POLITECNICO DI TORINO

Master's Degree Course

in Mathematical Engineering

Master Thesis

Signature development for the detection of Pulmonary Hypertension

Supervisor

Prof. Mauro Gasparini

External Supervisor - Actelion Janssen

Matthieu Villeneuve

Candidate

Luca Semeraro

Academic Year 2020-2021

Summary

This thesis follows and assembles what I did during a six-months traineeship at Actelion Janssen. Once the company's therapeutic areas of interest are briefly described, the emphasis in this work is on one of them: Pulmonary Hypertension (PH). Then, a WHO classification and available diagnostic tests for this disease are examined and two clinical studies in PH labelled TRACE and CIPHER are outlined: the first is a multicenter, double-blind, placebo-controlled, Phase-4 study to evaluate the effect of Uptravi® treatment on the daily life physical activity of patients with pulmonary arterial hypertension, while the second one is a prospective, multicenter study designed to identify a biomarker signature for the early detection of PH and another signature to distinguish two sub-classes of this pathology from the others.

The problem presented in the latter belongs to the challenging field of genetic analysis and is here addressed by developing a process for the identification of blood-based biomarker signatures for any disease through the implementation on R of different statistical methods such as Gradient Boosting, Support Vector Machines, GLM with elastic-net regularization and resampling techniques like Cross-Validation and Nested Cross-Validation.

Thus, after summarizing what is available in the literature concerning biomarkers involved in the processes and mechanisms of the disease under investigation, the approach is applied to the dataset extracted from the proof-of-concept study used to design CIPHER, to search for a PH signature with microRNAs as biomarkers. Moreover, the resulting model is evaluated through some metrics such as sensitivity, specificity, and precision, and it is analyzed in detail together with its most influential biomarkers and the results separately by the different groups of the disease classification. Finally, the signature is compared with the standard non-invasive diagnostic method by means of the 95% level Wilson confidence region of sensitivity and specificity, and possible ways to improve its performance are proposed in the conclusion.

Acknowledgements

Coming to the end of this wonderful adventure full of experiences, friendships, and memories, here I would like to dedicate some words to all the people who have made extraordinary this chapter of my life. It's with them that I want to share this goal, although it is not easy to find the right words.

First of all, my heartfelt thanks go to prof. Mauro Gasparini, supervisor of this thesis, for his trust and availability as well as for proposing me the traineeship in Janssen and for giving me precious suggestions and corrections to compose this work.

I have to express warm gratitude to Janssen and in particular to Lilla Di Scala for allowing the realization of my aspiration to carry out the traineeship in a pharmaceutical company and for working so hard to make this possible. I also have to give special thanks to Matthieu Villeneuve for the availability he demonstrated to me, for his professionalism and expertise, and for scrupulous follow-up during the six months in Actelion Janssen despite my remote participation but also in all the moments I disturbed him in the subsequent weeks. With him, I extend this sincere gratitude to Adele Morganti and her great team for welcoming me among them and for sharing experiences. I would like to thank all of them because this experience has certainly enriched my background.

I wish to thank also my parents who trusted me and allowed me to reach this important goal, and to whom I dedicate this thesis. They have been with me through it all, by their financial and emotional support over these years, their constant help, and their encouragement in difficult circumstances.

Together with them, I would like to extend my appreciation to all my relatives for their support, despite the distance preventing us from meeting often.

Special thanks are deserved to my brother Andrea, an essential point of reference in these years: without him I would have probably never decided to move to Turin for studying, and without

his frequent invitations to lunch or dinner I would have come back to Puglia for my vacation periods more and more diminished and thin.

I also would like to address my thanks to all the people I have met in these years with whom I shared traits of this journey full of unforgettable memories and who became my second family: my flatmates who supported me but above all put up with me for four years, my university colleagues for making interesting even the most boring lessons, and all the others I spent fantastic nights with around the city sharing anxieties, joys but also some difficulties. I send my thanks to them for all the memories and the shared moments, conscious that each of them has left a mark on my life and that some of the grown friendships will surely persist regardless of the future that is ahead of us.

Last but not least, I would like to sincerely express my gratitude to all my friends in Cisternino who, although in recent years I have only managed to spend the few weeks of my vacations with them, have never neglected their friendship and support. I appreciate them because, despite being disregarded when I was in Turin, at any latitude they have always been a safe haven, ready to listen to my confidences and help me in the most critical moments (even if I almost always did the opposite of what was advised), as well as available for chats of any kind.

So, I wish to sincerely thank all the people who have been an integral part of these fantastic years and who I met along this route because in their own small way they helped to make this journey extraordinary. At the same time, thanks to myself because if I hadn't chosen Turin as the city for my studies, I have no idea if I would have been able to experience these emotions, but also for the determination and commitment, for believing in it, for never giving up and for succeeding in reaching the goal.

Contents

List of Tables

List of Figures

Abbreviations

6MWD	Six-Minute Walk Distance
6MWT	Six-Minute Walk Test
AC	Activity Count
AUC	Area Under the Curve
BMPR2	Bone Morphogenetic Protein Receptor type-2
BNP	Brain Natriuretic Peptide
CI	Confidence Interval
CTED	Chronic Thromboembolic Disease
CTEPH	Chronic Thromboembolic Pulmonary Hypertension
CV	Cross-Validation
DLPA	Daily Life Physical Activity
ECG	Electrocardiogram
eCRF	Electronic Case Report Form
EMA	European Medicines Agency
EOT	End-of-Treatment
FN	False Negative
FP	False Positive

FPR	False Positive Rate
FDA	Food and Drug Administration
GLM	Generalized Linear Model
GLMM	Generalized Linear Mixed Model
HFpEF	Heart Failure with Preserved Ejection Fraction
HFrEF	Heart Failure with Reduced Ejection Fraction
HPAH	Heritable Pulmonary Arterial Hypertension
IPAH	Idiopathic Pulmonary Arterial Hypertension
k**-NN**	k-Nearest Neighbors
LASSO	Least Absolute Shrinkage and Selection Operator
miRNA	micro Ribonucleic Acid
mPAP	mean Pulmonary Artery Pressure
MSE	Mean Square Error
MVPA	Moderate-to-Vigorous Physical Activity
NT-proBNP	N-terminal pro-hormone Brain Natriuretic Peptide
PAH	Pulmonary Arterial Hypertension
PAP	Pulmonary Artery Pressure
PASMC	Pulmonary Arterial Smooth Muscle Cell
PAWP	Pulmonary Artery Wedge Pressure
PCR	Polymerase Chain Reaction
PH	Pulmonary Hypertension
PPV	Positive Predicted Value
PVR	Pulmonary Vascular Resistance

qPCR	Quantitative Polymerase Chain Reaction
RHC	Right Heart Catheterization
ROC	Receiver Operating Characteristic
RSV	Respiratory Syncytial Virus
SAP	Statistical Analysis Plan
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
TPR	True Positive Rate
TTE	Transthoracic Echocardiogram
VM	Vector Magnitude
VSMC	Vascular Smooth Muscle Cell
WHO	World Health Organization

Chapter 1

Introduction

This work collects and describes what I have done during a traineeship from October 2020 to April 2021 in Janssen-Cilag and through it in Actelion Janssen, a pharmaceutical and biotechnology company based near Basel in Switzerland and purchased in recent years by Janssen, the pharmaceutical companies of Johnson&Johnson. The internship was carried out under the supervision of Prof. Mauro Gasparini, full professor at the Department of Mathematical Sciences of the Politecnico di Torino, and Matthieu Villeneuve, scientific director at the company.

1.1 Janssen, the pharmaceutical companies of J&J

Johnson&Johnson is a multinational company with more than 130000 workers that has been operating for more than a century in three main business areas: Consumer Healthcare, Medical Devices, and Pharmaceuticals. Janssen is therefore the group of companies that includes Actelion

Figure 1.1: Janssen and Actelion Janssen

and focuses on the latter area of the business wherein more than 40000 employees in 150 countries

of all continents around the world work to treat, cure, stop, and prevent some of the most deadly and complex disorders from heart disease, HIV to cancer and Alzheimer's disease. The main purpose is to change the current approach to how diseases are managed and prevented through the contribution of scientific advances, the development of effective and affordable medicines, and collaborations with different experts at every stage of their design. J&J established a global network of innovation centers that are constantly working on new ideas to deliver life-saving and life-enhancing healthcare solutions to patients around the world, through huge investments but also the cooperation with start-ups, universities, scientific organizations, and government institutions. Such efforts contributed to the EU approval of seven pharmaceutical products in 2020, including the Ebola vaccine, and four more were in the process of being approved at the end of the same year [33].

At Janssen, researches and clinical studies concern six therapeutical areas or fields of medicine where the need is high and the opportunity to make a difference is great: pulmonary hypertension, cardiovascular and metabolism, immunology, infectious diseases and vaccines, neuroscience, and oncology and hematology [30]. Moreover, many of the diseases for which treatments or cures are sought belong to rare diseases.

Furthermore, several projects have been and still are implemented in some of these areas to help patients in their daily lives, as well as their families and doctors. Additionally, many websites and pages on the main social networks are available to promote awareness on the prevention of different disorders and to support patients in their treatment.

Let us now describe in detail the therapeutic areas and diseases on which the studies are focused.

1.2 Pulmonary Hypertension

The first therapeutic area consists of Pulmonary Hypertension (PH), a rare progressive disease with multiple causes but no cure whose incidence is increasing in some developing countries such as China and India [18, 50]. PH includes a heterogeneous group of clinical conditions associated with increased pulmonary arterial pressure, which can compromise most cardiovascular and respiratory functions.

This is the therapeutic area they deal with at Actelion, and it became part of Janssen's pipeline only after the purchase of Actelion itself in 2017, as the sixth medical area of interest at Janssen.

For 20 years Actelion has achieved the first oral pulmonary arterial hypertension (PAH) treatment, the first long-term outcomes data, and that on triple combination therapies. Now, capturing Actelion's expertise in PAH, which is the rarest form of PH, Actelion Janssen is a world leader in PAH and has expanded its focus to include other types of PH for changing the disease to a long-term manageable condition, resulting in 16 approved drugs on the WHO Essential Medicines List in 2016. This is achieved partly through collaboration with patient organizations but also through commercial partnerships that provide unique access to new diagnostic technologies and medical devices that aim to extend the benefits of its PH portfolio. Janssen is also leading the *PHocus360* project to raise awareness of this illness through digital channels and social networks, and to provide support for patients and families, especially for PAH.
Researchers seek to overcome the diagnostic gap and improve survival rates through new disease monitoring approaches and new treatments. Many studies are underway with the immediate objective of improving the quality of patients' lives, and understanding the challenges and risks associated with it. Considerable advances have already been achieved, as the median survival from PAH diagnosis has increased from 2.5 years to almost 10 years, but the most ambitious goal is to cure the disease. Furthermore, several drugs for PH have been developed in recent years, and there are currently four approved and marketed products in Europe [34].
Another important purpose is therefore the early identification and diagnosis of the disorder, which can be achieved by investing in diagnostics and biomarkers to find the disease signature. Indeed, it has been demonstrated the feasibility of developing microRNA-based biomarker signatures through machine learning methods to differentiate patients with and without PH. Hence, work is underway through the CIPHER study to exploit these miRNA signatures for the development of non-invasive tests for the early detection or the discrimination of different types of PH.
Thus, this thesis focused on this therapeutic field and precisely on two PH clinical protocols provided by Janssen: first of all, I investigated the possible effect of drug treatment in PAH patients on their daily life physical activity, and then I developed a procedure for finding biomarker PH signatures which, as already mentioned, can help in the early diagnosis of the disease. The latter is explored in detail together with a pathophysiological description of PH and its WHO classifications.

1.3 Other therapeutic areas in Janssen

Cardiovascular and Metabolism

Cardiovascular and metabolic disorders such as diabetes, obesity, and chronic kidney diseases, are rapidly rising, especially in recent years, so that cardiovascular illnesses are the leading cause of death for both men and women, while diabetes is the seventh most common killer worldwide [69]. Metabolic diseases, moreover, are closely associated with cardiovascular ones and, for example, nearly half of people with type-2 diabetes have chronic kidney disease while obesity may lead to increased rates of cardiovascular diseases and other complications.

This encourages further investigation in this field of study for searching compounds that provide real clinical benefits to patients. Several diseases in this therapeutic area are of interest at Janssen since they affect millions of people worldwide that need treatments and, if possible, cures. Meanwhile, many successes have been achieved over the years, such as the development of some treatments for type-2 diabetes and thrombosis.

Relevant cardiovascular diseases include deep vein thrombosis and pulmonary embolism, whose victims are at high risk of other major cardiovascular events, and whose current treatments are effective but often associated with marked bleeding risks. Thus, while searching for new methods to help patients and prevent thrombosis, the desire is to develop new generation anticoagulants both safer and equally effective.

In metabolic disorders, on the other hand, research ranges from chronic kidney disease, whose end-stage treatment is limited to dialysis and kidney transplantation and therefore associated with a short life expectancy, to obesity, which causes a multitude of cardiovascular complications and affects around 13% of the world's population, hence the need for effective weight reduction therapies for obese people at higher risk of the illness, to diabetes mellitus, which concerns 6.3% of the European population and whose prevalence over the next 25 years is expected to improve in countries worldwide, but more markedly in middle-income ones [34].

Lastly, this field includes retinal disorders, for which there are currently few treatments and several therapeutic limitations. By collaborating with other areas of the company and with external organizations and universities, the aspiration is to better study these conditions and develop new and innovative therapies that may change lives in a better way for many people.

Immunology

Immune-mediated diseases affect about one in ten European citizens but, in some cases, these are not quickly diagnosed and therefore people suffer for years before receiving treatment. Moreover, these disorders afflict the patient from their onset for his whole life as his immunity system also reacts abnormally against the organism itself [30]. However, more than two million people worldwide do not respond to currently available treatments, hence the need to recruit the best minds to develop new therapeutic options.

Over the past decades, Janssen developed the first approved monoclonal antibody to act directly on the immunological system, providing patients with exceptional efficacy and symptomatic benefit. After leading for twenty years the discovery, development, and commercialization of monoclonal antibodies that have changed the treatment of several diseases, people are now engaged in the discovery of promising novel oral and biological therapies, as well as new diagnostic solutions.

To accomplish this, it is fundamental to understand how the disease works at the molecular level and thus to discover the targets behind the autoimmune processes. The research area has also expanded in recent years to try to meet the needs of patients from early diagnosis to preventive care and treatment.

The company holds a strong pipeline of treatments for some of the most common immune-mediated inflammatory disorders, still expanding with innovative oral small-molecule or biological therapies. The key fields of interest are dermatology, gastroenterology, and rheumatology with diseases including psoriasis, rheumatoid arthritis, Chron's disease, and ulcerative colitis, and it is intended to use previous achievements to identify new treatments as well as opportunities for early prevention and long-term treatability of some of these illnesses.

Infectious Diseases and Vaccines

Infectious diseases are among the greatest threats to life and public health, and this condition has become more evident in recent months due to the COVID-19 pandemic.

Janssen's portfolio in this area is very strong and is continuously being enriched by working to develop innovative solutions to combat the most complex infectious disorder, find the right interventions, and improve the lives of millions of people through prevention and treatment, with the ambition of being able to prevent the spread of infectious diseases in the foreseeable future. Many clinical trials are underway to investigate the efficacy and safety of several treatments, and

advanced technologies enable researchers to look ahead and pursue next-generation therapies.
First, at Janssen, there is a deep-rooted tradition of treating respiratory infections since the development in 1918 of hypodermic needles and sterile gauze masks to overcome the Spanish influenza pandemic. This mission is still carried out today through the production and advancement of many products for daily respiratory care of diseases such as COVID-19, respiratory syncytial virus (RSV), and influenza: on the one hand, work is to address the serious damage caused by infections with the development of antivirals, while on the other hand, efforts are in progress to collaborate with diagnostic companies developing platforms that offer accurate, rapid, and affordable diagnostics for achieving early diagnosis and thus maximizing the benefit of antivirals.
Despite advancement made, also HIV is still an illness of interest that science has already transformed into a manageable critical condition and, at Janssen, researchers attempt to control the disease and improve the lives of patients for example through the *Moving Fourth* campaign that aims to address HIV-related health challenges such as associated physical and emotional comorbidities, and mental health issues.
Other focus disorders are chronic hepatitis B which afflicts about 300 million people in the world but only 20% of them know to have the disease and few are those who get treatment, bacterial infections that often develop drug resistance leading to a higher death rate, but also tuberculosis which causes 1.6 million deaths worldwide each year [34].
Furthermore, research is also focused on prevention with vaccine development through Janssen's AdVac® adenovirus vaccine technology platform, which was used to develop the Ebola vaccine approved by the European Commission in 2020 as well as to build vaccine candidates against HIV, RSV, and Zika. Over the past year, the platform was also used to develop the candidate vaccine to combat the COVID-19 pandemic, which is the first single-dose vaccine to be approved by FDA and EMA.

Neuroscience

Neuroscience is a critical area of medicine as one in four people experiences a mental health disorder in his or her lifetime, and that condition has a huge impact on the daily existence of patient and caregivers [70].
At Janssen, this field has been of interest since 1958 with the development of a pioneering therapy for schizophrenia while today efforts are ongoing to develop new solutions and innovations

that improve and transform the lives of people suffering from a mental illness through collaboration with about 30 groups and associations, and the employment of new technologies available in imaging, genomics, and biomarkers. Thus, the purpose is to improve the care standards of people with neuropsychiatric disorders by bringing together two of J&J's business areas such as pharmaceuticals, and medical and diagnostics.

Studies focus mainly on Alzheimer's disease, mood disorders, and schizophrenia and are directed to search for diagnostic and prognostic biomarkers of disease as well as the development of new diagnostic technologies using imaging, and drugs and treatments for the same illness. To accomplish these goals, researchers are exploring several properties of the nervous system such as synaptic plasticity and cellular resilience that are linked to learning, memory, and even mood, and whose abnormalities lead to neurodegenerative disorders.

Furthermore, Janssen has launched several campaigns aimed at raising awareness of the challenges of living with neurodegenerative disorders, such as *Breaking Depression* that has been developed for different types of depression with the support of the Global Alliance of Mental Illness Advocacy Networks-Europe (GAMIAN-Europe) and the European Federation of Associations of Families of People with Mental Illness (EUFAMI).

Oncology and Hematology

Lastly, despite the hard work made so far by the scientific community, oncology and hematology remain an area of medicine only partially explored and cancer causes 1 out of 4 deaths in Europe, where it is still the second cause of premature death [13].

For 30 years they have pioneered innovations in oncology with a very strong portfolio of novel therapies, and still today research is focused on the best-known types and subtypes of cancer and treatments for clinical needs currently unmet or with limited treatment options. The aspiration is certainly to make cancer a manageable and curable illness, and Janssen is engaged in research of drugs and treatments achieving the first positive results in 1988.

However, research is not limited to the treatment and cure of the disease, but also cancer interception and early detection through innovative collaborations with leading scientific institutions. Indeed, early diagnosis by detecting suspicious cells before they become malignant and resistant to treatment, and start to proliferate can allow people to be treated at the onset of the disease and thus to a positive outcome.

As a result of close collaboration between translational research and oncology diagnostics teams, there is continuous accommodation of new clinical findings to accelerate the development of

drugs and supporting diagnostics.
Among solid tumors, the attention is on those with a dramatic unmet need such as prostate, bladder, or lung cancers, while among hematological neoplasms, which represent 7% of cancers worldwide, rare blood cancers such as multiple myeloma and B-cell neoplasia are being analyzed, through the employment of new therapeutic approaches including precision medicine and immune-based therapy, like CAR-T [67].

1.4 Aim of the thesis

The initial goal of this thesis is to describe two protocols in PH developed at Janssen and my contributions regarding them. Moreover, a short description of the pathogenesis of PH and some classifications of the disease are provided to examine and discuss one of the protocols in detail. Thus, a multi-center, double-blind, placebo-controlled, Phase-4 study in patients with PAH is considered to study the effect of treatment with a drug already approved by regulatory agencies on patients' daily life physical activity.
In medicine, however, several diseases are not always easy to detect, but at the same time, it is essential to diagnose them early and start treatment as soon as possible. Over the last few decades, moreover, the usefulness of biomarkers found, for example, in the blood has been demonstrated for describing the processes involved in these illnesses and thus for developing diagnostic and prognostic tests.
This work aims to address the need of developing novel accurate and non-invasive diagnostic tests by setting up a procedure for identifying a blood-based biomarker signature of a disorder, through employing machine learning techniques. Once the procedure and the statistical methods included in it have been outlined in detail, they are applied to a real data collection provided by the company to look for a miRNA biomarker signature for the early detection of PH.
RStudio is the software employed for statistical processing in both studies, while database preparation in the first one has been accomplished by SAS via Citrix Receiver [5, 54].

Chapter 2

Pulmonary Hypertension

Pulmonary hypertension is a rare pathophysiological disorder with high blood pressure that affects the arteries in the lungs and the right side of the heart. Formally PH is defined as a mean resting pulmonary artery pressure mPAP $\geq$ 25mmHg, assessed by right heart catheterization (RHC) [18]. However, a retrospective study more recently suggested a definition based on exercises combined with mPAP and pulmonary vascular resistance (PVR) values, and the PH severity is defined by a WHO classification into four functional classes according to physical activity limitations [25].

Typically, the onset of the disease occurs between 20 and 60 years of age and it may involve multiple clinical conditions and aggravate many cardiovascular and respiratory diseases. According to the available literature, the prevalence of PH in the population is very low and, in the UK, for example, it is 97 cases per one million inhabitants with a female:male ratio of 1.8 while 1000 new cases occur each year in the US. On the other hand, it is increasing in incidence rate in developing countries like India and China [18].

Normally the mPAP at rest is 14 ± 3.3 mmHg with maximum values that can reach up to 20 mmHg [26, 39]. From the clinical point of view, the meaning of mPAP values between 21 and 24 mmHg is not very clear and should be investigated because patients with these values of pulmonary artery pressure could be at risk of developing PAH. For this reason, according to the most recent definition of PH, this condition is defined as mPAP $>$ 20 mmHg [59].

Hearts and lungs, in the normal run of things, work together to carry blood throughout the body: the right side of the heart receives de-oxygenated blood from the rest of the body and pumps it into the lungs, which replace the carbon dioxide with oxygen. At this point, the oxygenated

blood passes through the left side of the heart which pumps it throughout the body and thus a new cycle begins [50].
The condition of high blood pressure within the arteries of the lungs leads to damage and narrowing of the blood vessels in the lungs and therefore the heart has difficulty pumping blood into them. Consequently, the most affected organs are the lungs and heart.

2.1 WHO Groups

Combining the main definition of PH with pulmonary artery wedge pressure (PAWP) values measured at rest, it is possible to provide two hemodynamic definitions of the disease: pre-capillary if PAWP $\geq$ 15 mmHg, post-capillary PH otherwise, and in the latter, some sub-classes may be identified by including other clinical cardiac values [26, 65].
However, in 1998 the WHO has defined five clinical groups of PH, called PH WHO Groups, according to the clinical presentation, hemodynamic features, causes and symptoms, and treatment strategies of the disease [51, 58]. This was then updated in 2013 as more knowledge about PH was available, resulting in the following categories which are summarized in Figure 2.1:

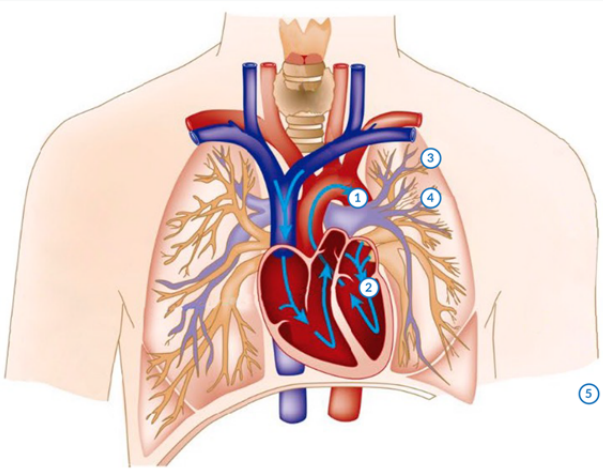


Figure 2.1: Circulatory system and locations of dysfunctions associated with each type of PH according to 2013 WHO classification [31]

- WHO Group 1: PAH

  It occurs when arteries in the lungs become stiffer or narrowed due to dysfunction of the cells in the walls of blood vessels, leading to increasing difficulty for the heart to pump blood through them. It is probably the group about which there is more knowledge on

both causes and possible treatments, and it may show up in different clinical conditions: Idiopathic PAH (IPAH) which is a sporadic form of the disease with no family history and Heritable PAH (HPAH) that is linked to inherited gene even if it is not common, but PAH may occur also in association with other illnesses such as heart and liver disorders, HIV or connective tissue diseases, and lastly the illness may be induced by past or present use of drugs such as methamphetamines and cocaine. Approximately half of PAH patients belong to the last subgroup and all types are treated similarly but the prognosis may be different. Most diagnoses are in women between 30 and 60 years old and, in Europe, PAH incidence is in the range of 5-10 cases per million per year and its prevalence is 15-60 subjects per million inhabitants [49]. For this reason, it is one of the rare disorders, as well as being the rarest type of PH.

- WHO Group 2: PH due to Left Heart Disease

  The disorder is caused by dysfunction in the left side of the heart regarding contraction and relaxation of the heart muscle or related to valves such as the mitral and aortic. Therefore, the left heart is no longer able to do its task properly leading to increased blood pressure in the lungs. It is post-capillary PH as the flow obstructions are at the post-capillary level, in contrast to PAH which is at the pre-capillary level. This is certainly the most common form of PH with growth in prevalence as the severity of dysfunctions and symptoms increases. Indeed, about 60% and 70% of subjects with heart failure with reduced ejection fraction (HFrEF) and heart failure with preserved ejection fraction (HFpEF), respectively, experience PH symptoms while the disorder is detected in almost all patients with symptomatic mitral valve disease and up to 65% of patients with severe symptomatic aortic stenosis [21].

- WHO Group 3: PH due to Lung Disease and/or Hypoxia

  The group includes PH associated with chronic obstructive and restrictive lung disease that leads to narrowing of the lung airways or difficulty expanding during inhalation. Some of the most common conditions comprise emphysema, chronic bronchitis but also sleep apnea that is widespread in overweight middle-aged adults. These disorders cause the blood in the lungs manage to reach only the areas with more oxygen and it implies an increase in blood pressure. In this group, severe PH is not commonly seen but has a high prevalence in individuals with emphysema or fibrosis [27].

- WHO Group 4: Chronic Thromboembolic PH (CTEPH) and other pulmonary artery obstructions

It is a rare form of PH caused by old blood clots in the lungs which block the flow of blood in the vessels and sometimes have no symptoms. In most cases, the use of anticoagulants leads to the recovery of normal blood flow to the lungs preventing the disease, but in some people it may still occur and require surgical removal of the clots. Recent studies, however, have shown that CTEPH may also develop through changes in small blood vessels in the lungs, like PAH. Moreover, some analyses indicate an incidence of about 5 people per million per year [18].

- WHO Group 5: PH due to unclear mechanisms
  In this last group, PH derives from other diseases through a mechanism that is still poorly understood and it may include many causes such as metabolic or systemic disorders but also sarcoidosis or anemias. It can be either pre-capillary or post-capillary PH, and patients identified in this group need a careful diagnosis and subsequent treatment of the disorder and only afterward of PH.

Among the two Janssen owned protocols described in these chapters, the first one (TRACE) considers only the first group, i.e patients with PAH, to examine if there is a treatment effect on their daily life physical activity, while the second study (CIPHER) concerns, in addition to healthy subjects, patients belonging to all PH groups for determining a biomarker signature of the disease, and then takes in analysis only PH patients for establishing a signature for discriminate subjects belonging to the first and fourth WHO groups, hence with PAH and CTEPH, from others.
Despite the different causes leading to the above classifications, all PH groups share common operating alterations, which are summarized in Figure 2.2.
The right ventricle, indeed, begins to face obstacles in pumping venous blood into the lungs and, to be able to carry out its function regularly, it dilates. However, this lasts only for a fairly limited time, because over time it loses effectiveness until cardiac decompensation occurs, meaning that it cannot pump enough blood to the lungs. As a result, the blood accumulates in the ventricle and veins, and it causes leakage of fluid through the vessel walls, edema, and ascites. Finally, it reduces the cardiac output in organs and tissues, and complicates even the most common daily activities, from housework to shopping and walking.
Usually, the therapy involves the combined use of several drugs, but their effect also depends on the time of diagnosis. However, this therapy is viable only for some PH groups such as PAH.

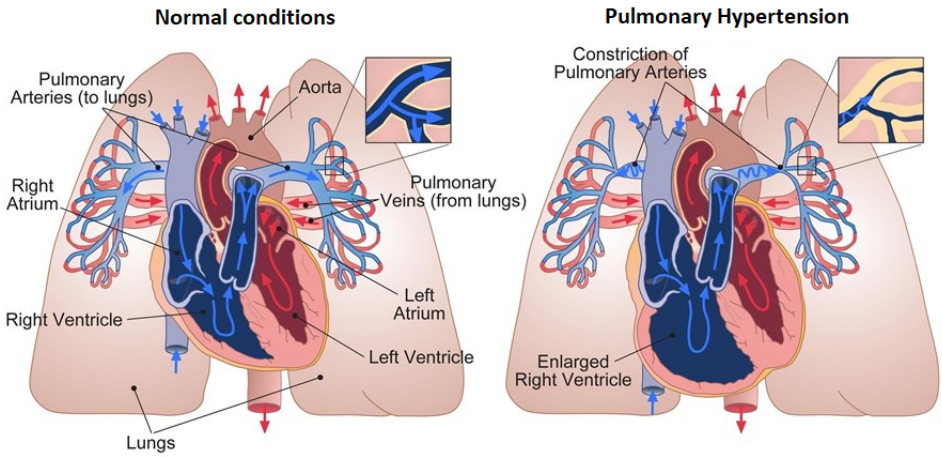


Figure 2.2: Heart and lungs in normal conditions (left) and with PH (right) [7]

2.2 Diagnosing Pulmonary Hypertension

Diagnosis of PH is not easy because its major symptoms such as breathlessness, fatigue, and dizziness are non-specific but are also associated with many other diseases. Less common manifestations, instead, are dry cough, nausea/vomiting after exercise, arrhythmia, and hoarseness. All symptoms are mainly induced by reduced effectiveness of the right ventricle in pumping blood to the lungs, but they also depend on the specific PH group and its pathophysiology.
Diagnosis requires a clinical suspicion grounded on symptoms and an investigation on personal and family medical history, as well as a physical examination and several diagnostic tests. Interpretation of these tests involves a discussion by a multidisciplinary team with a broad range of expertise, and if suspicion of PH is verified, RHC is performed to support the diagnosis. Besides, clinical group identification is a key aspect of the diagnosis and treatment pathway.
The first step in the diagnostic identification of PH is an echocardiographical assessment according to the procedures and recording parameters established by the European guidelines. Several laboratory examinations are then required to confirm the diagnosis and identify the clinical group of the patient suspected of suffering from pulmonary hypertension [18, 52]:

- Electrocardiogram (ECG), checks the electrical impulses of the heart. It is one of the best current practices for non-invasive diagnosis of PH despite it is not enough for a diagnosis since a negative result does not exclude the diagnosis of the illness;
- Echocardiography, is used to estimate PAP and to check the proper heart working through high-frequency sound waves. This test helps in finding the cause of the suspected PH and

it is a necessary but not sufficient examination for supporting the diagnosis. The transthoracic echocardiogram (TTE) is the most used non-invasive screening test which provides multiple parameters to distinguish between different types of PH;

- Chest radiography, may show cicatrices in the lungs but also enlarged ventricle or pulmonary arteries and thus may help in differentiating PH diagnoses. However, as well as for ECG, a negative test result doesn't exclude a positive diagnosis;

- Pulmonary function tests, may help in detecting the cause of the suspected PH by measuring the amount of air held by lungs, and that its move into and out of them, but also their ability to exchange oxygen;

- Blood tests, are required to identify some forms of PH in addition to organ damage. Serologic tests are sometimes performed to investigate the presence of N-terminal prohormone brain natriuretic peptide (NT-proBNP) that may abound in patients with the disorder;

- Exercise tolerance test, in which the patient is subject to a physical exercise like the 6-minute walk test (6MWT) to measures the degree of functional limitation of the subject;

- Ventilation and perfusion lung scan, is useful in PH patients to check for blood clots and blood flow in the lungs and then to look for CTEPH: it is conducted by breathing and injecting via vein into the lungs a small dose of radioactive material to observe their condition.

Despite the plentiful of tests available to confirm a suspected PH, these are not sufficient for the final verification. Indeed, the gold standard for PH diagnosis and confirmation is RHC, that is an invasive procedure, associated with low morbidity (1.1%) and mortality (0.055%) rates, but considered one of the most accurate methods to get a diagnosis of the disorder and the only one that measures the pressure directly inside the pulmonary arteries [29]. For this reason, it is essential for diagnosing PAH and CTEPH but may also help to assess the seriousness of PH.
The RHC procedure starts by inserting a catheter, an extended skinny rubber tube, into an oversized vein within the patient's groin or neck then passing it into the heart to take pressure within the right side of the heart and lungs. In some cases, left heart catheterization may be performed in addition to RHC to avoid misclassification of subjects with an unexpected high PAWP, by introducing the catheter in the heart through an artery.

2.3 Search for non-invasive diagnostics

Many recent studies highlight the necessity to find an accurate non-invasive test alternative to the RHC for diagnosing PH and to develop some treatments that allow patients to lead a normal life and longer survival. Indeed, despite its high accuracy, RHC is a test that may present some risks and adverse events related to the catheterized vessel but also infections or organ damage. Besides, the test should be able to differentiate the types of PH because a critical PH problem is that there is no cure for it, but the treatment, the causes, and symptoms change from patients to patients, and depend on the PH WHO group to which the disease belongs.
Finally, all forms of PH are frequently diagnosed when the illness is already in an advanced state, and that decreases the potential effectiveness of the treatment. For these reasons, it is necessary to build up methods to accelerate the diagnosis and then to develop new treatments to turn PH into a manageable long-term condition.
To accommodate all of these requirements, it would therefore be useful a non- or less-invasive diagnostics which would allow a diagnosis of the disease since the onset of symptoms and a quicker start of treatment, with the objective also to improve the outcomes of the test through the recognition of those subjects who would otherwise never be diagnosed. To do this one might employ new technologies and therefore images and artificial intelligence but also genetics, and a possible solution may be to identify blood or breath-based biomarkers associated with PH.

2.3.1 Blood-based biomarkers

A practical application of the latter goal is through genetic testing or gene expression analysis. These techniques allow finding a biomarker or an ensemble of biomarkers for diagnosing PH and for stratifying patients on the risk of the disorder.
Gene expression analysis is the study of how genes are transcribed to synthesize functional gene products and provides insights into normal cellular processes, such as differentiation and pathological processes. This has rapidly spread over the last decade, due to the impact of new technologies and the development of algorithms that model the cell interactions and many techniques have been developed to profile and quantify the gene expression such as DNA microarrays and quantitative Polymerase Chain Reaction (qPCR).
In recent years, a wide range of clinical studies have been conducted to discover and develop blood-based biomarkers for clinical research, diagnostics, and therapy monitoring, especially for

oncological and cardiovascular diseases, which have led to innovations in the therapies themselves, but also to a faster diagnosis of diseases. Indeed, it is only recently that the importance of microRNAs, which are small non-coding RNA molecules consisting of 18-25 nucleotides, has been discovered with the role they play in regulating gene expression both at the transcriptional and post-transcriptional levels, but also in RNA silencing by pairing their bases with complementary sequences in mRNA molecules. Due to their small size, miRNAs are very engaged in several fields of medicine and, also because of the low concentrations in biofluids, need specific techniques to be detected, amplified, and measured such as PCR which is certainly the most robust technology and that with the highest sensitivity.

Although the knowledge of miRNAs dates back at least a century ago, only in the 90's the first miRNA has been characterized while now there are about 30000 miRNAs of which more than 2000 in humans, and potentially these can target at least 60% of human protein-coding genes [11]. More recent studies indicate approximately 600 miRNAs encoded by the human genome [16].

Furthermore, miRNAs released in body fluids including blood are called circulating miRNAs and it is proved that they are involved in the pathologies of numerous illnesses, such as vascular, neurological but also pulmonary vascular.

Considering many biomarkers simultaneously yields disease signatures and this is confirmed for example in [71] where authors highlight that circulating miRNAs have potential applications as diagnostic biomarkers in heart failures also combined with NT-proBNP.

Thus, for any given disorder, once its diagnostic biomarkers are identified, a disease signature could be found for use as an alternative test to the current standard one. This approach is also being explored at Janssen, for example, to understand the mechanisms of a variety of neurodegenerative and cancer diseases and to develop drugs and treatments, but the use of biomarkers is also being extended to other therapeutic areas.

These considerations suggest that a signature for diagnosing pulmonary hypertension may be identified. Actelion Janssen's CIPHER protocol defines the first study to accomplish this and is explored further in subsequent chapters.

Chapter 3

Two protocols in Pulmonary Hypertension at Janssen

Actelion is mainly concerned with Janssen's sixth therapeutic area, i.e. PH. This chapter considers three datasets associated with clinical protocols or statistical analysis plans (SAPs) in PH from two different studies, TRACE and CIPHER [1, 2].

The first one is a double-blind, multicenter, placebo-controlled, Phase-4 study including patients with PAH to assess the effect of treatment with a drug on daily life physical activity (DLPA) as measured through a triaxial accelerometer worn on the non-dominant wrist. The drug under investigation is Uptravi®, a tablet approved by regulatory authorities for the treatment of adults with PAH to prevent worsening of the disease. After manipulating the provided tables with SAS, a linear mixed-effects model is chosen and used along each directional axis over 14-day intervals to extract random effects of each patient to be used as covariates with several machine learning techniques involving the level of physical inhibition as the response variable. Once the best method is chosen and the model to be considered is constructed with data from the 14 days before the randomization date, it is evaluated on the data from the last two weeks of treatment to analyze whether or not the treatment produces an improvement in the patient's physical condition.

The second study, on the other hand, is a prospective, multicenter, Phase 0 biomarker signature identification study for the early detection of PH, where biomarkers are indicators of physiological and pathological disease processes that might be identified, for example, by analysis of blood samples. This study doesn't involve any treatment and, unlike the previous one, include all PH

groups as well as healthy subjects to analyze the functioning of the signature on those patients as well. Moreover, the progression of the study also plans to identify a signature for differentiating patients with PAH and CTEPH from others as they usually exhibit the early symptoms of the illness years before the diagnosis and thus would benefit more from early detection. In this instance, however, the protocol is only partially respected as two datasets with two distinct families of biomarkers are provided and they are different from the one described in the protocol. Once the available data are preprocessed, a way of identifying a signature with biomarkers for a disease is built through machine learning and resampling techniques. Then, the achieved final model and signature are analyzed and tested on a collection of data independent of that with which the model itself was built.
Now, let us examine the two studies in detail and then take an in-depth analysis of the second one, skipping at this stage the definition of the statistical techniques and evaluation metrics applied.

3.1 TRACE study for PAH

3.1.1 Protocol description

As reported in [2], the study involves 108 patients with PAH which are first screened in a visit based on eCRF (Electronic Case Report Form) to collect data for the clinical study and then observed for 14 days, the so-called baseline, wearing the triaxial ActiGraph GT9X Link accelerometer on the non-dominant arm. This tool measures and collects the instantaneous accelerations along the three spatial directions in units of gravity at a sampling rate of 30 Hz, and the acceleration signal is first filtered by a band-pass filter at 0.25-2.5 Hz and then digitized using an 8-bit conversion, through algorithms licensed by ActiGraph and implemented by the ActiLife software. Finally, the digital signal is summed over specific intervals (epochs) of 60 seconds for each direction returning Activity Counts (ACs) indicating the average counts per minute [46]. These values allow the vector magnitude calculation

$$VM = \sqrt{AC_x^2 + AC_y^2 + AC_z^2}$$

to classify each epoch into different types of activity using Koster's algorithm or even, by appropriate transformations, Freedson's one; in TRACE, as stated by the protocol, the second is preferred [15, 38].

Furthermore, as reported in the clinical study protocol, DLPA is the activity performed while awake with a worn device, and such requirements are identified through the Troiano and Tudor-Locke algorithms, whereas ACs are collected first daily and then in 14-day windows for subsequent analysis [63, 64].
Next, at the baseline visit, patients are randomized in a 1:1 ratio to drug or placebo double-blind administration of 200 mcg oral tablet, where Uptravi® is the drug being analyzed containing the active substance selexipag, the only approved selective IP receptor against targeting the prostacyclin pathway in PAH. Thus, the study drug first is increased at weekly intervals to the maximum tolerable dose for 12 weeks and then the dose of selexipag/placebo is kept constant, except for safety adjustments, until week 24 and the end-of-treatment (EOT) visit.
The objectives of the study are first to evaluate the effect of selexipag on DLPA of PAH patients after 24 weeks, and then on symptoms, exercise capacity, and disease severity in participants. Besides, the safety and tolerability of the drug and its impact on patients' daily lives, along with potential associations between traditional efficacy outcomes and DLPA, are to be assessed. All actigraphy variables are summarized by periods of 14 days.

3.1.2 My own contributions: SAS for database preparation and R for statistical processing

The endpoint considered by me is the evaluation of the treatment effect on DLPA after 24 weeks, once that has been studied by machine learning methods over the 14 days of the baseline. The prognostic effect is the WHO variable FC which indicates the degree of PAH severity and physical activity limitation [25]: it has only two levels labeled as FC1 and FC2 for mild and marked physical activity limitations, respectively. This is accomplished by considering the ACs in all directions at the baseline while awake with a worn accelerometer, and the WHO FC ones.
First of all, SAS software has been employed for database preparation, through Citrix Receiver, an online workspace allowing employees to work and use many programs in remote widely adopted by companies, public administration but also universities [5].
Therefore, several tables have been provided in SAS7BDAT files among which there are:

- *epochsummary*, collecting for each patient the values of ACs measured through the ActiGraph accelerometer in the three axial directions in each one-minute epoch during all weeks of observation, from baseline to EOT, with the corresponding timestamp and thus information regarding the wearing of the device (i.e. whether worn or not) and the patient's

state (i.e. awake or sleep);

- *adsl*, containing data of all PAH patients involved in the study such as the randomization date and the treatment administered;
- *adbs*, including participants' data collected during the baseline visit as the WHO functional class and other clinical information.

From the previous tables, the following variables and records of interest are merged: the epochs with the device worn and patient awake in the first database, the randomization date for each participant in the second one, and finally the WHO FC of all patients in the last one.
Then, using data extracted from the first two tables, baseline ACs in each of the three axial directions are selected and averaged daily for each subject, and the baseline day number is calculated for the resulting records, as shown by the SAS code in Figure 3.1. The same procedure

```
/* Baseline data */
data baselinedata;
merge epoch randomdate;
by SUBJID;
run;
data baselinedata;
set baselinedata;
date = input(substr(Timestamp, 1, 10), yymmdd10.);
format date yymmdd10.;
if (RANDDT > date) AND (RANDDT < date+15);
run;
/* Daily averages and baseline date numbers */
proc sql;
create table bs_dailyavg as
select subjid, date, date-randdt as diffdate, avg(AxisXCounts) as avgx, avg(AxisYCounts) as avgy,
       avg(AxisZCounts) as avgz
from baselinedata
group by subjid, date, diffdate;
run;
data bs_dailyavg;
set bs_dailyavg;
diffdate = diffdate+15;
drop date;
run;
```

Figure 3.1: Some lines of SAS code for baseline data in TRACE [55]

is also repeated for data from the last weeks of treatment. Data regarding the patients' functional class, instead, are collected in a distinct table to be considered later at the end of the preprocessing phase.
Once the database preparation phase was completed, the software RStudio was introduced to proceed with the dataset transformation and the statistical processing. Starting with the baseline data, these are transformed through four different Generalized Linear Mixed Models (GLMMs)

to evaluate which approach to take in the analysis.
Let $u \sim \mathbb{N}(0, \Sigma)$ the vector of the random effects and $Y|u \sim H(\cdot)$, with H a distribution belonging to the exponential family, a GLMM can be defined in matrix form as

$$\eta = g(\mathbb{E}(Y|u)) = X\beta + Zu$$

where the two factors indicate the linear predictor for the fixed effects and the random effects terms, respectively, whence

$$\mathbb{E}(Y_i) = \mathbb{E}[\mathbb{E}(Y_i|u_i)] = \mathbb{E}[g^{-1}(X_i\beta + Z_i u_i)]$$
$$Var(Y_i) = \mathbb{E}[Var(Y_i|u_i)] + Var[g^{-1}(X_i\beta + Z_i u_i)]$$

In the specific scenario of normal distribution $Y_i|u_i \sim \mathbb{N}(\mu_i, \sigma^2 I)$ the model takes the form

$$y = X\beta + Zu + \epsilon = X\beta + \sum_{h=1}^{H} Z^h u^h + \epsilon, \quad \epsilon \sim \mathbb{N}(0, \sigma_\epsilon^2 I)$$

and

$$\mathbb{E}(Y_i) = X_i\beta$$
$$Var(Y_i) = \sigma^2 I + Z_i \Sigma_i Z_i^T$$

where the second term in the variance expression introduces dependence between observations.
In the conducted analysis, four models are considered and applied to the baseline data separately on each direction with the average daily ACs as the response variable y and the number of the day within the time interval as a covariate. For the first case, a simple GLM is chosen (i.e. linear regression) having as the only covariate the number of the day in the given 14 days interval. Subsequently, in the other models, the random intercepts for each patient, the random slopes, and finally also the random second-degree terms are added progressively.
Specifically, for each of these models a new database is obtained, collecting for each patient the slope values on the three axes for the linear regression, and those of the random effects in the remaining cases, but also the associated WHO FC.
For example, the GLMMs with random intercepts

$$y_{ij} = \beta_0 + \sum_{d=1}^{p} x_{ij,d}\beta_d + u_i = (\beta_0 + u_i) + \sum_{d=1}^{p} x_{ij,d}\beta_d$$

are built by using the *lmer* function in the *lme4* package for linear models and, for each axis, the mean value y_{ij} of ACs for the i-th patient on the day j-th of the given 14 days time interval is described by the model

$$y_{ij} = \beta_0 + \beta_1 \cdot diffdate_j + u_i = \beta^*_{0,i} + \beta_1 \cdot diffdate_j$$

where u_i denotes the random intercept effect associated with the i-th patient, and so $\beta^*_{0,i}$ is the intercept of the same patient, while $diffdate_j$ indicates the number of the day in the time interval and takes values from 1 to 14. The random effects u_i along each of the three axes are stored in the new database for each patient together with the associated functional class. Few lines of R code related to the application of the latter model on the baseline data are now shown, similarly to the other chosen methods.

```
set.seed(825)
dfavgdaily <- read.sas7bdat("bs_dailyavg.sas7bdat")
dfavgdaily$SUBJID <- as.factor(dfavgdaily$SUBJID)
# GLMM with random intercept - x-axis
glmm1 <- lmer(avgx ~ diffdate + (1|SUBJID), data = dfavgdaily)
rand1 <- data.frame(ranef(glmm1))[,c(3,4)]
names(rand1) <- c("SUBJID", "effx")
# GLMM with random intercept - y-axis
glmm2 <- lmer(avgy ~ diffdate + (1|SUBJID), data = dfavgdaily)
rand2 <- data.frame(ranef(glmm2))[,c(3,4)]
names(rand2) <- c("SUBJID", "effy")
# GLMM with random intercept - z-axis
glmm3 <- lmer(avgz ~ diffdate + (1|SUBJID), data = dfavgdaily)
rand3 <- data.frame(ranef(glmm3))[,c(3,4)]
names(rand3) <- c("SUBJID", "effz")
# Read file with WHO FC
whodf <- read.sas7bdat("whofc.sas7bdat")
whodf$WHOFC <- as.factor(ifelse(whodf$WHOFC==2,"FC1","FC2"))
whodf$SUBJID <- as.factor(whodf$SUBJID)
names(whodf)<-c("FC", "SUBJID")
# Merge data
dataset <- merge(merge(merge(rand1,rand2),rand3),whodf)[,-c(1)]
```

Now, only one out of the four datasets is chosen by considering several machine learning methods: gradient boosting, Support Vector Machines (SVMs) with linear, radial, and polynomial kernels, random forest, logistic regression, and GLM with elastic-net regularization. These and what follows are defined in more detail in the subsequent chapters.

Proceeding similarly for each dataset, each of these is randomly split into two parts where the first

one contains two out of three observations and is used for the construction of the classification models with slopes or random effects as covariates and the functional class as the response variable. Particularly, models are fitted to the first subset through a 3-fold Cross-Validation to optimize their hyper-parameters, using AUC score maximization as a criterion. The latter is also used to choose the best performing model for each dataset among those constructed, resulting in gradient boosting and GLM with elastic-net regularization for the first and third datasets respectively, and SVM with polynomial kernel for the remaining two. After calculating for each model the optimal cut-off probability on the first set, they are tested on the collection of observations not involved in their construction to predict their level of physical activity limitation and some resulting statistics such as the AUC score but also the sensitivity, specificity, and precision are compared.
From this analysis, the method associated with the dataset containing only random intercepts appears to be the best performing and is therefore selected together with the dataset itself.

	Validation AUC score
Gradient Boosting	0.6620
Linear SVM	0.6674
Radial SVM	0.7291
Polynomial SVM	0.8191
Random Forest	0.6154
Logistic Regression	0.7985
Elastic-net GLM	0.8018

Table 3.1: Validation AUC score for ML methods with dataset transformed by GLMMs with random intercepts

Looking at the validation AUCs shown in Table 3.1, however, the second best method is preferred, i.e. GLM with elastic-net regularization

$$\hat{\eta} = -0.8563 - 0.0056 \cdot eff_x - 0.0065 \cdot eff_y + 0.0085 \cdot eff_z$$

from which

$$\hat{\pi} = \mathbb{P}(\hat{y} = FC1|X) = \frac{\exp(\hat{\eta})}{1 + \exp(\hat{\eta})}$$

with random effects eff_x, eff_y, and eff_z as covariates, because a slight decrease in the value

of the statistic corresponds to a significant increase in model interpretability. The final choice is confirmed by comparing the predicted probabilities and the validation metrics already mentioned. Once the GLMM with random intercepts is chosen to transform patients' ACs over 14 days time periods and thus the classification model to be considered, the former is applied to average daily AC values from weeks 23-24 by deriving random effects as already done for baseline data, to assess whether there is a treatment effect and a change in physical activity limitations from the two weeks prior to randomization to the last two weeks of treatment from day 155th to day 168th. These are indeed chosen because they follow 12 weeks of progressive increase in the administered dose and 10 weeks of maintenance of the individual's maximum tolerated dose, and culminate in the EOT visit as well as the next 30 days safety follow-up phase.
Subsequently, the classification model built with the baseline data is refitted on the data got from the transformation of final weeks ACs, and the treatment effect is evaluated by comparing the estimated probabilities of physical limitation levels and its main statistics with those obtained by fitting it on the previous collection, separately by treatment. So, the aim is to check whether a change occurs in DLPA from baseline to week 24 through a variation in the predicted probabilities of both the WHO FCs.
However, it results that there is no clear benefit from taking selexipag and therefore this therapy seems not to work as there is no significant improvement in the limitations of physical state. Of course, this may also be due to the GLMM employed for data transformation, the length of the time intervals considered, or the classification method itself, but is certainly influenced by the low number of PAH patients involved in the clinical trial.

3.1.3 Progress of the study

The study is over but more in-depth analysis has been carried out considering more variables since the study is exploratory. Indeed, different types of physical activity have been taken into account, and Fredson's and Koster's algorithms and the thresholds established by them are used with VM to distinguish non-sedentary activity and moderate-to-vigorous physical activity (MVPA) [15, 38].
Thus, as described in the protocol, changes in DLPA are assessed by considering accelerometry-based endpoints such as time spent in non-sedentary activities. Moreover, in response to some FDA feedback for other Actelion studies, the volume of non-sedentary activity measured through the number of steps is added in the SAP, together with the daily time spent in MVPA as suggested by Bellerophon Therapeutics due to the positive outcome of a Phase 2b clinical trial with

this endpoint.
The variation for all these variables was analyzed using an analysis of covariance including terms for treatments and geographic region, and baseline values as covariates. Then, the differences in least square means between treatment and placebo, corresponding 2-sided 95% confidence interval and p-value was provided. The same analysis has been repeated also by considering other PAH-related variables such as PAH-SYMPACT® domain scores, i.e. cardiovascular and cardiopulmonary symptom domain scores or physical impact one, 6-minute walk distance (6MWD), Borg dyspnea index during a 6MWT, and NT-proBNP.

3.2 CIPHER study

3.2.1 Protocol description

As outlined in [1], the study involves the development of a PH signature using biomarkers from blood samples (50mL) in the absence of treatment. Being multicenter, patients are chosen from several countries such as Japan, the USA, and the UK to represent different areas of the world, and healthy subjects are also included to observe how the signature works on them as well control group. The analysis of the blood samples was undertaken by MiRXES, an external Singapore-headquartered biotechnology company, and participants are subjected to a transthoracic echocardiogram at the time of enrollment in the study to update the eCRF.
The dataset contains for each patient clinical data (which are neglected in the analysis) but also values of several miRNAs as well as proBNP (prohormone brain natriuretic peptide), a protein biomarker highly associated with PH. Thus, the collection of data is split into two distinct and disjoint sets, a discovery set and a testing one. In both collections, the categorical variable DANA indicates the diagnosis and has 11 levels (9 of diseased patients according to the WHO classification, one for healthy patients, and one of diseased subjects but not PH) which are then grouped into two levels to bring the study to a case-control one. Next, only PH subjects are considered and grouped differently for further analysis.
Moreover, a retrospective study is conducted to identify potential biomarker signatures using blood samples collected from patients with PH in multiple studies during the last 10 years, and then to select the potentially most useful biomarkers for diagnosis.
Thus, the protocol includes a procedure with machine learning and sampling methods to identify circulating miRNA biomarkers associated with PH in the collected blood samples, and to

achieve the identification, and subsequent validation through some evaluation metrics, of a disease biomarker signature to be used as a non-invasive test for PH diagnosis.

3.2.2 My own contributions

Based on the CIPHER protocol, the aim of my analysis is first to develop a procedure for building a disease signature with quantitative biomarkers and then to compare the diagnostic performance of the signature with that of the best available non-invasive diagnosis test. Thus, the procedure is applied to a real dataset to identify a miRNA biomarker signature for PH.

The strategy is created by combining different statistical and resampling methods to choose the best technique for a given collection of data. However, the built procedure is slightly different from the one described in the protocol for identifying a PH signature, as it is generalized for any disease and is also adjusted to the needs driven by the data provided.

Given a discovery set containing miRNA biomarkers as quantitative variables and a binary diagnostic variable, several machine learning techniques are considered and built on it with the diagnosis as the response variable and the others as covariates. Thus, the most suited and best performing method is first chosen via Nested Cross-Validations and its hyper-parameters are optimized through a simple Cross-Validation. Furthermore, the identified signature is analyzed and evaluated on some testing observations, separate from those contained in the discovery set, through metrics that will be described later, and then its performance is compared with that of the standard non-invasive method for diagnosing the disorder through Wilson's confidence regions of two evaluation metrics. Moreover, it is also planned to evaluate the model execution on the different subclasses of the illness and the impact of missing values on the signature.

At this point, it is desired to find a PH signature by applying the built procedure on the provided dataset split into a discovery and a testing sets with 1194 and 376 observations, respectively. Both of these include clinical data of healthy and PH subjects from several countries around the world and meet the conditions described as the presence of several quantitative variables for miRNA biomarkers and a binary variable of diagnosis to be used in the models as covariates and response variable, respectively. However, this data collection is different from the one described in the CIPHER clinical protocol because it has been extracted from the proof-of-concept study used to design the protocol itself. Moreover, a wide range of machine learning techniques is chosen, from the simple k-Nearest Neighbors (k-NN) to tree-based methods, and also more recent approaches such as neural networks.

In the following chapters, the procedure for identifying a biomarker signature for any disease is

described in detail with the statistical methods adopted and then the results of its application to the data collection provided by the company are shown to obtain a miRNA biomarker signature for early detection of PH.
Finally, the company also provided a dataset similar to the previous one but with metabolomic data. These are also pilot data not yet included in the CIPHER ones but structured similarly, with two sets of discovery and testing including 1221 and 359 patients respectively, for which data on 1522 metabolic biomarkers have been collected. Then, the developed procedure is also used to find the PH signature with metabolic biomarkers, similarly to that done with the other biomarkers. This, however, is not analyzed here.

3.2.3 Progress of the study

At the company, according to what defined in the protocol and its dataset, the plan is similar to what done with the first collection of data, which is to explore several machine learning techniques to identify a suitable PH signature with miRNA biomarkers and then to compare it to the standard non-invasive method of diagnosis. Indeed, the procedure defined in the CIPHER protocol is not very different from the one defined in this work. Furthermore, a second signature is intended to develop to differentiate PAH and CTEPH from other PH groups, using the ROC curve and AUC score.
At the moment, the CIPHER study is ongoing and no analysis has been performed so far: the end of the first phase, that of discovery, is approaching and it is expected to run models by the end of this year.

Chapter 4

miRNA biomarkers: studies and researches

It has been proved in several papers that circulating miRNAs are involved in the pathogenesis of several disorders. The purpose of this chapter is to briefly recap what is available in the literature regarding miRNAs concerned with the different processes of PH and then go on to describe a procedure for identifying a miRNA biomarker signature for any disease and its application to a real collection of data for PH.

First of all, in literature microRNA sequences are defined by concatenating *miR*, the number of the family it belongs to with eventually a letter for differentiation among multiple members of the same family, and then the tag of the double helix RNA it comes from, i.e. *3p* or *5p* if it comes from the 3' or the 5' arm respectively of the precursor. This sequence is then preceded by three letters indicating the species to which it belongs, and *hsa* is used for humans. An example is given by hsa-miR-148a-3p but there are some nomenclature exceptions for miRNAs discovered before the standard just described came into force such as hsa-let-7f-5p [22].

Due to their small size, miRNAs have many applications in medicine and one of the most important employments of genetic analysis is to isolate individual miRNA, amplify its signal, and identify biomarkers to assess disease risk but also to perform diagnosis or prognosis.

4.1 Fold change

Over the years, several statistical and computational methods have been used to investigate the differentiation of gene expression.

The simplest approach for selecting genes whose expression patterns differ by phenotype is to use the fold change criterion [40]. This, however, is not enough to assess the significance of gene expression differences and so it is combined with some statistical tests such as t-test and F-test. Due to the repetition of the selected test for each gene, that could lead to an increase in false positives and so the procedure of multiple tests provides an overall assessment of the test significance by controlling the first type error, i.e. false positives, but also the false discovery rate introduced by Benjamini and Hochberg and defined as the expected percentage of false positives among the claimed positives [4]. Another possible approach is the cluster analyses, still widely used although it cannot provide accurate disease predictions due to its characteristics [20], but in recent years many alternatives have been offered and applied for tumor classification such as the Support Vector Machine (SVM) and the naive Bayes method [17, 47].

These methods can thus be applied to separate biomarkers into up-regulated and down-regulated in response to an external stimulus, such as a disease or the processes involved in it: a variable or biomarker results up-regulated if it is more expressed in diseased subjects than in controls, while it is down-regulated in the opposite case.

The computation of the fold changes is the basic method to apply in the analysis of gene expression data for measuring variation in the expression level of a gene due to the possibility of combining it with the results of statistical tests through graphs. For a biomarker in a case-control study, the fold change is defined by the ratio of its mean value in cases and that in controls, and logarithmic transformation is often applied to its absolute value. In clinical studies, however, the base-2 logarithmic transformation is preferred because it better describes changes and variations between the two classes, and the log2 fold change for the i-th biomarker is defined as

$$log2\ fold\ change_i = \log_2 \left| \frac{x_{mean,case}}{x_{mean,control}} \right|$$

Thus, for a positive value it is up-regulated, down-regulated otherwise. Furthermore, a null value represents no change between the mean values of the two classes, +1 represents a doubling of the mean value in cases compared with that in controls while -1 represents a halving of the mean values in cases compared with controls.

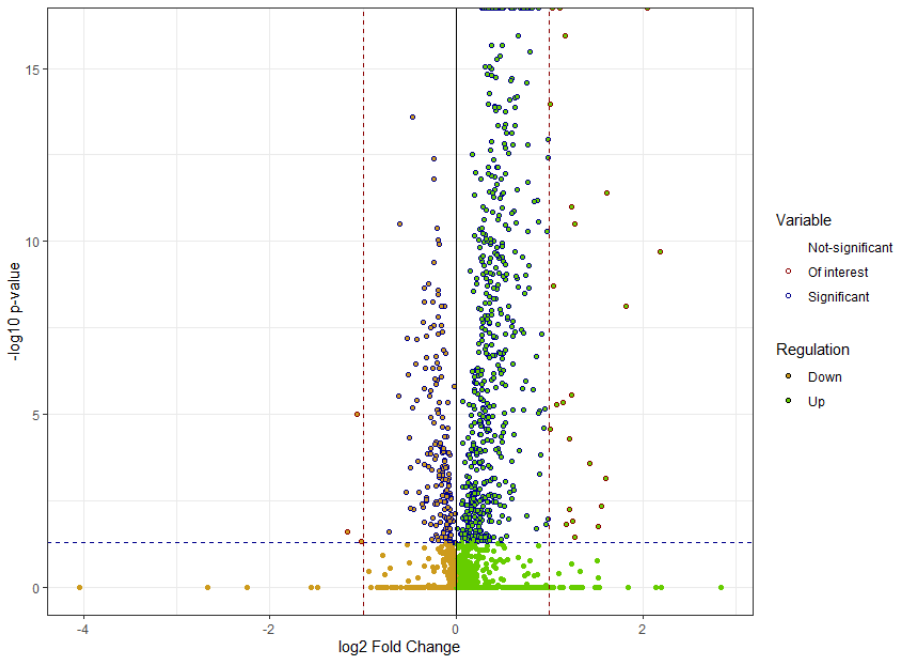


Figure 4.1: Volcano plot with the second dataset described in the CIPHER study: test significance and variable regularization

As already said, these values may also be shown in combination with the p-values of a hypothesis test with α significance level for identifying among the significant variables, i.e. for which $p - value \leq \alpha$, those that are also of interest, i.e. satisfying $|log2\ fold\ change| > 1$ or in other words those having an average value in cases at least double that of controls or vice versa. Results could be visualized for example by a volcano plot, as shown in Figure 4.1 where variables above the blue line are significant for the test, and those not in the area between the two red lines are of interest. At the same time, values on the x-axis allow the distinction between up-regulated and down-regulated biomarkers, depicted in green and orange, respectively.

4.2 miRNAs and pulmonary hypertension

The pathogenesis of PH involves several molecular mechanisms that cause thrombosis, vasoconstriction, vascular remodeling, and heart failure. Increased or decreased levels of miRNAs have been shown in several studies to be closely associated with the clinical phenotypes of PH and PAH and, at the same time, the role as markers of some genetic mutations causing PAH and PH has been demonstrated such as that of bone morphogenetic protein receptor type-2 (BMPR2)

found in 75% and 25% of IPAH and HPAH cases, respectively [6, 48]. These findings suggest that several circulating miRNAs are involved in regulating the expression of genes of pathological and physiological processes of PH and PAH.

Several papers have been written in recent years identifying possible miRNA biomarkers of PH or one of its types, and some of them associated with processes involved in the illness, through statistical tests such as the t-test and the F-test.

Miao et al [45] highlight the role of some circulating miRNAs in PH pathogenesis and focus primarily on their role in PAH, by suggesting some biomarkers as possible diagnostic and therapeutic targets of the disorder. Abnormal levels of some miRNAs are, for example, involved in the anomalous proliferation of pulmonary arterial smooth muscle cells (PASMCs). miR-34a inhibition leads to an increase in cell proliferation and a consequent rise in the expression of platelet-derived growth factor receptor α leading to PAH, but also the miR-17∼92 cluster because removal of its miRNAs from PASMCs of PAH patients drives to a relaxation of the disease. miR-145, instead, is important for the differentiation of vascular smooth muscle cells (VSMCs) but also plays a key role in the PAH pathogenesis as a rise in its expression causes over proliferation and invasion of VSMCs, whereas miR-143-3p is involved in pulmonary artery remodeling and its inhibition suppresses PH.

However, one of the most important mechanisms in PAH is right ventricular failure, and miR-126 expression is significantly lower in patients with this condition and significantly reduces capillary density, whereas miR-140 expression is up-regulated in hypertrophic right ventricles.

Fares et al [11], on the other hand, summarize several profiling studies regarding the differentiation of some miRNAs involved in PH pathogenesis and focus on miRNAs regularization in some PAH manifestations. The set of up-regulated miRNAs includes miR-143 and miR-145, which are abundantly expressed in cardiac and vascular smooth muscle cells and increase in humans with PH, miR-21 that is over-expressed in the lungs of PH patients, and miR-27a that is induced by hypoxia in pulmonary artery endothelial cells. Those down-regulated, instead, comprise miRNAs in the miR-17∼92 cluster that are essential for lungs development as well as miR-204, miR-424, and miR-503 that are positively regulated by apelin that results essential for pulmonary vascular homeostasis, and miR-126 that decreases in PAH skeletal muscles.

The three mechanisms that most regulate miRNAs in PAH are hypoxia, inflammation, and tyrosine kinase pathways. The first one activates some transcription factors such as HIF-α, which induces the expression of miR-210 in endothelial cells and pancreatic cancer cells, and miR-155. Inflammation, instead, releases numerous growth factors and cytokines by increasing miR-155

and miR-146, whereas other miRNAs like miR-204 and those in the miR-17∼92 cluster are regulated by tyrosine kinase and STAT3.

Furthermore, it has been proved that several circulating protein biomarkers are highly associated with PAH and PH and these are grouped into markers of vascular dysfunction, inflammation, tissue hypoxia, secondary organ damage, and myocardial stress. To this last class belong BNP and proBNP, and McGlinchey et al [44] emphasize their usefulness as equal to some circulating miRNAs and in combining them in multi-biomarker panels to increase their diagnostic power. Both peptides are not specific for PH but more generally for heart disease: the former appears to have a closer correlation with pulmonary hemodynamics and is less influenced by renal function, whereas the latter appears to be a powerful prognostic predictor. These also highlight the possibility of combining several biomarkers to get a possible disease signature.

Most of the statistical methods used so far, however, can identify highly differentiated genes but lead to a not so high biomarker accuracy. This is mainly due to all of these techniques not using classification accuracy to measure the miRNAs discriminating ability and therefore genes are not sorted by their accuracy in experiments. Indeed, many authors combine the genes that come out best from their analyses to obtain a better classifier [72]. To overcome all these limitations, methods commonly referred to as machine learning techniques can be employed.

4.3 miRNA biomarkers and machine learning techniques

The term *machine learning* has been coined around the beginning of the second half of the last century to denote the ability of machines or computers to learn from experience or, in other words, to be able to perform tasks only by processing input data with some algorithms. It involves several methods elaborated since the beginning of the last century such as linear regression, logistic regression, and the generalized linear model. As technology has evolved, non-linear models like tree-based methods have been developed, starting from classification and regression trees and proceeding to random forests and boosting. Finally, the rise of more computationally powerful software has led to more recent and complex techniques such as neural networks.

Nowadays, these methods and other more complex ones are employed by major multinationals such as Amazon, Google, Facebook, and Netflix to perform market research and segmentation (e.g. to understand what their audience likes and dislikes, and target the right content to them), to explore customer behavior and personalize recommendations (e.g. to get to know customers in-depth and make recommendations to them via advertisements) but also to predict trends for

businesses by using big data.

All methods used to understand data and learn from experience are grouped into supervised and unsupervised according to their output: the former obtain predictions as output (e.g. classification or regression analysis) while those in the second group analyze relationships and structures of input data (e.g. cluster analysis).

This work is an attempt to differentiate miRNA biomarkers and identify a biomarker signature for diagnosing the disease, employing supervised models. Moreover, since the response variable is the diagnosis, this is a classification problem. For achieving the established goal, it is of primary importance to find a classification method that both allows defining an accurate signature and guarantees good interpretability of the results. Thus, in the following chapters, eight supervised methods for binary classification are considered, and their main features are now described.

The first method is among the simplest machine learning algorithms, the k-Nearest Neighbors (k-NN). It is a non-parametric method that consists only in the discovery or training set and, once a value of k has been chosen, for a given test observation the most frequent class among the k nearest neighbors in the discovery set is assigned to. It is very fast even with large datasets and has a computational advantage, but it is susceptible to outliers.

Then, two techniques are chosen into the family of tree-based methods, which involve stratifying the predictor space into many regions to predict the diagnosis label for one or more observations and all the splitting rules may be displayed in trees. Decision trees are simple and easily interpreted but sometimes perform poorly than other supervised approaches. Thus, multiple trees can be combined in ensemble methods to obtain a single prediction improving performance significantly. Among these techniques, random forest and gradient boosting are chosen in this work.

Gradient boosting improves the decision tree predictions because it is built sequentially: starting with a simple model, at each step, a shrunken version of the new learner that minimizes a loss function is added and it is trained with data incorrectly classified in the previous step. Then, each learner uses information from the previous step with the functional gradient descent algorithm to improve the tree where it works worst. So, the final prediction function is obtained by combining all the built models.

Random forest, instead, involves several trees built on bootstrapped discovery samples at the same time. Furthermore, while building each tree, at each stage only a subset of predictors is considered to choose the cutting one, leading to different and decorrelated trees and preventing overfitting. For a classification problem, the output for each test observation corresponds to the

mode of the classes predicted by all trees. The model becomes less interpretable as the number of trees increases.

Another family of binary classification models is that of the Support Vector Machines (SVMs), which offers also several approaches to generalize them to the multilevel case. These methods infer the Support Vector Classifier (SVC) by accommodating non-linear boundaries through kernels, which quantify the similarity of pairs of observations. In the linear case, it finds a hyperplane that better splits the classes by allowing some misclassifications to avoid overfitting; here, polynomial and radial kernels are employed in addition to the linear one.

Another method is the penalized Generalized Linear Model (GLM) with elastic-net regularization which improves the logistic regression by adding a penalty term in the expression that is maximized to estimate coefficients. This approach represents a compromise between lasso and ridge regression, by overcoming their limitations and allowing variable selection by shrinkage. It is described and defined in more detail in Chapter 5.

To use the widest possible range of methods, it has been also decided to employ a technique developed in the 80s even though it requires a large amount of data to work well and so it is not expected to work well in the real case analyzed in this work. This last method is the neural network, a nonlinear parametric approach inspired by the neural networks of animal brains. It consists of many interconnected nodes distributed over several layers and the output for each test observation is obtained by combining linear and nonlinear functions. This method is difficult to interpret and therefore the simplest one, i.e. that with a single hidden layer, is chosen for demonstration purposes.

All these techniques are employed to identify a biomarker signature for a disease, achieving different performance and levels of interpretability. At the same time, they may be combined with resampling methods to improve their performance, which are defined as procedures that upgrade the data learning process and that have become an essential part of modern statistics. These methods require a set of observations and some hyper-parameters with which to build a model, and they involve repeated sampling from the dataset and fitting a model of interest to each sample to extract additional information about the fitted model. Some of these procedures will be described later.

Thus, by combining sampling techniques with the previously defined classification models, it is possible to choose the best method and to analyze it in detail, taking into account that in a clinical context it is preferable to understand the models and their results.

Proceeding in this way, a procedure is developed and described in Chapter 5 for building the signature of a general disease, and then applied to a real collection of data for PH in the succeeding chapter.

Chapter 5

Development of methods for biomarker signatures

In this chapter, a general procedure to determine a disorder blood-based biomarker signature from a dataset that meets some requirements is described.

Let's consider a collection of data containing clinical data from blood samples of patients enrolled in the study: a categorical variable indicating the diagnosis to be transformed into a binary one, with 1 meaning a diseased subject and 0 a healthy one, and several quantitative variables collecting biomarker values in blood samples. Other medical information from eCRF, however, is overlooked in the study. The database, if not yet, is split into two disjoint sets, a discovery set and a testing one, where the first is used to choose the best technique for developing the signature that is then evaluated on the second one.

After a preliminary step of data analysis, variable selection, and preprocessing, a retrospective case-control problem with biomarkers as covariates and the binary diagnosis as response variable is then outlined. The endpoint is the development of a procedure to build a signature of a disease with blood-based biomarkers and to evaluate its performance, also considering any sub-levels of the categorical variable.

This objective is accomplished through the implementation of classification methods mentioned and briefly described in the last section of the previous chapter such as GLM with elastic-net regularization and gradient boosting, together with some resampling techniques including k-fold Cross-Validation and Nested Cross-Validations. Specifically, the functionalities and properties of methods and evaluation metrics employed in the strategy and into the next case study are

analyzed in detail, including insights about their implementation on Rstudio. Moreover, the implemented code lines are shown in the appendix at the end of this thesis.

5.1 Univariate analysis and variable selection

Before going ahead with the procedure development, it is recommended to analyze the provided dataset and select variables by checking their power to discriminate healthy from diseased subjects, standardize them to be able to make comparisons and use some algorithms, and remove those that result to be highly correlated.

First, a retrospective study on banked blood samples of diseased patients from multiple sites around the world is carried out: this is an essential step because it allows to retrospectively select the variables that may be useful in the disease signature and thus to overlook those certainly having no discriminatory power between healthy and diseased subjects.

Furthermore, missing values in the discovery set are analyzed: if these are not so many, one could remove the not complete observations paying attention to the preservation of the proportions of the two classes, otherwise, appropriate imputation techniques would have to be used. This last step is not meticulously explored here as it is not a thesis topic.

A logarithmic transformation is applied to variables with a scale markedly different from the others according to boxplots but also the available literature, to stabilize the variance and to reduce multiplicative effects to additive ones. At the same time, this transformation limits the influence of outliers on the distributions.

Now, some statistical tests and a pre-processing step follow to better explore the provided data collection.

5.1.1 Normality test

To further investigate the distributions of the quantitative variables in the discovery set, one can start by checking their normality with the Shapiro-Wilk test whose results may be confirmed through some statistics such as skewness and kurtosis.

The normality check with the Shapiro-Wilk test is performed by comparing two alternative variance estimators: a non-parametric one defined as the optimal linear combination of the order statistics of a gaussian random variable, and the usual parametric estimator, that is the sample variance. The null hypothesis H_0 of the test is that the underlying distribution is normal.

Before testing a sample X containing n observations, its values are sorted in ascending order so

that $X_1 \leq X_2 \leq \cdots \leq X_n$. Furthermore, once $m = \lfloor \frac{n}{2} \rfloor$ is set, the test statistic is defined as

$$W = \frac{\left(\sum_{i=1}^{m} k_i (X_{n+1-i} - X_i)\right)^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

where k_i values are provided by specific tables to derive the linear combination of the order statistics of a variable with normal distribution, while the denominator denotes a multiple of the sample variance. The W statistic takes values between 0 and 1, and the closer it is to 1, the closer the distribution of X is to a normal distribution. Conversely, with small W values, the test rejects the null hypothesis.

One way to confirm the test results is the quantile-quantile plot that displays and compares the truth quantiles of a given distribution and those expected from a gaussian one: if the depicted points are arranged along a straight line, then the distribution closely approximates the normal one. However, with hundreds of features it is difficult and computationally expensive to explore all variables but you may choose those that result normal from the previous test or a subset of them.

Another excellent alternative is given by calculating two statistics for the distribution of quantitative variables such as skewness and kurtosis defined as

$$skew(X) = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mathbb{E}[(X-\mu)^3]}{(\mathbb{E}[(X-\mu)^2])^{3/2}} = \frac{\sum_{i=1}^{n} \frac{1}{n}(x_i - \bar{x})^3}{\sigma^3}$$

$$kurt(X) = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{\mathbb{E}[(X-\mu)^4]}{(\mathbb{E}[(X-\mu)^2])^2} = \frac{\sum_{i=1}^{n} \frac{1}{n}(x_i - \bar{x})^4}{\sigma^4}$$

On one hand, skewness is a measure of the asymmetry of the distribution, and 0 or values close to the latter indicate symmetric distributions and normally distributed data. For positive skewness values, instead, the distribution is positively skewed with most values below the mean and therefore concentrated on the left side of the distribution, while for negative values it is negatively skewed with values mostly above the average value.

On the other hand, kurtosis measures the distance of a distribution from the normal one with respect to which there is a greater flattening or greater elongation, and thus the peak shape. For that measure, the value for an approximately normal distribution should be close to 3. Instead, the data distribution is leptokurtic and shows a marked peak if the coefficient of kurtosis is greater than 3 while for smaller values the distribution is platykurtic and is more flattened on the axis. However, many functions often subtract 3 from the computed coefficient, resulting in

an approximately zero value for normal distributions or approximately such.
The Shapiro-Wilk test is performed in R through the *shapiro.test* function which returns the W-statistic and the associated p-value, while the quantile-quantile plot is drawn, for example, by implementing the *qqnorm* function. The package *e1071*, instead, contains *skewness* and *kurtosis* for computing these two statistics for a given sample.

5.1.2 Rank-based tests

An analysis of quantitative variables through tests on mean values may be helpful. A standard example is a one-sample t-test that is any statistical hypothesis test in which the test statistic follows a Student's t-distribution under the null hypothesis. It compares the mean $\bar{x}$ of the sample data to a theoretical value μ, and the t-statistic is calculated as

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

where n is the sample size and s is the sample standard deviation. Thus, the critical value of the Student's t-distribution with $n-1$ degrees of freedom related to the chosen test α significant level is found in proper tables also available in statistics manuals, with associated p-value denoting whether the difference is statistically significant or not. However, the t-test should not be used if the normality null hypothesis of the Shapiro-Wilk test is rejected for many biomarkers. So, a non-parametric test not influenced by possible outliers could be opted for.
Consider as an alternative the rank-based regression which has the same goal of the linear regression, that is to estimate the vector β of coefficients for which

$$y_i = \beta_0 + x_i^T \beta + \epsilon_i, \quad i = 1,2,\ldots,n$$

where y_i is the response variable, β_0 the model intercept, ϵ_i the error term, while x_i is the vector of explanatory variables [37]. However, this formula is resumed and explained better later to define the GLMs.
Unlike linear regression, the rank-based one uses the information of the observations ranking leading to estimates that are less susceptible to outliers. The rank-based estimator is obtained by employing as distance metric the convex Jaeckel's dispersion function

$$D(\beta) = ||y - X\beta||_\varphi$$

wherein $||\cdot||_\varphi$ is a norm defined as

$$||u||_\varphi = \sum_{i=1}^{n} a\big(R(u_i)\big)u_i$$

and R indicates the rank, $a(t) = \varphi(\frac{t}{n+1})$ while φ is a non-decreasing, square-integrable and standardized score function on the interval [0,1]. The Wilcoxon scores are the functions for rank-based fitting defined as

$$\varphi(u) = \sqrt{12}(u - 0.5),\ u \in [0,1]$$

that work better with most datasets generalizing the process to distributions that deviate from the Gaussian shape. Thus, the β estimate is obtained by

$$\hat{\beta}_\varphi = \arg\min||y - X\beta||_\varphi$$

and it is consistent and asymptotically normal, and defined as

$$\hat{\beta}_\varphi \sim \mathbb{N}\big(\beta, \tau_\varphi^2(X^T X)^{-1}\big)$$

where τ_φ depends on the errors' probability and the score functions, while an approximate $(1-\alpha)\times 100\%$ confidence interval for β_j is

$$\hat{\beta}_j \pm t_{1-\alpha/2,n-p-1}\hat{\tau}_\varphi(X^T X)^{-1}_{jj}$$

A Wald test associated to the model is

$$\begin{cases} H_0: & M\beta = 0 \\ H_1: & M\beta \neq 0 \end{cases}$$

and the null hypothesis H_0 is rejected if

$$\frac{(M\hat{\beta}_\varphi)^T[M(X^T X)^{-1}M^T]^{-1}(M\hat{\beta}_\varphi)/q}{\hat{\tau}_\varphi^2} > \chi^2_{1-\alpha}(q)$$

where $q = dim(M)$ is the number of parameters not included in the reduced model, and $\chi^2_{1-\alpha}(q)$ is the $(1-\alpha)\times 100\%$ percentile of a chi-square distribution with q degree of freedom. However,

simulations suggest using F critical value and so the null hypothesis is rejected if

$$\frac{(M\hat{\beta}_{\varphi})^T[M(X^TX)^{-1}M^T]^{-1}(M\hat{\beta}_{\varphi})/q}{\hat{\tau}_{\varphi}^2} > F_{1-\alpha}(q, n-p-1)$$

where $F_{1-\alpha}(q, n-p-1)$ is the $(1-\alpha)\times 100\%$ percentile of a F distribution with q and $n-p-1$ degrees of freedom.

Another useful rank-based test is the drop in dispersion one which is based on both the full and reduced model estimates. Let $\hat{\beta}_r$ be the rank-based coefficient estimate of the model constrained by H_0, i.e. the reduced one, and $D_W(\hat{\beta}_r)$ and $D_W(\hat{\beta}_f)$ that are the reduced and full model minimum dispersions with the Wilcoxon score functions. Then, the drop in dispersion test has an asymptotic chi-square distribution but, similarly to the Wald test, simulations suggest using F-percentiles, and the null hypothesis H_0 of the test is rejected if

$$\frac{[D_W(\hat{\beta}_r) - D_W(\hat{\beta}_f)]/q}{\hat{\tau}_{\varphi}/2} > F_{1-\alpha}(q, n-p-1)$$

Thus, applying it separately for each variable, the test tells whether there is a reduction in dispersion by switching from the model with only one biomarker to the null one and whether it is significant. This may also be a measure of the difference in biomarkers in the two classes.

The R software provides the *Rfit* package which uses linear models to make inference and meets the previous requirements. The main available function is *rfit* which has syntax and outputs similar to those of *lm* to fit linear model, and it has as one of the arguments the scores that can be those of Wilcoxon, normal but also bent, where the latter are recommended if the errors are skewed distributed. Thus, from its implementation come the rank-based coefficient estimate $\hat{\beta}_{\varphi}$ and then the Wald test outcomes with the p-values related to the t-statistic of each feature. Moreover, it shows the results associated with the drop in dispersion test which can be used to compare different models through the *drop.test* function.

The results of these tests can be shown using volcano plots as described earlier, identifying significant variables and among these those of interest. In this way, the variables are split into up-regulated and down-regulated according to the sign of log2 fold change and this may also be shown through boxplots.

5.1.3 Preprocessing

Before proceeding to apply the machine learning methods described in the previous chapter, the data analysis is completed with a preprocessing phase.

First, any variables with zero or near-zero variance that may cause instability and failures in the models must be identified and, if any, removed. The *caret* package contains the *nearZerVar* function to accomplish this task.

Then, the variables need to be standardized to make comparable even those with mean and standard deviation measured on a different scale: each variable is centered by subtracting its mean value and scaled dividing its values by its standard deviation

$$Z_i = \frac{X_i - \mu_i}{\sqrt{Var(X_i)}}$$

so that it has zero mean and unit variance. This is especially important for the analyses that follow, and thus for many classification methods. For example, standardization is crucial for SVMs since the kernels are distance-based, but also for regression analyses because the importance can be computed by sorting the absolute values of the standardized estimated coefficients in descending order. Once the operation on the discovery set is carried out, the same is reproduced on the testing observations by using means and standard deviations estimated with the data of the first set. This is executed in R by using the *preProcess* function of the *caret* package, with the *center* and *scale* methods.

Finally, it needs to explore the correlations of predictors through the *cor* function. Thus, while some models may benefit from the presence of correlated predictors, others conversely may gain from a reduced correlation between them. For these reasons, it is chosen to remove variables that are strongly correlated with others in the discovery set and that might negatively affect subsequent analyses: once a correlation threshold is fixed, the variables with at least one correlation in absolute value above it are selected and for each pair of correlated features, that having the highest mean value is removed from both the discovery and the testing sets.

At this point, one has the same two starting sets without any discovery observations with missing values and without the variables that turned out to be highly correlated or found to have no discriminatory power between cases and controls.

Therefore, one proceeds to search for the disease signature: the discovery set is used first to find the most suitable classification method for the dataset and the problem, and then to derive the final model and signature. The testing set, on the other hand, is used to test the built model on

a data collection independent of that of discovery, and this results to be a key step in evaluating the performance of the signature on a generic set and comparing it to that of the current standard non-invasive method.

5.2 Method selection

Now, let's proceed to choose the best method for searching a disease signature. Indeed, starting with any number of statistical techniques, it is first required to select one of them, and then to find the best hyper-parameters for it. Specifically, this work considers the eight methods briefly described in the last section of Chapter 4.
The first goal is achieved by exploiting the Nested Cross-Validation, which is a resampling technique improving simple Cross-Validation. The latter, instead, is employed for the construction of the final model and signature. Only the discovery set is taken into account at this stage.
The salient features of these two resampling techniques are now stated, and then it is explained how they are applied in the procedure being constructed.

5.2.1 k-fold Cross-Validation

k-fold Cross-Validation consists of splitting the discovery set into k non-overlapping subsets (folds) having the same number of observations, by preserving the class proportions of the starting collection. If k is set equal to the number of observations in the collection, it leads back to the Leave-One-Out Cross-Validation (LOOCV) that is computationally more expensive and doesn't allow sample stratification.
Once a validation criterion C is chosen, at each iteration, the k-th fold becomes the validation set for model evaluation while the residual subsets constitute the set of training for building the model itself. Thus, the model is fit on the training data and then validated on the remaining ones by returning a value C_i of the validation criterion. This procedure is repeated for each fold and results are finally averaged to provide an overall assessment of the model.
The validation measure changes according to the problem type: in regression problems usually the mean square error (MSE) is chosen, while in classification studies such as the one under investigation, the classification error or the AUC score are mainly preferred. So, in the first case, C_{mean} is the mean error estimate, while in the second one it is the mean misclassification error or the mean validation AUC score.
This technique is convenient for model selection or hyper-parameters tuning on a given collection

of data, and for choosing the appropriate level of model flexibility: it is repeated for all possible combinations of hyper-parameters, and the model for which the established criterion is maximized or minimized is picked. Thus, the chosen technique is retrained on the whole discovery set with the optimal hyper-parameters. The same holds for comparing several models.

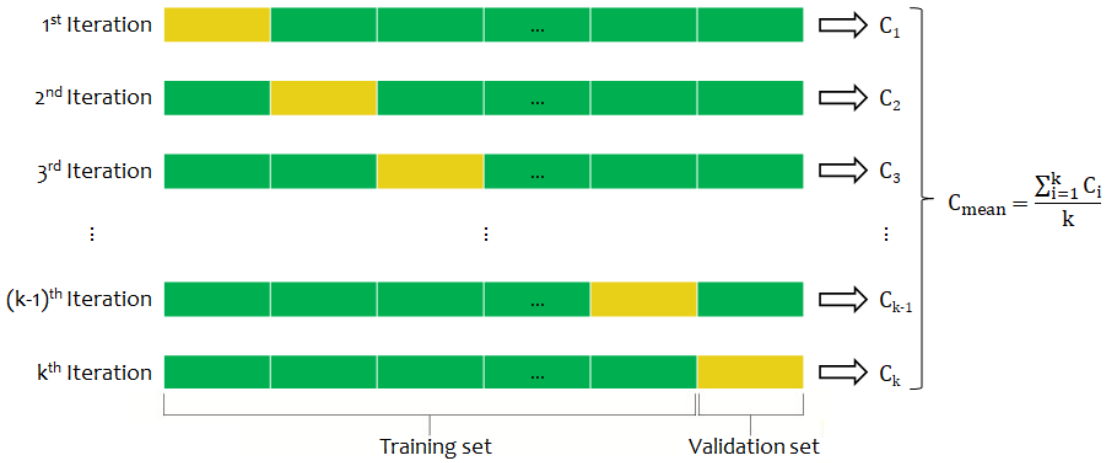


Figure 5.1: k-fold Cross-Validation

In this way, k-fold Cross-Validation avoids, for example, overfitting and selection bias by measuring how the model performs with a generic collection of data. Indeed, the objective of the technique is to simultaneously minimize the variance and the bias of the taken model, where the former refers to the working of the model itself on a set different from the training one, while the latter indicates an error introduced by the approximation of a complex real-life problem with a simpler model. However, these measures have different behaviors and when the first increases, the second decreases and vice versa. For this reason, it is usually referred to as a bias-variance trade-off since a balance of the two measures is necessary.

The choice of k-value is critical in this context, as it pushes the technique to an intermediate level between LOOCV and the simple validation set approach, where the former leads to high variance and computational time, and low bias, while the latter has high bias, low variance, and a reduced computational time. Usually, $k = 5$ or $k = 10$ are selected, but $k = n$ is preferred for very sparse dataset. However, this also depends on the size of the collection at hand such that each training and validation set of samples is large enough to statistically representative of the broader dataset.

5.2.2 $A \times B$ Nested Cross-Validation

When a procedure is required for both hyper-parameter tuning and model evaluation, it is not convenient to use a simple Cross-Validation. In this context, a good alternative may be the Nested Cross-Validation which is another resampling technique, defined from the k-fold Cross-Validation. It consists of two loops, one internal and one external, and each of them is associated with a k-fold Cross-Validation with different purposes: the first one is exploited for searching the best values of hyper-parameters and selecting the model, while into the external loop the built models are evaluated through some validation metrics and the comparison between different machine learning methods are allowed. For simplicity, it is denoted by $A \times B$ Nested Cross-Validation where the values of A and B are associated with the previous two loops, respectively. Let's consider a collection of discovery data, a classification method with an associated grid of hyper-parameter values and a given criterion C to be maximized, to describe in detail how the Nested Cross-Validation works. The following procedure, that is summarized in Figure 5.2, is the same when considering a criterion to be minimized.

In the outer loop, an A-fold Cross-Validation occurs: the collection of data is split into A subsets and, at the k-th step, the k-th subset becomes the validation set while the remaining $A-1$ folds constitutes the training data for the model building, which is performed through an inner B-fold Cross-Validation. In the inner resampling, indeed, the set is split into B subsets for searching the optimal hyper-parameters among the provided grid of values. Thus, for each hyper-parameters set a B-fold Cross-Validation is executed and the value C_i of the chosen criterion is calculated for each validation set of the inner loop. These measures are then averaged by obtaining C_{mean} and the set of hyper-parameter values which leads to the highest C_{mean} is selected.

Subsequently, the model with the chosen hyper-parameters is re-fitted on the whole training set of the external sampling and then validated on the k-th fold, returning $C_{VAL,k}$. All this is repeated for all A validation sets leading to A different models since they are built on different training sets even if they could have the same hyper-parameters. Additionally, each of these is associated with its value of the validation criterion.

For the overall model performance evaluation, one may estimate the mean performance of the method on all the validation sets of the outer resampling through the validation values $C_{VAL,1}, C_{VAL,2}, \ldots, C_{VAL,A}$ and the estimate of its mean value $C_{VAL,mean}$ with the associated 95% confidence interval (CI). Then, these results may be used for comparing many algorithms and selecting the best one.

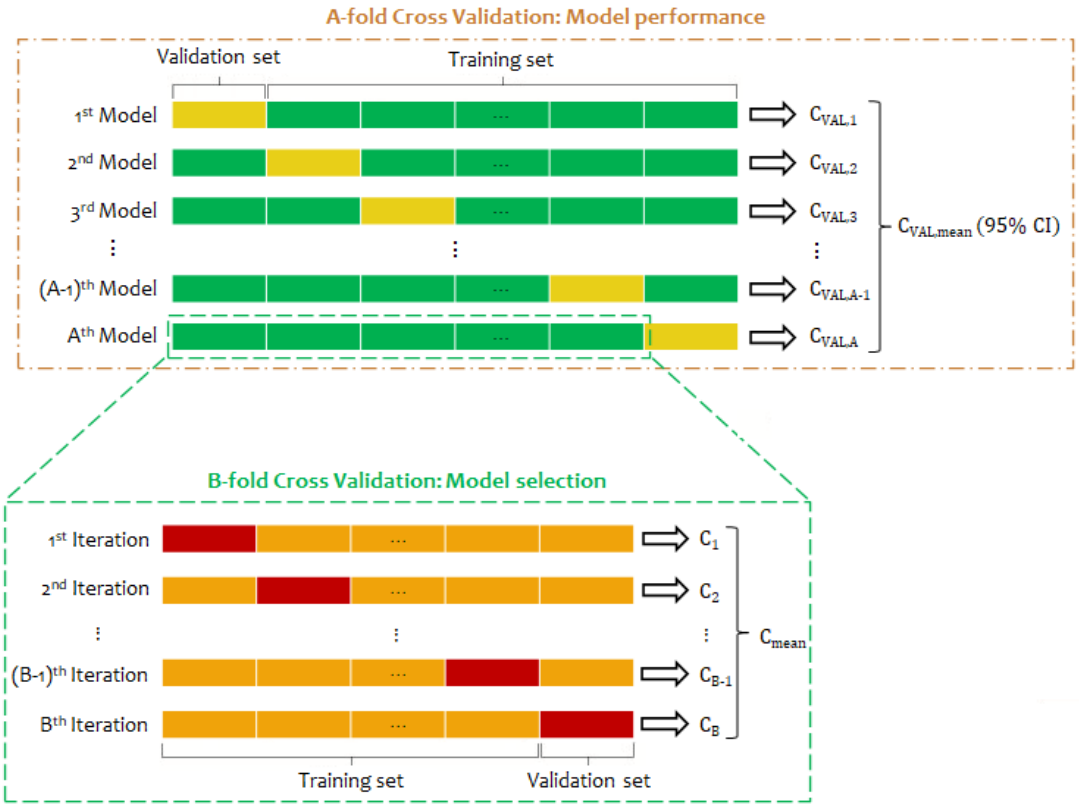


Figure 5.2: $A \times B$ Nested Cross-Validation

However, the newly defined $A \times B$ Nested Cross-Validation may be repeated several times to reach more stable and robust results by considering different partitions of the discovery set in the outer resampling. In this way, with k repetitions the number of built models and thus of validation values of the criterion increase to $k \times A$.

This approach is very efficient and reduces, if not delete, the risk of overfitting on the dataset because each model is validated only on a subset of the collection of data, and it also guarantees a less biased performance through the inner loop of the Nested Cross-Validation, in which model selection becomes an integrated component of the model fitting procedure.

A disadvantage concerning the $A \times B$ Nested Cross-Validation is certainly related to the computational cost. Indeed, given H sets of hyper-parameters of a selected method, $B \times H$ different models are built and validated in the inner loop, and this task is repeated A times, as many as the number of folds of the outer resampling. Finally, this is reiterated as many times as k, yielding a total cost of $A \times B \times H \times k$.

Therefore, the choices of A and B for the resampling and then the number k of repetitions and the size H of hyper-parameters set are crucial and should take into consideration the available data but also the overall cost required by each method. Here, to reduce the computational cost, the number of possible combinations of hyper-parameter values is limited at first. Then, a more careful search for the optimal hyper-parameters follows, but only with the resulting best performing method.

Coming back to the maximization criterion, as already remarked for k-fold Cross-Validation, accuracy is often used in a classification problem but in case-control studies, it is suggested to use the AUC score which measures the separation of two categories. These metrics are defined in the next sections together with others.

5.2.3 Application for biomarker signature

In the following case-control study, considering also the size of the dataset, the maximization criterion of the AUC score is employed in 5×5 and 5×10 Nested CVs, both without repetitions and with 3 repetitions, on the discovery set to select the best classification method for searching a biomarker signature.

RStudio offers several alternatives to implement these techniques, such as the *trainControl* function in the *caret* package which admits among its arguments the resampling technique with the associated numbers of sampled sets and repetitions. This function is then required as an argument of *train* once it is applied with a classifier.

To display and analyze the results at each iteration, since a *seed* is used in R to divide the discovery set, it is preferred to use a double *for* loop, one on repetitions and one for the external resampling, and then the employment of a simple B-fold Cross-Validation in the *train* function. To avoid these steps which obviously increase the computational cost, instead, you can use the *repeatedcv* method directly without setting any random seeds.

Taking up the models briefly defined in section 4.2, Table 5.1 summarizes the classifiers implemented in this procedure together with the associated methods for the *train* function.

A peculiarity of the strategy developed concerns the performance comparison of the chosen classifiers with all the resampling techniques employed. In this work, for each pair of classifier and Nested Cross-Validation, the mean value $C_{VAL,mean}$ of the validation AUC scores with the associated 95% confidence interval and the percentage of validation AUCs exceeding an acceptability threshold, set equal to 0.80, are computed. Then, the best method is chosen by looking at these results but also at the level of interpretability of the model itself.

Due to these conditions and the results shown in the next chapter for the case study being examined, the most interpretable and best-performing method is the Generalized Linear Model (GLM) with elastic-net regularization, a special case of GLMs, that is among the most widely adopted techniques in clinical trials for its high level of interpretability.

Classification model	**Method**
k-Nearest Neighbors	*knn*
Gradient Boosting	*gbm*
Random Forest	*ranger*
Support Vector Machines (SVMs)	*svmLinear* (linear kernel) *svmPoly* (polynomial kernel) *svmRadial* (radial kernel)
GLM with elastic-net regularization	*glmnet*
Neural Network	*nnet*

Table 5.1: Classification methods involved as arguments of the *train* function

5.3 Generalized Linear Models

A Generalized Linear Model (GLM) is obtained from the general linear model in which a response variable y is defined through the linear combination of p variables called predictors. In formula, for the i-th observation the general linear model can be written as

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p-1} + \epsilon_i$$

where the residual ϵ_i contains what the variables fail to explain and satisfies $\sum_{i=1}^{n} \epsilon_i = 0$. In matrix form, on the other hand, the previous expression for n observations becomes

$$y = X\beta + \epsilon$$

where X is a $n \times p$ matrix, $\beta = (\beta_0, \beta_1, \ldots, \beta_{p-1})^T$ a p-dimensional vector of coefficients, and $\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^T$ is a vector of n independent error terms with $\epsilon \sim \mathbb{N}_n(0_n, \sigma^2 1_n)$. The goal is to

estimate the vector $\hat{\beta}$ of coefficients which minimizes $||y - X\beta||^2$ and results $\hat{\beta} = (X^T X)^{-1} X^T y$. Turning to a GLM, the previous relationship is expressed by a link function g as follows:

$$\eta = g(\mu) = X\beta$$

where $\mu = \mathbb{E}(y)$, and y takes on an exponential family distribution. More specifically, a random variable y_i belongs to the exponential family if its density can be traced back to the following expression

$$f(y_i; \theta_i, \phi_i) = \exp\left\{ \frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)} + c(y_i, \phi_i) \right\}$$

where θ_i and ϕ_i are the natural and the dispersion parameters, respectively. Furthermore, it results

$$\begin{aligned} \mathbb{E}(y_i) =& b'(\theta_i) = \mu_i \\ Var(y_i) =& b''(\theta_i) a(\phi_i) \end{aligned}$$

The vector $\hat{\beta}$ of the GLM coefficients is estimated by looking at the log-likelihood function

$$\mathcal{L}(\beta, y) = \sum_{i=1}^{n} \log f(y_i; \theta_i, \phi_i) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)} + \sum_{i=1}^{n} c(y_i, \phi_i)$$

and, specifically, the score statistic

$$U = \frac{\partial \mathcal{L}(\beta; y)}{\partial \beta}$$

is set equal to zero, whence it follows that

$$U_j = \frac{\partial \mathcal{L}(\beta; y)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(y_i - b'(\theta_i))x_{ij}}{Var(y_i)} \cdot \frac{\partial \mu_i}{\partial \eta_i} = 0, \quad j = 1,2,\ldots,p$$

with $\eta_i = \sum_{j=1}^{p} \beta_i x_{ij} = g(\mu_i)$ for the link function g. If it is not possible to perform these calculations, numerical algorithms are often used.

According to the Law of Large Numbers, for large n values the estimator $\hat{\beta}$ has asymptotically a normal distribution with a covariance matrix that is the inverse of the same for the score statistic, called information matrix J, with elements

$$J_{jk} = \mathbb{E}\left[\frac{-\partial^2 \mathcal{L}(\beta)}{\partial \beta_j \partial \beta_k} \right]$$

Furthermore, recalling that y_i are independent and identically distributed in the exponential family, it follows that

$$\mathbb{E}\left[\frac{-\partial^2 \mathcal{L}_i}{\partial\beta_j \partial\beta_k}\right] = \mathbb{E}\left[\left(\frac{\partial \mathcal{L}_i}{\partial\beta_j}\right)\left(\frac{\partial \mathcal{L}_i}{\partial\beta_k}\right)\right]$$

whence

$$\begin{aligned} J_{jk} &= \mathbb{E}\left[\left(\sum_{i=1}^{n} \frac{(y_i-\mu_i)x_{ij}}{Var(y_i)} \cdot \frac{\partial\mu_i}{\partial\eta_i}\right)\left(\sum_{h=1}^{n} \frac{(y_h-\mu_h)x_{hk}}{Var(y_h)} \cdot \frac{\partial\mu_h}{\partial\eta_h}\right)\right] = \\ &= \sum_{i=1}^{n} \frac{\mathbb{E}(y_i-\mu_i)^2 x_{ij}x_{ik}}{(Var(y_i))^2} \cdot \left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 = \\ &= \sum_{i=1}^{n} \frac{x_{ij}x_{ik}}{Var(y_i)} \cdot \left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 \end{aligned}$$

where the latter step is due to the relation $\mathbb{E}(y_i-\mu_i)^2 = Var(y_i)$, and in matrix form it becomes

$$J = X^T W X$$

being W a diagonal matrix with main-diagonal elements

$$W_{ii} = \frac{1}{Var(y_i)} \cdot \left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2, \qquad i = 1,2,\ldots,n$$

From all these remarks, it follows that

$$\hat{\beta} \sim \mathbb{N}(\beta, J^{-1})$$

and you can test on individual β_j as well as β. Besides, one may proceed to estimate the confidence intervals of $\hat{\eta}_i$ and, if g is not the identity link function, those of $\hat{\mu}_i$ by applying g^{-1}. The log-likelihood may be also used to make tests on the model and, specifically, to compare the current model with the saturated one, which is the most accurate with as many parameters as possible and $\hat{y}_i = y_i$. Denoting by $\mathcal{L}(y;\hat{\beta}_{max})$ and $\mathcal{L}(y;\hat{\beta})$ the log-likelihood for the saturated model and that of the built one, respectively, the formula

$$D = 2\Big(\mathcal{L}(y;\hat{\beta}_{max}) - \mathcal{L}(y;\hat{\beta})\Big) \sim \chi^2_{m-p,\nu}$$

defines the deviance with non-centrality parameter $\nu = \mathcal{L}(y;\beta_{max}) - \mathcal{L}(y;\beta)$, and the model becomes better as the deviance decreases. Similarly, deviance is useful for comparing different sub-models.

5.3.1 GLM for binary data

One of the classification methods most frequently applied in clinical studies to make a diagnosis is the logistic regression, which is a special case of GLM wherein the response variables are generated by the binomial distribution

$$y_i \sim Binomial(\pi_i, n_i)$$

with y_i that indicates the number of successes in n_i trials, and for which $\mathbb{P}(y_i = 1) = \pi_i$ and $\mathbb{P}(y_i = 0) = 1 - \pi_i$. Thus,

$$\mathbb{E}(y_i) = \mu_i = n_i\pi_i$$
$$Var(y_i) = n_i\pi_i(1-\pi_i)$$

and the probability density function is

$$f(y_i; n_i, \pi_i) = \binom{n_i}{y_i}\pi^{y_i}(1-\pi_i)^{n_i-y_i} = \\ = \exp\left\{ y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + n_i \log(1-\pi_i) + \log\binom{n_i}{y_i}\right\}$$

with $\theta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right)$ and $\phi_i = 0$.
The link function takes on the log-odds form

$$g(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \sum_{j=1}^{p}\beta_j x_{ij} = x_i\beta, \qquad i = 1,2,\ldots,n$$

and then

$$\pi_i = \frac{\exp(x_i\beta)}{1+\exp(x_i\beta)}$$

Observing that

$$1 - \pi_i = \big(1+\exp(x_i\beta)\big)^{-1}$$

it results that π_i is monotone in each variable in accordance with the coefficients' sign: high probability of $y_i = 1$ are reached for large values of $x_i\beta$, while conversely, low values of $x_i\beta$ result in high probability values of $y_i = 0$. For determining the significance of the coefficients, the cases of a quantitative and a qualitative variable are analyzed separately.

For a quantitative variable, β_j indicates the slope of the straight-line tangent to the logistic curve at any point and so the instantaneous variation rate in π_i at that point. Indeed:

$$\frac{\partial \pi_i}{\partial x_{ij}} = \beta_j \frac{\exp(x_i\beta)}{(1+\exp(x_i\beta))^2} = \beta_j \pi_i (1-\pi_i)$$

and the maximum slope equal to $\frac{\beta_j}{4}$ is obtained for $\pi_i = \frac{1}{2}$, while it tends to 0 as π_i moves towards 0 or 1. Therefore, if $\beta_j > 0$, an increment of the value x_{ij} leads to a rise of π_i while if $\beta_j < 0$, on the contrary, to the same increment corresponds a decrease of π_i.

Moving on to qualitative variables, it is useful to consider a logistic model with p predictors in which the k-th predictor is qualitative and it is assumed that it takes the value 1 if it belongs to a class (e.g. diseased) and 0 otherwise. By supposing $x_{ik} = 1$ and $x_{hk} = 0$ for two observations i and h, while $x_{ij} = x_{hj}$ for $j = 1,2,\ldots,k-1,k+1,\ldots,p$, it results that

$$OR = \frac{\frac{\pi_i}{1-\pi_i}}{\frac{\pi_h}{1-\pi_h}} = \frac{\exp\left\{\beta_0 + \sum_{j=1}^{k-1}\beta_j x_{ij} + \beta_k + \sum_{j=k+1}^{p}\beta_j x_{ij}\right\}}{\exp\left\{\beta_0 + \sum_{j=1}^{k-1}\beta_j x_{ij} + \sum_{j=k+1}^{p}\beta_j x_{ij}\right\}} = e^{\beta_k}$$

and then the logarithm is

$$\log(OR) = logit(\pi_i) - logit(\pi_h) = \beta_k$$

So, all other values being equal, for $\beta_k > 0$ disease subjects have a higher probability π_i and moving from $x_{hk} = 0$ to $x_{hk} = 1$ the probability increases, whereas with $\beta_k < 0$ the situation is the opposite.

However, the log-likelihood in the logistic model results in

$$\mathcal{L}(y_i, \pi_i, n_i) = y_i \log(\pi_i) + (n_i - y_i)(1-\pi_i) + \log\binom{n_i}{y_i}$$

and asymptotically $\hat{\beta} \sim \mathbb{N}(\beta, J^{-1})$, with $J = X^T diag(n_i\pi_i(1-\pi_i))X$, while the deviance is

$$D = 2\sum_{i=1}^{n}\left[y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - \hat{y}_i}\right)\right]$$

where the first term refers to successes while the second one to failures. Thus, the latter is a sum over the success and failure totals multiplied by 2, and it takes the general form

$$D = 2\sum_{i=1}^{n}\left[obs\ success \times \log\left(\frac{obs\ success}{est\ success}\right) + obs\ failure \times \log\left(\frac{obs\ failure}{est\ failure}\right)\right]$$

5.3.2 GLM with elastic-net regularization

Starting with GLMs, one can move on to regularization methods by introducing bias in the model to gain a drop in variance and adding a penalty term $P(\beta)$ to the log-likelihood function to be maximized to estimate the β vector, as

$$\mathcal{L}^*(\beta) = \mathcal{L}(\beta) - P(\beta)$$

where only the log-likelihood $\mathcal{L}(\beta)$ is a data-depend term while the penalty $P(\beta)$ is a function of the model smoothness which doesn't include the model intercept. These are called penalized-likelihood methods and usually, they use the L_q-norm smoothing function

$$P(\beta) = \lambda\sum_{j=1}^{p}|\beta_j|^q$$

where $\lambda > 0$ is the smoothing or shrinkage parameter. The explanatory variables in this model should be standardized as they are equally treated in the shrinkage function.

Two of the most applied methods are the ridge regression and the lasso (least absolute shrinkage and selection operator) which employ the L_2 and L_1 norms, respectively. In the first one, the log-likelihood maximization is constrained to $\sum_{j=1}^{p}\beta_j^2 \leq \lambda^*$ where λ^* is the inverse value of λ, while in the second method it is subject to $\sum_{j=1}^{p}|\beta_j| \leq \lambda^*$. For both methods, as the shrinkage parameter λ increases, the least squares estimate shrinks 0 but in the lasso, as λ increases, more β_j are shrunken to 0 leading to variable selection and hence less complex models. In other words, only lasso includes feature selection and it is great when a simpler and more interpretable model is desired: it selects one of the correlated features while ridge regression shrinks correlated features together.

Thus, both techniques have the same formulation but different constraints: for the bivariate case with $p = 2$, these conditions reduce to a square in the lasso and a circle in the ridge regression both centered in the origin. Moreover, for linear normal models, the isovalue curves of the log-likelihood are elliptically centered on the $\hat{\beta}$ estimate as well as approximately for the other GLMs

with large n. Thus, the lasso estimate occurs when an ellipse touches the square constraint and sometimes results in some β_j set equal to 0, while in the ridge regression when it touches the circular constraint, as shown in Figure 5.3.

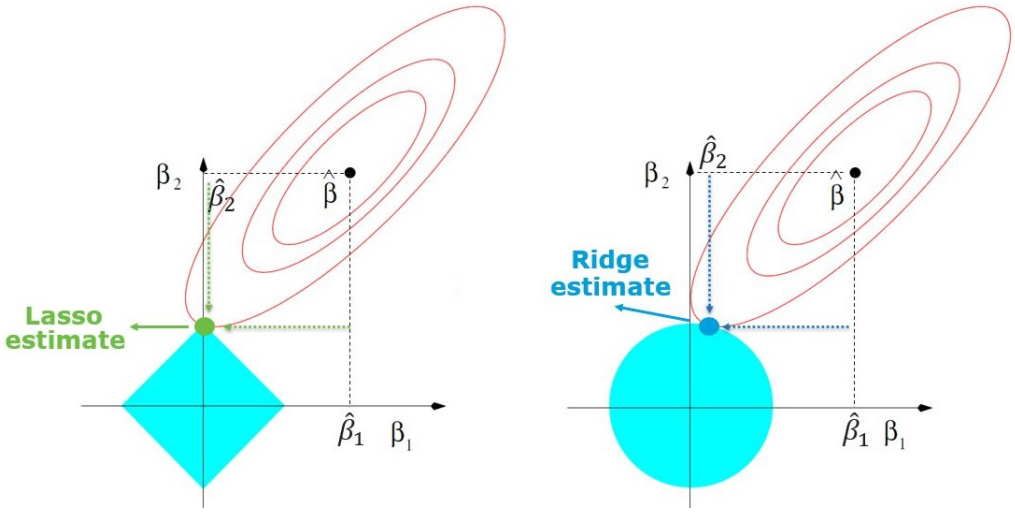


Figure 5.3: Geometrical interpretation of lasso (left) and ridge regression (right) for bivariate models [53]

A more complex method is the GLM with elastic-net regularization where the penalty term is defined as a linear combination of L_1 and L_2 norms. Thus, the vector of coefficients $\hat{\beta}$ is estimated by maximizing

$$\mathcal{L}^*(y;\beta) = \mathcal{L}(y;\hat{\beta}) - \lambda\left(\frac{1-\alpha}{2}\sum_{i=1}^{p}\beta_i^2 + \alpha\sum_{i=1}^{p}|\beta_i|\right) = \mathcal{L}(y;\hat{\beta}) - \lambda\left(\frac{1-\alpha}{2}||\beta||_2^2 + \alpha||\beta||_1\right)$$

where $\mathcal{L}(y;\hat{\beta})$ is the log-likelihood and α is the regularization parameter: if $\alpha = 1$ the method is reduced to the lasso, while if $\alpha = 0$ it leads to the ridge regression.
Considering only the L_2-norm penalty, the method includes all predictors because the smoothing term shrinks coefficients towards zero, without setting any of them equal to zero. On the other hand, the L_1-norm penalty allows for more interpretable model shrinking coefficients by setting some of them exactly equal to zero: it can build models with any number of variables through variable selection and some numerical methods should be applied to solve it.
The hyper-parameters optimization and then the variables selection can be execute through a k-fold Cross-Validation or a Nested Cross-Validation as an integrating section of the model building. However, it is very important not to forget the variable standardization.

Then, through the linear combination of the two different penalty terms, the elastic-net GLM attempts to overcome the limitations of the two extreme models: as the regularization parameter increases, the number of non-zero coefficients decreases. So, combining their strengths it involves feature selection and shrinks correlated features together.

5.3.3 R implementation

RStudio provides a broad variety of packages and functions for implementing GLMs and *glm* is probably the most used function to fit them.
Furthermore, the *glmnet* package contains efficient procedures for implementing lasso, ridge regression or elastic-net regularization in linear, logistic or multinomial regression models as well as in many other models. In particular, the *glmnet* function allows the fitting of all these models and the main hyper-parameters are:

- *alpha*, the elastic-net mixing parameter which takes values in the range [0,1], and the penalty with the extreme interval values reduces the model to ridge regression and lasso, respectively;
- *lambda*, the shrinkage parameter by which the penalty term is multiplied to be subtracted from the log-likelihood;
- *family*, a description of the link function to use in the model and, similarly to the *glm* function, *binomial* is the one required for the method just defined.

5.4 Model analysis

After selecting the best performing method at the previous step, the optimal set of hyper-parameters is found from a broader grid of values than those used in Nested Cross-Validations, through a k-fold Cross-Validation on the discovery set, i.e. the same set used so far. Once some optimal values have been found, a deeper analysis is done around these to derive the optimal hyper-parameters.
As already accomplished in the previous stage, the maximization of the AUC score is chosen as the criterion for the resampling techniques. Indeed, in a problem like this one where you want to make a diagnose, the AUC score is managed as a reference metric, since it measures the ability to discriminate the two classes of the categorical variable, and thus the ability to distinguish healthy and diseased subjects. This choice is also due to the deliberation that datasets are not

always balanced, and in the unbalanced case, i.e. when the set is not equally or nearly equally divided between the two different diagnoses, the accuracy has several limitations.

Thus, once the optimal hyper-parameters are chosen and the final model is re-fitted on the discovery set, an optimal probability threshold is found for the classification, among those that maximize or minimize some statistics, through the analysis of some evaluation metrics, before the assessment of the model performance is carried out on the testing set by using the same metrics.

The procedure outlined that is better described in the coming sections is valid not only for the previous method but for any binary classifier that results as the best technique from the Nested Cross-Validations. Nevertheless, due to their high interpretability, the logistic regression and the GLM model with elastic-net regularization are among the most widely employed methods in clinical studies, for example, to diagnose a disorder.

Now, some important tools like the confusion matrix and the ROC curve should be characterized to define the AUC score and other evaluation metrics. Then it is described how to reach the final model.

5.4.1 Confusion matrix and evaluation metrics

Let's start with the confusion matrix or classification table, which is a summary of prediction results of a classification model. For a binary problem into which observations are labeled as *Positive* (P or 1) or *Negative* (N or 0), it cross-classifies the binary response y and the prediction $\hat{y}$ as displayed in Figure 5.4, where the element (i, j) of the matrix indicates the number of patients in the j-th class and predicted as i.

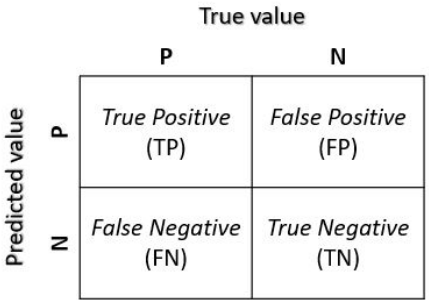


Figure 5.4: Confusion matrix for a binary classification problem

Thus, for a case-control clinical study it displays:

- True Positive (TP): diseased subjects correctly classified by the model;
- False Positive (FP): healthy patients mistakenly classified with positive diagnosis;
- False Negative (FN): diseased patients misclassified as healthy ones;
- True Negative (TN): healthy subjects correctly classified by the algorithm.

The cell entries strongly depend on a cut-off probability $\mathbb{P}_{th}$: for a given observation i, the prediction is $\hat{y}_i = 1$ if $\hat{\pi}_i > \mathbb{P}_{thr}$, otherwise it is 0. A widely used threshold value is $\mathbb{P}_{th} = 0.50$ while sometimes it is estimated by maximizing or minimizing certain statistics, as explained below. However, this matrix can be also extended to the generic case of classification into k classes.

Using what has just been explicated, many useful metrics can be defined to evaluate a binary classifier.

First of all, the accuracy indicates the fraction of correctly predicted samples

$$\widehat{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

or, in other words, the fraction of not misclassified observation. It is an effective reference metric in balanced problems, while in unbalanced ones this metric can be misleading and so it doesn't make much sense.

The predictive power of classifiers, therefore, is summarized by sensitivity and specificity: the first one, also known as recall, is defined as the fraction of diseased patients that are correctly classified as such

$$\widehat{Sensitivity} = \mathbb{P}(\hat{y}_i = 1 | y_i = 1) = \frac{TP}{TP + FN}$$

while the latter is the proportion of healthy patients that are correctly predicted as such by the model

$$\widehat{Specificity} = \mathbb{P}(\hat{y}_i = 0 | y_i = 0) = \frac{TN}{TN + FP}$$

In other words, the sensitivity is the true positive rate (TPR) while $(1 - \widehat{specificity})$ corresponds to the false positive rate (FPR).

When the dataset is unbalanced, there is a strong difference between the percentage of observed positive and negative diagnoses and a more appropriate metric is the precision or Positive Predictive Value (PPV) that indicates the fraction of truly positive diagnosis out of all the predicted

positive samples, and it is defined as

$$\widehat{Precision} = \mathbb{P}(y_i = 1|\hat{y}_i = 1) = \frac{TP}{TP + FP}$$

It may be expressed also as a function of the observed prevalence β, that is the proportion of observed positive diagnoses

$$\beta = \mathbb{P}(y_i = 1|X_i) = \frac{TP + FN}{TP + FP + TN + FN}$$

Then, the previous formula becomes

$$\widehat{Precision} = \frac{\widehat{sensitivity} \times \beta}{\widehat{sensitivity} \times \beta + (1 - \widehat{specificity}) \times (1 - \beta)}$$

Another useful metric for unbalanced data is the balanced accuracy $\hat{J}$ that follows the same principles of accuracy taking values between 0 and 1, and it is defined as the mean value of the estimated sensitivity and specificity:

$$\hat{J} = \widehat{Balanced\ Accuracy} = \frac{\widehat{sensitivity} + \widehat{specificity}}{2}$$

Finally, the F-measure is an assessment metric defined as the harmonic mean of sensitivity and precision

$$\hat{F}_1 = \frac{2}{\frac{1}{\widehat{precision}} + \frac{1}{\widehat{sensitivity}}} = 2\,\frac{\widehat{precision} \times \widehat{sensitivity}}{\widehat{precision} + \widehat{sensitivity}}$$

It takes values between 0 and 1 and the latter is obtained for both precision and sensitivity equals 1 while the minimum is achieved if at least one of the two statistics is equal to 0. That measure can be generalized for an unbalanced dataset as follows by using the observed prevalence β

$$\hat{F}_\beta = \frac{(1 + \beta^2) \times \widehat{precision} \times \widehat{sensitivity}}{\beta^2 \times \widehat{precision} + \widehat{sensitivity}}$$

and for $\beta = 0$ the previous is equal to the precision, while it is the recall as $\beta \to \infty$.
Out of the several functions available in R that generate the confusion matrix or the previously defined evaluation metrics, the *caret* package offers one called *confusionMatrix* which yields all of them simultaneously requiring only the vectors of the observed and predicted classes.

5.4.2 ROC curve and AUC score

Another tool for assessing performance is the Receiver Operating Characteristic (ROC) curve which is built with the probabilities predicted by the model on a given set of observations, and from which the AUC score is obtained. Specifically, the ROC curve illustrates the diagnostic ability of the model by changing its discrimination threshold and is drawn by plotting the sensitivity (or True Positive Rate) against the False Positive Rate (1− specificity), at various probability thresholds:

$$\widehat{ROC}(\cdot) = \Big\{(FPR(c), TPR(c)),\ c \in (-\infty, +\infty)\Big\}$$

Thus, it is a concave monotone increasing function mapping [0,1] to [0,1] that joins the points (0,0) and (1,1) and measures sensitivity and specificity as the cut-off probability thresholds changes.

To plot the curve, the probabilities $\mathbb{P}(\hat{y}_j = 1|X_j)$ of the considered observations are predicted and sorted in increasing order. Then, TP, TN, FP, and FN observations are counted and then the rates $\widehat{TPR}$ and $\widehat{FPR}$ are estimated for each probability value. Furthermore, by plotting the computed values of $\widehat{FPR}$ and $\widehat{TPR}$, and joining the points with alternate horizontal or vertical segments, you get the ROC curve.

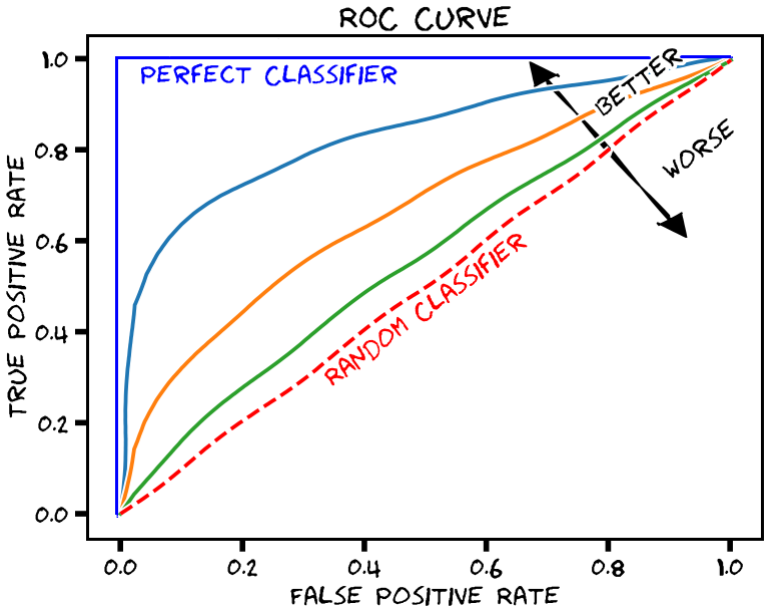


Figure 5.5: ROC curve: five example classifiers [68]

If the probability threshold tends towards 1, almost all samples have a negative prediction ($\hat{y}_i = 0$) while, on the other hand, if the same probability tends towards 0 almost all predictions are $\hat{y}_i = 1$. Moreover, as sensitivity increases, there is a reduction in specificity, and between two curves, one of these being equal, there is greater predictive power for higher values of the other metric.
In diagnostic medicine, the golden standard is achieved with $\widehat{FPR} = 0$ and $\widehat{TPR} = 1$: the ROC curve improves as it approaches the point (0,1) while it gets worse the more it becomes flattened on the diagonal line joining its two extreme points, as shown in Figure 5.5 which depicts the ROC curve for the random classifier as well as the perfect one and three middle-performing models. Thus, ROC curves are particularly useful for comparing different models and their discriminatory capacities.
Consequently, the area under the ROC curve, called AUC (Area Under the Curve) score, is a well-grounded metric to assess the classifier performance and is defined as

$$\widehat{AUC} = \int_0^1 \widehat{ROC}(t)dt$$

It takes on values between 0 and 1 and measures the degree of separability between diseased and healthy distributions: as it approaches 1, the model is able to distinguish better the two distributions, while these are overlapped if $\widehat{AUC} = 0.50$ (i.e. random guess). Thus, the greater the $\widehat{AUC}$, the better the model: as the value increases, the model performance improves, moving from an uninformative model to a highly accurate one if $0.9 < \widehat{AUC} < 1.0$ and to a perfect one with $\widehat{AUC} = 1.0$.
In other words, the AUC score is interpreted as the probability that a value extracted from the distribution of diseased people is higher than one pull out from the distribution of healthy subjects. Indeed, being FPR and TPR two function mapping a probability threshold to a value on the x and y axes respectively as

$$FPR(t) = x \quad and \quad TPR(t) = y(x)$$

it results that

$$\begin{aligned} AUC &= \int_0^1 y(x) = \\ &= \int_0^1 TPR(t)dx = \end{aligned}$$

$$= \int_0^1 TPR(FPR^{-1}(x))dx$$

Furthermore, being $f_1(x)$ and $f_0(x)$ the positive and negative probability distributions respectively, TPR and FPR in the integral form are

$$\begin{aligned} TPR(t) &= \int_t^\infty f_1(x)dx = \\ &= \int_{-\infty}^\infty \mathbb{I}(t' > t) f_1(t')dt' \end{aligned}$$

and

$$FPR(t) = \int_t^\infty f_0(x)dx$$

Applying Leibnitz's integration rule on the latter

$$\begin{aligned} FPR'(t)dt &= \frac{\partial}{\partial t} \int_t^\infty f_0(x)dx = \\ &= \frac{\partial}{\partial t} \lim_{c\to\infty} \int_t^c f_0(x)dx = \\ &= \lim_{c\to\infty} \frac{\partial}{\partial t} \int_t^c f_0(x)dx = \\ &= \lim_{c\to\infty} \frac{\partial f_0(c)}{\partial t} - f_0(t) = \\ &= -f_0(t) \end{aligned}$$

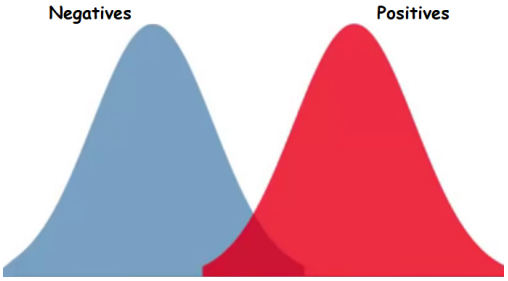
and from these considerations, one has:

$$\begin{aligned} AUC &= \int_{+\infty}^{-\infty} \int_{-\infty}^{+\infty} \mathbb{I}(t' > t) f_1(t')(-f_0(t))dt'dt = \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \mathbb{I}(t' > t) f_1(t') f_0(t)dtdt' = \\ &= \mathbb{P}(X_1 > X_0) \end{aligned}$$

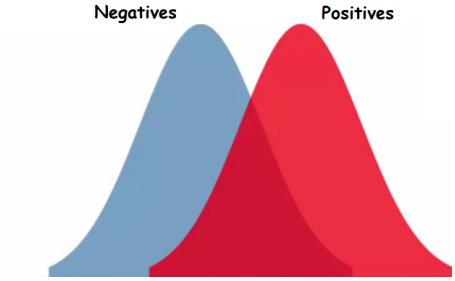
where X_1 and X_0 are extracted from the distribution of diseased people and that of healthy ones, respectively.

Thus, in clinical problems of disease diagnosis, the AUC score is an important criterion for discriminating group distributions and, specifically, in a case-control study for discriminating between cases and controls: if $AUC = 1$ the classifier is able to distinguish between all the positive and negative class, if $AUC = 0.5$ it is in the opposite situation, while if $0.5 < AUC < 1$

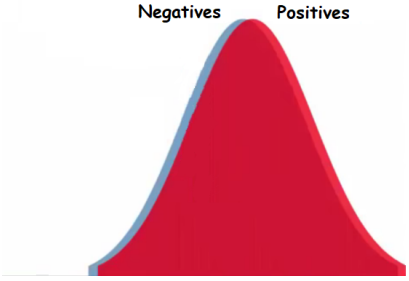
there is a high probability that the model can distinguish the positive class from the negative class and to detect more numbers of TP and TN than FP and FN (Figure 5.6).



(a) ROC curve very close to (0,1) and AUC approaching 1



(b) ROC curve at an intermediate level between perfect and random classifiers



(c) ROC curve flattening to that of random classifier and AUC approaching 0.50

Figure 5.6: Positive and negative distributions with different ROC curves and AUC scores [8]

This metric is scale-invariant because it doesn't look at predictions but how they are ranked, and classification-threshold-invariant because it measures the model's prediction quality regardless of the chosen probability threshold.

However, the AUC score is equivalent to a measure of the model predictive power, called the concordance index, that is defined as the proportion of pairwise predictions for which $\hat{\pi}_i > \hat{\pi}_j$, among all pairs of observations (i, j) having $y_i = 1$ and $y_j = 0$.

To sum up, through AUC score maximization, in method selection one chooses the classifier that better discriminates overall the two subject types on the different validation sets of external sampling associated with the Nested Cross-Validations. Similarly, in model fitting one chooses the set of hyper-parameter values that leads to a model that on average best distinguishes positive and negative observation distributions of the different validation sets of Cross-Validation.

Among the many R packages for plotting the ROC curve and computing the AUC score, *pROC* contains the function *roc* that accomplishes both tasks and requires the predicted probabilities,

the levels of the response variable and especially the positive one.

5.4.3 Final model

Once the ROC curve is built with the predictions of the discovery set, one may derive for each probability threshold the associated sensitivity and specificity. At the same time, as the threshold probability varies, one can identify correctly classified (TP and TN) but also misclassified observations (FP and FN).

So, after choosing the best method from the Nested Cross-Validations and then also the optimal set of hyper-parameters one should find an optimal probability threshold $\hat{\mathbb{P}}_{thr}$: in this way, for a binary model, the prediction for an observation i is $\hat{y}_i = 1$ if $\hat{\pi}_i > \hat{\mathbb{P}}_{thr}$, otherwise it is 0.

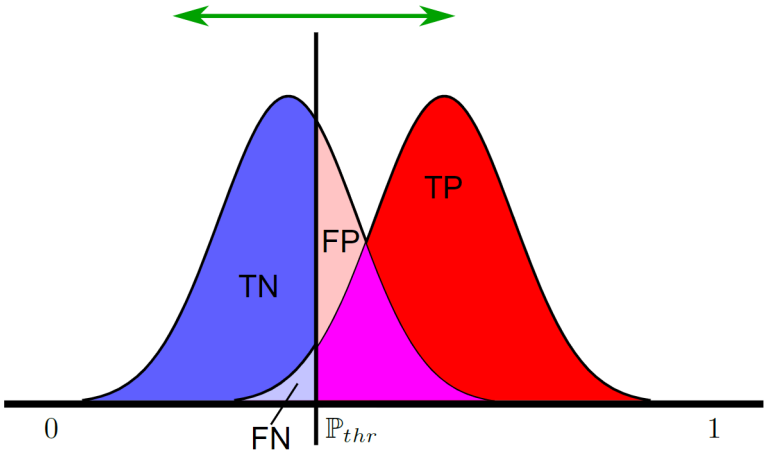


Figure 5.7: Model performances as the probability threshold changes [68]

The most frequently adopted threshold is $\mathbb{P}_{thr} = 0.50$ but sometimes, especially with unbalanced datasets, it is preferable to use other cut-off values obtained by the maximization of some statistics as the criterion.

To find an alternative cut-off probability, the built model is fitted on the discovery set and the associated ROC curve is constructed with the predicted probabilities. It would be wrong to do that with the testing set. In detail, after ordering the predicted values, some statistics are chosen and computed for each probability threshold. Then, those probabilities maximizing statistics are extracted to then choose only one by examining the estimates of accuracy, sensitivity, specificity, precision, and other assessment metrics.

In this work, the $\hat{J}$ and $\hat{F}_\beta$ statistics are considered and estimated at each probability threshold,

in order to find the probability values maximizing each of them. Thus, among these values is chosen the one with which the fitted method performs better on the discovery set.

5.5 Biomarker signature and model evaluation

Up to this point, the discovery set has been used both to choose the method and find the optimal hyper-parameters and thus the optimal probability threshold, but not to evaluate the model globally. Now, a testing set disjointed from that of discovery must be considered to give an overall model evaluation and study in detail the resulting signature for early disease detection exploiting the features of the model itself. If the provided database was not already partitioned, at the beginning of the analysis, it would be advisable to randomly extract from it the observations to be used to test the final model, keeping the others for its construction.
Resuming the GLM with elastic-net regularization previously defined, for each test observation $X_j = (x_{1j}, x_{2j}, \ldots, x_{pj})$ it calculates the probabilities

$$\mathbb{P}(\hat{Y}_j = 1|X_j) = \frac{\exp\left(\hat{\beta}_0 + \sum_{i=1}^{p} \hat{\beta}_i x_{ij}\right)}{1 + \exp\left(\hat{\beta}_0 + \sum_{i=1}^{p} \hat{\beta}_i x_{ij}\right)}$$

and the signature consists of the coefficients $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)$ estimated by the model itself. Thus, the fitted model and the associated optimal probability threshold $\hat{\mathbb{P}}_{thr}$ are employed to predict the labels of the testing observations: positive if $\mathbb{P}(\hat{Y}_j = 1|X_j) > \hat{\mathbb{P}}_{thr}$, negative otherwise. The procedure is repeated for all observations in the testing set and, finally, the performance evaluation is carried out considering the same metrics used before, such as sensitivity, specificity, and precision, and showing the confusion matrix and thus the AUC score.
Furthermore, if the binary response variable has many sublevels, the model could also be evaluated in each subgroup of cases and controls through accuracy. In this way, it can be assessed whether the performance on the whole testing set is the same in the various sub-levels, or whether the signature is better at identifying one category of diseased subjects than another.
At this point, the signature of the disease can be analyzed in detail and it is possible in GLMs by examining the estimated coefficients of the model. Indeed, all biomarkers with non-zero coefficients are included in the signature, while the others are excluded from it. At the same time, all biomarkers are quantitative variables and, for each one, the estimated coefficient indicates the slope of the straight-line tangent to the logistic curve at any point and thus the variation rate of $\mathbb{P}(\hat{Y}_j = 1|X_j)$. Then, the estimated coefficients measure the influence of each biomarker

in predictions.
Analysis of these coefficients, therefore, allows one to show which biomarkers play a more important role in disease diagnosis and which little or no role. Besides, one may study especially for the biomarkers found to be more influential, if there is any match in the available literature.

5.5.1 Model comparison

The signature may also be compared with the current best practice for non-invasive diagnosis the disease or any other tests through a statistical test on the estimated assessment metrics. To accomplish that, it is necessary to define the $(1-\alpha)$-level Wilson confidence interval.
Let's consider a sample of variables $X_1, X_2, \ldots, X_n$ independent and identically distributed (i.i.d.) as $X \sim Bernoulli(p)$. It follows that $\sum_{i=1}^{n} X_n \sim Binomial(n,p)$ and this asymptotically approximates with a normal distribution

$$\sum_{i=1}^{n} X_n \sim \mathbb{N}(np, np(1-p))$$

Furthermore, this can be rewritten as

$$\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim \mathbb{N}(0,1), \quad with \;\; \hat{p} = \frac{\sum_{i=1}^{n} X_n}{n}$$

and from it, approximate confidence intervals such as Wald's and Wilson's can be obtained. The latter, especially, asymptotically equals to the former but, for a fixed value of n, it gives better coverage, i.e. closer to the nominal constant $(1-\alpha)$.
Then, the $(1-\alpha)$-level Wilson confidence interval is defined from

$$\mathbb{P}\left(\left| \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \right| \leq z_{\alpha/2} \right) = 1-\alpha$$

where $z_{\alpha/2}$ is the $(1-\frac{\alpha}{2})$-quantile of $Z \sim \mathbb{N}(0,1)$, whence follows

$$\frac{\hat{p} + \frac{z_{\alpha/2}^2}{2n} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z_{\alpha/2}^2}{4n^2}}}{1+\frac{z_{\alpha/2}^2}{n}}$$

Starting with Wilson confidence intervals, one can move to the generic k-variate case. From statistics it is known that a k-dimensional $(1-\alpha)$-level confidence region of $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$ is

defined as the Cartesian product of k distinct one-dimensional $(1 - \frac{\alpha}{k})$-level confidence intervals related to $\theta_1, \theta_2, \ldots, \theta_k$.

Thus, in the bivariate case, one computes the Cartesian product of the $(1 - \frac{\alpha}{2})$-level Wilson confidence intervals of a pair of statistics to obtain a $(1 - \alpha)$-level Wilson confidence region of the two statistics. Therefore, considering two 95% level Wilson confidence intervals yields a confidence region of just over 90%, whereas it requires two 97.5% level confidence intervals to result in a confidence region with a coverage of about 95%.

Now, let's assume that the signature has sensitivity and specificity equivalent to those of a meta-analysis of the performance of the standard non-invasive test, setting as minimum acceptable values of sensitivity (TPR_0) and specificity ($1 - FPR_0$) the lower bounds of the confidence intervals of the same in the aforementioned meta-analysis.

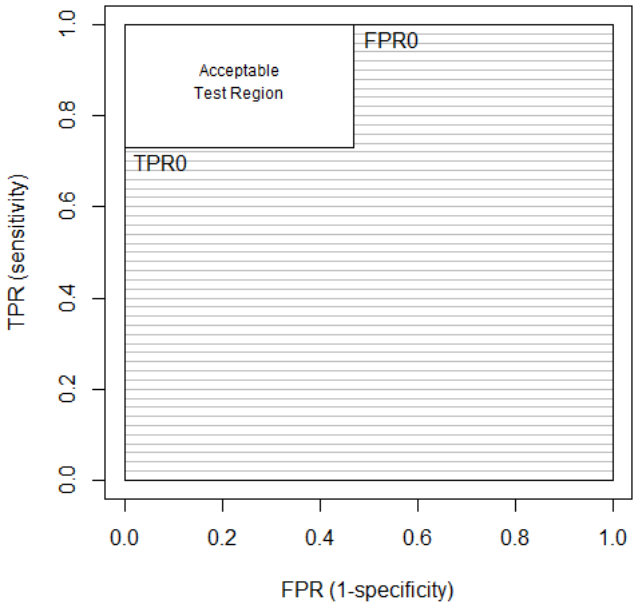


Figure 5.8: Model comparison test with the current non-invasive one

Then, the TPR and FPR 2-sided 95% Wilson confidence region for the signature is computed as the Cartesian product of the 97.5% Wilson confidence intervals of the two statistics, and if this region is entirely within the acceptable test region, the null hypothesis

$$H_0 : \{TPR \leq TPR_0 \quad or \quad FPR \geq FPR_0\}$$

meaning that the new model does not improve the performance of the considered test, is rejected and then the problem has a positive conclusion and the signature is a valid alternative to the standard non-invasive method for diagnosing the disease (Figure 5.8). Otherwise, statistics can be studied individually to try to improve diagnostic signatures later in subsequent studies.
This statistical test can then be repeated to compare different signatures but also these with several diagnostic tests already available in medicine.
Starting from the analysis undertaken so far, one could eventually proceed to a more in-depth investigation, but a team with different expertise, including clinicians and biologists, is required to do this.

Chapter 6

An application to the CIPHER protocol: miRNA biomarker signature for the early detection of PH

In this chapter, the strategy described in the previous chapter is applied to a real database in order to searching for a miRNA biomarker signature of PH. Thus, its performance is evaluated on a set of observations independent of the discovery one from which it is built, taking into account also the accuracy for each sub-level of the response variable, i.e. the different subpopulations and so the various types of diseases, healthy subjects, and those who are diseased but without PH. Finally, the signature and the biomarkers included in it are examined by comparing the results with what is already available in the literature.

6.1 Dataset description

Now let's consider the dataset provided by Actelion Janssen, extracted from the proof-of-concept study used to design CIPHER. Being the study multicenter, it is representative of various sets of populations all around the world and the values in the dataset are taken from the analysis of blood samples as described in the protocol [1].

The dataset is already divided into a discovery set with 1191 observations and a testing set with only 376 observations. Both collections have 596 variables, 40 of which are clinical data variables such as *age* and *gender*, or information for the enrollment of patients in the study such as the patient identification code (*PID*). These are excluded from the analysis at this stage as the focus is only on blood-based biomarker signatures. Other features are, however, the categorical *DANA* variable indicating the diagnoses, and 555 quantitative variables representing the miRNA biomarkers and proBNP.

Applying the built procedure, in classification models *DANA* is the response variable while the quantitative variables are the covariates. However, according to the WHO PH classification in [18], the *DANA* variable initially consists of 11 levels:

HC healthy subjects

PH0_NOPH diseased patients but no PH yet

PH1.1 IPAH patients

PH1.2 HPAH subjects

PH1.3 drugs and toxins induced PAH people

PH1.4 patients having PAH associated with other diseases

PH2 subjects with PH due to left heart disease

PH3 patients with PH due to lung disease and/or hypoxia

PH0_CTED subjects having chronic thrombo-embolic disease (CTED)

PH4 subjects with CTEPH and other pulmonary obstructions

PH5 patients with PH due to unclear mechanisms

These classes are organized into only two categories to apply the strategy for searching a biomarker signature for early detection of PH: controls denoted as NO_PH, include the first two levels, while the others are the cases indicated as PH. Notably, as suggested by the company, patients with CTED are included in cases even if CTED is characterized by symptoms and perfusion defects similar to CTEPH but without PH at rest, and there is no evidence that CTED evolves into CTEPH also because the natural history of the disorder is still unknown [36].

Thus, with reference to the only categorical variable in the dataset, both collections are made up as follows:

- Discovery set: 967 PH (81.19%) and 224 NO_PH (18.81%)
- Testing set: 270 PH (71.81%) and 106 PH (28.10%)

The two sets are unbalanced with proportions of the two classes not very different, but with a greater presence of cases and a lesser presence of controls, where the inclusion of the latter is important because it also allows you to achieve and analyze signature performance on subjects without PH.
Quantitative variables, on the other hand, indicate biomarker values obtained from 50 mL blood samples collected from patients enrolled in the study, using a qPCR-based miRNA assay technology developed by MiRXES that is a biotechnology company based in Singapore whose technologies and methodologies have broad applications in biomarker discovery and disease diagnosis.

6.2 Univariate analysis and variable selection

First, a retrospective study of miRNA biomarkers has been carried out to see which may be or not considered in the signature, by using banked blood samples collected from PH patients in several sites around the world during the last 10 years as UK, USA, and Japan. This leads to a reduction of covariates to consider in the algorithms and 334 miRNA biomarker variables with missing values in the testing set are excluded from the data collection being analyzed.
However, the discovery set turns out to have 298 missing values, corresponding to 0.11% of the values in the whole set, in 61 variables and 183 patients. Thus, it is opted to remove these observations, reducing the size of the set to 1008 observations but preserving the proportions of the two classes as in the starting one.
Furthermore, *proBNP* takes values in a very wide range (between 5 and 35000) and a logarithmic transformation is applied to it, even if keeping this name, to obtain a scale like other biomarkers. This transformation, suggested by the company, finds several supports in the available literature.
Then, one analyzes the dataset starting with the Shapiro-Wilk test: using 0.05 as test significance level, it follows that the normality null hypothesis can be rejected for 164 out of 219 quantitative variables while those significant become 130 by lowering the test significance level to 0.01. These conclusions are confirmed by analyzing the variables through skewness and kurtosis, as shown in Table 6.1, but also through quantile-quantile plot of which some examples are collected in Figure 6.1. Specifically, the picture depicts this plot for the three biomarkers with the highest p-values in the Shapiro-Wilk test and for those with the lowest value of the same: it shows that points

are arranged along a straight line for the variables in the first subset which are so approximately normal, while they go far from the line for high quantiles of biomarkers into the second subset for which the test rejects the null hypothesis of normality.

Biomarker	**p-value**	**Skewness**	**Kurtosis**
hsa.miR.625.5p	0.9294	0.0262	0.1618
hsa.miR.20b.5p	0.8937	0.0481	0.0972
hsa.miR.140.3p	0.8924	0.0453	−0.1069
hsa.miR.320c	$1.71 \cdot 10^{-14}$	0.8238	1.9102
hsa.miR.10a.3p	$5.02 \cdot 10^{-16}$	0.6268	2.9822
hsa.miR.150.5p	$1.80 \cdot 10^{-16}$	0.8246	2.6363

Table 6.1: Biomarkers with the highest and the lowest p-values in the Shapiro-Wilk test

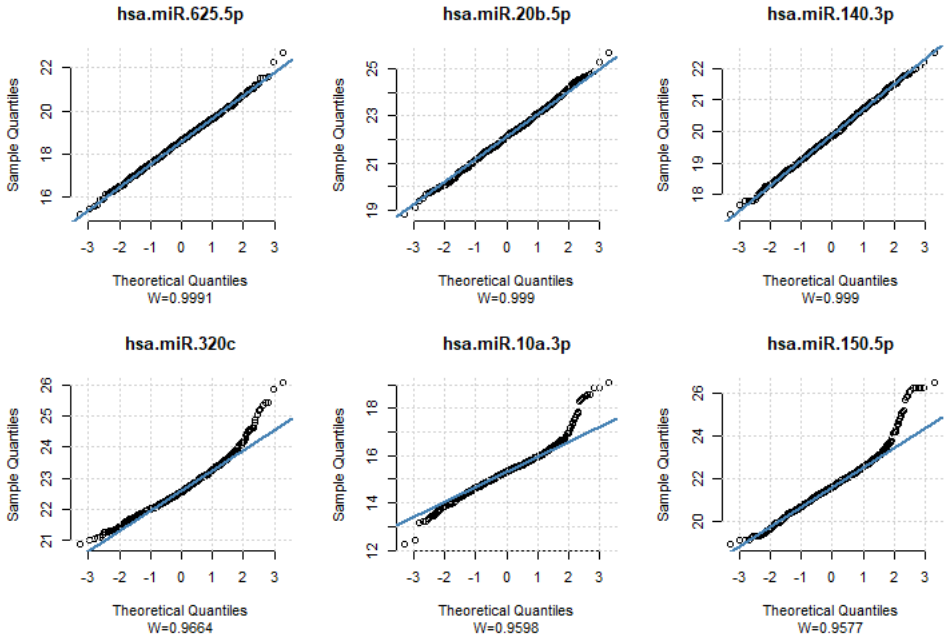


Figure 6.1: Q-Q plot for biomarkers with the highest and the lowest p-value in the Shapiro-Wilk test

Next, for each biomarker the rank-based regression

$$biomarker = \beta_0 + \beta_1 \cdot DANA$$

is fit on the discovery set, by estimating coefficients through the Wilcoxon score. Then, the drop in dispersion test associated with it is performed and the null hypothesis is rejected for 130 biomarkers if the test significance level is set equal to 0.05. Given the high number of variables and therefore the impossibility of using Bonferroni's method, by lowering the significance level of the test to 0.01 the statistically significant variables become 106.

Considering now the log2 fold change, 31 biomarkers are down-regulated while 190 are up-regulated, and thus most biomarkers have a higher mean value in cases than in controls. Combining these results with those of the previous test with the second significance level value, slightly less than half of the variables are significant: only 5 biomarkers are down-regulated while 101 are up-regulated.

	Down-regulated	**Up-regulated**	**Total**
All biomarkers	31	190	221
$p < 0.05$	9	121	130
$p < 0.01$	5	101	106

Table 6.2: Biomarkers regularization and drop in dispersion test

Furthermore, *proBNP* is the variable with the highest log2 fold change even though it is less than 1 and therefore no variable results of interest from the test, as depicted in Figure 6.2 where *proBNP* is indicated by a circle on the upper right edge.

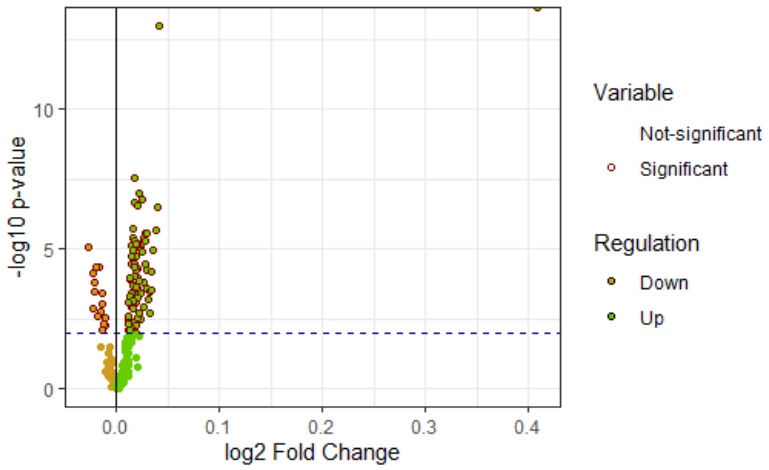


Figure 6.2: Volcano plot for the drop in dispersion test

Furthermore, the boxplots in Figure 6.3 show for each regularization group the two biomarkers with the highest log2 fold change in absolute value. From the graphs, it is clear that the variables need standardization as they are defined on different intervals, but only in *proBNP* there are clear differences between cases and controls while in the other biomarkers the distributions are less distinguishable. However, median biomarker values in subjects with PH are higher than the same in the first two boxplots, whereas an inverse behavior is observed in the remaining boxplots.

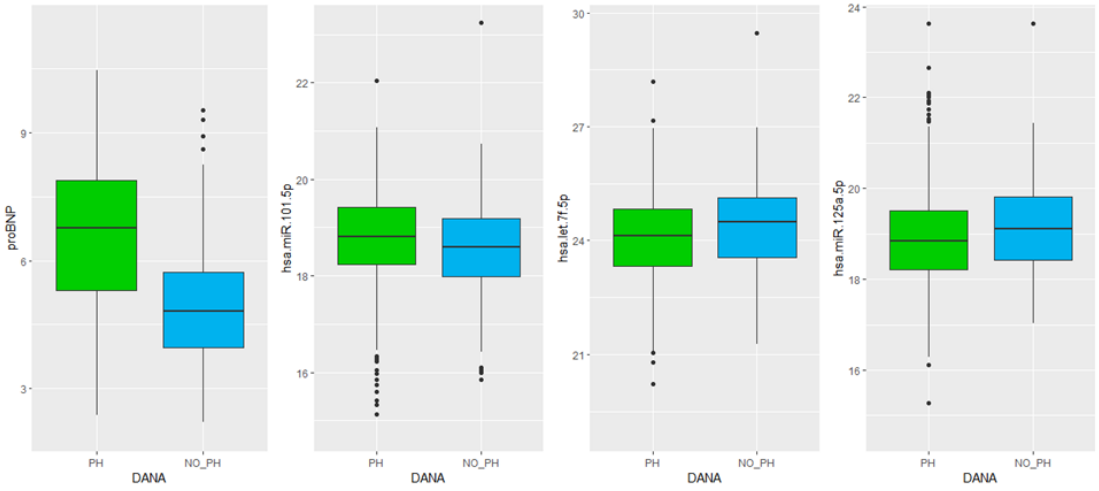


Figure 6.3: Boxplots of two up-regulated (left) and down-regulated (right) biomarkers

This analysis is now completed with a preprocessing step. First, variables are scaled and centered to be comparable, and then those highly correlated are also identified. For the latter, the threshold is set at 0.94 and three highly correlated variables are removed, reducing the total number of variables to 219. Specifically, the resulting highly correlated biomarkers are *hsa.miR.126.5p*, *hsa.miR.27a.3p*, and *hsa.miR.1285.5p*.

Coming to this point, the discovery and testing sets consist of 1008 and 376 observations, respectively. In addition, there are 219 selected variables, of which 217 are circulating miRNA biomarkers.

6.3 Method selection

After analyzing and preprocessing the dataset, let's move on to the choice of the best method by employing some Nested Cross-Validations on the discovery set.

Table 6.3 displays the results achieved by the eight selected machine learning techniques with four different Nested Cross-Validations, and thus for each of them the mean validation AUC score, the associated 95% level confidence interval, and the percentage of validation AUC above the established threshold. Specifically, the 5×5 and 5×10 Nested CVs are looked at either without repetitions or with three repetitions.

Results highlight that neural networks achieve the poorest performances with mean AUCs around 0.73, probably due to the size of the discovery set which is not large enough, and quite narrow confidence intervals especially in Nested CVs with repetitions. Furthermore, k-NN methods get pretty poor results with average values of the validation criterion around 0.70 and confidence intervals fairly larger than the other methods, with lower bounds also significantly below 0.70. Both these techniques share a weak behaviour and neural networks never manage to exceed the acceptability threshold of the validation criterion, while the other one succeeds in approaching it only in three resamplings out of four, but few times.

SVMs, on the other hand, perform better with average AUC scores around 0.75 and confidence intervals that are narrower and further narrowing as repetitions increase. Moreover, although changing the kernel does not yield a dramatic change in performance, the model with a linear kernel seems to perform slightly better than the others by taking into account all values in the table overall. Indeed, it succeeds in exceeding the acceptability threshold equal to 0.80 only once in Nested CVs with repetitions, while the SVM with radial kernel never succeeds in overpassing this threshold.

Similarly, for random forests the average value of the validation criterion is around 0.75, but with fairly wide ranges in cases without repetitions. The mean value rises both as the number of repetitions increases and as the number of folds of the internal sampling of the Nested CV grows, overcoming the acceptable threshold in all the resamplings but in no more than 20% of the validation sets.

Finally, the remaining methods are the two best performing with higher percentages of validation AUC scores above the threshold of acceptability.

Out of these, gradient boosting reaches average AUC values not very different from those seen for SVMs and random forests with very wide confidence intervals especially in resampling without repetitions, but always surpasses the acceptable cut-off up to 26.67% of acceptable validation AUCs in 3-times repeated 5×10 Nested CV.

GLM with elastic-net regularization, instead, turns out as the best method for all values shown in Table 6.3. Indeed, the mean validation AUC approaches the established threshold in all Nested

Methods	5 × 5 Nested CV		3-times repeated 5 × 5 Nested CV		5 × 10 Nested CV		3-times repeated 5 × 10 Nested CV	
	Mean AUC (95% CI)	> 0.80	Mean AUC (95% CI)	> 0.80	Mean AUC (95% CI)	> 0.80	Mean AUC (95% CI)	> 0.80
Gradient Boosting	0.7670 [0.6872,0.8468]	20.00%	0.7657 [0.7400,0.7914]	20.00%	0.7497 [0.6783,0.8211]	20.00%	0.7544 [0.7284,0.7804]	26.67%
Linear SVM	0.7603 [0.7361,0.7845]	0.00%	0.7620 [0.7456,0.7784]	6.67%	0.7605 [0.7361,0.7849]	0.00%	0.7609 [0.7448,0.7770]	6.67%
Radial SVM	0.7598 [0.7282,0.7913]	0.00%	0.7555 [0.7428,0.7682]	0.00%	0.7506 [0.6893,0.8119]	0.00%	0.7529 [0.7350,0.7708]	0.00%
Polynomial SVM	0.7562 [0.7252,0.7871]	0.00%	0.7578 [0.7423,0.7733]	6.67%	0.7487 [0.7186,0.7789]	0.00%	0.7563 [0.7423,0.7703]	0.00%
Elastic-net GLM	0.7894 [0.7085,0.8703]	20.00%	0.7882 [0.7648,0.8117]	40.00%	0.7907 [0.7114,0.8700]	20.00%	0.7922 [0.7671,0.8174]	46.67%
Random Forest	0.7471 [0.6967,0.7974]	20.00%	0.7547 [0.7357,0.7738]	13.33%	0.7464 [0.6815,0.8113]	20.00%	0.7569 [0.7362,0.7776]	13.33%
k-Nearest Neighbors	0.7070 [0.6369,0.7771]	20.00%	0.7004 [0.6702,0.7307]	13.33%	0.6961 [0.6362,0.7560]	0.00%	0.7007 [0.6729,0.7286]	6.67%
Neural Network	0.7245 [0.6991,0.7499]	0.00%	0.7308 [0.7163,0.7454]	0.00%	0.7241 [0.6868,0.7615]	0.00%	0.7293 [0.7133,0.7453]	0.00%

Table 6.3: Mean AUC score with 95% CI and percentage of validation AUCs over the acceptability threshold separately by Nested CV for all the considered methods

CVs and it is the only technique where the confidence intervals, which are very wide in sampling without repetitions, are always partially above the same threshold. This method gains a better discrimination capacity of the two classes of patients both as the number of repetitions increases and as the number of sets to be considered in the internal sampling grows, up to halving almost one out of two AUC scores above the established cut-off in the 3-times repeated Nested CV. Summarizing the previous findings, the GLM with elastic-net regularization performs far better than all the other methods and so it is the technique to be considered in the next analysis to look for the PH signature. This method is easily interpretable as its results, which are a leading factor in clinical studies as model analysis and interpretability are both important.

6.4 Model analysis

Next, one switches to determining the optimal hyper-parameters for the chosen method, reminding that the main arguments of method *glmnet* for fitting penalized GLMs with elastic-net regularization are the elastic mixing parameter *alpha* and the shrinkage parameter *lambda*.
This task is fulfilled by a 20-fold Cross-Validation on the discovery set, where the value of the folds is chosen considering the size of the dataset. First, as displayed in Figure 6.4, from an initial analysis the biggest AUC scores are achieved with $lambda = 0.01$ and *alpha* values higher than 0.50. Upon further analysis of the AUC scores obtained with these values, the optimal set of hyper-parameters achieves an AUC just below 0.92 and results in $alpha = 0.74$ and $lambda = 0.01$.

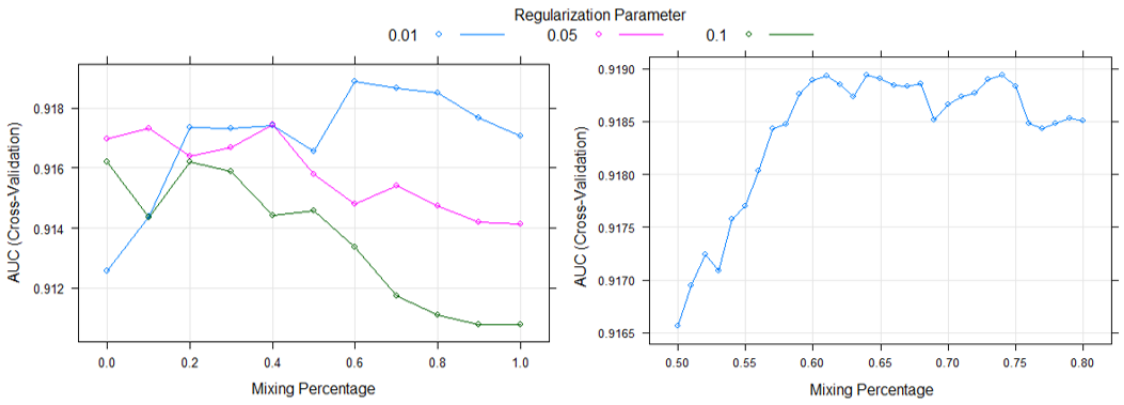


Figure 6.4: Tuning hyper-parameters by maximizing validation AUC score over a wide grid (left), then zooming in for some *alpha* values (right)

Building the ROC curve with the probability predictions of the model on the discovery set, it is

observable that this tends to approach the top-left vertex reaching an AUC score on the whole discovery set just below 0.85 and therefore the model gains a good overall ability to discriminate the distributions of cases and controls.

Now, before analyzing the model and evaluating it on the testing set in detail, it needs to find the optimal probability threshold $\mathbb{P}_{thr}$ for the classification into *DANA*'s levels, using the discovery set. The two optimal probability thresholds which maximize the J and F_β statistics are $\hat{\mathbb{P}}_{thr,J} = 0.8425$ and $\hat{\mathbb{P}}_{thr,F_\beta} = 0.5507$, respectively, as shown in Figure 6.5. These values are far

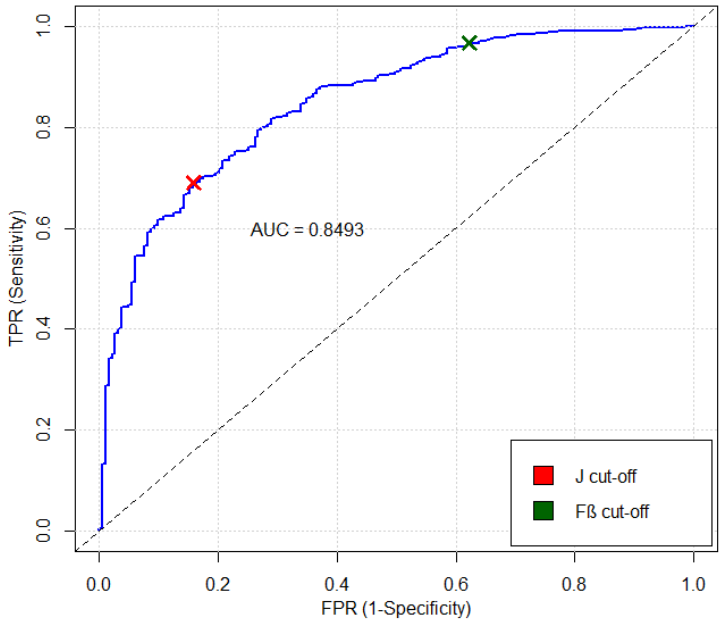


Figure 6.5: ROC curve for discovery observations, with the optimal cut-off points associated with the statistics J and F_β

away from each other and to the first one corresponds a low FPR while to the second one a high TPR. It implies different results of evaluation metrics summarized in Table 6.4.

What immediately stands out among the results of the metrics on the discovery set are the sensitivity and specificity values. With the second threshold, the model achieves a sensitivity almost close to 1 but a very low specificity, less than 0.40. With the other threshold, conversely, there is a considerable decrease in sensitivity up to 0.69 but an increase in specificity, which is around 0.84. This is also reflected in the accuracy and balanced accuracy which reach similar values with the first threshold while the balanced one is about 0.20 lower than the other in the model with the second probability cut-off. Moreover, the first model achieves almost the highest possible precision value while the second one reaches a good value even if lower than the former.

Metric	J cut-off	F_β cut-off
$\hat{\mathbb{P}}_{thr}$	0.8425	0.5507
Statistic	0.7656	0.9098
Accuracy	0.7173	0.8601
Balanced Accuracy	0.7656	0.6722
Sensitivity	0.6897	0.9673
Specificity	0.8415	0.3770
Precision	0.9515	0.8750

Table 6.4: Model evaluation statistics for the optimal probability thresholds on the discovery set

This is confirmed by confusion matrices in Figures 6.6 and 6.7: with the former, there is a high number of FNs but a low number of FPs, whereas the situation is the reverse with the latter, with a significant increase in TPs and more than a halving of TNs.

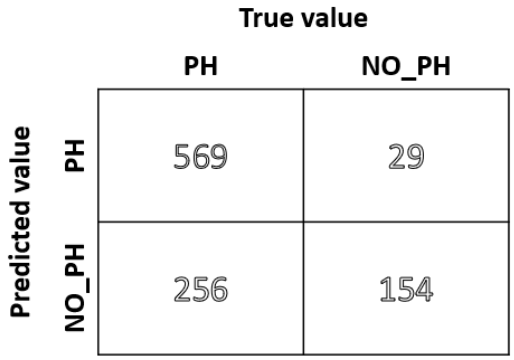


Figure 6.6: Confusion matrix for the model on the discovery set with $\hat{\mathbb{P}}_{thr,J}$

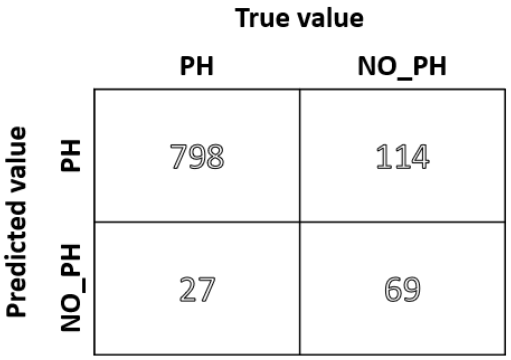


Figure 6.7: Confusion matrix for the model on the discovery set with $\hat{\mathbb{P}}_{thr,F_\beta}$

What just shown and described indicates that the same model with two different thresholds behaves markedly differently. On the one hand, the model with the first probability threshold succeeds in correctly identifying around 84% of healthy subjects but just over 2 out of 3 diseased, even if more than 95% of the predicted positive diagnoses are true positive and less than 5% are false positives. On the other hand, with the second probability threshold, the same model correctly recognizes almost all the PH subjects but less than half of the healthy subjects, and one out of eight predicted positives are false positives.

Precisely for all these findings, the choice of the optimal threshold is not very difficult, and therefore in the following analysis $\hat{\mathbb{P}}_{thr} = 0.8425$ is considered.

6.5 Biomarker signature and model evaluation

The model built so far and the optimal threshold $\mathbb{P}_{thr}$ are now considered and applied on the testing set for an overall evaluation of it. In addition, these results are compared with those of the current best non-invasive method for diagnosing PH, while the signature and the biomarkers included in it are analyzed in detail.

Reconsidering that the response variable *DANA* has the two levels PH and NO_PH, once the model is built and the vector of coefficients $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_{218})$ is estimated, it predicts for a test observation $X_j = (x_{j1}, x_{j2}, \ldots, x_{j218})$ a positive diagnosis, i.e. PH, if

$$\mathbb{P}(\widehat{DANA}_j = PH|X_j) = \frac{\exp\{\hat{\beta}_0 + \sum_{i=1}^{218} \hat{\beta}_i x_{ji}\}}{1 + \exp\{\hat{\beta}_0 + \sum_{i=1}^{218} \hat{\beta}_i x_{ji}\}} > \hat{\mathbb{P}}_{thr} = 0.8425$$

negative otherwise.

Let's start with the classification of the third test observation as an example, whose predictions are shown in the Table 6.5: the predicted PH probability is just over the optimal probability threshold and so the associated prediction is PH as the true *DANA* label.

odds$[\mathbb{P}(\widehat{DANA}_j = PH\|X_j)]$	**logit**$[\mathbb{P}(\widehat{DANA}_j = PH\|X_j)]$	$\mathbb{P}(\widehat{DANA}_j = PH\|X_j)$
1.8492	6.3539	0.8640

Table 6.5: Third test observation's predictions

Thus, the model proceeds in this way for all test observations. Since the testing set is independent of the discovery one, it is very gainful to estimate the evaluation metrics on the obtained predictions to have an overall assessment of the running of the fitted model and therefore of the signature. In summary, the results shown in Table 6.6 are similar to those previously obtained with the discovery set, with decreases due to the independence of the testing set from that of discovery. As shown also in the confusion matrix depicted in Figure 6.8, just over two out of three of the 376 test observations are correctly classified by the signature with differences in the two classes. Indeed, about 75% of the healthy subjects are correctly identified as such while 67% of the PH patients are recognized as such. Moreover, as observed in the discovery set, the

accuracy is slightly lower than that balanced. At the same time, there is a not too high number of FPs, which makes it possible to reach a precision slightly higher than 0.87. Finally, the ability to discriminate between the two classes is described by the value of the AUC score which is estimated to be around 0.81, just below that of the discovery set.

Evaluation metric	**Value**
AUC Score	0.8176
Accuracy	0.6941
Balanced Accuracy	0.7126
Sensitivity	0.6704
Specificity	0.7547
Precision	0.8744

Table 6.6: Model evaluation statistics for the optimal probability threshold on the testing set

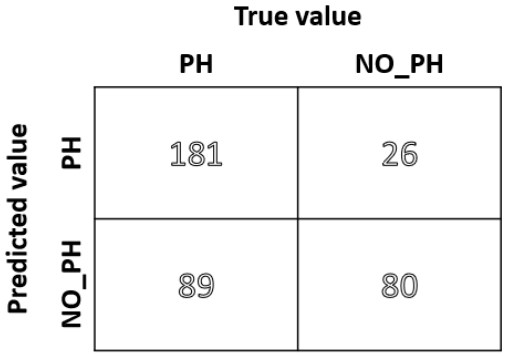


Figure 6.8: Confusion matrix on testing set with the optimal probability threshold

In other words, the signature is able to fulfill a fair distinction between the distributions of cases and controls with a greater ability to correctly identify healthy subjects. However, the high precision of the model is important as it reduces the probability of misdiagnosing PH in healthy subjects.

Given the non-excellent results of the model, it may be useful to analyze the accuracy of the model in each *DANA* sub-level as summarized in Tables 6.7 and 6.8.

With regard to healthy subjects, the accuracy is maximal for the actually healthy subjects (HC) while it is fair (68.29%) in the largest group, i.e. the group of subjects who are diseased but not

yet PH (PH0_NOPH).

	Accuracy	**N.Obs.**
PH0_NOPH	68.29%	82
HC	100.00%	24
PH1.1	82.35%	51
PH1.2	33.33%	6
PH1.3	100.00%	1
PH1.4	62.50%	40
PH2	76.19%	42
PH3	61.29%	31
PH0_CTED	27.27%	22
PH4	68.06%	72
PH5	100.00%	5

Table 6.7: Biomarker signature accuracy separately by starting levels of *DANA*

	Accuracy	**N.Obs.**
NO_PH	75.47%	106
PH1	71.43%	98
PH2	76.19%	42
PH3	61.29%	31
PH0_CTED	27.27%	22
PH4	68.06%	72
PH5	100.00%	5

Table 6.8: Biomarker signature accuracy separately by 7 sub-levels of *DANA*

Let's now turn to the PH sub-levels. First, the signature is very weak for patients with CTED with an accuracy below 30%, from which it could be suggested that this class might have been included in controls and not cases. Furthermore, different scores are achieved in groups of PAH patients: the signature is good for IPAH subjects (PH1.1) while it drops by about 20 percentage points in the PH1.4 level of people with PAH associated with other diseases. The only patient with drug-induced PAH (PH1.3) is also correctly identified as a PH but this is not statistically significant, as are those in PH1.2, due to the small size of the group. What has just been observed may be repeated for PH subjects with unclear mechanisms (PH5) since, despite an accuracy of 100%, these are only 5. On the other hand, results are statistically significant for the three remaining groups: while the signature correctly identifies about three out of four PH subjects due to left heart disease (PH2), this proportion decreases in CTEPH patients (PH4) until it reaches that of three out of five patients with PH due to lung disease and/or hypoxia correctly classified as PH.

To summarize, the built model correctly identifies a large proportion of healthy subjects and a smaller proportion of diseased ones but with different performances in the 11 sub-levels. Indeed,

it performs perfectly in completely healthy subjects and those with PH due to unclear mechanisms. Moreover, in diseased people, the signature works better in subjects with PH due to left heart disease (which is also the most common type of PH) and in PAH patients, especially in idiopathic PAH and drug or toxin induced PAH subjects.

Let's now analyze the PH signature in detail. As already said, GLMs with elastic-net regularization include variable selection by shrinkage and, being the optimized elastic-net mixing parameter equal to 0.74, the built model is a mixture between the lasso and the ridge regression. Thus, the signature includes only 41 biomarkers by setting the coefficients of the remaining 177 equal to zero. Moreover, among the non-zero coefficients, 24 are positive and 17 are negative while the estimated intercept is positive and equal to $\hat{\beta}_0 = 1.9462$. Taking up the regularization of biomarkers, it is mentioned that the majority were up-regulated, and this finding is reflected in the estimated coefficients: 23 of 24 estimated positive coefficients are associated with biomarkers that were found to be up-regulated in the discovery set, while about half of those with negative coefficients were down-regulated.

As shown in Figure 6.9, given $alpha = 0.74$, increasing the value of *lambda* rises the number of biomarkers with non-zero coefficients until the ridge regression is reached. In the case under analysis, however, less than one biomarker out of five has a non-zero coefficient resulting in $||\hat{\beta}||_1 = 4.42$.

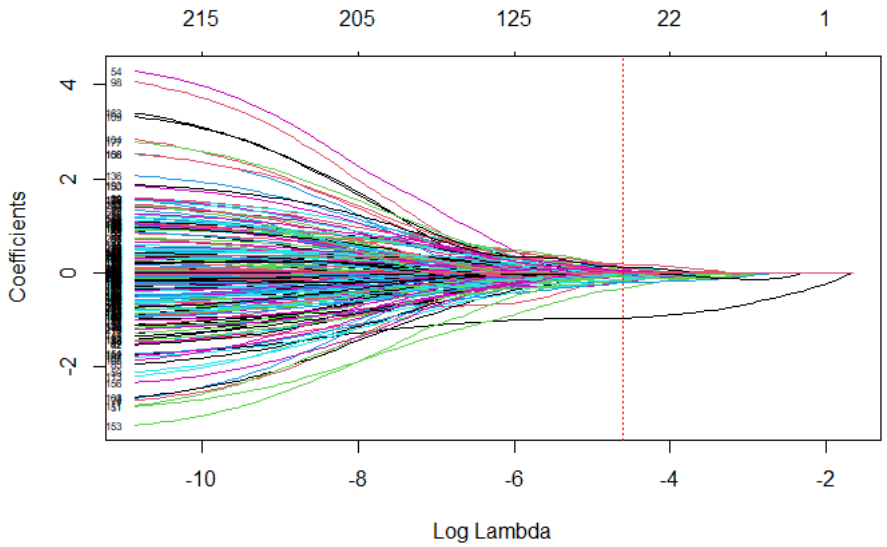


Figure 6.9: Coefficients of the model with $alpha = 0.74$ as *lambda* changes

For the optimal *lambda* value, there is an estimated parameter with a coefficient that is much

larger than the others, namely the one represented by a black line: it corresponds to *proBNP* and the coefficient is just over 0.95, about three times as large as the second largest associated to *hsa.miR.30a.5p* represented by a light green line. All other biomarkers, instead, have coefficients in the range $(0.19, +0.19)$: among biomarkers with non-zero parameters, only 3 have $|\hat{\beta}_j| < 0.01$ while 15 features are added to the two mentioned above with $|\hat{\beta}_j| > 0.01$. Biomarkers with the largest coefficient in modulus are collected in Tables 6.9 and 6.10, separately by sign.

Biomarker	**Coefficient**
proBNP	0.9573
hsa.miR.30a.5p	0.3202
hsa.miR.93.5p	0.1883
hsa.miR.140.3p	0.1875
hsa.miR.192.5p	0.1855
hsa.miR.148a.3p	0.1684
hsa.miR.18a.3p	0.1383
hsa.miR.1825	0.1321

Table 6.9: Biomarkers with the highest coefficients in the built model

Biomarker	**Coefficient**
hsa.let.7f.5p	−0.1874
hsa.miR.501.3p	−0.1436
hsa.miR.151a.5p	−0.1352
hsa.miR.374a.3p	−0.1334
hsa.miR.34a.5p	−0.1317
hsa.miR.132.3p	−0.1257
hsa.miR.142.3p	−0.1204
hsa.miR.29c.5p	−0.1125

Table 6.10: Biomarkers with the lowest coefficients in the built model

Since all predictors are quantitative variables, the coefficients can be understood as rates of change in the probability $\mathbb{P}(\widehat{DANA}_j = PH|X_j)$. Then, positive coefficients indicate rates of growth of the probability, while those negative ones indicate rates of decrease of the same one. In other words, a unit increase in the value of *proBNP* corresponds to a growth rate of $\mathbb{P}(\widehat{DANA}_j = PH|X_j)$ equal to three times that obtained with the same rise in *hsa.miR.30a.5p*. At the same time, the rate of decrease associated with the same increase in the value of *hsa.let.7f.5p* is nearly equal in modulus but of opposite sign to the growth rate observed with a unit increase in *hsa.miR.93.5p*. Furthermore, several biomarkers have an associated zero coefficient: it means that these make no contribution to the prediction of diagnosis and that therefore these biomarkers are excluded from the PH signature.

What is just described may be expressed through the variable importance, which corresponds to the absolute value of the associated estimated coefficient. It is therefore confirmed that *proBNP* is by far the most influential biomarker in the model followed by *hsa.miR.30a.5p* and then by a large group of features with influence around 0.15. Moreover, the four most influential biomarkers

all have positive coefficients: in other words, as their values increase, the probability of diagnosing PH increases.

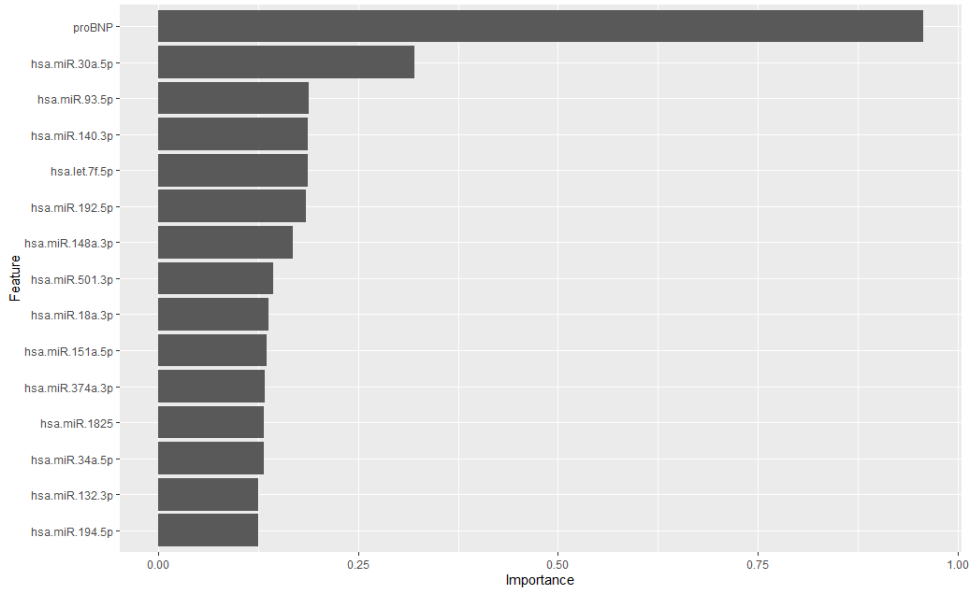


Figure 6.10: Importance plot

At this point, it would require the help of some experts to know whether biomarkers in the signature are significant: in these studies, it would be useful to have not only biostatisticians but also clinicians and biologists. Now, however, in this work it is simply seen if there is any match in the literature.

Let's start with the two biomarkers with the highest coefficients. *proBNP* is a natriuretic peptide that reflects right ventricle structure and function in PH and has been shown to be useful and accurate for diagnosis and categorization into subtypes of heart failures, with preserved and reduced ejection fraction [71]. *hsa.miR.30a.5p*, instead, is a sequence of 22 nucleotides that is proven to be a diagnostic and prognostic biomarker of certain diseases such as left ventricular dysfunction after myocardial infarction and glioma, but also a suppressor gene of various tumors (e.g. gastric cancer, renal cell carcinoma, and oral cancer) and thus of their proliferation and invasion [32, 61, 66]. Both of these are biomarkers of vascular dysfunction or myocardial stress, and as their value increases, there is a growth in the probability of a positive diagnosis.

For completeness, the feature with the greatest influence among those with negative coefficients is also analyzed. *hsa.let.7f.5p* is a miRNA sequence of 22 nucleotides that promotes bone marrow mesenchymal stem cells survival in Alzheimer's disease patients, and it serves as a diagnostic

biomarker of carcinoid tumors of the lung and as pro-angiogenic miRNA [9, 23]. *hsa.miR.501.3p*, instead, suppresses metastasis and progression of hepatocellular carcinoma and it is identified as a novel serum biomarker for Alzheimer's disease [24, 42]. Thus, for these biomarkers, as the values increase, there is a decrease in the probability of having PH.

Now, to conclude this analysis, the above results are compared with those of the best current non-invasive method for PH diagnosis.

Specifically, the signature is assumed to have true sensitivity and specificity equivalent to the summary sensitivity and specificity of a meta-analysis of echocardiogram performance, i.e. $TPR = 0.83$ and $FPR = 0.28$. Furthermore, the minimum acceptable values of the two cited statistics are set to the lower bounds of the confidence intervals of the summary sensitivity and specificity of the echocardiogram in [29], and so $TPR_0 = 0.73$ and $FPR_0 = 0.47$. Then, the TPR and FPR 2-sided 95% Wilson confidence region for sensitivity and specificity of the signature is computed and if it is entirely within the threshold of acceptability, the null hypothesis

$$H_0 : \{TPR \leq 0.73 \quad or \quad FPR \geq 0.47\}$$

is rejected and then the problem has a positive conclusion and the signature is a valid alternative to the standard non-invasive diagnostic method.

To built a 95% level Wilson confidence region for sensitivity and specificity, the 97.5% level confidence intervals of the two statistics mean values should be defined, and these in the present case turn out to be

$$Sensitivity: \ 0.6704 \ \ [0.5898, 0.7447]$$
$$Specificity: \ 0.7547 \ \ [0.6378, 0.8486]$$

So, the TPR and FPR 2-sided 95% Wilson confidence region for the signature is defined as the following Cartesian product [0.1514,0.3622] $\times$ [0.5898,0.7447] which is not entirely in the acceptability region but is only minimally so, as shown in Figure 6.11.

It means that the null hypothesis cannot be rejected and that the signature does not seem to improve the performance of the best standard non-invasive test. Thus, the obtained signature cannot be considered as a diagnostic method for PH but should be improved to be treated as a viable alternative to the best current non-invasive one.

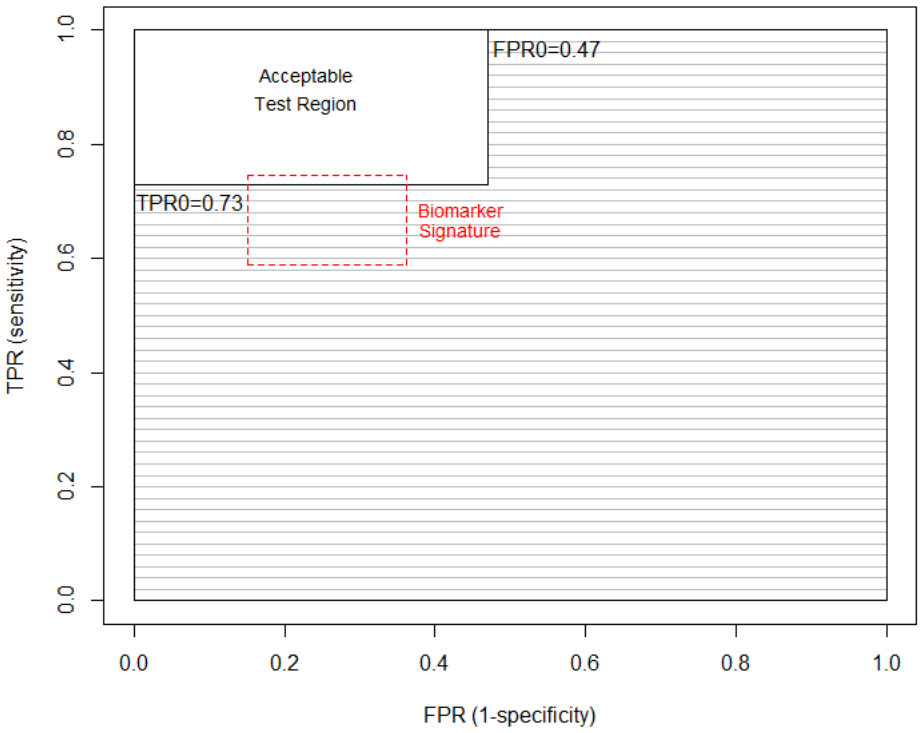


Figure 6.11: Acceptable ad unacceptable test regions, with the TPR and FPR 2-sided 95% Wilson confidence region for the signature

Chapter 7

Conclusions

Genetic expression analysis is increasingly an innovative and useful tool for the diagnosis of a wide range of disorders. Thanks to their small size, indeed, miRNAs are discovered to play an important role as biomarkers for the disease diagnosis or subtyping, with applications in different fields of medicine from oncology, to neurological diseases, to respiratory disorders, and many research proves that miRNAs are involved in the pathophysiological processes of many illnesses. At the same time, however, the discipline is still evolving and looks promising in view of developing new diagnostic tests.

Up to now, many statistical tests such as t-tests and F-tests have been employed to perform this challenge, but in the last decades, the development of innovative machine learning statistical methods has opened new horizons for the identification of disease biomarkers. In this work, the goal of developing a procedure for identifying a blood-based biomarker signature for any disease through the employment and the implementation of several statistical and sampling methods in a case-control problem is met: starting from the protocol of Janssen's CIPHER study and looking at eight different methods, first the most interpretable and best performing one in some Nested CVs on a discovery set is chosen and its hyper-parameters are optimized by a k-fold CV, and then the obtained signature is evaluated on a testing set independent of the previous one.

This strategy may be applied in several medical fields to find biomarker signatures of disorders, and in this work, pulmonary hypertension, which constitutes the sixth therapeutic area of Janssen, has been considered because its symptoms are common to several respiratory and circulatory illnesses while its diagnosis requires many diagnostic tests and a confirmatory invasive examination characterized by low morbidity and mortality. Next comes the challenge of

finding new non-invasive diagnostic approaches for faster detection such as biomarker signature of the disease to explain the several processes involved in the illness itself. This task has already been partly discussed in many scientific papers that enabled the identification of some miRNA biomarkers describing the mechanisms implicated in PH but also demonstrated the convenience of combining several biomarkers for diagnostic purposes.

Here, an attempt has been made to fulfill this requirement by looking at a dataset provided by Actelion Janssen extracted from the proof-of-concept study used to design CIPHER, consisting of discovery and testing sets with 1191 and 376 observations, respectively. First, the GLM with elastic-net regularization results as the best performing method from some Nested CVs. Then, by means of a 20-fold CV, a signature consisting of only 41 out of 555 starting biomarkers is generated: among these, *proBNP* and *hsa.miR.30a.5p* are widely the most influential, also according to the studies in the literature that indicate the role of biomarkers of processes involved in PH. The resulting model has a quite satisfactory discriminatory capacity between positive and negative diagnoses with an AUC score in the testing set just below 0.82. At the same time, the signature turns out to perform better in the identification of healthy rather than diseased subjects, with sensitivity and specificity equal to 0.6704 (97.5% Wilson CI [0.5898,0.7447]) and 0.7547 (97.5% Wilson CI [0.6378,0.8486]), respectively. These values, however, change when different subgroups are considered, also because the several types of PH present diverse causes and characteristics. On the one hand, indeed, the signature correctly detects the totality of healthy subjects, 76% of patients with PH due to left heart disease, 71% of those with PAH, and among them about 82% of those having IPAH. On the other hand, the signature recognizes less than 50% of the subjects of some sub-levels and in particular 27.27% of the patients with CTED and 33.33% of those having HPAH. Finally, the signature does not seem to improve the performance of the best current non-invasive method for the diagnosis of PH by resulting in a 95% level Wilson confidence region of TPR and FPR not wholly within the acceptability region of the test: while the specificity confidence interval is entirely above the lower bound of that of the echocardiogram, for sensitivity it only slightly exceeds the lower bound of that of the current method.

7.1 Future outlook

The work carried out in this thesis may represent a starting point to identify biomarker signatures of some illnesses, especially of those currently without cures or more lethal, and then make

early diagnoses and start treatments as soon as possible.

First of all, the proposed approach may be improved especially in the dataset analysis and pre-processing stage and in the choice of the machine learning methods to be investigated. Indeed, retrospective studies, executed to select biomarkers that could constitute the signature itself and to exclude from the analysis those that do not seem to explain the processes involved in the targeted disease, are essential for identifying accurate signatures. At the same time, if a wider variety of statistical methods are adopted, perhaps also sacrificing a bit the interpretability of the results, it would be possible to find models and signatures with greater ability to detect both healthy subjects and those diseased. Moreover, one may evaluate employing imputation techniques especially when there are several missing values for biomarkers that could be influential in the signature or were revealed as such in several scientific articles.

About the PH signature got in the previous chapter, it does not seem to perform better than the current standard non-invasive method but it could be improved through a collaboration between specialists with different expertise, such as clinicians, biologists, and biostatisticians, but also through the inclusion of other biomarkers such as those metabolic or clinical and demographic indicators that may contribute to the onset of the disease, like the age since the first symptoms occur in patients between 20 and 60 years. Additionally, because of the very low accuracy in detecting ill subjects with CTED, it might be thought to include this sub-level in healthy ones as already done for diseased subjects but not PH since, despite similar symptoms, it is not proven that this disorder evolves into CTEPH.

Furthermore, due to the different clinical presentation of the many forms of PH, as well as their different hemodynamic characteristics, causes, and symptoms, it could be helpful and strategical to find signatures to detect individual subgroups of PH or batches of them having analogous features, similar to the intention of CIPHER. This may bring, especially for patients with PAH and CTEPH, a stronger advantage and a faster treatment administration, as their symptoms usually emerge on average two years before the diagnosis is made.

To sum up, all these remarks may help to improve the developed methodology and the performance of the resulting signature, to design new ones for PH and other diseases, and thus to explore new horizons in the field of medicine and diagnostics.

R Code

This chapter collects all the functions, with input arguments and output values, defined and involved in the whole process for searching a biomarkers signature of a disease, starting with the univariate analysis and variable selection, then continuing with the method selection, the final model specification, and its evaluation on the testing set as described in Chapter 5. Subsequently, the full code run for the case study described in Chapter 6 is shown.

First, the simple function *missingValues* is defined for identifying and quantifying missing values in a given data collection and, if present, detecting the set of variables containing them.

```
# INPUT:
# - data: data collection
# - label: name of the set (i.e. "Discovery")
# OUTPUT:
# - print: number and percentage of missing values
# - indexes: indexes of variables with NAs
missingValues <- function(data, label){
  # Number (NAs) and percentage (percNAs) of missing values
  NAs <- sum(rowSums(ifelse(is.na(data),1,0)))
  percNAs <- round(NAs/(dim(data)[1]*dim(data)[2])*100,2)
  print(paste(label, " set: ", as.character(NAs), " NA (",
        as.character(percNAs), "%)", sep=""))
  if(NAs>0){
      # Percentage of missing values for each variable, sorted by
      # decreasing order
      misValues <- data.frame(perc=round(colSums(ifelse(is.na(data),
                              1,0))/dim(data)[1]*100,2))
      misValues <- data.frame(sortPerc=misValues[order(misValues$perc,
                              decreasing=TRUE),], row.names=
                              rownames(misValues)[order(misValues$perc,
                              decreasing=TRUE)])
      # Variables having missing values and their indexes
      set <- rownames(misValues)[1:sum(misValues$sortPerc>0)]
      indexes <- data.frame(name=names(data), i=seq(1,dim(data)[2]),
                            row.names=names(data))[set,2]
```

```
26        return(indexes)
27    }
28  }
```

Thus, the *ShapiroWilkTest* function consists of the first step of the univariate analysis, that is the Shapiro-Wilk test on each quantitative variable with associated Q-Q plots for those variables which result in the highest and lowest values of the test statistic.

```
1  # INPUT:
2  # - dataset: collection of data (only quantitative variables)
3  # - y: categorical variable
4  # OUTPUT:
5  # - shapRes: Shapiro-Wilk test results (W-statistic and p-value), and
6  #            values of skewness and kurtosis
7  # - print: number of significant variable for levels 0.05 and 0.01
8  # - plot: quantile-quantile plot for six variables
9  ShapiroWilkTest <- function(dataset, y){
10   shapRes <- data.frame(W=c(), pvalue=c(), skewness=c(), kurtosis=c())
11   # For each quantitative variable
12   for(i in c(1:dim(dataset)[2])){
13       variable <- names(dataset)[i]
14       # Shapiro-Wilk test
15       test <- shapiro.test(na.omit(dataset[,names(dataset)[i]]))
16       # Skewness and Kurtosis
17       skew <- skewness(na.omit(dataset[,names(dataset)[i]]))
18       kurt <-  kurtosis(na.omit(dataset[,names(dataset)[i]]))
19       shapRes <- rbind(shapRes, data.frame(W=test$statistic,
20                        pvalue=test$p.value, row.names=variable,
21                        skewness=skew, kurtosis=kurt))
22   }
23   # Sort result by p-values in decreasing order
24   shapRes <- shapRes[order(shapRes$pvalue, decreasing=TRUE),]
25   # Number of significant variables with levels 0.05 and 0.01
26   print(data.frame(signVariable=c(sum(shapRes$pvalue>0.05),
27         sum(shapRes$pvalue>0.01)), row.names=c("p>0.05","p>0.01")))
28   # q-q plot 6 features (3 with the highest W + 3 with the lowest W)
29   features <- shapRes[c(1:3, (dim(shapRes)[1]-2):dim(shapRes)[1]),]
30   par(mfrow=c(2,3))
31   for(i in c(1:6)){
32     qqnorm(miPresentData[,row.names(features)[i]], pch=1, frame=FALSE,
33            main=row.names(features)[i],
34            sub=paste("W=", as.character(round(features[i,1],4)), sep=""))
35     grid()
36     qqline(miPresentData[,row.names(features)[i]], col="steelblue", lwd=2)
37   }
38   par(mfrow=c(1,1))
39   return(shapRes)
```

```
}
```

Hence, *RankBasedRegression* follows to fit rank-based regression for each biomarker as described in Chapter 6, and analyze the associated Wald and drop in dispersion tests, whose results can be shown through *volcanoPlot* within volcano plots together with log2 fold change to identify the statistically significant variables for the test and, among them, those of interest.

```
# INPUT:
# - dataset: data collection where the first variable is binary
# - levels: levels of the categorical variable (case and control)
# OUTPUT:
# - testRes: dataframe with p-values associated with Wald and drop in
#            dispersion tests, and log2 fold change for each quantitative
#            variable
RankBasedRegression <- function(dataset, levels){
  testRes <- data.frame(Wald=c(), Dispersion=c(), logfold=c())
  # For each quantitative variable
  variable <- names(dataset[,-c(1)])
  for(i in c(1:length(variable))){
    variables <- dataset[,variable[i]]
    # Rank-based regression: variable ~ categorical variable
    fit <- rfit(variables~dataset[,1], data=dataset)
    # Means in cases and controls, and log2 fold change
    meanCase <- mean(na.omit(dataset[dataset[,1]==levels[1], variable[i]]))
    meanControl <- mean(na.omit(dataset[dataset[,1]!=levels[1],
                          variable[i]]))
    logFoldChange <- log2(abs(meanCase/meanControl))
    testRes <- rbind(testRes,
                     data.frame(Wald=summary(fit)$coefficients[2,4],
                     Dispersion=summary(fit)$droppval[1],
                     logfold=logFoldChange, row.names=variable[i]))
  }
  return(testRes)
}
```

```
# INPUT:
# - rbr: dataset with results of the RankBasedRegression function
# - pvalues: names of columns with pvalues to show ("Wald" or "Dispersion")
# - alpha: significance level of the test (default value = 0.05)
# OUTPUT:
# - p: volcano plot
# - downUp: number of down and up regulated variables, those significant
#           for the test and of interest
volcanoPlot <- function(rbr, pvalues, alpha=0.05){
  # Select data for the plot
  data <- data.frame(logfold=rbr[,"logfold"], pvalue=rbr[,pvalues],
```

```
                    row.names=rownames(rbr))
  data[,"Regulation"] <- as.factor(ifelse(data$logfold>0, "Up", "Down"))
  data[,"Variable"] <- ifelse(data$pvalue<alpha, "Significant",
                              "Not-significant")
  data[data$Variable=="Significant","Variable"] <- ifelse(abs(
                       data[data$Variable=="Significant","logfold"])>=1,
                       "Of interest","Significant")
  data$Variable <- as.factor(data$Variable)
  # Volcano plot
  p <-  ggplot(data = data, aes(x=logfold,y=-log10(pvalue)))+
          geom_point(aes(fill=Regulation, color=Variable), shape=21,
                     size=1.5) +
          scale_fill_manual(values = c("goldenrod3","chartreuse3")) +
          scale_color_manual(values = c("NA", "darkred","darkblue"))+
          xlab("log2 Fold Change")+
          ylab("-log10 p-value")+
          geom_hline(yintercept=-log10(alpha),col='darkblue',lty=2)+
          geom_vline(xintercept=0, col='black')+
          theme_bw()
  if(max(abs(data$logfold)>1)){
    p <- p + geom_vline(xintercept=-1, col='darkred',lty=2)+
                geom_vline(xintercept=+1, col='darkred',lty=2)
  }
  # down-regulated and up-regulated variables
  downReg <- c(dim(data[data$Regulation=="Down",])[1],
               sum(data[data$Regulation=="Down","Variable"]=="Significant"),
               sum(data[data$Regulation=="Down","Variable"]=="Of interest"))
  upReg <- c(dim(data[data$Regulation=="Up",])[1],
             sum(data[data$Regulation=="Up","Variable"]=="Significant"),
             sum(data[data$Regulation=="Up","Variable"]=="Of interest"))
  downUp <- data.frame(downRegulated=downReg,upRegulated=upReg,
            tot=downReg+upReg,row.names = c('All variables',paste("p<",
            as.character(alpha),sep=""),'Variables of interest'))
  return(list(p,downUp))
}
```

Furthermore, up and down regulated variables identified by the volcano plot may also be displayed through boxplots as outlined in the *boxplotVariable* function.

```
# INPUT:
# - data: collection of data (the categorical variable and the chosen
#         quantitative one)
# - posLevel: positive label of the categorical variable
# OUTPUT:
# - p: boxplot of the quantitative variable separately by the categorical
#      one
boxplotVariable <- function(data, posLevel){
```

```
  data <- na.omit(dataset)
  # Compute lower and upper whiskers
  ylimCase <- boxplot.stats(data[data[,1]==posLevel,2])$stats[c(1,5)]
  ylimControl <- boxplot.stats(data[data[,2]!=posLevel,2])$stats[c(1,5)]
  ylim <- c(min(ylimCase[1], ylimControl[1]),
            max(ylimCase[2], ylimControl[2]))
  ylim[1] <- ifelse(ylim[1]>0, ylim[1]*0.9, ylim[1]*1.1)
  ylim[2] <- ifelse(ylim[2]>0, ylim[2]*1.1, ylim[2]*0.9)
  p <- ggplot(data=data, aes(x=data[,1], y=data[,2], fill=data[,1]))+
           geom_boxplot() +
           scale_fill_manual(values=c("green3", "deepskyblue2"))+
           theme(legend.position="none") +
           xlab(names(data)[1]) + ylab(names(data)[2]) +
           coord_cartesian(ylim=ylim)
  return(p)
}
```

The univariate analysis and variable selection section ends with the pre-processing step included in the *preProcessing* function, which involves removing near-zero-variance variables, standardizing, and dropping highly-correlated features for both discovery and testing sets.

```
# INPUT:
# - discovery: discovery set (only the first variable is categorical)
# - test: testing set (only the first variable is categorical)
# - corrCut: cut-off correlation value to identify highly-correlated
#            features
# OUTPUT:
# - discovery: pre-processed discovery set
# - test: pre-processed testing set
# - print: number of near-zero-variance features
preProcessing <- function(discovery, test, corrCut){
  # Remove near-zero-variance variables
  nzv <- nearZeroVar(discovery)
  print(paste("Near-zero-variance variables:", as.character(length(nzv))))
  if(length(nzv)>0){
    discovery <- discovery[,-c(nzv)]
    test <- test[,-nzv]
  }
  # Scale and center
  standardization <- preProcess(discovery[,-c(1)], method=c("center",
                                "scale"))
  discovery <- predict(standardization, discovery)
  test <- predict(standardization, test)
  # Correlation analysis
  corrmat <- cor(discovery[,-c(1)])
  var <- c()
  for(i in c(2:(dim(corrmat)[1]-1))){
```

```
    for(j in c(1:(i-1))){
      # Pair of variables with correlation higher than the cut-off value
      if(corrmat[i,j]>corrCut){
        # Select the variable with the highest mean correlation
        var <- c(var, ifelse(mean(corrmat[,row.names(corrmat)[i]])>
                  mean(corrmat[,row.names(corrmat)[j]]), i+1, j+1))
      }
    }
  }
  # Remove highly-correlated variables
  discovery <- discovery[,-var]
  test <- test[,-var]
  return(list(discovery, test))
}
```

Once completed the first step, proceeds with the function *NestedCV* for implementing 3-times repeated $5 \times B$ Nested Cross-Validations on the discovery set and choosing the best-performing method. Specifically, as stated earlier, a double *for* loop was chosen in the function rather than the *repeatedcv* method to see results step by step. Moreover, since not all the chosen techniques require the same hyper-parameters, the *if-else* construct was selected to include several classifiers, even if it is not very computationally convenient.

```
# INPUT:
# - dataset: discovery set (DANA is the response variable)
# - method: method to fit
# - grid: grid of hyper-parameter values
# - ninnerfolds: folds inner CV
# - times: number of repetitions (default value = 3)
# OUTPUT:
# - print: validation AUC scores and t-test without and with 3 repetitions
NestedCV <- function(dataset, method, grid, ninnerfolds, rep=3){
  # Outer partitions
  evaluationSets <- list()
  for(i in c(1:rep)){
    set.seed(824+i)
    folds <- createFolds(dataset$DANA, k=5)
    evaluationSets[[i]] <- list(folds$Fold1, folds$Fold2, folds$Fold3,
                                folds$Fold4, folds$Fold5)
  }
  # rep-times 5Xninnerfolds Nested Cross-Validation
  AUCs <- c()
  for(j in c(1:rep)){
    # Outer Cross-Validation
    for(i in c(1:5)){
      # Split discovery set in training and testing ones
```

```
24        training <- dataset[-evaluationSets[[j]][[i]],]
25        validation <- dataset[evaluationSets[[j]][[i]],]
26        # Inner Cross-Validation
27        set.seed(825)
28        innerCV <- trainControl(method="cv", number=ninnerfolds,
29                                classProbs=TRUE, savePredictions=TRUE,
30                                summaryFunction=prSummary)
31        # Methods:
32        # - k-Nearest Neighbors: method="knn"
33        # - Gradient Boosting: method="gbm"
34        # - Elastic-net GLM: method="glmnet", family="binomial"
35        # - Random Forest: method="ranger"
36        # - SVMs: method=c("svmLinear", "svmRadial", "svmPoly")
37        # - Neural Network: method="nnet"
38        if(method=="ranger"){
39          # Random Forest
40          methodFit <- train(DANA~., data=training, method=method,
41                             importance="permutation",
42                             trControl=innerCV, verbose=FALSE,
43                             tuneGrid=grid, metric="AUC", maximize=TRUE)
44        }else if(method=="nnet" || method=="knn"){
45          # Neural Network and k-Nearest Neighbors
46          methodFit <- train(DANA~., data=training, method=method,
47                             trControl=innerCV, tuneGrid=grid, metric="AUC",
48                             maximize=TRUE)
49        }else{
50          # Gradient Boosting, GLM with Elastic-net regularization, and
51          # Support Vector Machines
52          methodFit <- train(DANA~., data=training, method=method,
53                             trControl=innerCV, verbose=FALSE,
54                             tuneGrid=grid, metric="AUC", maximize=TRUE)
55        }
56        # AUC score
57        auc <- roc(predictor=predict(methodFit, newdata=validation,
58                   type="prob")[,1], response=validation$DANA,
59                   positive=levels(validation$DANA)[1])$auc
60        print(paste("Rep.", as.character(j), "VALIDATION SET n.",
61              as.character(i), "- Outer Validation AUC:",
62              as.character(round(auc,4))))
63        AUCs <- c(AUCs, auc)
64      }
65    }
66    # t-test without and with repetitions
67    print("t-test (no repetitions):")
68    print(t.test(AUCs[1:5]))
69    print(paste("t-test (", as.character(rep)," repetitions)", sep=""))
```

```
  print(t.test(AUCs))
}
```

Then, the *kCV* function is defined to choose the optimal hyper-parameters and fit the best GLM with elastic-net regularization model through a k-fold Cross-Validation on the discovery set.

```
# INPUT:
# - dataset: discovery set
# - grid: grid of hyper-parameters
# - nfolds: k (default value = 20)
# OUTPUT:
# - p: plot AUCs for each set of values
# - methodFit: fitted model
# - print: optimal hyper-parameters
kCV <- function(dataset, grid, nfolds=20){
  # nfolds-fold CV
  set.seed(825)
  CV <- trainControl(method="cv", number=nfolds, classProbs=TRUE,
                     savePredictions=TRUE, summaryFunction=prSummary)
  methodFit <- train(x=dataset[,-c(1)], y=dataset[,1], method="glmnet",
                     family="binomial", trControl=CV, verbose=FALSE,
                     tuneGrid=grid, metric="AUC", maximize=TRUE,
                     case=levels(dataset[,1])[1],
                     control=levels(dataset[,1])[2])
  # Plot and best hyper-parameters
  p <- plot(methodFit)
  print("Best hyper-parameters")
  print(methodFit$bestTune)
  return(list(p, methodFit))
}
```

Once the best model is built, function *cutoffs* is defined to identify the probability thresholds that maximize the J and F_β statistics by fitting the chosen model on the discovery set, which are then shown on the ROC curve constructed with the same collection of data by implementing the *cutROC* function.

```
# INPUT:
# - data: discovery set
# - methodFit: fitted model
# - var: response variable
# - levels: levels of categorical variable
# OUTPUT:
# - Jmax: threshold with the highest J-statistic, and associated
          evaluation metrics
# - Fbetamax: threshold with the highest F-measure, and associated
               evaluation metrics
```

```
# - plot: plot of J and Fbeta statistics
cutoffs <- function(data, methodFit, var, levels){
  # Evaluation metrics for each probability threshold
  roc_test <- roc(data[,var], predict(methodFit, newdata=data,
                  type="prob")[,1], levels=levels, plot=FALSE)
  prevalence <- table(data[,var])[1]/(table(data[,var])[1]+
                table(data[,var])[2])
  actualCase <- table(data[,var])[1]
  actualControl <- table(data[,var])[2]
  tp <- roc_test$sensitivities*actualCase
  tn <- roc_test$specificities*actualControl
  fp <- actualControl-tn
  precision <- tp/(tp+fp)
  cutoff <- data.frame(threshold=roc_test$thresholds,
                       sensitivity=roc_test$sensitivities,
                       specificity=roc_test$specificities,
                       precision=precision)
  cutoff <- cutoff[-c(1, dim(cutoff)[1]),]
  # J-statistic
  cutoff[,"J"] <- (cutoff$sensitivity+cutoff$specificity)/2
  # Fbeta-statistic
  cutoff[,"Fbeta"] <- ((1+prevalence^2)*cutoff$sensitivity*
                      cutoff$precision)/((prevalence^2)*cutoff$precision+
                      cutoff$sensitivity)
  # Maximization of the statistics and plot
  Jmax <- cutoff[cutoff$J==max(cutoff$J),]
  Fbetamax <- cutoff[cutoff$Fbeta==max(cutoff$Fbeta),]
  plotJ <- ggplot(data=cutoff, aes(x=threshold, y=J))+
              geom_line(col="blue")+
              geom_point(data=cutoff[cutoff$J==max(cutoff$J),],
                         col="red", cex=2.5)+
              ggtitle("J-statistic")
  plotF <- ggplot(data=cutoff, aes(x=threshold, y=Fbeta))+
              geom_line(col="blue")+
              geom_point(data=cutoff[cutoff$Fbeta==max(cutoff$Fbeta),],
              col="red", cex=2.5)+
              ggtitle("Fbeta-statistic")
  pl <- plot_grid(plotJ, plotF)
  return(list(Jmax, Fbetamax, pl))
}
```

```
# INPUT:
# - data: set of data
# - Jmax: evaluation metric for the threshold with the highest
#         J-statistic value
# - Fbetamax: evaluation metric for the threshold with the highest
#             F-measure
```

```
# - methodFit: fitted model
# - var: response variable
# - levels: levels of response variable
# OUTPUT:
# - p: ROC curve and optimal cut-off points
cutROC <- function(dataset,Jmax,Fbetamax,methodFit,var='DANA',
                   levels=c('NO_PH','PH')){
   # Predicted probabilities
  roc_test <- roc(dataset[,var], predict(methodFit, newdata=dataset,
                  type="prob")[,1], levels=levels, plot=FALSE)
  par(mfrow=c(1,1))
  pred_final <- prediction(predict(methodFit, newdata=dataset,
                        type="prob")[,1], ifelse(dataset$DANA=="PH",1,0))
  perf_final <- performance(pred_final,"tpr","fpr")

  # ROC curve
  p <- plot(perf_final@x.values[[1]], perf_final@y.values[[1]],
            col='blue', type="l", xlim=c(0,1), ylim=c(0,1), lwd=2,
            xlab='FPR (1-Specificity)', ylab='TPR (Sensitivity)',
            main='ROC curve and cut-off points')
  p <- p + grid()
  p <- p + abline(a=0, b=1, col='black', lty=2)

  # Optimal probability thresholds
  p <- p + points(1-Jmax$specificity, Jmax$sensitivity, pch=4,
                  col='red', lwd=3, cex=1.5)
  p <- p + points(1-Fbetamax$specificity, Fbetamax$sensitivity,
                  pch=4, col='darkgreen', lwd=3, cex=1.5)
  p <- p + legend(legend=c('J cut-off', 'F   cut-off'),
                  fill=c('red','darkgreen'),'bottomright',inset=0.01)
  p <- p + text(x=0.35, y=0.6, paste('AUC = ',
                as.character(round(roc_test$auc[1],4)),sep=''))
  return(p)
}
```

Then, the fitted model with a given probability threshold can be evaluated by the function *confMatrix* on a set of observations through confusion matrix and evaluation metrics such as specificity, sensitivity, and AUC score.

```
# INPUT:
# - data: dataset
# - thr: threshold probability
# - methodFit: fitted method
# - var: categorical variable
# - levels: levels of the response variable
# OUTPUT:
# - confusionMat: confusion matrix and evaluation metrics
```

```
# - aucScore: AUC score
confMatrix <- function(data, thr, methodFit, var, levels){
  # Predictions
  pred <- as.factor(ifelse(predict(methodFit, newdata=data,
                   type="prob")[,1]>thr, levels[2],levels[1]))
  # Confusion matrix and evaluation metrics
  confusionMat <- confusionMatrix(relevel(pred,levels[2]), data[,var])
  # AUC score
  aucScore <- roc(predictor=predict(methodFit, newdata=data,
                  type="prob")[,1], response=data[,var],
                  levels=rev(levels(as.factor(data[,var]))),
                  positive=levels[2])$auc
  return(list(aucScore, confusionMat))
}
```

By employing the properties defined in Chapter 5 of the GLM with elastic-net regularization, the function *GLMNETexample* allows one to manually compute the logit, odds, and probability $P(\hat{Y}_j = 1|X_j)$ for each test observation and compare the latter with the same predicted by the fitted model.

```
# INPUT:
# - dataset: test data
# - methodFit: the fitted model
# - posLevels: label of positive diagnosis
# OUTPUT:
# - example: logit, odds, Pcomputed and Ppredicted P(Y=1) for each test
#            observation
GLMNETexample <- function(dataset, methodFit, posLevel){
  # Extract intercept and coefficients
  intercept <- coef(methodFit$finalModel, methodFit$bestTune$lambda)@x[1]
  coefs <- coef(methodFit$finalModel, methodFit$bestTune$lambda)@x[-1]
  if(levels(dataset[,1])[1]==posLevel){
    intercept <- -intercept
    coefs <- -coefs
  }
  coefsindex <- coef(methodFit$finalModel,
                     methodFit$bestTune$lambda)@i[-1]+1
  coefs <- data.frame(coefs, row.names=colnames(dataset[,coefsindex]))
  # logit, odds and probability
  logitTest <- as.matrix(dataset[,coefsindex])%*%as.matrix(coefs)+
               as.matrix(data.frame((intercept*rep(1,dim(dataset)[1])),
               row.names=rownames(dataset)))
  oddsTest <- exp(logitTest)
  PTrain <- oddsTest/(1+oddsTest)
  Ppred <- predict(methodFit, newdata=dataset, type = "prob")[,posLevel]
  example <- data.frame(logit=logitTest, odds=oddsTest,
```

```
                          P(Y=1)_computed=PTrain, P(Y=1)_predicted=Ppred)
  return(example)
}
```

Thus, the *groupAnalysis* function evaluates the built model through the accuracy in each sub-levels of the response variable *DANA*. Specifically, the function considers the categorical variable indicating the PH diagnosis, as depicted in the case study in Chapter 6.

```
# INPUT:
# - datasetStart: starting testing set (DANA with 11 levels)
# - datasetMod: preprocessed testing set (DANA with 2 levels)
# - pred: predictions for testing set
# OUTPUT:
# - accuracy11: accuracy separately by 11 levels
# - accuracy7: accuracy separately by 7 levels
groupAnalysis <- function(datasetStart, datasetMod, pred){
  # Correctly classified or not
  diagnosis <- data.frame(Diagnosis=datasetStart$DANA,
                          Truth=datasetMod$DANA, Prediction=pred)
  diagnosis[,"Correct"] <- ifelse(diagnosis$Truth==diagnosis$Prediction,
                                  1,0)
  # 11 levels: PH0_CTED, PH_NOPH, PH1.1, PH1.2, PH1.3, PH1.4, PH2, PH3,
  #            PH4, PH5, HC
  res <- aggregate(x=diagnosis$Correct, by=list(diagnosis$Diagnosis),
                   FUN=mean)$x*100
  names(res) <- c("Diagn", "Accuracy")
  accuracy11 <- cbind(res, data.frame(count=summary(diagnosis$Diagnosis)))
  # 7 levels: PH0_CTED, NO_PH, PH1, PH2, PH3, PH4, PH5
  PHdiagn <- diagnosis
  levels(PHdiagn$Diagnosis)<- c("PH0_CTED","NO_PH","PH1","PH1","PH1",
                                "PH1","PH2","PH3","PH4","PH5","NO_PH")
  res2 <- aggregate(x=PHdiagn$Correct, by=list(PHdiagn$Diagnosis),
                    FUN=mean)$x*100
  names(res2) <- c("Diagn", "Accuracy")
  accuracy7 <- cbind(res2, data.frame(count=summary(PHdiagn$Diagnosis)))
  return(list(accuracy11, accuracy7))
}
```

So, the *coeffGLMNET* and *importancePlot* functions are useful to analyze the identified signature and thus the estimated coefficients for the several features.

```
# INPUT:
# - methodFit: fitted model
# - levels: levels of the response variable
# - posLevel: label of the positive class
# OUTPUT:
# - print: numbers of positive/negative coefficients and some of them
```

```
# - coefficients: estimated coefficients
# - p: plot of coefficients as lambda changes
coeffGLMNET <- function(methodFit, levels, posLevel){
  # Extract coefficients for P(Y=1|X)
  if(levels[2]==posLevel){
    coeff <- coef(methodFit$finalModel, methodFit$bestTune$lambda)
  }else{
    coeff <- -coef(methodFit$finalModel, methodFit$bestTune$lambda)
  }
  coefficients <- data.frame("Coeff"=coeff[is.na(coeff)==0],
                             row.names=rownames(coeff))
  coefficients["Importance"] <- abs(coefficients$Coeff)
  coefficients <- coefficients[order(coefficients$Importance,
                                     decreasing=TRUE),]
  # Split into positive and negative coefficients
  posCoeff <- coefficients[coefficients$Coeff>0,]
  posCoeff <- posCoeff[order(posCoeff$Coeff, decreasing=TRUE),]
  negCoeff <- coefficients[coefficients$Coeff<0,]
  negCoeff <- negCoeff[order(negCoeff$Coeff, decreasing=FALSE),]
  print(paste("Positive coefficients:", as.character(dim(posCoeff)[1])))
  print(head(posCoeff, 10))
  print(paste("Negative coefficients:", as.character(dim(negCoeff)[1])))
  print(head(negCoeff, 10))
  print(paste("Zero coeffients:", as.character(dim(coefficients)[1]-
        dim(posCoeff)[1]-dim(negCoeff)[1])))
  # Coefficients plot
  p <- plot(methodFit$finalModel, lwd=1.5, label=TRUE, "lambda")
  p <- p+abline(v=log(methodFit$bestTune$lambda), lwd=1, lty=3, col="red")
  return(list(coefficients, p))
}
```

```
# INPUT:
# - methodFit: fitted model
# - top: number of top variables to show (default value = 15)
# OUTPUT:
# - varPlot: variable importance plot
importancePlot <- function(methodFit, top=15){
  varimp <- varImp(methodFit, scale=FALSE)
  varPlot <- ggplot(varimp, top=top)
  return(varPlot)
}
```

Finally, the comparison of the built signature performance with the current best non-invasive diagnostic method is made by the construction of the 2-sided TPR and FPR Wilson confidence region with *plotRegion* and, before that, the identification of the confidence intervals of sensitivity and specificity through *WilsonInterval*.

```
# INPUT:
# - stat: statistic
# - level: confidence level
# - n: number of observations
# OUTPUT:
# - LB: lower bound CI
# - UB: upper bound CI
WilsonInterval <- function(stat, level, n){
  # Quantile
  q <- qnorm(level+(1-level)/2)
  # Upper and lower bounds
  UB <- (stat+q^2/(2*n)+q*sqrt(stat(1-stat)/n+q^2/(4*n^2)))/(1+q^2/n)
  LB <- (stat+q^2/(2*n)-q*sqrt(stat(1-stat)/n+q^2/(4*n^2)))/(1+q^2/n)
  return(c(LB,UB))
}
```

```
# INPUT:
# - sens: sensitivity CI (1st statistic)
# - spec: specificity CI (2nd statistic)
# - TPR0: TPR for the current best non-invasive test
# - FPR0: FPR for the current best non-invasive test
# OUTPUT:
# - p: plot confidence region and comparison with the current best
#      non-invasive test
plotRegion <- function(sens, spec, TPR0, FPR0){
  # Best non-invasive diagnostic test
  p <- plot(c(0,1), c(0,1), type="n", xlab="FPR (1-specificity)",
            ylab="TPR (sensitivity)")
  for(i in seq(0.02, floor(TPR0*100/2)*0.02, 0.02)){
    p <- p + segments(0, i, 1, i, col="grey")
  }
  for(i in seq(ceiling(TPR0*100/2)*0.02, 0.98, 0.02)){
    p <- p + segments(FPR0, i, 1, i, col="grey")
  }
  p <- p + segments(0, 0, 1, 0)
  p <- p + segments(0, 0, 0, 1)
  p <- p + segments(1, 0, 1, 1)
  p <- p + segments(0, 1, 1, 1)
  p <- p + segments(0, TPR0, FPR0, TPR0)
  p <- p + segments(FPR0, TPR0, FPR0, 1)
  # Wilson Confidence Region of the signature
  p <- p + segments(1-spec[2], sens[1], 1-spec[1], sens[1], col="red",
                    lty="dashed")
  p <- p + segments(1-spec[2], sens[1], 1-spec[2], sens[2], col="red",
                    lty="dashed")
  p <- p + segments(1-spec[1], sens[1], 1-spec[1], sens[2], col="red",
                    lty="dashed")
```

```
  p <- p + segments(1-spec[2], sens[2], 1-spec[1], sens[2], col="red",
                    lty="dashed")
  p <- p + text(x=0.075, y=TPR0-0.03,
                labels=paste("TPR0=",as.character(TPR0), sep=""))
  p <- p + text(x=FPR0+0.08, y=0.97,
                labels=paste("FPR0=",as.character(FPR0), sep=""))
  p <- p + text(x=floor(FPR0*100/2)*0.01,
                y=(TPR0+ceiling((1-TPR0)*100/2)*0.01+0.05),
                labels="Acceptable", cex=0.75)
  p <- p + text(x=floor(FPR0*100/2)*0.01,
                y=(TPR0+ceiling((1-TPR0)*100/2)*0.01),
                labels="Test Region", cex=0.75)
  p <- p + text(x=(1-specWilson[1])*1.2,
                y=sensWilson[2]-(sensWilson[2]-sensWilson[1])/2*0.75,
                labels="Biomarker", cex=0.90, col="red")
  p <- p + text(x=(1-specWilson[1])*1.2,
                y=sensWilson[2]-(sensWilson[2]-sensWilson[1])/2*1.25,
                labels="Signature", cex=0.90, col="red")
  return(p)
}
```

All these functions are employed to search for a PH signature with biomarkers as described by the following lines of code, where *TRAINING* and *VALIDATION* are the discovery and testing sets provided by the company, respectively. Results are gathered and explained in Chapter 6.

```
#### #### Univariate analysis and variable selection #### ####
miPresentData <- TRAINING       # Discovery set
dim(miPresentData)
names(miPresentData)
miFeaturesData <- VALIDATION    # Testing set
dim(miFeaturesData)
# Transformation of DANA into a binary variable
levels(miPresentData$DANA)<- c("PH", "NO_PH", "PH", "PH", "PH", "PH",
                               "PH", "PH", "PH", "PH", "NO_PH")
summary(miPresentData$DANA)
levels(miFeaturesData$DANA)<- c("PH", "NO_PH", "PH", "PH", "PH", "PH",
                                "PH", "PH", "PH", "PH", "NO_PH")
summary(miFeaturesData$DANA)
# Drop of not useful variables
miPresentData <- miPresentData[,-c(1:4, 6:38, 40:42)]
miFeaturesData <- miFeaturesData[,-c(1:4, 6:38, 40:42)]
# Missing values analysis
setDiscovery <- missingValues(data=miPresentData, label="Discovery")
setTesting <- missingValues(data=miFeaturesData, label="Testing")
# Remove variables with NA values in testing set
miPresentData <- miPresentData[,-setTesting]
miFeaturesData <- miFeaturesData[,-setTesting]
```

```
dim(miPresentData)
dim(miFeaturesData)
# Omit missing values in discovery set
miPresentData <- na.omit(miPresentData)
summary(miPresentData$DANA)
# Log-transformation of proBNP
miPresentData$proBNP <- log(miPresentData$proBNP)
miFeaturesData$proBNP <- log(miFeaturesData$proBNP)
# Shapiro-Wilk test
shapiroResults <- ShapiroWilkTest(dataset=miPresentData[,-c(1)],
                                  y=miPresentData$DANA)
head(shapiroResults)
# Rank-based regression and associated tests
rankBasedRes <- RankBasedRegression(dataset=miPresentData,
                                    levels=c("PH","NO_PH"))
head(rankBasedRes)
# Volcano plots
volcanoPlot(rbr=rankBasedRes, pvalues="Dispersion")
volcanoPlot(rbr=rankBasedRes, pvalues="Dispersion", alpha = 0.01)
# Boxplots: up and down regulated biomarkers
up1 <- boxplotVariable(data=miPresentData[,c("DANA","proBNP")],
                       posLevel="PH")
up2 <- boxplotVariable(data=miPresentData[,c("DANA","hsa.miR.101.5p")],
                       posLevel="PH")
down1 <- boxplotVariable(data=miPresentData[,c("DANA","hsa.let.7f.5p")],
                         posLevel="PH")
down2 <- boxplotVariable(data=miPresentData[,c("DANA","hsa.miR.125a.5p")],
                         posLevel="PH")
plot_grid(up1, up2, down1, down2, nrow=1)
# Pre-processing
preProc <- preProcessing(discovery=miPresentData, test=miFeaturesData,
                         corrCut=0.94)
miPresentData <- preProc[[1]]
miFeaturesData <- preProc[[2]]
#### #### Method selection #### ####
# Gradient Boosting
gbmGrid <- expand.grid(n.trees=c(70,100,140,150), interaction.depth=(2:4),
                       shrinkage=c(0.1,0.01), n.minobsinnode=10)
NestedCV(dataset=miPresentData, method='gbm', grid=gbmGrid, ninnerfolds=5)
NestedCV(dataset=miPresentData, method='gbm', grid=gbmGrid,
         ninnerfolds=10)
# Linear SVM
svmLinGrid <- expand.grid(C=c(0.001,0.005,0.01))
NestedCV(dataset=miPresentData, method='svmLinear', grid=svmLinGrid,
         ninnerfolds=5)
NestedCV(dataset=miPresentData, method='svmLinear', grid=svmLinGrid,
```

```
69          ninnerfolds=10)
70  # Radial SVM
71  svmRadialGrid <- expand.grid(sigma=c(0.001,0.005,0.01),
72                               C=c(0.001,0.005,0.01,0.1))
73  NestedCV(dataset=miPresentData, method='svmRadial', grid=svmRadialGrid,
74          ninnerfolds=5)
75  NestedCV(dataset=miPresentData, method='svmRadial', grid=svmRadialGrid,
76          ninnerfolds=10)
77  # Polynomial SVM
78  svmPolyGrid <- expand.grid(C=c(0.001,0.005,0.01,0.1), degree=c(2,3),
79                             scale=c(0.001,0.05,0.01))
80  NestedCV(dataset=miPresentData, method='svmPoly', grid=svmPolyGrid,
81          ninnerfolds=5)
82  NestedCV(dataset=miPresentData, method='svmPoly', grid=svmPolyGrid,
83          ninnerfolds=10)
84  # Elastic-net GLM
85  glmnetGrid <- expand.grid(alpha=c(0,0.1,0.25,0.5,0.75,0.9,1),
86                            lambda=c(0.1,0.01))
87  NestedCV(dataset=miPresentData, method='glmnet', grid=glmnetGrid,
88          ninnerfolds=5)
89  NestedCV(dataset=miPresentData, method='glmnet', grid=glmnetGrid,
90          ninnerfolds=10)
91  # Random Forest
92  rfGrid <- expand.grid(mtry=c(10,20,40), splitrule=c('gini','extratrees'),
93                        min.node.size=c(1,3))
94  NestedCV(dataset=miPresentData, method='ranger', grid=rfGrid,
95          ninnerfolds=5)
96  NestedCV(dataset=miPresentData, method='ranger', grid=rfGrid,
97          ninnerfolds=10)
98  # k-NN
99  knnGrid <- expand.grid(k=c(46,51))
100 NestedCV(dataset=miPresentData, method='knn', grid=knnGrid,
101         ninnerfolds=5)
102 NestedCV(dataset=miPresentData, method='knn', grid=knnGrid,
103         ninnerfolds=10)
104 # Neural Network
105 nnetGrid <- expand.grid(decay=0.2, size=c(3,5))
106 NestedCV(dataset=miPresentData, method='nnet', grid=nnetGrid,
107         ninnerfolds=5)
108 NestedCV(dataset=miPresentData, method='nnet', grid=nnetGrid,
109         ninnerfolds=10)
110 #### #### Model analysis #### ####
111 glmnetGrid <- expand.grid(alpha=seq(0,1,0.1), lambda=c(0.1,0.05,0.01))
112 methodFit <- kCV(dataset=miPresentData, grid=glmnetGrid)
113 # Zoom for searching the optimal hyper-parameters
114 glmnetGrid <- expand.grid(alpha=seq(0.5,0.8,0.01), lambda=0.01)
```

```
methodFit <- kCV(dataset=miPresentData, grid=glmnetGrid)
methodFit <- methodFit[[2]]
methodFit$results
# Probability cut-off points
cutoff <- cutoffs(data=miPresentData, methodFit, var="DANA",
                  levels=c("NO_PH","PH"))
Jmax <- cutoff[[1]]
Fbetamax <- cutoff[[2]]
# ROC curve with the two optimal probability thresholds
cutROC(data=miPresentData, Jmax, Fbetamax, methodFit, var="DANA",
       levels=c("NO_PH","PH")
# Evaluation of the two optimal cut-off points on the discovery set:
# confusion matrix, evaluation metrics, and AUC score
confMatrix(data=miPresentData, thr=Jmax$threshold, methodFit,
           var="DANA", levels=c("NO_PH","PH"))
confMatrix(data=miPresentData, thr=Fbetamax$threshold, methodFit,
           var="DANA", levels=c("NO_PH","PH"))
#### #### Biomarker signature and model evaluation #### ####
# Test examples
testExample <- GLMNETexample(dataset=miFeaturesData, methodFit,
                             posLevel="PH")
head(testExample)
# Evaluation on testing set: confusion metrics, evaluation metrics, and
# AUC score
confMatrix(data=miFeaturesData, thr=Jmax$threshold, methodFit, var="DANA",
           levels=c("NO_PH","PH"))
# Accuracy on testing observations, separately by subgroup
predJtest <- as.factor(ifelse(predict(methodFit, newdata=miFeaturesData,
                        type="prob")[,1]>Jmax$threshold, "PH", "NO_PH"))
acc <- groupAnalysis(datasetStart=VALIDATION, datasetMod=miFeaturesData,
                     pred=predJtest)
accuracy11 <- acc[[1]]
accuracy7 <- acc[[2]]
# Model coefficients
coeff <- coeffGLMNET(methodFit, levels=levels(miPresentData$DANA),
                     posLevel="PH")
importancePlot(methodFit)
# 97.5% Wilson confidence intervals for sensitivity and specificity
sensWilson <- WilsonInterval(stat=sensitivity(predJtest,
                             miFeaturesData$DANA), level=0.975, n=summary(
                             miFeaturesData$DANA)['PH'][[1]])
specWilson <- WilsonInterval(stat=specificity(predJtest,
                             miFeaturesData$DANA), level=0.975, n=summary(
                             miFeaturesData$DANA)['NO_PH'][[1]])
# 95% Wilson confidence region
plotRegion(sens=sensWilson, spec=specWilson, TPR0=0.73, FPR0=0.47)
```

Bibliography

[1] Actelion Pharmaceutical Ltd, a Janssen Pharmaceutical Company of Johnson&Johnson, A Prospective Multicenter Study for the Identification of Biomarker Signatures for early detection of Pulmonary Hypertension (PH), Protocol NAPUH0001: Cipher Phase 0, 2020, March 25.

[2] Actelion Pharmaceutical Ltd, a Janssen Pharmaceutical Company of Johnson&Johnson, Statistical Analysis Plan for CSR of Protocol AC-065A404. 2020, February 27.

[3] Agresti A., *Foundations of linear and generalized linear models*, John Wiley & Sons, Inc., Hoboken, New Jersey, 2015.

[4] Benjamini Y., Hochberg Y., Controlling the false discovery rate: a practical and powerful approach to multiple test, *Journal of the Royal Statistical Society: Series B (Methodological)*, 1995; 57(1): 289-300.

[5] Citrix Receiver, Available: https://www.citrix.com/

[6] Cogan J.D. et al, High frequency of BMPR2 exonic deletions/duplications in familial pulmonary arterial hypertension, *American Journal of Respiratory and Critical Care Medicine*, 2006, September 1; 174(5): 590-598.

[7] Columbus Ohio Adult Congenital Heart Disease Program at Nationwide Children's Hospital Heart Center, Columbus, Ohio, Pulmonary Hypertension, Available: https://www.nationwidechildrens.org/conditions/pulmonary-hypertension (Accessed: April 2021)

[8] Data School, ROC Curves and Area Under the Curve (AUC) Explained, YouTube, Available: https://www.youtube.com/watch?v=OAl6eAyP-yo (Accessed: April 2021)

[9] Di Fazio P., Maass M., Roth S. et al, Expression of hsa-let-7b-5p, hsa-let-7f-5p, and hsa-miR-222-3p and their putative targets HMGA2 and CDKN1B in typical and atypical carcinoid tumors of the lung, *Tumour Biology*, 2017, October; 39(10): 1-10.

[10] Internation Diabetes Federation, IDF Diabetes Atlas 9th edition 2019, Available: https://diabetesatlas.org/en/ (Accessed: April 2021)

[11] Fares W.H., Pandit K.V., Kaminski N., Novel mechanisms of disease: Network biology and microRNA signaling in pulmonary hypertension, In: Maron B., Zamanian R., Waxman A. (eds) *Pulmonary Hypertension*, Springer, Cham, 2016.

[12] Feldman S., You Should Probably Be Doing Nested Cross-Validation | PyData Miami 2019, YouTube, Available: https://www.youtube.com/watch?v=DuDtXtKNpZs, 2019, June 17.

[13] Ferlay J., Colombet M., Soerjomataram I., Dyba T., Randi G., Bettio M. et al, Cancer incidence and mortality patterns in Europe: Estimates for 40 countries and 25 major cancers in 2018, *European Journal of Cancer*, 2018, November; 103: 356-387.

[14] Fontana R., Bonino D., Appunti del Corso di Laurea in Pianificazione Territoriale, Urbanistica e Paesaggistico-Ambientale, Politecnico di Torino, A.A. 2015/2016.

[15] Freedson P.S., Melason E., Sirard J., Calibration of the Computer Science and Applications, Inc. accelerometer, *Medicine & Science in Sports & Exercise*, 1998, May; 30(5): 777-781.

[16] Fromm B., Billipp T., Peck L.E., Johansen M., Tarven J.E., King B.L. et al, A Uniform System for the Annotation of Vertebrate microRNA Genes and the Evolution of the Human microRNAome, *Annual Review of Genetics*, 2015, October 14; 49: 213-242.

[17] Furey T., Cristianini N., Duffy N., Bednarski D., Schummer M., Haussler D., Support vector machines classification and validation of cancer tissue samples using microarray expression data, *Bioinformatics*, 2000, October; 16(10): 906-914.

[18] Galié N., Humbert M., Vachiery J.L., Gibbs S., Lang I., Torbicki A. et al, ESR/ECR Guidelines for the diagnosis and treatment of pulmonary hypertension, *European Respiratory Journal*, 2015; 46: 903-75.

[19] Gasparini M., *Modelli probabilistici e statistici*, CLUT, 2006.

[20] Getz T.R., Levine E., Domany D., Coupled two-way clustering analysis of gene microarray data, *Proceedings of the National Academy of Sciences*, 2000, October 24; 97(22): 12079-12084.

[21] Goldberg A.B., Mazur W., Karla D.K., Pulmonary hypertension: diagnosis, imaging techniques and novel therapies, *Cardiovascular Diagnosis & Therapy*, 2017, August; 7(4): 405-417.

[22] Griffiths-Jones S., Grocock R.J., van Dongen S., et al, miRBase: microRNA sequences, targets and gene nomenclature, *Nucleid Acids Research*, 2006, January 1; 34(1 Suppl): D140-D144.

[23] Han L., Zhou Y., Zhang R. et al, MicroRNA Let-7f-5p Promotes Bone Marrow Mesenchymal Stem Cells Survival by Targeting Caspase-3 in Alzheimer Disease Model, *Frontiers in Neuroscience*, 2018, May 22; 12: 333.

[24] Hara N., Kikuchi M., Miyashita M. et al, Serum microRNA miR-501-3p as a potential biomarker related to the progression of Alzheimer's disease, *Acta Neuropathologica Communications*, 2017, January 31; 5(1): 10.

[25] Herve P., Lau E., Sitbon O. et al, Criteria for diagnosis of exercise pulmonary hypertension, *European Respiratory Journal*, 2015, May 28; 46: 728-737.

[26] Hoeper M.M., Bogaard H.J., Condliffe R. et al, Definitions and diagnosis of pulmonary hypertension, *Journal of the American College of Cardiology*, 2013, December 24; 62(25 Suppl): D42-D50.

[27] Hurdman J., Condliffe R. et al, Pulmonary Hypertension in COPD: Results from the ASPIRE registry, *European Respiratory Journal*, 2013 June; 41(6): 1292-1301.

[28] James G., Witten D., Hastie T., Tibshiran R., *An Introduction to Statistical Learning with Applications in R*, Springer, 2014.

[29] Janda S., Shahidi N., Gin K., Swiston J., Diagnostic accuracy of echocardiography for pulmonary hypertension: a systematic review and meta-analysis, *Heart*, 2011, April; 97(8): 612-622.

[30] Janssen, About us and our therapy areas, Available: https://www.janssen.com/emea/our-company/about-janssen (Accessed: April 2021)

[31] Janssen-Cilag SpA, Janssen Medical Cloud^{TM}, Ipertensione polmonare, Available: https://www.janssenmedicalcloud.it/ipertensione-polmonare-pagina-pubblica (Accessed: April 2021)

[32] Jia Z., Wang K., Wang G. et al, MiR-30a-5p antisense oligonucleotide suppresses glioma cell growth by targeting SEPT7, *PLoS One*, 2013; 8(1): e55008.

[33] Johnson&Johnson, 2020 Annual Report, Available: https://www.investor.jnj.com/annual-meeting-materials/2020-annual-report (Accessed: April 2021)

[34] Johnson&Johnson, Pharmaceutical Products, Available: https://www.jnj.com/healthcare-products/prescription (Accessed: April 2021)

[35] Khun M., The caret package, Available: https://www.topepo.github.io/caret (Accessed: April 2021)

[36] Kim N.H., Delcroix M., Jais X. et al, Chronic Thromboembolic Pulmonary Hypertension, *European Respiratory Journal*, 2019, January 24; 53(1).

[37] Kloke J.D., McKean J.W., Rfit: Rank-based estimation for linear models, *The R Journal*, 2012, December; 4(2): 57-64.

[38] Koster A., Shiroma E.J., Caserotti P., Matthews C.E., Chen N.Y., Glynn N.W., Harris T.B., Comparison of sedentary estimated between actiPAL and hip and wrist-worn ActiGraph, *Medicine & Science in Sports & Exercise*, 2016, August; 48(8): 1514-1522.

[39] Kovacs G., Berghold A., Scheidi S. et al, Pulmonary artery pressure during rest and exercise in healthy subjects: a systematic review, *European Respiratory Journal*, 2009, October; 34(4): 888-894.

[40] Kuyuk S., Ercan I., Commonly used statistical method for detecting differential gene expression in microarray experiments, *Biostatistics and Epidemiology International Journal*, 2017, December; 0(0): 1-8.

[41] Li X., Li X., *Regularization Achieving on Optimized and Balanced Solution*, Janssen Pharmaceutical Companies of Johnson&Johnson, 2020, October 15.

[42] Luo C., Yin D., Zhan H. et al, microRNA-501-3p suppresses metastasis and progression of hepatocellular carcinoma through targeting LIN7A, *Cell Death & Disease*, 2018, May 10; 9: 535.

[43] Marciejak A., Kostarska-Srokosz E., Gierlak W. et al, Circulating miR-30a-5p as a prognostic biomarker of left ventricular dysfunction after acute myocardial infarction, *Scientific Reports*, 2018, June 29; 8: 9883.

[44] McGlinchey N., Johnson M.K., Novel serum biomarkers in pulmonary arterial hypertension, *Biomarkers in Medicine*, 2014; 8(8): 1001-1011.

[45] Miao C., Chang J., Zhang G., Recent research progress of microRNA in hypertension pathogenesis, with a focus on the roles of miRNAs in pulmonary arterial hypertension, *Molecular biology reports*, 2018, December; 45(6): 2883-2896.

[46] Miller J., Accelerometer Technologies, Specifications and Limitations, ActiGraph, *International Conference on Ambulatory Monitoring of Physical Activity and Movement (ICAMPAM)*, 2013, June 17.

[47] Moler E.J., Chow M.L., Mian I.S., Analysis of molecular profile data using generative and discriminative methods, *Physiological Genomics*, 2000, December 18; 4(2): 109-126.

[48] Newman J.H., Wheeler L., Lane K.B. et al, Mutation in the gene for bone morphogenetic protein receptor II as a cause of primary pulmonary hypertension in large kindred, *The New England Journal of Medicine*, 2001, August 2; 345(5): 319-324.

[49] Peacock A.J., Murphy N.F., McMurray J.J.V. et al, An epidemiological study of pulmonary

arterial hypertension, *European respiratory Journal*, 2007, July; 30(1): 104-109.

[50] Pulmonary Hypertension Association, About Pulmonary Hypertension, Available: https://phassociation.org/patients/aboutph/ (Accessed: April 2021)

[51] Pulmonary Hypertension Association, About Pulmonary Hypertension: The WHO Groups, Available: https://phassociation.org/types-pulmonary-hypertension-groups/ (Accessed: April 2021)

[52] Pulmonary Hypertension Association, Diagnosing Pulmonary Hypertension, Available: https://phassociation.org/patients/diagnosis/ (Accessed: April 2021)

[53] Putri D., De Troyer E., Classification: Constrained Regression, A short guide on how to penalize your regression when it is not properly behaving, Janssen Pharmaceutical Companies of Johnson&Johnson, 2020, September 17.

[54] RStudio, Open source & professional software for data science teams, Available: https://www.rstudio.com/

[55] SAS® OnDemand for Academics, SAS Italy Analytics Software & Solutions, Available: https://www.sas.com/it_it/software/on-demand-for-academics.html

[56] Shai S.S., Shai B.D., *Understanding Machine Learning: From Theory to Algorithms*, Cambridge University Press, 2014.

[57] Simonneau G., Galié N., Rubin L.J. et al, Clinical classification of pulmonary hypertension, *Journal of the American College of Cardiology*, 2004, June 16; 43(12 Suppl S): S5-S12.

[58] Simonneau G., Gatzoulis M., Adatia I. et al, Updated clinical classification of pulmonary hypertension, *Journal of the American College of Cardiology.* 2013, December 24; 62(25 Suppl): D34-D41.

[59] Simonneau G., Montani D., Celermajer D.S., Denton C.P., Gatzoulis M.A., Kwowka M. et al, Haemodynamic definitions and updated clinical classification of pulmonary hypertension, *European Respiratory Journal*, 2019, January 24; 53(1).

[60] Starmer J., StatQuest with John Starmer, YouTube, Available: https://www.youtube.com/user/joshstarmer (Accessed: November 2020)

[61] Sun Y., Yang B., Lin M. et al, Identification of serum miR-30a-5p as a diagnostic and prognostic biomarker in colorectal cancer, *Cancer Biomarkers*, 2019, January 1; 24(3): 299-305.

[62] Terpstra J.T., McKean J.W., Rank-Based Analysis of Linear Models Using R, *Journal of Statistical Software*, 2005, July; 14(7).

[63] Troiano R.P., Berrigad D., Dodd K.W., Masse L.C., Tilert T., McDowell M., Physical

activity in the United States measured by accelerometer, *Medicine & Science in Sports & Exercise*, 2008, January; 40(1): 181-188.

[64] Tudor-Locke C., Barreira T.V., Schuna J.M., Mire E.F., Katzmarzyk P.T., Fully automated wrist-worn accelerometer algorithm for detecting children's sleep-period time separate from 24-h physical activity or sedentary behaviors, *Applied Physiology, Nutrition and Metabolism*, 2014, January; 39(1): 53-57.

[65] Vachiery J.L., Adir Y., Barbera J.A. et al, Pulmonary hypertension due to left heart disease, *Journal of the American College of Cardiology*, 2013, December 24; 62(25 Suppl): D100-D108.

[66] Wang C., Cai L., Liu J. et al, MicroRNA-30a-5p inhibits the Growth of Renal Cell Carcinoma by Modulating GRP78 Expression, *Cellular Physiology and Biochemistry*, 2017; 43(6): 2405-2419.

[67] World Cancer Research Fund, Worldwide cancer data: Global cancer statistics for the most common cancers, Available: https://www.wcrf.org/dietandcancer/cancer-trends/worldwide-cancer-data (Accessed: April 2021)

[68] Wikipedia, L'enciclopedia libera, Receiver operating characteristic, Available: https://en.wikipedia.org/wiki/Receiver_operating_characteristic (Accessed: April 2021)

[69] World Health Organisation, The top 10 causes of death, 2020, December 9, Available: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death (Accessed: April 2021)

[70] World Health Organisation, The World Health Report 2001: Mental disorders affect one in four people, 2001, September 9, Available: https://www.who.int/whr/2001/media_centre/press_release/en/ (Accessed: April 2021)

[71] Wong L.L., Zou R., Xhou L. et al, Combining microRNA and NT-proBNP to detect and categorize heart failure subtypes, *Journal of the American College of Cardiology*, 2019, March 26; 73(11): 1300-1313.

[72] Xiong M., Fang X., Zhao J., Biomarker Identification by Feature Wrappers, *Genome Research*, 2001, November; 11(11): 1878-1887.