1): p. 3-55.
42. Baethge, C., S. Goldbeck-Wood, and S. Mertens, SANRA—a scale for the quality assessment of narrative review articles. Research integrity and peer review, 2019. 4(1): p. 5.
43. Pauling, L., et al., Quantitative analysis of urine vapor and breath by gas-liquid partition chromatography. Proceedings of the National Academy of Sciences, 1971. 68(10): p. 2374-2376.
44. Schubert, J.K., et al., Breath analysis in critically ill patients: potential and limitations. Expert review of molecular diagnostics, 2004. 4(5): p. 619-629.
45. O'Hara, M., et al., Limonene in exhaled breath is elevated in hepatic encephalopathy. Journal of breath research, 2016. 10(4): p. 046010.
46. Hiroshi, K., et al., Evaluation of volatile sulfur compounds in the expired alveolar gas in patients with liver cirrhosis. Clinica Chimica Acta, 1978. 85(3): p. 279-284.
47. Tangerman, A., M.T. Meuwese-Arends, and J.H. van Tongeren, A new sensitive assay for measuring volatile sulphur compounds in human breath by Tenax trapping and gas chromatography and its application in liver cirrhosis. Clinica Chimica Acta, 1983. 130(1): p. 103-110.
48. Friedman, M.I., et al., Limonene in expired lung air of patients with liver disease. Digestive diseases and sciences, 1994. 39(8): p. 1672-1676.
49. Dadamio, J., et al., Breath biomarkers of liver cirrhosis. Journal of Chromatography B, 2012. 905: p. 17-22.
50. Del Río, R.F., et al., Volatile biomarkers in breath associated with liver cirrhosis—comparisons of pre-and post-liver transplant breath samples. EBioMedicine, 2015. 2(9): p. 1243-1250.
51. Pijls, K.E., et al., A profile of volatile organic compounds in exhaled air as a potential non-invasive biomarker for liver cirrhosis. Scientific reports, 2016. 6: p. 19903.

52. De Vincentis, A., et al., Breath-print analysis by e-nose for classifying and monitoring chronic liver disease: a proof-of-concept study. Scientific reports, 2016. 6: p. 25337.
53. De Vincentis, A., et al., Breath-print analysis by e-nose may refine risk stratification for adverse outcomes in cirrhotic patients. Liver International, 2017. 37(2): p. 242-250.
54. D'Amico, A., et al., An investigation on electronic nose diagnosis of lung cancer. Lung cancer, 2010. 68(2): p. 170-176.
55. Dragonieri, S., et al., An electronic nose in the discrimination of patients with non-small cell lung cancer and COPD. Lung cancer, 2009. 64(2): p. 166-170.
56. Eng, K., et al., Analysis of breath volatile organic compounds in children with chronic liver disease compared to healthy controls. Journal of breath research, 2015. 9(2): p. 026002.
57. LeBel, M., et al., Benzyl alcohol metabolism and elimination in neonates. Developmental pharmacology and therapeutics, 1988. 11: p. 347-356.
58. Miyazawa, M., M. Shindo, and T. Shimada, Metabolism of (+)-and (−)-limonenes to respective carveols and perillyl alcohols by CYP2C9 and CYP2C19 in human liver microsomes. Drug metabolism and disposition, 2002. 30(5): p. 602-607.
59. Burk, R.F., et al., Plasma selenium in patients with cirrhosis. Hepatology, 1998. 27(3): p. 794-798.
60. Alkhouri, N., et al., Isoprene in the exhaled breath is a novel biomarker for advanced fibrosis in patients with chronic liver disease: a pilot study. Clinical and translational gastroenterology, 2015. 6(9): p. e112.
61. Khalid, T.Y., et al., Breath volatile analysis from patients diagnosed with harmful drinking, cirrhosis and hepatic encephalopathy: a pilot study. Metabolomics, 2013. 9(5): p. 938-948.
62. Qin, T., et al., The screening of volatile markers for hepatocellular carcinoma. Cancer Epidemiology and Prevention Biomarkers, 2010. 19(9): p. 2247-2253.
63. Ferrandino, G., et al., Breath Biopsy Assessment of Liver Disease Using an Exogenous Volatile Organic Compound—Toward Improved Detection of Liver Impairment. Clinical and translational gastroenterology, 2020. 11(9).
64. Miller-Atkins, G., et al., Breath Metabolomics Provides an Accurate and Noninvasive Approach for Screening Cirrhosis, Primary, and Secondary Liver Tumors. Hepatology communications, 2020. 4(7): p. 1041-1055.
65. Arasaradnam, R., et al., Breathomics—exhaled volatile organic compound analysis to detect hepatic encephalopathy: a pilot study. Journal of breath research, 2016. 10(1): p. 016012.
66. Verdam, F.J., et al., Non-alcoholic steatohepatitis: a non-invasive diagnosis by analysis of exhaled breath. Journal of hepatology, 2013. 58(3): p. 543-548.
67. Alkhouri, N., et al., Analysis of breath volatile organic compounds as a noninvasive tool to diagnose nonalcoholic fatty liver disease in children. European journal of gastroenterology & hepatology, 2014. 26(1): p. 82-87.
68. Sinha, R., et al., Volatomic analysis identifies compounds that can stratify non-alcoholic fatty liver disease. JHEP Reports, 2020. 2(5): p. 100137.
69. Letteron, P., et al., Increased ethane exhalation, an in vivo index of lipid peroxidation, in alcohol-abusers. Gut, 1993. 34(3): p. 409-414.
70. Sukul, P., et al., Deficiency and absence of endogenous isoprene in adults, disqualified its putative origin. Heliyon, 2021. 7(1): p. e05922.
71. King, J., et al., Isoprene and acetone concentration profiles during exercise on an ergometer. Journal of breath research, 2009. 3(2): p. 027006.
72. Raman, M., et al., Fecal microbiome and volatile organic compound metabolome in obese humans with nonalcoholic fatty liver disease. Clinical Gastroenterology and Hepatology, 2013. 11(7): p. 868-875. e3.
73. Navaneethan, U., et al., Volatile organic compounds in urine for noninvasive diagnosis of malignant biliary strictures: a pilot study. Digestive diseases and sciences, 2015. 60(7): p. 2150-2157.
74. Arasaradnam, R.P., et al., Non-invasive distinction of non-alcoholic fatty liver disease using urinary volatile organic compound analysis: early results. Journal of Gastrointestinal & Liver Diseases, 2015. 24(2).
75. Bannaga, A.S., et al., Exploratory Study Using Urinary Volatile Organic Compounds for the Detection of Hepatocellular Carcinoma. Molecules, 2021. 26(9): p. 2447.

Supplementary materials

TABLE S1: ADDED QUESTIONS TO FURTHER AND MORE STRICTLY EVALUATE THE PAPERS INCLUDED IN THE PRESENT REVIEW.

Added questions
1. Did the authors correct or control for confounding factors such as diet, smoking, alcohol, medication, and place/time of sample collection?
2. Was the sample size sufficient?
3. Did the authors perform any statistical analysis? If so, did they correct for multiple testing? Also, did the authors validate their results? If so, did they perform internal (within the same cohort) or even external (new independent cohort) validation or both?
4. Where the study classes balanced?
5. Did the authors give the biological or non-biological origin of the compounds they found significant in their study?

TABLE S2: ADDED QUESTION COMMENTS AS TO WHY EACH PUBLICATION WAS GIVEN ITS RESPECTIVE QUALITY VALUE. IF NO EXPLANATION IS GIVEN, IT MEANS THAT ALL ISSUES RAISED IN THE ADDED QUESTIONS WERE ADDRESSED BY THAT PARTICULAR STUDY.

Publication	Quality	Explanation
Friedman et al. 1994 [1]	Medium	They did not mitigate for confounding factors; small sample size; they did not establish the origin of the found compounds; no statistical modelling was performed
Hiroshi et al. 1978 [2]	Low	They did not mitigate for confounding factors; small sample size; they did not establish the origin of the found compounds; no statistical modelling was performed
Letteron et al. 1993 [3]	Medium	They did not validate their results in terms of statistical modelling; imbalanced study classes; they did not establish the origin of the found compounds
Van den Velde et al. 2008 [4]	High	-
Dadamio et al. 2012 [5]	High	-
Pijls et al. 2016 [6]	High	-
Morisco et al. 2013 [7]	High	They did not validate their results in terms of statistical modelling
Del Rio et al. 2015 [8]	High	They did not validate their results in terms of statistical modelling
Eng et al. 2015 [9]	High	They did not validate their results in terms of statistical modelling
Alkhouri et al. 2015 [10]	High	They did not mitigate for confounding factors; imbalanced study classes
De Vincentis et al. 2016 [11]	Medium	They did not validate their results in terms of statistical modelling; no compounds were identified (due to the e-nose technology itself)
Khalid et al. 2013 [12]	Medium	They did not mitigate for confounding factors; imbalanced study classes

Publication	Quality	Explanation
O'Hara et al. 2016 [13]	High	They did not validate their results in terms of statistical modelling
Arasaradnam et al. 2015 [14]	Medium	Small sample size; imbalanced study classes; they did not establish the origin of the found compounds
Solga et al. 2006 [15]	Medium	They did not mitigate for confounding factors; they did not validate their results in terms of statistical modelling
Verdam et al. 2013 [16]	Medium	They did not mitigate for confounding factors; they did not validate their results in terms of statistical modelling
Alkhouri et al. 2013 [17]	Medium	They did not mitigate for confounding factors
Millonig et al. 2010 [18]	High	-
Hanouneh et al. 2014 [19]	High	-
Qin et al. 2010 [20]	Medium	They did not validate their results in terms of statistical modelling; imbalanced study classes
Sinha et al. 2019 [21]	Medium	Small sample size
Ferrandino et al. 2020 [22]	Medium	They did not mitigate for confounding factors
Miller-Atkins et al. [23]	High	-
Raman et al. 2013 [24]	Medium	They did not mitigate for confounding factors; they did not validate their results in terms of statistical modelling
Navaneethan et al. 2015 [25]	Medium	Small sample size; they did not validate their results in terms of statistical modelling
Navaneethan et al. 2015 [26]	Medium	They did not validate their results in terms of statistical modelling
Arasaradnam et al. 2012 [27]	Medium	They did not mitigate for confounding factors; small sample size; no compounds were identified (due to the e-nose technology itself)
Bannaga et al. 2021 [28]	Medium	They did not mitigate for confounding factors; small sample size

CHAPTER 3

Preprocessing and analysis of volatilome data

Georgios Stavropoulos, Dalia Salman, Yaser Alkhalifah, Frederik-Jan van Schooten, Agnieszka Smolinska

Breathborne Biomarkers and the Human Volatilome (Second Edition), 2020, Pages 633-647, doi: https://doi.org/10.1016/C2018-0-04980-4

Abstract

Biomarker discovery, i.e., finding disease or condition-specific markers, is a crucial aspect of biomedical research. Volatile organic compounds (VOCs) are excreted by various biofluids, cells and tissues, and bacteria, and these have been investigated extensively for their potential as markers of malfunctioning status in human. The number of VOCs excreted by those media and detected using sophisticated analytical instrumentations are numerically large and biologically complex. Therefore, data pre-processing and analysis are crucial for successful identification of valid VOC markers for their application in clinic practice. This chapter provides an overview of various pre-processing approaches suitable for volatilome data of diverse nature. The importance of normalization and scaling, often neglected in the field, is discussed. The most common and promising machine learning techniques are presented and discussed, including unsupervised and supervised approaches, followed by a rarely used strategy in volatilome field, data fusion. The chapter aims to equip the reader with a basic overview of suitable techniques for treating and successfully exploiting volatilome data.

Keywords: multivariate analysis, volatile organic compounds (VOCs), machine learning, supervised, unsupervised, data fusion

Overview

Volatilome research is a strongly emerging field that represents a new frontier in metabolomics. Research explores qualitative and/or quantitative changes in volatile organic compounds (VOCs) present within various biofluids, such as breath, blood, urine, saliva, feces, or excreted by cells or bacteria, and attempts to link these changes to health status. Consequently, the main objective in volatilome research is to discover patterns of VOCs that relate to deviant metabolic processes (for instance, inflammation) occurring in the human body. The emergence and continued development of high-throughput, high-resolution analytical platforms for VOC analysis has resulted in numerically large and biologically complex datasets for which sophisticated tools are required for their comprehensive processing and successful exploitation.

Volatilome data typically contain a mixture of endogenous compounds related to physiological, biochemical and metabolic processes, and exogenous compounds derived from environmental exposure, bacteria and viruses, amongst others (see chapter 1). This leads to data with different sources of variance, i.e., variance of interest as well as biologically irrelevant information or noise. The challenge in volatilome analysis is to separate those two sources of variation, by focusing on the compounds related to the studied problem and neglecting the redundant and irrelevant information in volatilome datasets. An array of sophisticated machine learning methods can be utilized to find the relevant set of VOCs. VOCs in volatilome data frequently correlate with one another, often because various metabolic pathways are involved that connect the measured VOCs. Since machine learning techniques are based solely on finding sets of important compounds, rather than single compounds, their usage should be at the core of statistical analysis of volatilome data.

This chapter presents and discusses the most common and promising machine learning techniques applicable for volatilome datasets, with the aim of providing a broad picture of the different methods as well as an overview of their pros and cons. Many different data processing techniques exist, but a comprehensive overview of these is beyond the scope of this chapter. Moreover, there is no golden rule or clear instructions in machine learning as to which techniques should be applied to which datasets. As such, only the most important aspects of data pre-processing and analysis are described herein. The chapter commences with an illustration of different data preprocessing approaches that are suitable for a broad range of volatilome analytical instrumentation, followed by the relevant (occasionally abandoned) aspects of scaling, normalization, and transformation. Unsupervised (also known as explorative or descriptive) [29] and supervised approaches are presented, together with the importance of model validation [30]. The chapter concludes with an introduction and discussion on data fusion strategies, which are a relatively new but unmet area in volatilomics.

Data pre-processing

Data pre-processing is a crucial step in every data mining approach since it helps to eliminate possible instrumental artifacts that may occur during the analysis, and most importantly, it improves and simplifies the data analysis. Note that diverse pre-processing procedures exist as either commercially or freely available packages, yet they all share very similar structure. The nature of pre-processing steps depend on the source of data, i.e., the instrumentation used in sample analyses, although many computations are common to all datasets. Pre-processing of analytical data typically comprises noise and baseline removal, correcting peak shift due to column ageing (relevant to gas chromatography-mass spectrometry; GC-MS), temperature drift or biochemical interaction, and peak picking. The first step of noise removal can be done by wavelet transformations [31] followed by baseline correction via P-splines [32]. Noise removal and baseline correction are essential because, nowadays, analytical techniques such as GC-MS (see chapter 15) or high-resolution MS (see chapter 16) have become highly sensitive; consequently, they capture non-biological information that ideally should be removed from datasets. The correction of peaks shift can be achieved via various techniques, as is reviewed in depth in the literature [33]. Peak picking, or binning, is the next step in pre-processing, which is a means of preliminary dimensionality reduction (i.e., reduction of the number of data points within each sample) by combining data points that relate to the same compound. The simplest method of peak picking is to sum up or average consecutive data points (e.g., 5-10) over the whole range of points in the samples. More advanced approaches, such as peak picking via peak correlation [33], are implemented when coupled techniques are used (e.g. GC-MS). In such cases, local minima, maxima, and peak areas are calculated for every peak in each sample, and these are assigned as originating from the same compound based on their signal-to-noise ratio (S/N) and retention time. Ultimately, these peaks are represented as a single value.

Recently, Alkhalifah *at al.* has suggested a procedure, named VOCCluster [34], for peak picking solely for GC-MS data. This procedure groups the peaks, i.e., creates clusters, in GC-MS data arising from the same VOC using similarity measures based on cosine angular separation. The important aspect of the VOCCluster technique is that each VOC could change cluster membership as the algorithm progresses and could be re-clustered into a different cluster, depending on the cosine similarity measurement of each VOC to the other VOCs. This re-clustering enhances the accuracy of the clusters and does not depend on the order of samples. In addition, the VOCCluster approach calculates the similarity threshold (epsilon) that is used as an input parameter for the clustering process. This is helpful for untargeted volatilome studies as it allows such process to be data driven rather than operator dependent.

Another important aspect of the VOCCluster procedure is its ability to take into account both retention index (RI) and mass spectrometric variations. For RI variation, the technique divides the data into different segments and calculates the RI variation for each segment. This leads to a RI range for each VOC within a sample, which is used to cluster VOCs into similar groups within that RI range. In the case of mass spectra variations, the technique takes into account the chemical nature of the VOC, signal intensity, and the number and/or order of extracted ions. The mass spectra variations are especially relevant for co-eluting VOCs.

A widely applied software in metabolomics but not yet extensively used in volatilomics, is XCMS, which is a freely available software for data pre-processing as a means of untargeted metabolite profiling between two sample groups [35]. The software was originally developed to pre-process liquid chromatography-mass spectrometry (LC-MS) data, but the developers state that it is equally applicable to GC-MS data. XCMS is an acronym for various (X) forms of chromatography-mass spectrometry techniques. The software has gained much attention in metabolomic analyses due to its simplicity and flexibility, as it allows for peak detection, peak matching (i.e., peak picking), misalignment correction, and metabolite identification in a semi-automated way and with minimal user interference. Baseline correction, or background subtraction, is not performed in XCMS due to the danger of this potentially adding more noise to the data and since overall the background typically remains constant from one run to the next, thus its impact is assumed minimal. Peak detection is carried out by employing second-order derivatives and S/N. Importantly, XCMS deals with a common problem for most of the peak detection algorithms (i.e., to miss out peaks even though they are present) due to the inherent uncertainty close to the S/N cut-off value. In particular, XCMS fills in intensity values for every peak that is not detected by the peak detection algorithm by using information gathered from the raw data. Peak matching follows once all the peaks have been identified, and it is performed based on mass drifts (i.e., m/z values) rather than retention time. Peaks that are not matched in at least half of the samples are discarded, although different tie-breaking criteria can be chosen depending on the application. If it is known, for example, that the samples were obtained from two different groups (i.e., healthy and disease), then peaks that are not present in at least half of the samples of one of the groups are removed from that particular group. Finally, retention time alignment is performed. The algorithm does not use a reference signal to align the rest of the data, but rather works purely with peak data. In fact, ‘well-behaved’ peak regions are identified; these regions are areas in the chromatograms where very few samples have no peaks, and very few samples have more than one peak present. These ‘well-behaved’ regions are more likely to be adequately matched and are therefore used as temporary standards (i.e., alignment references). XCMS permits for the identification of important metabolites by employing a univariate t-test, and ranks the metabolites based on their p-values once all the pre-processing steps are finished. At the same time, it also allows for metabolite profiling based on exact masses, since it is linked to a metabolite database.

Regardless of the data pre-processing approach, the volatilome data can be described as a data matrix consisting of each measured observation in the row and detected VOCs in columns. In the volatilome, the majority of compounds (i.e., VOCs) are not present in all samples; therefore, only compounds that are present at a certain percentage (e.g., 10-30%) of the samples are retained for further analysis.

Data normalization, transformation, and scaling

Data normalization, transformation, and scaling are often considered as data pre-processing steps; nonetheless, they are treated here as an additional step due to their significant influence in the outcome of the analysis. The issue of lognormality in biological datasets is treated in the previous chapter (chapter 37). Normalization is typically performed to remove effect size, i.e., unwanted variations between measured samples, and generates volatilome data with samples presented in an adequate and consistent way. Note that incorrect normalization might jeopardize putative differences between investigated cases. Data normalization can be implemented in a variety of ways. It is often performed before or after peak picking, and it is usually done by calculating a normalization factor per sample. Probabilistic quotient normalization and total area normalization (TAN) have been used in volatilome data [36, 37]. TAN uses the assumption that the total profile that is measured is directly proportional to the total concentration of the sample and that total area is constant between samples. This is rarely the case, however, therefore TAN might lead to spurious correlations between VOCs [38].

Data transformation follows the exclusion of zeros, and it accounts for correcting for data heteroscedasticity and skewness; both logarithmic and power transformation are the most broadly used [39]. As far as data scaling is concerned, many techniques exist, each of which exploits different aspects of the data at hand [39]. Scaling accounts for giving all the variables the same importance in the model; autoscaling and pareto scaling are the most common approaches. Moreover, it is essential to note that most of the multivariate approaches require the data to be scaled before running the algorithm. A few examples of powerful multivariate techniques which require prior data scaling are principal component analysis (PCA) [40], the robust principal component analysis (R-PCA) [41], clustering, the partial least squares (PLS) analysis [42, 43], as well as the support vector machines analysis (SVM) [44]. Some multivariate approaches do not require prior scaling, including random forest (RF) [45], unsupervised random forest (URF) [46], and adaptive boosting analyses [47].

Machine learning approaches

Unsupervised approaches

The human volatilome is characterized as a rather complicated, high-dimensionality data matrix. Often, conventional univariate statistical approaches such as t-test do not suffice when it comes to biomarker identification; therefore, multivariate statistical approaches are preferred. Multivariate approaches can be divided into two categories: unsupervised and supervised approaches. Unsupervised approaches are used for exploratory purposes, whereas supervised approaches are used for classification and regression purposes. The most common unsupervised approach is principal component analysis (PCA) [40]. PCA captures most of the variation in the data by creating new variables, the *principal components* (PCs), which represent the original variables of the data linearly. This means that the PCs are linear combinations of the original variables, and by definition, they are orthogonal to each other. The first PC (PC 1) captures the largest possible variation in the data, and each subsequent PC (PC 2, PC 3, etc.) captures most of the remaining orthogonal variation. PCA returns two matrices, the score and the loading matrix. The scores are simply the coordinates of the samples in the PCA space, whereas loadings show the importance/contribution of every single original variable in making each PC. Consequently, both score and loading figures can be generated to identify clusters in the data and to explore which of the original variables play an essential role in getting these clusters, and to what extent. These two figures can be also combined to generate what is called a bi-plot (figure 38.1), which offers a more comprehensive picture – and thereby a better understanding – of the data. In a bi-plot, each sample is represented as a circle and each VOC as an arrow. In the PCA bi-plot it is possible to visually appraise the relationship between VOCs by looking at the angle between them. A small angel between VOCs indicates high positive correlation (e.g., isoprene and benzene in figure 38.1) whereas obtuse angles close to 180° (e.g., phenol and p-benzoquinone in figure 38.1) indicate an anti-correlation. The two compounds are uncorrelated if the angle is close to 90° (e.g., acetic acid and p-benzoquinone in figure 38.1). Additional useful information delivered from a bi-plot is the possibility of defining the quantitative change in VOCs due to different class membership. The relative VOC concentration is elevated if the sign of the PCs for samples and compound is the same (i.e., either both negative or positive). Obviously, if the sign of the PCs for samples and VOC is opposite (i.e., positive and negative) the concentration is reduced in these samples.

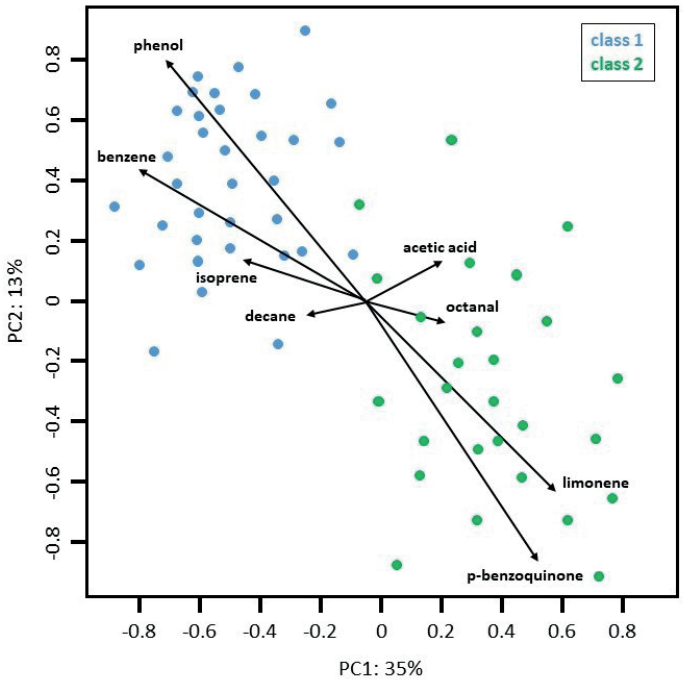


FIGURE 38.3 EXAMPLE OF A PCA BI-PLOT ILLUSTRATING TWO RANDOM CLASSES AND COMPOUNDS. THE DOTS ARE THE SCORES, WHICH REPRESENT THE SAMPLES, AND THE ARROWS ARE THE LOADINGS, WHICH REPRESENT THE COMPOUNDS PRESENT IN THE SAMPLES. THE COMPOUNDS THAT POINT TOWARDS A PARTICULAR GROUP ARE MORE IMPORTANT FOR THAT PARTICULAR GROUP. PC1 EXPLAINS 35% OF THE ORIGINAL INFORMATION, WHEREAS PC2 EXPLAINS 13% OF THE REMAINING ORIGINAL INFORMATION.

PCA is also useful in helping to detect outliers that are observable in the PCA space but are far away from the cloud of the sample observations. This type of outliers are called good leverage outliers. Frequently, however, the data also have what is called orthogonal outliers, which cannot be detected by PCA, and can therefore profoundly influence the analysis results and lead to false conclusions. These orthogonal outliers have orthogonal distances from the PCA space and it is mostly impossible to detect them by projecting them onto the PCA space. To overcome this issue, robust-PCA (R-PCA) was developed [41]. PCA maximizes variance by decomposing covariance in the data. R-PCA does the same, but replaces the covariance matrix with a robust covariance estimator. PCA creates one PC at the time to capture most of the variation in the data, whereas R-PCA creates consecutive possible 'interesting' PCs and then selects the one that best describes the data.

Unsupervised random forest (URF) [46] was introduced as a powerful unsupervised approach. URF is robust to any outliers as it examines variables, rather than the samples, as is done in PCA, R-PCA, and clustering. URF hypothesizes that if there is any hidden structure in the data, it should be possible to separate the data from a randomly generated version of themselves. Superiority of URF over other techniques is that it does not require any scaling, it provides variable importance, and most importantly, it returns a proximity matrix of the original data. Proximity matrices are excellent tools for visualization purposes and for direct comparison of the samples at hand.

Another unsupervised approach that is widely used is clustering [48], which consists of determining similarity measures such as correlation or distance of the whole volatilome dataset. Several different approaches exist, including hierarchical cluster analysis (HCA), k-means and c-means. HCA is based on iterative calculation of distance and combining the closes samples into one cluster. The procedure continues until all samples belong to the same cluster. The most important aspect of HCA is the selection of similarity measure (e.g., Euclidean, Manhattan, Mahalanobis, Minkowski distance or correlation) and the way the clusters are created, i.e., average linkage (average distance), single (minimum distance), complete linkage (maximum distance), and Ward's method [49]. The similarity/dissimilarity of samples is then represented as a dendrogram. An example of a dendogram is shown in Figure 2. As can be seen for a selected specific threshold on the distance measure (indicated as horizontal line on value 1000 on y-axis) two main groups are present. In addition, few samples are clearly included as separate clusters, suggesting that they are outliers (marked in the circles). HCA is the most suitable if a clear hierarchical structure is present in the data. An entirely different approach is taken by k-means [50], which selects a number of centroids (i.e., clusters), defined by a user, in such a way that the overall distance of all samples to the centroids is minimized. Each sample is assigned to the closest centroid. The centroids are updated each time a new sample is assigned. The assignment of the samples to the centroids stops when convergence (i.e., no more changes in cluster content) is reached. An adaptation of k-means is c-means, which does not give a definitive assignment of each sample to a cluster but provides a probability of belonging to each cluster.

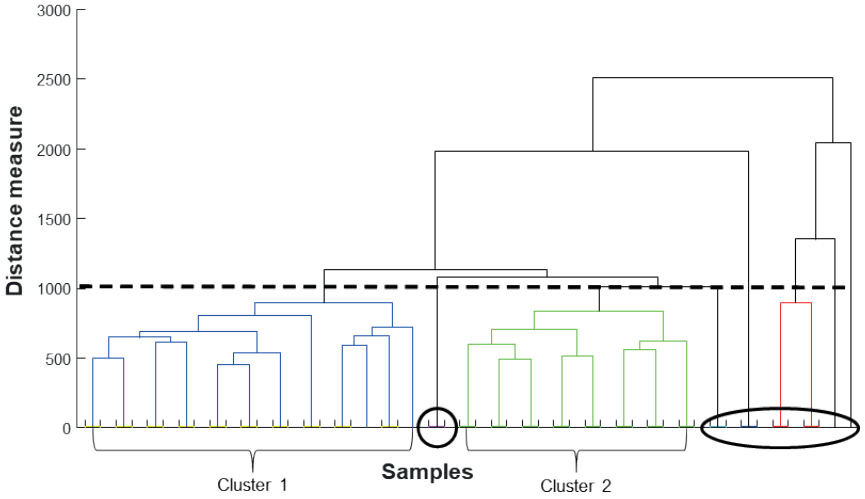


FIGURE 38.2. AN EXAMPLE OF DENDROGRAM ILLUSTRATING TWO MAIN CLUSTERS OF SAMPLES. SAMPLES INCLUDED IN THE CIRCLES ARE ALLOCATED AS OUTLIERS. A DRAWBACK OF CLUSTERING TECHNIQUES IS THAT THEY DO NOT DELIVER INFORMATION ABOUT COMPOUNDS THAT ARE RESPONSIBLE FOR THE RESULTING CLUSTERS. MOREOVER, THEY ARE VERY SENSITIVE TO OUTLIERS AND PROCESSING BECOME VERY TIME-CONSUMING FOR LARGE DATASETS.

Supervised approaches

The most common supervised approach is partial least squares (PLS) analysis [42, 43]. PLS was initially developed to deal with regression problems, but later it was extended to deal with classification problems, too. Similar to PCA, PLS creates new variables, the *latent variables* (LVs), that try to capture most of the information in the data with respect to a response/class vector Y in a linear way. Consequently, these LVs are linear combinations of the original variables, which have incorporated the response/class information of every sample, too. As with every supervised approach, PLS models must first be optimized (i.e., the optimal number of LVs to be used) then validated before any final predictions/classifications are made. An overview of validation methods that can be used is given in the next section. PLS can be applied not only on binary but also on multiple class problems, although the latter show inferior prediction accuracy when compared to results from 2-class or even multiple binary PLS models (i.e., one class against all the rest) [33]. PLS exhibits high accuracy results when applied to collinear data but has tendency for overfitting. Human volatilome data, however, often demonstrate rather nonlinear relationships among variables and therefore it may be preferable to implement supervised approaches that consider both linear and nonlinear relationships in the data; ensemble techniques are such approaches [51]. The most renowned and broadly used ensemble techniques are random forest [45], adaptive boosting (AdaBoost) [47] and gradient boosting, all of which are tree-based methods.

Random forest is a so-called *bagging* or *aggregative bootstrapping*, algorithm that builds fully-grown independent trees. Furthermore, each tree is built on a different subset of observations of the training dataset and a different subset of variables of these selected observations; therefore, different information is seen by different trees, thereby forcing the correlation among different outcomes to decrease. The observations that are not used to build a tree (i.e., out-of-bag observations) are used to assess this tree performance. The overall performance of the forest is assessed by the out-of-bag error (i.e., wrongly predicted/classified out-of-bag observations) of all the trees present in the forest, and in the end the forest is evaluated by using a validation set. Finally, it is worth mentioning that the more trees that are added to the forest, the merrier.

AdaBoost is representative of a boosting algorithm, which creates stumps (i.e., a tree with two leaves) instead of fully-grown trees. Most importantly, the stumps are built sequentially, so the mistakes a stump makes influence the way the next stump is built; thus, all stumps are dependent on each other. Each stump is built on a different subset of observations of the training dataset and a different subset of variables of these selected observations. Although the different subsets of observations are sampled without replacement, in RF they are sampled with replacement. This means that the same sample may be used in building more than one stump. The overall performance of the forest of stumps is evaluated with a validation set, and it should be mentioned that the more stumps that are added to the forest, the higher the chances to over-fit. A method between random RF and AdaBoost is gradient boosting, which makes use of decision trees as its weak classifiers that are constructed in a greedy manner [52]. In comparison to AdaBoost and RF that use stumps and fully-grown trees, respectively, gradient boosting builds trees of four to eight levels (i.e., splits).

As a final note on the supervised approaches, they can all provide importance/contribution of the variables in the models, allowing for potential biomarker discovery [33]. An overview of various aspects of the multivariate methods, unsupervised and supervised, described here is presented in table 38.1. The ensemble of techniques presented here are capable of finding linear and non-linear relationships between compounds in volatilome data.

TABLE 38.1. THE MAIN CHARACTERISTICS OF THE MOST COMMON MULTIVARIATE METHODS.

Method	linear; non-linear	Characteristics				
		sensitive to			compounds importance	
		outliers	scaling	normal distribution		
Unsupervised						
PCA	linear	Yes	Yes	No		Yes
R-PCA	linear	No	Yes	No		Yes
URF	Non-linear	No	No	No		Yes[#]
HCA	linear	No	No	No		No
k-means; c-means	linear	Yes	No	No		No
Supervised						Overfitting
PLS/PLS-DA	linear	Yes	Yes	No	Yes	Yes
AdaBoost	non-linear	No	No	No	Yes	Yes
RF	non-linear	No	No	No	Yes	No
Gradient boosting	non-linear	No	No	No	Yes	No
SVM	non-linear	Yes*	No	No	Yes[#]	No
K-PLS/K-PLS-DA	non-linear	Yes	Yes	No	Yes[#]	Yes
ANN/deep learning	non-linear	No	No	No	No	No

#POSSIBLE AFTER APPLYING PSEUDO-SAMPLES PRINCIPLES [53].
PCA: PRINCIPAL COMPONENT ANALYSIS
R-PCA: ROBUST PRINCIPAL COMPONENT ANALYSIS
URF: UNSUPERVISED RANDOM FOREST
HCA: HIERARCHICAL CLUSTERING
PLS/DA: PARTIAL LEAST SQUARE/DISCRIMINANT ANALYSIS
ADABOOST: ADAPTING BOOSTING
RF: RANDOM FOREST
SVM: SUPPORT VECTOR MACHINES

Another group of techniques capable of capturing non-linear information in volatilome data are kernel-based techniques. The most common methods are support vector machines (SVM) and kernel-PLS-and kernel-PLS-discriminant analysis (DA). The crucial step of kernel-based methods is transformation of the data via specific functions called kernel. This step allows mapping the non-linear problem in the original data into a higher-dimensional feature space, in a way that the problem becomes linear and thereby easily solvable. SVM was originally developed for two class classification problems. The performance of the technique depends highly on kernel function used to transform the data. Kernel-PLS-DA and SVM have comparable prediction ability [54] and both techniques have their disadvantage of losing compound information,

since after kernel transformation the obtained data have dimension of samples. Despite this, both techniques have been applied in various human volatilome data treatment endeavors [55, 56]. Note, that every supervised technique presented so far can be used for regression as well as classification problems.

The growing importance of artificial neural networks (ANN) and consequently deep learning in medical diagnostics has become evident from various applications and publications. Both techniques are very powerful to model complex, non-linear problems but require large sample size (in case of deep learning, tens of thousands) and the interpretation of the results is difficult. ANN has been applied to human volatilome data [57], but deep learning has not, to date.

Data fusion

Data fusion, or data concatenation, refers to the process of combining data coming from different data platforms (e.g., GC-MS, nuclear magnetic resonance, 16S ribosomal RNA sequencing, etc.). The principle behind data fusion is that different types of samples (e.g., breath, blood, feces) are complementary to each other when a particular cohort is examined. The same holds when different techniques measure the same samples (e.g., breath samples measured by GC-MS and multi-capillary-column-ion mobility spectrometry; MCC-IMS) because the different strengths and weaknesses of each technique with respect to their detection of compounds are exploited. Consequently, a better profile or increased prediction accuracy can be achieved for the cohort at hand by combining the data of these complementary samples and/or techniques. Data fusion is widely used in many research fields, and holds promise for human volatilome research, too.

Data fusion can be performed at three different levels: low-level, mid-level, and high-level fusion [58]. Low-level fusion is the simplest of all since it concatenates the data by placing them next to each other as they are from different platforms without any prior analysis. This means that the fused matrix to be used for further analysis will consist of as many rows as the number of samples measured, and as many columns as the number of all the compounds measured by all different data platforms. Low-level fusion is not usually employed because concatenating all these hundreds or even thousands of variables in a single matrix increases the dimensionality of the data too much, thus making it difficult to analyze them.

Mid-level fusion fuses either variables or features; therefore, it requires prior analysis on each data platform separately. On the one hand, important variables per platform can be found, for example, via random forest [45] or significance multivariate correlation [59] and then, all the important variables from the platforms are placed next to each other to create the fused matrix to be used for further analysis. On the other hand,

important features per platform can be found, for example, by implementing PCA [40] (and make use of the PCs) or PLS [42, 43] (and make use of the LVs). Then, all the features from all the platforms are placed side by side to create the fused matrix. Mid-level fusion is the level of fusion that is most widely applied.

High-level fusion requires prior separate analysis on each data platform, and it is quite different from both low-level and mid-level because it combines outcomes rather than actual data (i.e., variables or features) of the platforms. The most common approach for combining outcomes is majority voting [60]. For instance, if a sample of interest is classified as class 1 from the majority of the models (i.e., those built on the individual platforms), and as class 2 from the minority of the models, then the high-level fusion outcome will be class 1 for this sample of interest. High-level fusion can demonstrate excellent prediction results, which is to be expected because of the nature of this kind of fusion. Nonetheless, the possibility to discover potential biomarkers and to relate compound outputs from different platforms to each other is lost when high-level fusion is employed. This is result is to be expected because high-level fusion does not work with actual data but with outcomes. Recently, a new and more advanced way of data fusion that uses kernels was proposed [53]. The authors named their approach multiple kernel learning data fusion because each platform is mapped onto kernel space once variable selection is performed individually on every platform. Next, all different kernels are combined in a linear way by using a weighted sum to obtain the fused matrix for use in further analysis. Using kernels is synonymous with using samples; thus, theoretically, investigation of the original variables (e.g., how they behave in the samples or how important they are) might be challenging. In practice this problem can be overcome by using the pseudo-sample principle [53]. To conclude, data fusion can positively affect the outcome of a study, yet there is not a direct answer to identify which data fusion approach is the best, since this highly depends on the aim of each study.

Validation of supervised techniques

A crucial part of supervised methods is validation, which is a step that ensures the certainty of the findings. There are various ways of assessing the predictive ability of the constructed supervised model [61-63], as outlined below.

If a studied problem consists of a sufficient number of samples, the data are divided into a training set to train the supervised model, a validation set to optimize the supervised model, and a test set to assess the prediction power of the model. The most straightforward approach consists of randomly assigning samples into those three sets. Another approach allows selecting representative training and/or test sets [33]. The most commonly applied strategy is cross-validation (CV). This approach is based on retaining a predefined number of samples and building supervised model

on the remaining samples. The simplest CV procedure is leave-one-out (LOO), where one sample is excluded and used as a test set, and n-1 samples are used to create a training set (where 'n' is the total number of samples). In LOO CV, each sample becomes a test sample once, and therefore there are 'n' rounds of testing. LOO CV is suitable when the number of available samples is relatively low, e.g., 20-30 samples. If more samples are available, k-fold CV can be used, where k is defined as a number, e.g., 5 or 10, or percentage of total samples, e.g., 10%. In case of k-fold CV, k number of samples is removed from the data and used as test set. The important aspect of CV is its overestimation of the predictive power of the supervised model. Hence, for data with rather small number of samples it is recommended to apply double CV (called nested CV) or bootstrapping [61]. The final prediction ability of the supervised model can be assessed by a permutation test, which checks whether division into defined classes is significantly better than any random division [62]. The permutation test consist of random rearrangement of the class labels in the training set and the construction of a classification model using these randomly permuted classes. In the subsequent step, the prediction for real test set samples is obtained and the number of misclassifications is obtained. The entire procedure is repeated many times, for instance 1000 iterations. The assumption of the permutation test is that the test set should be wrongly predicted for randomly permuted classes in the training set.

A very important aspect of supervised techniques is the scaling of the data after division into the training, validation, and test sets. It is of great importance that samples used for validating the performance of the machine learning model should be always scaled (e.g., by autoscaling or pareto scaling) using parameters obtained from the training samples. Although there exist various approaches to test quality of the predictive model, the ultimate manner of assessing prediction ability of a statistical model is to use an independent test set consisting of completely new samples from an independently sampled population.

Summary

The main steps involved in data pre-processing and multivariate analysis of volatilome data is reviewed in this chapter. Data pre-processing is a crucial step when dealing with numerically complex volatilome data. This procedure can be carried out via various approaches, such as VOCCluster or XCMS (used for data pre-processing of GC-MS related data), and additionally involves the use of techniques for noise and baseline reduction that can be applied to different MS-based volatilome datasets. The importance of normalization and scaling of volatilome data has been emphasized in this chapter. The machine learning techniques covered in this chapter are limited to the most commonly applied approaches that have the highest potential for the future applications. Moreover, the techniques presented here are reasonably easy to apply and potentially insightful. Strategies for data fusion are currently underused but

should become more common in volatolomics in future. Because of the increased complexity of the studied problems, it is favourable to combine multiple sources of information to gain a better understanding of the underlining problem. Therefore, combining multiple types of measurements, e.g., VOCs in breath as well as feces, in a single statistical analysis, might not only increase the prediction accuracy but also interpretability and comprehensiveness of the results. Appropriate utilization of machine learning techniques in the field of human volatilome research offer the potential for more robust discoveries of relevant VOC biomarkers.

References

1. Krzanowski WJ. Principles of Multivariate Analysis (Revised edn). New York2000.
2. Vandeginste BGM, Massart DL, Buydens LMC, Jong SD, Lewi PJ, Smeyers-Verbeke J. Handbook of Chemometrics and Qualimerics: Part A. Amsterdam: Elsevier; 1998.
3. Walczak B. Wavelets in Chemistry. Data Handling in Science and Technology: Elsevier Science & Technology; 2000.
4. Eilers PH. A perfect smoother. Analytical chemistry. 2003;75(14):3631-6.
5. Smolinska A, Hauschild A-C, Fijten R, Dallinga J, Baumbach J, Van Schooten F. Current breathomics—a review on data pre-processing techniques and machine learning in metabolomics breath analysis. Journal of breath research. 2014;8(2):027105.
6. Alkhalifah Y, Phillips I, Soltoggio A, Darnley K, Nailon WH, McLaren D, et al. VOCCluster: Untargeted Metabolomics Feature Clustering Approach for Clinical Breath Gas Chromatography/Mass Spectrometry Data. Anal Chem. 2020;92(4):2937-45.
7. Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Analytical chemistry. 2006;78(3):779-87.
8. Smolinska A, Klaassen EM, Dallinga JW, van de Kant KD, Jobsis Q, Moonen EJ, et al. Profiling of volatile organic compounds in exhaled breath as a strategy to find early predictive signatures of asthma in children. PLoS One. 2014;9(4):e95668.
9. Pijls KE, Smolinska A, Jonkers DM, Dallinga JW, Masclee AA, Koek GH, et al. A profile of volatile organic compounds in exhaled air as a potential non-invasive biomarker for liver cirrhosis. Sci Rep. 2016;6:19903.
10. Filzmoser P, Walczak B. What can go wrong at the data normalization step for identification of biomarkers? J Chromatogr A. 2014;1362:194-205.
11. van den Berg RA, Hoefsloot HC, Westerhuis JA, Smilde AK, van der Werf MJ. Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genomics. 2006;7:142.
12. Bro R, Smilde AK. Principal component analysis. Anal Methods. 2014;6(9):2812-31.
13. Hubert M, Rousseeuw PJ, Vanden Branden K. ROBPCA: A New Approach to Robust Principal Component Analysis. Technometrics. 2005;47(1):64-79.
14. Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. Chemometrics and intelligent laboratory systems. 2001;58(2):109-30.
15. Gromski PS, Muhamadali H, Ellis DI, Xu Y, Correa E, Turner ML, et al. A tutorial review: Metabolomics and partial least squares-discriminant analysis--a marriage of convenience or a shotgun wedding. Anal Chim Acta. 2015;879:10-23.
16. Drucker H, Wu D, Vapnik VN. Support vector machines for spam categorization. IEEE Trans Neural Netw. 1999;10(5):1048-54.
17. Breiman L. Random Forest. Machine Learning. 2001;45:5-32.
18. Afanador NL, Smolinska A, Tran TN, Blanchet L. Unsupervised random forest: a tutorial with case studies. Journal of Chemometrics. 2016;30(5):232-41.
19. Freund Y, Schapire RE, editors. Experiments with a new boosting algorithm. icml; 1996: Citeseer.
20. Xu R, Wunsch D. Clustering: John Wiley & Sons; 2008.
21. T. Hastie RT, J. H. Friedman, editor. The Elements of Statistical Learning. Second Edition ed: Springer.
22. Sinha A, Desiraju K, Aggarwal K, Kutum R, Roy S, Lodha R, et al. Exhaled breath condensate metabolome clusters for endotype discovery in asthma. J Transl Med. 2017;15(1):262.
23. Dietterich TG, editor Ensemble methods in machine learning. International workshop on multiple classifier systems; 2000: Springer.
24. Friedman JH. Greedy function approximation: a gradient boosting machine. Annals of statistics. 2001:1189-232.

25. Smolinska A, Blanchet L, Coulier L, Ampt KA, Luider T, Hintzen RQ, et al. Interpretation and visualization of non-linear data fusion in kernel space: study on metabolomic characterization of progression of multiple sclerosis. PLoS One. 2012;7(6).

26. Smolinska A, Blanchet L, Coulier L, Ampt KAM, Luider T, Hintzen RQ, et al. Interpretation and Visualization of Non-Linear Data Fusion in Kernel Space: Study on Metabolomic Characterization of Progression of Multiple Sclerosis. PLoS One. 2012;7(6).

27. Purcaro G, Rees CA, Wieland-Alter WF, Schneider MJ, Wang X, Stefanuto PH, et al. Volatile fingerprinting of human respiratory viruses from cell culture. J Breath Res. 2018;12(2):026015.

28. Smolinska A, Bodelier AG, Dallinga JW, Masclee AA, Jonkers DM, van Schooten FJ, et al. The potential of volatile organic compounds for the detection of active disease in patients with ulcerative colitis. Aliment Pharmacol Ther. 2017;45(9):1244-54.

29. Kort S, Brusse-Keizer M, Gerritsen JW, van der Palen J. Data analysis of electronic nose technology in lung cancer: generating prediction models by means of Aethena. J Breath Res. 2017;11(2):026006.

30. Borràs E, Ferré J, Boqué R, Mestres M, Aceña L, Busto O. Data fusion methodologies for food and beverage authentication and quality assessment-A review. Analytica Chimica Acta. 2015;891:1-14.

31. Tran TN, Afanador NL, Buydens LM, Blanchet L. Interpretation of variable importance in partial least squares with significance multivariate correlation (sMC). Chemometrics and Intelligent Laboratory Systems. 2014;138:153-60.

32. Penrose LS. The elementary statistics of majority voting. Journal of the Royal Statistical Society. 1946;109(1):53-7.

33. Anderssen E, Dyrstad K, Westad F, Martens H. Reducing over-optimism in variable selection by cross-model validation. Chemometrics and Intelligent Laboratory Systems. 2006;84(1-2):69-74.

34. Westerhuis JA, Hoefsloot HCJ, Smit S, Vis DJ, Smilde AK, van Velzen EJJ, et al. Assessment of PLSDA cross validation. Metabolomics. 2008;4(1):81-9.

35. Szymanska E, Saccenti E, Smilde AK, Westerhuis JA. Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies. Metabolomics. 2012;8(1):S3-S16.

CHAPTER 4

Implementation of quality controls is essential to prevent batch effects in breathomics data and allow for cross-study comparisons

Georgios Stavropoulos, Daisy M. A. E. Jonkers, Zlatan Mujagic, Ger H. Koek, Ad A. M. Masclee, Marieke J. Pierik, Jan W. Dallinga, Frederik-Jan van Schooten, Agnieszka Smolinska

J. Breath Res. 14 026012,
doi: 10.1088/1752-7163/ab7b8d

Abstract

Introduction: Exhaled breath analysis has become a promising monitoring tool for various ailments by identifying volatile organic compounds (VOCs) as indicative biomarkers excreted in the human body. Throughout the process of sampling, measuring, and data processing, non-biological variations are introduced in the data leading to batch effects. Algorithmic approaches have been developed to cope with within-study batch effects. Batch differences, however, may occur among different studies too, and up-to-date, ways to correct for cross-study batch effects are lacking; ultimately, cross-study comparisons to verify the uniqueness of found VOC profiles for a specific disease may be challenging. This study applies within-study batch-effect-correction approaches to correct for cross-study batch effects; suggestions are made that may help prevent the introduction of cross-study variations.

Methods: Three batch-effect-correction algorithms were investigated: zero-centering, combat, and the analysis of covariance framework. The breath samples were collected from inflammatory bowel disease (n=213), chronic liver disease (n=189), and irritable bowel syndrome (n=261) patients at different periods, and they were analysed via gas chromatography-mass spectrometry. Multivariate statistics were used to visualise and verify the results.

Results: The visualisation of the data before any batch-effect-correction technique was applied showed a clear distinction due to probable batch effects among the datasets of the three cohorts. The visualisation of the three datasets after implementing all three correction techniques showed that the batch effects were still present in the data. Predictions made using partial least squares discriminant analysis and random forest confirmed this observation.

Conclusion: The within-study batch-effect-correction approaches fail to correct for cross-study batch effects present in the data. The present study proposes a framework for systematically standardising future breathomics data by using internal standards or quality control samples at regular analysis intervals. Further knowledge regarding the nature of the unsolicited variations among cross-study batches must be obtained to move the field further.

Keywords: exhaled breath, volatile organic compounds, VOCs, data analysis, batch effects, IBD, IBS, liver cirrhosis

Introduction

Breath analysis has recently emerged as a promising, non-invasive diagnostic and monitoring tool for a diversity of diseases [36, 37, 56, 64-69]. Volatile organic compounds (VOCs) have been identified in exhaled breath as indicative biomarkers among ailments. The rationale behind breath analysis is driven by the fact that disease-affected organs produce and therefore, release different VOCs in the human bloodstream. Eventually, those VOCs, due to their volatility, are then excreted through the body's air pathways [70]. Identified VOCs in human breath are not necessarily unique and specific for one particular illness; in some studies, similar VOCs have been assigned to different types of diseases [36, 37]. Thus, the distinction among those diseases becomes difficult or even impossible based on these single VOCs. Nonetheless, discriminatory power increases when searching for sets of compounds (i.e. VOC profiles) instead, and researchers have managed to find VOC profiles that aim to delineate different diseases based on a plethora of volatile metabolites found in exhaled breath [36, 37, 56, 64-69]. To prove that a set of compounds is indeed disease-specific, ideally, one would have to analyse the performance of a disease-specific VOC profile among other disease-related populations. A possibility to check for specificity and sensitivity of putative disease-specific VOC profiles is to use datasets from studies that have been generated over the years. However, these studies have been performed under different clinical settings, sampling periods, and sessional or instrumental conditions and as a result, caution is warranted when, retrospectively, datasets coming from multiple cohorts are used as input to validate biomarker performance since cross-study batch effects can be expected [71]. By definition, batch effects are sources of variation unrelated to the examined samples, or inter- or intra-sample class differences [71]. Environmental or methodological differences can cause batch effects during sample collection, chemical analysis, and data handling. 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, and this is because analytical techniques such as gas chromatography-mass spectrometry or nuclear magnetic resonance have become highly sophisticated and sensitive resulting in capturing both biological and non-biological variations [72]. Therefore, the reusability of existing exhaled breath datasets of VOCs for future biomarker-discovery and validation of studies might be challenging.

When combining data coming from multiple cohorts, it is crucial to detect and correct for any non-biological variations to prevent a compromised or even jeopardised analysis [71]. As such, in chemometrics, several within-study batch-effect-correction techniques are available to detect and correct for batch-induced variations, while retaining the biological information [73]. To the best of the authors' knowledge, no

cross-study batch-effect-correction algorithms have been reported in the literature. These algorithms have been successfully applied to different kinds of biological data, including metabolomics and predominantly gene microarray data [74-76]. The batch-effect-correction techniques can be divided into two major categories: the variable-wise correction and the sample-wise correction techniques. Zero-centering (i.e. mean-centering), and the analysis of covariance (ANCOVA) framework are variable-wise correction techniques that subtract predefined values from each variable [73, 76]. Combat and surrogate variables analysis (SVA) represent sample-wise correction techniques since they pool information across variables with similar expression characteristics in every batch [74, 75]. Finally, apart from the aforementioned techniques, a few more exist, but they are not examined in this study due to their less common implementation on biological data [77, 78].

To summarise, the dominance of batch effects in the data may hamper the discovery and subsequent validation of VOC profiles as disease-specific biomarkers. The present study aims to demonstrate the performance of currently available within-study batch-effect-correction algorithms to correct for cross-study batch effects. For that purpose, datasets of inflammatory bowel disease (IBD), irritable bowel syndrome (IBS), and liver cirrhosis (CIR) patients' samples were examined [37, 56, 64, 66]. These datasets were collected during different periods and used previously to identify discriminative VOCs for each of the diseases. As the final step of the current study, recommendations for future cross-study comparisons are provided to help overcome these cross-study challenges.

Materials and Methods

Data used

Three different datasets of exhaled breath samples were used, and they were obtained from patients suffering from gut and liver diseases. The content of exhaled breath in each dataset was chemically analysed utilising thermal desorption gas chromatography *time-of-flight* mass spectrometry (GC-*tof*-MS), at the same location (Maastricht University Medical Centre +, Maastricht, The Netherlands). The first dataset was collected and measured between 2009 and 2014, while the second dataset was sampled and measured between 2009 and 2012. The third one was measured between 2010 and 2012. The first dataset that was used was exhaled breath samples of IBS patients; detailed information on the IBS cohort can be found in Baranska et al. [64]. The second dataset consisted of exhaled breath samples of ulcerative colitis (UC) [56], and Crohn's disease (CD) patients [66], which both represent IBD cases. Finally, the third dataset was used for the identification of potential volatile biomarkers in human breath for liver cirrhosis [37]. To eliminate within-study batch effects as much as possible, all three datasets had the same volume, were measured with the

same instrument and with the same method; however, the analyst who conducted the experiment differed. Moreover, all three dataset samples were sampled by using sorbent tubes.

In the present study, 213 IBD (86 UC and 127 CD), 189 chronic liver disease (CLD), and 261 IBS exhaled breath samples were investigated. All patient studies were performed according to the revised version of the declaration of Helsinki, and they were approved by the local medical ethics committee. A detailed description of each dataset can be found elsewhere [37, 56, 64, 66], as well as the definition of the active and in remission states of the diseases are reported in the literature [56].

Data pre-processing

Initially, all the raw chromatograms were pre-processed before the actual analysis took place. Data pre-processing diminishes the effect of possible instrumental artefacts that can occur during the analysis. Data pre-processing, firstly, consisted of removing of the beginning and end of each chromatogram (i.e. retention time: < 1.3 and > 23 min) due to noisy mass spectra and column bleeding. Secondly, it was followed by noise removal via wavelets [31], baseline correction via P-splines [32], peak picking by combining the corresponding compounds based on their retention times and their mass spectra, and normalisation through probabilistic quotient normalisation [79]. A detailed description of the data pre-processing scheme is described in Smolinska et al. [80]. The majority of VOCs usually occurs only in a few samples [36]; consequently, only compounds that were detected in at least 10% of the samples were kept for further analysis. This led to a reduction in the total number of VOCs, from 7781 data points to 200 individual VOCs [33]. As a final pre-processing step, the data were logarithmically transformed [81]. The log transformation accounts for high skewness in the data. Pareto scaling [82] was also performed after every batch-effect correction attempt. Scaling accounts for giving all variables the same importance in the models.

Batch-effect-correction techniques

In the current study, three different within-study batch-effect-correction algorithms were examined: zero-centering [73], combat [74], and ANCOVA [76]. The selection of these three techniques was two-fold: it was based on their universal applicability in various fields of research, as well as on the way they correct for batch effects. Zero-centering was chosen as it is accepted as the first method of choice for correcting for batch effects [73]. Combat has been developed to process biological data, and it has proved that it maintains biological information in the data [74]. It is also considered the most commonly applied one due to its perceived high performance [83]. ANCOVA has been successfully utilised to correct for batch effects in metabolomics data while maintaining biological variability too [76].

In more detail, zero-centering removes unwanted variance within different batches by subtracting the mean of each measured parameter/variable from each measured parameter; thus, shifting the data of each batch to the origin (i.e. zero). Combat has been originally developed to deal with non-biological variation within different microarray experiments [74]. Based on either parametric or non-parametric empirical Bayes (EB) framework [84-86], combat finds EB estimates that robustly adjust the data. Parametric EB framework considers that a finite number of parameters defines the data distribution, whereas non-parametric EB framework considers that an infinite number of parameters defines the data distribution. Moreover, these EB estimates represent both the additive and multiplicative batch effect in the data assuming that these effects satisfy specific distributional forms. Combat was designed to remove unwanted variation in the data when the corresponding sources of variation are known in advance (e.g. different periods of measuring) [74]. The last correction technique investigated here was the ANCOVA framework [76], which consists of two main steps. In the first step, each parameter (i.e. VOC) of a dataset was transformed by subtracting the predicted value of that parameter in a given sample from its observed value. The predicted value of each particular parameter was obtained by using linear regression analysis [87]; the remaining parameters of the dataset are used as independent variables (i.e. predictors) in the regression analysis to get the predictions. In the second step, the mean value of the parameter across all samples was added to the predicted value. Note that those two steps are repeated for every single parameter in the data separately. The ANCOVA framework can be expressed as follows:

$$step\ 1:\ x_{inter,i} = x_{or,i} - \widehat{x_i}$$
$$step\ 2:\ x_{f,i} = x_{inter,i} + \overline{x_i}$$

where $x_{inter,i}$ is the intermediate value of parameter i and $x_{or,i}$ is the uncorrected/original value of parameter i. $\widehat{x_i}$ is the predicted value of parameter i, $\overline{x_i}$ is the mean value of parameter i across all samples, and $x_{f,i}$ is the corrected/final value of parameter i.

Also, all three aforementioned techniques and, in general, all correction techniques available are based on linear programming; only batch information that is represented in linear relation among parameters is removed from the data.

Inspection techniques

Three different methods were used to inspect and visualise the result of the batch-effect-correction techniques, namely, principal component analysis (PCA) [40], robust-PCA [41], and unsupervised random forest (URF) [46]. PCA and R-PCA discover eventual patterns and trends in the data by taking linear combinations of the original data. They both are used to explain the variance-covariance structure of the data; although the difference between them is that R-PCA is robust to outliers, while PCA

is highly affected by them. This is because PCA creates only one new dimension/direction at a time to capture most of the information in the original data. R-PCA, though, obtains consecutive possible "interesting" directions on which the original data are projected, and then, it selects the one that characterises the data the best. URF is based on the assumption that if there is any hidden linear or nonlinear structure in the data, it should be possible to distinguish them from a randomly generated version of themselves. It is worth mentioning that when one talks about linearities in the data, they refer to quantities, or variables in this case, which are proportional to each other. For example, if a variable increases, then another one either increases or decreases at a constant rate. Whatever falls outside this linearity definition is considered as a nonlinear relation [88].

Next to the visualisation techniques, the so-called Bhattacharyya distance [89], and surrogate variables analysis (SVA) method [75] were used to verify and compare the performance of the three batch correction techniques. Bhattacharyya distance is a quantitative way of assessing how similar (i.e. a measure of similarity) the examined datasets are by measuring the average distance between two normally distributed datasets:

$$D_B = \frac{1}{8} \times (\mu_1 - \mu_2)^T \times \Sigma^{-1} \times (\mu_1 - \mu_2) + \frac{1}{2} \times \left(\frac{\det \Sigma}{\sqrt{det\Sigma_1 \times det\Sigma_2}}\right)$$

where μ_1 and μ_2 are the means of the two datasets. $\det(\Sigma_1)$ and $\det(\Sigma_2)$ are the determinants of the covariance matrices of those two datasets, whereas $\det(\Sigma) = \det\left(\frac{\Sigma_1 + \Sigma_2}{2}\right)$. The D_B is an extension of the Mahalanobis distance [90] and it is considered more reliable because the Mahalanobis distance checks the similarity of the two classes when their standard deviation is the same, whereas D_B assumes that their standard deviations are different. The closer to zero the distance value is, the more similar the two datasets are.

SVA was particularly developed to solve heterogeneity problems in gene expression data, and the principle behind it is that it captures underlying information in the data. SVA assumes that some parameters (i.e. primary parameters) dominate over others during statistical modelling; therefore, valuable underlying information from the non-dominant/secondary parameters is lost (i.e. not captured). In short, SVA removes the signal from the primary variables to obtain a residuals matrix, and then it decomposes it. By doing so, SVA identifies subsets of parameters that significantly represent more variation than expected by chance and, for each subset of parameters, it creates a surrogate variable (SV). These SVs capture this so-called underlying information in the data, allowing for full parameters expression when put alongside the original parameters. In the present study, SVA was implemented as a means to confirm whether batch effects were only embedded in the dominant variables.

Validation

The last step of the batch-effect-correction techniques comparison and inspection section was to build prediction models by using supervised approaches such as random forest (RF) analysis [45], and partial least squares discriminant analysis (PLS-DA) [91]. The selection of those two techniques was two-fold: it was based on their successful implementation on biological studies [37, 56, 69, 80, 92], and on the way they consider the parameters of the data to be related to each other. RF considers both linear and nonlinear combinations of the data, while PLS-DA assumes only linearities in the data. For the PLS-DA and RF analyses, the three datasets were split into training and validation sets (one for each dataset) by using the Kennard-Stone algorithm [93]. In the present study, 80% of the samples of each dataset was used as a training set, while the remaining 20% was used as an internal independent validation set. Classification models via PLS-DA technique were performed on 2-class problem since the performance of such a model is better than a 3-class model [94]. The model optimisation (i.e. the optimal number of compounds and latent variables (LVs)) was achieved by the leave-out-cross-validation procedure [95]. More specifically, the optimal number of LVs was found by keeping out 10% of the training set at each cross-validation iteration, and the process was repeated 30 times. Moreover, when RF models were built, a significant feature extraction process was performed. RF provides the ability to identify significant variables for data classification/prediction. Based on this RF asset, several consecutive RF models were built, and for each one of them, the significant variables responsible for the data classification were identified. The first one would determine the most significant variables that discriminated the classes. Then, these variables were excluded, and predictions with the rest of the variables were made. Next, the second model would do the same, would identify the most important variables of the remaining, and it would give predictions. This procedure was repeated until only a few variables (e.g. 5-10) were left. The visualisation of all the models performance was achieved by using the so-called precision-recall (PR) curve [96]. The PR curve was used instead of the receiver operating characteristic (ROC) curve because the number of samples in the validation set is unbalanced. Both PLS and RF provide, as outcomes, probabilities of a sample being either IBD or CLD or IBS. Then, by using different thresholds (e.g. 0.1, 0.3, 0.7 etc.), a confusion matrix was calculated [97]. For every threshold, different sensitivities, specificities, and precisions are found. A PR curve plots, for every threshold, the pair of precision of the model (i.e. $\frac{\text{True Positives}}{True\ Positives+False\ Positives}$) against the recall (i.e. sensitivity) of the model (i.e. $\frac{\text{True Positives}}{True\ Positives+False\ Negatives}$) . The key characteristic of the PR curve is that it does not make use of the true negatives and therefore, it is only concerned about the correct prediction of the minority class. The minority class (i.e. negatives) differs in every model since it depends on which datasets (e.g. IBD vs. IBS) are used in the PLS case, and which datasets are compared against which class (e.g. IBD vs. CLD + IBS) in the RF case. A common feature between the PR and the ROC curves is that they both show a baseline threshold (i.e. 50% accuracy) above of which the performance of the model is considered better than random, whereas below this threshold the model

performance is random. In a PR curve, the baseline is determined as the number of positive samples over the total number of the training samples (i.e. the proportion of the positive samples in the dataset). All the analyses were performed in MatLab2016a except for the ANCOVA framework, which was performed in RStudio v1.1.453.

Results

Visualisation of the uncorrected data

The raw GC-*tof*-MS data of the 663 samples consisted of a total number of 7781 data points, which corresponds to more than a few hundreds of individual VOCs. After data pre-processing and data reduction steps, 200 VOCs were further examined. Three different batch-effect-correction methods were then applied to the pre-processed data and inspected. First, the uncorrected data were evaluated using visualisation techniques. Figure 1 illustrates the scores plots obtained from PCA, R-PCA, and URF. In all three cases, there were clear differences between the IBD and IBS patient cohorts indicated by dots and triangles, respectively, and the IBS cohort partially overlapped with the CLD cohort indicated by the squares. Furthermore, the groups were better separated in the R-PCA (Figure 1B) scores plot compared to the PCA scores plot (Figure 1A). The first three components in Figure 1 explained between 20% and 80% of the variance, indicating that the differences seen corresponded to the main variance in the data.

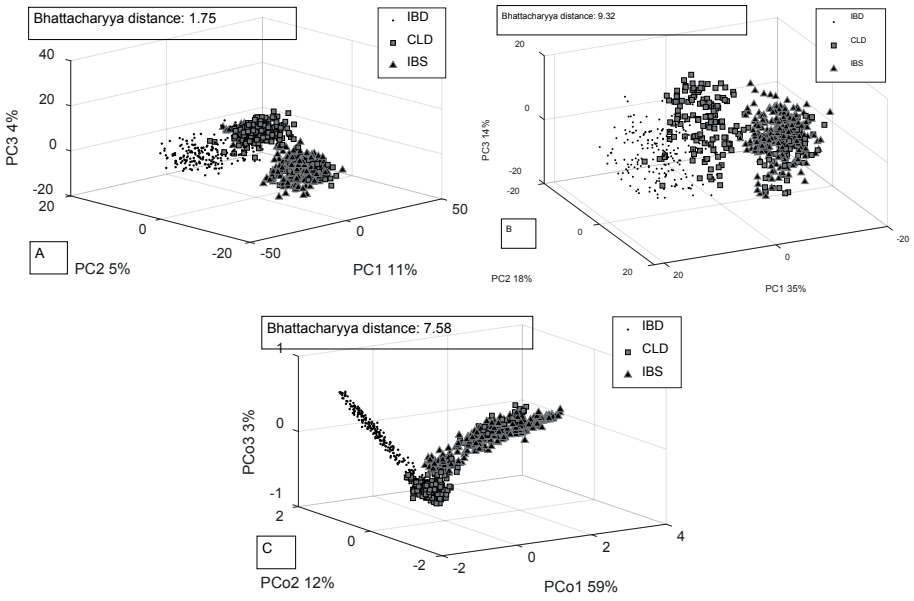


FIGURE 1: SCORES PLOTS OF THE THREE DATASETS OBTAINED BY (A) PCA, (B) R-PCA, AND (C) URF BEFORE ANY BATCH CORRECTION. THE PERCENTAGES INDICATE THE EXPLAINED INFORMATION BY EACH PC/PCO. THE DOTS REPRESENT IBD PATIENTS; THE SQUARES REPRESENT CLD PATIENTS; THE TRIANGLES REPRESENT IBS PATIENTS.

Implementation of the correction approaches and evaluation of their performance

For each one of the correction techniques, PCA, R-PCA, as well as URF were utilised to visualise the results. Figure 2 shows the corresponding scores plots after correcting for batch effects by employing the ANCOVA framework. The correction with ANCOVA led to a uniform cloud of data points with no visible classes in the cases of PCA and R-PCA (Figures 2A and 2B). The opposite trend was observed in the URF scores plot (Figure 2C). The average Bhattacharyya distances for the ANCOVA corrected data matrix of the classes were 6.89, 12.45, and 19.33, respectively; for the uncorrected data matrix of the classes, they were 1.75, 9.32, and 7.58, respectively, indicating that the batch effect had not been removed. Instead, loss of biological information had possibly happened since the data point clouds had shrunk, and the distance values had increased. Correcting the data by employing zero-centering and combat led to a similar outcome (Figure 1S and Figure 2S).

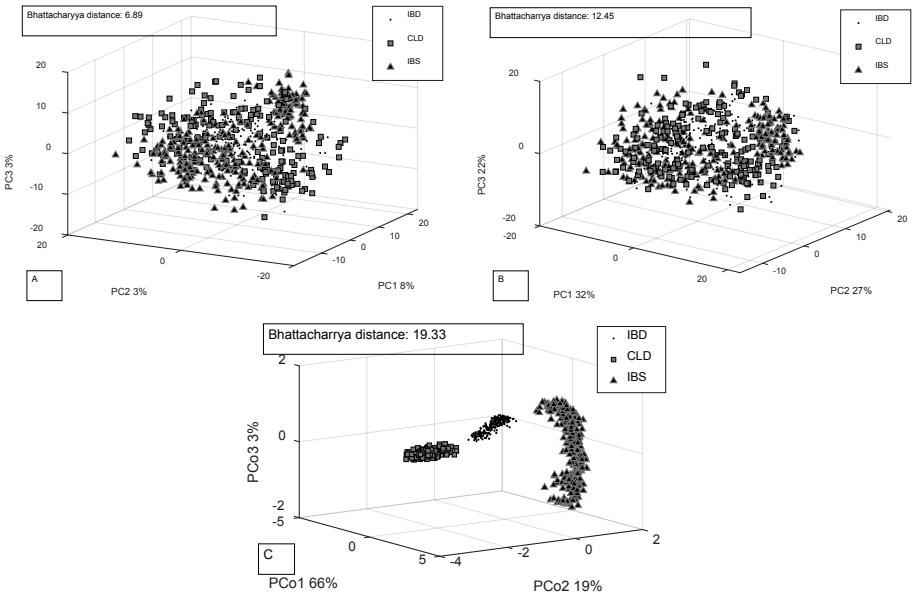


FIGURE 2: SCORES PLOTS OF THE THREE DATASETS OBTAINED BY (A) PCA, (B) R-PCA, AND (C) URF AFTER IMPLEMENTING ANCOVA. THE PERCENTAGES INDICATE THE EXPLAINED INFORMATION BY EACH PC/PCO. THE DOTS REPRESENT IBD; THE SQUARES REPRESENT CLD; THE TRIANGLES REPRESENT IBS.

Figure 3 illustrates the SVA implementation of the ANCOVA corrected data. Figures 3A and 3B show the scores of the SVs of each class obtained by PCA and R-PCA, respectively. Both figures lack groupings, suggesting that the batch effect had been removed. Similar to the previous case, the URF scores plot (Figure 3C) illustrates clear clusters, indicating that the batch effect was still present. The zero-centering and combat results can be found in the supplementary material (Figure 3S and Figure 4S). To further support the conclusion that the batch effect remained and it might be embedded in the data in nonlinear ways after applying several correction techniques, supervised approaches (i.e. PLS-DA and RF) were used to determine prediction accuracy of these three classes.

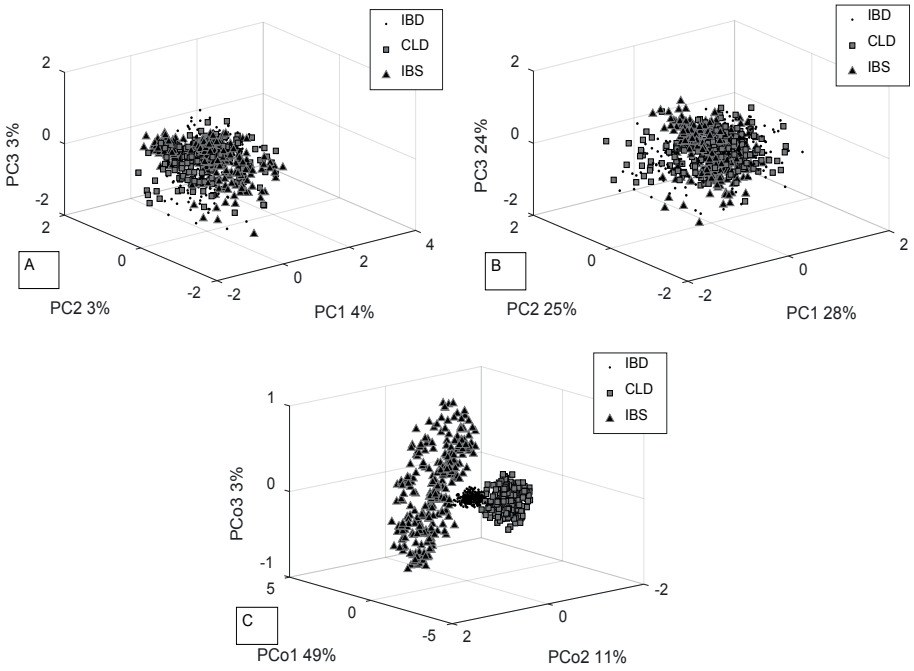


FIGURE 4: SCORES PLOTS OF THE THREE DATASETS OBTAINED BY (A) PCA, (B) R-PCA, AND (C) URF ON THE SVS AFTER IMPLEMENTING ANCOVA. THE PERCENTAGES INDICATE THE EXPLAINED INFORMATION BY EACH PC/PCO. THE DOTS REPRESENT IBD; THE SQUARES REPRESENT CLD; THE TRIANGLES REPRESENT IBS.

Implementation of supervised approaches to determine the prediction accuracy of the classes

As a training set, 152 samples of each class were used. The remaining samples of each class were used as a validation set. Therefore, the complete training set consisted of 456 samples, while the validation set consisted of 207 samples: 61 IBD, 37 CLD, and 109 IBS.

For PLS-DA, three different 2-class models were built (i.e. IBD vs IBS, IBD vs CLD, and CLD vs IBS). For clarity, only the PLS-DA results of the three 2-class models after implementing ANCOVA are illustrated in the form of PR curves (Figure 4). The results of PLS-DA after applying zero-centering and combat can be found in the supplementary materials (Figure 5S and Figure 6S).

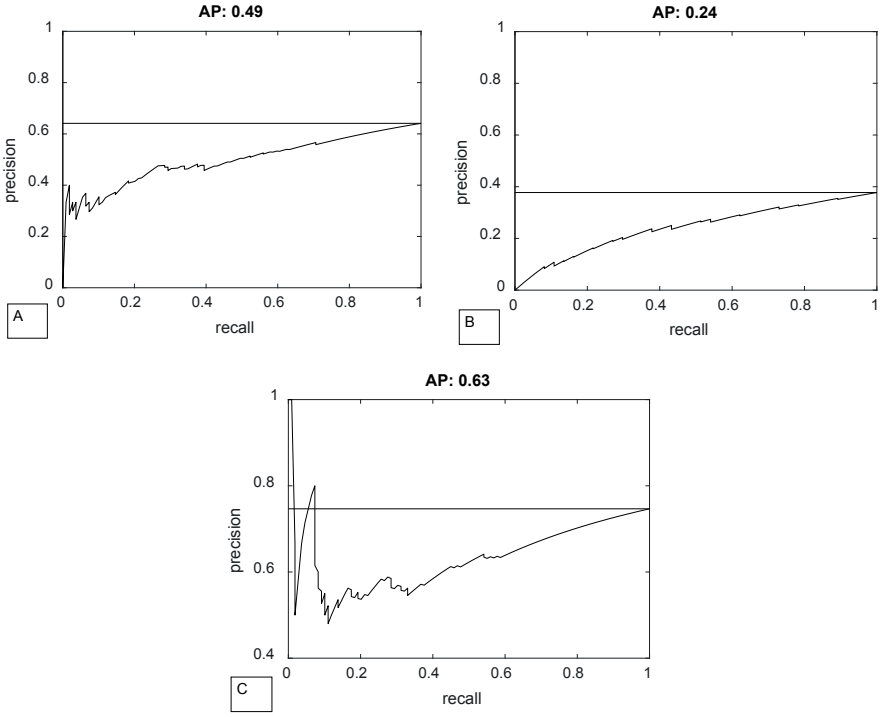


FIGURE 5: PR CURVES OF THE VALIDATION SET GIVEN BY PLS-DA OF THE (A) IBD VS IBS MODEL WITH IBS BEING THE MAJORITY CLASS, (B) IBD VS CLD MODEL WITH IBD BEING THE MAJORITY CLASS, AND (C) CLD VS IBS MODEL WITH IBS BEING THE MAJORITY CLASS AFTER IMPLEMENTING ANCOVA. THE HORIZONTAL LINE IS THE BASELINE; THE AREA ABOVE THE BASELINE IS THE GOOD PERFORMANCE AREA, WHEREAS THE AREA BELOW THE BASELINE IS THE POOR PERFORMANCE AREA.

Figure 4 shows that all three 2-class PLS-DA models, IBD vs IBS, IBD vs CLD, and CLD vs IBS, performed poorly with average precision (AP) values of 0.49, 0.24, and 0.63, respectively. Also, 2-class PLS-DA prediction models were made on the SVs adjusted data. The PR curves on the SVs adjusted data for all three-correction attempts and all three 2-class models looked similar to those obtained from the corrected data and therefore, they are not shown. For comparison purposes, 2-class PLS-DA models were also constructed for the uncorrected data. The corresponding PR curves obtained for the validation sets are seen in Figure 5.

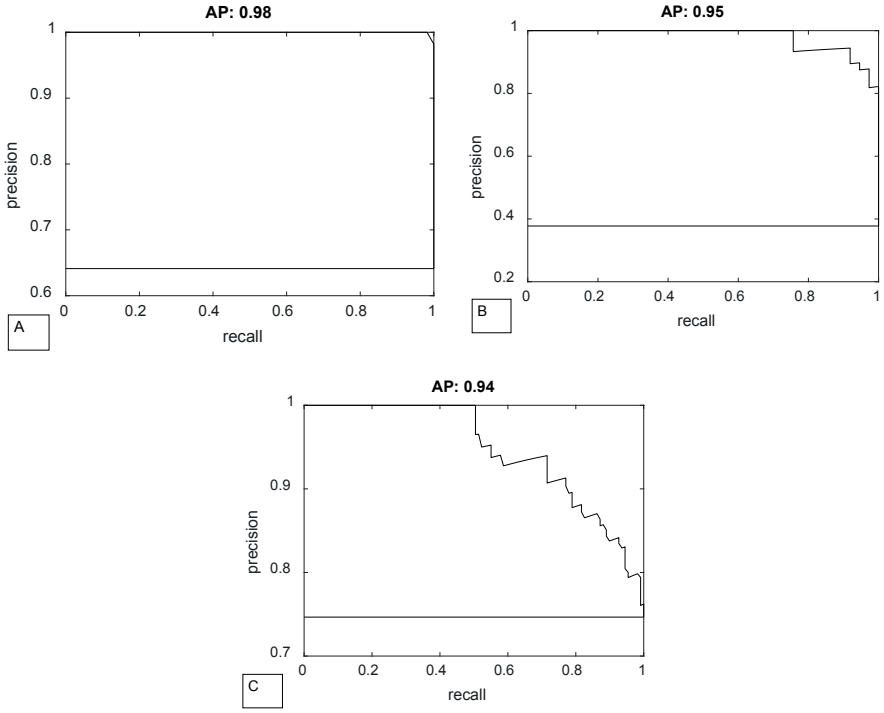


FIGURE 6: PR CURVES OF THE VALIDATION SET GIVEN BY PLS-DA OF THE (A) IBD VS IBS MODEL WITH IBS BEING THE MAJORITY CLASS, (B) IBD VS CLD MODEL WITH IBD BEING THE MAJORITY CLASS, AND (C) CLD VS IBS MODEL WITH IBS BEING THE MAJORITY CLASS WITHOUT ANY BATCH EFFECT CORRECTION. THE HORIZONTAL LINE IS THE BASELINE; THE AREA ABOVE THE BASELINE IS THE GOOD PERFORMANCE AREA, WHEREAS THE AREA BELOW THE BASELINE IS THE POOR PERFORMANCE AREA.

Achieving 0.98, 0.95, and 0.94 AP values for the IBD vs IBS, IBD vs CLD, and CLD vs IBS models, respectively, indicated that the batch effect was nonlinearly embedded in the data. RF was implemented by using 1000 trees, and it resulted in 100% accuracy (i.e. perfect classifier) for all three classes after applying all three-correction techniques. Furthermore, RF was also implemented on the SVs of each class and once more, 100% accuracy was achieved for all three classes.

Discussion

The present study demonstrated the occurrence of batch effects in breathomics data obtained across independently performed studies by using untargeted GC-*tof*-MS analysis. The effect of various within-study batch correction techniques was evaluated, and their outcomes were presented by using case studies of IBD, IBS, and CLD. In particular, three different correction techniques were applied: zero-centering, combat, and ANCOVA. When applying the correction algorithms, visualisation by PCA and R-PCA showed removal of batch effects, whereas when the results were visualised by URF, they displayed otherwise. Bhattacharyya distance values and SVA confirmed that batch effects still remained after correction.

To comprehend and interpret the results of this study, the pathology of the diseases that were examined has to be taken into account. IBD is a gastrointestinal (GI) tract disease that causes inflammation of the colon in the UC case and of any part of the GI tract in the CD case [98]. IBS is a disorder of the gut-brain interaction without clinically relevant organic pathology of the GI tract [99]. Clinically, ileo-colonoscopy and histology are the gold standards that can distinguish these two conditions [100]. Liver cirrhosis is the end-stage of CLD characterised by abnormal structure and function of the liver, and many liver disease patients also suffer from either IBD or IBS [101]. Hence, these three disorders share various symptoms and some pathophysiological mechanisms, and as a result, an explorative analysis should reveal some overlap or at least, not such apparent distinction among samples obtained from those three diseases in a score plot space. However, the first results demonstrated indicated an apparent separation along the first two PCs, suggesting that variation unrelated to the biological class differences is, probably, present in the data. Another reason that supports the presence of non-biological information in the data is the fact that, reportedly, several endogenous compounds such as aldehydes, short fatty acids, and branched-chain alkenes have been linked to all three diseases [102, 103]. Therefore, such an apparent distinction among the samples would not be realistic. This non-biological variation could be caused by several factors such as environmental, instrumental or differences at the periods at which the samples were measured or by the person that conducted the clinical sampling or chemical analysis. In this study, the three datasets were measured at different periods and at a different location, which probably led to the differences among the batches, and explains why these three datasets created distinct data point clouds (Figure 1).

Initially, PCA and R-PCA were used to visualise the results obtained after trying to correct for the batch effects; their scores plots (Figures 2A and 2B) demonstrated that the batch effects were diminished since the data point clouds overlapped. However, the Bhattacharyya distance values for the classes have increased compared to the uncorrected data distance values of the classes, meaning that the datasets have not

become more similar as expected. This is because to calculate the Bhattacharyya distance values all the PCA/R-PCA scores were used, whereas, in the score plots, only the first three scores are used to visualise the samples. When a more powerful tool was applied (i.e. URF), not only the Bhattacharyya distance values increased but also the visualisation of the first three scores showed that the three classes were differentiated (Figure 2C). The Bhattacharyya distance, in the URF case, has increased and the data point clouds have shrunk, confirming that there is a probable loss of biological information instead of elimination of batch effects. In particular, the IBD and CLD data point clouds shrunk (Figures 2C and 3C), and the fact that the data points came so close to each other illustrates that, most likely, there is no biological variation left among these particular samples. The reason why PCA and R-PCA failed to show the batch effect in the corrected data is that they search for linear combinations among the measured parameters of every sample [40, 41], whereas URF searches for both linear and nonlinear combinations among the measured parameters of every sample [46]. This indicates that the batch effect might be nonlinearly embedded in the data. Apart from the distance measure, SVA was also applied to the corrected data. Visualisation of the SVs confirmed that the batch effect is still present even though the primary information in the data was deducted, and the secondary or underlying information was brought to the surface (Figure 3). In practice, SVA is used to create the SVs that are meant to be put next to the original variables to express the samples fully. However, in this case, the batch effect was embedded in the original variables, and it dominated the primary information of the data. Therefore, the SVs were plotted alone (Figure 3) to visualise the samples rather than putting the SVs next to the original variables. To further confirm this conclusion, classification models were built to predict the three classes using PLS-DA and RF. PLS-DA performed on the corrected data demonstrated that discrimination of the classes failed (Figure 4), while it gave almost perfect prediction when applied to the uncorrected data (Figure 5). This means that the batch effect is not embedded in the data in linear ways, and that loss of biological information happens instead because achieving almost 100% accuracy is not realistic. The RF classification models revealed an overall accuracy of 100% when all the sample variables were used. Moreover, in this RF implementation, the significant feature extraction process was tried. Once again, 100% accuracy for the three classes was achieved throughout this backwards variable elimination process (results are not shown). In general, it is possible to achieve very good prediction accuracy models even with a small number of variables (e.g. 5-10); although, if it were not for the batch effect, the accuracy percentage would fluctuate.

Recently, Nakhleh et al. [104] conducted a large study aiming to classify patients of 17 different diseases by identifying VOCs in exhaled breath. By using discriminant factor analysis, they reported an average classification accuracy for all diseases of 86%, including IBD and IBS patients. For the IBD and IBS classification, in particular, they documented an approximately 80% accuracy while for some of the other diseases,

they reported 100% accuracy. The data used in their study were collected at different periods, different laboratories or even different countries, and they were generated via NanoArray and GC-MS technologies; however, nothing was reported regarding possible cross-study or even within-study batch effects that may have influenced their study results. Fijten et al. [80] established that it is difficult to achieve the same model accuracy even for the same disease when external validation sets are used. More specifically, Fijten et al. [80] identified a set of VOCs that differentiated sarcoidosis patients from healthy cohorts with an accuracy of almost 80%. Although, several years later, new sarcoidosis patients were recruited to validate the existing model, and an accuracy of 53% was achieved. The authors attributed this large decrease in their model accuracy to probable batch effects. Bearing that in mind, it is understood that batch effects affect the classification of different diseases and thus, confirming the results of the present study.

All three batch-effect-correction techniques demonstrated here failed because they look for linearities in the data and, as URF results suggested (Figure 2C and 3C), the batch effect may be nonlinearly embedded in the data. Furthermore, in the IBS uncorrected dataset, two small subgroups were also observed. When PCA, R-PCA, and URF were implemented after every correction technique, only the URF score plot showed that these small subgroups are still present. As a result, it may be assumed that within-study batch effects in breathomics data may be nonlinearly related to biological variations. To the best of the authors' knowledge, sufficient nonlinear ways of correcting for batch effects have not been reported in the literature yet. Recently, Shaham et al. [105] reported a nonlinear way of correcting for batch effects based on the residuals of neural networks, which outperformed zero-centering and combat. Their approach was developed for and applied to replicate samples of the same object (e.g. patient). This means that their approach is only suitable for data containing samples whose multivariate distributions are close to each other. This is, however, not the case here, where multiple batches with multiple samples (i.e. numerous instead of a single individual) are present, and therefore, the technique was not applied. Shaham et al. [105] indicated that a nonlinear approach which could be applied on multiple batches with multiple samples each needs to be developed. Such a nonlinear development may be challenging, though, because of the trade-off between biological information of samples and non-biological variation due to external influences. Even if a more advanced, nonlinear correction technique were developed, which would remove external influences accurately, it may come at the cost of eliminating relevant biological information. Another reason, which makes such a development challenging, is the complexity of tracing back nonlinearly transformed variables. Supposedly, a complex nonlinear batch-effect-correction model is developed. Then, the batch effect corrected data are used for further analysis and biomarkers discovery. At the end of this biomarkers-discovery process, a few variables are found as being important in classifying individuals into different groups (e.g. healthy against diseased). It may

then be challenging to trace back these important variables to the original ones due to the nonlinear transformation they have gone through and therefore, to identify which compounds were responsible for the classification. Additionally, little is known about effective batch effect mitigation. Another cause for the failure of these particular correction techniques may be the existence of subpopulations within the classes. It has been reported that subpopulations within a dataset may affect batch corrections [71]. In this study, IBD consisted of UC and CD samples, while CLD consisted of patients with and without cirrhosis and different underlying aetiologies of liver disease. More importantly, the most probable cause of failure of these particular correction techniques is that they are developed to deal with within-study batch effects, not with across studies external variations. Such a technique seems rather challenging to be developed since it should take into account many factors in correcting the data; nevertheless, it is believed that this problem may be overcome.

Currently, a standardisation framework for breath analysis research is lacking, and to achieve standardised protocols for sampling and measuring, several initiatives are ongoing within both the International Association of Breath Research (IABR) [106, 107] and the European Respiratory Society [108]. More specifically, the development and dissemination of a method for evaluating breath sampling and analysis techniques, with accompanying benchmark data, was prioritised by the IABR joint task force in 2016; as a result, the multi-centre study, the so-called “Peppermint” study, was created [109]. In this study, an oral administration of capsules with peppermint is used to monitor the perturbation in human breath over time. In the “Peppermint” consortium, benchmark data will be created by various breath sampling as well as analytical approaches, giving scientists in the breath community the possibility to monitor their analytical pipeline. To help develop such a standardisation framework, in future breathomic studies, the injection of internal standards in the breath samples and/or the inclusion of quality control (QC) samples at regular intervals throughout the measuring process is highly recommended. Such a standardisation process would make use of both QC information as well as batch labels and injection sequence information. Metabolomic studies have indicated that the use of QCs can help eliminate within-study batch effects, and ultimately, make the data suitable for cross-study comparisons too [110, 111]. In the metabolomics world, scientists can make use of pooled samples of, for example, urine or blood to use as their QC samples; however, pooling breath samples is not feasible. To overcome this issue, breath research QC mixtures should reflect the content of the measured breath samples, and preliminary ongoing QC applications point that QCs can indeed help correct for non-biological variations. An ideal QC mixture should contain compounds such as alkanes, alkenes, aldehydes, ketones, alcohols, and acids since these are stable and always present in breath samples [112]. For example, when putative markers are known due to *a priori* knowledge, such as dimethyl sulphide and limonene in liver disease [1, 37], these compounds can be used in the QC mixture. It should also be noted that a QC mixture

should contain established amounts of these known compounds at concentrations close to those that are known to be present in the subject samples. Nonetheless, several aspects have to be carefully considered when QCs are implemented: possible interaction of the internal standard with the analysis samples, and the optimal number of QC samples to be used, to name a few. This is because there is not a straightforward process as to what is optimal each time; it depends on different aspects, such as the stability of the compounds or the stability of the analytical system [76]. To the best of the authors' knowledge, a similar study that investigates within-study and cross-study batch effects in the breath-omics field has not been previously performed. The need for developing more advanced, nonlinear ways for batch effect removal, as well as the fact that dealing with batch effects is neither a straightforward nor an easy task to control, support the novelty of this study.

Conclusion

In conclusion, the present study revealed that batch effect challenges arise in untargeted VOC analysis with GC-*tof*-MS, and the current ways of correcting them with algorithmic techniques do not suffice. Attention should be paid on developing more advanced, nonlinear batch-effect-correction algorithmic methods. Most urgently, however, the need for a standardisation framework is of paramount importance; therefore, the use of QCs in future breath analyses should become a common practice since correcting for confounding influences afterwards seems to be challenging at the moment

.

Acknowledgements

The present study was supported by the VENI grant, Netherlands organisation for scientific research (NWO) no. 016 VENI 178.064.

References

1. Baranska, A., et al., *Volatile organic compounds in breath as markers for irritable bowel syndrome: a metabolomic approach.* Aliment Pharmacol Ther, 2016. **44**(1): p. 45-56.
2. Baranska, A., et al., *Profile of volatile organic compounds in exhaled breath changes as a result of gluten-free diet.* J Breath Res, 2013. **7**(3): p. 037104.
3. Bodelier, A.G., et al., *Volatile Organic Compounds in Exhaled Air as Novel Marker for Disease Activity in Crohn's Disease: A Metabolomic Approach.* Inflamm Bowel Dis, 2015. **21**(8): p. 1776-85.
4. Mujagic, Z., et al., *A novel biomarker panel for irritable bowel syndrome and the application in the general population.* Sci Rep, 2016. **6**: p. 26420.
5. Pijls, K.E., et al., *A profile of volatile organic compounds in exhaled air as a potential non-invasive biomarker for liver cirrhosis.* Sci Rep, 2016. **6**: p. 19903.
6. Schnabel, R., et al., *Analysis of volatile organic compounds in exhaled breath to diagnose ventilator-associated pneumonia.* Sci Rep, 2015. **5**: p. 17179.
7. Smolinska, A., et al., *The potential of volatile organic compounds for the detection of active disease in patients with ulcerative colitis.* Aliment Pharmacol Ther, 2017. **45**(9): p. 1244-1254.
8. van Vliet, D., et al., *Can exhaled volatile organic compounds predict asthma exacerbations in children?* J Breath Res, 2017. **11**(1): p. 016016.
9. Smolinska, A., et al., *Profiling of volatile organic compounds in exhaled breath as a strategy to find early predictive signatures of asthma in children.* PloS one, 2014. **9**(4): p. e95668.
10. Miekisch, W., J.K. Schubert, and G.F. Noeldge-Schomburg, *Diagnostic potential of breath analysis—focus on volatile organic compounds.* Clinica chimica acta, 2004. **347**(1-2): p. 25-39.
11. Goh, W.W.B., W. Wang, and L. Wong, *Why Batch Effects Matter in Omics Data, and How to Avoid Them.* Trends Biotechnol, 2017. **35**(6): p. 498-507.
12. Leek, J.T., et al., *Tackling the widespread and critical impact of batch effects in high-throughput data.* Nature Reviews Genetics, 2010. **11**: p. 733.
13. Nygaard, V., E.A. Rodland, and E. Hovig, *Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses.* Biostatistics, 2016. **17**(1): p. 29-39.
14. Johnson, W.E., C. Li, and A. Rabinovic, *Adjusting batch effects in microarray expression data using empirical Bayes methods.* Biostatistics, 2007. **8**(1): p. 118-27.
15. Leek, J.T. and J.D. Storey, *Capturing Heterogeneity in Gene Expression Studies by "Surrogate Variable Analysis".* PLoS Genetics, 2005. **preprint**(2007).
16. Wehrens, R., et al., *Improved batch correction in untargeted MS-based metabolomics.* Metabolomics, 2016. **12**: p. 88.
17. Luo, J., et al., *A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data.* The pharmacogenomics journal, 2010. **10**(4): p. 278.
18. Benito, M., et al., *Adjustment of systematic microarray data biases.* Bioinformatics, 2004. **20**(1): p. 105-114.
19. Walczak, B., *Wavelets in Chemistry. Data Handling in Science and Technology*. 2000: Elsevier Science & Technology.
20. Eilers, P.H., *A perfect smoother.* Analytical chemistry, 2003. **75**(14): p. 3631-3636.
21. Dieterle, F., et al., *Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics.* Analytical chemistry, 2006. **78**(13): p. 4281-4290.
22. Fijten, R.R.R., et al., *The necessity of external validation in exhaled breath research: a case study of sarcoidosis.* J Breath Res, 2017. **12**(1): p. 016004.
23. Smolinska, A., et al., *Current breathomics—a review on data pre-processing techniques and machine learning in metabolomics breath analysis.* Journal of breath research, 2014. **8**(2): p. 027105.
24. Changyong, F., et al., *Log-transformation and its implications for data analysis.* Shanghai archives of psychiatry, 2014. **26**(2): p. 105.

25. van den Berg, R.A., et al., *Centering, scaling, and transformations: improving the biological information content of metabolomics data.* BMC genomics, 2006. **7**(1): p. 142.
26. Kaderali, L., et al., *Removing Batch Effects from Longitudinal Gene Expression - Quantile Normalization Plus ComBat as Best Approach for Microarray Transcriptome Data.* Plos One, 2016. **11**(6).
27. Casella, G., *An introduction to empirical Bayes data analysis.* The American Statistician, 1985. **39**(2): p. 83-87.
28. Johns Jr, M., *Non-parametric empirical Bayes procedures.* The Annals of Mathematical Statistics, 1957: p. 649-669.
29. Morris, C.N., *Parametric empirical Bayes inference: theory and applications.* Journal of the American Statistical Association, 1983. **78**(381): p. 47-55.
30. Montgomery, D.C., E.A. Peck, and G.G. Vining, *Introduction to linear regression analysis*. Vol. 821. 2012: John Wiley & Sons.
31. Bro, R. and A.K. Smilde, *Principal component analysis.* Anal. Methods, 2014. **6**(9): p. 2812-2831.
32. Hubert, M., P.J. Rousseeuw, and K. Vanden Branden, *ROBPCA: A New Approach to Robust Principal Component Analysis.* Technometrics, 2005. **47**(1): p. 64-79.
33. Afanador, N.L., et al., *Unsupervised random forest: a tutorial with case studies.* Journal of Chemometrics, 2016. **30**(5): p. 232-241.
34. Sivapalan, M., C. Jothityangkoon, and M. Menabde, *Linearity and nonlinearity of basin response as a function of scale: discussion of alternative definitions.* Water Resources Research, 2002. **38**(2): p. 4-1-4-5.
35. Bhattacharyya, A.K., *A Measure of Divergence between Two Statistical Populations Defined by Their Probability Distributions.* Mathematical Society, 1943(35): p. 99-109.
36. Mahalanobis, P., *On the generalised distance in statistics (Vol. 2, pp. 49–55).* Proceedings National Institute of Science, India. Retrieved from http://ir. isical. ac. in/dspace/handle/1/1268, 1936.
37. Breiman, L., *Random Forest.* Machine Learning, 2001. **45**: p. 5-32.
38. Barker, M. and W. Rayens, *Partial least squares for discrimination.* Journal of Chemometrics, 2003. **17**(3): p. 166-173.
39. Perez-Enciso, M. and M. Tenenhaus, *Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (PLS-DA) approach.* Hum Genet, 2003(112): p. 581-592.
40. Kennard, R.W. and L.A. Stone, *Computer aided design of experiments.* Technometrics, 1969. **11**(1): p. 137-148.
41. Smolinska, A., et al., *NMR and pattern recognition can distinguish neuroinflammation and peripheral inflammation.* Journal of proteome research, 2011. **10**(10): p. 4428-4438.
42. Kohavi, R. *A study of cross-validation and bootstrap for accuracy estimation and model selection*. in *Ijcai*. 1995. Montreal, Canada.
43. Saito, T. and M. Rehmsmeier, *The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets.* PloS one, 2015. **10**(3): p. e0118432.
44. Stehman, S.V., *Selecting and interpreting measures of thematic classification accuracy.* Remote sensing of Environment, 1997. **62**(1): p. 77-89.
45. Podolsky, D.K., *Inflammatory bowel disease.* New England Journal of Medicine, 1991. **325**(13): p. 928-937.
46. Whitehead, W.E., B.T. Engel, and M.M. Schuster, *Irritable bowel syndrome.* Digestive diseases and sciences, 1980. **25**(6): p. 404-413.
47. Schoepfer, A.M., et al., *Discriminating IBD from IBS: comparison of the test performance of fecal markers, blood leukocytes, CRP, and IBD antibodies.* Inflammatory bowel diseases, 2007. **14**(1): p. 32-39.
48. Pinzani, M., M. Rosselli, and M. Zuckermann, *Liver cirrhosis.* Best practice & research Clinical gastroenterology, 2011. **25**(2): p. 281-290.
49. Bannaga, A.S., A. Farrugia, and R.P. Arasaradnam, *Diagnosing Inflammatory bowel disease using noninvasive applications of volatile organic compounds: a systematic review.* Expert review of gastroenterology & hepatology, 2019. **13**(11): p. 1113.
50. Markar, S.R., et al., *Exhaled breath analysis for the diagnosis and assessment of endoluminal gastrointestinal diseases.* Journal of clinical gastroenterology, 2015. **49**(1): p. 1-8.

51. Nakhleh, M.K., et al., *Diagnosis and Classification of 17 Diseases from 1404 Subjects via Pattern Analysis of Exhaled Molecules.* ACS Nano, 2017. **11**(1): p. 112-125.
52. Shaham, U., et al., *Removal of batch effects using distribution-matching residual networks.* Bioinformatics, 2017. **33**(16): p. 2539-2546.
53. Herbig, J. and J. Beauchamp, *Towards standardization in the analysis of breath gas volatiles.* Journal of breath research, 2014. **8**(3): p. 037101.
54. Gaude, E., et al., *Targeted breath analysis: exogenous volatile organic compounds (EVOC) as metabolic pathway-specific probes.* Journal of breath research, 2019. **13**(3): p. 032001.
55. Horváth, I., et al., *A European Respiratory Society technical standard: exhaled biomarkers in lung disease.* European Respiratory Journal, 2017. **49**(4): p. 1600965.
56. Malásková, M., et al., *Proton transfer reaction time-of-flight mass spectrometric measurements of volatile compounds contained in peppermint oil capsules of relevance to real-time pharmacokinetic breath studies.* Journal of breath research, 2019. **13**(4): p. 046009.
57. Sánchez-Illana, Á., et al., *Evaluation of batch effect elimination using quality control replicates in LC-MS metabolite profiling.* Analytica chimica acta, 2018. **1019**: p. 38-48.
58. Sanchez-Illana, A., et al., *Model selection for within-batch effect correction in UPLC-MS metabolomics using quality control-Support vector regression.* Analytica chimica acta, 2018. **1026**: p. 62-68.
59. Amann, A., et al., *The human volatilome: volatile organic compounds (VOCs) in exhaled breath, skin emanations, urine, feces and saliva.* Journal of breath research, 2014. **8**(3): p. 034001.
60. Friedman, M.I., et al., *Limonene in expired lung air of patients with liver disease.* Digestive diseases and sciences, 1994. **39**(8): p. 1672-1676.

Supplementary materials

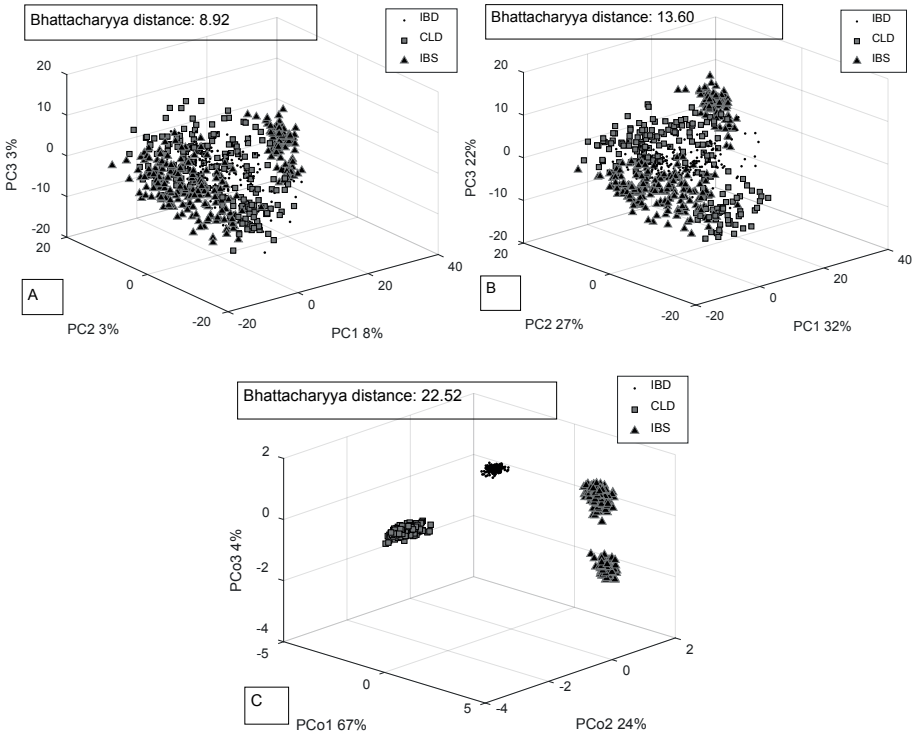


FIGURE 1S: SCORES PLOTS OF THE THREE DATASETS OBTAINED BY (A) PCA, (B) R-PCA, AND (C) URF AFTER IMPLEMENTING ZERO-CENTERING. THE BHATTACHARYYA DISTANCE FOR THE ZERO-CENTERING CASE IS 8.92, 13.60, AND 22.52, RESPECTIVELY. THE PERCENTAGES INDICATE THE EXPLAINED INFORMATION BY EACH PC/PCO. THE DOTS REPRESENT IBD; THE SQUARES REPRESENT CLD; THE TRIANGLES REPRESENT IBS. THE GROUPING OBSERVED IN THE IBS CASES IS ALONG PCO3 WHICH ONLY CAPTURES 4% OF THE EXPLAINED INFORMATION, AND THEREFORE, IT CAN BE IGNORED OR CONSIDERED IRRELEVANT.

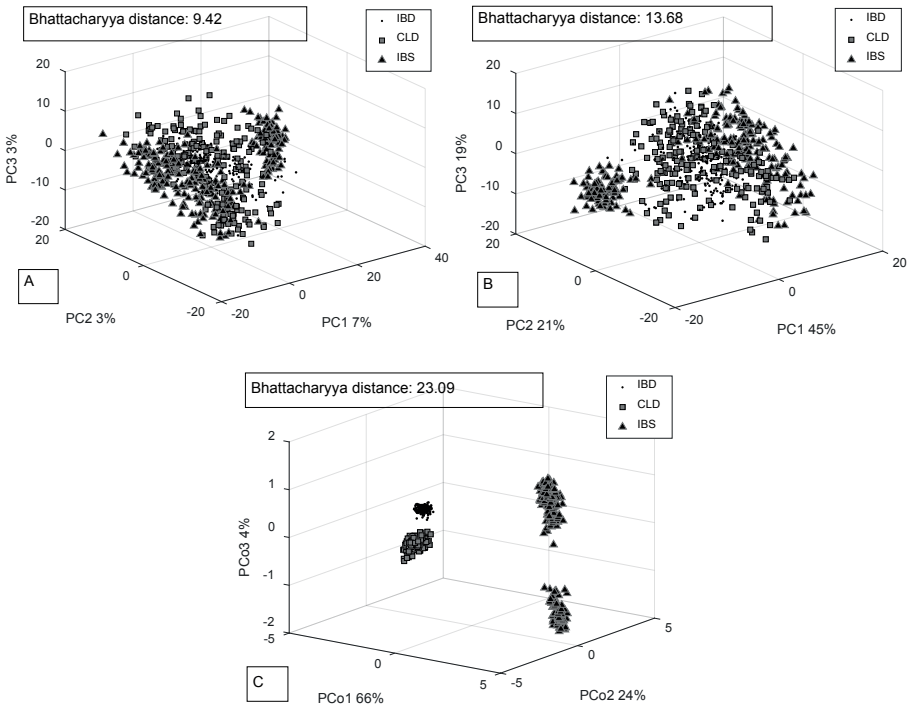


FIGURE 2S: SCORES PLOTS OF THE THREE DATASETS OBTAINED BY (A) PCA, (B) R-PCA, AND (C) URF AFTER IMPLEMENTING COMBAT. THE BHATTACHARYYA DISTANCE FOR THE COMBAT CASE IS 9.42, 13.68, AND 23.09, RESPECTIVELY. THE PERCENTAGES INDICATE THE EXPLAINED INFORMATION BY EACH PC/PCO. THE DOTS REPRESENT IBD; THE SQUARES REPRESENT CLD; THE TRIANGLES REPRESENT IBS. THE GROUPING OBSERVED IN THE IBS CASES IS ALONG PCO3 WHICH ONLY CAPTURES 4% OF THE EXPLAINED INFORMATION, AND THEREFORE, IT CAN BE IGNORED OR CONSIDERED IRRELEVANT.

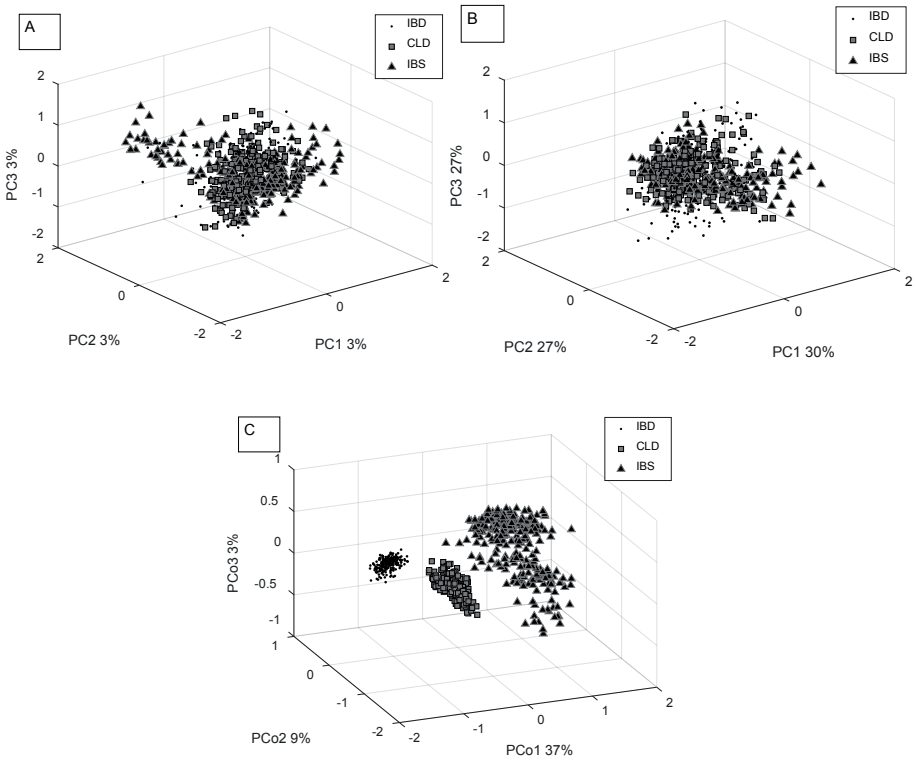


FIGURE 3S: FIGURE 3S: SCORES PLOTS OF THE THREE DATASETS OBTAINED BY (A) PCA, (B) R-PCA, AND (C) URF ON THE SVS AFTER IMPLEMENTING ZERO-CENTERING. THE PERCENTAGES INDICATE THE EXPLAINED INFORMATION BY EACH PC/PCO. THE DOTS REPRESENT IBD; THE SQUARES REPRESENT CLD; THE TRIANGLES REPRESENT IBS.

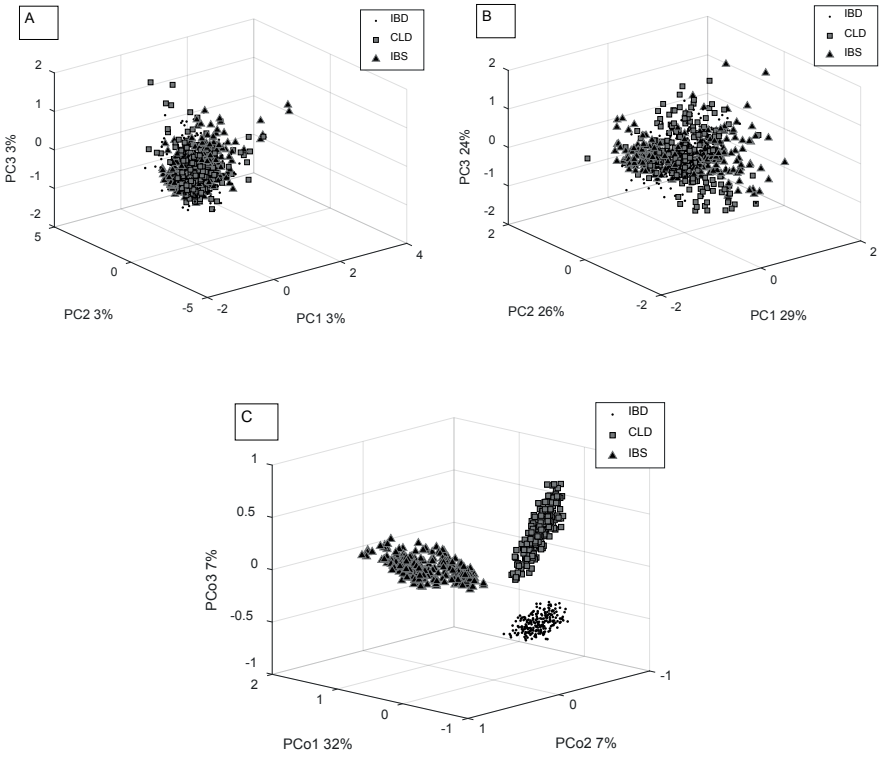


FIGURE 4S: FIGURE 4S: SCORES PLOTS OF THE THREE DATASETS OBTAINED BY (A) PCA, (B) R-PCA, AND (C) URF ON THE SVS AFTER IMPLEMENTING COMBAT. THE PERCENTAGES INDICATE THE EXPLAINED INFORMATION BY EACH PC/PCO. THE DOTS REPRESENT IBD; THE SQUARES REPRESENT CLD; THE TRIANGLES REPRESENT IBS.

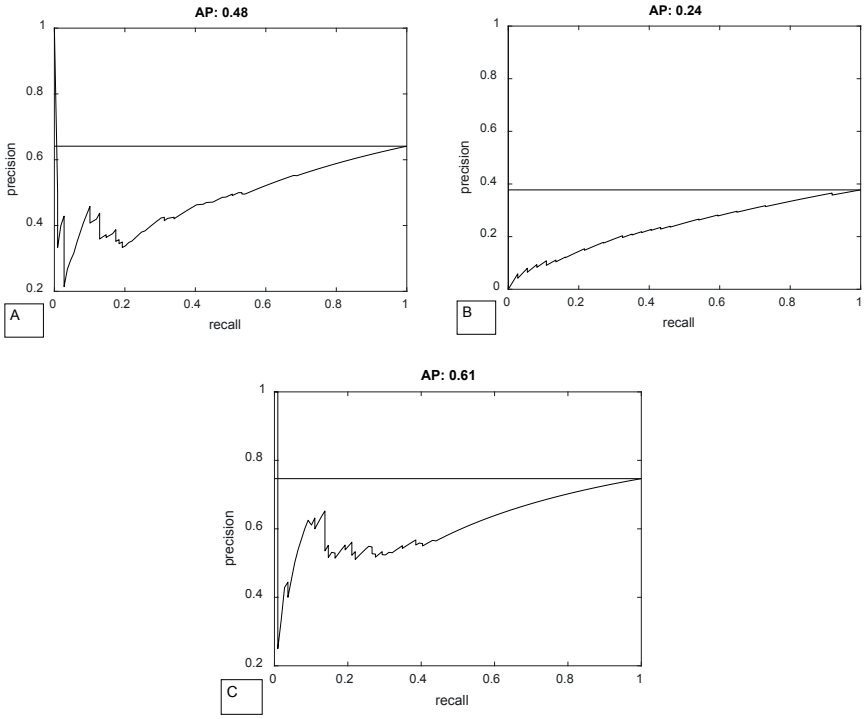


FIGURE 5S: PR CURVES OF THE VALIDATION SET GIVEN BY PLS-DA OF THE (A) IBD VS IBS MODEL WITH IBS BEING THE MAJORITY CLASS, (B) IBD VS CLD MODEL WITH IBD BEING THE MAJORITY CLASS, AND (C) CLD VS IBS MODEL WITH IBS BEING THE MAJORITY CLASS AFTER IMPLEMENTING ZERO-CENTERING. THE HORIZONTAL LINE IS THE BASELINE; THE AREA ABOVE THE BASELINE IS THE GOOD PERFORMANCE AREA, WHEREAS THE AREA BELOW THE BASELINE IS THE POOR PERFORMANCE AREA.

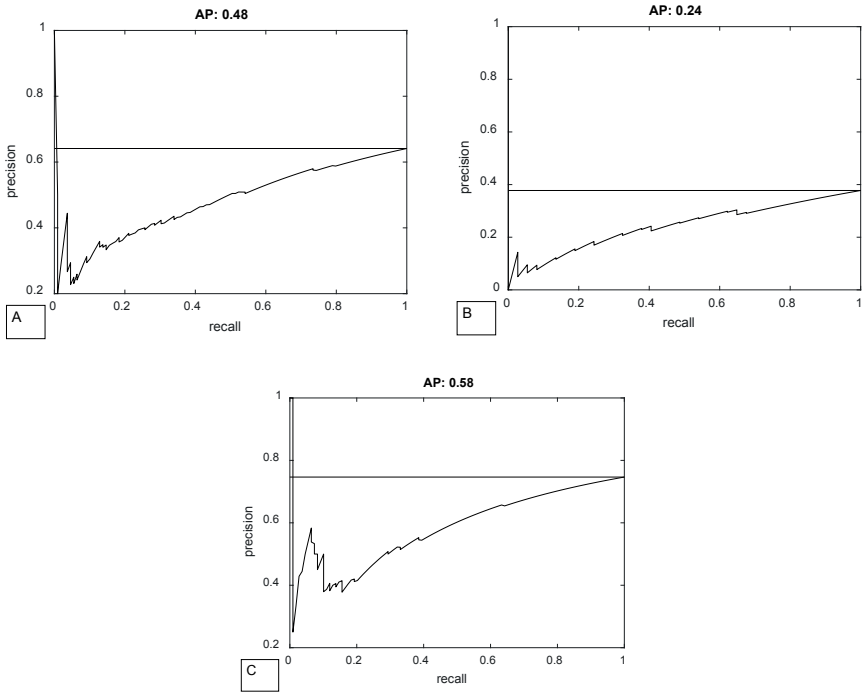


FIGURE 6S: PR CURVES OF THE VALIDATION SET GIVEN BY PLS-DA OF THE (A) IBD VS IBS MODEL WITH IBS BEING THE MAJORITY CLASS, (B) IBD VS CLD MODEL WITH IBD BEING THE MAJORITY CLASS, AND (C) CLD VS IBS MODEL WITH IBS BEING THE MAJORITY CLASS AFTER IMPLEMENTING COMBAT. THE HORIZONTAL LINE IS THE BASELINE; THE AREA ABOVE THE BASELINE IS THE GOOD PERFORMANCE AREA, WHEREAS THE AREA BELOW THE BASELINE IS THE POOR PERFORMANCE AREA.

MatLab codes for Combat and zero-centering:

```
%% combat

% mod matrix
mod = [];

[gamma_star,delta_star,nnn] = combat(X_toBeUsed',classes,mod);
X_toBeUsed_corr = nnn';
X_toBeUsed_corr_final = pareto(X_toBeUsed_corr);

%% zero-centering

[m, ~] = size(X_toBeUsed(IBD_idx,:));
MC = mean(X_toBeUsed(IBD_idx,:));
MCX = X_toBeUsed(IBD_idx,:) - ones(m,1)*MC;

[n, ~] = size(X_toBeUsed(liver_idx,:));
MC2 = mean(X_toBeUsed(liver_idx,:));
MCX2 = X_toBeUsed(liver_idx,:) - ones(n,1)*MC2;

[k, ~] = size(X_toBeUsed(IBS_idx,:));
MC3 = mean(X_toBeUsed(IBS_idx,:));
MCX3 = X_toBeUsed(IBS_idx,:) - ones(k,1)*MC3;

X_toBeUsed_corr = [MCX; MCX2; MCX3];
X_toBeUsed_corr_final = pareto(X_toBeUsed_corr);
```

The combat function can be downloaded from here:
https://github.com/Jfortin1/ComBatHarmonization/tree/master/Matlab

R code for ANCOVA:

```
X_toBeUsed <- read.csv("~/R/X_toBeUsed", header=FALSE)
batch.idx <- matrix(0, nrow = 663, ncol = 1)
batch.idx[214:402] <- 1
batch.idx[403:nrow(batch.idx)] <- 2
seq.idx <- sample(1:663, 663, replace = FALSE, NULL)
ref.idx <- seq.idx

results <- matrix(, nrow = 663, ncol = 200)
```

```
for (i in 1:200){
  results[1:663,i] <- doBC(X_toBeUsed[1:nrow(X_toBeUsed),i], ref.idx, batch.idx, seq.
idx,
                           result = c("correctedX"), method = c("lm"), correctionFormula =
formula("X ~ S * B"),
                       minBsamp = NULL, imputeVal = NULL)
}
results
```

The ANCOVA function can be downloaded from here:
https://github.com/rwehrens/BatchCorrMetabolomics

Visualisation of the three individual datasets without any correction and with combat correction:

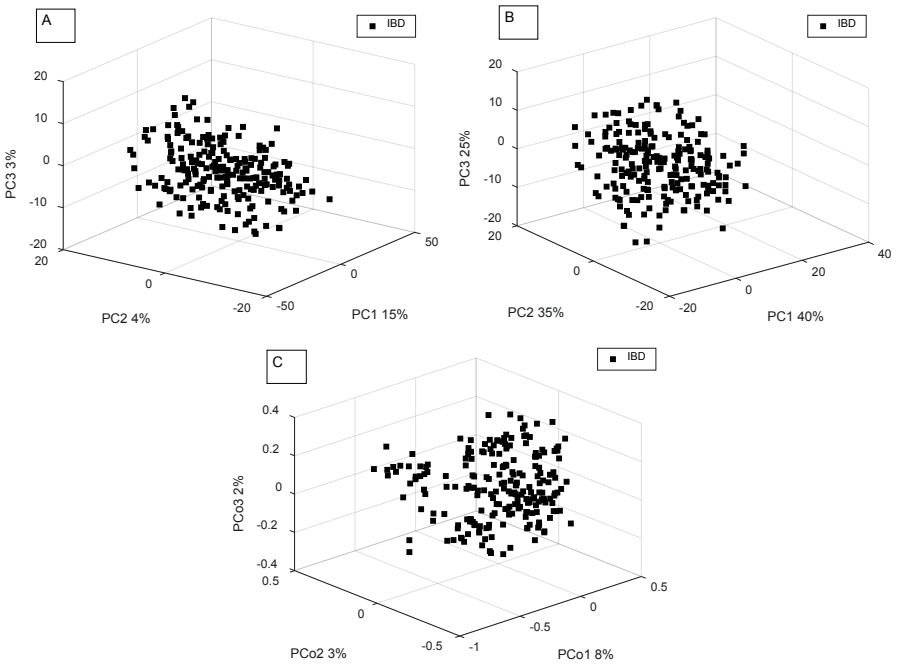


FIGURE 7S: VISUALISATION OF THE IBD DATASET VIA (A) PCA, (B) R-PCA, AND (C) URF WITHOUT ANY CORRECTION ATTEMPT.

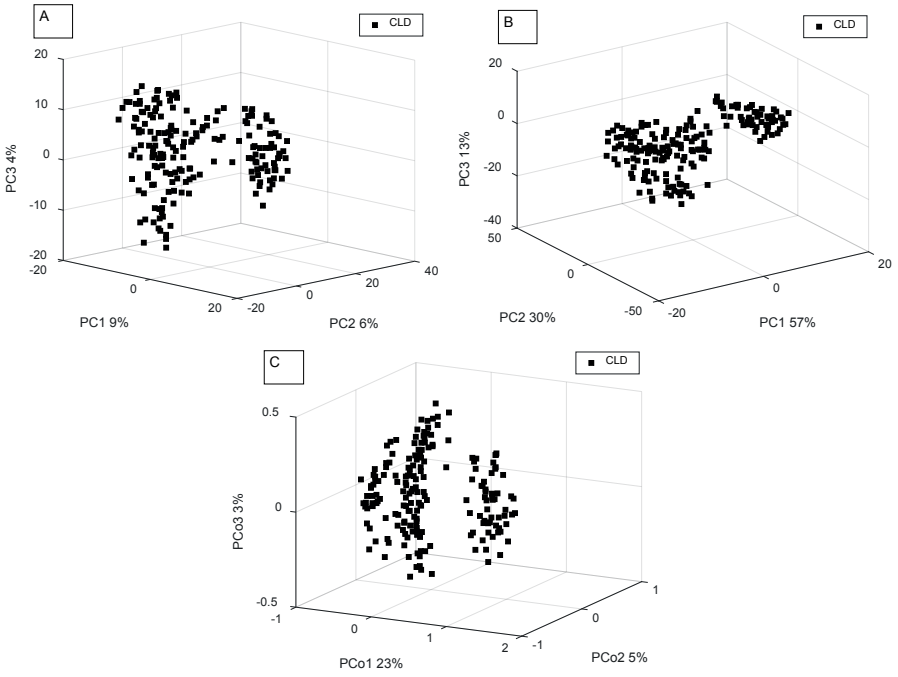


FIGURE 8S: VISUALISATION OF THE CLD DATASET VIA (A) PCA, (B) R-PCA, AND (C) URF WITHOUT ANY CORRECTION ATTEMPT.

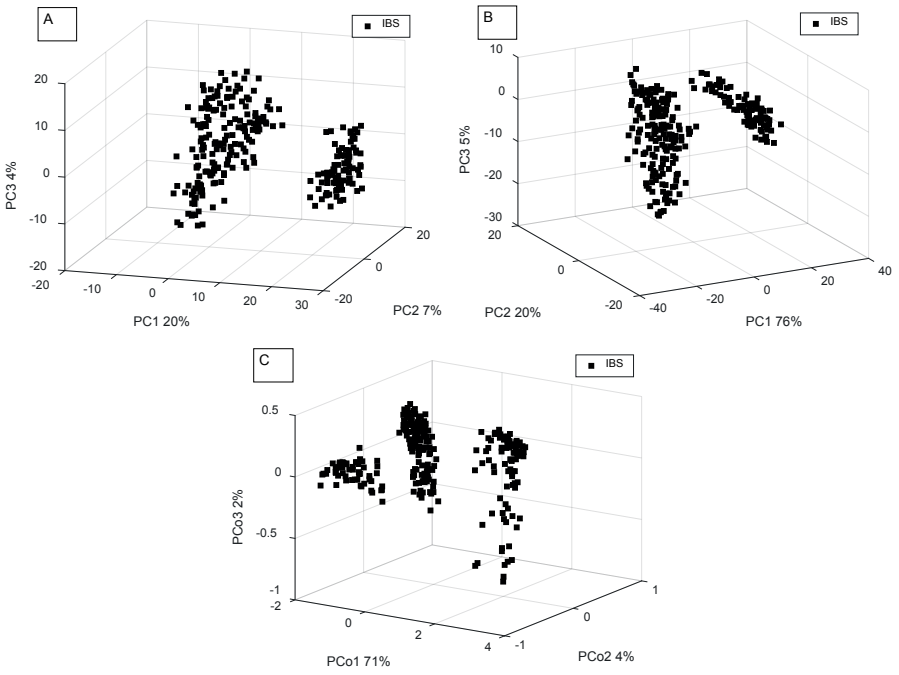


FIGURE 9S: VISUALISATION OF THE IBS DATASET VIA (A) PCA, (B) R-PCA, AND (C) URF WITHOUT ANY CORRECTION ATTEMPT.

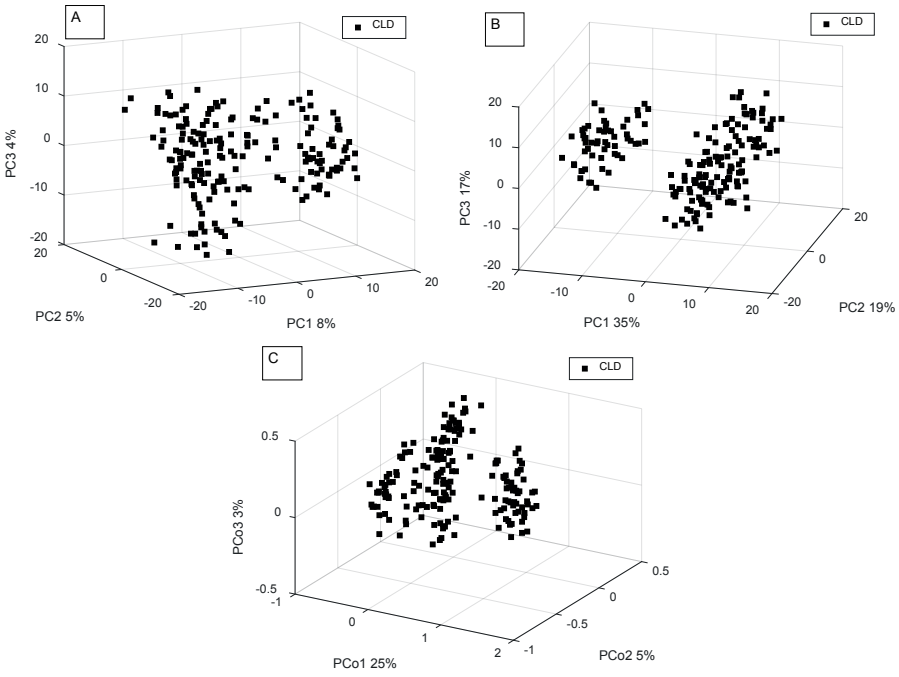


FIGURE 10S: VISUALISATION OF THE CLD DATASET VIA (A) PCA, (B) R-PCA, AND (C) URF AFTER IMPLEMENTING ZERO-CENTERING.

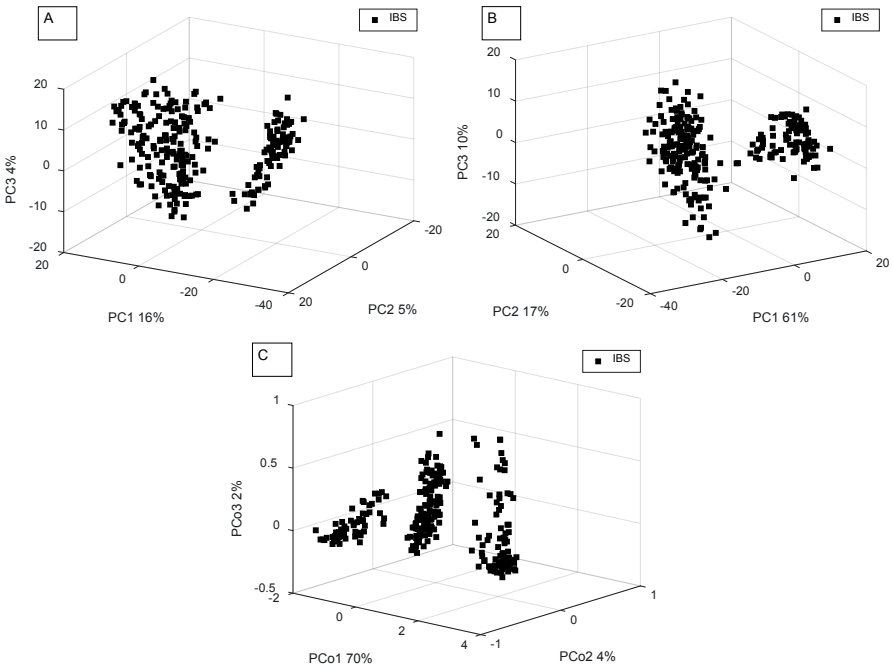


FIGURE 11S: VISUALISATION OF THE IBS DATASET VIA (A) PCA, (B) R-PCA, AND (C) URF AFTER IMPLEMENTING ZERO-CENTERING.

CHAPTER 5

Random forest and ensemble methods

George Stavropoulos, Robert van Voorstenbosch, Frederik-Jan van Schooten, Agnieszka Smolinska

Comprehensive Chemometrics (Second Edition), Chemical and Biochemical Data Analysis 2020, Pages 661-672,
doi: https://doi.org/10.1016/B978-0-12-409547-2.14589-5

Abstract

Recent expansions of technology led to growth and availability of different types of data. This, thus gave various opportunities for the machine learning, data mining, chemometrics and data science fields. Both fields have been consequently developing new approaches and algorithms in a wide range of applications in biomedical, medical, -omics but also from daily-life to national security areas. Ensemble techniques become the backbone of the machine learning field. The phrase refers to an approach in which multiple, independent, aka uncorrelated, predictive models are combined. Those multiple models can be combined for instance by simple averaging or voting. The advantage of ensemble techniques is their ability to yield very high performance model. The use of ensemble techniques is present in our daily lives. We tend to ask or check the opinion of several specialists before making the final decision for instance before purchasing an item or before hiring a new employee we search for judgment of several referees. In this book chapter, the theoretical and practical demonstration of three ensembles techniques, adaptive boosting, random forest and gradient boosting are shown. Each technique is discussed from its theoretical perspective followed by presentation of pro and cons of each method. The last part of the chapter is focused on the comparison between the techniques using two simulated data sets.

Introduction

Ensemble techniques have gained a lot of attention in machine learning the past decade [56, 64, 65, 113-115]. When predicting a target value in any machine-learning domain, the main causes of the difference between the actual and predicted values are variance, bias, and noise [116]. Except for noise, which is an irreducible error, ensemble techniques help reduce bias and variance. An ensemble of classifiers (i.e. strong learner) consists of a collection of classifiers (i.e. weak learners), and the principle behind an ensemble is that if many classifiers try to predict the same target, they will perform better than any single classifier alone [117]. A weak learner is an algorithm with predicting probability of error slightly better than random guessing. A strong learner is an algorithm which, given enough training data, can yield classifiers with arbitrarily small error probability. The concepts of weak and strong learners derive from the theory of probably approximately correct (PAC) learning, and they can be found elsewhere [118]. By correctly combining weak learners, the strong learner tends to be more flexible (i.e. less bias), and less data-sensitive (i.e. less variance) [117]. The ensemble idea originates back in the late seventies when two linear models were combined. The first linear model was fitted to the original data, and the second was fitted to the residuals of the first model [117]. However, it was only until the late nineties that the ensemble idea was revolutionised. Hansen et al. [119] proposed an ensemble of neural networks to achieve a better prediction accuracy than a single neural network. At the same time, Schapire [120] proved the strength of weak learnability, and as such, paved the way for the development of the adaptive boosting (AdaBoost) algorithm, the first strong classifier in the PAC sense [47]. AdaBoost led ideas such as bootstrapping [121] and stacking (or stacked generalisation) [122] to advance in the ensemble techniques domain as well, and thus, organising the ensemble techniques, as known up-to-date, into three main categories: boosting [47], bagging (or aggregative bootstrapping) [123], and stacking [122].

A strong classifier requires a proper selection of which weak classifiers (i.e. decision trees [124], regressors [125], neural networks [126], support vector machines [127]) will be used, and a proper way of how these weak classifiers will be combined (i.e. sequential or parallel). Weak classifiers can be either homogeneous or heterogeneous, and therefore, they result in homogeneous or heterogeneous ensembles since the way of which the weak classifiers are combined has to be, preferably, coherent. Boosting and bagging consider homogeneous weak classifiers (e.g. decision trees, regressors), whereas stacking considers, mostly, heterogeneous weak classifiers (e.g. different ensemble models). Furthermore, stacking may also be seen as a means of building a meta-model to improve prediction accuracy of various weak models predictions. The present chapter focuses on boosting and bagging ensemble models; therefore, readers interested in stacking are referred to [122, 128-130].

Boosting and bagging are fundamentally different even though they share the idea of combining homogeneous weak classifiers to create a strong classifier, and they both can be used for classification and regression purposes. On one hand, boosting is a sequential ensemble approach because the weak classifiers are built one after the other. The mistakes one classifier makes influences the way the next classifier is built; thus, making all classifiers dependent on each other [47]. The key point of every boosting ensemble is that each new classifier is built on new subset containing the components that were misclassified by the preceding models. [131]. As a result, boosting ensembles are meant to decrease bias. On the other hand, bagging is a parallel ensemble approach because the weak classifiers are built independently from each other. The key point of every bagging ensemble is that each classifier is built on a subset of training samples sampled with replacement [131]. Consequently, bagging ensembles are meant to decrease variance. Boosting and bagging ensembles also differ on what is known as training loss (or objective function). For example, some boosting ensembles try to minimise particular loss functions, such as $L1_{norm}$, $L2_{norm}$, Huber loss, logarithmic loss [52, 132], and assign weights to misclassified training samples [47], whereas some bagging ensembles try to either minimise the binary cross-entropy [133] or use the Gini impurity index [45]. Boosting ensembles also differ on the criterion for updating the weights of their training samples. Furthermore, another important aspect that makes boosting and bagging differ is the way they get to their conclusion of the sample of interest (i.e. the sample that needs to be predicted). Boosting ensembles use weighted sums, whereas bagging ensembles use either majority voting or averaging [117]. All this plurality of options in different aspects of boosting and bagging ensembles has led to the dominance of these techniques in the machine-learning domain. However, this dominance is owed, primarily, to AdaBoost [47], random forest [45], and gradient boosting [52]. AdaBoost represents the boosting ensembles, while random forest represents the bagging ensembles. As far as gradient boosting is concerned, it may as well be seen as a hybrid technique since it exploits aspects of both types of ensembles even though, in principle, it belongs to the boosting category. The chapter provides a brief encounter and summary on the ensemble techniques by discussing the basics and putting them in the perspective of their pro and cons.

Ensembles techniques

Adaptive boosting (AdaBoost)

AdaBoost is the most applied sequential ensemble [134-137]; it has found application in various scientific domains, from weather forecast to ailments prediction and monitoring, or even to daily human activities behaviour. AdaBoost, mostly, uses decision trees as weak classifiers, and in particular, it uses stumps (i.e. decision trees with only one split node); however, AdaBoost implementations that use regressors or pruned decision trees have also been documented [138-140]. AdaBoost is the only boosting algorithm that tries to minimise the exponential loss function, and at every iteration, it assigns weights to every training sample. In particular, misclassified samples are higher weighted in order the next classifier (i.e. stump) to focus more on correctly classifying the misclassified ones. Assigning weights changes the training sample distribution, thus, forcing the algorithm to emphasize more on the misclassified samples. The AdaBoost procedure can be seen in Figure 1. Initially, all training samples are given the same weight, and the first classifier is fitted to the training data by minimising the loss function.

Input:

Training samples $\{x_n, t_n\}$, $n = 1, 2, \ldots, N$, and responses $t_n \in \{-1, 1\}$

WeakLearn: learning procedure that produces classifier $y_m(x)$

Initial sample weights: $w_n^m(x) = 1/N$

Do:

For $m = 1{:}\,M$

1. $y_m(x) = WeakLearn(\{x\}, t, w)$, by minimising the cost function $J_m = \sum_{n=1}^{N} w_n^m [y_m(x_n) \neq t_n]$
2. $\varepsilon_m = \sum_{n=1}^{N} w_n^m [y_m(x_n) \neq t_n] / \sum_{n=1}^{N} w_n^m$
3. $\alpha_m = \log(1 - \varepsilon_m / \varepsilon_m)$
4. ${w_n}^{m+1} = {w_n}^m \times \exp\{\alpha_m [y_m(x_n) \neq t_n]\}$

End

FIGURE 7 THE ADAPTIVE BOOSTING ALGORITHM (ADABOOST).

In the consequent step, the classification error ε is calculated, and the classifier coefficient α is evaluated. The classifier coefficient is an evaluation metric that shows how much this weak classifier should be taken into account in the ensemble model; the higher the α, the more this weak classifier contributes to the outcome. Consequently, the ensemble is updated by adding this new classifier multiplied by its updated coefficient. The final step of the loop includes the computation of the new observation weights that express which samples the next classifier should focus on. For wrongly classified samples, their weights increase, while for correctly classified samples their weights decrease. At the end of the M iterations, the ensemble $Y_M(x)$ consists of M, sequentially built, weak classifiers (called here stump), which are then aggregated into a linear combination weighted by the coefficients α. A graphical illustration of AdaBoost is shown in Figure 2, where each stump (representing a weak classifier) is built in a sequential manner till M different classifiers are obtained.

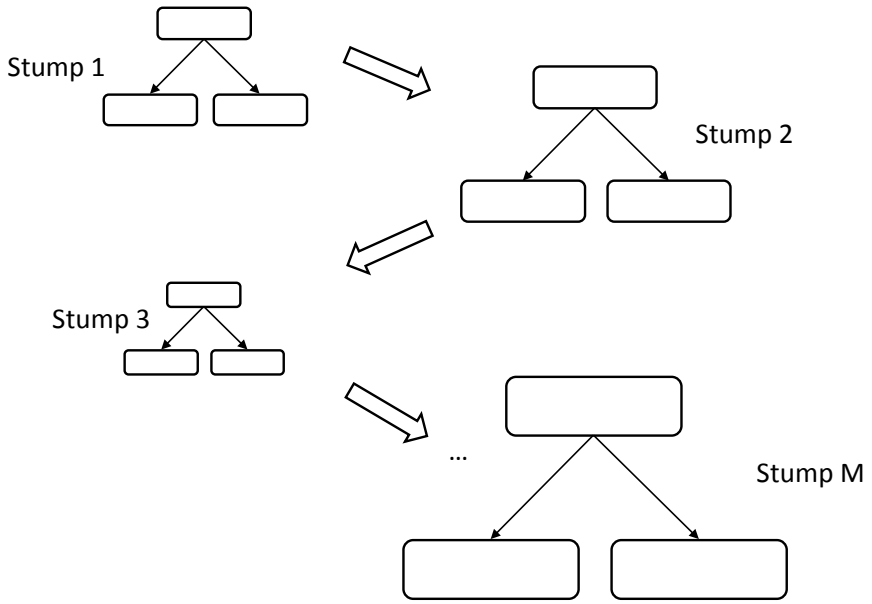


FIGURE 2 THE SEQUENTIAL BUILDING OF STUMPS IN ADABOOST. THE BIGGER THE SIZE OF THE STUMP, THE MORE IMPORTANT THIS STUMP IS IN THE ENSEMBLE MODEL.

Random forest

The most successful representative of bagging ensembles is random forest since its conception [45] with numerous implementations on every kind of scientific domain [37, 141-145]. Random forest uses fully-grown decision trees (i.e. trees that in their terminal/leaf nodes contain only one class); although, random forests with pruned

decision trees have also been documented [146, 147]. In random forest, every tree is built independently and on a different subset of training data (i.e. bootstrapped training datasets), and it uses a random subset of variables to be split in its nodes too. This means that every tree is built on different observations, and seeing different variables of the chosen observations; therefore, different information is seen by each tree in an attempt to obtain uncorrelated trees. Notably, from the total training observations, these that were not included in the bootstrapped training datasets are used to evaluate the performance of the forest, and they are called the out-of-bag observations. The lower the out-of-bag error is (i.e. misclassification of the out-of-bag samples), the better the performance of the forest is. In classification problems, random forest uses the Gini impurity index to select the optimal variable (from a randomly made subset of variables) to split at every node. In regression problems, it uses the mean squared error (MSE). The random forest procedure is described, briefly, in Figure 3. Initially, a bootstrapped training dataset is made, and for every variable randomly selected to be examined at the root node, the Gini index or the MSE is calculated. Then, the root node is split into two child nodes (also called internal nodes). The splitting continues until there are only terminal child nodes in the tree. At the end of the M iterations, the ensemble $Y_m(x)$ consists of M, built in parallel, weak classifiers (i.e. trees). Noteworthy, depending on the problem at hand (i.e. classification or regression), the way the outcome is found varies.

Input:

Training samples $\{x_n, t_n\}$, $n = 1, 2, \ldots, N$, and responses $t_n \in \{-1, 1\}$

WeakLearn: learning procedure that produces decision tree $y_m(x)$

Training subset size $\mu < n$

Number of variables randomly selected at every node $P < t$

Do:

For $m = 1{:}\,M$

1. Draw a bootstrap sample $\{x_\mu, t_P\}$ from training samples $\{x_n, t_n\}$

1.1 For $j = 1{:}\,P$

1.2 $y_m(x) = WeakLearn(\{x\}, t_j)$

1.3 Split the internal node into two child nodes

End

End

Final model:

$Y_M(x) = \frac{1}{N} \times \sum_{m=1}^{N} y_m(x)$ for regression, and $Y_M(x) = arg_x \max[card(m|y_m(x) = j)]$ for classification

FIGURE 3 THE RANDOM FOREST ALGORITHM.

In regression problems, the outcome is the average value over all tree results for the sample of interest. In classification problems, the outcome can be found by either taking the majority of the votes that the sample of interest received (this is also called hard-voting) or by considering the probabilities for both classes that the sample of interest received and average them (this is also called soft-voting). Finally, averages or majority votes can either be simple or weighted in case any relevant weights can be used [148, 149]. A graphical representation of a random forest is shown in Figure 4.

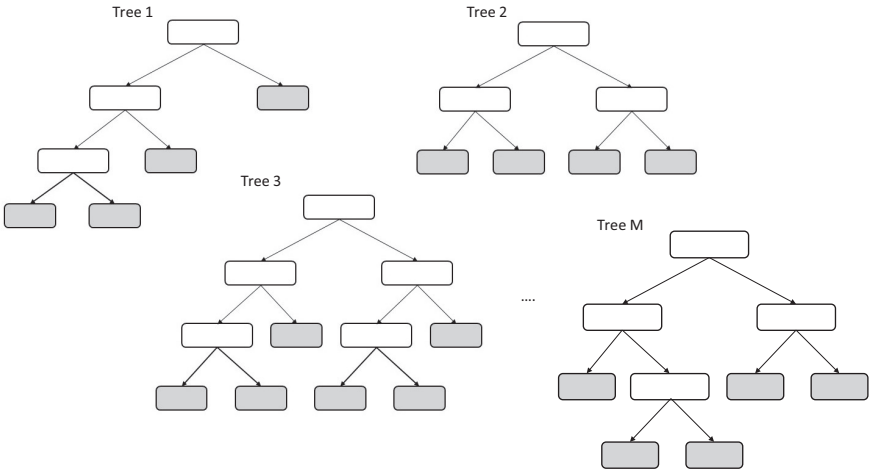


FIGURE 4 THE PARALLEL BUILDING OF FULLY-GROWN TREES IN RANDOM FOREST. THE GREY BLOCKS INDICATE TERMINAL/LEAF NODES.

Gradient boosting

Gradient boosting is the third on the line of succession, and as the two algorithms mentioned above (i.e. AdaBoost and random forest), it has found application in every kind of scientific domain [113, 150-153]. Gradient boosting is also known as gradient boosting trees or gradient boosted trees or even boosted trees since it makes use of decision trees as its weak classifiers, which are constructed in a greedy manner [52]. In opposition to AdaBoost and random forest that use stumps and fully-grown trees, respectively, gradient boosting makes use of trees of four to eight levels (i.e. splits). Gradient boosting can be called a generic algorithm since it is agnostic of the type of loss function. Rather than trying to minimise a single loss function (e.g. AdaBoost minimises the exponential loss function, and random forest, in regression cases, minimises the MSE), gradient boosting can use and therefore, minimise any differentiable loss function. Many standard loss functions (e.g. $L1_{norm}$, $L2_{norm}$, Huber loss, cross entropy loss) can be used, but the user can define their loss function too. The only requirement is that it must be a convex function (i.e. bowl-shaped) [154],

which is due to the way gradient boosting minimises the loss functions. Gradient boosting uses partial derivatives (i.e. gradients), and in particular, it uses the steepest gradient descent optimisation method [155]. This is why gradient boosting is called "gradient". More importantly, gradient boosting calculates the gradients of the loss function at every iteration with respect to the predictions of the current model instead of the variables, which is the case in AdaBoost and random forest. In mathematical terms, it can be written: $F_m(x) = F_{m-1}(x) + h_m(x)$, where $F_m(x)$ is an iterative boost of the ensemble, $F_{m-1}(x)$ is the previous iteration of the ensemble, and $h_m(x)$ is a decision tree trained on the residuals (also known as pseudo-residuals) of the $F_{m-1}(x)$ (i.e. $h_m(x) = y - F_{m-1}(x)$). One may also state that gradient boosting, repetitively, takes advantage of patterns in residuals, and tries to strengthen a model with weak predictions. The modelling of the residuals stops when a stage, where residuals do not have any patterns to be modelled, has been reached. The gradient boosting procedure is illustrated in Figure 5. Initially, for a given loss function, the pseudo-residuals are set equal to the observation values. In the first iteration, the best possible weak classifier is fitted to the pseudo-residuals.

Input:

Training samples $\{x_n, t_n\}$, $n = 1, 2, \ldots, N$, and responses $t_n \in \{-1, 1\}$

WeakLearn: learning procedure that produces decision tree $h_m(x)$

ρ_m : step or "boost" the steepest descent takes

$F_0(x) = argmin_\rho \sum_{i=1}^{N} L(y_i, \rho)$

Do:

For $m = 1: M$

1. $\tilde{y}_i = -[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}]_{F(x)=F_{m-1}(x)}$, $i = 1, \ldots, N$
2. $a_m = argmin_{a,\beta} \sum_{i=1}^{N} [\tilde{y}_i - \beta \times WeakLearn(x_i|; a)]^2$
3. $\rho_m = argmin_\rho \sum_{i=1}^{N} L(y_i, F_{m-1}(x_i) + \rho \times WeakLearn(x_i; a_m))$
4. $F_m(x) = F_{m-1}(x) + \rho_m \times WeakLearn(x; a_m)$

End

Final model:

$F(x; \{\beta_m, a_m\}_1^M) = \sum_{m=1}^{M} \beta_m \times WeakLearn(x;\ a_m)$

FIGURE 5 THE GRADIENT BOOSTING ALGORITHM.

The value of the optimal step size that defines by how much the ensemble model needs to be updated is calculated, and then, the ensemble is updated by adding the new weak classifier multiplied by the step size. Then again, new pseudo-residuals are calculated, and the whole process is repeated M times. At the end of the M iterations,

the ensemble $F_m(x)$ consists of M, sequentially built, weak classifiers. Furthermore, it is to be mentioned that a lot of research and development has been done on the gradient boosting algorithm since its conception, and therefore, gradient boosting can be divided into the following algorithms: gradient boosting [52], extreme gradient boosting (XGBoost) [132], light gradient boosting (LightGBM) [156], and CatBoost [157]. XGBoost, LightGBM, and CatBoost show modifications that, in their way, help get better, faster and generalizable results. A graphical representation of gradient boosting is shown in Figure 6.

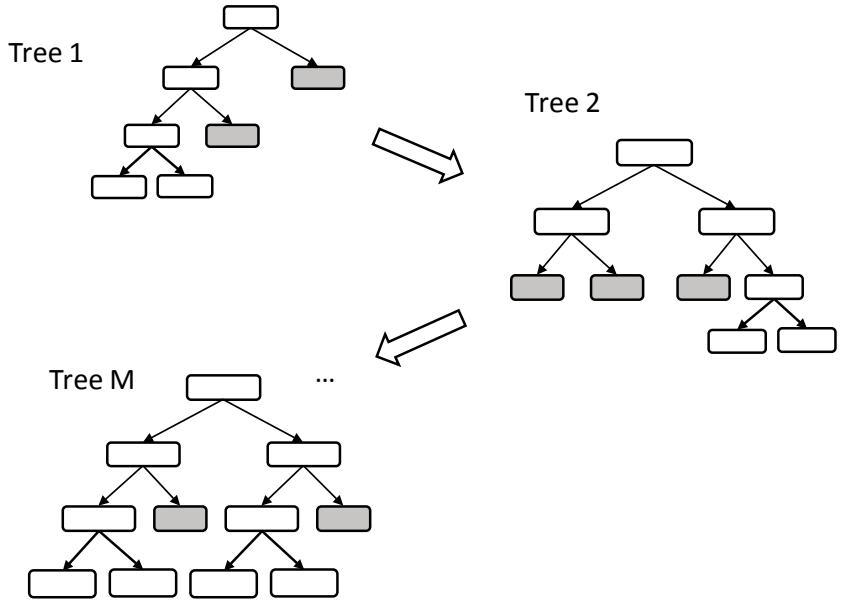


FIGURE 6 THE SEQUENTIAL BUILDING OF PRUNED TREES IN GRADIENT BOOSTING. EACH TREE CONSISTS OF FOUR TO EIGHT SPLIT. THE BIGGER THE SIZE (NOT THE DEPTH) OF THE PRUNED TREE, THE MORE THIS TREE WILL BE TAKEN INTO ACCOUNT IN THE ENSEMBLE. THE GREY BLOCKS INDICATE TERMINAL/ LEAF NODES. NOTE: FOR VISUALISATION COHERENCE, THREE-LEVEL PRUNED TREES ARE SHOWN.

Comparison of ensembles techniques

AdaBoost, random forest and gradient boosting have been successfully applied in almost every scientific field, and consequently, there is not a clear guideline for saying which ensemble outperforms the others. The choice of the ensemble one should use for their analysis, purely, depends on the problem at hand and possible analysis characteristics (e.g. analysis time or type of data). The two major advantages of AdaBoost are its complexity and its parameters tuning. AdaBoost only requires the number of stumps to be tuned and since stumps are the least complex classifiers,

it shows the lowest model complexity of all three ensembles [47]. The major disadvantage of AdaBoost is that it is prone to overfitting. The more stumps are added to the ensemble, the higher the chance of overfitting exist. Nevertheless, this can be controlled via proper optimization and validation of the model by for instance cross-validation. At the same time, AdaBoost builds its classifiers sequentially, and it also examines all variables to split at every node [47]; AdaBoost per definition is computationally not extensive, since stumps are very small trees. However, since it always use all variables to find the best split, for high-dimensional data with many variables it may be time-consuming.

Random forest presents two major advantages over AdaBoost and gradient boosting: it does not overfit, and it provides a proximity matrix of the training samples [45]. Random forest grows full trees, which inevitably will overfit, although, every tree overfits a different part of the training data and in a different way. In the end, via majority voting or averaging, this overfitting is cancelled out. In random forest, the more trees that are added, the merrier, whereas in AdaBoost and gradient boosting, the more stumps/trees are added, the higher the chances to overfit. This will increase the running time of the algorithm; nonetheless, it is not seen as a drawback because random forest builds its classifiers in parallel, which by definition makes random forest faster than AdaBoost and gradient boosting. As far as the proximity matrix is concerned, the random forest algorithm calculates a proximity matrix of the training samples once the forest is built. The term proximity means “closeness” or “nearness” between pairs [158]. In general, proximity matrices are square distance matrices that show how similar or dissimilar the data are. Proximity matrices are calculated by using traditional distance measures such as Euclidean distance; in random forest, however, the proximity matrix is not calculated by using a distance measure [45]. More specifically, for every pair of samples, the proximity indicates the percentage of the times these two samples ended up in the same terminal node. For instance, if the random forest consists of 1000 trees and the pair of samples ended up in the same terminal node in 100 of the 1000 trees, then the proximity for this pair of samples is 100/1000 = 0.1. A toy example of how the random forest proximity is obtained is shown in Figure 7.

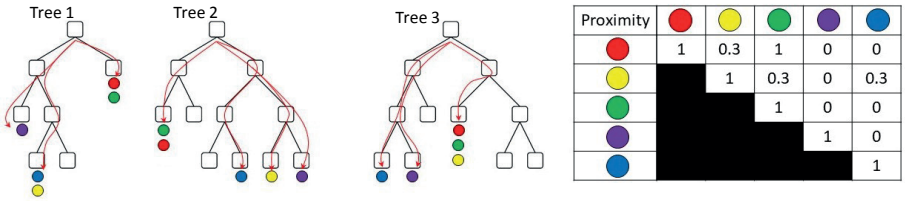


FIGURE 7 IN A RANDOM FOREST CONSISTING OF THREE TREES, ALL TRAINING SAMPLES ARE RUN THROUGH ALL THE TREES ONCE THE FOREST IS BUILT. THEN, IT IS CHECKED HOW MANY TIMES EVERY POSSIBLE PAIR OF SAMPLES ENDED UP IN THE SAME TERMINAL NODE, AND IN THE END, THIS NUMBER IS DIVIDED BY THE TOTAL NUMBER OF TREES IN THE FOREST TO GET THE PROXIMITIES.

Consequently, the higher the proximity, the more similar the samples are. This kind of proximity matrix is particularly useful when datasets of non-numeric variables are used, for instance. Conventional distance measures (e.g. Euclidean distance) are designed for numeric variables only; therefore, finding a distance measure for other kinds of data (e.g. categorical, ranks) may be challenging. Furthermore, proximities can also be used for identifying outliers, replacing missing values or visualising the data [45]. The major disadvantage of random forest is its complexity. Random forest is a rather complex ensemble since it grows full trees, which also means that it is computationally expensive. At the same time, the prediction process using random forest is more time-consuming than AdaBoost and gradient boosting.

The major advantage of gradient boosting over AdaBoost and random forest is its immense flexibility. Every differentiable and convex loss function can be used; moreover, the user can define their loss function too. In theory, this means that gradient boosting can solve almost every classification or regression problem, something that is proved by the big success of gradient boosting in Kaggle competitions. Additionally, the fact that the user can tune several hyper-parameters (e.g. number of observation per node, tree depth, learning rate), it is considered more as an advantage rather than as a drawback. Constraining the trees in all these different ways ensures that the trees remain weak; therefore, the model complexity does not increase too much as in random forest. When it comes to gradient boosting disadvantages, the biggest one is that it is prone to overfitting. Although, this can be controlled via cross-validation. Gradient boosting is also time-consuming because it builds the trees sequentially and it requires a large grid search during tuning. Table 1 shows an overview of the pros and cons of the three ensembles.

TABLE 5 OVERVIEW OF THE PROS AND CONS OF THE THREE ENSEMBLES.

	AdaBoost	Random forest	Gradient boosting
Affected by noisy data	+	-	+
Affected by outliers	+	-	+
Analysis time	Classifiers are built sequentially	Classifiers are built in parallel	Classifiers are built sequentially
Can be applied to multi-class problems	+	+	+
Data preparation (e.g. scaling, transformation)	-	-	-
Deals with missing values	+	+	+
Loss function	Exponential loss	MSE for regression, Gini index for classification	Multiple loss functions
Parameters tuning	Number of stumps	Number of trees, and number variables to be examined at every split	Number of trees, depth of the trees, number of nodes or leaves or number of observations per split
Prone to overfitting	+	-	+
Proximities	-	+	-
Requires large datasets	+	-	+
Variables importance	+	+	+
Works with continuous and categorical variables	+	+	+

A common tool that has received a lot of attention in –omics related fields and industry, and it has been very successful in multivariate data analysis is partial least squares (PLS) analysis [42, 91]. PLS originates back at the beginning of the 20th century, several years before AdaBoost, random forest, and gradient boosting were developed. PLS was initially designed to deal with regression problems [42], but later on, a variant of PLS able to deal with classification problems was proposed too [91]; thus, making PLS suitable for both data analysis cases. A comparison of PLS with the three ensembles techniques is worth making due to its popularity even though PLS is not an ensemble technique. The profound difference between PLS and the three ensembles shown here is that original version of PLS considers linear relations between variables in the data only, whereas the ensembles consider both linear and nonlinear relations in the data [42, 45, 47, 52, 91]. Nowadays, most of the data that are generated are very complex, and therefore, they often show nonlinearities amongst them; thus, making the ensembles, most of the times, the first choice of use. PLS requires the data to be scaled, while the tree-based ensemble techniques, though, do not require any data preparation. Moreover, generally PLS can be applied to numerical

and continuous data, while the ensembles can be applied to any data. Nevertheless, PLS has a major advantage over the three ensembles, and this is that PLS is notably faster and less complex than the three ensembles since it creates a single model. It is also easier to interpret the model itself and its results compared to the ensembles since they, in general, are still regarded as black boxes. This is why the number of PLS-related publications keeps increasing [159-162]. To conclude, ensemble techniques are powerful, and they can deal with various problems with very little parameters tuning; however, the complexity they show and the understanding they require to from the user, they might not always be the first choice.

Practical demonstration of the ensemble techniques

PLS models remained extremely popular in various fields thanks to the simplicity of their interpretation [43, 163-165]. Ensemble techniques are not typically employed, although they are now increasingly common [166-171]. Indeed, PLS discriminant analysis (PLS-DA) is the most well-known and common tool to implement classification and regression in metabolomics as suggested by Gromski at al. [43]. The predominance of PLS-DA in data analysis is so extensive that researchers often forget about other techniques. To demonstrate the power and the advantages of ensemble techniques, some examples based on simulated data are presented here to provide the reader with a sense of when exactly the choice for ensemble techniques can be beneficial. For sake of simplicity, the focus is on classification tasks. Yet, the choice of when a specific algorithm is left to the readers own creativity, as this is always dependent on a multitude of factors, such as type of data, complexity of the problem, time available, education, experience and experimental design.

This part of the chapter focuses on a simplified demonstration of ensemble-based techniques taking into account their ability to deal will non-linear data, Followed by variable selection and the general performance of the three ensemble techniques.

Methods and simulated data

In order to demonstrate the performance of the different techniques, specific architectures were used. In random forest the training set was iteratively split into an internal validation and training set to find the important variables in a traditional validation procedure. The description of the different variable importance measures and their performance can be found elsewhere [172]. The variable selection is an embedded process in random forest and thus often seems a rigid and objective. However, on applying cross validation on variable selection, one obtains measures which are tuneable to be stricter on variable inclusion.

Variable importance can be expressed as the frequency at which the variable was chosen within AdaBoost, random forest or gradient boosting technique. It is possible to put a threshold α on this measure, to select the most important ones per iteration within a k-fold validation loop. Following over the numerous iterations (i.e. within k-fold validation loop), the frequency can be measured at which a variable was selected as important over the different models. The final number of variables can be selected by using β threshold, i.e. the frequency of variables being selected over various models. This translates to including those variables that were found to be important in at least β% of all models. Finally, the most important selected variables were used to build a model with the entire training set. The resulting model was tested on the external independent test set.

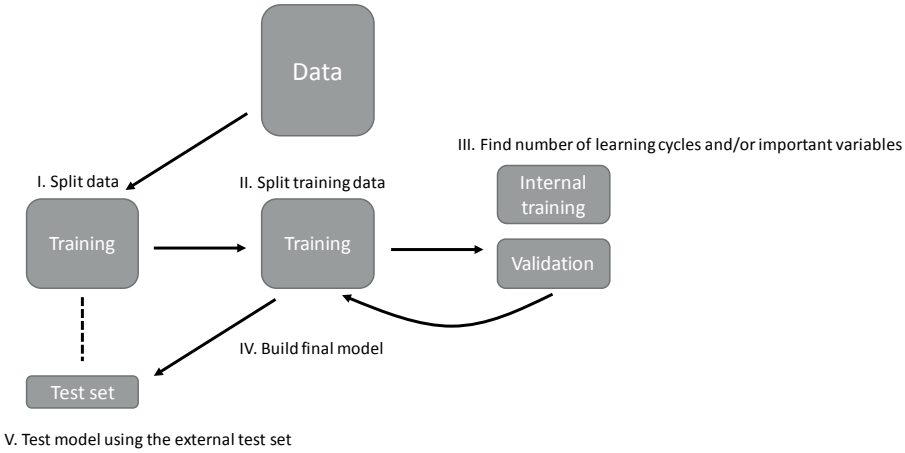


FIGURE 8 TRAINING AND TESTING ARCHITECTURE USED FOR THE RANDOM FOREST, ADABOOST AND GRADIENT BOOSTING. THE VARIABLE SELECTION WAS PERFORMED FOR EACH TECHNIQUE, WHILE THE SELECTION OF NUMBER OF LEARNING CYCLE WAS DONE IN ADABOOST AND GRADIENT BOOSTING.

In case of PLS-DA, a very similar set-up was used, where the internal training and validation set were applied to find the optimal number of variables and model complexity. Here Significance Multivariate Correlation were applied as variable selection and importance measures [173]. Next, using the most important variables a model was created using the entire training set and subsequently tested using the external independent test set.

The modelling approaches described for random forest, AdaBoost, gradient boosting and PLS-DA can be seen in in Figures 8-9, respectively. Note that the training and independent test sets were identical among the four different classification methods.

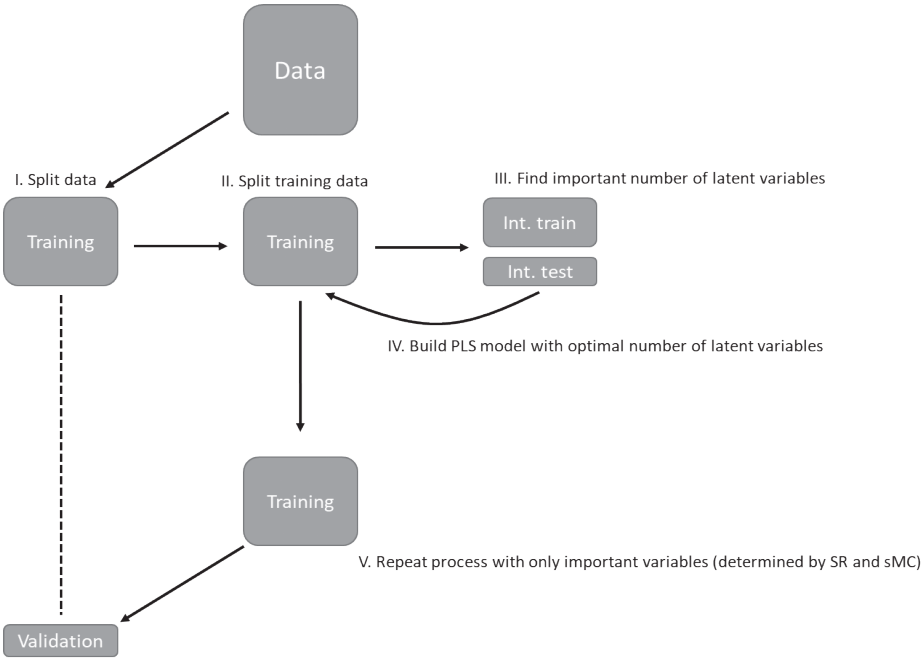


FIGURE 9 TRAINING AND TESTING THE PARTIAL LEAST SQUARES DISCRIMINANT ANALYSIS MODEL. HERE VARIABLE SELECTION WAS CALCULATED USING SELECTIVITY RATIO AND SIGNIFICANCE MULTIVARIATE CORRELATION ON AN ITERATIVE INTERNAL TRAINING AND TEST SET.

The performance of each classification model by means of random forest, AdaBoost, gradient boosting and PLS-DA was assessed using the Geometric mean (G-mean), which is calculated by taking the square of multiplication of sensitivity and specificity [174, 175].

The performance of the classification techniques has been tested using two sets of simulated data, using an approached presented by Wojciechowski et al. [176]. The first data set consisted of two groups, class 1 and class 2 both consisted of 300 samples and only five variables of which all were informative. This examples was a simplified type of data, since no uninformative variables were present. The aim of this example was to demonstrate that all 3 ensembles techniques shown here can deal with complex data exhibiting high non-linearities. The simulated data is shown in Figure 10, using the first three variables.

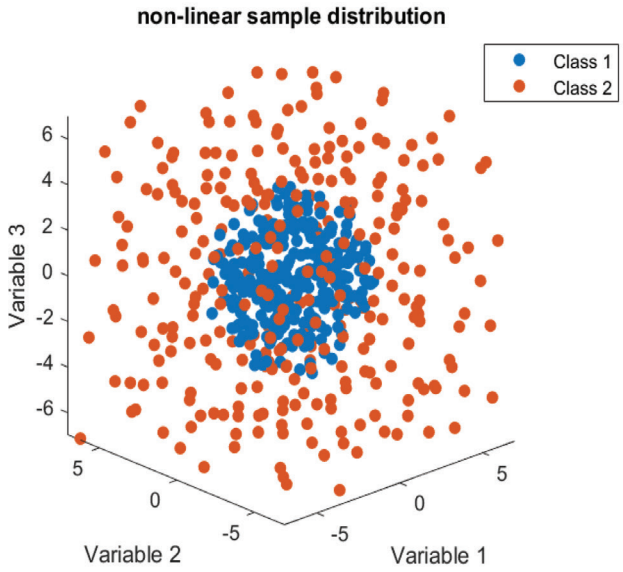


FIGURE 10 AN EXAMPLE OF SIMULATED DATA CONSISTING OF TOTAL OF FIVE VARIABLES. A 3-DIMENSIONAL SPHERE REPRESENTING A NON-LINEAR DISTRIBUTION OF TWO CLASSES IS SHOWN USING THE FIRST THREE VARIABLES.

This situation is obviously not realistic, as analytical challenges often involve far more than five variables, including informative and non-informative or redundant variables. Therefore, to mimic untargeted 'omics' datasets, 1149 uninformative random variables with different level of noise as normally distributed pseudorandom numbers were added to the set containing five informative variables. For both simulated data sets, 50 samples were randomly drawn to represent an independent test set for each group, leading to the training set consisting of 500 samples and the test set of 100 samples.

Based on the two simulated data the different ensemble based techniques and PLS-DA were examined. The performance of each optimized classification model was verified using an independent test set. The overview of the results by mean of G-means is shown in Table 2.

TABLE 2 THE PERFORMANCE FOR INDEPENDENT TEST SET FOR THE THREE ENSEMBLES TECHNIQUES AND PLS USING TWO SIMULATED DATA SETS EXPRESSED AS $Gmean = \sqrt{(Sensitivity\ x\ Specificity)}$.

Classification technique	Data set 1	Data set 2
	G-mean	
AdaBoost	94.0	65.52
Random forest	96.99	85.91
Gradient boosting	94.87	80.94
PLS-DA	50.91	48.54

The results presented in Table 2 clearly demonstrated that the ensemble techniques were able to cope with the non-linear structure of the data sets. The values of G-means were comparable between gradient boosting and random forest, while AdaBoost underperformed the two ensembles techniques. The performance of PLS was close to random classifier. When comparing which informative variables were selected in data set 2 by three ensembles technique, again random forest and gradient boosting selected all 5 informative variables, while AdaBoost missed one variable. Interestingly, for data set 2 each ensemble technique selected noisy variables. The number of uninformative variables included in the final model was depended on the threshold selected for the variable importance measure. Obviously, the stricter threshold led to exclusion majority of the uninformative variables but at the same time 1 informative variable was excluded as well. This might obscure or complicate biological interpretation of the selected variables later on, as the uninformative variables might cloud metabolic pathway analysis.

Discussion

The goal of the current chapter was to demonstrate the theoretical and practical application of ensembles techniques, more specifically, adaptive boosting, random forest and gradient boosting. Historically speaking, the first work on ensemble technique was shown in 1979 by Dasarathy and Sheela [177]. In their paper they suggested the use of an ensemble technique for partitioning the feature space using two or more classifier. The next occurrence of an ensemble technique has been shown by Hansen and Salamon [178], who demonstrated the variance reduction property of an ensemble system. However, it took some years before Schapire [120] introduced ensemble technique as a center of machine learning field. He has proven that strong classifier in probably approximately correct sense can be generated by combining weak classifiers through a procedure he called boosting, which has become the precursor of AdaBoost techniques. As the results the expansion of ensemble techniques in the literature has become very rapid [120, 122, 179, 180] with diverse strategies used for the classification and the way the individual models are combined.

The AdaBoost was for a long time as the very good example of the black box technique, which could be used by a practitioner without any need for parameters optimization. The idea behind AdaBoost was compared by its authors, Freund and Shapire, to group of friends betting on the horses going to race track. One of the person decided to develop a method of betting a part of his money taking into account his friend decision and adjusting the fractions based on the results. In that way, his performance over time reached the performance of most winning friend. Similarly with the boosting, the main goal is to improve the prediction performance. AdaBoost was the most commonly used version of the boosting, called by Leo Brieman in 1996 “the “best off-the-shelf classifier in the world” [181] and over the years it outperformed various classification techniques due its margins and boosting as the way of optimizing an exponential likelihood function.

Although, all those algorithms differ substantially the main idea behind them is to minimize variance and bias of the prediction model. With the introduction of Random Forests (RF) in 2001 by Leo Brieman [182] the popularity of ensemble techniques boosted even more. Although, RF is characterized with similar accuracy then AdaBoost, yet it is more robust to outliers and experimental error [183]. RF has become a method of choice for Kaggle competitions and nowadays it has exceeded the popularity of AdaBoost.

Till now various modification of AdaBoost [184-187] and RF exist [188-191]. Yet, they all share the same principle and reasons of the utilization of ensembles techniques [192, 193]. The most obvious reasons is that the performance of the best classifier might or might not be outperformed by averaging, yet it diminishes the danger of making a poor overall decision. Obviously, good performance on the training samples does not correspond to good generalization of the classification model on the test set. Therefore, combing the output of several classifier may reduce this risk. A second advantage is to make possible the analysis of large volume of data. Particularly, in some fields available data can be too large to be effectively used by one classifier, therefore portioning the data and using each of this part in a classification model and subsequently combining using intelligent rules, proves to be more efficient. Similarly, small number of data might jeopardize the robustness and application of the final classifier. Therefore, resampling techniques can be used to train different classifier. The last mentioned reason for usage of ensembles techniques is their capability of defining complex boundaries, which cannot often be found by a single classifier.

The techniques presented here exhibited the inherently good performance in predicting the classes in the independent test sets for both simulated data sets. Moreover, a strong asset of ensemble techniques is their embedded variable selection. The different mechanisms of each of ensemble techniques resulted in selection of different set of the informative variables. When the different prediction models

were tested on all 1154 variables, similar performances were obtained as the ones described above. The slight differences between the different methods can mainly be attributed to (1) diversity of the model due to the number of trees and/or learning cycles and (2) to the weight of outlying samples that might strongly influencing the model. Although performances were almost identical, the correctly selected important variables did differ among the different models. The simulations shown here exhibited typical non-linear properties. Therefore, it is not surprising that ensemble methods outperformed PLS-DA. Nevertheless, PLS-DA remains a method of choice when the linear combination of variables leads to the discrimination between groups of interest. If non-linarites are expected, PLS-DA has to be transferred to its nonlinear version by for instance a kernel mapping [56, 194].

References

1. Tian, D., et al. *An accurate eye pupil localization approach based on adaptive gradient boosting decision tree*. in *2016 Visual Communications and Image Processing (VCIP)*. 2016. IEEE.
2. Rubin, J., et al., *An ensemble boosting model for predicting transfer to the pediatric intensive care unit.* International journal of medical informatics, 2018. **112**: p. 15-20.
3. Baranska, A., et al., *Volatile organic compounds in breath as markers for irritable bowel syndrome: a metabolomic approach.* Aliment Pharmacol Ther, 2016. **44**(1): p. 45-56.
4. Baranska, A., et al., *Profile of volatile organic compounds in exhaled breath changes as a result of gluten-free diet.* J Breath Res, 2013. **7**(3): p. 037104.
5. Smolinska, A., et al., *The potential of volatile organic compounds for the detection of active disease in patients with ulcerative colitis.* Aliment Pharmacol Ther, 2017. **45**(9): p. 1244-1254.
6. Divina, F., et al., *Stacking ensemble learning for short-term electricity consumption forecasting.* Energies, 2018. **11**(4): p. 949.
7. Domingos, P. *A unified bias-variance decomposition*. in *Proceedings of 17th International Conference on Machine Learning*. 2000.
8. Rokach, L., *Ensemble-based classifiers.* Artificial Intelligence Review, 2010. **33**(1-2): p. 1-39.
9. Decatur, S. *Statistical queries and faulty PAC oracles*. in *Proceedings of the Sixth Workshop on Computational Learning Theory*. 1993. Citeseer.
10. Hansen, L.K. and P. Salamon, *Neural network ensembles.* IEEE Transactions on Pattern Analysis & Machine Intelligence, 1990(10): p. 993-1001.
11. Schapire, R.E., *The strength of weak learnability.* Machine learning, 1990. **5**(2): p. 197-227.
12. Freund, Y. and R.E. Schapire. *Experiments with a new boosting algorithm*. in *icml*. 1996. Citeseer.
13. Freedman, D.A., *Bootstrapping regression models.* The Annals of Statistics, 1981. **9**(6): p. 1218-1228.
14. Wolpert, D.H., *Stacked generalization.* Neural networks, 1992. **5**(2): p. 241-259.
15. Breiman, L., *Bagging predictors.* Machine learning, 1996. **24**(2): p. 123-140.
16. Rokach, L. and O. Maimon, *Decision trees*, in *Data mining and knowledge discovery handbook*. 2005, Springer. p. 165-192.
17. Amemiya, T., *Selection of regressors.* International economic review, 1980: p. 331-354.
18. Haykin, S., *Neural networks: a comprehensive foundation*. 1994: Prentice Hall PTR.
19. Scholkopf, B. and A.J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. 2001: MIT press.
20. Breiman, L., *Stacked regressions.* Machine learning, 1996. **24**(1): p. 49-64.
21. Ozay, M. and F.T.Y. Vural, *A new fuzzy stacked generalization technique and analysis of its performance.* arXiv preprint arXiv:1204.0171, 2012.
22. Smyth, P. and D. Wolpert, *Linearly combining density estimators via stacking.* Machine Learning, 1999. **36**(1-2): p. 59-83.
23. Basu, D., *On sampling with and without replacement.* Sankhyā: The Indian Journal of Statistics, 1958: p. 287-294.
24. Chen, T. and C. Guestrin. *Xgboost: A scalable tree boosting system*. in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016. ACM.
25. Friedman, J.H., *Greedy function approximation: a gradient boosting machine.* Annals of statistics, 2001: p. 1189-1232.
26. Kline, D.M. and V.L. Berardi, *Revisiting squared-error and cross-entropy functions for training neural network classifiers.* Neural Computing & Applications, 2005. **14**(4): p. 310-318.
27. Breiman, L., *Random Forest.* Machine Learning, 2001. **45**: p. 5-32.
28. Hassan, A.R. and M.I.H. Bhuiyan, *An automated method for sleep staging from EEG signals using normal inverse Gaussian parameters and adaptive boosting.* Neurocomputing, 2017. **219**: p. 76-87.

29. Qi, J., et al., *Light-driven transformable optical agent with adaptive functions for boosting cancer surgery outcomes.* Nature communications, 2018. **9**(1): p. 1848.
30. Herfeh, M.P., A. Shahbahrami, and F.P. Miandehi. *Detecting earthquake damage levels using adaptive boosting*. in *2013 8th Iranian Conference on Machine Vision and Image Processing (MVIP)*. 2013. IEEE.
31. Ram, S., et al., *Predicting asthma-related emergency department visits using big data.* IEEE journal of biomedical and health informatics, 2015. **19**(4): p. 1216-1223.
32. Mishra, S., D. Mishra, and G.H. Santra, *Adaptive boosting of weak regressors for forecasting of crop production considering climatic variability: An empirical assessment.* Journal of King Saud University-Computer and Information Sciences, 2017.
33. Patil, S., A. Patil, and V.M. Phalle, *Life Prediction of Bearing by using Adaboost Regressor.* Available at SSRN 3398399, 2018.
34. González, S., F. Herrera, and S. García. *Managing monotonicity in classification by a pruned adaboost*. in *International Conference on Hybrid Artificial Intelligence Systems*. 2016. Springer.
35. Rodriguez-Galiano, V.F., et al., *An assessment of the effectiveness of a random forest classifier for land-cover classification.* ISPRS Journal of Photogrammetry and Remote Sensing, 2012. **67**: p. 93-104.
36. Pijls, K.E., et al., *A profile of volatile organic compounds in exhaled air as a potential non-invasive biomarker for liver cirrhosis.* Sci Rep, 2016. **6**: p. 19903.
37. Tedjo, D.I., et al., *The fecal microbiota as a biomarker for disease activity in Crohn's disease.* Sci Rep, 2016. **6**: p. 35216.
38. Gray, K.R., et al., *Random forest-based similarity measures for multi-modal classification of Alzheimer's disease.* NeuroImage, 2013. **65**: p. 167-175.
39. Wang, Z., et al., *Flood hazard risk assessment model based on random forest.* Journal of Hydrology, 2015. **527**: p. 1130-1141.
40. Fraiwan, L., et al., *Automated sleep stage identification system based on time–frequency analysis of a single EEG channel and random forest classifier.* Computer methods and programs in biomedicine, 2012. **108**(1): p. 10-19.
41. Nan, F., J. Wang, and V. Saligrama. *Pruning random forests for prediction on a budget*. in *Advances in neural information processing systems*. 2016.
42. González, S., F. Herrera, and S. García, *Monotonic random forest with an ensemble pruning mechanism based on the degree of monotonicity.* New Generation Computing, 2015. **33**(4): p. 367-388.
43. Winham, S.J., R.R. Freimuth, and J.M. Biernacka, *A weighted random forests approach to improve predictive performance.* Statistical Analysis and Data Mining: The ASA Data Science Journal, 2013. **6**(6): p. 496-505.
44. Daho, M.E.H., et al. *Weighted vote for trees aggregation in random forest*. in *2014 International Conference on Multimedia Computing and Systems (ICMCS)*. 2014. IEEE.
45. Guelman, L., *Gradient boosting trees for auto insurance loss cost modeling and prediction.* Expert Systems with Applications, 2012. **39**(3): p. 3659-3667.
46. Semanjski, I. and S. Gautama, *Smart city mobility application—Gradient boosting trees for mobility prediction and analysis based on crowdsourced data.* Sensors, 2015. **15**(7): p. 15974-15987.
47. Touzani, S., J. Granderson, and S. Fernandes, *Gradient boosting machine for modeling the energy consumption of commercial buildings.* Energy and Buildings, 2018. **158**: p. 1533-1543.
48. Fan, C., et al. *PredRSA: a gradient boosted regression trees approach for predicting protein solvent accessibility*. in *Bmc Bioinformatics*. 2016. BioMed Central.
49. Krasnosel'skii, M. and Y.B. Rutickii, *Convex Functions and Orlicz Spaces, P. Noordhoff Ltd.* Groningen-The Netherlands, 1961.
50. Boyd, S. and L. Vandenberghe, *Convex optimization*. 2004: Cambridge university press.
51. Ke, G., et al. *Lightgbm: A highly efficient gradient boosting decision tree*. in *Advances in Neural Information Processing Systems*. 2017.
52. Dorogush, A.V., V. Ershov, and A. Gulin, *CatBoost: gradient boosting with categorical features support.* arXiv preprint arXiv:1810.11363, 2018.

53. Upton, G. and I. Cook, *A Dictionary of Statistics 2 rev.* 2008.
54. Wold, S., M. Sjöström, and L. Eriksson, *PLS-regression: a basic tool of chemometrics.* Chemometrics and intelligent laboratory systems, 2001. **58**(2): p. 109-130.
55. Barker, M. and W. Rayens, *Partial least squares for discrimination.* Journal of Chemometrics, 2003. **17**(3): p. 166-173.
56. Sinkovics, R.R., et al., *Testing measurement invariance of composites using partial least squares.* International marketing review, 2016.
57. Richter, N.F., et al., *European management research using partial least squares structural equation modeling (PLS-SEM).* European Management Journal, 34 (6), 589-597., 2016.
58. Teo, A.-C., et al., *Why consumers adopt mobile payment? A partial least squares structural equation modelling (PLS-SEM) approach.* International Journal of Mobile Communications, 2015. **13**(5): p. 478-497.
59. Khedher, L., et al., *Early diagnosis of Alzheimer's disease based on partial least squares, principal component analysis and support vector machine using segmented MRI images.* Neurocomputing, 2015. **151**: p. 139-150.
60. Fedoseeva, L.A., et al., *Molecular determinants of the adrenal gland functioning related to stress-sensitive hypertension in ISIAH rats.* BMC Genomics, 2016. **17**(Suppl 14): p. 989.
61. Kroes, A., et al., *Brevicoryne brassicae aphids interfere with transcriptome responses of Arabidopsis thaliana to feeding by Plutella xylostella caterpillars in a density-dependent manner.* Oecologia, 2017. **183**(1): p. 107-120.
62. Lee, L.C., C.Y. Liong, and A.A. Jemain, *Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: a review of contemporary practice strategies and knowledge gaps.* Analyst, 2018. **143**(15): p. 3526-3539.
63. Gromski, P.S., et al., *A tutorial review: Metabolomics and partial least squares-discriminant analysis--a marriage of convenience or a shotgun wedding.* Anal Chim Acta, 2015. **879**: p. 10-23.
64. Tenori, L., et al., *Serum metabolomic profiles evaluated after surgery may identify patients with oestrogen receptor negative early breast cancer at increased risk of disease recurrence. Results from a retrospective study.* Molecular Oncology, 2015. **9**(1): p. 128-139.
65. Zacharias, H.U., et al., *Identification of Plasma Metabolites Prognostic of Acute Kidney Injury after Cardiac Surgery with Cardiopulmonary Bypass.* Journal of Proteome Research, 2015. **14**(7): p. 2897-2905.
66. Domingo, C. and O. Watanabe, *Scaling up a boosting-based learner via adaptive sampling.* Knowledge Discovery and Data Mining, Proceedings, 2000. **1805**: p. 317-328.
67. Davis, D.J., C. Burlak, and N.P. Money, *Osmotic pressure of fungal compatible osmolytes.* Mycological Research, 2000. **104**: p. 800-804.
68. Bouwmeester, R., L. Martens, and S. Degroeve, *Comprehensive and Empirical Evaluation of Machine Learning Algorithms for Small Molecule LC Retention Time Prediction.* Analytical Chemistry, 2019. **91**(5): p. 3694-3703.
69. Lee, M.Y. and T. Hu, *Computational Methods for the Discovery of Metabolic Markers of Complex Traits.* Metabolites, 2019. **9**(4).
70. Degenhardt, F., S. Seifert, and S. Szymczak, *Evaluation of variable selection methods for random forests and omics data sets.* Brief Bioinform, 2019. **20**(2): p. 492-503.
71. Tran, T.N., et al., *Interpretation of variable importance in Partial Least Squares with Significance Multivariate Correlation.* Chemometrics and intelligent laboratory systems, 2014. **138**(15): p. 153-160.
72. Baratloo, A., et al., *Part 1: Simple Definition and Calculation of Accuracy, Sensitivity and Specificity.* Emerg (Tehran), 2015. **3**(2): p. 48-9.
73. Kubat, M., R.C. Holte, and S. Matwin, *Machine Learning for the Detection of Oil Spills in Satellite Radar Images.* Machine Learning, 1998. **30**: p. 195-215.
74. Wojciechowski, S. and S. Wilk, *Difficulty Factors and Preprocessing in Imbalanced Data Sets: An Experimental Study on Artificial Data.* Foundations of Computing and Decision Sciences, 2018. **42**(2): p. 149–176.
75. Dasarathy, B.V. and B.V. Sheela, *Composite classifier system design:concepts and me-thodology.* Proceedings of the IEEE, 1979. **67**(5): p. 708-713.

76. Hansen, L.K. and P. Salamon, *Neural network ensembles* IEEE Transactions on Pattern Analysis and Machine Intelligence, 1990. **12**(10): p. 993-1001.
77. Jordan, M.J. and R.A. Jacobs, *Hierarchical mixtures of experts and the EM algorithm.* Neural Computation, 1994. **6**(2): p. 181-214.
78. Jacobs, R.A., et al., *Adaptive mixtures of local ex-perts.* Neural Computation, 1991. **3**: p. 79-87.
79. Friedman, J., T. Hastie, and R. Tibshirani, *Additive Logistic Regression: A Statistical View of Boosting. Boosting.* The Annals of Statistics, 2000. **28**(2): p. 337–407.
80. Breiman, L., *Random forests.* Machine Learning, 2001. **45**(1): p. 5-32.
81. Han, J., M. Kamber, and J. Pei, *Classification: Basic Concepts*, in *Data mining*, T.M.K.S.i.D.M. Systems, Editor. 2012, Elsevier. p. 327-391.
82. Lu, Y.M., et al., *DELTA: A Distal Enhancer Locating Tool Based on AdaBoost Algorithm and Shape Features of Chromatin Modifications.* Plos One, 2015. **10**(6).
83. Yu, Q.Z., *Weighted bagging: a modification of AdaBoost from the perspective of importance sampling.* Journal of Applied Statistics, 2011. **38**(3): p. 451-463.
84. Domingo, C. and O. Watanabe, *MadaBoost: A Modification of AdaBoost.* COLT, 2000: p. 180–189.
85. Freund, Y., *An adaptive version of the boost by majority algorithm.* Machine Learning, 2001. **43**(3): p. 293-318.
86. Canovas-Garcia, F., et al., *Modification of the random forest algorithm to avoid statistical dependence problems when classifying remote sensing imagery.* Computers & Geosciences, 2017. **103**: p. 1-11.
87. Shipway, N.J., et al., *Performance Based Modifications of Random Forest to Perform Automated Defect Detection for Fluorescent Penetrant Inspection.* Journal of Nondestructive Evaluation, 2019. **38**(2).
88. Strempel, S., et al., *Using Conditional Inference Trees and Random Forests to Predict the Bioaccumulation Potential of Organic Chemicals.* Environmental Toxicology and Chemistry, 2013. **32**(5): p. 1187-1195.
89. Dasgupta, S. and Y. Freund, *Random Projection Trees and Low Dimensional Manifolds.* Stoc'08: Proceedings of the 2008 Acm International Symposium on Theory of Computing, 2008: p. 537-546.
90. Polikar, R., *Ensemble based systems in decision making.* IEEE Circuits and Systems Magazine, 2006. **6**(3): p. 21-45.
91. Polikar, R., *Bootstrap inspired techniques in computational intelligence: ensemble of classifiers, incremental learning, data fusion and missing features.* IEEE Signal Processing Magazine, 2007. **24**(4): p. 59-72.
92. Mendez, K.M., S.N. Reinke, and D.I. Broadhurst, *A comparative evaluation of the generalised predictive ability of eight machine learning algorithms across ten clinical metabolomics data sets for binary classification.* Metabolomics, 2019. **15**(12): p. 150.

CHAPTER 6

Advanced data fusion: Random forest proximities and pseudo-sample principle towards increased prediction accuracy and variable interpretation

Georgios Stavropoulos, Robert van Voorstenbosch, Daisy M.A.E. Jonkers, John Penders, Jane E. Hill, Frederik-Jan van Schooten, Agnieszka Smolinska

Analytica Chimica Acta, Volume 1183, 23 October 2021, 339001, doi: 10.1016/j.aca.2021.339001

Abstract

Data fusion has gained much attention in the field of life sciences, and this is because analysis of biological samples may require the use of data coming from multiple complementary sources to express the samples fully. Data fusion lies in the idea that different data platforms detect different biological entities. Therefore, if these different biological compounds are then combined, they can provide comprehensive profiling and understanding of the research question in hand. Data fusion can be performed in three different traditional ways: low-level, mid-level, and high-level data fusion. However, the increasing complexity and amount of generated data require the development of more sophisticated fusion approaches. In that regard, the current study presents an advanced data fusion approach (i.e. proximities stacking) based on random forest proximities coupled with the pseudo-sample principle. Four different data platforms of 130 samples each (faecal microbiome, blood, blood headspace, and exhaled breath samples of patients who have Crohn's disease) were used to demonstrate the classification performance of this new approach. More specifically, 104 samples were used to train and validate the models, whereas the remaining 26 samples were used to validate the models externally. Mid-level, high-level, as well as individual platform classification predictions, were made and compared against the proximities stacking approach. The performance of each approach was assessed by calculating the sensitivity and specificity of each model for the external test set, and visualised by performing principal component analysis on the proximity matrices of the training samples to then, subsequently, project the test samples onto that space. The implementation of pseudo-samples allowed for the identification of the most important variables per platform, finding relations among variables of the different data platforms, and the examination of how variables behave in the samples. The proximities stacking approach outperforms both mid-level and high-level fusion approaches, as well as all individual platform predictions. Concurrently, it tackles significant bottlenecks of the traditional ways of fusion and of another advanced fusion way discussed in the paper, and finally, it contradicts the general belief that the more data, the merrier the result, and therefore, considerations have to be taken into account before any data fusion analysis is conducted.

Keywords: Data fusion; proximities; stacking; variable behaviour; Crohn's disease; classification

Introduction

Data fusion has gained much attention in the field of, among others, life sciences [1-10], and this is because analysis of biological samples may require the use of data coming from multiple complementary sources to express the samples fully. The principle behind data fusion lies in the idea that different data platforms, such as gas chromatography-mass spectrometry (GC-MS) and nuclear magnetic resonance (NMR) detect different biological entities. Therefore, if these different biological compounds are then combined, they can provide comprehensive profiling and understanding of the research question in hand [2]. Theoretically, one would imagine that the more data generated per biological sample, the merrier since different data platforms demonstrate different strengths. Practically, this is not always the case; considerations have to be made regarding the research question, and the nature of the samples before any data fusion analysis is conducted. Data fusion can be performed in three different ways: low-level, mid-level, and high-level data fusion [5]. At the low-level, the various data platforms are fused at a data level, whereas in the mid-level, the platforms are fused at a data level of selected variables or features of the original data. At the high-level, the platforms are fused at a prediction level, meaning that each platform gives predictions individually and then, these individual predictions are combined to get the final prediction.

Recently, a more sophisticated way of data fusion was introduced that can also be seen as a modified version of mid-level fusion [1]. Smolinska et al. introduced the fusion of kernels of the individual platforms rather than the important variables, features or latent variables of the platforms. More specifically, they mapped each platform to a higher-dimensional feature space with the use of a kernel function, and they then fused all the individual kernels by using a weighted sum. Kernel functions transform the data in such a way that they result in non-negative square matrices, and these matrices can be seen as measures of similarity/dissimilarity of samples; therefore, when one works with kernels, they work with samples rather than variables. This approach holds great potential when it comes to unravelling trends in data or getting predictions of data since it considers both linear and nonlinear relations amongst data, and most of the biological systems reveal nonlinear characteristics [1]. Another advantage of working with kernels, and therefore samples, rather than variables/features is that scaling issues are overcome. For example, in a mid-level fusion approach, scaling of the original variables is required before any data from different sources are concatenated since the magnitude of the data coming from different sources is most likely different. To find the optimal scaling parameter that would suit all the data might be not an easy task to perform, and on top of that, if the data being concatenated are of different type (i.e. quantitative or discrete), then this issue gets even more challenging. The major disadvantage of working with kernels is that information about the importance/contribution of variables of the dataset in

the model performance is lost due to the transformation of variables to distance or similarity measures among samples, and it can be challenging to trace back these variables. Nonlinear bi-plots introduced by Gower at al. have been further modified and developed the idea of pseudo-samples by Krooshof et al. [11,12] and Smolinska et al. [13], to overcome this bottleneck. The pseudo-sample principle uses the transformed data (i.e. the square matrices) to illustrate not only the importance of the original variables but also the original variable trajectory (i.e. how the variables behave amount-wise) in the samples of interest, which are both essential assets when it comes to drawing safe conclusions on the study results.

Proximity matrices are actual measures of similarity/dissimilarity of samples, and they are non-negative square matrices [14]. Originally, the term proximity means "closeness" or "nearness" between pairs, and it is calculated by using traditional distance measures such as Euclidean distance or Gaussian distance. The closer to zero the proximity of two samples is, the more similar these two samples are; this is why the diagonal of a proximity matrix always consists of zeros. The square matrix has a size of $n \times n$ (where n is the number of samples in the original dataset) since proximities imply similarities amongst samples. Moreover, proximities do not consist of transformed data, which is the case with kernels (e.g. the original dataset is transformed using the radial basis function), but instead of newly generated data (i.e. distances in space among samples). Random forest (RF) also returns a proximity matrix of the data that it is run on; although, the proximity matrix here is calculated differently [15]. The RF proximity matrix is indicative of the number of times that samples ended up in the same terminal node rather than a demonstration of the actual distance in the space of samples. More details on how the proximity via RF is calculated are shown in the materials and methods section. Recently, Blanchet et al. [16] published a tutorial where they illustrate the successful implementation of the RF proximities along with the pseudo-sample principle to visualize variable importance. However, to the best of the authors' knowledge, proximity matrices, and mainly RF proximity matrices, have not been examined before in terms of data fusion to check their performance on predicting and investigating complex biological samples.

In this research, Crohn's disease (CD) serves as a case study to demonstrate the utility of data fusion using proximities. CD is a complex biological metabolic disorder. CD is a chronic inflammatory process with no known cause (idiopathic) that can affect any part of the gastrointestinal tract, from the mouth to the anus [17]. More specifically, CD causes muscle hypertrophy, it changes the colon to a cobblestone appearance, it creates fissures in the colon, and it also covers the colon with fat. Colonoscopy has been the gold standard to diagnose and monitor the disease activity; therefore, alternative ways (e.g. biological biomarkers) to diagnose and monitor the disease activity are needed since colonoscopy is a considerably invasive and costly technique. Previous research focused on identifying CD biomarkers in either human blood (i.e.

metabolites) or faeces (i.e. bacterial species) [18-21] to diagnose and monitor the disease activity. All studies demonstrated promising results as far as prediction accuracy is concerned; although, each of these studies examined one data platform only to draw their conclusions. Consequently, the aim of the present study is to propose a new, advanced fusion approach based on RF proximities, as well as to see whether prediction accuracy of CD can be increased if more data are concatenated along with potential biomarker behaviour examination using pseudo-sample principle. To illustrate that this new fusion approach performs well, it is advantageous over the currently existing fusion approaches, and that it can be implemented in biomedical data, it is compared against the current ways of data fusion and the individual platforms used in the present study.

Materials and Methods

Data used and data preprocessing

Four different data platforms were used: faecal microbiome, blood, blood headspace, and exhaled breath samples from patients suffering from CD. The CD patients were categorized into two classes based on the disease activity: remission and active cases of CD. The criteria used to classify the patients as being in either remission or active stage can be found elsewhere [19]. In the present study, 130 CD patients were sampled, of which 66 were patients in the remission stage of the disease, and the remaining 64 were patients in the active stage of the disease. Initially, all the raw data were preprocessed before the actual analysis took place. Data preprocessing diminishes the effect of possible instrumental artefacts that can occur during the analysis. Each data platform followed a different preprocessing strategy.

The faecal microbiome samples were treated and sampled as described elsewhere [18], and they were analysed by employing 16S ribosomal RNA pyrosequencing. The faecal microbiome was analysed in terms of operational taxonomic units (OTUs). The raw microbiome pyrosequencing reads were, first, preprocessed by means of quality filters to reduce the error rate, and de-multiplexed and clustered into OTUs based on a 97% similarity—the entire prepossessing procedure that was followed is described elsewhere [18]. Then, they were transformed into continuous data. This is because preprocessing of the pyrosequencing reads results in data counts (i.e. OTUs per sample) which cannot be used for multivariate analysis purposes; the transformation was done by employing the inverse hyperbolic sine [22]. Next, the exclusion of zeros followed. The majority of bacterial species (OTUs) is not present in all the samples; consequently, only those that are present in a specified per cent of the samples are kept. Here, species that were found in at least 35% [18] of the samples were retained. As a final preprocessing step, microbiome data were logarithmically transformed since the log transformation accounts for high skewness in the data.

The blood was treated and sampled as described elsewhere [23], and the blood sample metabolites were analysed by using NMR Bruker 600 MHz with a cryoprobe. In the blood NMR data, first, the water peak was removed, and then, baseline correction via P-splines [24], misalignment correction via correlation optimized warping [25], and peak picking in the form of binning via adaptive intelligent binning [26] were performed. Moreover, normalization via a reference peak (i.e. trimethylsilyl-propanoic acid–TSP) as well as via probabilistic quotient normalization [27] followed. Normalization via the TSP peak is done to enhance the signal comparison among the samples, whereas probabilistic quotient normalization accounts for dilution effects, effect size, among the samples. Finally, the blood data were logarithmically transformed.

The blood headspace was treated and sampled as described elsewhere [28]; in short, the blood headspace samples were measured by utilizing gas chromatography/gas chromatography-*time-of-flight*-mass spectrometry (GC×GC-*tof*-MS; Pegasus 4D, LECO Corporation, St Joseph, MI, USA). Blood headspace was analysed in terms of volatile organic compounds (VOCs). The blood headspace data were initially preprocessed as discussed elsewhere [28], and in the end, the exclusion of zeros followed. As with the microbiome data, the majority of VOCs does not occur in all the samples; therefore, only those found present in at least 20% [29] of the samples coming from the same class were kept for further analysis. In the end, a logarithmic transformation was performed.

Finally, the exhaled breath was captured as described elsewhere [19], and the exhaled breath samples were analysed by using GC-*tof*-MS. Breath was analysed in terms of VOCs as well. The exhaled breath data were preprocessed as described elsewhere [19], and as an extra preprocessing step, these data underwent exclusion of zeros (compounds found in at least 20% of each class [29] of the samples were retained) and logarithmic transformation.

Data fusion approaches

Data platforms can traditionally be fused at three different levels: low-level, mid-level, and high-level data fusion [5]. Low-level data fusion refers to concatenation of the whole data platforms, sample-wise, into a single matrix that consists of as many rows as the number of samples, and as many columns as the total number of variables from all different data platforms. Low-level fusion attempts were not tried here because this would affect the degrees of freedom of the data, and thus, making the concatenated matrix challenging to deal with and the analysis results untrustworthy; readers interested in low-level fusion applications are referred to [5].

Mid-level fusion

Mid-level data fusion can be divided into two categories: the concatenation of either important/significant variables or features of the different platforms. A variety of ways exists to find important variables or features. For example, variables can be found by using, among others, RF [15], partial least squares based variable selection [30], or even significance multivariate correlation [31], whereas features (or latent variable space) can be found by implementing principal component analysis (PCA) (and use the principal components) [32], recursive feature elimination [33], or partial least squares analysis (and use the latent variables) [34, 35]. Then, all these variables or features are concatenated, sample-wise, to create the single fused matrix to be used for further analysis. In the present study, RF was used to find the most important variables per platform. A schematic representation of the mid-level fusion approach is given in Figure 1.

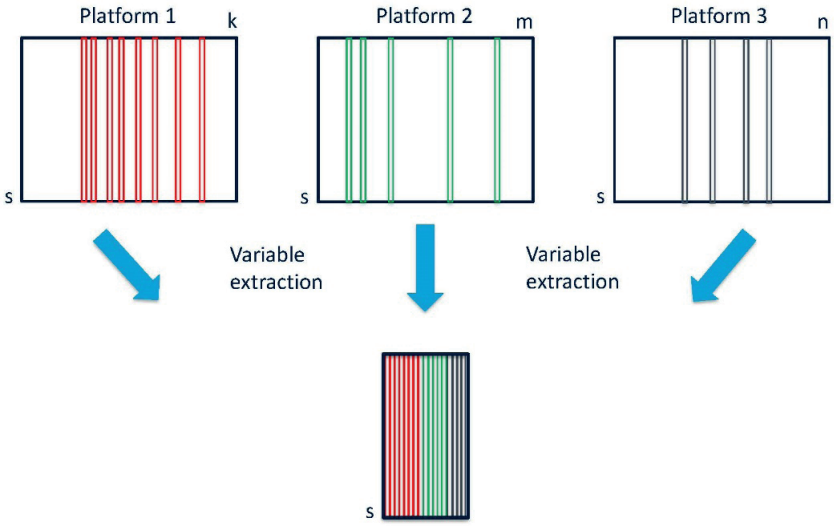


FIGURE 1: SCHEMATIC REPRESENTATION OF THE MID-LEVEL FUSION APPROACH OF THREE DATASETS. RF IS RUN ON EACH OF THE DATASETS TO GET THEIR MOST IMPORTANT VARIABLES. THEN, ALL THE IMPORTANT VARIABLES OF ALL THREE DATASETS ARE FUSED SAMPLE-WISE TO GET THE FINAL FUSED MATRIX.

High-level fusion

High-level data fusion refers to a combination of the outcome of the individual platforms; this is why it is also called as decision-level fusion. Specifically, a classification or regression model is built for each one of the available data platforms, and the results from each model are combined to obtain the final decision for every sample of interest. The outcome of each model is given as either a class label or a set of probabilities; therefore, one can choose to either use majority voting [36] or adjusted probabilities to get the

final decision for the samples of interest. In the current study, adjusted probabilities via the Bayes' theorem [37] were used to get the final decisions, and the optimal decision threshold was found from a loop of 100 cross-validation iterations; in every iteration, the data were randomly split into training and validation sets, and therefore, each iteration used different training and validation samples. Bayes' theorem is also called Bayesian integration because it provides the ability to define probability models for disparate or independent types of data. More specifically, RF was used on every single platform to get the sets of initial likelihood probabilities (i.e. prior probabilities) for every sample of interest, and then, these probabilities were transformed into posterior probabilities. A detailed description of the implementation of the Bayes' theorem in biological data can be found elsewhere [38]. A schematic representation of the high-level fusion approach is shown in Figure 2.

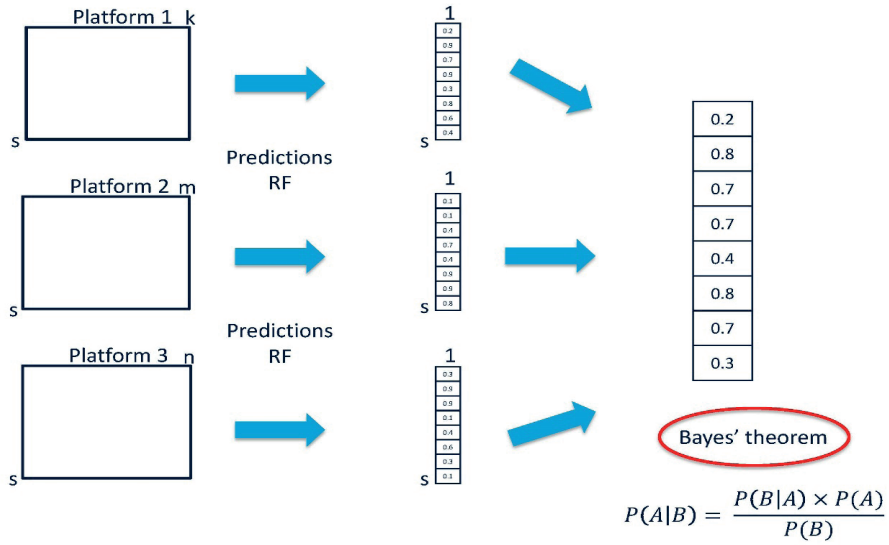


FIGURE 2: SCHEMATIC REPRESENTATION OF THE HIGH-LEVEL FUSION APPROACH OF THREE DATASETS. RF IS RUN ON EACH OF THE DATASETS TO GET THEIR PREDICTIONS (I.E. CLASSIFICATION PROBABILITIES HERE), WHICH ARE THEN ADJUSTED VIA THE BAYES' THEOREM TO GET THE OUTCOME. THE BAYES' THEOREM FORMULA IS DEPICTED AT THE BOTTOM RIGHT CORNER OF THE FIGURE, WHERE P(A) AND P(B) ARE THE PROBABILITIES OF OBSERVING THE EVENTS A AND B RESPECTIVELY, P(B|A) IS THE PROBABILITY OF EVENT B OCCURRING GIVEN THAT EVENT A IS TRUE, AND P(A|B) IS THE PROBABILITY OF EVENT A OCCURRING GIVEN THAT EVENT B IS TRUE.

Proximities stacking

The current study implemented a modified version of mid-level fusion. This approach makes use of proximity matrices ($\boldsymbol{P}_i$) of the original platforms, which are then arranged one on top of each other; consequently, this approach was called as proximities stacking. The proposed approach consists of two steps: the creation of the ($\boldsymbol{P}_i$)

matrices of each used data platforms, and the discovery of an optimal set of weights w with which the platforms are combined in a weighted linear parameterized way to create a new single proximity matrix $\boldsymbol{K}$ that is used for further analysis.

$\boldsymbol{P}_i$ matrices are square distance matrices that show how similar or dissimilar the data are. RF returns $\boldsymbol{P}_i$ matrices of the data that the algorithm was run on, and these proximities were used here; however, these $\boldsymbol{P}_i$ matrices are not calculated by using a distance measure [15]. More specifically, for every pair of samples, the proximity indicates the percentage of the times these two samples ended up in the same terminal node. For instance, if the RF consists of 1000 trees and the pair of samples ended up in the same terminal node in 100 of the 1000 trees, then the proximity for this pair of samples is 100/1000 = 0.1. As a result, the higher the proximity, the more similar the objects are. This means that the diagonal of the RF $\boldsymbol{P}_i$ matrix is filled with ones rather than zeros (as an actual proximity matrix); therefore, the RF $\boldsymbol{P}_i$ matrix is subtracted from one to, ultimately, transform it to an actual distance/proximity matrix ($\boldsymbol{P}_{trans,i}$). Equation (1) depicts a toy example of such transformation.

$$\boldsymbol{P}_{trans,i} = 1 - \boldsymbol{P}_i = 1 - \begin{pmatrix} 1 & x_{1,2} & x_{1,3} \\ x_{2,1} & 1 & x_{2,3} \\ x_{3,1} & x_{3,2} & 1 \end{pmatrix} = \begin{pmatrix} 0 & 1-x_{1,2} & 1-x_{1,3} \\ 1-x_{2,1} & 0 & 1-x_{2,3} \\ 1-x_{3,1} & 1-x_{3,2} & 0 \end{pmatrix} \quad (1)$$

Subsequently, the $\boldsymbol{P}_{trans,i}$ matrices of the present study were combined in a weighted linear parameterized combination to create the new single proximity matrix $\boldsymbol{K}$ that was used for further analysis. This linear combination can be expressed as follows:

$$\boldsymbol{K} = \sum_{i=1}^{m} w_i \times \boldsymbol{P}_{trans,i} \quad (2)$$

where m is the total number of $\boldsymbol{P}_{trans,i}$ matrices (here, m equals four), and w_i is the weight or importance of the $\boldsymbol{P}_{trans,i}$ matrix in the new $\boldsymbol{K}$ matrix. The set of weights w can be found by applying regularisation methods such as L_1 or L_2 norm. Regularisation methods are processes that introduce additional information to prevent over-fitting. L_2 norm is applied when the data platforms are complementary to each other because it avoids the possibility of shrinking the importance of any of the platforms; L_2 norm [1] was used here, and it is expressed as follows:

$$\|w\| = \sqrt{\sum_{i=1}^{m} w_i^2} = 1 \quad (3)$$

where m is the total number of $\boldsymbol{P}_{trans,i}$ matrices, and w_i is the weight of the $\boldsymbol{P}_{trans,i}$ matrix. The optimal set of weights was selected in two steps approach. In the first step, ten sets of numbers that fulfilled the equation (3) were generated via grid search. Then, the weight values of every w_i were shuffled to create a total of 40 different possible combinations of w since there were four data platforms available in this study. The $w_{optimal}$ that maximizes classification accuracy of the model was found from a loop of 100 cross-validation iterations. A schematic representation of the proximities stacking fusion is pictured in Figure 3.

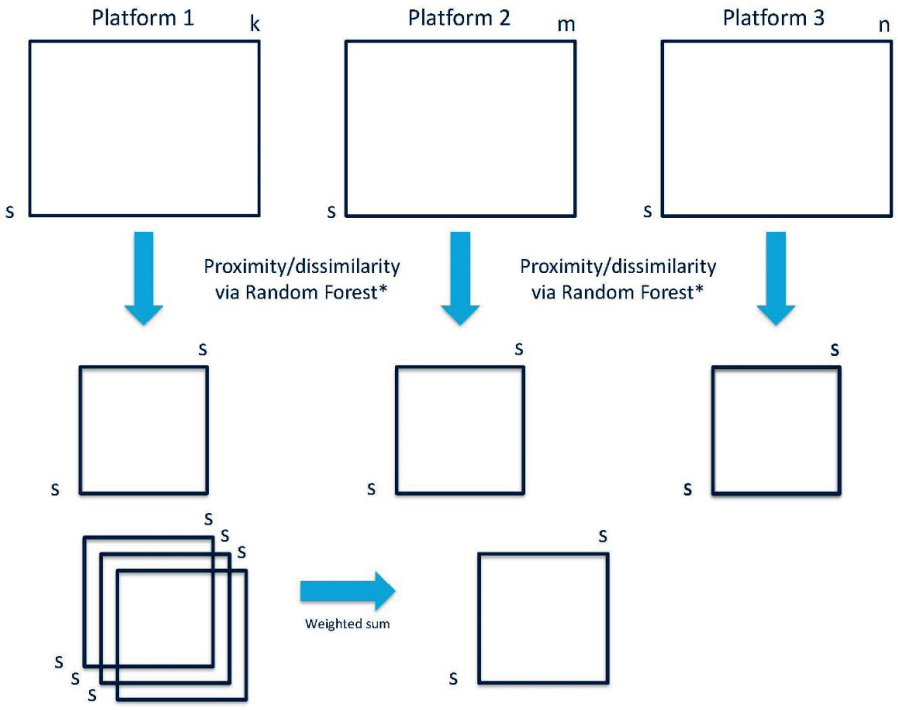


FIGURE 3: SCHEMATIC REPRESENTATION OF THE PROXIMITIES STACKING FUSION APPROACH OF THREE DATASETS. RF IS RUN ON EACH OF THE DATASETS TO GET THEIR PROXIMITY MATRIX. THEN, ALL THREE PROXIMITY MATRICES ARE STACKED ONE ON TOP OF EACH OTHER, AND VIA A WEIGHTED SUM, THEY CREATE THE FINAL SINGLE PROXIMITY MATRIX K. *THE PROXIMITY/DISSIMILARITY MATRICES CAN ALSO BE CREATED VIA UNSUPERVISED RANDOM FOREST, AND THESE PROXIMITIES WERE USED IN THE PRESENT STUDY. MORE DETAILS ON THE MATTER CAN BE FOUND IN THE SECTION 2.4.3.

Pseudo-sample principle

The pseudo-sample principle was employed to explore the behaviour and importance of the original variables (i.e. bacterial species, metabolites, and VOCs) in the final classification model in the proximities stacking fusion approach [11]. A pseudo-

sample is a matrix that has the values of one particular variable from an entire dataset (e.g. $\boldsymbol{A} = (n \times p)$, where n is the number of samples, and p is the number of variables) sorted out in one column, and the rest of its columns are filled in with zeros. For every original variable in the $\boldsymbol{A}$ matrix, a $\boldsymbol{B} = (k \times p)$ pseudo-sample matrix is created, where k is the number of points that one chooses to spread the range of the values of that particular variable on. Based on existing literature [1, 11], k usually ranges from 20 to 40, to properly represent the range of the values of each variable—the present study used 40 points. Then, this $\boldsymbol{B}$ matrix is predicted using RF, which results in obtaining its corresponding pseudo-sample proximity matrix. In the end, one gets as many pseudo-sample proximity matrices as the total number of the original variables (i.e. p pseudo-sample proximity matrices to be analysed, in total). A graphical illustration of how a single pseudo-sample proximity matrix is created is shown in Figure 4. As a final step, principal coordinate analysis (PCoA) is run on the proximity matrix of the original dataset, and subsequently, all the pseudo-sample proximity matrices are projected onto the PCoA space of the proximity of the original dataset since they can be treated as any other subject/patient sample.

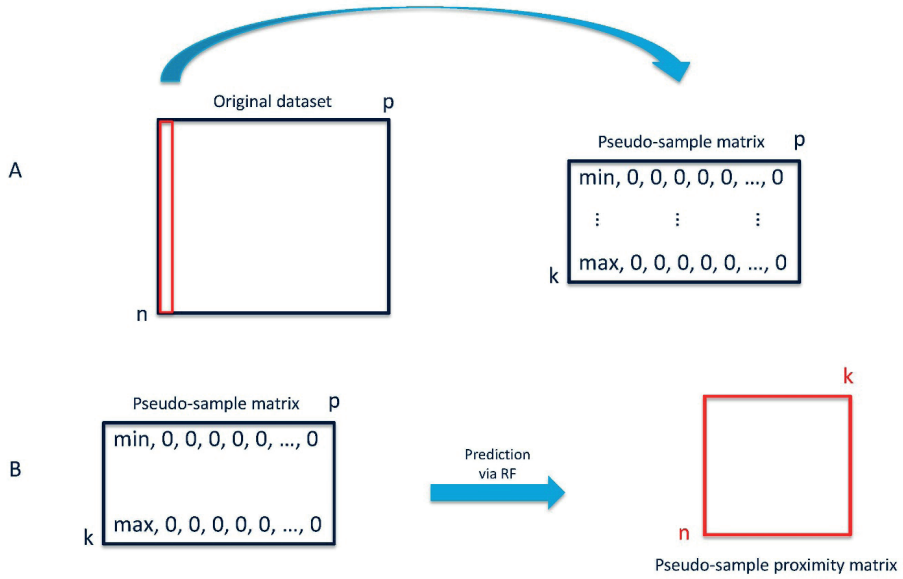


FIGURE 4: GRAPHICAL REPRESENTATION OF HOW THE PSEUDO-SAMPLE PROXIMITY MATRIX OF THE VERY FIRST ORIGINAL VARIABLE IS CREATED. A) FIRST, ALL THE VALUES OF VARIABLE ONE ARE SORTED OUT AND PLACED IN COLUMN ONE OF THE PSEUDO-SAMPLE MATRIX, WHEREAS THE REMAINING OF THE COLUMNS ARE FILLED IN WITH ZEROS. B) THEN, RF IS RUN ON THE PSEUDO-SAMPLE MATRIX TO OBTAIN THE PSEUDO-SAMPLE PROXIMITY MATRIX, WHICH ULTIMATELY HOLDS INFORMATION OF THE VERY FIRST VARIABLE ONLY. N IS THE NUMBER OF SAMPLES, P IS THE NUMBER OF THE ORIGINAL VARIABLES, AND K IS THE NUMBER OF POINTS THAT THE USER CHOOSES TO SPREAD THE RANGE OF THE VALUES OF VARIABLE ONE ON.

Conceptual flowchart and fusion approach optimization

Variable selection and RF model optimization

Each data platform underwent preprocessing, and then, its samples were divided into the training and validation samples (i.e. 104 samples, of which 57 were remission and 47 were active), and independent internal test set samples (i.e. 26 samples, of which 11 were remission and 15 were active). The division between the training and validation, and the independent internal test samples was achieved by employing the Duplex algorithm [39] since Duplex algorithm aims to maintain a comparable diversity between the sets. The URF/RF model parameters (i.e. number of trees, predictors, and samples per tree terminal leaf per RF model), as well as to the number of variables to be kept per platform were optimized within a 1000-iteration loop—the number of samples per tree terminal leaf accounts for overfitting minimization and model complexity reduction.

For each iteration, the training and validation set samples were randomly split (80% of the 104 samples were used as training samples, and the remaining 20% of the 104 samples were used as validation samples), an RF model was built, and the importance of every variable was found. By default, a variable is considered important if its importance value is positive; however, here, a variable was considered as important if its importance value was equal or higher than 30% of the amount of the highest variable importance value found in the RF model. Next, the number of times that every single variable had been found as important in all the 1000 RF models was calculated (i.e. counts per variables), and in the end, the variables that had the most counts were kept. The threshold which determined the optimal number of variable counts to be kept for further analysis differed per platform since the data platforms contained different types of data. For each of the 1000 iterations, a one-by-one backwards variable elimination procedure was performed, and every time a variable was eliminated, the root-mean-square-error-prediction (RMSEP) value was calculated. The number of variables that gave the lowest RMSEP value was considered as optimal. Each of the 1000 iterations gave its own optimal number of variables, and by averaging them out, the optimal number of variables per platform (i.e. counts per variable) was found. The RF model parameters were optimized with a similar way too. The RF model optimization, i.e. number of trees and the number of samples to be kept per tree terminal leaf, was done using the out-of-bag error of the model. As far as the number of predictors to be used in the bootstrapping procedure goes, the square root of the total number of predictors present in the data was used. Finally, a 1000-iteration permutation test was run to confirm that the selection of the RF parameters was indeed optimized. In the present study, 4000 trees per model were used, and at the same time, the minimum number of samples per tree leaf for every tree in each model was set to eight. Ultimately, a new optimized RF model was built by using the 104 training and validation samples to predict the independent internal test set samples. Its performance was assessed by calculating the sensitivity and specificity for the independent internal test set.

Mid-level and high-level fusion

In the mid-level fusion case (Figure 1), the variables with the most counts (found as described in section 2.4.1) from all the platforms were fused, sample-wise, and then, a single optimized RF model was built by using all the 104 samples. Its performance was assessed by calculating the sensitivity and specificity for the independent internal test set and visualized by subsequently performing PCA on the RF proximity matrix of the training samples, where then the independent internal test samples were also projected.

In the high-level fusion case (Figure 2), optimization of the classification probability threshold within a 100-iteration loop followed the optimization of the variable selection and the RF model parameters (section 2.4.1). For each of the 100 iterations, the 104 samples were randomly split into training and validation sets, and individual platform predictions were made. Then, the individual platform classification probabilities were adjusted via the Bayes' theorem, and the receiver operating characteristic (ROC) curve was plotted to find the classification probability threshold that maximized both sensitivity and specificity of the model. The average of all the optimal thresholds of all the 100 models was calculated, and this threshold was then considered optimal. In the end, all 104 samples were used once again to build the final optimized RF model, whose performance was then assessed by predicting the independent internal test set, in terms of sensitivity and specificity.

Proximities stacking fusion

As mentioned already in the data fusion approaches paragraph (paragraph 2.2), first, ten sets of numbers that fulfilled the equation (3) were found. Then, these sets were shuffled to give 40 different possible combinations of sets. A table with all the sets of weights w can be found in the supplementary materials.

For each of the 100 iterations, the 104 samples were randomly split into training and validation sets, and the proximity matrices of these sets and for all the data platforms were obtained by unsupervised random forest (URF) (i.e. four training and four validation proximity matrices) [40]. URF is the unsupervised version of RF that assumes that if there is any structure hidden in the data, it should be possible to distinguish them from a randomly generated version of themselves. URF was employed to get the proximity matrices instead of RF to limit possible overfitting when a small number of samples is used (Figure 3). The use of RF proximities may result in possible overfitting even though an optimization of the RF model has been performed due to the supervised nature of RF. The sample classes of the training data are embedded in the RF model by definition, and when the number of samples is small, it can lead to overfitting and to unnecessarily complex or nonflexible models. The use of URF proximities should suffice for improving classification accuracy; however, if the user does not achieve a fair classification accuracy by using the URF proximities, RF proximities may also

be used. In each iteration and for every set of weights (i.e. w, found via equation (3)), the training proximities were stacked as well as the validation proximities (Figure 3). This resulted in 40 training proximities with their corresponding validation proximities. PCA was applied to every training proximity, and its related validation proximity was projected onto its training proximity PCA space, and based on how well the validation samples were projected on the training sample PCA space, the best set of weights w for this particular set of training and validation samples was found. The number of times that every set of weights w_i was found as the optimal one out of all the 100 iterations was calculated. In brief, an AUC was calculated for every set of weights w to find the optimal one per iteration; the PCo1 scores of each iteration validation set were used to calculate these AUCs. The set of weights w that gave the highest AUC was considered as the optimal. The final classification model was assessed by calculating the sensitivity and specificity for the independent internal test set.

All data analyses were performed by using MatLab R2016b version—the Statistics and Machine Learning Toolbox. For the RF models, the TreeBagger function was used, whereas for the URF models, the code was found elsewhere [40].

Results

The raw microbiome data consisted of 6629 variables, whereas the raw blood data consisted of 32768 variables. The raw blood headspace data consisted of 2549 variables, while the raw exhaled breath data consisted of 545 variables. After data preprocessing and data reduction steps, microbiome matrix was left with 734 variables, blood matrix with 423, blood headspace matrix with 531, and exhaled breath matrix with 256. The optimal number of variables per platform (found via the platform optimization process described in section 2.5.1) to be used for both individual and fused matrices predictions were 58 for the microbiome (the threshold was 50%, meaning that the variables that found as important in more than 50% of the total number of iterations were kept), 19 for blood (with a threshold of 35%), 14 for blood headspace (with a threshold of 40%), and 16 for exhaled breath (with a threshold of 40%). At the same time, all four data platforms consisted of 130 samples, of which 66 were remission cases, and the remaining 64 were active cases of the disease. Notably, 104 samples were used to build and validate the models, whereas the remaining 26 were used to validate the models independently.

Mid-level, high-level, proximities stacking data fusion, as well as individual platform RF models were built, and their performance was assessed by calculating the sensitivity and specificity for the independent internal test set. Furthermore, for the individual platform cases, the mid-level, and the proximities stacking fusion cases, PCA was performed on the training sample proximity matrices, where the independent internal test samples were projected for visualization purposes. The mid-level case gave a

sensitivity of 67% and a specificity of 91% (Table 1) and its corresponding score plot can be seen in Figure 5, while the high-level case gave a sensitivity of 27% and a specificity of 100% (Table 1). In the proximities stacking attempt, the optimal set of weights w was [m_1=0.900 m_2=0.200 m_3=0.100 m_4=0.3872], which shows the contribution of the microbiome, blood, blood headspace, and exhaled breath in the final RF model, respectively; the final RF model gave a sensitivity of 93% and a specificity of 100% (Table 1). The proximities stacking corresponding score plot is illustrated in Figure 6. The sensitivities and specificities of the individual platforms are summarised in Table 1, and their corresponding score plots can be found in the supplementary materials. The proximities stacking approach outperformed both the mid-level and high-level fusion approaches, as well as all the individual platform results in terms of sensitivity and specificity except for the microbiome, which performed equally well.

TABLE 1: SENSITIVITIES AND SPECIFICITIES OF ALL THE FUSION AND ALL THE INDIVIDUAL PLATFORM CASES FOR THE EXTERNAL TEST SET. THE NUMBERS IN THE PARENTHESES SHOW THE ACTUAL NUMBER OF THE CORRECTLY CLASSIFIED PATIENTS; THE NUMBER OF PATIENTS IN EACH INDIVIDUAL PLATFORM DIFFERED FROM EITHER OTHER AND FROM THE NUMBER OF SAMPLES PRESENT IN THE EXTERNAL TEST SET IN THE FUSED CASES. THIS IS BECAUSE SOME PATIENTS PROVIDED ALL THREE SAMPLES (I.E. FAECES, BLOOD, BREATH), WHEREAS SOME OTHERS ONLY PROVIDED ONE (MEANING EITHER ONLY BREATH, OR FAECES, OR BLOOD) OR TWO SAMPLES (MEANING EITHER BLOOD AND FAECES, OR FAECES AND BREATH, OR BREATH AND BLOOD).

	Sensitivity	**Specificity**
Mid-level fusion	67% (10/15)	91% (10/11)
High-level fusion	27% (4/15)	100% (11/11)
Proximities stacking fusion	93% (14/15)	100% (11/11)
Microbiome	95% (19/20)	94% (15/16)
Blood	21% (3/14)	93% (13/14)
Blood headspace	35% (6/17)	47% (8/17)
Exhaled breath	85% (17/20)	50% (8/16)

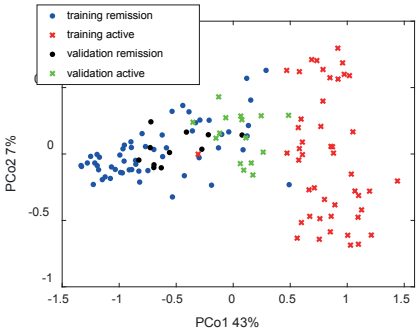


FIGURE 5: SCORE PLOT OF THE TRAINING (I.E. 104) AND VALIDATION (I.E. 26) SAMPLES OF THE RF MODEL IN THE MID-LEVEL FUSION CASE. THE BLUE DOTS REPRESENT THE REMISSION TRAINING SAMPLES, WHEREAS THE BLACK DOTS REPRESENT THE REMISSION VALIDATION SAMPLES. THE RED CROSSES REPRESENT THE ACTIVE TRAINING SAMPLES, WHILE THE GREENS CROSSES REPRESENT THE ACTIVE VALIDATION SAMPLES.

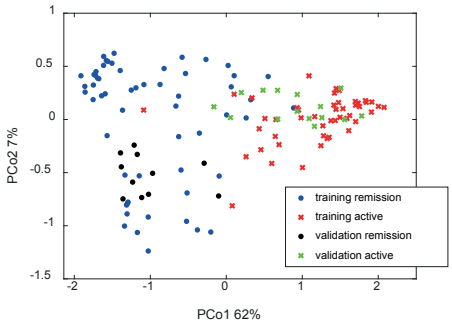


FIGURE 8: SCORE PLOT OF THE TRAINING (I.E. 104) AND VALIDATION (I.E. 26) SAMPLES OF THE RF MODEL IN THE PROXIMITIES STACKING FUSION CASE. THE BLUE DOTS REPRESENT THE REMISSION TRAINING SAMPLES, WHEREAS THE BLACK DOTS REPRESENT THE REMISSION VALIDATION SAMPLES. THE RED CROSSES REPRESENT THE ACTIVE TRAINING SAMPLES, WHILE THE GREENS CROSSES REPRESENT THE ACTIVE VALIDATION SAMPLES.

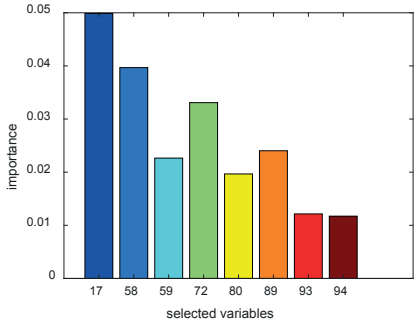


FIGURE 9: BAR PLOT DEPICTING THE IMPORTANCE OF THE TWO MOST IMPORTANT VARIABLES PER PLATFORM (IN TOTAL, THERE WERE 107 FUSED VARIABLES). THE VARIABLES 17 AND 58 COME FROM THE MICROBIOME, THE VARIABLES 59 AND 72 COME FROM BLOOD, THE VARIABLES 80 AND 89 FROM BLOOD HEADSPACE, AND THE VARIABLES 93 AND 94 COME FROM EXHALED BREATH. THE VARIABLE INDICES COME FROM THE RF MODEL, AND THEY REPRESENT THE POSITION OF EACH VARIABLE IN THE DATASET. THE DIFFERENT COLOURS ARE USED FOR ILLUSTRATIVE PURPOSES ONLY.

The results of the pseudo-sample principle applied in the proximities stacking case are shown in Figure 7 and Figure 8. In particular, Figure 7 shows the importance of the two most important variables per platform: the first two variables (i.e., variables number 17 and 58 out of all the 107 that were fused) come from the microbiome, the next two variables (i.e., variables number 59 and 72) come from blood, the following two (i.e., 80 and 89) variables come from blood headspace, and the last two (i.e., 93 and 94) variables come from exhaled breath. It should be mentioned here that the variable numbers represent the position of each variable in the original variable concatenated dataset, and that the importance of each variable was found via the pseudo-samples projected onto the PCoA space, and it is calculated by using the maximum absolute value of the loadings of the original variables trajectories. Figure 8 represents the trajectory plot of two selected variables (i.e. those with the highest importance) per data platform. The variables are colour-coded with the same colours in both figures to provide better illustrative comparisons. More specifically, Figure 8 shows the relation between the top two variables per platform and their relative amount change in the active and remission groups. One can see that the relative amounts of variables 59 and 72 exhibit downregulation in the remission group in comparison to the active group. The other way around holds for the other six variables (i.e., 17, 58, 80, 89, 93, and 94) coming from the microbiome, blood headspace, and exhaled breath. These particular variables are present in very low relative abundance amounts in active cases of the disease, but when these cases become remission, these variables show their highest relative abundance amount.

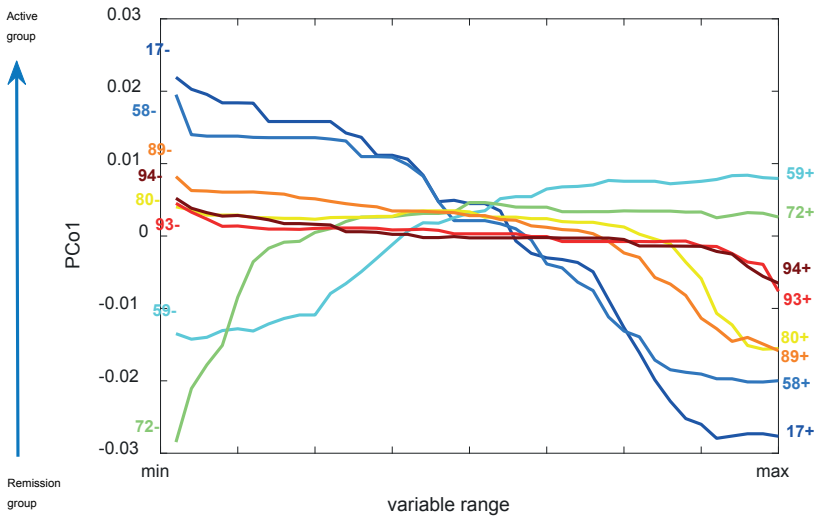


FIGURE 10: TRAJECTORY PLOT OF THE TWO MOST IMPORTANT VARIABLES PER PLATFORM. MORE SPECIFICALLY, VARIABLES 17 AND 58 COME FROM MICROBIOME AND IN ACTIVE GROUPS, THEY ARE PRESENT IN VERY LOW RELATIVE ABUNDANCES, WHILE IN REMISSION CASES, THEY SHOW THEIR HIGHEST RELATIVE ABUNDANCES. VARIABLES 80 AND 89 COME FROM BLOOD HEADSPACE, AND

THEY SHOW THE SAME TREND AS THE ONES COMING FROM MICROBIOME. THE SAME HOLDS FOR VARIABLES 93 AND 94 THAT COME FROM EXHALED BREATH, WHEREAS VARIABLES 59 AND 72 THAT COME FROM BLOOD, THEY ARE PRESENT IN VERY LOW RELATIVE CONCENTRATIONS IN REMISSION GROUPS; WHEN THESE GROUPS BECOME ACTIVE, THESE VARIABLES SHOW THEIR HIGHEST RELATIVE ABUNDANCES. THE DIFFERENT COLOURS ARE USED FOR ILLUSTRATIVE PURPOSES ONLY.

Discussion

The current study investigated the potential of fusing RF proximities of various datasets (i.e. proximities stacking) to ultimately increase the prediction accuracy of disease activity in CD cases, and compared its performance against traditional ways of data fusion in terms of sensitivity and specificity of an external test set. Proximities stacking demonstrated an excellent classification of the independent internal test samples (Figure 6), whereas mid-level fusion (Figure 5) gave a fair classification accuracy of the independent internal test samples. Proximities stacking significantly outperformed all individual platform results as well except for the microbiome case, which performed equally well (Table 1 and supplementary materials). Concurrently, this study also applied the pseudo-sample principle that helped discover and examine possible biomarker behaviour in CD patients in the proximities stacking fusion case (Figure 7 and Figure 8).

Data fusion has proved to be a valuable asset not only in computer science domains but also in life science fields (e.g. metabolomics) too [1-10] as a result of the vast amount of data that are generated nowadays. High-level fusion is rightfully considered as, perhaps, the most potent traditional way of data fusion when it comes to high prediction accuracy due to the way it is defined: many models are combined to get the final predictions instead of one model. The various model outcomes can be combined by using either class labels (i.e. majority voting [36]) or adjusted probabilities. The substantial advantage of choosing adjusted probabilities over majority voting is that one can find how sure the individual models are about their decisions on the samples of interest. Another advantage of high-level fusion is that if a new dataset for the problem in hand becomes available, it can be used to improve the versatility of the decision process too. The major disadvantage, however, of high-level fusion is that it does not give any information about variables/compounds that are important in classifying/predicting samples since it only works with outcomes and not variables. However, in the present study, the high-level fusion results (via the Bayes' theorem) did not demonstrate the best performance with a sensitivity and specificity of 27% and 100%, respectively, which may be due to the limited number of platforms and therefore models that were combined to get the fused outcome. Mid-level fusion can, possibly, increase prediction accuracy when compared to individual platform predictions, as well as it gives the ability to biomarker discovery since it works with either variables or features. In life science fields, and the metabolomic world more specifically, an at least fair prediction accuracy along with biomarker identification are

sought; this is why mid-level fusion has become the most broadly implemented fusion approach. Here, the mid-level fusion results (Figure 5) were inferior to the proximities stacking results (Figure 6), and superior to both the high-level fusion and the individual platform results (Figures S1-S4) except for the microbiome case, achieving a sensitivity and specificity of 67% and 91%, respectively. Further variable importance in the mid-level fusion results (e.g. compound behaviour in the CD samples) was not conducted. Low-level fusion is the least applied approach in the metabolomic world, and as it was mentioned in the 2.3 section already, the degrees of freedom of the data play a crucial role in this. In low-level fusion approach, the error degrees of freedom is negative since the number of variables is almost always a lot bigger than the number of samples; leading to challenges in proper model optimization and development. Metabolomic data are high-dimensionality data on their own (i.e. the number of variables far exceeds the number of samples); therefore, fusing already high-dimensionality data creates matrices of hundreds or thousands of variables which are challenging to be dealt with. This is why low-level fusion was not applied in the present study. Furthermore, all individual platform results (Figures S1-S4) were inferior to the proximities stacking fusion results (Figure 6), except for the microbiome case and superior to the high-level fusion results. The sensitivities and specificities of the individual platforms are summarised in Table 1—the microbiome was the only platform that outperformed the mid-level fusion results.

The fusion of RF/URF proximities by using a weighted sum (i.e. proximities stacking) has not been performed before to the best of the authors' knowledge, and the current study results showed that they could be successfully implemented in complex biological samples, such as CD cases. In particular, proximities stacking demonstrated excellent performance in classifying the external CD cases (Figure 6). The optimal set of weights w was [m_1=0.900 m_2=0.200 m_3=0.0100 m_4=0.3872], which shows the contribution of every platform in the final model. On the one hand, the microbiome contributed the most, and then breath and blood followed. On the other hand, blood headspace contribution was the least. The low contribution of blood headspace contradicts the general belief that the more data, the merrier the result, and as it has been stated already in the introduction, considerations have to be taken before any data fusion analysis is conducted. For example, if the aim of a study is to explore the biology of a system, then the more data gathered would be beneficial; however, if the aim is biomarker discovery, the more data gathered is not always beneficial. The contribution of each platform provided by the set of weights w was to be expected given the individual platform performances. The pseudo-sample principle results illustrated the importance of the original variables in classifying the CD cases (Figure 7), as well as the original variable behaviour in the samples for two selected variables per platform (Figure 8). Figure 7 supports the optimal set of weights w since one can see in the figure that the most high-importance variables are the microbiome variables. In Figure 8, one can see that the blood selected variables are present in very low relative abundances in

the remission cases of CD, and they reach their highest relative abundances in the active CD cases—the other way around holds for the microbiome, blood, and exhaled breath selected variables. Most importantly, Figure 8 helps demonstrate changes that occur in the variable relative abundances. For example, the breath variables (i.e. variable numbers 93 and 94) exhibit an instant increase in their relative abundance when going from active to remission. The same holds for the blood headspace variables (i.e. variable numbers 80 and 89) as well; however, the blood headspace variables exhibit a slower pace increase right before they reach their highest relative abundances. This similar behaviour amongst the blood headspace and exhaled breath variables indicates a connection of these four compounds coming from different sources, and therefore, it can also help dive deeper into the CD pathophysiology.

URF/RF proximities, in terms of fusion, would be of added value in the field of metabolomics and data science, in general. This is because the URF/RF proximities stacking, combined with the pseudo-sample principle approach, has several strengths to show over the other traditional ways of fusion. First of all, it proved that it significantly outperforms the other traditional fusion ways in terms of sample classification, and when compared against the mid-level fusion, it also solves the variable scaling problem since proximities make use of samples rather than variables [5]. Moreover, when compared against the high-level fusion, it solves the variable examination problem that occurs since high-level only uses model outcomes rather than variables [5]. Most importantly, URF/RF proximities stacking, via the weighted sum, also demonstrates the contribution of every platform in the final model, something that no other traditional fusion approach does. The proximities stacking approach also permits the fusion of any type of data (i.e. continuous or discrete), which has proved to be an issue when different data sources are used for a question in hand. It should also be noted here that the URF/RF proximities stacking approach illustrates an essential advantage over the approach reported by Smolinska et al. [1] as well. Smolinska et al. [1] fused kernels instead of proximities. Their approach was successfully applied in metabolomics data, however, finding the optimal kernel for the analysis in hand might be a challenging task to conduct because it requires variable scaling beforehand, and a rather extensive optimisation process. In a fused kernel approach, the user has to select and optimize the type of the kernel and the corresponding parameters, such as the polynomial order if the kernel used is the polynomial or the distribution width if the kernel used is the radial basis function. Finally, it has to be mentioned that in the proximities stacking approach, the final fused matrix (i.e. all the individual proximities combined via the weighted sum) can be used for visualisation purposes of the data as well by directly applying PCA, for instance. In the present study, this fused matrix was used for classification purposes of the independent internal test set samples instead (Figure 6). Linear supervised approaches such as partial-least-squares (PLS) [35] analysis may also be used for either classification or visualisation purposes.

The present study validated its results by using an independent internal test set, thus strengthening its validity even more; nonetheless, the present study also demonstrates some limitations that have to be addressed. The current study did not perform the low-level fusion. Although, it is considered highly unlikely that low-level fusion would have been of any added value to the study since the dimensionally of the data was high, and low-level fusion cannot cope with high dimensionality data. One can also argue that the present analysis lacks variable/compound identification since the pseudo-sample principle permits for compound identification. This was not performed due to the nature of the paper, which is to present the proximities stacking approach rather than identify biomarkers for the disease activity. Lastly, the authors acknowledge the fact that the study results might be seen as accidental since the proposed fusion approach was applied only on one disease data; to prove that the presented approach works on other datasets as well, a simulation analysis was also performed, and it can be found in the supplementary materials. Briefly, four data platforms consisting of 250 samples and 50 variables each were generated. Proximities stacking achieved the best classification results, and the contribution of each simulated platform provided by the set of weights w was to be expected given the individual simulated platform performances. Nevertheless, the proposed fusion approach should be tried on other real data fusion occasions as well to further confirm its strength over the currently available fusion ways.

Conclusion

In conclusion, URF/RF proximities stacking fusion coupled with the pseudo-sample principle approach proved to outperform the traditional ways of fusion significantly, overcame essential drawbacks of the current fusion methods, and helped examine variable behaviour and relations; therefore, establishing itself as a new, powerful data fusion tool that can be implemented in any scientific domain. Data fusion keeps gaining a lot of attention in various scientific fields since combining different types of data can yield higher model performance. However, this is not always the case, and considerations have to be taken into account before any analysis is conducted based on the type of study and the ultimate analysis aim. For example, the data have to be complementary for data fusion to work successfully, and as the present study demonstrates, the more data used or fused does not necessarily mean the merrier the result. The traditional ways of fusion (i.e., low-level, mid-level, and high-level) have been successfully implemented [1-10] so far, but as complexity and amount of data increase along with the complexity of the question in hand, more advanced and sophisticated fusion ways are needed.

Declaration of competing interest

The authors declare that they have no competing financial interests or personal relationships that could appear to influence the work presented in this paper.

Acknowledgements

The present study was supported by the VENI grant, Netherlands organization for scientific research (NWO) no. 016 VENI 178.064.

References

1. Smolinska, A., et al., *Interpretation and visualization of non-linear data fusion in kernel space: study on metabolomic characterization of progression of multiple sclerosis.* PLoS One, 2012. **7**(6).
2. Acar, E., et al., *Forecasting chronic diseases using data fusion.* Journal of proteome research, 2017. **16**(7): p. 2435-2444.
3. Blanchet, L., et al., *Fusion of metabolomics and proteomics data for biomarkers discovery: case study on the experimental autoimmune encephalomyelitis.* BMC bioinformatics, 2011. **12**(1): p. 254.
4. Smilde, A.K., et al., *Fusion of mass spectrometry-based metabolomics data.* Analytical chemistry, 2005. **77**(20): p. 6729-6736.
5. Borràs, E., et al., *Data fusion methodologies for food and beverage authentication and quality assessment–A review.* Analytica Chimica Acta, 2015. **891**: p. 1-14.
6. Silvestri, M., et al., *A mid level data fusion strategy for the Varietal Classification of Lambrusco PDO wines.* Chemometrics and Intelligent Laboratory Systems, 2014. **137**: p. 181-189.
7. Vera, L., et al., *Discrimination and sensory description of beers through data fusion.* Talanta, 2011. **87**: p. 136-142.
8. Sun, W., et al., *Data fusion of near-infrared and mid-infrared spectra for identification of rhubarb.* Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2017. **171**: p. 72-79.
9. Malegori, C., et al., *A modified mid-level data fusion approach on electronic nose and FT-NIR data for evaluating the effect of different storage conditions on rice germ shelf life.* Talanta, 2020. **206**: p. 120208.
10. Márquez, C., et al., *FT-Raman and NIR spectroscopy data fusion strategy for multivariate qualitative analysis of food fraud.* Talanta, 2016. **161**: p. 80-86.
11. Krooshof, P.W., et al., *Visualization and recovery of the (bio) chemical interesting variables in data analysis with support vector machine classification.* Analytical chemistry, 2010. **82**(16): p. 7000-7007.
12. Postma, G., P. Krooshof, and L. Buydens, *Opening the kernel of kernel partial least squares and support vector machines.* Analytica chimica acta, 2011. **705**(1-2): p. 123-134.
13. Gower, J. and S. Harding, *Nonlinear biplots.* Biometrika, 1988. **75**(3): p. 445-455.
14. Upton, G. and I. Cook, *A Dictionary of Statistics 2 rev.* 2008.
15. Breiman, L., *Random Forest.* Machine Learning, 2001. **45**: p. 5-32.
16. Blanchet, L., et al., *Constructing bi-plots for Random Forest: tutorial.* Analytica Chimica Acta, 2020.
17. Baumgart, D.C. and W.J. Sandborn, *Crohn's disease.* The Lancet, 2012. **380**(9853): p. 1590-1605.
18. Tedjo, D.I., et al., *The fecal microbiota as a biomarker for disease activity in Crohn's disease.* Sci Rep, 2016. **6**: p. 35216.
19. Bodelier, A.G., et al., *Volatile Organic Compounds in Exhaled Air as Novel Marker for Disease Activity in Crohn's Disease: A Metabolomic Approach.* Inflamm Bowel Dis, 2015. **21**(8): p. 1776-85.
20. Jansson, J., et al., *Metabolomics reveals metabolic biomarkers of Crohn's disease.* PloS one, 2009. **4**(7).
21. Daniluk, U., et al., *Untargeted metabolomics and inflammatory markers profiling in children with Crohn's disease and ulcerative colitis—A preliminary study.* Inflammatory bowel diseases, 2019. **25**(7): p. 1120-1128.
22. Burbidge, J.B., L. Magee, and A.L. Robb, *Alternative transformations to handle extreme values of the dependent variable.* Journal of the American Statistical Association, 1988. **83**(401): p. 123-127.
23. Smolinska, A., et al., *Simultaneous analysis of plasma and CSF by NMR and hierarchical models fusion.* Analytical and bioanalytical chemistry, 2012. **403**(4): p. 947-959.
24. Eilers, P.H., *A perfect smoother.* Analytical chemistry, 2003. **75**(14): p. 3631-3636.
25. Tomasi, G., F. Van Den Berg, and C. Andersson, *Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data.* Journal of Chemometrics: A Journal of the Chemometrics Society, 2004. **18**(5): p. 231-241.
26. De Meyer, T., et al., *NMR-based characterization of metabolic alterations in hypertension using an adaptive, intelligent binning algorithm.* Analytical chemistry, 2008. **80**(10): p. 3783-3790.

27. Dieterle, F., et al., *Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics.* Analytical chemistry, 2006. **78**(13): p. 4281-4290.
28. Rees, C.A., A. Smolinska, and J.E. Hill, *The volatile metabolome of Klebsiella pneumoniae in human blood.* Journal of breath research, 2016. **10**(2): p. 027101.
29. Smolinska, A., et al., *Current breathomics—a review on data pre-processing techniques and machine learning in metabolomics breath analysis.* Journal of breath research, 2014. **8**(2): p. 027105.
30. Wang, Z.X., Q.P. He, and J. Wang, *Comparison of variable selection methods for PLS-based soft sensor modeling.* Journal of Process Control, 2015. **26**: p. 56-72.
31. Tran, T.N., et al., *Interpretation of variable importance in partial least squares with significance multivariate correlation (sMC).* Chemometrics and Intelligent Laboratory Systems, 2014. **138**: p. 153-160.
32. Bro, R. and A.K. Smilde, *Principal component analysis.* Anal. Methods, 2014. **6**(9): p. 2812-2831.
33. Guyon, I., et al., *Gene selection for cancer classification using support vector machines.* Machine learning, 2002. **46**(1-3): p. 389-422.
34. Wold, S., et al., *The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses.* SIAM Journal on Scientific and Statistical Computing, 1984. **5**(3): p. 735-743.
35. Barker, M. and W. Rayens, *Partial least squares for discrimination.* Journal of Chemometrics, 2003. **17**(3): p. 166-173.
36. Penrose, L.S., *The elementary statistics of majority voting.* Journal of the Royal Statistical Society, 1946. **109**(1): p. 53-57.
37. Lindley, D.V., *Fiducial distributions and Bayes' theorem.* Journal of the Royal Statistical Society. Series B (Methodological), 1958: p. 102-107.
38. Webb-Robertson, B.-J.M., et al., *A Bayesian integration model of high-throughput proteomics and metabolomics data for improved early detection of microbial infections*, in *Biocomputing 2009*. 2009, World Scientific. p. 451-463.
39. Wu, W., et al., *A benchmarking approach for comparing data splitting methods for modeling water resources parameters using artificial neural networks.* Water Resources Research, 2013. **49**(11): p. 7598-7614.
40. Afanador, N.L., et al., *Unsupervised random forest: a tutorial with case studies.* Journal of Chemometrics, 2016. **30**(5): p. 232-241.

Supplementary materials

TABLE S1: ALL DIFFERENT SETS OF WEIGHTS USED IN THE PROXIMITIES STACKING FUSION ATTEMPT [MICROBIOME BLOOD BLOOD-HEADSPACE BREATH]. THE OPTIMAL SET OF WEIGHTS FOUND THROUGH THE 100-CROSS-VALIDATION PROCESS IS HIGHLIGHTED.

[1,0000 0,0000 0,0000 0,0000]	[0,0000 1,0000 0,0000 0,0000]	[0,0000 0,0000 1,0000 0,0000]	[0,0000 0,0000 0,0000 1,0000]
[0,5000 0,4000 0,300 0,7071]	[0,4000 0,5000 0,7071 0,3000]	[0,3000 0,7071 0,5000 0,4000]	[0,7071 0,3000 0,4000 0,5000]
[0,7681 0,1000 0,4000 0,4899]	[0,1000 0,7681 0,4899 0,4000]	[0,4000 0,4899 0,7681 0,1000]	[0,4899 0,4000 0,1000 0,7681]
[0,3872 0,0100 0,2000 0,9000]	[0,0100 0,3872 0,9000 0,2000]	[0,2000 0,9000 0,3872 0,0100]	**[0,9000 0,2000 0,0100 0,3872]**
[0,4000 0,7000 0,0100 0,5916]	[0,7000 0,4000 0,5916 0,0100]	[0,0100 0,5916 0,4000 0,7000]	[0,5916 0,0100 0,7000 0,4000]
[0,5000 0,5000 0,5000 0,5000]	[0,5000 0,5000 0,5000 0,5000]	[0,5000 0,5000 0,5000 0,5000]	[0,5000 0,5000 0,5000 0,5000]
[0,6000 0,1000 0,7000 0,3742]	[0,1000 0,6000 0,3742 0,7000]	[0,7000 0,3742 0,6000 0,1000]	[0,3742 0,7000 0,1000 0,6000]
[0,8000 0,0500 0,5000 0,3279]	[0,0500 0,8000 0,3279 0,5000]	[0,5000 0,3279 0,8000 0,0500]	[0,3279 0,5000 0,0500 0,8000]
[0,2000 0,9000 0,3000 0,2449]	[0,9000 0,2000 0,2449 0,3000]	[0,3000 0,2449 0,2000 0,9000]	[0,2449 0,3000 0,9000 0,2000]
[0,2400 0,8900 0,3877 0,0000]	[0,8900 0,2400 0,0000 0,3877]	[0,3877 0,0000 0,2400 0,8900]	[0,0000 0,3877 0,8900 0,2400]

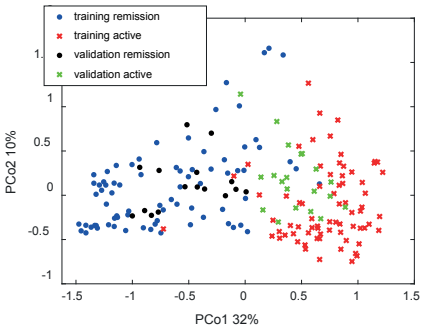


FIGURE S1: SCORE PLOT OF THE TRAINING (I.E. 104) AND VALIDATION (I.E. 26) SAMPLES OF THE RF MODEL IN THE MICROBIOME CASE. THE BLUE DOTS REPRESENT THE REMISSION TRAINING SAMPLES, WHEREAS THE BLACK DOTS REPRESENT THE REMISSION VALIDATION SAMPLES. THE RED CROSSES REPRESENT THE ACTIVE TRAINING SAMPLES, WHILE THE GREENS CROSSES REPRESENT THE ACTIVE VALIDATION SAMPLES.

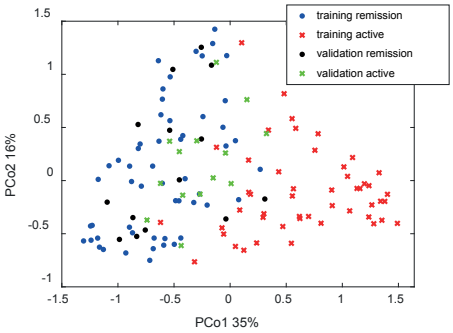


FIGURE S2: SCORE PLOT OF THE TRAINING (I.E. 104) AND VALIDATION (I.E. 26) SAMPLES OF THE RF MODEL IN THE BLOOD CASE. THE BLUE DOTS REPRESENT THE REMISSION TRAINING SAMPLES, WHEREAS THE BLACK DOTS REPRESENT THE REMISSION VALIDATION SAMPLES. THE RED CROSSES REPRESENT THE ACTIVE TRAINING SAMPLES, WHILE THE GREENS CROSSES REPRESENT THE ACTIVE VALIDATION SAMPLES.

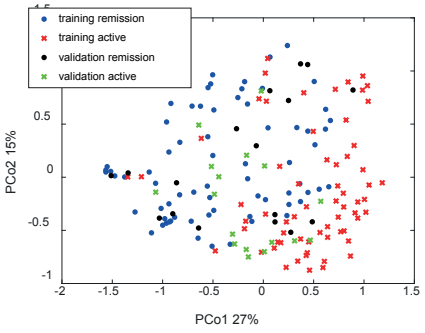


FIGURE S3: SCORE PLOT OF THE TRAINING (I.E. 104) AND VALIDATION (I.E. 26) SAMPLES OF THE RF MODEL IN THE BLOOD HEADSPACE CASE. THE BLUE DOTS REPRESENT THE REMISSION TRAINING SAMPLES, WHEREAS THE BLACK DOTS REPRESENT THE REMISSION VALIDATION SAMPLES. THE RED CROSSES REPRESENT THE ACTIVE TRAINING SAMPLES, WHILE THE GREENS CROSSES REPRESENT THE ACTIVE VALIDATION SAMPLES.

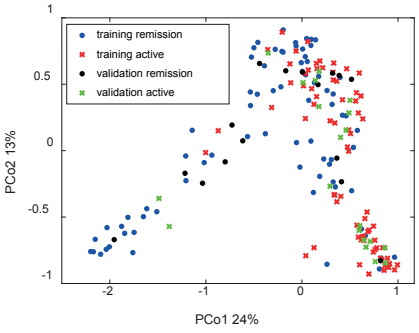


FIGURE S4: SCORE PLOT OF THE TRAINING (I.E. 104) AND VALIDATION (I.E. 26) SAMPLES OF THE RF MODEL IN THE EXHALED BREATH CASE. THE BLUE DOTS REPRESENT THE REMISSION TRAINING SAMPLES, WHEREAS THE BLACK DOTS REPRESENT THE REMISSION VALIDATION SAMPLES. THE RED CROSSES REPRESENT THE ACTIVE TRAINING SAMPLES, WHILE THE GREENS CROSSES REPRESENT THE ACTIVE VALIDATION SAMPLES.

Simulated data

1. Generation, visualization, and optimization of the simulated data

Four data platforms were generated for the simulation analysis; each platform consisted of 250 samples (i.e. 125 samples per class) and 50 variables, and it was generated as follows. Initially, scores data were generated from normally distributed data, given centroid coordinates and specified dispersion (i.e. sigma) values for each class. Then, loadings data were generated from a given range (i.e. [-0.5 0.5] for the first data platform, [-0.2 0.2] for the second data platform, [-0.1 0.1] for the third data platform, and [-0.05 0.05] for the forth data platform), which were then multiplied with the scores data to generate the data platform. Random homoscedastic noise per data platform was also introduced by multiplying each data platform with a random number and then adding this "new" platform to the original one to get the final data platform to be used for the simulation analysis. Finally, five original variables were added at each data platform, which were randomly selected from the four real data platform (i.e. microbiome, breath, blood, blood headspace). Figure S5 A-D illustrates the score plots obtained for all the platforms from principal component analysis (PCA).

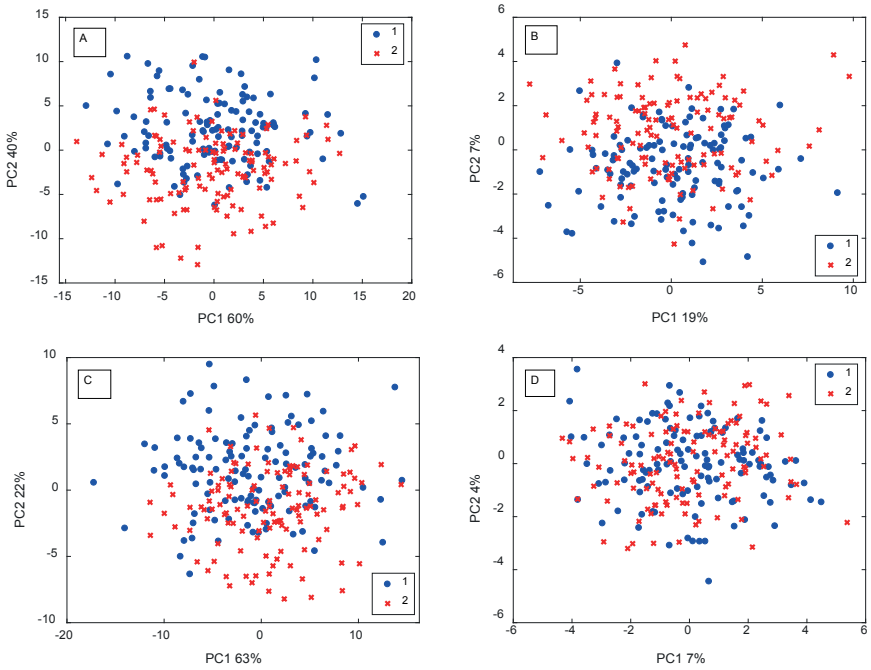


FIGURE S5: SCORE PLOTS OF ALL THE FOUR SIMULATED DATASETS: A) SCORE PLOT OF SIMULATED DATASET 1; B) SCORE PLOT OF SIMULATED DATASET 2; C) SCORE PLOT OF SIMULATED DATASET 3; AND D) SCORE PLOT OF SIMULATED DATASET 4. THE BLUE DOTS REPRESENT CLASS ONE, WHEREAS THE RED CROSSES REPRESENT CLASS TWO. NO REAL GROUPINGS ARE OBSERVED.

Each platform was split into training and validation samples (i.e. 200 samples, of which 100 were class one and the other 100 were class two), and independent internal test set samples (i.e. 50 samples, of which 26 were class one and the remaining 24 were class two). The division between the training and validation, and the independent internal test set samples was achieved by employing the Duplex algorithm. The random forest (RF) model parameters (RF proximities were used in the simulation analysis) and the number of variables kept per platform were optimised as described in section 2.4.1. The statistical methodology applied to the simulated data was exactly the same as the one used for the real world data (sections 2.4.2 and 2.4.3).

2. Results

The optimal number of variables per platform to be used for both individual and fused matrices predictions were 24 for simulated platform one, eight for simulated platform two, 15 for simulated platform three, and nine for simulated platform four. In all four cases, the threshold was set to 50%. Table S2 summarises the classification results of all the RF models built.

TABLE S2: SENSITIVITIES AND SPECIFICITIES OF ALL THE FUSION AND ALL THE INDIVIDUAL PLATFORM CASES FOR THE INDEPENDENT INTERNAL TEST SET. THE NUMBERS IN THE PARENTHESES SHOW THE ACTUAL NUMBER OF THE CORRECTLY CLASSIFIED SAMPLES.

	Sensitivity	**Specificity**
Simulated platform 1	79% (19/24)	80% (21/26)
Simulated platform 2	70% (17/24)	65% (17/26)
Simulated platform 3	67% (14/24)	58% (15/26)
Simulated platform 4	54% (13/24)	50% (13/26)
Mid-level fusion	83% (20/24)	80% (21/26)
High-level fusion	75% (18/24)	85% (22/26)
Proximities stacking fusion	88% (21/24)	85% (22/26)

Figure S6 illustrates the score plots of the training (i.e. 200) and validation (i.e. 50) samples of the RF models in the individual as well as the fused cases. As it can be seen from both Table S2 and Figure S6, individual platforms performed poorly except for platform 1 that demonstrated a good classification accuracy, whereas all three fusion cases performed well. In particular, the proximities stacking fusion approach performed the best with a sensitivity and specificity of 88% and 85%, respectively. Mid-level fusion performed slightly worse with a sensitivity and specificity of 83% and 80%, respectively, and then, high-level fusion followed with a sensitivity and specificity of 75% and 85%, respectively.

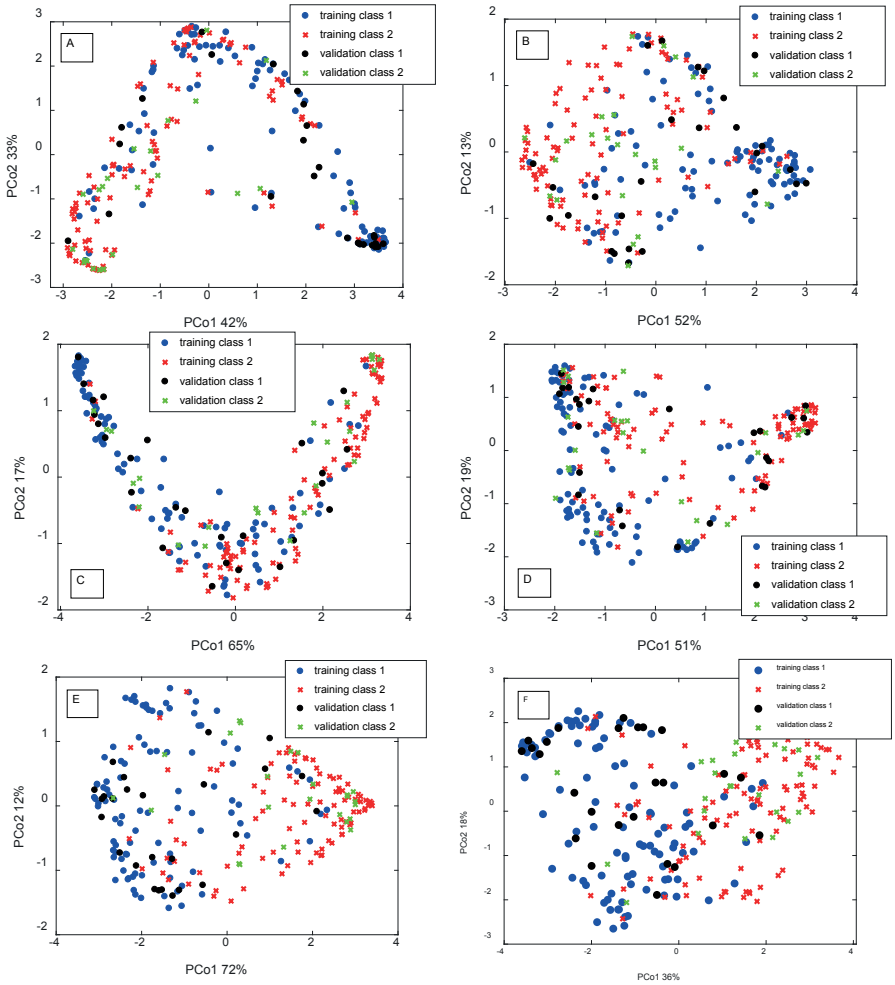


FIGURE S6: A) SCORE PLOT OF THE TRAINING (I.E. 200) AND VALIDATION (I.E. 50) SAMPLES OF THE SIMULATED DATASET 1 RF MODEL; B) SCORE PLOT OF THE TRAINING (I.E. 200) AND VALIDATION (I.E. 50) SAMPLES OF THE SIMULATED DATASET 2 RF MODEL; C) SCORE PLOT OF THE TRAINING (I.E. 200) AND VALIDATION (I.E. 50) SAMPLES OF THE SIMULATED DATASET 3 RF MODEL; D) SCORE PLOT OF THE TRAINING (I.E. 200) AND VALIDATION (I.E. 50) SAMPLES OF THE SIMULATED DATASET 4 RF MODEL; E) SCORE PLOT OF THE TRAINING (I.E. 200) AND VALIDATION (I.E. 50) SAMPLES OF THE MID-LEVEL FUSION CASE RF MODEL; AND F) SCORE PLOT OF THE TRAINING (I.E. 200) AND VALIDATION (I.E. 50) SAMPLES OF THE PROXIMITIES STACKING FUSION CASE RF MODEL. THE BLUE DOTS REPRESENT THE CLASS 1 TRAINING SAMPLES, WHEREAS THE BLACK DOTS REPRESENT THE CLASS 1 VALIDATION SAMPLES. THE RED CROSSES REPRESENT THE CLASS 2 TRAINING SAMPLES, WHILE THE GREENS CROSSES REPRESENT THE CLASS 2 VALIDATION SAMPLES.

The optimal set of weights in the proximities stacking fusion approach weights w was [m_1=0.7000 m_2=0.3742 m_3=0.6000 m_4=0.1000], which shows the contribution of the simulated platforms in the final RF model, respectively. The contribution of each simulated platform provided by the set of weights w was to be expected given the individual platform performances.

CHAPTER 7

Exploring the potential of exhaled breath implementation as a means to diagnose primary sclerosing cholangitis

Georgios Stavropoulos, Kim van Munster, Frederik-Jan van Schooten, Cyriel Ponsioen, Agnieszka Smolinska

In preparation

Abstract

Introduction: Primary sclerosing cholangitis (PSC) is a chronic cholestatic liver disease with multiple stenosis and segmental dilatations of the bile ducts. PSC is a complex phenotype disease with a largely unknown aetiology. It is also an orphan disease, and the only available treatment is liver transplantation. Furthermore, a strong association among PSC and inflammatory bowel disease (IBD) patients has been established. Exhaled breath analysis has gained a lot of attention the last couple of decades, and volatile organic compounds (VOCs) have been identified in exhaled breath as biomarkers for a variety of diseases. Since early detection of PSC is critical to the patient's clinical prospect, non-invasive diagnostic tools are urgently needed. As a first step for such a diagnostic tool, the present study aims to discover VOC biomarkers in exhaled breath that can help distinguish PSC cases from IBD cases.

Methods: In total, 16 PSC, 47 PSC with IBD, and 53 IBD patients were included in the study, and breath and blood samples were acquired at every outpatient clinic visit. The breath samples were analysed for VOCs by using thermal desorption gas chromatography-*time of flight*-mass spectrometry, whereas the blood samples were used to measure blood parameters (i.e. alkaline phosphatase, aspartate aminotransferase, and alanine aminotransferase) that are clinically monitored in PSC patients as liver function indicators. Multivariate statistics were used to conduct the analyses, and model performance was assessed by calculating the sensitivity and specificity for a test set.

Results: Twenty VOCs were used to build a predictive model, which demonstrated a good classification performance by achieving a sensitivity of 77%, and a specificity of 83% for the test set. Combing the 20 VOCs and the three serum parameters yielded a sensitivity of 77% and a specificity of 86%. The 20 VOCs were categorised into four main categories: alkanes, alkenes, ketones, and aldehydes.

Conclusion: The present study demonstrates that exhaled breath can distinguish PSC cases from IBD cases, with a reasonable accuracy and has potential as a non-invasive diagnostic approach. The needs for new advances in the field of PSC alongside the present study results and the latest developments in the field of exhaled breath analysis could stimulate further analyses in the research field of PSC that could potentially lead to a clinical breath test for PSC.

Introduction

Primary sclerosing cholangitis (PSC) is a chronic cholestatic liver disease with multiple stenosis and segmental dilatations of the bile ducts. It is characterized by inflammation and fibrosis of both the intrahepatic and extrahepatic bile ducts that lead to formation of multifocal bile duct strictures. Ultimately, PSC can lead to cirrhosis, liver failure, malignancy, and death [1]. Moreover, PSC is a complex phenotype disease caused by an interplay between genetics and the environment, and its aetiology remains largely unknown. It is also an orphan disease affecting roughly 60.000 individuals in the western world, and it is rather underdiagnosed. Currently, the only available treatment is liver transplantation; oftentimes, side effects occur and a second liver transplantation maybe needed. Furthermore, a strong association among PSC and inflammatory bowel disease (IBD) patients has been established—particularly ulcerative colitis (UC) patients that demonstrate a clinically distinct phenotype [2]. Approximately 80% of PSC patients suffer from IBD as well; however, the opposite does not necessarily hold true since only 5% of IBD patients can develop PSC. On the one hand, this makes the IBD diagnosis easier since the IBD occurrence, most of the times, proceeds the PSC occurrence. On the other hand, this does not help with the PSC diagnosis, also because PSC with IBD cases, typically, show mild to no IBD symptoms, thus leading to PSC under-diagnosis [3,4]. The high prevalence of IBD in PSC patients calls for speculation that IBD present in PSC patients might be a different disease than UC and/or Crohn's disease (CD) that together, they constitute IBD [2]. Loftus et al. [5] suggested that PSC-IBD maybe a distinct IBD phenotype. More specifically, they reported that PSC-IBD is characterised by a high prevalence of pancolitis with rectal sparing and backwash ileitis, and that PSC-IBD patients are at higher risk for colorectal neoplasia. In their analysis, where they adjusted for age, calendar year, and duration of IBD, they found that the presence of PSC was a significant independent risk factor for the development of cancer alone, and that it was also suggestive for the development of colorectal neoplasia. Furthermore, in the same analysis, Loftus et al. also reported that PSC-IBD patients showed a reduced survival rate compared to IBD patients since the presence of PSC was found as a significant independent risk factor for death after adjusting for several potential confounders. Additionally, it is known that gut microorganisms can positively reflect on the development of UC or CD; although it cannot be excluded the fact that these bacterial products or gut-derived bacteria in general, play a direct role in the aetiology of PSC. The high prevalence of IBD in confirmed PSC cases makes it easier to detect IBD, but the very limited prevalence of PSC in confirmed IBD cases makes it more difficult to detect PSC. This results in late treatment for the PSC patients; late diagnosis and start of medication treatment can result in the need for liver transplantation, or eventual death. PSC diagnosis is achieved via a showing of elevated serum parameters (especially alkaline phosphatase) in combination with imaging findings (magnetic resonance cholangiography (MRCP)), where characteristic strictures and/or beading

of either intra- and/or extrahepatic bile ducts are illustrated [3, 6]. The disadvantage of using MRCP is that the bile duct lesions must have progressed to macroscopic morphological abnormalities to become detectable via MRCP, and as a result, MRCP cannot be used to detect early stage of PSC. This makes apparent the need for new, ideally non-invasive and cost-effective, diagnostic tools, where early PSC can be detected and the role of the microbiome present in the GI track can be examined.

Exhaled breath analysis has gained a lot of attention the last couple of decades in various fields of research, and especially, in the research field of medicine, due to its highly promising use as a non-invasive, cost-effective, and easy-to-use diagnostic and monitoring tool [7-10]. More specifically, volatile organic compounds (VOCs) have been identified in exhaled breath as biomarkers for a variety of diseases [10, 11]. It is believed that disease-affected organs (e.g. via inflammation) produce VOCs, which then, due to their volatility, are released into the bloodstream, thus leading to their excretion through the air pathways in breath. The release of these VOCs in other bodily excretions such as urine, blood, or faeces is also possible [11]. These VOCs are called endogenous since they originate from inside the body. Exogenous VOCs, they originate from outside the body, can also be detected. A known example is limonene, which originates from foods, and it has been frequently reported in the literature as a found VOC in breath [12]. Exhaled breath VOC analysis is still in its infancy since no actual VOC breath tests have been implemented in the clinics yet; only the so-called C13 breath tests have been implemented, where C13 isotopes are administered to patients, and then, their breath is captured to measure the emission of the isotope-labelled carbon dioxide [10]. The lack of clinical implementation is, among others, because of standardisation issues that arise when it comes to sampling and analysing breath that can cause batch effects in the data [7] or due to the fact that almost all breath VOC analyses that have been conducted were of small sample size or proof of concept studies. Moreover, an exhaled breath profile can be influenced by many cofounding factors such as smoking, diet or the environment, thus making it even more challenging to generate reproducible and trustworthy results. Technological developments, however, have fuelled the hype for further research in the exhaled breath field since they have allowed for a better sampling (e.g. ReCIVA [13]), storing (e.g. Tenax tubes [13]), and analysing (e.g. gas chromatography-mass spectrometry) of breath samples [11].

Liver diseases have been investigated by means of breath VOC analysis, and various biomarkers have been reported [10]. PSC, however, has not been one of the investigated diseases; it has only been examined by means of VOC analysis in the bile [14] and urine [15], where it showed promising results. Therefore, exhaled breath might be a possibility when it comes to diagnosing early stages of PSC cases and differentiating them from IBD cases. It is hypothesized that on-going inflammation, probably originating in the colon (i.e. the “leaky-gut” theory [16]), supports bile duct

inflammation in every PSC case (e.g. from no or minor symptoms to severe cholestasis and/or portal hypertension). Consequently, particular molecules may appear in breath samples, faeces samples, as well as blood samples. In parallel, gut microbiome changes may occur too and thus, system approaches implemented on these samples might prove themselves useful in obtaining scientific insight of the disease at first, and developing screening tools for early detection of the disease at second. Such detectable compounds may be either sulphur-based compounds in breath and faeces or metabolites in faecal microbiome and blood. Possible identification of these compounds may lead to the development of such a tool for PSC disease. Consequently, the present study aims to identify VOC biomarkers in exhaled breath that can help distinguish PSC cases (i.e. either PSC or PSC with IBD cases) from IBD cases, and to compare these VOCs performance to blood parameters that are also currently used in the diagnosis of PSC, such as alkaline phosphatase. Confounding factors that can influence the breath profile of the subjects such as age, smoking, diet, supplements, medication, and gender are also examined.

Materials and methods

Patient inclusion

The present study recruited individuals that suffer from PSC, PSC and IBD, and IBD during a one-year period at the Amsterdam Medical Centre (AMC) in Amsterdam, the Netherlands. The PSC and PSC with IBD groups were recruited at the PSC expertise centrum at the AMC, whereas the IBD group was recruited at the outpatient clinic at the AMC. The IBD group was consisted of both UC and CD cases, and it was used as the control group since the aim was to discriminate any PSC case from IBD cases. In total, 16 PSC, 47 PSC with IBD, and 53 IBD patients were included in the study. Inclusion criteria for the PSC, PSC with IBD, and IBD patients were an established PSC diagnosis based on the EASL criteria [17] for the PSC and PSC with IBD groups, an established UC diagnosis based on ECCO [18] guidelines for the IBD group, an age range from 18 to 65 years old, and a BMI range from 19 to 30. Exclusion criteria for all three groups were: unable to provide informed consent, the presence of any disease that compromises the immune system such as HIV positive or organ transplantation, the presence of any other liver disease, the presence of active or untreated tuberculosis, the presence of ileo-anal pouch, and the use of chemotherapy agents. For the IBD group, an extra exclusion criterion was applied, which was abnormality in liver tests such as elevated alkaline phosphatase or transaminases. The study was approved by the Institutional Review Board (IRB) of the Amsterdam University Medical Centre (AUMC).

Sampling and data acquisition

Breath and blood samples were acquired at every outpatient clinic visit. For the breath sampling, the ReCIVA (Owlstone Medical, Cambridge, UK) breath sampler connected to a CASPER air pump (Owlstone Medical, Cambridge, UK) was used [13]. Briefly, the CASPER pump takes ambient air and passes it through an air filter unit before supplying it via a plastic tube to the ReCIVA; the subject breathes in filtered air. That way the level of background VOCs from the surroundings that are present in the collected sample are significantly reduced. ReCIVA is able to sample breath fractions (e.g. alveolar air) based on the subject's breathing profile. Each subject had to breathe for approximately five minutes in the device, and two samples per subject were collected into airtight-capped stainless steel carbon-filled sorption tubes Tenax/Carbograph-5TD TD tubes (Markes International Ltd, Llantrisant, UK) [13]. Two hours before sampling, the subjects were overnight fasted. All patients were sampled in duplicates at the same location to prevent background bias, and all samples were stored at 5°C until the analysis took place. The breath samples were analysed in terms of VOCs, and the analysis was conducted by using thermal desorption gas chromatography-*time of flight*-mass spectrometry (TD-GC-*tof*-MS). In short, VOCs are separated via GC, and then, they are identified via *tof*-MS. The experimental settings of the GC-*tof*-MS are described elsewhere [19], and the VOC identification was achieved by using the NIST Mass Spectral Search Program v2.3. Moreover, it should be noted that an internal standard (i.e. Bromobenzene-D5) was injected in every sample before measuring and quality controls (SUPELCO Analytical; reference number 44589) were also run in between the breath samples throughout the GC-*tof*-MS runs to ensure and monitor a good analysis quality [17].

Participants donated blood as part of their regular blood testing procedure. These blood samples were used to measure blood parameters such as alkaline phosphatase, aspartate aminotransferase, and alanine aminotransferase; these parameters are clinically monitored in PSC patients because they are used as liver function indicators. Bilirubin was also measured in the IBD group. Lastly, patients also filled in a questionnaire regarding BMI, smoking, diet, supplements, medication, disease activity for PSC (Amsterdam cholestatic complains score), UC (simple clinical colitis activity index), and CD (Harvey Bradshaw index). The demographic data were used to check their influence, if any, on the subjects' breath composition and on the found VOCs—all patient characteristics are summarised in Table 1.

TABLE 6: PATIENT CHARACTERISTICS INCLUDED IN THE PRESENT STUDY. THE PSC AND PSC/IBD PATIENT WERE CONSIDERED AS ONE CLASS AND THE IBD AS THE OTHER. FOR ALP, AST, AND ALT, THE MEDIAN VALUES ARE SHOWN. NO SIGNIFICANCE WAS FOUND REGARDING PATIENTS CHARACTERISTICS AND COHORTS.

	PSC (N = 16)	PSC/IBD (N = 47)	IBD (N = 53)
Age (mean)	50	47	46.6
Gender	8/8 (F/M)	14/33 (F/M)	25/28 (F/M)
Smoking	10/0/6 (no/yes/quit < 2 years ago)	35/1/11 (no/yes/quit < 2 years ago)	44/4/5 (no/yes/quit < 2 years ago)
Diet	12/4 (no/yes)	42/5 (no/yes)	44/9 (no/yes)
Medication	2/14 (no/yes)	2/45 (no/yes)	5/48 (no/yes)
Ursodiol	1/13/2 (no/yes/NA)	12/32/3 (no/yes/NA)	46/0/7 (no/yes/NA)
Corticosteroids	14/0/2 (no/yes/NA)	40/3/4 (no/yes/NA)	44/4/5 (no/yes/NA)
Thiopurines	14/0/2 (no/yes/NA)	40/3/4 (no/yes/NA)	44/4/5 (no/yes/NA)
Biologicals	14/0/2 (no/yes/NA)	37/6/4 (no/yes/NA)	30/18/5 (no/yes/NA)
Supplements	9/7 (no/yes)	33/14 (no/yes)	32/21 (no/yes)
Alkane phosphatase (ALP)	128 (65 – 539)	151 (60 – 1104)	78 (32 – 109)
Aspartate aminotransferase (AST)	39 (17 – 91)	29 (18 – 266)	24 (16 – 72)
Alanine aminotransferase (ALT)	58 (12 – 141)	36 (16 – 455)	25 (11 – 75)

Data handling and statistical modelling

The present study followed a so-called semi-targeted approach rather than an untargeted, which is the approach of choice when it comes to VOC analysis. Untargeted approach means that one uses the whole chromatograph and blindly tries to find VOCs that might be of interest for the question at hand, whereas targeted means that one targets specific compounds in the chromatographs that are known already to be of interest. Semi-targeted is defined in the present study and considered the approach where one targets specific compounds that, based on a priori knowledge or hypothesis, might be of interest for the question at hand. The present study focused on the compounds that have been reported in the literature to be related to liver impairment—Table 2 shows all the VOCs that were selected for examination. Recently, Stavropoulos et al. [10] reviewed all the available literature on liver diseases examined by means of VOC analysis, and they reported all the compounds that have been found to be significantly related to liver impairment. More specifically, they published a table that reports all the VOCs that have been found in more than one studies (e.g. dimethyl-sulphide); the present study based its compound selection on the Stavropoulos et al. results. Moreover, the present study examined some aldehydes (mentioned in Table 2) that were not reported by Stavropoulos et al. [10]. These aldehydes are products of lipid peroxidation due to reactive oxygen species production and inflammation, and they been previously related to liver impairment [10].

The selected compound peaks were integrated through Xcalibur v2.2 SP1.48 once all the chromatographs were generated. An integration method was created by manually going through all the chromatograms and locating the peaks, and then, pinpointing where the compounds of interests were (i.e. within a thirty-second time range). Peak areas of each compound of interest was then obtained using characteristic mass fractions. The integration method was then used to preprocess the raw chromatographs and generate an Excel spreadsheet that contained the peak values of each VOC for every sample. Next, the data were normalised with the internal standard peak, and then, the internal standard peak was removed from the dataset and a logarithmic transformation of the data followed before the statistical modelling process begun. Logarithmic transformation accounts for data heteroscedasticity and skewness [20].

TABLE 7: COMPOUNDS SELECTED AS POSSIBLE TARGETS TO DISTINGUISH PSC AND PSC/IBD FROM IBD PATIENTS.

Acetaldehyde	2-Butanone	2-Nonene	Undecane
Ethanol	Hexane	2-Octanone	Nonanal
Acetone	Benzene	Heptanal	Dodecane
Pentane	Pentanal	Beta-pinene	Decanal
Isoprene	2-Pentanone	Alpha-pinene	Tridecane
2-Propanol	Hexanal	Benzaldehyde	Indole
Dimethyl-sulphide	Octane	Decane	Undecanal
Carbon-disulphide	Nonane	Octanal	
Butanal	Styrene	Limonene	

The statistical modelling process included, first, data exploration via unsupervised random forest (URF) [21] to see whether any groups existed within the data, and second, sample classification via random forest (RF) [22] to find compounds enabling differentiating PSC cases from IBD cases. The data were split into training (i.e. 80% of the data) and test (i.e. the remaining 20% of the data) sets before any supervised analysis was conducted—the kenstone algorithm [23] was used to split the data since it selects objects to model sets such that they are uniformly scattered over the whole experimental space. The test set was only used in the end when the final RF model was built to validate it. The RF model parameters (i.e. number of trees, predictors, number of splits per tree and samples per tree terminal leaf per RF model), as well as to the number of variables/VOCs to be kept for the final classification model were optimized within a 1000-iteration loop. A detailed description of how this 1000-iteration loop is performed is described elsewhere [24]. In the present study, 1000 trees per model were used, the minimum number of samples per tree leaf for every tree in each model was set to six, and the maximum number of splits was set to five. The variables that

appeared as important in at least 10 from the 1000 iterations were kept for further analysis. The kept variables were also checked for possible significance between the two classes (i.e. PSC and PSC/IBD vs IBD) by using the Wilcoxon signed rank test; it tests the null hypothesis that data in the two populations are samples coming from continuous distributions with equal medians. Ultimately, the final and optimized RF model was built by using the training samples to predict the test set samples. Its performance was assessed by calculating the sensitivity and specificity for the test set; a receiver operating characteristic (ROC) curve was also plotted to calculate the area under the curve (AUC) and to visualise the model performance.

Additionally, for comparison purposes, a classification RF model was also built by using the three serum parameters (alkaline phosphatase, aspartate aminotransferase, and alanine aminotransferase) only, as well as a RF model that used both the serum parameters and the selected VOCs that were found as described above. No further optimisation was done for these two extra models; the settings used for these two models were the same as the ones found above, and their performance was assessed by calculating the sensitivity and specificity for the test set. It should be mentioned, however, that for some of the patients, the serum parameter values were missing. These missing values were dealt with by using surrogate splits, where the algorithm sends the sample to the left or right child node of the missing variable using the best surrogate predictor. Furthermore, regularised multivariate analysis of variance (rMANOVA) [236] was implemented to test for significance of confounding factors (i.e. smoking, diet, medication, age, supplements, and gender) considering the two classes (i.e. PSC and PSC/IBD vs. IBD). All data analyses were performed by using MatLab R2016b version—the Statistics and Machine Learning Toolbox. For the RF models, the TreeBagger function was used, whereas for the URF model and rMANOVA, the codes were found elsewhere [21, 25].

Results

VOC analysis

In total, 16 PSC, 47 PSC with IBD, and 53 IBD patients were sampled and included in the study analysis. The PSC and PSC/IBD patient samples were considered as one class and the IBD as the other. Each patient was sampled in duplicates, and therefore, the total number of measurements to be used in the analysis was 234—a couple of samples were accidently measured twice in the GC-*tof*-MS. From these 234 measurements, 173 measurements (93 PSC and PSC/IBD and 80 IBD) were used as the training set, and the remaining 61 (35 PSC and PSC/IBD and 26 IBD) were used as the test set—measurements coming from the same patient were kept in either the training or the test set to avoid overestimation of the model. The number of variables/VOCs in the dataset was 34 (Table 2). The optimal number of variables (see section 2.3)

to be used for building the final classification RF model was 20 (Table 3). These 20 compounds were also checked for possible significance between the two classes by using the Wilcoxon signed rank test (Table 3); Bonferroni correction for multiple testing was also performed by dividing the given from the test p-values for each compound by the number of samples in the training set. Eight compounds were significantly different: acetone, hexanal, octane, 2-octanone, decane, undecane, dodecane, and decanal. The concentration of acetone and hexanal decreased in the IBD class, whereas the concentration of the other six compounds increased in the IBD class.

TABLE 8: OVERVIEW OF THE VOCS FOUND AS IMPORTANT IN THE 1000-ITERATION PROCEDURE TO BE USED FOR BUILDING THE FINAL RF CLASSIFICATION MODEL. THE ASTERISK NEXT TO A COMPOUND INDICATES WHETHER THE COMPOUND IS SIGNIFICANTLY DIFFERENT BETWEEN THE TWO CLASSES; BONFERRONI CORRECTION FOR MULTIPLE TESTING WAS PERFORMED. FOR THE SIGNIFICANT COMPOUNDS, IT IS ALSO INDICATED WHETHER THEIR RELATIVE ABUNDANCE INCREASED OR DECREASED WHEN COMPARED TO ITS PSC AND PSC/IBD GROUP ABUNDANCE.

Ethanol	↓ 2-Octanone*
↑ Acetone*	Alpha-pinene
Pentane	Benzaldehyde
Isoprene	↓ Decane*
Carbon-disulphide	Limonene
Pentanal	↓ Undecane*
2-Pentanone	↓ Dodecane*
↑ Hexanal*	↓ Decanal*
↓ Octane*	Tridecane
Nonane	Undecanal

Exploratory analysis of the dataset via URF did not show any underlying groupings in the data (results not shown), whereas the final RF predictive model demonstrated a good classification performance by achieving a sensitivity of 77%, a specificity of 83%, and an AUC of 0.8352 for the test set. Figure 1 shows the ROC curve of the model for the test set, and Figure 2 shows the score plot based on the proximities of the samples.

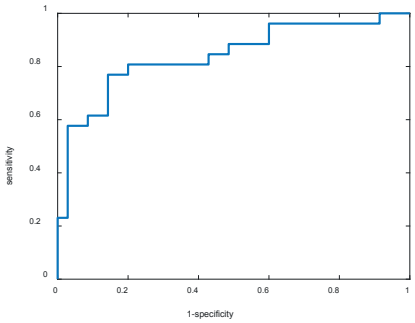


FIGURE 11: ROC CURVE OF THE TEST SET; AUC = 0.8352, SENSITIVITY = 77%, AND SPECIFICITY = 83%.

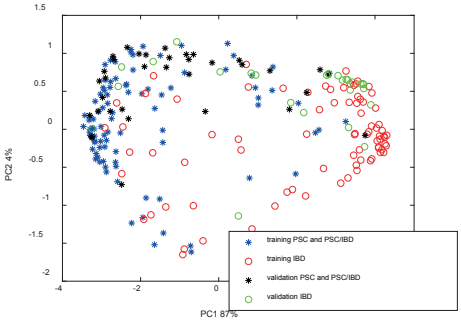


FIGURE 12: SCORE PLOT BASED ON THE PROXIMITIES OF THE TRAINING AND TEST SETS.

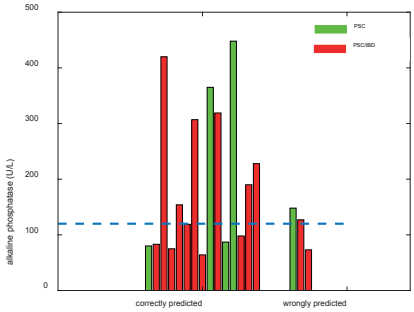


FIGURE 13: VISUALIZATION OF THE FACT THAT THE CHOSEN BREATH PROFILE CAN ALSO CORRECTLY PREDICT SAMPLES THAT HAVE ALKALINE PHOSPHATASE VALUES BELOW THE 120 U/L THRESHOLD USED IN THE CLINICS AS AN INDICATION FOR THE PSC PRESENCE.

VOC, serum, and cofounding factor analysis

Furthermore, an RF classification model using the three serum parameters (using the same training and test samples) yielded a sensitivity of 38% and a specificity of 63%, respectively for the test set samples. Finally, an RF model that used both the 20 VOCs and the three serum parameters yielded a sensitivity of 77% and a specificity of 86%. Since the serum samples were not in duplicates, and therefore, the breath measurements had to be averaged out per patients to match the number of the serum samples before the data fusion. A further investigation on the PSC and PSC/IBD test set samples with respect to their alkaline phosphatase serum parameter value was also performed, and as it can be seen in Figure 3, the VOC profile correctly predicted samples that had alkaline phosphatase value either below or above 120 U/L.

Possible significance of the confounding factors that might have interfered with the study results was examined. No significant differences were found for any of the confounding factors with respect to the two classes except for the ursodiol

medication. However, this result was attributed to the classes (i.e. PSC and PSC/IBD vs IBD)—to statistically confirm this, rMANOVA was run considering only the two classes, and then again, the result was significant.

Discussion

The present study investigated the potential clinical application of the exhaled breath in identifying PSC cases amongst IBD cases. A breath VOC profile was able to identify PSC cases amongst IBD cases; it also outperformed the predictive performance of three serum parameters that are currently used in the clinics to diagnose the presence of PSC. Possible statistical significance of confounding factors that may influence the breath VOC profile of the PSC cases was also checked.

A breath profile of 20 VOCs (Table 3) was able to identify PSC cases amongst IBD cases with a sensitivity of 77% and a specificity of 83% for an independent test set (Figure 1 and Figure 2). A predictive model that used only the three serum parameters achieved a sensitivity of 38% and a specificity of 63% for the same independent test set. At the same time, when both the VOCs and the serum parameters were combined/fused to build another predictive model, they yielded a sensitivity of 77% and a specificity of 86% for the same independent test set. As it has been highlighted in the literature already, separating PSC cases from IBD cases is of crucial importance because, oftentimes, the PSC cases remain under-diagnosed [3, 4]. The high prevalence of IBD in confirmed PSC cases makes it easier to detect IBD, but the very limited prevalence of PSC in confirmed IBD cases makes it more difficult to detect PSC. This results in late treatment for the PSC patients; late diagnosis and start of medication treatment can result in the need for liver transplantation, or eventual death. The found breath VOC profile could be used in clinical settings to detect PSC when confirmed IBD cases come to the outpatient clinic as a first screening tool, and thus, leading to an earlier PSC detection. A further investigation on the PSC and PSC/IBD test set samples with respect to their alkaline phosphatase value was also performed since its value is clinically connected to the presence of PSC. A common practice in clinical settings [26] is the use of a threshold of 120 units per litre (U/L) to indicate PSC presence—below 120 U/L indicates no PSC presence (normal liver function), whereas above 120 U/L indicates PSC presence (abnormal liver function). However, oftentimes, the alkaline phosphatase is below 120 U/L even though the individual suffers from PSC [16]. Figure 3 shows that PSC and PSC/IBD samples with lower than 120 U/L alkaline phosphatase value were also correctly predicted by the VOC model, proving that the found breath VOC profile could potentially serve as a better screening tool than alkaline phosphatase. Additionally, sampling breath is more patient-friendly and non-invasive, whereas sampling blood might cause patient discomfort since it is invasive.

The 20 VOCs identified in the present study can be categorised into four main categories: alkanes, alkenes, ketones, and aldehydes. No striking findings were seen here since all the VOCs used in the present study are known to be connected to liver impairment [10]; however, the absence of dimethyl-sulphide can be characterised as a striking one since it has been repeatedly reported in VOC analyses that examined liver diseases as a result of incomplete metabolism of sulphur-containing amino acids in the transamination pathway. The presence of limonene can be characterised as something to be expected since it originates from foods and drinks. More specifically, limonene is metabolised by the P450 enzymes CYP2C9 and CYP2C19 into other compounds such as perillyl alcohol, trans-isoperitenol, and trans-carveol [10]; in liver impairment, these enzymes are reduced, leading to increased amounts of limonene in the body. Finally, another compound that deserves to be named separately is isoprene, and this is because literature is conflicted when it comes to the isoprene origin; it can either originate from impairment in the cholesterol biosynthesis pathway or it can be the result of disturbed colon flora or it can be the result of exercise [10]. The present study participants were deprived from exercising two hours before sampling; however, this does not necessarily exclude the fact that isoprene might have originated from exercising since it can still be stored in muscle compartments and get released later [27]. As far as the remaining of the VOCs identified here is concerned, it has generally been proposed that lipid peroxidation, a process triggered by increased inflammation in diseased liver, generates alkanes and long-chain aldehydes, and both groups can be converted into alcohols or ketones by CYPs or aldo-keto reductases, respectively [10]. Additionally, these 20 compounds were also tested for possible significant differences between the two classes in a univariate way by using the Wilcoxon signed rank test. It was found that acetone and hexanal had significantly increased abundance in the PSC class, whereas octane, 2-octanone, decane, undecane, dodecane, and decanal had significantly decreased abundance in the PSC class. Hexanal is considered a stable breakdown product of lipid peroxidation, which is formed because of oxygen free radical (OFR) activity. OFRs are considered responsible for liver damage, and hexanal has also been confirmed as cytotoxic to most cells [28]. Acetone is one that has been repeatedly reported in the literature. The increased acetone abundance found in the present study coincides with what has been reported in almost all cases where acetone was identified as related to liver impairment [10]; hepatic insulin resistance is believed to lead to increased triglycerides, free fatty acids, and ketones such as acetone. VOC by-products of lipid peroxidation such as alkanes and aldehydes have also been linked, in increased abundances, to IBD patient classification [8]. The fact that six VOCs were significantly increased in the IBD group could suggest that it is a VOC pattern or profile rather than a single VOC that could represent metabolic changes.

Exhaled breath profile can be influenced by a variety of factors, and therefore, these factors should be accounted or corrected for to ensure an unbiased VOC profile. Such confounding factors are age, gender, smoking, diet, medication, and supplements.

Additionally, environmental or instrumental artefacts can influence the breath profiles as well. The present study controlled as much as possible for environmental artefacts by sampling all the participants at the same location, using the same equipment, and by the same personnel. The instrumental artefacts were controlled by using the same equipment to measure the samples, the analyses were run by the same personnel, and by using quality controls in regular intervals throughout the measuring process of the samples in the GC-*tof*-MS runs. The use of quality controls shows whether there is non-biological variation introduced in the data due to temperature shift or column aging that may have caused peak shifting [7]. As far as age, gender, smoking, diet, medication, and supplements are concerned, rMANOVA was used to check whether there is a significant difference in the study population. No significant differences were found for any of the aforementioned possible confounding factors (Table S1) except for the ursodiol medication. This result was to be expected since the vast majority of the IBD patients do not take this medication, and the vast majority of the PSC and PSC/IBD patients do take this medication. To statistically confirm this, an rMANOVA was also run considering only the two classes. The result was significant, and thus, it confirmed that the ursodiol medication significant result was indeed due to the classes, and not due to the medication itself. Therefore, no significance for any of the confounding factors was found.

The present study demonstrates that exhaled breath can be potentially implemented in the clinics as a means of diagnosing PSC cases from IBD cases, something that, to date, remains a challenge in clinical settings, and this supports the novelty of the study since, to the best of the authors' knowledge, such a study has not been performed before. Another strength of the present study is the fact that care was taken to prevent all possible non-biological variations to bias the study data, and the fact that the study results were validated by an independent test set. To date, most of the exhaled breath VOC analyses that have been conducted in liver diseases were not independently validated due to their small sample size or they did not account for various confounding factors that could have influenced their results [10]. All these alongside the positive study results can stimulate further research to be conducted on PSC examined via breath VOC, which may, eventually, lead to a clinically applicable breath test for PSC patients. The present study, however, also demonstrates some limitations that have to be addressed here too. The study population should have been larger even though the population used cannot be considered small. Moreover, further test of the model and the found VOC profile with an external independent test set that would have been measured at a later stage after the study samples were taken could have been performed too. Ideally, the present study should have also examined PSC cases (without IBD) against IBD cases and PSC cases against PSC/IBD cases. However, these comparisons were not possible to perform due to the limited number of PSC samples. These comparisons along with a further test of the found VOC profile should be considered for further research in the topic.

In conclusion, the present study demonstrates the possibility of using exhaled breath to diagnose PSC patients from IBD patients, something that has been challenging clinicians to date. This has led to under-diagnosis of PSC, and therefore, mistreatment of the PSC patients. Breath analysis is still in its infancy and far from being implemented in the clinics yet. However, the apparent needs for new advances in the field of PSC alongside the present study results and the latest developments in the field of exhaled breath analysis could be the catalyst to stimulate further analyses in the research field of PSC that could potentially lead to a clinical breath test for PSC.

Declaration of competing interest

The authors declare that they have no competing financial interests or personal relationships that could appear to influence the work presented in this paper.

Acknowledgements

The present study was supported by the VENI grant, Netherlands organization for scientific research (NWO) no. 016 VENI 178.064.

References

1. Williamson, K.D. and R.W. Chapman, *Primary sclerosing cholangitis: a clinical update.* British medical bulletin, 2015. **114**(1): p. 53-64.
2. Tsaitas, C., A. Semertzidou, and E. Sinakos, *Update on inflammatory bowel disease in patients with primary sclerosing cholangitis.* World journal of hepatology, 2014. **6**(4): p. 178.
3. Sirpal, S. and N. Chandok, *Primary sclerosing cholangitis: diagnostic and management challenges.* Clinical and experimental gastroenterology, 2017. **10**: p. 265.
4. Palmela, C., et al., *Inflammatory bowel disease and primary sclerosing cholangitis: a review of the phenotype and associated specific features.* Gut and liver, 2018. **12**(1): p. 17.
5. Loftus, E., et al., *PSC-IBD: a unique form of inflammatory bowel disease associated with primary sclerosing cholangitis.* Gut, 2005. **54**(1): p. 91-96.
6. Dave, M., et al., *Primary sclerosing cholangitis: meta-analysis of diagnostic performance of MR cholangiopancreatography.* Radiology, 2010. **256**(2): p. 387-396.
7. Stavropoulos, G., et al., *Implementation of quality controls is essential to prevent batch effects in breathomics data and allow for cross-study comparisons.* Journal of Breath Research, 2020. **14**(2): p. 026012.
8. Bodelier, A.G., et al., *Volatile Organic Compounds in Exhaled Air as Novel Marker for Disease Activity in Crohn's Disease: A Metabolomic Approach.* Inflamm Bowel Dis, 2015. **21**(8): p. 1776-85.
9. Smolinska, A., et al., *The potential of volatile organic compounds for the detection of active disease in patients with ulcerative colitis.* Aliment Pharmacol Ther, 2017. **45**(9): p. 1244-1254.
10. Stavropoulos, G., et al., *Liver Impairment—The Potential Application of Volatile Organic Compounds in Hepatology.* Metabolites, 2021. **11**(9): p. 618.
11. Beauchamp, J., C. Davis, and J. Pleil, *Breathborne biomarkers and the human volatilome.* 2020.
12. Ferrandino, G., et al., *Breath Biopsy Assessment of Liver Disease Using an Exogenous Volatile Organic Compound—Toward Improved Detection of Liver Impairment.* Clinical and translational gastroenterology, 2020. **11**(9).
13. Doran, S.L., A. Romano, and G.B. Hanna, *Optimisation of sampling parameters for standardised exhaled breath sampling.* Journal of breath research, 2017. **12**(1): p. 016007.
14. Navaneethan, U., et al., *Volatile organic compounds in bile for early diagnosis of cholangiocarcinoma in patients with primary sclerosing cholangitis: a pilot study.* Gastrointestinal endoscopy, 2015. **81**(4): p. 943-949. e1.
15. Navaneethan, U., et al., *Volatile organic compounds in urine for noninvasive diagnosis of malignant biliary strictures: a pilot study.* Digestive diseases and sciences, 2015. **60**(7): p. 2150-2157.
16. Tabibian, J.H. and K.D. Lindor, *Primary sclerosing cholangitis: a review and update on therapeutic developments.* Expert review of gastroenterology & hepatology, 2013. **7**(2): p. 103-114.
17. Liver, E.A.F.T.S.O.T., *EASL Clinical Practice Guidelines: management of cholestatic liver diseases.* Journal of hepatology, 2009. **51**(2): p. 237-267.
18. Magro, F., et al., *Third European Evidence-based Consensus on Diagnosis and Management of Ulcerative Colitis. Part 1: Definitions, Diagnosis, Extra-intestinal Manifestations, Pregnancy, Cancer Surveillance, Surgery, and Ileo-anal Pouch Disorders.* Journal of Crohn's and Colitis, 2017. **11**(6): p. 649-670.
19. Fijten, R.R.R., et al., *The necessity of external validation in exhaled breath research: a case study of sarcoidosis.* J Breath Res, 2017. **12**(1): p. 016004.
20. Stavropoulos, G., et al., *Preprocessing and analysis of volatilome data*, in *Breathborne Biomarkers and the Human Volatilome*. 2020, Elsevier. p. 633-647.
21. Afanador, N.L., et al., *Unsupervised random forest: a tutorial with case studies.* Journal of Chemometrics, 2016. **30**(5): p. 232-241.
22. Breiman, L., *Random Forest.* Machine Learning, 2001. **45**: p. 5-32.
23. Kennard, R.W. and L.A. Stone, *Computer aided design of experiments.* Technometrics, 1969. **11**(1): p. 137-148.

24. Stavropoulos, G., et al., *Advanced data fusion: Random forest proximities and pseudo-sample principle towards increased prediction accuracy and variable interpretation.* Analytica Chimica Acta, 2021: p. 339001.
25. Engel, J., et al., *Regularized MANOVA (rMANOVA) in untargeted metabolomics.* Analytica chimica acta, 2015. **899**: p. 1-12.
26. Cappello, M., et al., *Liver function test abnormalities in patients with inflammatory bowel diseases: a hospital-based survey.* Clinical Medicine Insights: Gastroenterology, 2014. **7**: p. CGast. S13125.
27. Sukul, P., et al., *Deficiency and absence of endogenous isoprene in adults, disqualified its putative origin.* Heliyon, 2021. **7**(1): p. e05922.

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.

References

1. Amundsen, T., et al., Can dogs smell lung cancer? First study using exhaled breath and urine screening in unselected patients with suspected lung cancer. Acta oncologica, 2014. 53(3): p. 307-315.
2. Cornu, J.-N., et al., Olfactory detection of prostate cancer by dogs sniffing urine: a step forward in early diagnosis. European urology, 2011. 59(2): p. 197-201.
3. Willis, C.M., et al., Olfactory detection of human bladder cancer by dogs: proof of principle study. Bmj, 2004. 329(7468): p. 712.
4. Cambau, E. and M. Poljak, Sniffing animals as a diagnostic tool in infectious diseases. Clinical Microbiology and Infection, 2020. 26(4): p. 431-435.
5. George, A., The woman who can smell Parkinson's. New Scientist, 2019. 241(3220): p. 40-41.
6. Ajibola, O.A., et al., Effects of dietary nutrients on volatile breath metabolites. Journal of nutritional science, 2013. 2.
7. Blanchet, L., et al., Factors that influence the volatile organic compound content in human breath. Journal of breath research, 2017. 11(1): p. 016013.
8. Stavropoulos, G., et al., Preprocessing and analysis of volatilome data, in Breathborne Biomarkers and the Human Volatilome. 2020, Elsevier. p. 633-647.
9. Bogen, E., The diagnosis of drunkenness—a quantitative study of acute alcoholic intoxication. California and western medicine, 1927. 26(6): p. 778.
10. Haworth, J.J., et al., Breathing new life into clinical testing and diagnostics: perspectives on volatile biomarkers from breath. Critical Reviews in Clinical Laboratory Sciences, 2022: p. 1-20.
11. Ghoshal, U.C., How to interpret hydrogen breath tests. Journal of neurogastroenterology and motility, 2011. 17(3): p. 312.
12. Abdullah, T. and A. Ahmet. Genomics analyser: a big data framework for analysing genomics data. in Proceedings of the Fourth IEEE/ACM International Conference on Big Data Computing, Applications and Technologies. 2017.
13. Federico, A., et al., Transcriptomics in toxicogenomics, part II: preprocessing and differential expression analysis for high quality data. Nanomaterials, 2020. 10(5): p. 903.
14. Kinaret, P.A.S., et al., Transcriptomics in toxicogenomics, part I: experimental design, technologies, publicly available data, and regulatory aspects. Nanomaterials, 2020. 10(4): p. 750.
15. Serra, A., et al., Transcriptomics in toxicogenomics, part III: data modelling for risk assessment. Nanomaterials, 2020. 10(4): p. 708.
16. Efstathiou, G., et al., ProteoSign: an end-user online differential proteomics statistical analysis platform. Nucleic acids research, 2017. 45(W1): p. W300-W306.
17. Gowda, H., et al., Interactive XCMS Online: simplifying advanced metabolomic data processing and subsequent statistical analyses. Analytical chemistry, 2014. 86(14): p. 6931-6939.
18. Stavropoulos, G., et al., Implementation of quality controls is essential to prevent batch effects in breathomics data and allow for cross-study comparisons. Journal of Breath Research, 2020. 14(2): p. 026012.
19. Stavropoulos, G., et al., Advanced data fusion: Random forest proximities and pseudo-sample principle towards increased prediction accuracy and variable interpretation. Analytica Chimica Acta, 2021: p. 339001.
20. Athey, S., The impact of machine learning on economics, in The economics of artificial intelligence: An agenda. 2018, University of Chicago Press. p. 507-547.
21. Ghoddusi, H., G.G. Creamer, and N. Rafizadeh, Machine learning in energy economics and finance: A review. Energy Economics, 2019. 81: p. 709-727.
22. Albert, K., et al., Politics of adversarial machine learning. arXiv preprint arXiv:2002.05648, 2020.
23. Patil, A.P., et al. Applying Machine Learning Techniques for Sentiment Analysis in the Case Study of Indian Politics. in International Symposium on Signal Processing and Intelligent Recognition Systems. 2017. Springer.

24. Wiley, E.O., et al., Niche modeling perspective on geographic range predictions in the marine environment using a machine-learning algorithm. 2003.
25. Kang, Y., et al., Understanding house price appreciation using multi-source big geo-data and machine learning. Land Use Policy, 2021. 111: p. 104919.
26. Phelps, D.L., et al., The surgical intelligent knife distinguishes normal, borderline and malignant gynaecological tissues using rapid evaporative ionisation mass spectrometry (REIMS). British journal of cancer, 2018. 118(10): p. 1349-1358.
27. St John, E.R., et al., Rapid evaporative ionisation mass spectrometry of electrosurgical vapours for the identification of breast pathology: towards an intelligent knife for breast cancer surgery. Breast Cancer Research, 2017. 19(1): p. 59.
28. Bajwa, J., et al., Artificial intelligence in healthcare: transforming the practice of medicine. Future Healthcare Journal, 2021. 8(2): p. e188.
29. Stavropoulos, G., et al., Liver Impairment—The Potential Application of Volatile Organic Compounds in Hepatology. Metabolites, 2021. 11(9): p. 618.
30. Smolinska, A., et al., Current breathomics—a review on data preprocessing techniques and machine learning in metabolomics breath analysis. Journal of breath research, 2014. 8(2): p. 027105.
31. Herbig, J. and J. Beauchamp, Towards standardization in the analysis of breath gas volatiles. Journal of breath research, 2014. 8(3): p. 037101.
32. Gaude, E., et al., Targeted breath analysis: exogenous volatile organic compounds (EVOC) as metabolic pathway-specific probes. Journal of breath research, 2019. 13(3): p. 032001.
33. Horváth, I., et al., A European Respiratory Society technical standard: exhaled biomarkers in lung disease. European Respiratory Journal, 2017. 49(4): p. 1600965.
34. Drabińska, N., et al., A literature survey of all volatiles from healthy human breath and bodily fluids: the human volatilome. Journal of breath research, 2021. 15(3): p. 034001.
35. Wehrens, R., et al., Improved batch correction in untargeted MS-based metabolomics. Metabolomics, 2016. 12: p. 88.
36. Stavropoulos, G., et al., Random Forest and Ensemble Methods, in Comprehensive Chemometrics: Chemical and Biochemical Data Analysis. 2020, Elsevier BV. p. 661-672.
37. Fraiwan, L., et al., Automated sleep stage identification system based on time–frequency analysis of a single EEG channel and random forest classifier. Computer methods and programs in biomedicine, 2012. 108(1): p. 10-19.
38. Herfeh, M.P., A. Shahbahrami, and F.P. Miandehi. Detecting earthquake damage levels using adaptive boosting. in 2013 8th Iranian Conference on Machine Vision and Image Processing (MVIP). 2013. IEEE.
39. Ram, S., et al., Predicting asthma-related emergency department visits using big data. IEEE journal of biomedical and health informatics, 2015. 19(4): p. 1216-1223.
40. Rodriguez-Galiano, V.F., et al., An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS Journal of Photogrammetry and Remote Sensing, 2012. 67: p. 93-104.
41. Pijls, K.E., et al., A profile of volatile organic compounds in exhaled air as a potential non-invasive biomarker for liver cirrhosis. Sci Rep, 2016. 6: p. 19903.
42. Wang, Z., et al., Flood hazard risk assessment model based on random forest. Journal of Hydrology, 2015. 527: p. 1130-1141.
43. Guelman, L., Gradient boosting trees for auto insurance loss cost modeling and prediction. Expert Systems with Applications, 2012. 39(3): p. 3659-3667.
44. Semanjski, I. and S. Gautama, Smart city mobility application—Gradient boosting trees for mobility prediction and analysis based on crowdsourced data. Sensors, 2015. 15(7): p. 15974-15987.
45. Smolinska, A., et al., The potential of volatile organic compounds for the detection of active disease in patients with ulcerative colitis. Aliment Pharmacol Ther, 2017. 45(9): p. 1244-1254.
46. Blanchet, L., et al., Constructing bi-plots for Random Forest: tutorial. Analytica Chimica Acta, 2020.
47. Smolinska, A., et al., Interpretation and visualization of nonlinear data fusion in kernel space: study on metabolomic characterization of progression of multiple sclerosis. PLoS One, 2012. 7(6).

48. Camilleri, M., Leaky gut: mechanisms, measurement and clinical implications in humans. Gut, 2019. 68(8): p. 1516-1526.
49. Friedman, M.I., et al., Limonene in expired lung air of patients with liver disease. Digestive diseases and sciences, 1994. 39(8): p. 1672-1676.
50. Van den Velde, S., et al., GC–MS analysis of breath odor compounds in liver patients. Journal of Chromatography B, 2008. 875(2): p. 344-348.
51. Dadamio, J., et al., Breath biomarkers of liver cirrhosis. Journal of Chromatography B, 2012. 905: p. 17-22.
52. Smolinska, A., et al., Profiling of volatile organic compounds in exhaled breath as a strategy to find early predictive signatures of asthma in children. PLoS One, 2014. 9(4): p. e95668.
53. van Vliet, D., et al., Can exhaled volatile organic compounds predict asthma exacerbations in children? J Breath Res, 2017. 11(1): p. 016016.
54. Doran, S.L., A. Romano, and G.B. Hanna, Optimisation of sampling parameters for standardised exhaled breath sampling. Journal of breath research, 2017. 12(1): p. 016007.

IMPACT PARAGRAPH

The present thesis analysed exhaled breath volatile organic compounds (VOCs) and their implementation in clinical settings regarding the diagnosis and prognosis of gastrointestinal diseases. Liver diseases, such as primary sclerosing cholangitis (PSC), are life threatening since no proper early diagnostic tools exist. Lack of early diagnostic tools leads to late treatment, which often results in a liver transplantation. In 2017, an estimated of 1,5 billion cases of chronic liver diseased individuals were reported worldwide [1]. Cirrhosis, the end-stage of liver impairment, accounted for approximately 1,32 million deaths in 2017. In the United States alone, estimated healthcare expenditures regarding hospitalizations reach $81,1 billions [2]. Moreover, the global liver disease treatment market size was valued at $20,673.70 millions in 2020, and it is estimated to reach $36,455.70 millions by 2030, growing at a compound annual groth of 5,7% from 2021 to 2030 [3]. PSC is a rare liver condition with unclear etiopathogenesis; it affects roughly 70.000 individuals in the western world. Nevertheless, it still remains the fifth most common indication for liver transplantation in the United States, and it remains a leading indication in several other countries as well [4]. Liver diseases are currently diagnosed through liver biopsy. Its invasiveness, costs, and relatively low diagnostic accuracy require new techniques to be sought. Colon diseases, such as inflammatory bowel disease (IBD), have dramatically increased over the years. In 2018, there were more than 36,8 million ambulatory visits for gastrointestinal symptoms and 43,4 million ambulatory visits with a primary gastrointestinal diagnosis in the United States [5]. IBD alone affects as many as 1,6 million Americans; 70.000 new cases are reported each year. In 2018, gastrointestinal disease healthcare expenditure totalled $119,6 billions; the estimated financial burden of IBD in the United States is more than $31 billions [6-8]. Colon diseases are presently diagnosed through colonoscopy, which has been the gold standard for diagnosing and monitoring disease activity. Alternative ways to diagnose and monitor disease activity are needed since colonoscopy is a considerably invasive and costly technique.

In human research, VOCs arise from different body matrices such as breath, faeces, urine, bile, breast milk, and blood. Based on research conducted in the present thesis, VOC analysis might greatly benefit gastrointestinal disease diagnosis and prognosis due to its promising use as a non-invasive, cost-effective, and easy-to-use diagnostic and monitoring tool. Exhaled breath VOC technology aims at replacing the current costly and invasive diagnostics with a noninvasive approach, using powerful algorithms, which can identify VOCs for accurate monitoring and diagnosis. Achieving this would reduce diagnosis and monitoring costs since exhaled breath analysis is cost-effective. At the same time, it would drastically improve patient treatment because it is patient-friendly due to its non-invasiveness. Moreover, it would also be convenient for clinicians since it can be applied directly at the point of care due to its potential portability. As discussed in the present thesis, bringing exhaled breath analysis into daily clinical settings would highly benefit research and treatment of gastrointestinal diseases since they require highly invasive, expensive, and often not very accurate

tools (e.g. biopsy or colonoscopy). Exhaled breath clinical implementation will have an immense impact on health insurance companies and hospitals because it will substantially decrease healthcare costs. Early diagnosis and proper monitoring will reduce, for example, the need for liver transplantations or the need for costly endoscopic equipment. In both cases, clinicians could also save up time to devote it into performing other medical care duties that might be lagging behind. Furthermore, it is believed that exhaled breath clinical implementation will allow patients to control the disease more efficiently; they will have to go to the hospital less often, which will save them energy and time. In return, this will make them more productive and more active members of the society; it will allow them to improve socially and financially, and make life more enjoyable for them and their families too.

However, implementation of the VOC analysis in gastrointestinal clinical practices is not ready yet for routine applications since more research is required in various aspects. Chapter 2 demonstrated that most VOC studies are either proof-of-concept studies or of a small sample size. Many studies did not perform any internal or external validation of their findings. The correction of possible confounding factors was also not considered, which might have affected the study results. Furthermore, most breath research has focused on endogenous VOC untargeted analysis; Chapter 2 showed that scientific interest should also shift towards exogenous VOC targeted analysis. Chapter 4 raised awareness regarding batch effects in exhaled breath VOC studies that do not allow for across-study comparisons. Chapters 2 and 4 showed that 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 have hampered the exhaled breath clinical implementation. Therefore, Chapters 3 and 5 aimed to provide an overview of various pre-processing approaches suitable for volatilome data of diverse nature and to equip the reader with a basic overview of suitable techniques for treating and successfully exploiting volatilome data. Furthermore, from a VOC analysis standpoint, a diseased organ could release VOCs via the bloodstream in breath and other body excretion means (e.g. faeces, urine). Chapter 6 showed that fusing this complementary information could result in higher accuracy breath diagnostic tests. Given the complexity and size of volatilome data, more advanced fusion methods might be needed; Chapter 6 proposes such a method. Chapter 7 performed a case study that took into account knowledge gained in the present thesis (i.e. Chapters 2-6) and tested the assumption of using exhaled breath to differentiate primary sclerosing cholangitis patients from inflammatory bowel disease patients. The study results confirmed that assumption.

The findings of the present thesis might contribute to scientific advancement in several ways. Chapters 2 and 4 summarise and raise awareness regarding the 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. Chapters 3 and 5 follow up on the lack of consensus in data handling and statistical tool availability and use by thoroughly discussing and proposing how volatilome data might be approached and analysed regarding VOC biomarker discovery. Chapter 6 provides insight on the concept of data fusion and why and how this concept can be applied to volatilome data. Data fusion is a known concept in computer sciences; however, little seems to be known and published regarding applications of data fusion in the field of exhaled breath research and VOC analysis as a whole. Chapter 7 shows that exhaled breath can be used to diagnose and monitor primary sclerosing cholangitis patients, which has been challenging clinicians to date. Additionally, the present thesis results and previous and future study results might help bring exhaled human breath research into daily clinical setups. Implementing exhaled breath analysis in the clinics would benefit not only gastrointestinal research but other medical fields since the same principle can be applied in any medical field when it comes to using VOC analysis. Therefore, the present thesis can interest scientific researchers in various fields aside from breath, and the presented statistical tools and ideas can be considered general guidelines for researchers who perform statistical modelling with complex biomedical data.

Finally, the work and results in the present thesis have been shared with other researchers since they have been presented at several international scientific conferences through poster and oral presentations. The work in Chapter 4 regarding the implementation of quality controls to prevent batch effects in breathomics data and allow for cross-study comparisons was awarded the Best Poster Prize by the Journal of Breath Research during the Breath Summit 2018 (June 17–20, 2018, Maastricht, The Netherlands). The work in Chapter 6 regarding advanced data fusion by using random forest proximities and the pseudo-sample principle was awarded the Best Presentation Prize at the 42nd Chromatographic Symposium (June 4-7, 2019, Szczyrk, Poland). All of the results have been or will be documented through scientific publications: Chapters 2-6 have been published, whereas Chapter 7 is in preparation.

References

1. Vento, S. and F. Cainelli, Chronic liver diseases must be reduced worldwide: it is time to act. The Lancet Global Health, 2022. 10(4): p. e471-e472.
2. Hirode, G., S. Saab, and R.J. Wong, Trends in the burden of chronic liver disease among hospitalized US adults. JAMA network open, 2020. 3(4): p. e201997-e201997.
3. Shraddha Mali, M.D., Onkar Sumant, Liver Disease Treatment Market by Treatment Type (Antiviral Drugs, Immunosuppressants, Vaccines, Immunoglobulins, Corticosteroids, Targeted Therapy, Chemotherapy Drugs) and Disease Type (Hepatitis, Autoimmune Diseases, Non-alcoholic Fatty Liver Disease (NAFLD), Cancer, Genetic Disorders and Others): Global Opportunity Analysis and Industry Forecast, 2021–2030. 2021: Allied Market Research. p. 254.
4. Mehta, T.I., et al., Global incidence, prevalence and features of primary sclerosing cholangitis: A systematic review and meta-analysis. Liver International, 2021. 41(10): p. 2418-2426.
5. Peery, A.F., et al., Burden and cost of gastrointestinal, liver, and pancreatic diseases in the United States: update 2021. Gastroenterology, 2022. 162(2): p. 621-644.
6. Kappelman, M.D., et al., Direct health care costs of Crohn's disease and ulcerative colitis in US children and adults. Gastroenterology, 2008. 135(6): p. 1907-1913.
7. Gibson, T.B., et al., The direct and indirect cost burden of Crohn's disease and ulcerative colitis. Journal of Occupational and Environmental Medicine, 2008: p. 1261-1272.
8. Longobardi, T., P. Jacobs, and C.N. Bernstein, Work losses related to inflammatory bowel disease in the United States: results from the National Health Interview Survey. The American journal of gastroenterology, 2003. 98(5): p. 1064-1072.

ACKNOWLEDGMENTS

This has been it, then, the very last piece of the PhD thesis puzzle, the acknowledgement section. Throughout my PhD journey, I met, collaborated, and shared many beautiful moments with some very lovely people, all of which, in their way, helped me get through this challenging journey and reach the sweet end. My people, my family, also contributed tremendously to this achievement in their ways. Therefore, I would like to take a moment and use this last piece of the puzzle to thank them for everything they did these past few years.

First, I would like to express my sincerest gratitude to my promoter, **Frederik-Jan**, for his unwavering support, guidance, and encouragement throughout my PhD journey. His mentorship and guidance have been a constant source of inspiration and motivation; I would not have been able to complete this work without his guidance and support. He has been an incredible mentor, not only on a professional level but also on a personal one. I will never forget my first Christmas in Maastricht, where I was on my own, and he kindly invited me to his home to celebrate with his family. I could also never forget all the fun and amazing times we had at conferences or department gatherings. Thank you for everything you did for me, especially your support during these last challenging moments of the journey.

I would also like to extend my appreciation and gratitude to my supervisor, **Agi**, for her valuable insights and feedback. Her expertise and insights in the field of Exhaled Breath have been precious. I am grateful to her for allowing me to work on such an exciting project and challenging research field. Her guidance and suggestions significantly improved my research and thesis, and I am thankful for the time and effort she invested in my work.

Alex, my man, I have said this many times, but I will say it once more. I would not have gotten through the PhD journey had it not been for you. Because of you, I managed to stay on track, keep my sanity, and keep going when things became ugly. You taught me many things regarding academia, from how I should approach my PhD, get the most out of it and enjoy the ride, to how I should approach teaching and be a good tutor and supervisor to students. Most importantly, however, you taught me many things outside academia. I will not get into details, these are things between you and me anyway, but I will still say that I will never forget all these days and nights we spent (and keep spending) together, either at my place or yours or at Falstaff discussing whatever. Nor will I ever forget us going out and having fun. Likely, I am sticking around Maastricht a bit longer so we can continue having fun moments together. I am deeply thankful for getting to know you, my man, and for becoming such good friends. I would not have lasted long here were it not for you; I mean this. A simple thank you would not be enough; nonetheless, THANK YOU, my brother!

Christy, I do owe a special thank you to you, too. You were probably the best officemate I could have asked for. You were always open to listening when things went sideways, and my frustration hit red. You always shared a cup of coffee with me even though you were always crazy busy, running among all the labs and constantly being asked questions by everyone in the department. You are one of the most hard-working, friendly, polite, and kind individuals I have ever met. Keep being you, and I wish you great success in your future endeavours. Who knows, we might get to keep seeing each other somewhere in Utrecht or Amsterdam in the future.

I would also like to thank my other fellow PhDers for the lovely coffee breaks, lunch breaks, lab uitjes we spent together, and the fun we had. **Wenbo**, **Shan**, **Philippe**, and **Sven**, thank you. It was fun being around you; I learned a lot from you. The breath team, **Robert** and **Kim**, thank you for all the lovely memories, especially those we made at the conferences we attended.

I also owe a thank you to **Alex** and **Danielle**, who ran most of my lab analyses, ensuring that we would get high-quality data to work with even though the GCMS would not be the most helpful with that. They also taught me a lot about the GCMS and provided a helping hand with all the practicals we had to give in the lab for our bachelor students. I am thankful for all of that.

A very special thank you goes to **Ger**, too. Thanks, man, for all our fun and sometimes frustrating talks. Markets can be a nasty place to be part of sometimes, but we hang in there; good times are ahead! ;)

My boys, **Sybren**, **Wouter**, **Noury**, and **Connor**, you contributed a ton to making my PhD time a lot of fun. Those night outs at Falstaff, where we would also continue to Basilica and the rest will never be forgotten.

Μα και **μπα**, το μεγαλύτερο ευχαριστώ είναι σε εσάς. Πρώτα από όλα, δε θα ήμουν καν στην Ολλανδία αν εσείς δε με είχατε στηρίξει τόσο ψυχολογικά όσο και οικονομικά. Χάρη σε εσάς έκανα το μεταπτυχιακό μου στο 'Αμστερνταμ που μου έδωσε στη συνέχεια τη δυνατότητα να κάνω το διδακτορικό μου στο Μάαστριχτ. Χάρη σε εσάς άντεξα την αφερεμένη αυτά τα τέσσερα-πέντε χρόνια και χάρη σε εσάς και την υποστήριξή σας κατάφερα να φέρω εις πέρας το διδακτορικό. Μπορεί να μας χωρίζουν μερικές χιλιάδες χιλιόμετρα, αυτό όμως δε σημαίνει ότι δε μου λείπετε ή ότι δε σας σκέφτομαι καθημερινά. Σας αγαπώ πολύ!

Ολινού! Σαφώς και εσύ συνέβαλες τα μέγιστα στο να ξεπεράσω δύσκολες στιγμές καθ' όλη τη διάρκεια του διδακτορικού. Ήταν πάρα πολλές οι φορές εκείνες που μιλούσαμε μαζί και άκουγες υπομονετικά τον πόνο μου για την αφερεμένη και το πως κάνει τα νεύρα μου τσατάλια! Σε ευχαριστώ για όλη την υποστήριξη και αναμένω να σε δω στην ορκωμοσία ως παράνυφη. ☺ Σε αγαπώ πολύ!

Last but most definitely not least, mijn **meisje**! A huge thank you goes to you, too, meisje. You helped me enormously in getting through the PhD journey. You always patiently Listened and provided rational thinking when things would get annoying and I was fuming. God knows how many times I said “that’s it, it’s over, I’m not doing anything else, I’m done with this whole thing”, and you would always tell me to calm down, put things aside for a bit, reason, and continue later with a clear mind to achieve my goal. Your support in achieving my goal has been tremendous, and I will always be deeply thankful for that. So, here I am, writing these last words of that goal, the PhD thesis, and when this is all set and done, we have no other worries than looking the map, checking our lists, and finding where we are headed next. Oh, en wij moeten natuurlijk niet vergeten de toetjes en de koekjes. Hoeveel? Heel veel! Σε αγαπώ πολύ, meisje!

ABOUT THE AUTHOR

Georgios Stavropoulos was born on May 1, 1991 in Patras, Greece. After completing his secondary education (Lyceum) in 2009, he pursued a Bachelor of Science in Chemistry at the University of Patras, Department of Chemistry. He wrote his Bachelor's thesis on "Quantitative and Qualitative Analysis of Assays and Related Substances of Medicines with High-Performance Liquid Chromatography (HPLC)" and received his Bachelor's degree in 2014. Following his completion of military service in 2015, he moved to The Netherlands to pursue a Master of Science in Analytical Chemistry at the University of Amsterdam. During his 9-month Master's internship at AkzoNobel in Deventer, The Netherlands, he conducted data analysis and performed statistical modelling research to predict thermodynamic parameters of binary solvent mixtures from spectroscopic data.

In September 2017, Georgios began his PhD studies at the Department of Pharmacology & Toxicology at Maastricht University under the supervision of Prof Dr Frederik-Jan van Schooten and Dr Agnieszka Smolinska. Throughout his PhD research, he focused on applying Machine Learning and Chemometrics to human diseases using exhaled breath data. He was responsible for designing experiments, executing laboratory analyses, developing and applying machine learning techniques, and interpreting analysis outcomes. His research has been presented at scientific conferences and has been awarded first prizes for best poster from the Journal of Breath Research during the Breath Summit 2018 and best presentation during the 42nd Chromatographic Symposium 2019.

In addition to his research, Georgios has also had experience teaching, lecturing, and supervising biomedical bachelor students. He has coordinated and supervised practical courses, supervised interns with their bachelor theses, and guided problem-based learning groups. After completing his PhD, Georgios began working for Rabobank as a Compliance Model Validator in Utrecht, The Netherlands.

A

LIST OF PUBLICATIONS

Stavropoulos, G., et al., Liver Impairment—The Potential Application of Volatile Organic Compounds in Hepatology. Metabolites, 2021. 11(9): p. 618.

Stavropoulos, G., et al., Preprocessing and analysis of volatilome data, in Breathborne Biomarkers and the Human Volatilome. 2020, Elsevier. p. 633-647.

Stavropoulos, G., et al., Implementation of quality controls is essential to prevent batch effects in breathomics data and allow for cross-study comparisons. Journal of Breath Research, 2020. 14(2): p. 026012.

Stavropoulos, G., et al., Random Forest and Ensemble Methods, in Comprehensive Chemometrics: Chemical and Biochemical Data Analysis. 2020, Elsevier BV. p. 661-672.

Stavropoulos, G., et al., Advanced data fusion: Random forest proximities and pseudo-sample principle towards increased prediction accuracy and variable interpretation. Analytica Chimica Acta, 2021: p. 339001.