Thesis for the degree of European Master in Quality in Analytical Laboratories

Free Fatty Acid Profiling of Metabolic Syndrome Blood Serum by Gas Chromatography-Mass Spectrometry

By Carlos Miguel Vicente Gonçalves

Supervisors:

Prof Dr Bjørn Grung (University of Bergen, Norway)

Prof Dr Hongmei Lu (Central South University, PR China)

Bergen, September 2015

Faculty of Mathematics and Natural Sciences
Department of Chemistry
University of Bergen, Norway

Research Centre of Modernization of Chinese Medicines, College of Chemistry and Chemical Engineering, Central South University, China

Thesis for the degree of European Master in Quality in Analytical Laboratories

Free Fatty Acid Profiling of Metabolic Syndrome Blood Serum by Gas Chromatography-Mass Spectrometry

By Carlos Miguel Vicente Gonçalves

Supervisors:

Prof Dr Bjørn Grung (University of Bergen, Norway)

Prof Dr Hongmei Lu (Central South University, PR China)

Bergen, September 2015

Faculty of Mathematics and Natural Sciences
Department of Chemistry
University of Bergen, Norway

Research Centre of Modernization of Chinese Medicines, College of Chemistry and Chemical Engineering, Central South University, China

Index

Acknowledgments

I want to thank to everyone involved in this research project, in particular to my supervisors Dr Hongmei Lu and Dr Bjørn Grung, which helped me with all the process since the selection of the research topic, analytical platforms and team coordination, availability for problem discussion (both academic and personal issues), thesis revision and project conclusion. I am grateful to Dr Yizeng Liang, Yonghuan Yun, Wei Fan, and Xuxia Long for being so welcoming and helpful during all my stay. I also want to thank to my research partner Lin Dai for all her hard work, I learned a lot from her. Also want to thank to all my colleagues at Central South University (CSU) which helped me during the significant lifestyle change with the move to Changsha, special thanks to Zhang Lin for her kindness and teaching me the Chinese lifestyle, and also to the colleagues both at the College of Chemistry and Chemical Engineering, Pedro Sousa, Kamila Smieszkol, Ming Wen, Xinyi Zhou and Candy, and all others from the same department or at the Pharmacy Department of the Xiangya Hospital.

I want to thank to all the people involved in the EMQAL projects, since all the lecturers that came from Bergen, Gdansk, Barcelona, Faro and other parts of the world to share their knowledge during my first year at Cadiz University (UCA). I want to acknowledge Dr Isabel Cavaco and Dr Miguel Palma for their hard work in coordinating the program. Special thanks to all my EMQAL colleagues that taught me about their origins and shared a little bit of their culture, and made this EMQAL program the best experience of my life. In particular I want to thank Kasahun Abeje for all his support during our journey in this program, in which he has taught me about motivation and goal achievement.

I want to express my gratitude also to all those that directly and indirectly contributed to my financial support, through this Erasmus Mundus scholarship or through working opportunities that allowed me to maintain myself. Also to thank to my working colleagues in Bergen, Christine Drage, Carlos Souza, Mondel Kfry, Magdalena Tomalska, Leandro Barretiri and all the others that in different ways showed their support.

I want to thank all the people I met during my stay in Bergen for their support and motivation, coming from different parts of the world, in particular to Yun Sikyin and Moritz Pohl during my first months of stay in Bergen. A very special thanks to Lucia Brhlikova for her continuous support, kindness and patience in the final stage of my research, and for teaching me about the spirit of adventure.

I would like to thank my family for all the support during this period of my life, for their acceptance in this first time that I moved from my hometown. Especially I want to thank my mother Preciosa Gonçalves for her love, strength and motivation; she always has been a good example of determination and hard work. I want to thank my sister Liliane Gonçalves and my father Joaquim Gonçalves, and all the family that in one way or another has shown me that wherever you are, they will always be there for you.

Carlos Miguel Vicente Gonçalves

Bergen, September 2015

Abbreviation List

AdA	Adrenic Acid
ATP	Adenosine Triphosphate
BMI	Body Mass Index
CAC	Citric Acid Cycle
CCA	Canonical Correlation Analysis
CDS	Chinese Diabetes Society
CV	Coefficient of Variation
CVD	Cardiovascular Diseases
DBP	Diastolic Blood Pressure
ECL	Equivalent Chain Length
EFA	Esterified Fatty Acids
FA	Fatty Acid
FAMEs	Fatty Acid Methyl Esters
FDA	USÃ Food and Drug Administration
FFA	Free Fatty Acids
FWER	Family-Wise Error Rate
GC	Gas Chromatography
GLC	Gas Liquid Chromatography
GSC	Gas Solid Chromatography
HC	Healthy Controls
HDL	High Density Lipoproteins

HDL-C	High Density Lipoprotein Cholesterol
IDF	International Diabetes Federation
IS	Internal Standard
IUPAC	International Union of Pure and Applied Chemistry
LDL	Low Density Lipoproteins
LLOQ	Lower Limit of Quantification
MDS	Multidimensional Scaling
MeOH	Methanol
MetS	Metabolic Syndrome
MS	Mass Spectroscopy
NCD	Non-communicable diseases
NCEP:ATPIII	National Cholesterol Education Program's Adult Treatment Panel III
NEFA	Non-Esterified Fatty Acids
NIST	National Institute of Standards and Technology
OOB	Out-of-Bag
PCs	Principal Components
PCA	Principal Component Analysis
QA	Quality Assurance
QC	Quality Control
QC-scan	Quality Control samples analysed in full scan mode

RAFFT	Recursive Alignment by Fast Fourier Transform
RF	Random Forests
RSD	Relative Standard Deviation
SBP	Systolic Blood Pressure
SIM	Selected Ion Monitoring
SNS	Sympathetic Nervous System
SOP	Standard Operating Procedure
TCMSys	Traditional Chinese Medicine Fingerprint Database System
TG	Triglycerides
TOF	Time-of-Flight
T2DM	Type 2 *Diabetes Melitus*
VLDL	Very Low Density Lipoproteins
WHO	World Health Organization
WC	Waist Circumference

Abstract

It is estimated that 20-25% of adults in the world have metabolic syndrome (MetS), and in Chinese population this has become an important public health problem. MetS is characterized as cluster of metabolic disorders that lead to cardiovascular diseases or type 2 *diabetes mellitus*, and is diagnosed through the identification of three out of five possible risk factors: central obesity, increased blood pressure, increased fasting blood glucose, increased serum triglycerides and decreased HDL-cholesterol. Although the diagnosis of MetS is possible, its underlying mechanisms are yet to unveil, and new approaches are needed to explore the metabolic disorders and provide better diagnosis methods. Metabolomics has proven to be a powerful tool in biomarker identification providing metabolic information in several clinical fields. In this study, a metabolomics approach is used for free fatty acid (FFA) profiling by gas chromatography-mass spectrometry (GC-MS) of MetS serum samples of Chinese patients, and a combined chemometrics strategy using principal component analysis (PCA), random forests (RF) and canonical correlation analysis (CCA) is applied for classification and biomarker identification. After optimizing the derivatization and program temperatures for GC-MS analysis, an FFA analysis based on experimental design and internal quality control procedures is performed, with demonstrated reliability though the continuous evaluation of quality control (QC) samples. The classification results from PCA and RF confirm the good performance of the QC samples, and the later method is able to classify MetS samples and healthy control (HC) samples efficiently, with an accuracy of 0.8632. From CCA the results show a canonical correlation coefficient of 0.9015, indicating a high correlation between the clinical parameters and FFA. The most relevant clinical parameters for this correlation are shown to be HDL-C, TG and FBG, while the FFA that have the most influence in the correlation are C16:1n-9c, C20:1n-9c, and C22:4n-6c, which are identified as potential biomarkers. Results suggest that targeted metabolic profiling of FFA by GC-MS coupled with chemometrics tools such as RF and CCA are able to provide important information for biomarker discovery and disease mechanism investigation on MetS.

1. Introduction

1.1 Objectives

This study aimed at the serum fatty acid profiling of Chinese Metabolic Syndrome (MetS) patients by gas chromatography (GC) coupled to mass spectrometry (MS), and the establishment of a predictive model for classification of MetS samples in the Chinese population. In order to fulfil such aims, secondary objectives were set:

1.1.1 The achievement of repeatability conditions through the improvement of the derivatization steps and chromatographic temperature programs.

1.1.2 The development of an internal quality control procedure including experimental design, the use of internal standards and the use quality control samples.

1.1.3 The determination of fatty acid profiles and achievement of a classification model using principal component analysis (PCA) and random forest (RF) machine learning algorithms after pre-processing.

1.1.4 To relate the disease classification patterns with its diagnosis factors by means of canonical correlation analysis (CCA).

1.1.5 To identify the potential biomarkers for the disease based on the classification results.

1.2 Literature Review

Health problems often have a high impact on the development of a country, affecting its social and economic status due to the increasing of poverty in the community and the decrease of active working citizens. Between 2000 and 2011 the world mortality increased almost 4%, with Non-communicable diseases (NCD) being the leading group, representing approximately 63% of the global death number in 2008[1,2]. NCD involve heart diseases, stroke, cancer, chronic respiratory diseases and diabetes, and have several identified modifiable risk factors. A particular cluster of some of these risk factors that lead to Cardiovascular Diseases (CVD) and type 2 *Diabetes Mellitus* (T2DM) is defined as Metabolic Syndrome[3–5]. In China, there have been studies reporting a high incidence of MetS with a potential of having a great impact on Chinese public health[6].

1.2.1 Metabolic Syndrome

MetS was firstly described in the 1920s by Kylin[7] as a cluster of hypertension, hyperglycaemia and gout. However, it took until 2005 for a consensus statement from the International Diabetes Federation (IDF) to be published. Meanwhile, the Chinese Diabetes Society (CDS) stated in 2004 that a MetS patient is an individual who satisfies any three or four risk factors among central obesity (BMI $\geq$ 25.0 kg m^{-2}), raised blood pressure (SBP/DBP $\geq$ 140/90 mmHg or hypertension patients), raised plasma glucose (fasting plasma glucose $\geq$ 6.1 mmol/L or diabetes) together with raised plasma triglycerides (TG $\geq$ 1.7 mmol/L) and/or reduced HDL-cholesterol (HDL-C < 0.9 mmol/L). The several international criteria (WHO, NCEP:ATPIII, CDS) in force present slight changes, and are also affected by ageing, physical inactivity, proinflammatory state, life style bias and country/ethnic group. It is estimated that about 20-25 percent in world's adults have MetS and they are twice as likely to die from and three times as likely to have a heart attack or stroke compared with people without this syndrome. In addition, people with MetS have a fivefold greater risk of developing T2DM. They would join to the 230 million people in the world who already have diabetes, one of the most common chronic diseases worldwide and the fourth or fifth primary cause of death in the developed world. Therefore IDF acknowledges the urgency for a better understanding of MetS and for long-term research for "measuring more precisely a wide range of variables"[3] and "establishing the most predictive markers"[3]. Despite the MetS criteria being well known, its variability in the diagnosis (any three of the risk factors can be taken into account to the MetS condition) leads to problems in consistency and questions about whether individuals that have one or two of the risk factors may be considered as MetS patients or healthy, and if healthy, how close they are to be MetS patients[3,8–13].

As previously described, MetS involves a particular number of risk factors that are well known to be also risk factors for obesity, T2DM and CVD individually, having their individual proposed biochemical pathways. As shown in figure 1, some proposals for the biochemical behaviours that occur in the presence of MetS have been done[5].

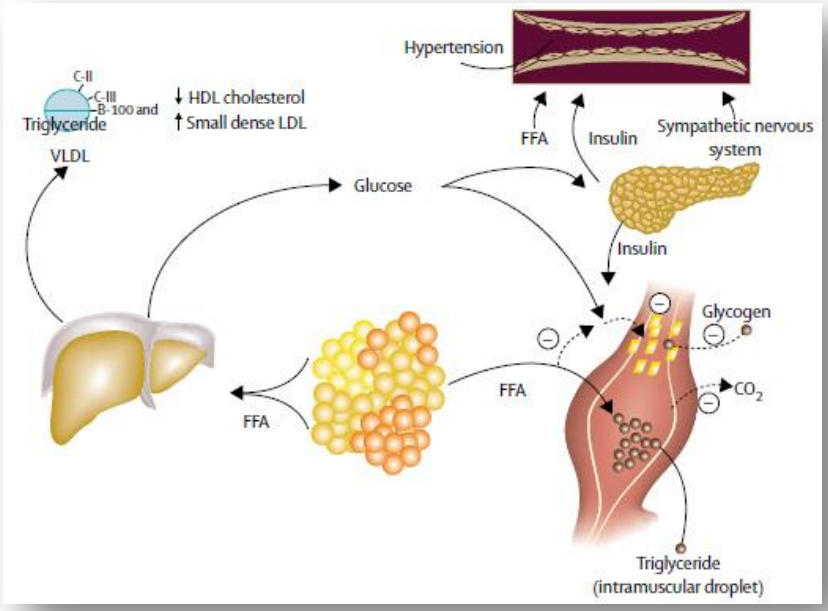

Figure 1: MetS proposed biochemical pathways[5].

When a significant amount of adipose tissue mass is accumulated in the human body there is a release of free fatty acids (FFA) which will act in the liver as promoters of glucose production, and being incorporated in TG and very low density lipoproteins (VLDL). In addition, HDL-C is decreased concomitantly with the increase of low density lipoproteins (LDL). The action of FFA is extended to the inhibition of insulin-mediated glucose uptake in the muscle, which consequently will stimulate its accumulation in the blood stream. As a parallel result, glycogen synthesis is inhibited due to the low concentration of glucose in the muscle and intramuscular droplets of TG are formed due to the excess of FFA. Nevertheless, FFA seem to stimulate the insulin secretion in the pancreas, leading to an hyperinsulinemia status which ultimately results in the increased activity of the sympathetic nervous system (SNS), which together with the FFA will lead to hypertension. Such mechanisms summarize the phenomena of insulin resistance, increased BMI/waist circumference (WC) due to intra-abdominal or visceral adipose tissue, dyslipidaemia (focusing on VLDL, HDL-C, LDL and FFA), glucose intolerance due to insulin dysregulation, and hypertension mediated by FFA[14] and SNS.

Much information on the specific pathways affected by the metabolomic changes in the organism have been gathered in recent years[5,9,11,15–17], however specific information on the FFA profile present in the serum samples of MetS patients is yet to unveil.

1.2.2 Fatty Acids and Serum Sampling

Commonly, a fatty acid (FA) is biochemically defined as a long chain of carbon atoms linked to a carboxylic acid functional group, and are the key constituents of lipids. The degree of unsaturation (number of double bonds) or saturation, and their number of carbons, are the main characteristics responsible for their nomenclature and simultaneously for their biochemical properties. As the example shown on figure 2, a 19 carbon length FA with an unsaturation on the 10th carbon is designated by Z-10 nonadecenoic acid or cis-10-nonadecenoic acid (using a systematic nomenclature accordingly to the International Union of Pure and Applied Chemistry (IUPAC) recommendations), but also other nomenclature ways may refer to the same compound as cis-Δ^{10} nonadecenoic acid (where the superscript number indicates the position of the double bond counting from the carboxylic acid group, and the prefix shows the structure conformation), as ω-9 nonadecenoic acid (where the number after the omega is the location of the double bond counting from the methyl group located next to the carbonyl carbon), or simply as C19:1(cis-10), describing in a very straightforward way the number of carbons and unsaturations as well as the position and conformation of the double bond. Nevertheless, some “trivial” or historical names may be considered in scientific writing depending upon convenience[18,19].

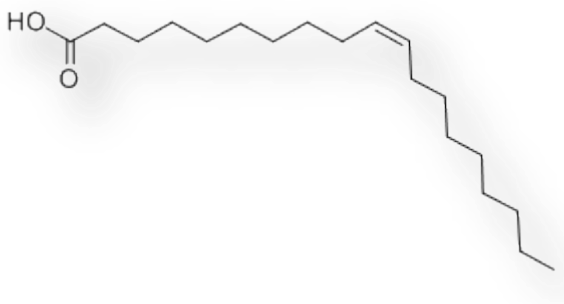

Figure 2: 10-nonadecenoic acid structure[20]

In nature, FAs can be found in several types of sources which contain either fats or oils[19]. In the human body these molecules are involved in many processes and can be found incorporated in phospholipids (generally one or more FA linked to a glycerol,

which has a phosphate-alcohol group attached to it) and consequently forming biological membranes. Their amphipathic properties allow a selective permeability (hydrophilic polar glycerol side versus the hydrophobic apolar FA part) in membranes, a diverse organization of the several phospholipids in the membrane depending on their unsaturation degrees (allowing a dynamic fluidity) and relative percentage of each, and also a tolerance to the integration of molecules with biological functions (glycolipids, channel proteins, cholesterol and others). Indeed, FAs play an important role in the human body, being also a source of energy while stored as TG (three FA chains bonded to a glycerol molecule), often in fat deposits or originated from dietary fats. These molecules have as main characteristic being nonpolar, and their main storage location is the adipose tissue, being mobilized to other tissues when energy is needed, in the form of FFA. In the biological pathways, this process is carried on by specific lipases that do the cleavage by hydrolysis. In the blood plasma, these FFA are then transported by the circulating albumin being finally assimilated in the target cells by specific membrane proteins, and oxidized by the respective metabolic pathways which will origin energy in the form of ATP (Adenosine Triphosphate) via the citric acid cycle (CAC), as shown in figure 3[18,19].

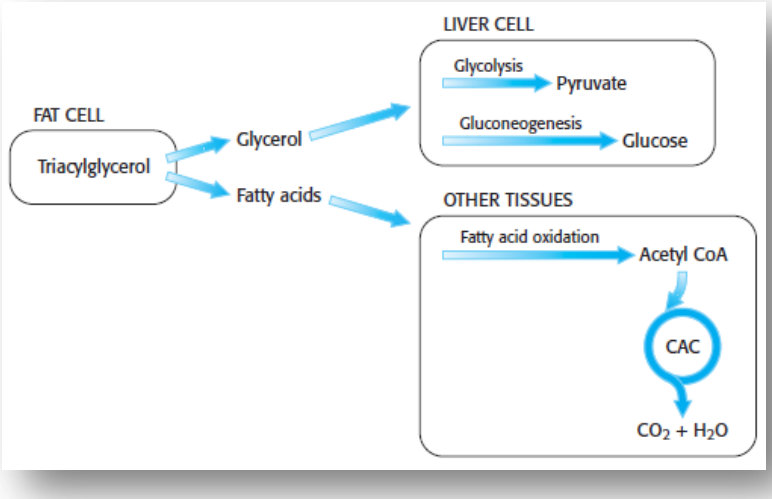

Figure 3: Triglyceride metabolic pathways and their location[18]

Some common analytical techniques such as GC-MS require the FAs to be derivatized after the extraction step, making them volatile in order to allow their elution through the chromatographic column. These derivatives are often fatty acid methyl

esters (FAMEs), which can be easily obtained by common methylation techniques involving the addiction of an acid-methanol mixture (e.g. H_2SO_4, HCl, BF_3) as described in figure 4.

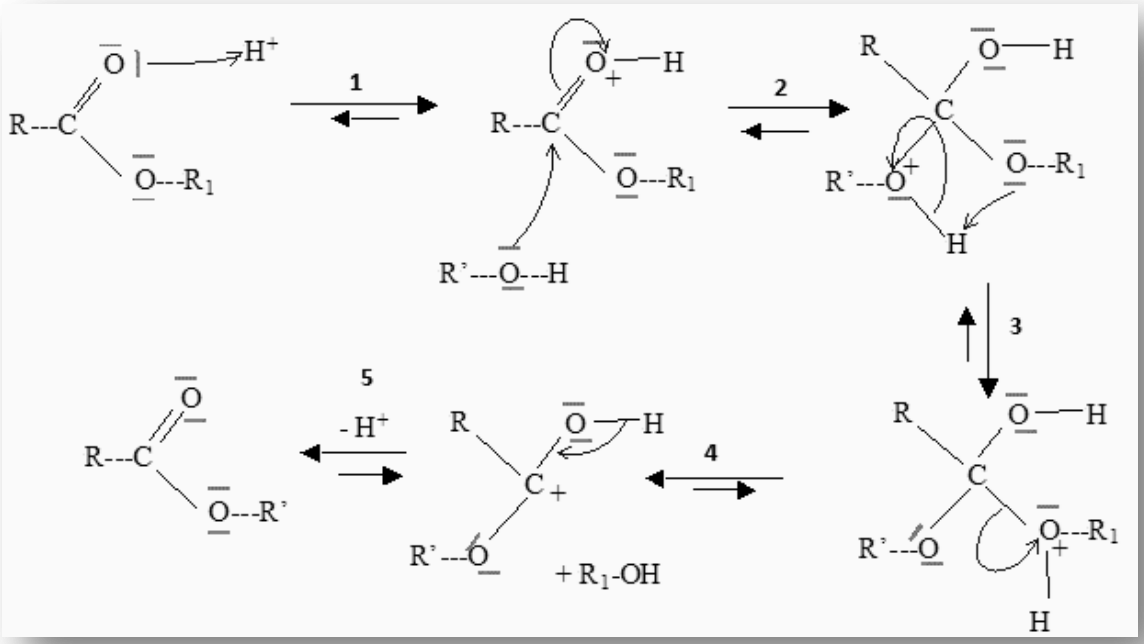

Figure 4: Five-step basic mechanism of FA methylation

In a first step there is the attack of the acidic proton (from the H_2SO_4) by the non-bonding electrons of the carbonyl group. The carbon of the carbonyl group will then have a reduced electron density that will stimulate a nucleophilic attack from the non-bonding electrons of the oxygen in the methanol (MeOH). After step 2, a four-bond carbon is formed; the MeOH oxygen will attract the electrons that establish its bond with hydrogen, making it vulnerable to the attack of any of the other two adjacent hydroxyl groups. As a result a leaving group is formed (H_2O) and an unstable carbocation is generated, leading ultimately to the formation of a double bond with the hydroxyl group and simultaneously to the release of the labile hydrogen of the same functional group. The final outcome will then be the establishment of a FAME compound[21].

Based on this procedure several approaches were described in literature, leading to the use of different kinds of acids and solvents, and specifying the type of FA to be methylated, both in plasma[19,22,23] and serum samples[24]; EFA (esterified fatty acids) or

NEFA (non-esterified fatty acids). While EFA exist in the biological context as phospholipids, glycolipids and lipid sterols, NEFA are generally found linked to albumin molecules. Each of these FA types have a specific derivatization before analysis because NEFA need to be separated from the albumin, while EFA need to be cleaved from their associated polar moieties[22,23]. The latter are not bioactive since they require more complex reactions to be released into the blood stream. In fact, after some research on the influence of FA on the classification of metabolic-related disease groups, it was noted that EFA show no significant effects on the discrimination of T2DM and control groups[25]. Considering this, our study developed a derivatization procedure focused only on NEFA, as shown in figure 5.

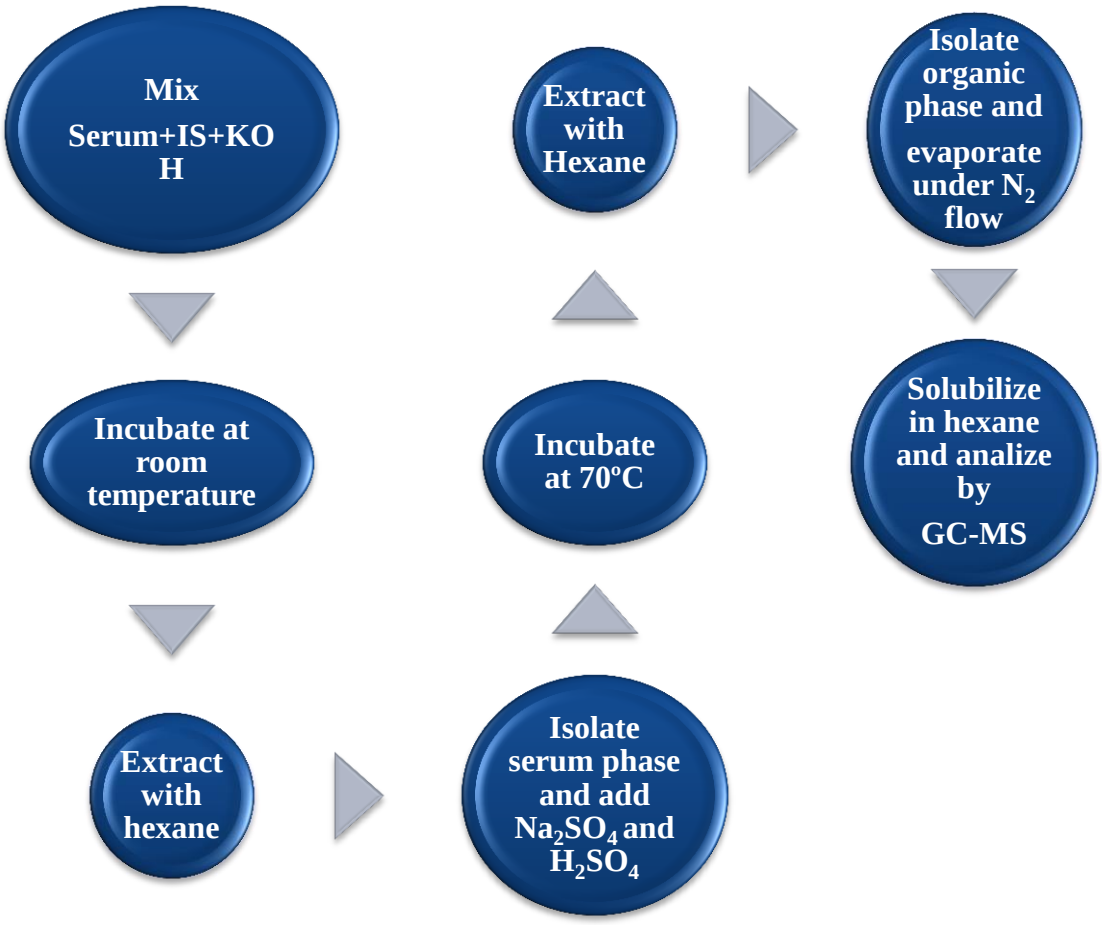

Figure 5: General methodology for FA extraction

Prior to FA extraction, one has to carry out the biological sampling which is a critical step in the analytical process. In the quality assurance cycle, three phases can be identified: pre-analytic, analytic and post-analytic. In the pre-analytic stage there is the

sample collection, test evaluations, sample transport (in case the sampling is done in a different place from the one where is the storage) and sample storage. The analytic stage is where the actual analysis is performed by doing the testing and quality control. Lastly, the post-analytic stage is where the results are recorded, processed by doing the data analysis and ultimately the final version of the results are reported. In an accredited laboratory, the pre-analytical phase is defined by ISO 15189:2012 standard for laboratory accreditation as the "processes that start, in chronological order, from the clinician's request and include the examination request, preparation and identification of the patient, collection of the primary sample(s), and transportation to and within the laboratory, and end when the analytical examination begins" [26]. Such a definition helps us to understand the need to control and understand the processes that happen during the initial phase of the analytical process. It is estimated that approximately 60% of the clinical laboratory errors happen in the pre-analytic stage. The main causes of these are: specimen collection tube not filled properly; patient ID error, inappropriate specimen collection tube; test request error: and empty collection tube[27]. These errors will ultimately lead to problems related with contaminations, sample degradation or wrong result delivery, which can be minimized by the use of standard operating procedures (SOP) and taking corrective actions for continuous improvement. In the specific case of FA, avoiding hydrolysis and autoxidation is a concerning factor. When working with biological tissues, and particularly with serum samples, the blood collection is followed by an incubation period and a centrifugation step, resulting in the isolation of the serum phase containing the compounds present in the blood stream. This phase shall be stored at temperatures below -70ºC in a nitrogen atmosphere due to the possibility of enzymatic activity and consequent degradation of FA[19,28].

As described before, several errors may arise from inaccurate experimental handling and specimen collection, influencing laboratory measurements (e.g. high-negative-pressure blood draws, non-fasting patients, insufficient tube filling, bacterial contamination, among others). Unlike plasma, serum samples are allowed to pass through the coagulation process avoiding the presence of clotting factors, but potentially masking the levels of certain biomarkers. These two blood-derived matrixes are commonly used interchangeably; however it is known that proteomes and other biomarkers present different profiles depending on the chosen matrix[19,28]. Due to the lower protein concentration on serum, lower demands on sample pre-treatment are required and a lower risk of treatment may influence the results (e.g. incubation

affecting analyte peak areas) [29]. On the other hand, the differences between serum and plasma seem not to be of clinical relevance, despite the statistical significance presented for compounds other than FA[30].

Accordingly to previous knowledge, in order to reduce changes in the serum analytes, each sample shall use the same kind of collection tubes relatively to the manufacturer and material type (significant difference found in growth factors, insulin, hormones and antigens) [31–33]. As stated before, serum is usually allowed to coagulate and therefore a waiting time is required before centrifugation. Obviously this shall be a controlled time if one needs to maintain repeatability and reduced sample variability conditions. This time is estimated to be between 30 and 60 min and will allow the precipitation of cellular components that may affect the downstream analysis, but in some cases more time is need due to the presence of anti-coagulant agents in the sample (e.g. previous drug treatment applied in the patient) [34,35]. After the coagulation process, immediate freezing storage is needed at the previously mentioned temperature of at least -70ºC in order to maximize sample quality[36,37]. Despite some works have shown a reliability limit of 1 serum freeze-thaw cycle[38], there are evidences that FA are able to stand up to 5 cycles with low relative standard deviation (RSD) values, particularly for EFA rather than NEFA[23].

1.2.3 Gas Chromatography and Mass Spectrometry

The MetS diagnosis proposals by the several national and international organizations are based on simple tools preferred by clinicians to evaluate the biochemical status of the patient, facilitating the clinical practice[5]. In the same way, GC-MS is among the most widely used analytical techniques for high-throughput analysis, allowing the analysis of large amount of samples in a relatively short time by researchers around the world.

Chromatography is defined by IUPAC as “a physical method of separation in which the components to be separated are distributed between two phases, one of which is stationary (stationary phase) while the other (the mobile phase) moves in a definite direction”[39]. This is a very clear definition that describes a method initially developed by Mikhail Tswett in 1903 for the separations of pigments in green leaves[40], but established later by Martin and Synge in 1941 during the development of the “Theory of Chromatography”[41], having the later been awarded a Nobel Prize in Chemistry (1952) “for their invention of partition chromatography” [42]. As shown in figure 6, the

chromatographic process is based on a successive series of equilibria established between the two phases, the mobile phase (also called eluent) which contains the analyte, and the stationary phase (in the case of a packed column, it is referred as the packing material, otherwise it is referred as the coating of the column). These equilibria are based on the composition of each of the phases, and their affinity to each other. For instance, a certain analyte which possesses polar characteristics will be more retained in a more polar stationary phase, eluting later in the chromatogram (right side of figure 6 (b)) than another analyte, which possesses apolar characteristics and shows greater affinity for the mobile phase. Such interactions can be easily manipulated by using different stationary and mobile phases, promoting faster and more efficient separations between the analytes of interest.

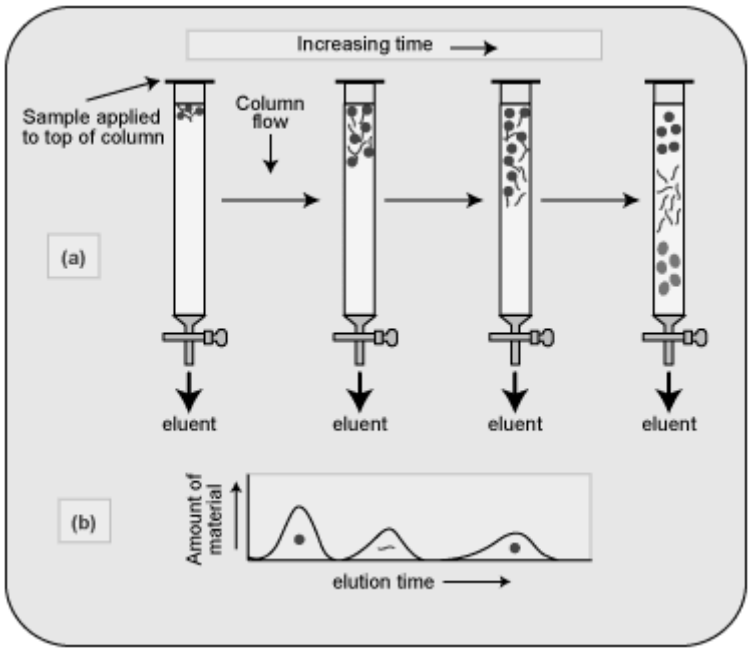

Figure 6: (a) General chromatographic process (b) Chromatogram example[43]

Chromatography can be divided into two main types: planar and column chromatography. In the case of planar chromatography, there is a planar support (a plate) which is coated with the stationary phase, and where the analytes are placed. Afterwards, these supports are put in contact with the mobile phase which is a liquid phase that adsorbs to one of the edges of the plate and moves by either capillarity of gravity, crossing all the plate and resulting on the migration of each of the analytes through the stationary phase. On the other hand, column chromatography is performed in a column/tube (as illustrated in figure 6) that can be packed with the stationary phase

or simply coated. In the case of a packed column, the mobile phase moves by gravity or pressure, and the analyte is introduced at the top of the column, migrating though the column accordingly with the type of chromatography (normal phase, ion exchange, reversed-phase, affinity, size exclusion…). Thus, the column packing will define the relationship between the analyte and the stationary phase, and consequently influence the interaction of the analyte with the mobile phase. Independently of the column packing, the chromatographic process will always produce a chromatogram that results on the signal detection along the time. A more retained analyte will have a larger fraction of time spent in the stationary phase, then a less retained analyte, leading to the generation of different bands in the chromatogram that can afterwards be analysed quantitatively and qualitatively. In general, the analytes are diluted in the chromatographic process due to diffusion, which means that more sensitive detectors are needed for this kind of analysis. Many factors may be adjusted in order to improve the separation of the analytes. One of these is column performance, which is increased by the increase in the column length. This will provide a better peak separation, however peak broadening also happens, thus it becomes difficult to adjust this parameter. Ideally a good separation needs to increase band separation and decrease band spreading, and these factors depend directly on the rate of distribution of each analyte between the mobile and stationary phase. This means that, by selection of an appropriate stationary and mobile phase, this separation can be manipulated. In this way, we can access an approximation of the column efficiency by analysing the peak width in a chromatogram.

When talking about column efficiency and band broadening it is mandatory to refer to the Van Deemter equation[44], which describes the general mechanisms of band broadening in linear nonideal chromatography. This equation considers factors such as Eddy diffusion, longitudinal diffusion and mass transfer, and relates these with the plate theory of Martin and Synge[41]. In this theory, the column is theoretically divided in a number of layers called theoretical plates, as shown in figure 7, in which the mobile phase is able to reach a series of equilibriums in each theoretical plate.

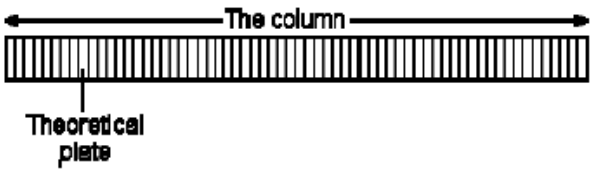

Figure 7: illustration of theoretical plate division of chromatographic columns

In this model, when the number of theoretical plates (N) increases, the efficiency of the separation increases as well. In the same way, the separation efficiency increases when the height (H) of each theoretical plate decreases. These variables can be related with the column length (L) using equation 1:

$$(1) \qquad N = {}^{L}/_{H}$$

Knowing this, it is possible to understand the assumptions of the Van Deemter equation (equation 2):

$$(2) \qquad H = A + {}^{B}/_{u} + C \cdot u$$

in which A is the Eddy diffusion coefficient, B is the longitudinal diffusion coefficient, C is the mass transfer coefficient and u is the linear velocity of the mobile phase. The coefficient A represents the channelling (presence of preferential paths), and this can be minimized by using uniform stationary phases and low particle diameter. The B coefficient shows the spreading effect of the analytes along the analysis, which can be minimized by increasing the flow rate (a critical factor in gas chromatography due to the high gas diffusion). The C coefficient considers the difference in the movement of particles in the mobile phase and in the stationary phase, due to the equilibrium phenomenon, and is more significant in liquid chromatography. The later can be reduced by increasing the column temperature.

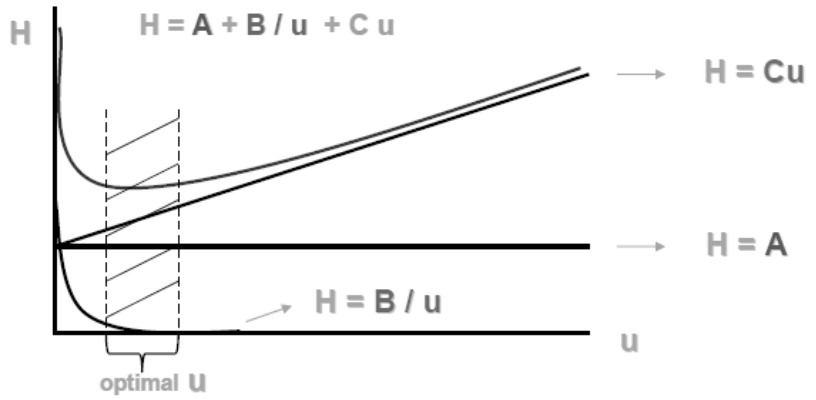

Figure 8: variation of theoretical plate height (H) with linear velocity of mobile phase (u)

As shown in figure 8, it is possible to estimate an optimal flow rate of the mobile phase in order to minimize H, by adjusting the 3 mentioned coefficients.

Apart from the column contribution for band broadening, there is also an outside of column contribution which is related to other parts of the chromatographic system. Components like the injector, the connecting lines to the column, connecting lines to the detector, and also the geometry and characteristics of the detector, are significant points to consider.

Applications of chromatography are many, as this separation method can be used for both qualitative and quantitative analysis. The retention times shown in the chromatograms and the manipulation of mobile and stationary phases together with the different elution temperatures, provide a characteristic data set for a single sample type, allowing identification. On the other hand, the measurement of peak areas or heights, and their comparison with standard references, makes possible the quantitative analysis, in which it is extremely important to control peak broadening factors. The use of peak areas is in fact less influenced by these effects and is the preferred method of quantitation.

Gas chromatography is a fast technique that uses a vaporized sample, which is partitioned between the mobile phase (an inert gas) and the stationary phase (a solid or liquid layer coating the column). It is not necessary large amounts of sample, and after the injection of the sample in the column head, the mobile phase carries the analytes without interacting with it, thus allowing a separation based on the interaction between the analytes and the stationary phase. The two types of GC are Gas-solid

chromatography (GSC) and Gas-liquid chromatography (GLC), depending on the phase of the stationary component. In the first type, there is the risk of semi-permanent adsorption of the analytes to the stationary phase, which will result in peak tailing, or column saturation, which is an important limitation. In GLC, commonly named GC, there is a thin liquid coating layer on the capillary tubing, avoiding the limitations of the previous chromatographic type, making this a widely used method. When comparing the Van Deemter curves for LC and GC, it is possible to see that optimal flows for GC are higher than for LC, making GC a faster method. In addition, it is generally found that H are typically higher for GC (also due to the possibility of having much longer columns) leading to a superior efficiency. The fact that GC uses high temperatures and volatile samples makes it easier to couple it with MS detectors (as shown in figure 9), since MS also requires volatile samples.

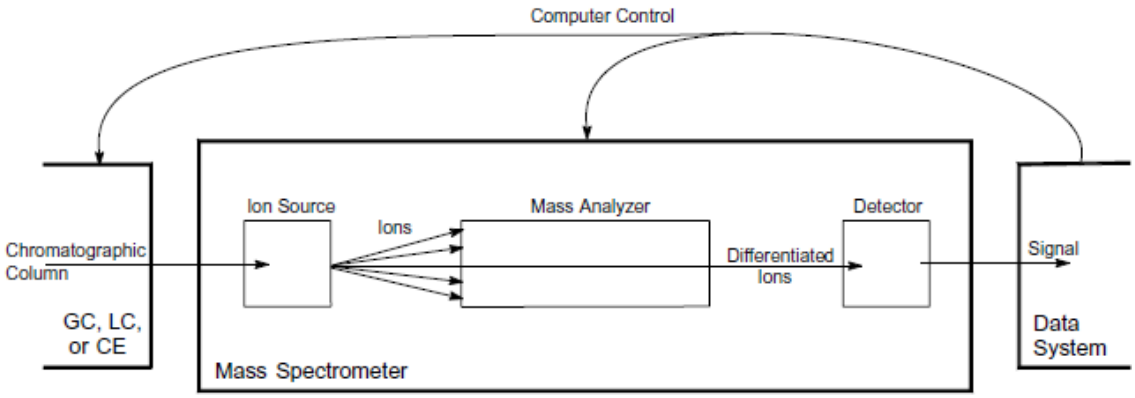

Figure 9: diagram of GC-MS-computer coupled system[45]

Among other types of detectors, MS is a powerful technique able to identify a compound based on its fragmentation pattern, by evaluating the mass-to-charge ratio (m/z) of the charged particles generated in the fragmentation process. Consequently it is a destructive method, not allowing the recovery of the sample after analysis. The particles can be either molecular ions, isotope ions or even fragment ions, but the resultant detection is based on mass, charge and speed.

In general this coupled system works by using an autosampler to load the sample into the GC column inlet. The sample then passes through the column with the help of the carrier gas and reaches the transfer line between the chromatograph and the mass spectrometer. A rubber or silicone septum is located at the inlet, ensuring an isolation of this section during the sample injection (as shown in figure 10).

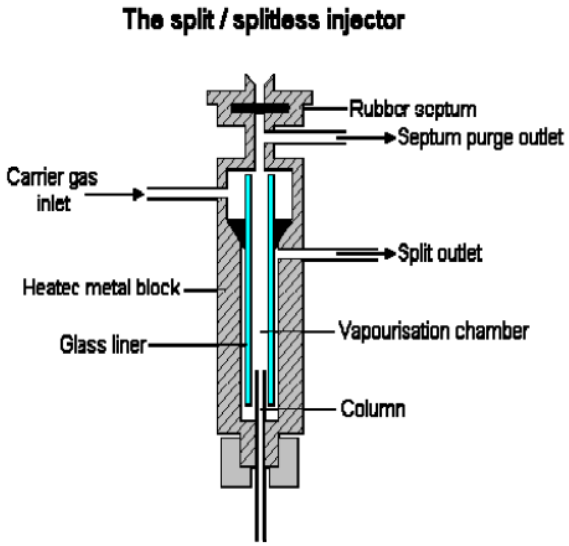

Figure 10: Sample port diagram

The sample port should be 50ºC higher than the boiling point of the least volatile component of the sample. The process of injection may happen in different modes: split, splitless, on column or programmed temperature vaporising. In this work it was used a split injection mode, where the sample is introduced into a heated injection chamber (where the evaporation occurs) and mixed with the carrier gas. This mix is then vented into the split exit and only a part of this mix is allowed into the column. By adjusting the split ratio, it is possible to control the amount of sample that enters the column (ex: 1:5 split ratio means that in each 5 parts injected, only 1 part enters the column). This mode is not suitable for trace analysis, since a low sample volume is analysed, however it helps to avoid column saturation and to obtain narrow bands. Moreover, in order to avoid the reaction of the analyte with the metal surface of the injector, it is necessary to use liners, which most of the times have a glass wool filling that blocks the passage of pieces of the septum or non-volatile particles of the sample.

The column is placed in an oven that will control the temperature during the process, while the mass spectrometer is placed in a different section under vacuum conditions. Nowadays, most applications use capillary columns (open tubular) due to their higher efficiency. Their length makes it easier to place them in the oven as coils. In most analyses, a column temperature program is necessary, due to the different boiling points of the components of the sample and the need to avoid exceeding the temperature limit of the column. At low temperatures better resolutions are often found, however the analysis time is longer.

After passing through the column, the sample is directly transferred to the mass spectrometer (in the case of capillary columns), where it is received in the ion source. At this point, the sample is in a volatile state and undertakes electron ionization by the accelerated electron beam generated by the instrument, as following:

$$M + e^- \rightarrow M^+ \cdot + 2e^-$$

where M is the original molecule.

In order to obtain a good chromatographic profile, the full scan range should be set to a higher m/z than the one of the molecular cluster ions (allowing inclusion of all the fragments). After this step, only the charged species will proceed to the mass analyser, passing through the focusing lens. The most common analysers are Quadrupoles, Ion Traps and Time-of-Flight (TOF). Quadrupoles consist of 4 electrodes placed in parallel, building scanning lines that work as selective m/z filters, operating at relatively low vacuum and being easier to couple with other analysers (tandem analysis). In Ion Traps, the ions reach a ring electrode in a chamber in which the ion motion is damped by a buffer gas (generally helium). Presenting a higher cost, TOF has a wide m/z and high scanning speed, using also a high vacuum. This last technique uses a non-linear ion route, making the ions travel from the ion source through a magnetic field called Reflectron, spreading the ions according to their energies, allowing a better selection in the detector. In this work, a single quadrupole MS was used for analysis.

As a result, the GC-MS system is able to provide a 3-dimensional data set of retention times, m/z and signal, as shown in figure 11.

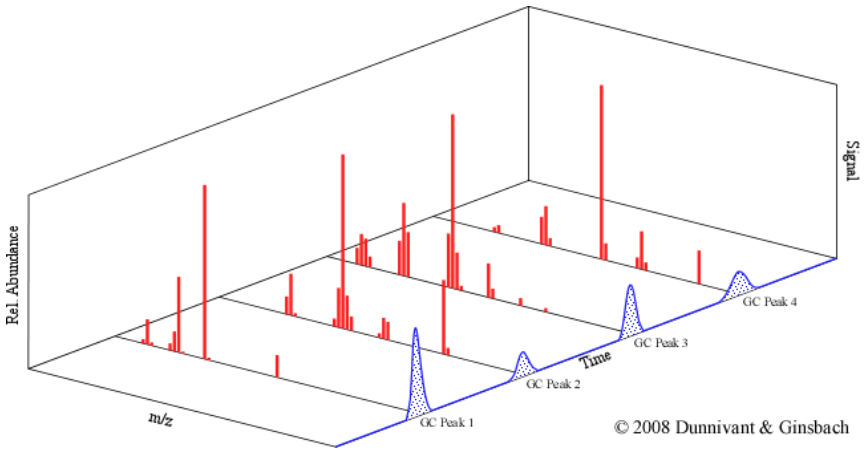

Figure 11: Gas Chromatography-Mass Spectroscopy data set illustration[46]

Such data sets can be obtained in two ways: full scan or SIM (Selected Ion Monitoring). In short, a full scan mode will register all the fragments obtained during the analysis time, making this detection type useful for the determination of unknown compounds or for identifying fingerprints prior to SIM mode. On the other hand, SIM mode allows the detection of the intensity of few previously selected ions along the analysis time, providing a lower detection limit (targeted analysis) and less matrix interference, thus allowing quantitative analysis to be carried out. Nevertheless, it is important to always use reference standards to ensure identification of the mass fragmentation patterns. Generally, an internal standard is mixed with the sample, in order to increase the precision of analysis and to overcome the high instability between successive injections (e.g.: imprecision in sample volume due to small volumes and evaporation from injection needle). The internal standard used should be pure and not react with other components in the sample, have a similar retention time (same structural family), well resolved elution peak from other components, and obviously not be present in the original sample.

Before data analysis, it is important that the GC-MS parameters are optimized so that good resolution and separation is achieved. However, there are external factors that may also influence the experimental results. Among these factors are the high temperatures in several components of the system, carrier gas flow, stationary phase age and column dimensions.

1.2.4 Metabonomics and Metabolomics

In systems biology, one of the tools to control and explain biological complexity is metabolomics, providing information about molecular processes in cells, tissues, organs or organisms, focusing on metabolites and their profiles (leading to organism phenotypes). Fiehn states that metabolomics is "a comprehensive and quantitative analysis of all metabolites in a system"[47], while Nicholson uses another commonly used interchangeable term, metabonomics, described as the profiling of the whole metabolite composition in a living system, accounting for their changes due to environmental conditions and genetic background[48]. Independently of the term used, metabolomics studies offer the opportunity to quickly identify changes in the metabolome (range of metabolites present in a certain organism) due to the close relationship between the metabolic profile and the phenotype (observable properties of a certain organism). Metabolites are low-molecular-weight compounds that are involved in several

biochemical pathways, often found as reagents or products in metabolic processes, or as part of structures such as proteins, nucleic acids or organelles (membranes, cell walls…), having sometimes importance in signalling mechanisms. Thus, such molecules are fundamental in pathological studies and biomarker discovery, and their analysis may lead to the better understanding of diseases and consequently, to better and faster diagnosis methods or ultimately to drug development. When compared with other "-omics" studies, metabolomics has as main advantage the possibility of observing the direct consequence of system perturbations though the metabolome composition[49].

Metabolomics profiling is one of the strategies that can be used as a first approach to access the metabolome of an organism, and is carried out with the use of powerful detectors such as MS. Coupled with GC, this technique is one of the most widely used due to their adaptability to high-throughput analysis (typically less than 30 minutes per sample), low instrument cost (comparing with NMR) low sample cost and volume, good sensitivity and reproducibility, high reliability, ease of use, and the its transferable library (independent of the manufacturer). However, some limitations may be found since there is a need of sample pre-treatment (volatile samples are needed), it is a destructive method (sample cannot be recovered), and some maintenance is needed between runs. Ideally, in order to acquire the highest range of metabolites in the matrix, it would be necessary to involve other analytical methods (such as NMR or UPLC-MS, as example), but in the present study there is a special concern in the analysis of FFA in serum samples, which reduces significantly the range to analyse (35-450 m/z in our case) [49,50].

In order to develop the targeted analysis of FFAs in serum, derivatization is needed. However, the use of GC-MS is still vulnerable to surface contamination, mass calibration, and other effects that influence mass resolution and accuracy. In order to avoid this, repeatability conditions should be assured by the use of quality control (QC) samples. According to ISO 15189 "The laboratory shall use quality control materials that react to the examining system in a manner as close as possible to patient samples" and "Quality control data shall be reviewed at regular intervals to detect trends in examination performance that may indicate problems in the examination system."[26]. These guidelines provide the basis for QC of our method, allowing repeatability conditions for each obtained metabolic feature (chromatographic peak that results from retention index and mass spectrum), and consequently for each detected metabolite

(identified with one or more metabolic features). To reduce the bias in the data analysis, a randomization of the samples in each run and in between runs is necessary. In this way, systematic errors can be avoided and the proper performance of the analytical system can be evaluated by the continuous monitoring of the QC samples. In our work, these QC samples are made of a pooled mix of all the samples to be analysed in the study and undergo the same derivatization process as the study samples, being also spiked with internal standard (allowing to account for variability in sample processing and in the analytical system). In this way, it is ensured that the QC samples will behave in a similar way as the study samples.

In the quality assurance (QA) process, a continuous monitoring of QC samples will allow for outlier removal when significant drift in signal, mass or retention time is observed. Quality assurance is evaluated by establishing several parameters. If the process complies with these parameters, it can be considered under control; otherwise we face an out-of-control situation. The first parameter is to identify if the analytical platform is "balanced" (producing a reproducible signal) before the real samples are analysed, confirming that the system has all active sites saturated with the sample matrix. This is obtained by injecting a number of QC samples (approximately 5) in the beginning of each batch[51]. Secondly, there is the estimation of technical precision. International guidelines for bioanalytical method validation such as the ones from USA Food and Drug Administration (FDA) recommends a general value of 15% for the coefficient of variation (CV) and a CV of 20% near the lower limit of quantification (LLOQ) [52], but other studies mention that due to the derivatization process needed and losses in injection in GC-MS, 30% is an acceptable limit for technical precision[49]. In practical terms, CV and RSD are the same as long as RSD is represented as a percentage, and can be calculated through the ratio between the standard deviation (σ) and the mean (μ), as shown in equation 3:

$$(3)\ CV = RSD \times 100\% = {}^{\sigma}/_{\mu} \times 100\%$$

Nevertheless, a qualitative step was introduced by analysing QC samples in full scan mode (QC-scan) in each run, ensuring that metabolic features would be stable. In this work, technical precision was set to be lower than 20% RSD, and the QC-scan was compared with a standard FAMEs mixture. For the EFA from the serum samples that were not in the standard mixture, equivalent chain length (ECL) values were estimated

by calculation, and posterior identification was carried on by library comparison[53]. In this way is it possible to design the experimental QA process for the analytical method used.

1.2.5 Chemometrics: PCA, CCA and RF

Chemometrics is defined by Svante Wold as "how to get chemical relevant information out of measured chemical data, how to represent and display this information, and how to get such information into data". Massart adds that chemometrics aims at designing or selecting optimal measurement procedures and experiments and at extracting maximum information from chemical data[54,55]. These tools are fundamental for enhancing the achievements of metabolomics studies, however they should be applied to pre-processed data rather than to raw data formats. After running a GC-MS analysis, data is obtained as a 3D matrix of m/z versus retention time versus intensity, as shown in figure 11, in a format supplied by the manufacturer. Due to the lack of a universal standardized format for GC-MS data, the original format is usually converted into a format that is compatible with pre-processing software. Pre-processing is needed in order to remove irrelevant information/variables caused by several factors such as column aging, experimental variations or sample stability, and the most common procedures are baseline correction, chromatogram alignment, normalisation, smoothing and peak area integration[49,56–58].

Several types of normalisation procedures have been described in literature[59], but in biomarker research it is usually done by setting a default value to the value of internal standard (e.g. setting it to 1) and consequently establishing relative values to other metabolic features by dividing their original values by the value of the internal standard. In this way the relative magnitude of the signal will be revealed and if needed, concentrations can be calculated. Smoothing procedures are applied to increase the signal-to-noise ratio, promoting the detection of small signals that can be easily overlapped by noise. There are several methods used for smoothing, such as moving-window average smoothing or Savitsky-Golay filter (based on polynomial regression) [55]. Baseline correction/background correction is often needed due to swamping of the signals by the baseline drift, and this is done using algorithms like penalized least squares or adaptive iteratively reweighted procedure[55]. Nevertheless, chromatogram alignment is essential since it will correct for systematic errors that may occur during the analytical run and eventual peak shifts on the chromatogram. One of the most

accurate alignment methods is the recursive alignment by fast Fourier transform (RAFFT), where a global scale alignment is performed prior to a local scale alignment, having as reference the correlation values between test chromatograms and reference chromatograms. As a final alignment step, the optimal warping path is achieved by shifting the points of test chromatogram[60]. Peak integration can often be done on the analytical platform software but several alternatives can be found. This should be the last pre-processing procedure, after normalising, smoothing, correcting baseline and aligning chromatograms.

After pre-processing and prior to data analysis, it is always necessary to proceed to an identification of the metabolic features for FA, a frequently challenging step in metabolomics. Most of the identification methods are based on the comparison of mass spectra and retention indices with established libraries (e.g. National Institute of Standards and Technology (NIST)) or with authentic standards. However, libraries often differ between research groups, and results are not so accurate. On the other hand, the use of a large amount of standard compounds increases the costs of analysis. One alternative strategy is the equivalent chain length (ECL) values determination[53]. In this method, a special database is developed using the ECL values and mass spectra characteristics. The latter were used to recognize saturated FAMEs and consequently calculate the ECL values for unsaturated FAMEs in samples. This strategy overcomes the proximity (or overlapping) between retention indices and mass spectra of the isomers[55,61–63]. First, after doing the data pre-processing, SIM mode chromatograms are extracted in order to identify the expected saturated FAMEs. The retention times of these compounds are converted into ECL (the ECL value is defined as the number of carbon atoms in the FA carbon chain). Finally, a calibration is done using the retention times of the saturated FAMEs, and ECL values for each FA are calculated using the saturated FAMEs eluting before and after the compound of interest, as following:

$$ECL(x) = n + \frac{RT(x) - RT(n)}{RT(n+1) - RT(n)}$$

where n and n+1 are the number of carbons of the saturated FAME eluting before and after the target compound, respectively. Finally, the ECL value is then compared with the ones in the custom database for identification[53].

In general, metabolomics studies provide large datasets with several variables being measured for each sample, where the complexity demands more advanced statistical tools. Classic statistics may in some cases be insufficient to provide information about if a certain metabolite is discriminant or it is simply a statistical artefact[64]. When evaluating if FFAs or even if the clinical parameters are statistically different in healthy control samples when compared with MetS patients, we face a multiple comparison problem due to the simultaneous comparison of several variables. To find statistical differences is particularly useful in the way that allows the researcher to understand if, in the case of clinical parameters, the diagnosis of a disease makes sense and there are mathematical differences between the healthy people and the patients. It is also useful for the identification of potential biomarkers, as in the case of finding that a particular FFA as significantly different value in between MetS patients and healthy controls. These statistical differences are analysed by testing the so called null hypothesis, indicated as H_0, which refers to the case in which the values of each group are not different. A statistical test is then applied in order to decide to accept or reject H_0, resulting in a p-value. A typically used statistical test is the t-test which compares the mean values of each group, and sets a α threshold for the confidence level. Typically α is set to 0.05, and the respective confidence level is 1- α which in this case is 0.95, also represented as a percentage as 95%. This means that H_0 can be rejected with a confidence level of 95%. However, there is still the 5% probability of wrongly rejecting H_0, and this is called a type I error (false positive). On the other hand, a type II error may happen when H_0 is accepted and in fact it should be rejected (false negative). As mentioned before, in multiple comparisons these classical approaches are not applicable. The probability of finding significant differences, by chance, increases. One of the solutions for this problem is based on the principle of family-wise error rate (FWER), which refers that if two classes are not different, their variables should not show statistical significant difference. In order to control FWER, Holm proposed an improved method to Bonferroni approach known as Bonferroni-Holm correction[65], in which a step-wise approach based on the p values is used, as shown in figure 12. This method uses a different α for each variable, being more restrictive with variables that show a smaller p value. The smallest p is compared with α_{UV}/m (Bonferroni limit) in which α_{UV} is the probability of error for each individual test and m is the number of individual tests. In each step, this limit is made less conservative[66].

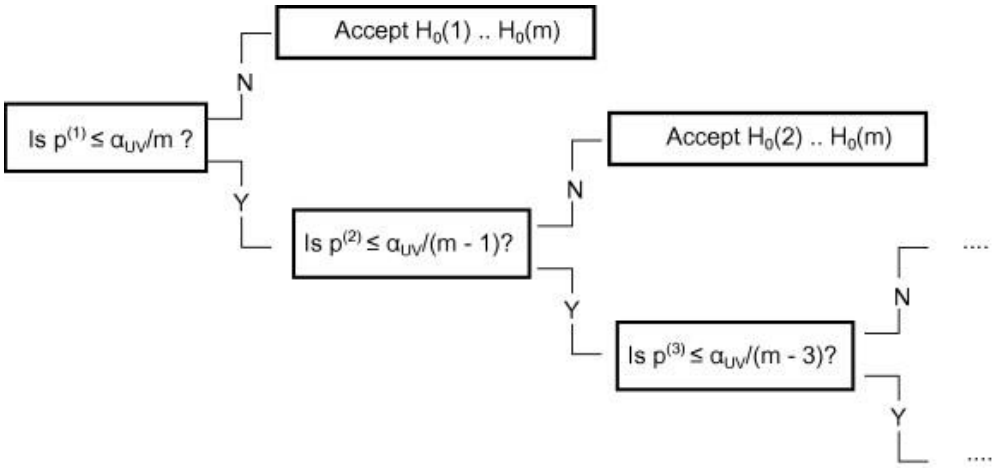

Figure 12: Holm-Bonferroni step-wise approach to control FWER [66]

When testing the correlation of two sets of variables (like FFAs and clinical parameters), one must variance is equal in both variable sets. One way of doing this is by applying Bartlett's test, making an approximation to the chi-squared distribution[67].

This is where chemometrics, including multivariate data analysis come into play. In a traditional case, each object (sample) can be represented graphically with its variables (measurements), as shown in figure 13. By representing the values the two variables x and y in space, it is possible to obtain a vector r from the origin until the point matching the variables values. When several objects have similar measurement values, they will be located graphically close to each other forming a cluster. Problems with simple representations happen with the number of variables increases, necessitating the use of computer analysis and three dimensional plots. Nevertheless, a high number of variables increase the chance of having unnecessary information. Multivariate data analysis provides methods that enable data reduction through the elimination of redundant information[68].

Principal component analysis (PCA) is one of the most popular techniques used as a starting point in metabolomics exploratory data analysis, and generates principal components (PCs) as linear combinations of the initial variables[50,68].

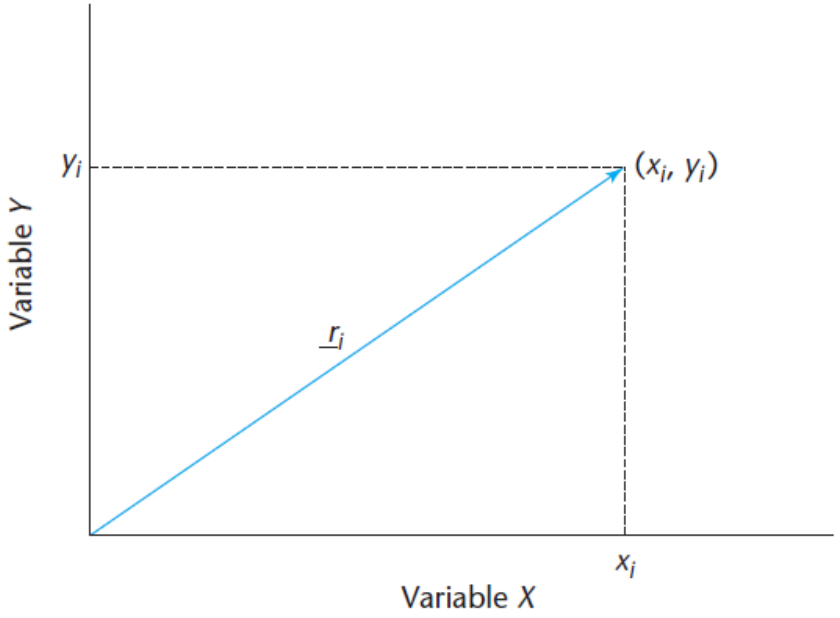

Figure 13: representation of variables x_i and y_i projecting the vector $\underline{r}_i$[68]

As shown in figure 13, the vector $\underline{r}$ can be projected taking coordinates from the original variables x and y. Each specimen X is derived from these original variables and can generate the following PCs by linear combination:

$$PC_1 = a_{11}X_1 + a_{12}X_2 + \ldots a_{1n}X_n$$

$$PC_2 = a_{21}X_1 + a_{22}X_2 + \ldots a_{2n}X_n$$

etc.

where a_{11}, a_{12}, a_{21}, a_{22}... are the coefficients of the independent new variables X_1, X_2 and X_n. Each PC will contain the variance on the initial dataset, being the first PC the one that contains most of the variance, and the last PC with the least variance. By definition, each PC is orthogonal to each other (PCA is an orthogonal transformation), as shown in figure 14(a), and accordingly they can also be displayed as axes as in figure 14(b). A representation of data points onto these axes will show clearly where most of the variance resides. From the simultaneous variance of two variables it is possible to calculate the coefficient of variation (CV), and a CV matrix will then allow obtaining the eigenvectors (each eigenvector is a PC) and their respective eigenvalues (dataset variance explained by each PC). On the other hand, the eigenvectors are described by a set of coefficients that when multiplied by the variable values of a sample will result in scores. When a good PCA is possible, the scores plot often helps to reach conclusions

about classification of the variables and their interaction, as shown in figure 15, due to their relative location in the plot, making this a good method for visualization of trends[58,68].

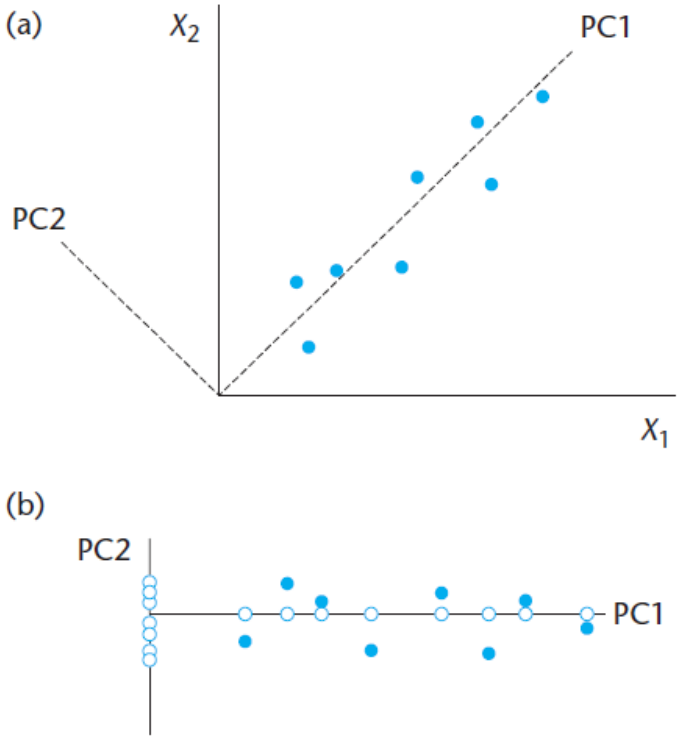

Figure 14: representation of Principal Component 1 and Principal Component 2 (a) relatively to x_1 and x_2, (b) with points projected on Principal Component axis, where balls are data points and circles are their projection in each axis[68]

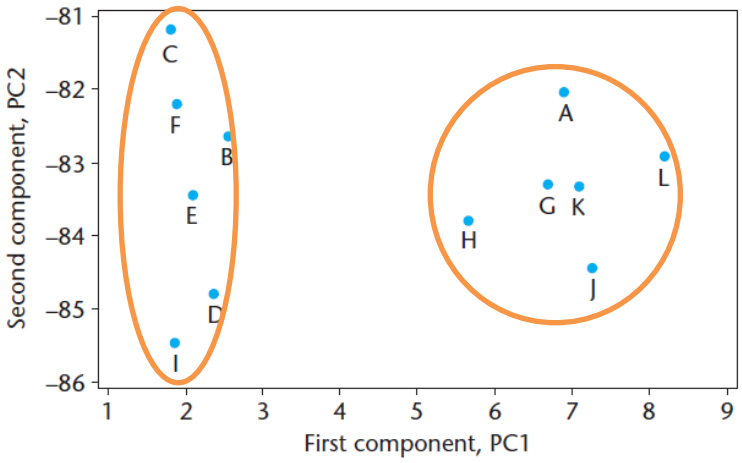

Figure 15: scores plot of Principal Component 1 and Principal Component 2[68]

In recent studies it is possible to find novel classification methods based on machine learning such as random forests (RF)[69]. Introduced firstly by Breiman in 2001, this method is able to classify a dataset by generating random vectors by independent sampling. Each vector defines a tree selected from a forest (origin of the method's name), having the same distribution as all the other vectors existing in the forest. This tree is then called a predictor. The RF method is based on two distinct techniques: decision tree learning and tree bagging. Decision tree learning is the process of generating a model that predicts the value of a target variable based on the input, and afterwards splitting this input in smaller groups based on the value of the test. This splitting is repeated iteratively until the value of the smaller group equals to the value of the target variable or when the split shows no change in the value of the predictors. It is independent of scaling and some other transformations, which makes it a robust method, but the main drawback of this procedure is the low accuracy and the possibility of over-fitting the training set (low bias). In other hand, tree bagging was developed to add accuracy to the previous method and avoid over-fitting. This last process consists of the generation of new training sets from the original data set, and consequent uniform sampling with replacement. By definition, this type of sampling will generate a bootstrap sample, and each one of them will then lead to a model, which will ultimately be voted to provide classification. In short, the more classification trees are built, the more accurate becomes the classification process. The estimation of this accuracy is done internally by using out-of-bag (OOB) estimation, avoiding the need for cross-validation. Cross validation in many cases requires the removal of at least one sample from the dataset prior to the model construction (PC calculation) and then verify how well the removed sample can be predicted by the model[70].

Proximities among samples are also possible to be calculated. The number of times two samples are classified in the same terminal node of a tree, divided by the total number of trees is the value of proximity among these samples. The graphical representation of these proximities (similarities and dissimilarities) is done by MDS (multidimensional scaling) plots, which represent the proximity set of points in Euclidean space based in nonmetric information[71,72]. The RF algorithm for classification is introduced as[57,73,74]:

a. Draw n_{tree} bootstrap samples from the original dataset. (n_{tree} is the number of trees in the forest. In our study it was set n_{tree}=600)

b. For each of the bootstrap samples, grow an un-pruned classification tree, and modify with the following procedure: at each node, randomly select m_{try} variables and choose the best split from among those variables. (For classification, the default value of m_{try} is equal to sqrt(m_{all}) where m_{all} is the total number of the variables in the original dataset. In this study, m_{all}=28, thus, m_{try}=5).
c. Predict new data by aggregating the predictions of the n_{tree} trees.

An example of the process in decision trees is shown on figure 16, from the study of Worachartcheewan et al[75]. In this case the tree is designed for classification of MetS status, and the root node is the TG diagnosis factor. Depending if the sample has a value higher than 149 mg/dLor not, it will lead to a different internal node. Each decision may lead to another question or to a terminal leaf node for decision.

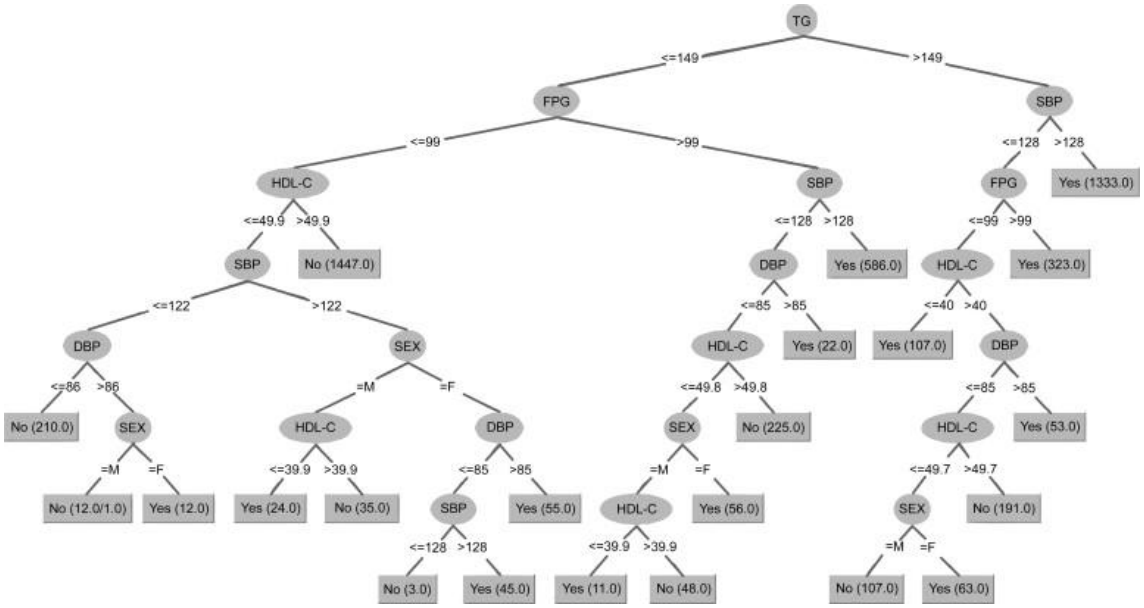

Figure 16: Metabolic syndrome status decision tree classification, in which Yes means MetS presence, No means non-MetS. The final output of Yes or No is called terminal leaf node and contains one or two numbers in parentheses, which represent the number of correctly and incorrectly classified samples, respectively[75].

As mentioned before, RF algorithm draws a number of tree bootstrap samples for classification, but at each node it randomly selects variables and chooses the best split, and afterwards aggregates the predictions of all the trees for predicting new data. From the final results for the RF model, it is possible to calculate its sensitivity, specificity

and accuracy. Classification accuracy is related with the overall percentage of correct classification, while sensitivity describes the percentage of correctly classified variable in a certain group, and specificity represents the correct classification of a variable out of a certain group.[76]

After obtaining the analyte profiles in the samples, a correlation between these profiles and diagnosis is necessary for producing significant results and knowledge about the disease mechanisms. A classical method of performing this kind of data analysis is canonical correlation analysis (CCA)[73,77,78]. This method works through cross-covariance matrices, finding the maximum correlation by generating linear combinations of the studied variables (in our case x and y would be the studied variables matching FFA and clinical parameters, respectively). After mean centring the data, the calculated covariance matrix can be represented as:

$$C = \begin{bmatrix} C_{xx} & C_{xy} \\ C_{yx} & C_{yy} \end{bmatrix} = E\left[\begin{pmatrix} x \\ y \end{pmatrix} \quad \begin{pmatrix} x \\ y \end{pmatrix}^T \right]$$

where C_{xx} and C_{yy} are the within-sets covariance matrixes of x and y, and $C_{xy} = {C_{yx}}^T$ which represent the between-sets covariance matrix. The CCA will provide then the eigenvalue equations that will result in the correlation values:

$$\begin{cases} {C_{xx}}^{-1} C_{xy} {C_{yy}}^{-1} C_{yx} \hat{a}_x = \rho^2 \hat{a}_x \\ {C_{yy}}^{-1} C_{yx} {C_{xx}}^{-1} C_{xy} \hat{b}_y = \rho^2 \hat{b}_y \end{cases}$$

Where ρ^2 is the square of canonical correlations (eigenvalues), and the eigenvectors $\hat{a}_x$ and $\hat{b}_y$ are the normalized canonical correlation basis vectors. The smallest dimensionalities of x and y will ultimately define the number of possible solutions to these equations. Finally, the result is given as the canonical variables U and V, with U=Xa and V=Yb.

2. Experimental

2.1 Materials and Reagents

In the determination of the FAs, a Supelco FAME mix (No. 47885-U) of 37 component, methyl 10-nonadecenoate and a cis-10-nonadecenoic acid (C19:1n-9c, ≥99% purity) internal standards (IS), methyl heptadecanoic acid and heptadecanoic acid (C17:0, ≥99% purity) IS, were purchased from Sigma-Aldrich (St. Louis, MO, USA). For the derivatization steps, a 5% H_2SO_4/CH_3OH solution was freshly prepared through dilution of H_2SO_4 (98,0% purity, Zhuzhou Chemical Industry Institute, Zhuzhou, China) with chromatographic grade methanol (Tedia Company, Inc, Fairfield, USA), and a 0.4 M KOH/CH_3OH solution was freshly prepared by dissolving analytical grade KOH in chromatographic grade methanol. Analytical grade n-hexane and anhydrous sodium sulphate were commercially obtained from Tianjin Yongda Chemical Reagent Co, Ltd (China) and Tianjin Kernel Chemical Reagent Co, Ltd (China), respectively. All the procedures were performed using appropriate routine laboratory glassware, washed accordingly to routine methods established by the laboratory (including rinsing with tap water for large particle removal, methanol immersion and incubation for 30 minutes at 25 ºC in an ultrasound cleaner, washed with de-ionized water and dried overnight in an oven incubator).

2.2 Serum Samples

2.2.1 Sample collection

The serum samples were collected under pre-analytical overnight fasting requirements for blood collection[79], and according to the established standard protocols[6], from 100 Chinese subjects, aged between 25 and 70 years, at the Xiangya Hospital of Central South University in Hunan province (China), from February 2012 to March 2013. A sample of venous blood (2 mL) was collected in a blank tube without anticoagulant or preservative, and was allowed to clot at 4 ºC for 1 h in order to obtain the serum. The MetS diagnosis was done following the guidelines on prevention and treatment of dyslipidaemia in Chinese adults[80] adapted from the most recent definition of MetS given by CDS in 2004[81] and by IDF in 2005[82]. Accordingly, MetS patients were defined whenever three of the following risk factors were fulfilled: central obesity (waist circumference > 90 cm in men, > 85 cm in women), increased blood pressure

(SBP/DBP ≥ 130/85 mmHg or previously diagnosed hypertension), increased fasting blood glucose (FBG ≥ 6.1 mmol/L or previously diagnosed type 2 diabetes), increased serum triglycerides (TG ≥ 1.70 mmol/L) and decreased HDL-cholesterol (HDL-C ≤ 1.04 mmol/L). The subjects were divided into two groups: healthy controls (HC) group where the subjects had 0 risk factors and MetS patients group where the subjects had at least 3 of the 5 risk factors. All participants have written an informed consent and Xiangya Institutional Human Subjects Committee approved all clinical experiments.

2.2.2 Sample preparation and derivatization

After collection, the serum samples were stored at -80 ºC, since previous studies in the research group shown that higher temperatures do not allow enough FA stability[83]. Sample preparation was carried out according to an adaptation of the procedures established by the research group in previous works for FA analysis by GC-MS, as shown in figure 5[25]. Initially, a pooling QC sample was prepared by addition of 10 µL of each of the serum samples in our study, followed by 1 min vortex mixing[49,84]. Aliquots (50 µL) of serum were spiked with 10 µL of 0.1 mg/mL C19:1n-9c IS (in n-hexane), added to 0.5 mL of 0.4 M of KOH/CH_3OH, 30 s vortex mixed and incubated for 10 min at room temperature. Afterwards, 0.5 mL n-hexane was added to the previous mixture and vortex mixed for 30 s, followed by isolation of the serum phase. This step was repeated using the remaining n-hexane phase, and both serum phases were collected to the same glass tube. Anhydrous sodium sulphate was added to the resulting serum phase followed by vortex mixing, addition of 0.5 mL of 5% H_2SO_4/CH_3OH and vortex mixed for 30 s. The mixture was kept in a 70 ºC water bath for 30 min and then placed at room temperature for temperature stabilization. To the room-temperature mixture was added 1 mL of n-hexane twice with 30 s vortex mixing after each addition, the n-hexane phase being isolated each time in the same glass tube. The resulting solution was evaporated to dryness under N_2 flow, and the remaining sediments (FAMEs) were reconstituted with 100 µL of n-hexane for GC-MS analysis.

2.3 GC-MS analysis

2.3.1 Optimization and Repeatability

In order to study the repeatability of the method, a first approach included the derivatization and analysis at the facilities of the Research Centre of Modernization of Chinese Medicines at the College of Chemistry and Chemical Engineering of Central

South University, using a Shimadzu GC2010A (Kyoto, Japan) GC instrument coupled to a GCMS-QP2010 MS (Compaq-Pro Linear data system, class 5K software). A DB-23 capillary column (30 m x 0.25 mm id, film thickness 0.25 µm, Agilent Technologies, USA) was used for the FA analysis. Test serum samples from previous approved studies in the research group were used in this initial phase and derivatization methods were developed according to the previous described method. Repeatability was carried on by ensuring the execution of each derivatization step following the established procedures and checking the RSD of the spectra obtained with this analytical system. In this initial approach IS was C17:0 and C17:0 methyl ester. Optimizations in this process included changes in reagents and sample quantity without losing significant signal intensity. The optimized temperature program in this GC-MS system is described in table 1, achieving a final GC run time of 16.5 min. With an inlet temperature of 250 ºC, the injection was performed in split mode with a split ratio of 1:10 in the case of methyl ester FA and 1:5 for non-methylated FA, for 2 µL of injected sample. The Helium (99.999% purity) carrier gas was kept at a constant flow rate of 1.0 mL/min, while the MS conditions were: ion source, quadrupole and interface temperature of 230 ºC, 150 ºC and 250 ºC, respectively, detector voltage of 1 kV, MS program starting at 4 min and finishing at 16.5 min, a velocity of 0.2 s/scan, and in SIM mode the selected characteristic m/z for FA profiling were 55, 67, 74 and 79, while in full-scan mode the range was 35-450 m/z.

Table 1: GC-MS optimized temperature program on the Shimadzu analytical system

Start Time (min)	Rate (ºC/min)	End Time (min)	Final Temperature (ºC)	Hold Time (Min)
0	0	0	70	1
1	20	4	150	0
4	6	9	180	0
9	20	11.5	230	5

2.3.2 Internal Quality Control and Quality Assurance

Some internal quality control procedures were implemented in order to ensure data reliability. N-hexane was selected as the reagent blank (B) and QC samples were in each analytical run. Every batch consisted of 12 study samples randomly selected, preceded by a minimum of 3 QC sample injections and one QC-scan, including a QC

sample every fourth study sample injected and ended with two consecutive QC sample injections as shown in the experimental design in table 2. Apart from the QC-scan, all other injections were acquired in SIM mode.

Table 2: Experimental design of sample order for each batch

Injection Nr	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
Sample Type	QC	QC	QC	QC-Scan	S_1	S_2	S_3	B	QC	S_4	S_5	S_6	B

Injection Nr	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**
Sample Type	QC	S_7	S_8	S_9	B	QC	S_{10}	S_{11}	S_{12}	QC	QC	B	B

Nevertheless, after data acquisition and pre-processing, only metabolic features that fulfil the quality assurance (QA) requirement of peak area RSD < 20% were able to proceed for subsequent data analysis.

2.3.3 Serum FA profiling of study samples

After doing an initial optimization of the Shimadzu system, studying the method's repeatability and establishing the experimental design, an optimization of all processes in the new analytical system at the facilities of the Xiangya Hospital of Central South University was carried out. The GC-MS analyses used an Agilent 7890A GC coupled to an Agilent 5975C quadrupole mass spectrometer, using the same capillary column as in the preliminary testing with the Shimadzu system. The injector temperature was 250 ºC, and it was operated in split mode at a ratio of 1:5 for a sample injection volume of 2 µL. Helium (99.999% purity) was used as carrier gas at a 1.0 mL/min flow rate. The resulting optimized temperature program is shown in table 2, with a total program time of 16.83 min.

Table 3: GC-MS optimized temperature program on the Agilent analytical system

Start Time (min)	Rate (ºC/min)	End Time (min)	Final Temperature (ºC)	Hold Time (Min)
0	0	0	100	1
1	20	3	160	0
3	6	6.33	180	3
9.33	20	11.83	230	5

In the MS component, the transfer line, quadrupole and ion source temperatures were 250 ºC, 150 ºC and 230 ºC, respectively, and the electron impact (EI) energy was of 70 eV. In SIM mode the selected characteristic m/z for FA profiling were 55, 67, 74 and 79, and in full-scan mode the range was 35-450 m/z. In this phase the IS was switched to C19:1n-9c, and only FFA were considered for analysis and further identification.

2.4 FFA identification

The serum FFA identification was achieved using a combination of a 37 component FAME standard mixture with a previous developed method in our research group[53]. The method included the automatic recognition of the expected saturated FAMEs through their mass spectra, followed by the use of the respective retention times to calculate their ECL values and consequently being able to interpolate the ECL for the unsaturated FAMEs in the samples. The resulting ECL values and characteristic ions were then used for library identification.

2.5 Data Analysis

After GC-MS analysis, it was necessary to convert the raw .D format data files into NetCDF format, allowing to do the pre-processing using the Traditional Chinese Medicine Fingerprint Database System (TCMSys v1.0.2) developed by the research group. Using this software, smoothing, baseline correction, peak shift alignment, normalization and peak area integration was carried out. The normalization step was done using the IS, so the metabolite peak areas were always represented as relative intensities to the IS peak area. After QA, the resulting data matrix contained 42 HC, 58 MetS patients and 49 QC samples, which were afterwards imported into MATLAB 7.10

(The MathWorks, Inc., Natick, MA) for multivariate statistical analysis using custom scripts.

With the pre-processed autoscaled data, it was possible to proceed to the PCA where the general data trends were observed by the generation of a score plot of PC1, PC2 and PC3. Then, RF was applied for classification and obtaining a prediction model for MetS, setting the parameters of n_{tree}=600 and m_{try}=5, and building an MDS plot of the results. Finally, CCA was carried on for unveiling the relationship between FFA and the clinical parameters. A scatter plot was obtained for all dataset, and absolute value of coefficients for both FFA and clinical parameters were represented in a bar plot for identifying potential biomarkers and important clinical parameters in disease diagnosis.

3. Results and Discussion

3.1 MetS clinical parameters

Metabolic syndrome has been reported as cluster of risk factors that lead to cardiovascular diseases and type 2 *Diabetes Mellitus*. These risk factors are currently used for the diagnosis of this syndrome and their critical limits are established differently around the globe, being most definitions published by IDF, WHO, NCEP:ATPIII and CDS. Previous works on the general Chinese population have reported the prevalence of MetS using as reference ten city regions and ten rural counties, and a four-stage stratified sampling method. This survey accounted with a representative sample of 15540 people examined under NCEP:ATPIII MetS definition and concluded an incidence of 9.8% in men and 17.8% in women, being the higher incidence in northern China and in urban areas[6]. In our study[73], the population undergoing analysis was original from Hunan province in China, and used the CDS definition with some adjustments. In comparison with Gu et al (2005), the clinical parameters for MetS diagnosis were the same with the exception for the limits of HDL-C, which were not so strict in our study keeping a value of 1.04 mmol/L for both genders while in the former study these values can go up to 1.3 mmol/L in women, being the latter more restrictive. For the limits of waist circumference were more restrictive with values of 90 cm for men and 85 cm for woman in our study, and being allowed to reach 102 cm in men and 88 cm for women in Gu et al (2005). The use of these waist circumference limits in the present study is seen here as more suitable for the study population, and this becomes evident when taking into account that different populations have characteristic genetics and different diets that lead to a change in the anthropometric values that are considered as normal. At the Xiangya Hospital of Central South University, the sample collection was preformed and MetS was diagnosed whenever patients had three of the following five risk factors: waist circumference (WC) > 90 cm in men or > 85 cm in women, SBP/DBP ≥ 130/85 mmHg or previously diagnosed hypertension, FBG ≥ 6.1 mmol/L or previously diagnosed type 2 diabetes, TG ≥ 1.70 mmol/L, and HDL-C 1.04 mmol/L. As a result, it was possible to calculate the average values for these risk factors for the MetS and HC individuals, as shown in table 4.

Table 4: Average values for clinical parameters of the study individuals

Clinical parameters	HC (n = 42)	MetS (n = 58)	P values
WC (cm)	76.76 ± 6.87	91.19 ± 4.93	1.1315e-20
SBP (mmHg)	111.33 ± 9.90	133.00 ± 14.25	9.9910e-13
DBP (mmHg)	70.95 ± 6.89	85.03 ± 11.00	2.1114e-10
TG (mmol/L)	0.89 ± 0.41	3.60 ± 2.03	9.9910e-13
HDL-C (mmol/L)	1.39 ± 0.28	1.06 ± 0.23	4.1058e-09
FBG (mmol/L)	4.89 ± 0.42	6.91 ± 2.57	2.1097e-06

values given as mean ± SD for a confidence level of 95% (P values < 0.05) from T-test and adjusted by Bonferroni-Holm correction for multiple comparisons.

These values suggest that there was a significant difference between average values for healthy controls and metabolic syndrome patients. In the case of WC, the values seem to present the most significant difference in this table, but may not be suitable for comparison with the limits established by the risk factors, since these are different between genders. It can clearly be observed here that TG have the values that most exceed the critical risk factor limits, with a value of 3.60±2.03 mmol/L exceeding the limit of 1.7 mmol/L, and concomitantly presenting a high significance level. Triglycerides play an important role in dyslipidaemia processes, being storage molecules for FFA. A large quantity of TG may indicate also a large amount of circulating FFAs involved in the metabolic disturbance caused by MetS.

3.2 GC-MS profiling

After diagnosis, sample collection and storage at -70ºC, sample derivatization took place. In a first approach to achieve repeatability and optimized conditions, sample derivatization was performed at the Research Centre of Modernization of Chinese Medicines. This derivatization was based in previous studies developed in the research group[25], using a Shimadzu GC-MS system. After the first round of tests it was possible to confirm that C17:0 methyl ester and C17:0 were not present in the sample and were available in the laboratory stock, thus it was decided to use these FA as IS. On the other hand, the initial approach in this study aimed at analysing both NEFA and EFA from the serum samples in order to confirm previous findings that state the inability of EFA to discriminate between disease and control groups in dyslipidaemia processes[25]. Accordingly, the initial procedure included also the analysis of the first extract obtained

in the derivatization step instead of discarding it. In the end of this first period, experimental conditions allowed to achieve repeatability of the method and optimization of the temperature program for FA analysis.

When moving to the Xiangya Hospital laboratories, a new Agilent GC-MS system was employed in the analysis. Thus it was necessary to also test the method's repeatability and optimization of temperature program. In this period it was necessary to clean the MS detector according to the manufacturer instructions, replace and clean the liner as a routine procedure, replace injector septum regularly, replace the gold seal at the beginning of the experiment, install the column that was used in also in the Shimadzu system and preform the tuning of the MS detector. After these procedures, the analytical system showed good behaviour during the experiment; however two main problems were identified. The first one was identified from the tuning report, which has shown high levels of air and water. After checking all the possible leaks, it was observed that the graphite ferrule, responsible for the isolation of the connection between the column and the transfer line that leads to the detector, was broken, having half of it around the column and half of it inside the transfer line. This problem happened due to an incorrect adjustment of this connection, most probably due to over tightening the nut. The high temperatures in these parts of the system lead to the blockage of the transfer line by the melt ferrule, problem that was solved by inserting a small tool to push this part out. The tuning preformed after the correction of this situation confirmed the inexistence of any leakages. The second problem was a large overlapping band on the right side of the spectra that was identified as Cholesterol. After the study of this problem, it was concluded that the liner cleaning should be done in a more frequent manner, and in order to correct this problem it was necessary to run several blank samples (n-hexane) for column "washing". After fixing these problems, and confirming the appropriate performance of the equipment, the experiment was started.

During the optimization period with the Agilent system there was the analysis of some real samples without IS. These samples revealed an interesting result, which was not shown with the Shimadzu system, probably due to the higher resolution capacity of the Agilent system. The real samples without IS contained C17:0 in their profile. This fact required the change of the selected IS to C19:1n-9c, not present in the real samples. An example of the FFA profile is shown in figure 17.

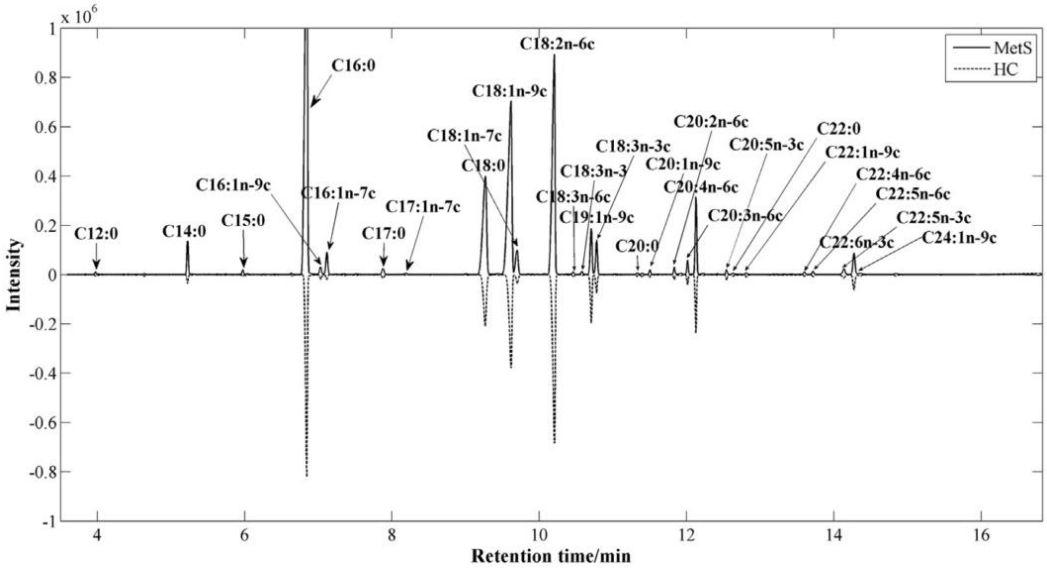

Figure 17: Free fatty acid methyl ester profiles by GC-MS from MetS patients (plotted upwards) and healthy controls (plotted downwards)

As can be seen from the previous figure, it is included C19:1n-9c, spiked as an IS. MetS patients possess a wide variety of FFA in comparison with HC, an expected result since this disease has a characteristic lipid dysregulation. This spectrum also supports the results from the clinical parameters in table 4, in which the most relevant risk factor for MetS was the high TG values.

At this point, profiling was only done with NEFA due to time restrictions and consequently, EFA were discarded and not used for analysis. This decision was also taken considering that EFA analysis was not part of the main objectives of this study and taking into account the fact that previous works have proven their inefficiency in discrimination of metabolic dyslipidaemia-related diseases[25]. The profiling resulted in the identification and quantification of 28 FFAs, as shown in table 5, 22 of them being validated with reference standards.

Table 5: Free fatty acids profile of MetS patients and healthy control individuals.

RT (min)	FFA	HC (n=42) (µg/mL)	MetS (n=58) (µg/mL)	QC RSD (%)	P value
3.971	C12:0	0.40±0.23	1.02±0.57	9.51	8.9558e-07
5.217	C14:0	2.59±0.94	5.94±4.13	8.52	2.7682e-05
5.966	C15:0	0.67±0.14	0.96±0.52	8.91	0.0040
6.832	C16:0	89.14±18.86	142.96±57.47	5.45	1.7110e-06
7.023	C16:1n-9c	1.35±0.40	2.31±0.73	9.44	3.6508e-10
7.103	C16:1n-7c	2.88±1.33	5.49±3.03	6.36	2.2527e-05
7.862	C17:0	1.46±0.30	1.89±0.58	6.20	3.8969e-04
8.184	C17:1n-7c	0.20±0.07	0.40±0.30	11.49	8.6202e-04
9.242	C18:0	37.39±6.78	49.78±16.99	5.09	3.1527e-04
9.601	C18:1n-9c	56.12±17.59	95.00±39.14	5.14	8.9558e-07
9.682	C18:1n-7c	4.30±1.31	7.73±3.56	5.33	1.0329e-06
10.186	C18:2n-6c	72.34±14.68	102.76±45.08	5.54	7.3969e-04
10.452	C18:3n-6c	0.45±0.23	0.81±0.46	7.29	1.3319e-04
10.571	C18:3n-3	0.40±0.19	0.69±0.45	6.24	0.0020
10.764	C18:3n-3c	4.12±1.87	7.39±4.97	6.63	0.0011
11.315	C20:0	0.47±0.15	0.71±0.51	9.18	0.0322
11.487	C20:1n-9c	0.74±0.35	1.24±0.61	5.78	1.3243e-04
11.822	C20:2n-6c	1.33±0.31	1.88±0.63	4.37	2.4112e-05
12.003	C20:3n-6c	2.83±1.00	4.27±1.77	6.79	1.3905e-04
12.117	C20:4n-6c	15.08±2.89	17.55±5.43	6.40	0.0462
12.533	C20:5n-3c	1.07±0.72	1.49±1.14	5.93	0.0948
12.622	C22:0	0.54±0.15	0.67±0.37	7.76	0.0948
12.801	C22:1n-9c	0.66±0.20	0.85±0.42	7.35	0.0462
13.588	C22:4n-6c	0.49±0.11	0.74±0.32	5.79	1.3905e-04
13.702	C22:5n-6c	0.67±0.23	0.96±0.45	5.94	0.0027
14.119	C22:5n-3c	1.80±0.41	2.38±0.73	4.45	1.5782e-04
14.257	C22:6n-3c	6.23±2.19	7.26±2.62	7.83	0.0740
14.345	C24:1n-9c	0.49±0.16	0.56±0.16	7.53	0.0688

concentrations reported as mean±SD, for a confidence level of 95% (P value < 0.05), adjusted by Bonferroni-Holm correction for multiple comparisons. Underlined FFAs showen no statistical significance.

From the 28 identified peaks it is possible to observe that 4 of them did not present significantly different values (underlined in table 5) between MetS patients and HC, with P values for Bonferroni-Holm correction exceeding 0.05. From the remaining 24 peaks, 22 were identified using reference standards and 2 were identified using the ECL method based in the retention indices/retention times and mass spectra information[53]. As mentioned before, MetS patients have displayed a wider diversity of FFAs than HC, and when present in both groups of samples, FFAs often presented higher values for MetS patients.

3.3 Multivariate data analysis: PCA, RF and CCA

In order to confirm the general trend that shows a higher diversity and quantity of FFAs in MetS patients, it was developed a PCA using a MATLAB script. Before PCA, the data was autoscaled (a process of standardisation of the variables in order to minimize the dominant effect of metabolites with higher intensity, establishing a similar significance among metabolite changes)[70] and a data matrix of 149 samples (42 HC, 58 MetS and 49 QC) and 28 variables (the detected metabolic features shown in table 5) were considered for further analysis. The score plot of the first 3 PCs shown in figure 18 was obtained. Accordingly, the first 3 PCs were able to explain 78.25% of all variance in data (PC1 63.68%, PC2 9.48% and PC3 5.09%). Although PC1 represents most of the variance, it is possible to observe that the separation between the clusters of the 3 groups of samples happens in the PC2 direction, with a distribution of the MetS samples in the negative side, the HC samples in the positive side and the QC samples concentrated in the origin. This analysis helps to understand some differences between groups of samples; however their classification is not clear enough.

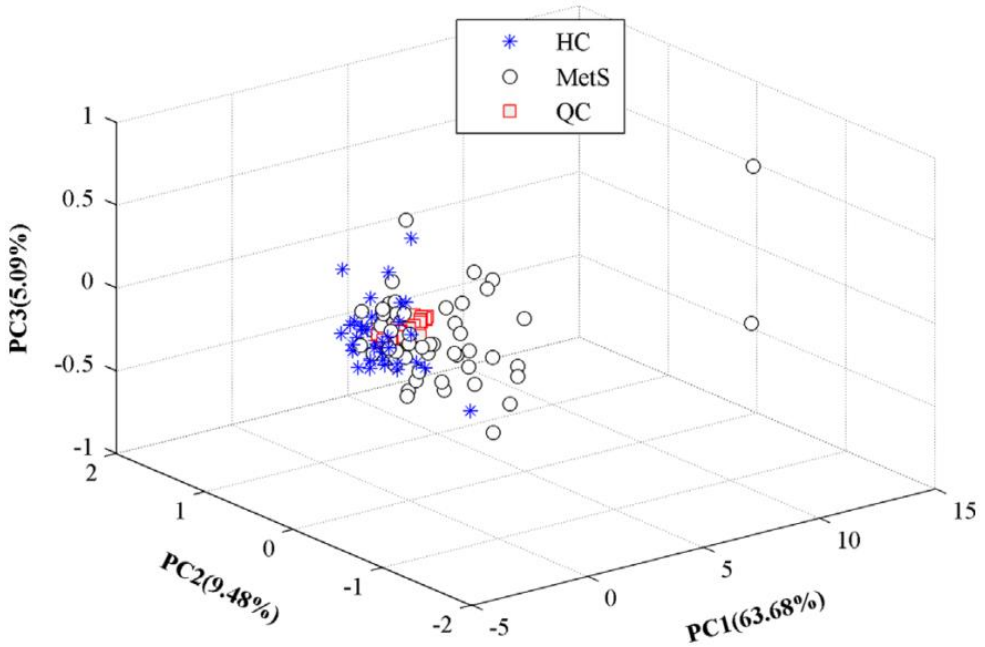

Figure 18: PCA scores plot of MetS, HC and QC samples from serum FFA profiles.

After analysing the initial trends in the dataset, it was applied a more efficient classification method: random forests. This methods main advantage can be the exemption of cross-validation, allowing an accurate classification and to achieve prediction models. As described before, RF is based in the iterative process of tree building based on bootstrap samples, and afterwards to test the fitness of the data in each tree, calculating the proximities between data from each branch of the tree. This machine learning technique was applied using also a MATLAB script on the dataset, and an MDS plot was built to represent the proximity matrix in a lower-dimensional space, as shown in figure 19.

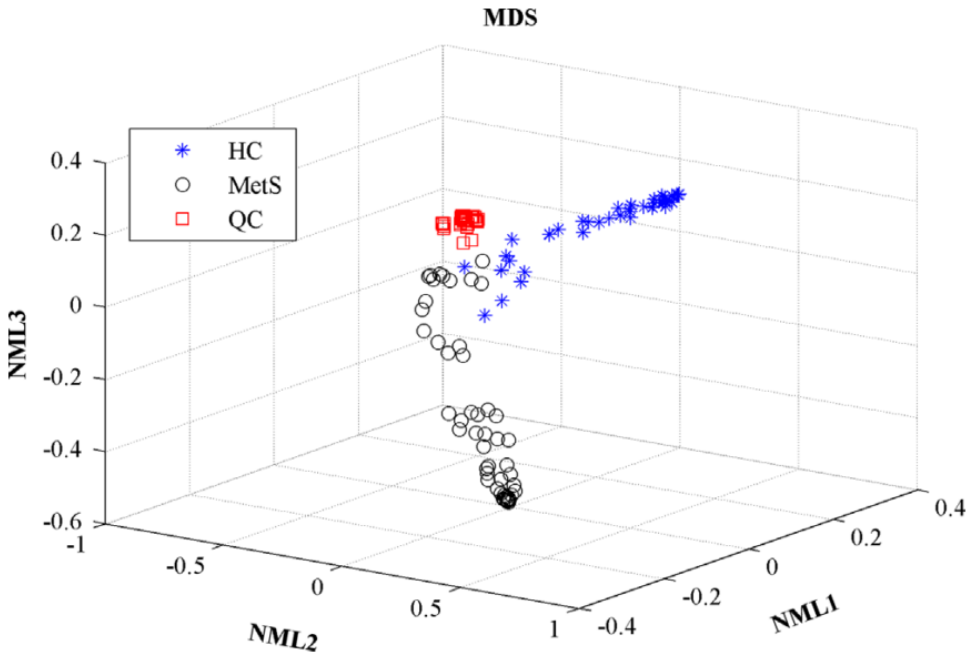

Figure 19: MDS plot of HC, MetS and QC samples from serum FFA, presented in 3 normalized maximum likelihood (NML) axis

By observing the MSD plot, it is possible to confirm that a better classification was achieved in comparison with the previous PCA method. In both methods, QC samples are able to cluster tightly together, which means that along the process the repeatability conditions were kept under control, with minimum bias. Furthermore, both groups of samples HC and MetS were able to cluster together with no major dispersion, obtaining a clear classification. The prediction model developed with RF resulted in a sensitivity of 0.8611, specificity of 0.8644 and accuracy of 0.8632.

In a third data analysis step, it was explored the importance of each of the clinical parameters and each of the FFA in MetS by CCA. This method allows to access the correlation between the two sets of variables (clinical parameters and FFA), by generating linear combinations of these. As shown in table 4, five risk factors take part in the diagnosis of MetS (SBP and DBP are considered as the same clinical factor, although they have their own limit values) and, as discussed previously, these are often related with general dyslipidaemia processes which involve also FFA. In this work, CCA is applied in the attempt of unveiling which degree of correlation may exist between these, and to identify potential biomarkers in this interaction. CCA general workflow passes through the analysis of correlation between covariance matrices of two datasets (in this case FFA and clinical parameters), resulting in the generation of two

canonical variables which in our case the first two were U1 (as the FFA) and V1 (as the clinical parameters). By plotting our data as function of these two variables, as shown in figure 20, it can be seen a demonstration of a high correlation with a canonical correlation coefficient of 0.9015.

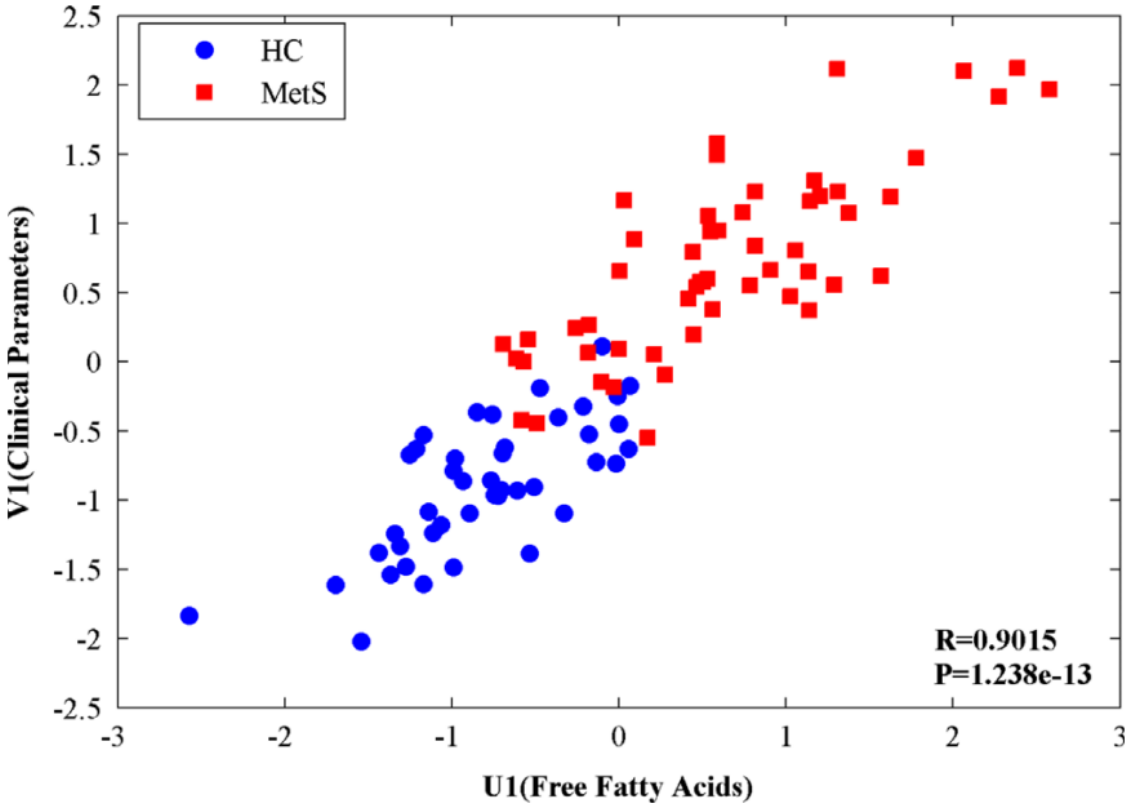

Figure 20: Scores plot of HC and MetS samples as function of canonical variables U1 and V1. P value of right-tail is calculated from Bartlett's approximate chi-square for H_0

These results confirm the initial expectations that FFAs are in fact highly correlated with the clinical parameters. As it can be observed in the plot the P value is very low, which demonstrates the high level of significance of these results when tested by Bartlett's approximate chi-squared statistics for null hypothesis. Nevertheless, CCA can also be used to estimate the variables that have most of the covariance responsible for the correlation. This is done by taking the coefficients of the canonical variables and plotting the absolute values of each study variable. In the case of clinical parameters, it is possible to observe in figure 21 that HDL-C, TG and FBG assume most of the correlation with the FFAs, showing high absolute values for their coefficients. When coupling these results with the ones on table 4, it is possible to observe that FBG and TG not only are the parameters that are more intensively expressed in the overall

observations for the MetS patients group, but are also the ones that are mostly correlated with FFAs. On the other hand, HDL-C shows a much higher correlation with FFAs although it initially does not seem to have such a strong effect in the disease diagnosis.

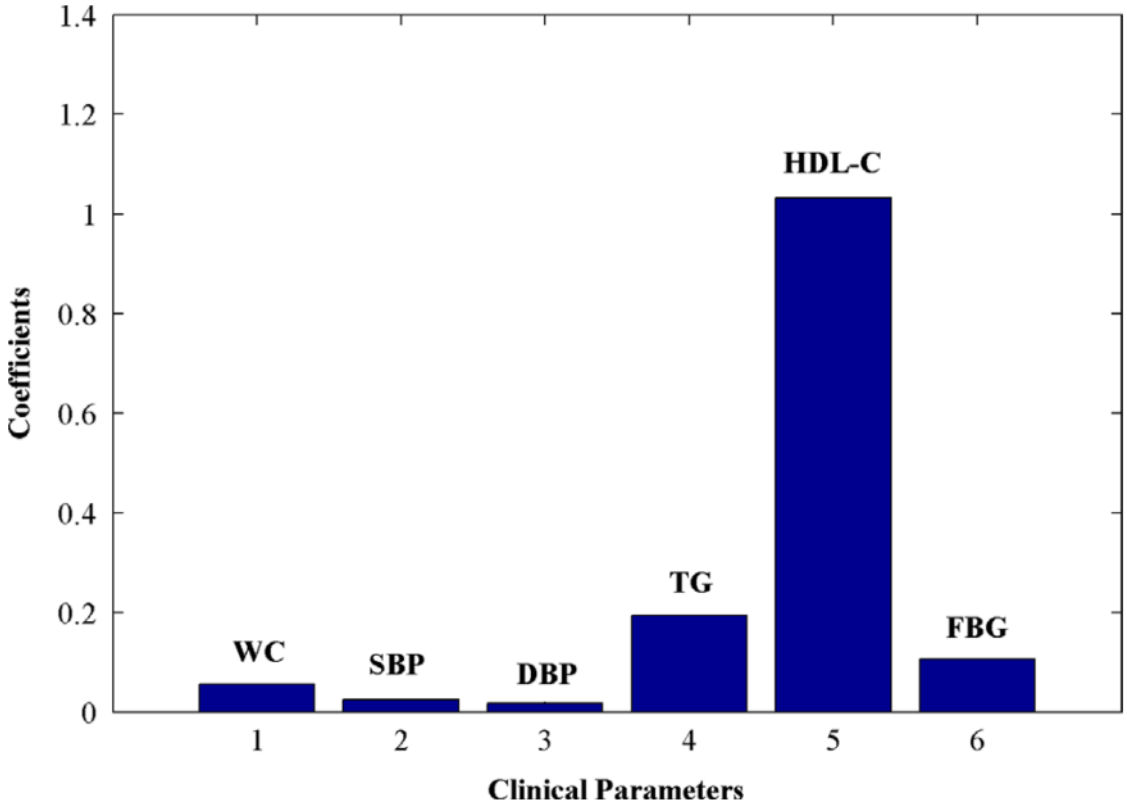

Figure 21: Absolute values of canonical coefficients for each clinical parameter

When plotting the FFAs absolute canonical coefficient values, it can be observed from figure 22 that C16:1n-9c (palmitoleic acid), C20:1n-9c (11-eicosenoic acid), C22:0 (docosanoic acid) and C22:4n-6c (adrenic acid), are the main responsible for the correlation between FFAs and clinical parameters. Recalling the results from table 5 it can be noted that C22:0 has no significant difference between MetS and HC samples, and consequently should be disregarded as potential biomarker. This effect seems to be due to the high canonical correlation of this FFA with the clinical parameters, which is possible, although it is not determinant to be a biomarker, since its value is not statistical significant in between the HC and the MetS group, being this FFA equal important in the absence of clinical risk factors.

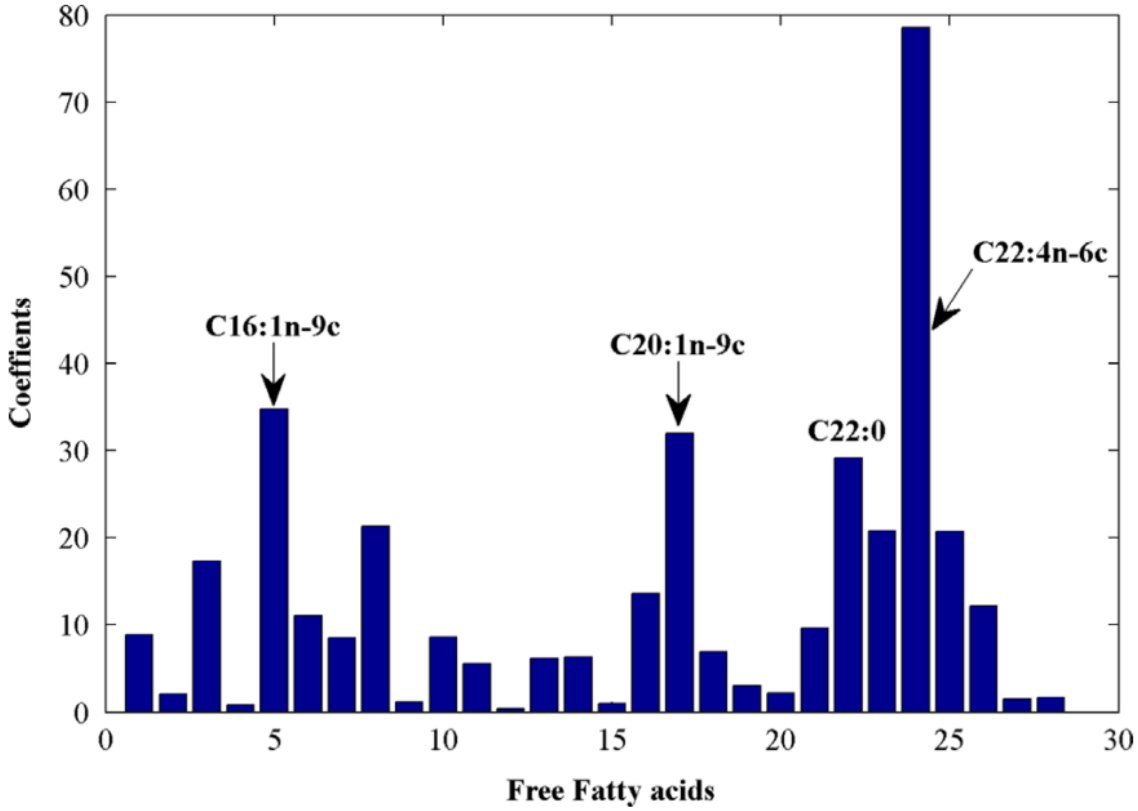

Figure 22: Absolute values of canonical coefficients for FFAs

As potential biomarkers, palmitoleic acid, 11-eicosenoic acid and adrenic acid (AdA) should be considered not only by their presence in the MetS patients serum and their high correlation with the clinical parameters that allow MetS diagnosis, but also by their possible role within the biological pathways that can influence direct or indirectly the disease mechanisms of MetS. Metabolomics studies often face challenging result interpretations due to the complex web of metabolic pathways that allow biological systems to work. Particularly FFAs are involved in lipid metabolism, and the interactions which involve this process. Palmitoleic acid has its origin in the β-oxidation of oleic acid, and it has been reported as being associated with obese prepubertal children incidence of MetS[85]. Its importance has also been recognized due to its significant higher concentration in an *a priori* analysis of individuals that have developed MetS. This retrospective study compared MetS patients with control individuals, considering palmitoleic acid as the most discriminant FFA in the RF classification[86]. On the other hand, 11-eicosenoic acid is involved in the conversion of oleic acid to nervonic acid, the latter being connected to negative effects on the risk factors that lead to obesity-related metabolic disorders[87]. Moreover, previous studies also report the higher concentration of 11-eicosenoic acid in diabetic rats[88], which are in

agreement with our findings that this FFA can be involved in the disease mechanism of MetS, a metabolic disorder that involves risk factors for T2DM. Finally, AdA has its origin in arachidonic acid, which apart from being its precursor can also be its end product, in case AdA undergoes a β-oxidation reaction[89]. This FAA has not been reported to be involved particularly in metabolic disorders, however it is strongly associated with essential physiological processes[90], and consequently may play also a role in MetS mechanisms.

4. Conclusions

In order to explore the relationship of serum FFA profiles of Chinese individuals with MetS, a metabolomics strategy was developed; coupling GC-MS metabolic profiling with chemometrics strategies involving PCA, RF and CCA.

The development of repeatability conditions was essential to achieve reliable results and quality data analysis. Such conditions were achieved through the development of preliminary training with derivatization methods and study of the test samples, allowing optimized temperature programs for GC-MS analysis to be achieved. Moreover, the internal QC was possible by establishing an adequate IS, C19:1n-9c, and the use of QC samples developed from a pool of the test samples which ensured that QC samples will contain the common characteristics of the test samples. A continuous assessment of QC samples allowed tracking the stability of the analysis conditions, and the establishment of RSD limits for FFA profiling completed the QA before applying chemometrics methods. The GC-MS FFA profiling is a fast method, with a total run time of 16.83 min, and results show that using an experimental design for internal quality control and quality assessment of data allow keeping relatively low losses in data (only four out of 28 metabolic features did not pass the QA requirements). Furthermore, metabolic feature identification was possible by using a combination of reference standards and ECL method.

The results of the PCA showed the main trends in the obtained dataset, revealing the reliability of the analysis from the close cluster of the QC samples in a scores plot of 3PCs explaining a total variance of 78.25%. However, the MetS and HC groups did not show a clear separation and clustering. From RF was possible to improve the previous classification results, being able to describe a better discrimination of the MetS and HC group, and also showing a close cluster of the QC samples, achieving a classification accuracy of 0.8632. These results suggest that RF can provide good classification results for MetS serum FFA profiling samples, and can be a potential tool for disease diagnosis.

Correlation between FFA profiles and clinical parameters for MetS diagnosis was estimated by CCA. A high correlation was found, with a canonical correlation coefficient of 0.9015. Results from clinical parameter diagnosis of MetS patients

showed a preliminary dominance of TG in the overall risk factor analysis, showing a high difference between HC and MetS patients. However, CCA demonstrated that not only TG, but also HDL-C and FBG had the major influence in the correlation between clinical parameters and FFA. In fact, HDL-C was identified as the main responsible by this correlation with the highest canonical coefficient. It is clear from the results that mechanisms involving lipid metabolism are responsible for the causal effects of MetS.

In other hand, CCA of FFA shown that C16:1n-9c, C20:1n-9c, and C22:4n-6c were the highest correlated FFA with the clinical parameters, having the later the highest canonical coefficient value. These FFA have been described in literature as relevant in metabolic dyslipidaemia disorders and are in this study identified as potential biomarkers for MetS.

Finally, this approach brings novel information about GC-MS analysis of serum FFA profiling of Chinese MetS patients, and results suggest that it can been very effective for biomarker identification, metabolic dysregulation analysis and for assisting in disease diagnosis and clinical information.

5. Future Work

Further research in metabolic syndrome mechanisms and causal effects is clearly needed. From this study it is suggested that lipids and particularly FFAs are the key points to unveil the main disease pathways.

In this study, the general analysis of clinical parameters was made and conclusions were taken from the main responsibles for the correlation between clinical parameters and FFAs. It would be interesting to confirm if the establishment of a new MetS dataset including only patients with HDL-C, TG and FBG risk factors would still be significantly different from the HC, and if we would still obtain the same FFA for potential biomarkers.

RF is able to provide feedback about variable importance measures, and consequently this can be estimated and compared with the results of CCA analysis, allowing to confirm if the FFAs that are highly correlated with the clinical parameters are the same as the ones that are the main responsibles for the discrimination between the HC and MetS group in RF classification.

6. References

(1) World Health Organization. Number of deaths: WORLD By cause http://apps.who.int/gho/data/node.main.CODWORLD?lang=en (accessed Apr 7, 2014).

(2) Kordalewska, M.; Markuszewski, M. J. *J. Pharm. Biomed. Anal.* **2015**.

(3) Alberti, K. G. M. M.; Zimmet, P.; Shaw, J. *Diabet. Med.* **2006**, *23* (5), 469–480.

(4) Després, J.-P.; Lemieux, I. *Nature* **2006**, *444* (7121), 881–887.

(5) Eckel, R. H.; Alberti, K. G. M. M.; Grundy, S. M.; Zimmet, P. Z. *Lancet* **2005**, *365*, 1415–1428.

(6) Gu, D.; Reynolds, K.; Wu, X.; Chen, J.; Duan, X.; Reynolds, R. F.; Whelton, P. K.; He, J. *Lancet* **2005**, *365* (9468), 1398–1405.

(7) Kylin, E. *Zentralbl Inn Med* **1923**, *44*, 105–127.

(8) Balkau, B.; Valensi, P.; Eschwège, E.; Slama, G. *Diabetes Metab.* **2007**, *33* (6), 405–413.

(9) Wang, M. In *Metabolic Syndrome: Underlying Mechanisms and Drug Therapies*; John Wiley & Sons, Inc.: New Jersey, 2011; pp 179–197.

(10) Xiang, S.; Ji, L.; Xiang, H. *Chinese J. Diabetes* **2004**, No. 12, 156–161.

(11) Lusis, A. J.; Attie, A. D.; Reue, K. *Nat. Rev. Genet.* **2008**, *9* (11), 819–830.

(12) Beilby, J. *Clin. Biochem. Rev.* **2004**, *25*, 195–198.

(13) Grundy, S. M.; Cleeman, J. I.; Daniels, S. R.; Donato, K. a; Eckel, R. H.; Franklin, B. a; Gordon, D. J.; Krauss, R. M.; Savage, P. J.; Smith, S. C.; Spertus, J. a; Costa, F. *Circulation* **2005**, *112* (17), 2735–2752.

(14) Tripathy, D.; Mohanty, P.; Dhindsa, S.; Syed, T.; Ghanim, H.; Aljada, A.; Dandona, P. *Diabetes* **2003**, *52* (12), 2882–2887.

(15) Soares, R.; Costa, C. *Oxidative Stress, Inflammation and Angiogenesis in the Metabolic Syndrome*; Springer, 2009.

(16) Roberts, C. K.; Sindhu, K. K. *Life Sci.* **2009**, *84* (21-22), 705–712.

(17) Croze, M. L.; Soulage, C. O. *Biochimie* **2013**, *95* (10), 1811–1827.

(18) Berg, J. M.; Tymoczko, J. L.; Stryer, L.; Gatto, G. J. *Biochemistry*; W. H. Freeman and Company, 2012.

(19) Ratnayake, W. M. N.; Galli, C. *Ann. Nutr. Metab.* **2009**, *55* (1-3), 8–43.

(20) NCBI PubChem Compound. cis-10-nonadecenoic acid http://pubchem.ncbi.nlm.nih.gov/summary/summary.cgi?sid=24897674&viewopt=PubChem (accessed Apr 8, 2014).

(21) Christie, W. W. In *Advances in Lipid Methodology - Two*; Christie, W. W., Ed.; Oily Press: Dundee, 1993; pp 69–111.

(22) Han, L.-D.; Xia, J.-F.; Liang, Q.-L.; Wang, Y.; Wang, Y.-M.; Hu, P.; Li, P.; Luo, G.-A. *Anal. Chim. Acta* **2011**, *689* (1), 85–91.

(23) Yi, L.; He, J.; Liang, Y.; Yuan, D.; Gao, H.; Zhou, H. *Chem. Phys. Lipids* **2007**, *150* (2), 204–216.

(24) Sánchez-Avila, N.; Mata-Granados, J. M.; Ruiz-Jiménez, J.; Luque de Castro, M. D. *J. Chromatogr. A* **2009**, *1216* (40), 6864–6872.

(25) Yi, L.; He, J.; Liang, Y.; Yuan, D.; Chau, F. *FEBS Lett.* **2006**, *580*, 6837–6845.

(26) ISO (Internacional Organization for Standardization). In *ISO 15189*; Geneva - Switzerland, 2012; p 3.

(27) Carraro, P.; Plebani, M. *Clin. Chem.* **2007**, *53* (7), 1338–1342.

(28) Gillio-Meina, C.; Cepinskas, G.; Cecchini, E. L.; Fraser, D. D. *Pediatrics* **2013**, *131* (4), 754–766.

(29) Liu, L.; Aa, J.; Wang, G.; Yan, B.; Zhang, Y.; Wang, X.; Zhao, C.; Cao, B.; Shi, J.; Li, M.; Zheng, T.; Zheng, Y.; Hao, G.; Zhou, F.; Sun, J.; Wu, Z. *Anal. Biochem.* **2010**, *406* (2), 105–112.

(30) Franco, K. H.; Hoover, J. P.; Backues, K. a.; Payton, M. E. *J. Exot. Pet Med.* **2010**, *19* (2), 169–176.

(31) Yucel, A.; Karakus, R.; Cemalettin, A. *J. Immunoassay Immunochem.* **2007**, *28* (1), 47–60.

(32) Preissner, C. M.; Reilly, W. M.; Cyr, R. C.; O'Kane, D. J.; Singh, R. J.; Grebe, S. K. G. *Clin. Chem.* **2004**, *50* (7), 1245–1247.

(33) Randall, S. a; McKay, M. J.; Baker, M. S.; Molloy, M. P. *Proteomics* **2010**, *10* (10), 2050–2056.

(34) Tuck, M. K.; Chan, D. W.; Chia, D.; Godwin, A. K.; Grizzle, W. E.; Krueger, K. E.; Rom, W.; Sanda, M.; Sorbara, L.; Stass, S.; Brenner, D. E. *J. Proteome Res.* **2009**, *8* (1), 113–117.

(35) Tammen, H. *Methods Mol. Biol.* **2008**, *428*, 35–42.

(36) Schrohl, A.-S.; Würtz, S.; Kohn, E.; Banks, R. E.; Nielsen, H. J.; Sweep, F. C. G. J.; Brünner, N. *Mol. Cell. proteomics* **2008**, *7* (10), 2061–2066.

(37) Lewis, M. R.; Callas, P. W.; Jenny, N. S.; Tracy, R. P. *Thromb. Haemost.* **2001**, *86* (6), 1495–1500.

(38) Mitchell, B. L.; Yasui, Y.; Li, C. I.; Fitzpatrick, A. L.; Lampe, P. D. *Cancer Inform.* **2005**, *1*, 98–104.

(39) IUPAC. *Pure Appl. Chem.* **1993**, *65* (4), 819–872.

(40) Ettre, L. S.; Sakodynskii, K. I. *Chromatographia* **1993**, *35* (3-4), 223–231.

(41) Martin, A. J. P.; Synge, R. L. M. *Biochem. J.* **1941**, *35* (12), 1358–1368.

(42) Nobel Media AB. The Nobel Prize in Chemistry 1952 http://www.nobelprize.org/nobel_prizes/chemistry/laureates/1952/.

(43) Biotechnology Project at MATC. Microbial Life Educational Resouces.

(44) Van Deemter, J. J.; Zuiderweg, F. J.; Klinkenberg, a. *Chem. Eng. Sci.* **1995**, *50*, 3869–3882.

(45) Smith, R. M. *Understanding Mass Spectra: A Basic Approach*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2004.

(46) Dunnivant, F. M. 3D Aspects of GC-MS http://people.whitman.edu/~dunnivfm/C_MS_Ebook/CH5/5_7.html (accessed Sep 7, 2015).

(47) Fiehn, O. *Comp. Funct. Genomics* **2001**, *2* (3), 155–168.

(48) Nicholson, J. K.; Lindon, J. C.; Holmes, E. *Xenobiotica.* **1999**, *29* (11), 1181–1189.

(49) Dunn, W. B.; Broadhurst, D.; Begley, P.; Zelena, E.; Francis-McIntyre, S.; Anderson, N.; Brown, M.; Knowles, J. D.; Halsall, A.; Haselden, J. N.; Nicholls, A. W.; Wilson, I. D.; Kell, D. B.; Goodacre, R. *Nat. Protoc.* **2011**, *6* (7), 1060–1083.

(50) Nagana Gowda, G. A.; Djukovic, D. *Methods Mol. Biol.* **2014**, *1198*, 3–12.

(51) Begley, P.; Francis-McIntyre, S.; Dunn, W. B.; Broadhurst, D. I.; Halsall, A.; Tseng, A.; Knowles, J.; Goodacre, R.; Kell, D. B. *Anal. Chem.* **2009**, *81* (16), 7038–7046.

(52) Food and Drug Administration. *Draft Guidance for Industry Bioanalytical Method Validation*; 2013.

(53) Zhang, L.; Tan, B.; Zeng, M.; Lu, H.; Liang, Y. *Talanta* **2012**, *88*, 311–317.

(54) Massart, L.; Vandeginste, B. **1991**, 519–556.

(55) Liang, Y.; Yi, L. *Fundamentals of Chemometrics*, 1st ed.; East China University of Technology Press, 2010.

(56) Jonsson, P.; Johansson, A. I.; Gullberg, J.; Trygg, J.; Jiye, a; Marklund, S.; Sjstrm, M.; Antti, H.; Moritz, T.; Grung, B.; Sjo, M. *Anal. Chem.* **2005**, *77* (17), 5635–5642.

(57) Lin, Z.; Vicente Gonçalves, C. M.; Dai, L.; Lu, H. M.; Huang, J. H.; Ji, H.; Wang, D. S.; Yi, L. Z.; Liang, Y. Z. *Anal. Chim. Acta* **2014**, *827*, 22–27.

(58) Theodoridis, G.; Gika, H. G.; Wilson, I. D. *Mass Spectrom. Rev.* **2011**, *30*, 884–906.

(59) Filzmoser, P.; Walczak, B. *J. Chromatogr. A* **2014**, *1362*, 194–205.

(60) Jiang, W.; Zhang, Z.-M.; Yun, Y.; Zhan, D.-J.; Zheng, Y.-B.; Liang, Y.-Z.; Yang, Z. Y.; Yu, L. *Chromatographia* **2013**, *76* (17-18), 1067–1078.

(61) Tan, B.; Liang, Y.; Yi, L.; Li, H.; Zhou, Z.; Ji, X.; Deng, J. *Metabolomics* **2009**, *6* (2), 219–228.

(62) Liang, Y. Z.; Kvalheim, O. M.; Keller, H. R.; Massart, D. L.; Kiechle, P.; Erni, F. *Anal. Chem.* **1992**, *64* (8), 946–953.

(63) Kvalheim, O. M.; Liang, Y. *Anal. Chem.* **1992**, *64* (8), 936–946.

(64) Broadhurst, D.; Kell, D. *Metabolomics* **2006**, *2* (4), 171–196.

(65) Holm, S. *Scand. J. Stat.* **1979**, *6* (2), 65–70.

(66) Franceschi, P.; Giordan, M.; Wehrens, R. *TrAC - Trends Anal. Chem.* **2013**, *50*, 11–21.

(67) Bartlett, M. S. *Proc. R. Soc. Lond. A. Math. Phys. Sci.* **1937**, *160* (901), 268–282.

(68) Miller, J. N.; Miller, J. C. *Statistics and Chemometrics for Analytical Chemistry*, Fifth.; Pearson Education Limited: Great Britain, 2005.

(69) Gromski, P. S.; Muhamadali, H.; Ellis, D. I.; Xu, Y.; Correa, E.; Turner, M. L.; Goodacre, R. *Anal. Chim. Acta* **2015**, *879*, 10–23.

(70) Brereton, R. G. *Chemometrics: Data Analysis for the Laboratory and Chemical Plant*; 2003; Vol. 8.

(71) Shepard, R. N. *Psychometrika* **1962**, *27* (2), 125–140.

(72) Buja, A.; Swayne, D. F.; Littman, M. L.; Dean, N.; Hofmann, H.; Chen, L. *J. Comput. Graph. Stat.* **2008**, *17* (2), 444–472.

(73) Dai, L.; Gonçalves, C. M. V.; Lin, Z.; Huang, J.; Lu, H.; Yi, L.; Liang, Y.; Wang, D.; An, D. *Talanta* **2015**, *135*, 108–114.

(74) Breiman, L. *Mach. Learn.* **2001**, *45*, 5–32.

(75) Worachartcheewan, A.; Nantasenamat, C.; Isarankura-Na-Ayudhya, C.; Pidetcha, P.; Prachayasittikul, V. *Diabetes Res. Clin. Pract.* **2010**, *90* (1), e15–e18.

(76) Cutler, D. R.; Edwards, T. C.; Beard, K. H.; Cutler, A.; Hess, K. T.; Gibson, J.; Lawler, J. J. *Ecology* **2007**, *88* (11), 2783–2792.

(77) Hotelling, H. *Break. Stat.* **1992**, No. Chapter 14, 162–190.

(78) Zeng, M.; Liang, Y.; Li, H.; Wang, M.; Wang, B.; Chen, X.; Zhou, N.; Cao, D.; Wu, J. *J. Pharm. Biomed. Anal.* **2010**, *52* (2), 265–272.

(79) Simundic, a. M.; Cornes, M.; Grankvist, K.; Lippi, G.; Nybo, M. *Clin. Chim. Acta* **2014**, *432*, 33–37.

(80) *Chinese J. Cardiol.* **2007**, *35* (5), 390–419.

(81) *Chinese J. Diabetes* **2004**, *12*, 156–161.

(82) Alberti, K. G. M. M.; Zimmet, P.; Shaw, J. *Lancet* **2005**, *366* (9491), 1059–1062.

(83) Yi, L.; He, J.; Liang, Y.; Yuan, D.; Gao, H.; Zhou, H. *Chem. Phys. Lipids* **2007**, *150* (2), 204–216.

(84) Sangster, T.; Major, H.; Plumb, R.; Wilson, A. J.; Wilson, I. D. *Analyst* **2006**, *131* (10), 1075–1078.

(85) Gil-Campos, M.; Del Carmen Ramírez-Tortosa, M.; Larqué, E.; Linde, J.; Aguilera, C. M.; Cañete, R.; Gil, A. *Lipids* **2008**, *43* (8), 723–732.

(86) De Edelenyi, F. S.; Goumidi, L.; Bertrais, S.; Phillips, C.; MacManus, R.; Roche, H.; Planells, R.; Lairon, D. *Genes Nutr.* **2008**, *3* (3-4), 173–176.

(87) Oda, E.; Hatada, K.; Kimura, J.; Aizawa, Y.; Thanikachalam, P. V.; Watanabe, K. *Int. Heart J.* **2005**, *46* (6), 975–985.

(88) Malaisse, W. J.; Portois, L.; Sener, A.; Carpentier, Y. a. *Int. J. Mol. Med.* **2008**, *22* (1), 133–137.

(89) Mann, J.; Kaduce, T. L.; Spector, A.; Figard, H.; Heath, C. **1986**, *244* (2), 813–823.

(90) Guijas, C.; Astudillo, A. M.; Gil-De-Gómez, L.; Rubio, J. M.; Balboa, M. a.; Balsinde, J. *Biochim. Biophys. Acta - Mol. Cell Biol. Lipids* **2012**, *1821* (11), 1386–1393.