The University Of Sheffield.

Machine learning approaches for uncovering miRNAs as biomarkers of Pulmonary Hypertension

By:

Niamh Errington

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

The University of Sheffield

Faculty of Medicine, Dentistry and Health

Department of Neuroscience

Submission Date: October 2022

Abstract

Pulmonary hypertension (PH) covers a broad spectrum of diseases with a variety of pathobiological mechanisms, phenotypes and aetiologies. The current clinical classification is based on invasive haemodynamics and disease aetiology categorised by 5 groups, including two treatable subtypes; pulmonary arterial hypertension (PAH) and chronic thromboembolic pulmonary hypertension (CTEPH). Using microRNAs, small, non-coding molecules of RNA previously shown to be dysregulated in PH, we investigate the molecular classification of PH patients through machine learning models.

Initially, we applied four supervised machine learning methods to microRNA expression profiles to distinguish between 64 patients with PAH and 43 disease and healthy controls. Twenty microRNAs were identified as putative biomarkers by consensus from all four methods and examined the targets of individual microRNAs. We identified two consensus microRNAs (miR-636 and miR-187-5p) which predict PAH diagnosis with high accuracy (AUC 0.78 and 0.80 respectively).

We then applied these methods to a larger cohort of 1150 PH patients and 334 disease controls, developing microRNA panels of nine miRNAs to distinguish both between PH subtypes, and PH from disease controls. These panels compared favourably with the current standard clinical biomarker, N-terminal pro-brain natriuretic peptide (NT-proBNP) in detecting PH in a disease cohort, and outperformed NT-proBNP in identifying PAH and CTEPH from other forms of PH. A microRNA signature for PAH, validated in the independent cohort, appeared in patients diagnosed with PH-left heart disease (PH-LHD) and PH-lung, suggesting overlapping pathology or misclassification. Unsupervised learning of microRNAs assigned to a mixed cohort of PAH, PH-LHD and PH-lung patients identified six distinct molecular clusters that displayed differences in survival, haemodynamics, NT-proBNP and 6-minute walking distance, as well as different molecular pathway perturbation.

Circulating microRNAs offer greater insight into the heterogeneity of PH than clinical phenotyping alone and may have potential in diagnosis and better targeting of treatments.

Acknowledgements

This thesis would not exist without my two supervisors - Dennis Wang and Allan Lawrie. I couldn't have asked for more enthusiastic, approachable, and helpful supervisors. I would never have discovered this world (or stuck at it) without your guidance and support, and I will be eternally grateful. Thank you also to the rest of the Wang and Lawrie labs, in particular Mark and Sokratis for your input and encouragement.

A special thanks to my parents, for opening the doors to the world of science and providing unending support. I couldn't mention Mum and Dad without mentioning my big sister Sinead (not forgetting Laurence and Aidan). You've always been my biggest cheerleaders and I am deeply grateful to all of you.

To my Sheffield family; especially Tom, Michael and Charlotte who always pick me up when I'm down. Whether it's dinner, bikes, dogs or a hug, you're always there when I need you.

Last, but by no means least, Mike. You have supported me in every way you know how, thank you for everything you do. I couldn't have asked for better lockdown partners than you and Rocky.

My sincere appreciation to all the unnamed patients who gave the gift of data.

Publications and manuscripts

Christopher J. Rhodes*, Pablo Otero-Núñez*, John Wharton, Emilia M. Swietlik, Sokratis Kariotis, Lars Harbaum, Mark J. Dunning, Jason M. Elinoff, **Niamh Errington**, A. A. Roger Thompson, James Iremonger, J. Gerry Coghlan, Paul A. Corris, Luke S. Howard, David G. Kiely, Colin Church, Joanna Pepke-Zaba, Mark Toshner, Stephen J. Wort, Ankit A. Desai, Marc Humbert, William C. Nichols, Laura Southgate, David-Alexandre Trégouët , Richard C. Trembath, Inga Prokopenko, Stefan Gräf , Nicholas W. Morrell, Dennis Wang, Allan Lawrie, and Martin R. Wilkins (2020) Whole-Blood RNA Profiles Associated with Pulmonary Arterial Hypertension and Clinical Outcome. Am J Respir Crit Care Medicine

Niamh Errington*, James Iremonger*, Josephine A. Pickworth, Sokratis Kariotis, Christopher J. Rhodes, Alexander MK Rothman, Robin Condliffe, Charles A. Elliot, David G. Kiely, Luke S. Howard, John Wharton, A. A. Roger Thompson, Nicholas W Morrell, Martin R. Wilkins, Dennis Wang#, and Allan Lawrie# (2021) A diagnostic miRNA signature for pulmonary arterial hypertension using a consensus machine learning approach. EbioMedicine

Sokratis Kariotis*, Emmanuel Jammeh*, Emilia M Swietlik*, Josephine A. Pickworth, Christopher J Rhodes , Pablo Otero , John Wharton, James Iremonger , Mark J. Dunning , Divya Pandya, Thomas S Mascarenhas, **Niamh Errington**, A. A. Roger Thompson , Casey E. Romanoski , Franz Rischard , Joe G.N. Garcia, Jason X.-J. Yuan, Tae-Hwi Schwantes An, Ankit A. Desai, Gerry Coghlan, Jim Lordan, Prof Paul Corris, Luke S Howard, Robin A. Condliffe, Prof David G. Kiely, Colin Church, Joanna Pepke-Zaba, Mark Toshner, Stephen Wort, Stefan Gräf, Prof Nicholas W Morrell, Prof Martin R Wilkins , Prof Allan Lawrie#, Dennis Wang# (2021) Biological heterogeneity in idiopathic pulmonary arterial hypertension identified through unsupervised transcriptomic profiling of whole blood. Nature communications

Niamh Errington, Sokratis Kariotis, Chris Rhodes, Emmanuel Jammeh, Yiu-Lian Fong, Zhou Lihan, Cheng He, Timothy Jatkoe, Tatiana Vener, John Wharton, Roger Thompson, Robin Condliffe, David Kiely, Mark Toshner, Luke Howard, Eileen Harder, Aaron Waxman, Dennis Wang*, Allan Lawrie* Martin R Wilkins* (2021) Diagnostic miRNA signatures for treatable forms of pulmonary hypertension highlight challenges with clinical classification. *Manuscript in preparation*

List of Figures

List of tables

List of abbreviations

6MWT	6-minute walk test
AI	Artificial Intelligence
AIC	Akaike Information Criterion
ANN	Artificial neural networks
APAH	Associated Pulmonary Arterial Hypertension
AUC	Area Under the Curve
BNP	Brain natriuretic peptide
BWH	Brigham and Women's Hospital
cDNA	Complementary DNA
CI	Cardiac Index
CO	Cardiac Output
CT	High-resolution computed tomography
CTED	Chronic thromboembolic disease
CTEPH	Chronic thromboembolic pulmonary hypertension
DC	Disease Control
DNA	Deoxyribosenucleic acid
ECG	Electrocardiogram
eGFR	Estimated glomerular filtration rate
Echo	Echocardiogram
EPH10	EmPHasis-10
ESC	European society of Cardiology
ERA	Endothelin receptor antagonists
ERK	Extracellular signal-regulated kinase
ERS	European Respiratory Society
ESWT	Endurance shuttle walk test
FC	Functional class
FN	False negative

FP	False positive
FVC	Forced vital capacity
FVCP	Predicted forced vital capacity
GLM	Generalised Linear Model
HC	Healthy control
HPAH	Heritable Pulmonary Arterial Hypertension
HR	Hazard Ratio
IPAH	Idiopathic Pulmonary Arterial Hypertension
ISWT	Incremental shuttle walk test
KNN	K-Nearest Neighbour
LASSO	Least Absolute Shrinkage and Selection Operator
LDL	Low-density lipoprotein
LR	Logistic regression
LVEF	Left ventricular ejection fraction
MI	Multiple imputation
MICE	Multiple imputation using multivariate imputation by chained equations
miRNA	microRNA
mRNA	Messenger RNA
ML	Machine learning
mPAP	Mean pulmonary arterial pressure
mRAP	Mean right atrial pressure
NGS	next-generation sequencing
NPV	Negative predicted value
NT-proBNP	N-Terminal Pro-Brain Natriuretic Peptide
OLS	Ordinary least squares
PAH	Pulmonary Arterial Hypertension
PAWP	Pulmonary arterial wedge pressure
PCA	Principal component analysis

PDE5	Phosphodiesterase type 5 inhibitors
PEA	Pulmonary endarterectomy
PH	Pulmonary Hypertension
PPV	Positive predicted value
PSA	Prostate-specific antigen
PVOD	Pulmonary veno-occlusive disease
PVR	Pulmonary vascular resistance
qPCR	Quantitative-Polymerase chain reaction
REVEAL	Registry to Evaluate Early and Long-Term PAH Disease Management
RFE	Recursive feature elimination
RHC	right heart catheterisation
RNA	Ribonucleic acid
ROC	Receiver operator characteristic
STAT3	Signal transducer and activator of transcription 3
SSc	Systemic sclerosis
SVM	Support Vector Machine
TAVI	Transcatheter aortic valve implantation
TLCO	Transfer factor for carbon monoxide
TN	True negative
TP	True positive
t-SNE	t-Distributed Stochastic Neighbour Embedding
WHO	World Health Organisation
WU	Wood unit

TABLE OF CONTENTS

Chapter 1: Introduction

1.1 Pulmonary Hypertension

Pulmonary Hypertension (PH) is a cardiopulmonary condition, characterised by a high mean pulmonary arterial pressure (defined as ≥20 mm Hg as evaluated by right heart catheterization), leading to failure of the right ventricle and premature death (Simonneau et al. 2019). Since 2004, the World Health Organisation (WHO) has categorised PH into five subgroups, each with their own subcategories (Table 1.1); these remain the current clinical classification guidelines (Simonneau et al. 2019).

Pulmonary arterial hypertension (PAH) is a subgroup of patients with PH, characterised hemodynamically by the incidence of pre-capillary PH. PAH has a multifactorial pathobiology including increased pulmonary vascular resistance caused by vasoconstriction and thrombosis, endothelial cell dysfunction and vascular cell proliferation. PAH can be further sub-categorised into four groups: Idiopathic PAH (IPAH), heritable PAH (HPAH), drug and toxin induced, and PAH associated with other systemic diseases (APAH) (Simonneau et al. 2019). The pathogenesis of PAH involves the interplay of a predisposed genetic background, and epigenetic state. Regardless of the cause, PH reduces life expectancy and impacts quality of life.

Table 1.1*: Updated clinical classification of pulmonary hypertension (PH). Adapted from (Simonneau et al. 2019)*

1 Pulmonary Arterial Hypertension (PAH)
- 1.1 Idiopathic PAH
- 1.2 Heritable PAH
- 1.3 Drug and toxin induced PAH
- 1.4 PAH associated with:
  - 1.4.1 Connective tissue disease
  - 1.4.2 HIV infection
  - 1.4.3 Portal hypertension
  - 1.4.4 Congenital heart disease
  - 1.4.5 Schistosomiasis
- 1.5 PAH long-term responders to calcium channel blockers
- 1.6 PAH with overt features of venous/ capillaries (PVOD/PCH) involvement
- 1.7 Persistent PH of the newborn syndrome

2 PH due to left heart disease (PH-LHD)
- 2.1 PH due to heart failure with preserved LVEF
- 2.2 PH due to heart failure with reduced LVEF
- 2.3 Valvular heart disease
- 2.4 Congenital/acquired cardiovascular conditions leading to post capillary PH

3 PH due to lung disease and / or hypoxia (PH-lung)
- 3.1 Obstructive lung disease
- 3.2 Restrictive lung disease

3.3 Other lung disease with mixed restrictive/obstructive pattern
3.4 Hypoxia without lung disease
3.5 Developmental lung disease
4 PH due to pulmonary artery obstruction
4.1 Chronic thromboembolic PH
4.2 Other pulmonary artery obstructions
5 PH with unclear and/or multifactorial mechanisms
5.1 Haematological disorders
5.2 Systemic and metabolic disorders
5.3 Others
5.4 Complex congenital heart disease

1.1.1 Diagnosis of PH

A range of tests can be used to indicate a diagnosis of PH, including electrocardiogram (ECG), chest X-rays, echocardiogram ('echo'), lung function tests, ventilation-perfusion scanning, high-resolution computed tomography (CT), MRI scanning or pulmonary angiography. However, right heart catheterisation (RHC) is the 'gold standard' for diagnosing PH and is essential to conclusively diagnose PAH (Rosenkranz and Preston 2015). A RHC is an invasive procedure, involving a special catheter being guided to the right side of the heart, then passed into the pulmonary artery, with pressure measurements taken along the way (Table 1.2).

Table 1.2: *Haemodynamic definitions of pulmonary hypertension.*

Pre / Post capillary PH	**Classification**	**mPAP**	**PVR**	**PAWP**
Non-PH	Non-PH	≤ 20 mm Hg		
Pre-capillary PH	PAH	> 20 mm Hg	≥ 3 WU	≤ 15 mm Hg
	PH-lung	> 20 mm Hg	≥ 3 WU	≤ 15 mm Hg
	CTEPH	> 20 mm Hg	≥ 3 WU	≤ 15 mm Hg
	Miscellaneous PH	> 20 mm Hg	≥ 3 WU	≤ 15 mm Hg
Isolated post-capillary PH	PH-LHD	> 20 mm Hg	< 3 WU	> 15 mm Hg
	Miscellaneous PH	> 20 mm Hg	< 3 WU	> 15 mm Hg
Combined pre- and post-capillary PH	PH-LHD	> 20 mm Hg	≥ 3 WU	> 15 mm Hg
	Miscellaneous PH	> 20 mm Hg	≥ 3 WU	> 15 mm Hg

A recent study of PH patients in the UK found that for 48% of patients, it took over a year from the onset of symptoms to a diagnosis (Armstrong et al. 2019), a result in keeping with previous studies and seen worldwide. The typical journey a patient may go on can be seen in Figure 1.1, and highlights the roundabout patients can find themselves on along the way to diagnosis, driven in part by the generic symptoms. Considering for example, the predominant symptom of PAH is dyspnea on exertion. As PH is a life-limiting disease which deteriorates over time, the faster a patient can be diagnosed and treatment started, the better the disease outcome.

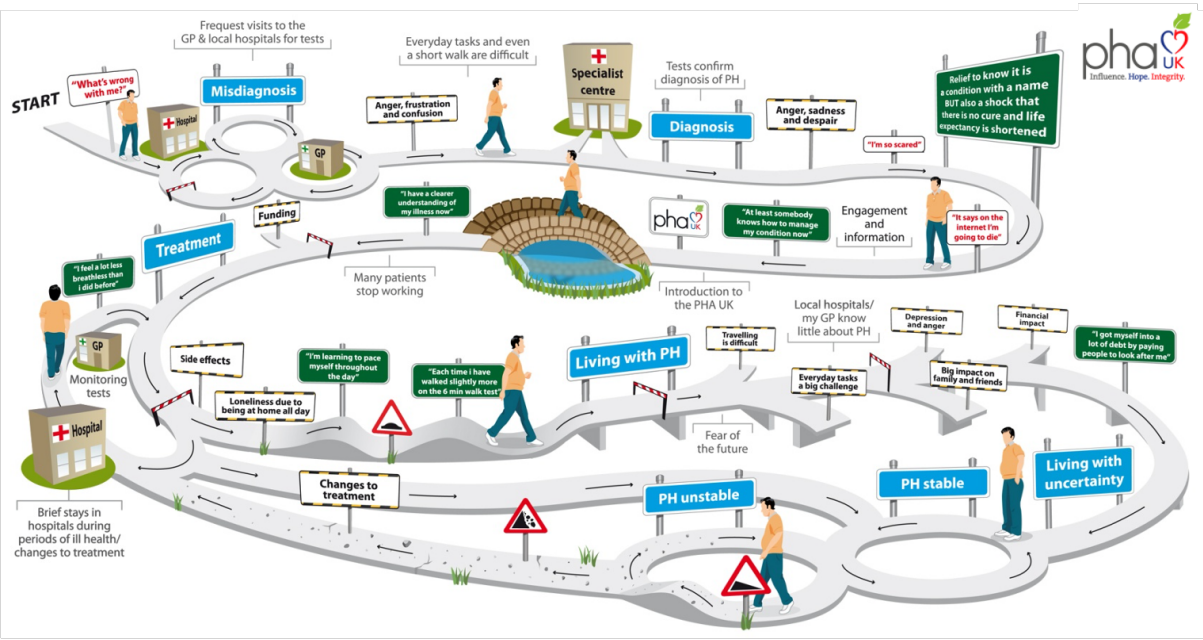


***Figure 1.1**: The clinical journey of a patient with PH, taken from symptom onset through to living with the disease. Figure produced by PHA UK.*

1.1.1.1 Risk assessment and progression monitoring in PH

There are a range of tools available to measure a patient's quality of life with PH, and to gauge how a patient's symptoms affect their day-to-day life. Two such examples are the WHO functional class, and the EmPHasis-10 questionnaire. EmPHasis-10 (EPH-10) questionnaire is a tool developed to track how PH is affecting patient's lives by measuring health-related quality of life (Lewis et al. 2021).The WHO functional classes describe the severity of a patient's PH symptoms. There are four classes:

- Class I: Symptom free when physically active or resting
- Class II: No symptoms at rest, but normal activities may cause discomfort or shortness of breath
- Class III: Patient may be symptom free at rest, but normal chores around the house are greatly limited due to shortness of breath or tiredness
- Class IV: Symptomatic at rest, and severe symptoms with activity

Walking tests are also used to evaluate functional exercise capacity and assess prognosis. There are three commonly used exercises, the 6 minute walk test (6MWT), incremental shuttle walk test (ISWT) and endurance shuttle walk test (ESWT), with all three found to be valid and reliable (Singh et al. 2014). In the 6MWT, patients walk as far as they can in 6 minutes along a flat corridor, recording the distance in metres. In the ISWT, the participant must walk faster, at a rate controlled by pre-recorded signals. The test continues until the participant cannot

keep up or can no longer continue, with a maximum duration of 20 mins. The ESWT is derived from the ISWT, where patients walk as long as possible at a predetermined rate based on performance in the ISWT. All three test measurements reflect treatment and rehabilitation of patients.

The European Society of Cardiology (ESC) and the European Respiratory Society (ERS) recommend the frequent use of risk assessment tools in PAH to inform treatment decisions and potentially improve morbidity and mortality. The Registry to Evaluate Early and Long-Term PAH Disease Management (REVEAL) risk calculator is a commonly used tool. The REVEAL risk score has been shown to predict survival outcomes in PAH populations, as well as offering sequential assessments. The recent version of this calculator, REVEAL 2.0, groups patients into one of three risk groups: low, intermediate and high-risk. The REVEAL group takes account of 12 variables; PAH subgroup, age and gender, comorbidities, WHO functional class, vital signs, all-cause hospitalizations in the past 6 months, 6-minute walk test distance, BNP, echocardiogram, pulmonary function test, and RHC findings (Benza et al. 2019)

1.1.2 Treatment in PH

Correctly diagnosing the sub-category of PH is especially important for patients with PAH or CTEPH as there are targeted treatments available for patients with PAH, and a potentially curable surgery as the gold standard of treatment for CTEPH patients. Treatments for PH-LHD and PH-lung, where pulmonary hypertension is a secondary condition, consist of treating the underlying disease. There is no standard treatment for Misc PH.

Aside from transplantation, there is no cure for PAH. However, a few medicines have been approved to ease symptoms. These can be classed into prostacyclin analogues (Epoprostenol, Iloprost, Treprostinil, Beraprost), endothelin receptor antagonists (ERAs; Bosentan, Macitentan, Ambrisentan), phosphodiesterase type 5 inhibitors (PDE5; Sildenafil, Tadalafil), and miscellaneous PH drugs (Riociguat) (Bazan and Fares 2015).

For CTEPH patients where an operation can be considered, a pulmonary endarterectomy (PEA) surgery is considered the treatment of choice (Wilkens et al. 2018). However, surgery is not an option for all patients; the thromboembolic obstruction may not be reachable through surgery, or the patient may have a comorbidity which suggests surgery is too high a risk, or patients may elect not to have the surgery. Additionally, up to one-third of patients who have surgery may continue to have CTEPH following the procedure (Hoeper 2015). For these groups of patients, current guidelines support the use of targeted therapies (Wilkens et al. 2018).

Medical management for CTEPH entails anticoagulants and diuretics, with continuing oxygen therapy for patients with hypoxaemia. Anticoagulants are also recommended for life, even following PEA (Galiè et al. 2015). CTEPH patients with inoperable blockages as well as those with enduring PH following a PEA have been the subject of several clinical trials looking at the efficacy of PH-target medical therapy; however, Riociguat is the only PH drug licensed for use in the UK.

PH is defined as a mean pulmonary arterial pressure (mPAP) > 20 mm Hg, compared with a normal mPAP of 14 ± 3.3 mm Hg. The RHC also allows the PH classification to be divided into

pre-capillary (PAH, PH-lung, CTEPH and occasionally some Misc PH), with pulmonary vascular resistance (PVR) ≥ 3 wood units (WU) and pulmonary arterial wedge pressure (PAWP) ≤ 15 mm Hg, or post-capillary (PH-LHD and most Misc PH), where the PAWP is > 15 mm Hg.

1.2 Diagnostic Biomarkers

Biomarkers (an amalgamation of biological markers), can be described as 'a defined characteristic that is measured as an indicator of normal biological processes, pathogenic processes, or biological responses to an exposure or intervention, including therapeutic interventions.' (FDA-NIH Biomarker Working Group 2016). A further subcategory of biomarkers, are diagnostic biomarkers, 'used to detect or confirm presence of a disease or condition of interest or to identify individuals with a subtype of the disease.' (FDA-NIH Biomarker Working Group 2016)

As human technologies have advanced, biomarkers have become more specific and reliable. Medical signs such as the pulse, and taste of urine have been used for thousands of years. In more recent times, the term biomarker tends to refer to certain molecules, such as proteins, detected across various bodily fluids, generally detected, and quantified under laboratory conditions. Blood based biomarkers have been detected in a range of diseases, for example including cardiovascular biomarkers in atrial fibrillation (Chua et al. 2019), proteins for tracking disease progression in Parkinson's Disease (Kitamura et al. 2018). Biomarkers come in a wide variety of molecules, including proteins, hormones, and enzymes.

Biomarkers can play a role in both the diagnosis and prognosis of disease. For example, troponin, a protein discovered in 1965 is used for the diagnosis of myocardial infarction, and the prostate-specific antigen (PSA) can be used for prostate cancer screening. Low-density lipoprotein (LDL) levels can be checked when cholesterol-lowering drugs are used, where it performs a role as both a monitoring biomarker and a prognostic biomarker. Patients with elevated LDL cholesterol are at a higher risk of both death or severe event and higher risk of developing atherosclerosis. Another example of biomarkers as monitors are CD4 antigen counts which are used in the monitoring of HIV infections.

As the shift away from the traditional one-size-fits-all approach continues towards precision medicine, there is an opportunity for biomarkers to play a key role in diagnostics. However, as the field of personalised medicine expands, new and accurate biomarkers must be determined. An ideal biomarker must fulfil three particular criteria. Firstly, it must be detected or measured through minimally invasive procedures so that it can be easily procured. Secondly, it is ideally detectable before the onset of clinical symptoms and will fluctuate depending on disease progression or treatment response. Finally, the biomarker should be adaptable from research to the clinical environment (Condrat et al. 2020).

There are both advantages and drawbacks to different biomarker types. For example, proteins have been found across a spectrum of biological fluids such as cerebrospinal fluid, blood and urine. They have already successfully been used as biomarkers, for example in Alzheimer's disease, a highly sensitive assay was developed to identify and quantify trace quantities of the proteins beta amyloid peptide, tau, and phosphorylated tau, accepted as potential biomarkers

for diagnosis (Chan et al. 2017). However, the cost of identifying novel proteins as biomarkers has proven to be a costly and time-consuming process, owing to the complex structures of proteins, which can include post-translational modifications. Additionally, there are low numbers of clinically significant proteins, and they can be tricky to quantify (Condrat et al. 2020).

1.2.1 Standard biomarkers for PH

A group of hormones principally secreted from the heart, kidney and brain which can result in vasodilation and natriuresis are known as natriuretic peptides. Brain-type natriuretic peptide (BNP) and its N-terminal fragment (NT-proBNP) are examples of these. BNP was initially described in 1988 after it was isolated from a porcine brain, however, it was soon discovered to be a cardiac hormone as it originates primarily from the heart (Weber and Hamm 2006).

NT-proBNP and BNP are released from the heart in response to three important characteristics of cardiac distress; myocardial hypoxia, myocyte stretch, and endocrine activation. Both BNP and NT-proBNP have been shown to correlate with a range of haemodynamic metrics linked to survival outcomes such as mPAP, PVR and RAP (Williams et al. 2006). Additionally, higher BNP levels in blood plasma have been related to worse survival outcomes in PAH patients. Subsequent falls in BNP levels correspond to the commencement of treatments then associated with improved mortality rates.

A hypothetical ideal biomarker to identify early disease or treatment response has yet to materialise for clinical use in PH (Hewes et al. 2020). At present, BNP and NT-proBNP, remain the only blood-based biomarkers which guidelines recommend for routine clinical use in PH (Galiè et al. 2015). However, there are several drawbacks to using NT-proBNP as a biomarker for PH. Firstly, NT-proBNP is not a PH specific biomarker but rather a marker of myocardial stress. Additionally, NT-proBNP levels may be affected by a range of different factors outside of the prevailing disease (Table 1.3).

The DETECT algorithm includes the biomarker NT-proBNP as part of the screening tool for PAH in patients with systemic sclerosis (SSc). PAH-SSc is a common form of associated PAH, with nearly one in five patients with SSc having associated PAH (Coghlan et al. 2014).

***Table 1.3**: Factors Influencing Natriuretic Peptide Levels Independent of Heart Failure. Reproduced with permission from (Brunner-La Rocca and Sanders-van Wijk 2019).*

Increase in natriuretic peptides
Cardiac
● Acute coronary syndrome
● Atrial fibrillation
● Valvular heart disease
● Cardiomyopathies
● Myocarditis
● Cardioversion
● Left ventricular hypertrophy
Noncardiac
● Age
● Female gender
● Renal impairment
● Pulmonary embolism
● Systemic bacterial infections (e.g. pneumonia, sepsis)
● Obstructive sleep apnea
● Critical Illness
● Severe burns
● Cancer chemotherapy
● Toxic and metabolic insults
Decrease in natriuretic peptides
● Obesity

1.3 Bioinformatic solutions for clinical diagnostics

Bioinformatics is an interdisciplinary field of biology and computational science, focusing on applying extracting and analysing information from biomolecules using computational methods. Proportions of bioinformatic analysis fall within the field of artificial intelligence (AI), the aim of which is to imitate human cognitive functions. Powered by a snowballing accumulation of healthcare and clinical data, as well as advances in analytic techniques, a move towards AI assistance is happening in healthcare. If directed by the right clinical

questions, powerful AI techniques can reveal pertinent clinical information from within substantial volumes of data, which in turn can aid the clinical decision process (Tekkeşin 2019).

The growing accrual of healthcare data has gone hand in hand with an increase in our understanding of genetics and genomics. Genetics is the study of genes, along with the part they play in inheritance, and genomics, the study of a person's genome and the way it interacts with the greater environment. With greater understanding has come the launch of precision medicine, starting with a focus on the fields of genetics and genomics, driven in part by the falling costs in time and money to conduct genetic testing. Precision medicine aims to improve personalised care by developing diagnostic and prognostic methods which take into account individual variability.

Clinical management is increasingly incorporating multigene messenger ribonucleic acid (mRNA) signature-based assays, with broad clinical applications in prognosis and diagnosis. Genomic diagnostics are used widely in cancer in order to improve decisions on treatment choices in the clinic, where patients presenting with similar symptoms but differing genomic backgrounds may be treated differently. For example, the drug Trastuzumab is only effective in tumours where the HER2 gene has been overexpressed, so a patient's tumour can be checked against a panel to determine if it could be used (Mao et al. 2021). Another example, AlloMap is a panel of 20 genes, using ribonucleic acid (RNA) gene expression to identify patients with lower probabilities of heart transplant rejection, which can be taken from a non-invasive blood test and has been in clinical use since 2005 (Starling et al. 2006). Another example is the Afirma gene expression classifier, which reduces unnecessary thyroid surgeries compared to management without gene expression classifier testing (Chudova et al. 2010).

1.4 Machine learning

There are several advantages to using AI within the medical field, which have been extensively discussed (Tekkeşin 2019; Murdoch and Detsky 2013). Machine learning (ML) approaches assist in the push for advances in the expansion of the wider AI field. The relationship between AI and ML can be seen in Figure 1.2. With the expansion of big data in the 21st century, the fields of ML and data science have exploded. The methods developed in these fields provide conceivable enhancements to both medical research as well as clinical care.

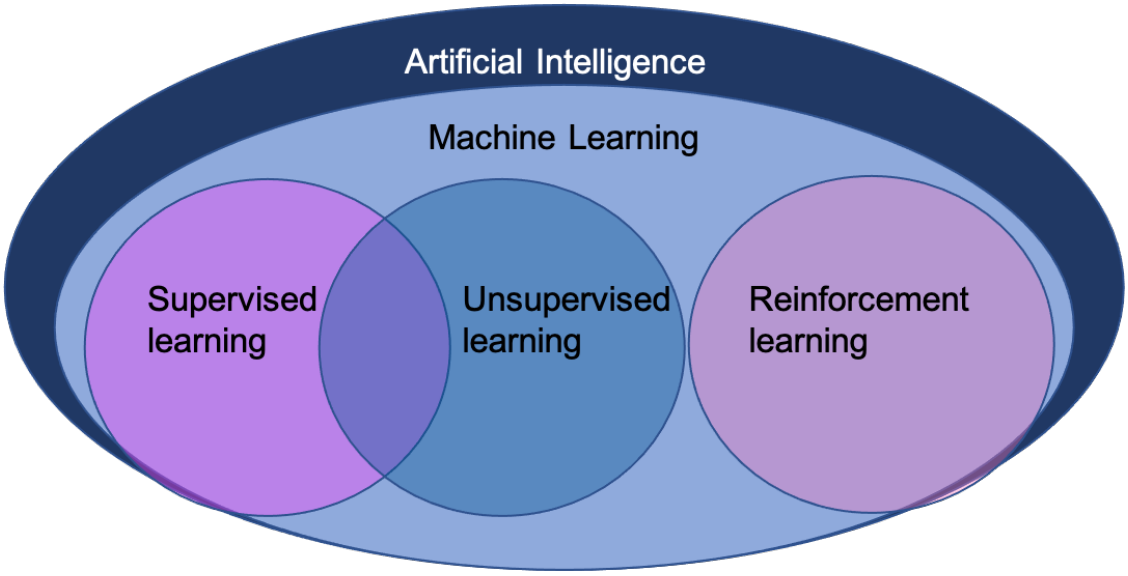


***Figure 1.2**: Relationship between artificial intelligence, machine learning and machine learning subsets.*

Two particular areas of the medical field which might find the use of ML methods advantageous are diagnosis and prognosis. ML algorithms have recently been successfully used to classify patients with breast cancer into triple negative and non-triple negative patients using gene expression data (Wu and Hicks 2021), and to predict the long term mortality risk in transcatheter aortic valve implantation (TAVI) patients (Wu and Hicks 2021; Penso et al. 2021). In this thesis I explore the application of ML to biomarker discovery in pulmonary hypertension (PH).

Machine learning can be defined as 'computational methods using experience to improve performance or to make accurate predictions' (Mohri, Rostamizadeh, and Talwalkar 2012). ML is one of the most frequently utilised forms of AI, and can largely take one of four different forms: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning.

ML methods have been effectively used on a widespread assortment of genomics data sets, overcoming the challenges of processing large dataset sizes and complex data with linear ML models or traditional statistical methods.

1.4.1 Unsupervised machine learning

The aim of unsupervised ML is to discover previously unknown patterns in data. It is applied without knowledge of the outcome variables. Unsupervised ML cannot be applied directly to regression or classification problems because the values for the output data are unknown. Unsupervised learning is best applied to exploring the underlying data structure instead. There are two main types of unsupervised learning applications; clustering and association mining. In clustering, objects are grouped together so that objects with the most similarities are in a group, and those objects have fewer similarities with the objects in another group. Commonalities between data objects are then used to categorise new objects. Association mining identifies sets of objects that frequently together occur in the data. Three common

methods in unsupervised clustering are k-means, hierarchical clustering, and spectral clustering, explained in brief below.

K-means divides n points into k clusters, using the distance between points as the divisor so that points with similar characteristics are clustered together. Initially, k points are randomly assigned as centroids. For all of the other points, the distance between each point and the centroids are measured, and the point assigned to whichever centroid is closer. Next, the central point for the clusters is determined, and the centroid repositioned to that point. The distances are then recalculated, using the new point as the centroid. The distance calculation and subsequent relocation steps are continued until the centroid points do not require repositioning and the clusters are stable.

Hierarchical clustering begins with each data point forming a cluster. The distance between each cluster is computed, and then the clusters with the shortest distance between them are combined to form a new cluster. Once a new cluster has been formed, the distances are again computed and the next two closest clusters are combined into a new cluster. This continues until all points have been combined into one large cluster containing all data points, and the clusters can be represented in a dendogram.

Spectral clustering clusters points based on how connected the points are. This can be done by constructing a distance matrix for the points, then calculating the Laplacian matrix. Using. The eigenvectors are then converted to form a matrix, which is normalised and used for clustering.

1.4.2 Supervised machine learning

Supervised learning involves providing the machine learning algorithm with labelled data. This is data from a known dataset which incorporates the required inputs and outputs, so that the algorithm can find a method to determine how to define both. The algorithm learns from the inputted observations and then identifies different patterns within the data, as well as making predictions based on the correct answers to the problem. Supervised learning can broadly be sub-categorised into classification tasks, regression tasks and forecasting.

1.4.2.1 Classifiers

One of the sub-categories of supervised learning is classification. In these types of tasks, the machine learning algorithm must use labelled values to draw conclusions, to determine the correct label for new observations. The following are common machine classifiers.

Decision trees

A simple non-parametric supervised machine learning method for classification is a decision tree. Decision trees are built from two different components: nodes and branches. Decision trees are sequential. At a decision node, an individual feature is assessed, and the observations split into two, mutually exclusive and collectively exhaustive decision branches. These decision nodes are places where a choice must be completed. The final result of combining these decisions and events are terminal nodes, found at the end of a branch.

1. Begin the tree with the best attribute in the dataset, this is the root of the tree.
2. Partition the training set into subsets, such that each subset has the same value for an attribute.
3. Repeat steps 1 and 2 on each subset until the class label (the leaf node) is found for each branch of the tree.

These trees can be built recursively in R using the rpart package. The training data is recursively split using the features which work best for the classification task, as measured by a chosen metric, such as Gini index or entropy. This allows for clear indications of which attribute should be at the root, and subsequent levels of the tree. The Gini Index calculates the probability of a specific feature that is wrongly classified when randomly selected. If all elements are connected with a single class, then the Gini Index is considered pure (Therneau and Atkinson 2018). Entropy is the measure of randomness or impurity within the data.

Random Forest

A method which builds on these decision trees is Random Forest. Random forest has previously successfully been used to diagnose PH from magnetic resonance imaging data (Lungu et al. 2016). As the name suggests, random forest models consist of large numbers of separate decision trees operating together. Each tree selects a class prediction, and the class with the most votes across the random forest becomes the model's prediction. The steps can be outlined as follows:

1. Randomly select “k” features from total “m” features, where k < m
2. Construct a decision tree for the selected features
3. Build the forest by repeating steps 1 and 2 to create ‘n’ number of trees
4. To use random forest as a classifier, use the each randomly created decision tree to predict and store the outcome from each tree
5. Calculate the votes for each predicted target and take the target with the highest number of votes as the final prediction.

Each tree will have a random selection of variables, so each tree is different. As random forests do not consider all features, dimensionality is less of a problem, additionally, as each tree is created independently, code can be parallelised to speed up computational time.

The random forest classification can be complemented with the Boruta algorithm. Boruta is an all-relevant feature selection wrapper algorithm.

1. Firstly, shuffled copies of all features (known as shadow features) are added to the data set to create randomness
2. A random forest classifier is trained on the extended data set, with a feature importance measure (Mean Decrease Accuracy as default) applied to evaluate the importance of each feature. The higher the measure, the more important the feature.
3. At each iteration, each feature is checked to see if it has a higher Z score than the maximum Z score of its shadow features. Alternatively, this can be thought of as testing to see if the feature importance score is more than the highest of its shadow features.
4. Features which are consistently deemed unimportant are removed
5. The algorithm stops either when all features have been confirmed or rejected, or when a specified limit of random forest runs is reached

6. The selected features can then be fed into the random forest classifier.

Extreme gradient boosting

Another method which builds on decision trees is the gradient boosting method XGBoost, which stands for eXtreme Gradient BOOSting. Boosting is an ensemble method, where several predictions from different models are combined into one. XGBoost is also an example of ensemble learning, combining the predictive power from multiple learners. Decision trees are again built, but here they are built successively so that each tree aims to shrink the errors of the preceding tree or trees. The residual errors are updated with each tree, so that the tree following in the series is learning from an updated version of residuals. The base learner in XGBoost is the weak learner decision tree. As in a random forest model, combining these weak learners results in a stronger learning model, with smaller bias and variance (T. Chen and Guestrin 2016). XGBoost has previously been utilised to develop a screening algorithm to identify patients at high risk of IPAH using routinely collected clinical data (Lungu et al. 2016; Kiely et al. 2019).

Least Absolute Shrinkage and Selection Operator

Least Absolute Shrinkage and Selection Operator (LASSO) is a regularisation technique for estimating generalised linear models (GLM), and is a modified version of linear regression. The linear regression equation can be expressed as Equation 1.

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_n x_n \quad \text{(Equation 1)}$$

Where:

y is the target variable
α is the intercept
$\beta_1, \beta_2, \beta_3, \ldots, \beta_n$ are the coefficients
$x_1, x_2, x_3, \ldots, x_n$ are the features

The parameters α and β are selected through the Ordinary least squares (OLS) method, which minimises the sum of squares of residuals, selecting coefficients for each variable minimising a loss function (Equation 2). Linear regression is not robust to outliers. Another drawback is overfitting because all predictors are considered. LASSO, ridge, and elastic net models are examples of regularised regression, which helps solve the problem of overfitting.

$$L = \Sigma (\hat{Y}i - Yi)^2 \quad \text{(Equation 2)}$$

Where:

L is the loss function
$\hat{Y}i$ are the predicted values
Yi are the actual values

LASSO assigns a penalty, λ, to coefficients in the linear model (Equation 3). This penalty reduces the value of many coefficients to 0.

$$L = \Sigma (\hat{Y}i - Yi)^2 + \lambda \Sigma |\beta| \quad \text{(Equation 3)}$$

LASSO models have previously been used to develop diagnostic models for example using an RNA-seq dataset to classify patients with idiopathic and heritable PAH from healthy controls (Rhodes et al. 2020).

Support Vector Machine (SVM)

SVMs are generated by creating a hyperplane between two classes that allows for the prediction of labels from one (or more) vectors. This decision boundary is orientated by maximising the distance from the closest data point for each class. These nearest points in turn are termed support vectors. Some advantages of SVM are the regularisation capabilities of the model, which reduce the risks of over-fitting. SVMs are also very stable to small changes in the data and can handle non-linear data efficiently. However, SVMs are also very complex algorithms that require a lot of memory to compute and are very computationally expensive. They are not easily interpreted and require scaling of variables before an SVM model can be applied.

K-Nearest Neighbour (KNN)

The KNN algorithm works under the assumption that things which exist in close proximity to each other are analogous. KNN utilises this idea of similarity (or distance) by calculating the distance between points on a graph. KNN is a ‘lazy learner’; there is no training period, and no functions are derived from the training data. The algorithm only learns once it makes a prediction, so can be calculated very rapidly. This has the associated advantage that new data can be added effortlessly, without affecting the accuracy. Finally, KNN is very easy to use as there are only two parameters (the value of K, and the distance function). Conversely, KNNs do not work well with large datasets, where the computational cost rises, as well as reducing model performance and speed. Similarly, KNNs do not perform well in high dimensional data. They are also sensitive to missing values, outliers and noise.

Comparing classification models

Deep learning techniques such as XGBoost present a specific challenge - although they are often high performing models, they are far less interpretable, often referred to as ‘black box’. For the end user, these complex interactions between variables can be difficult to understand. Understanding this trade-off between accuracy and interpretability (Figure 1.3) is essential when considering the most appropriate classification model.

***Figure 1.3**: The accuracy versus interpretability trade-off when selecting a classification method*

Several tools for comparing classifications can be used to help compare the performance of different models. Some of these use the binary measures of whether a subject is correctly classified. Using the example of classifying subjects with PAH vs healthy controls (HC) as an example, we can define:

- True positive (TP) - a patient with PAH is classified as having PAH
- True negative (TN) - a HC is classified as a HC
- False positive (FP) - a HC is incorrectly classified as having PAH
- False negative (FN) - a patient with PAH is incorrectly classified as a HC

With these definitions, we can further define performance metrics:

- Specificity = TN / (TN + FP)
- Sensitivity (aka Recall) = TP / (TP + FN)
- Negative predicted value (NPV) = TN / (TN + FN)
- Positive predictive value (PPV, aka Precision) = TP / (TP + FP)
- Correct classification rate (aka accuracy) = (TP + TN) / (TP + TN + FP + FN)
- Area under the receiver operator characteristic (ROC) curve (AUC); the confidence interval calculated using the method by Delong et al (E. R. DeLong, DeLong, and Clarke-Pearson 1988).

Diagnostic testing is a vital element in evidence-based patient care. Clinicians must weigh the risks and benefits of the test, as well as the diagnostic accuracy when deciding whether or not to use a diagnostic test. A high sensitivity is important where the test is used to identify a serious but treatable disease, for example cervical cancer. The cervical screening program is highly sensitive, so very few cases are missed; however it is not particularly specific – a high proportion of women with a positive smear are eventually found to have no underlying pathology.

For many clinical models, the sensitivity and NPV are the important metrics. For example, the Afirma test which uses an SVM classifier on mRNA expression data to reduce unnecessary thyroid surgeries must have a high sensitivity and a high NPV (Chudova et al. 2010). These metrics are again important in the PAM50 assay which uses 50 genes to create a risk model using a multivariate Cox model using ridge regression fit for breast cancer subtypes (Parker et al. 2009).

1.4.3 Feature selection

To reduce the computational burden of modelling, and or to improve the model's performance, the number of input variables used within a predictive model can be reduced. This is known as feature selection (Cai et al. 2018), or dimension reduction. The aim is to derive a subset of features from the original feature set, which retains the relevant features of the dataset. Feature selection has been successfully used to improve the performance of transcriptomic signatures in a range of classification problems, for example identifying a gene signature from RNA-seq data for malignant prostate cancer (Alkhateeb et al. 2019), as well as a part of a repeated cross-validated feature selection process to generate a 10 gene signature for paediatric sepsis mortality (Abbas and El-Manzalawy 2020).

The use of feature selection has several benefits. Firstly, it can speed up the training time for ML algorithms. Once the algorithm has been trained, the complexity is reduced due to the smaller number of variables. This can also help reduce overfitting and improve the overall accuracy of a model, provided the key important variables are selected.

1.4.4 Dealing with missing data

Missing data is a recurring problem across many fields of research (Raghunathan 2004), with the danger in leaving these missing data untreated established in 2002 (Schafer and Graham 2002). As with most statistical methods, most ML models require the training set to be complete, with no missing features. In order to perform statistical inference where data are missing, the missing data mechanism must be identified. Missing data can be classified into three categories relating to why the data are missing: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). When the data are MCAR, missing data are independent of both the observed and unobserved data (Little and Rubin 2019). In MAR, the data which are missing are systematically related to the observed but not the unobserved data (Little and Rubin 2019). When the data are MNAR, the missing data are systematically related to the unobserved data. It is impossible to distinguish between MAR and MNAR methods.

In medical literature, missing data is a common occurrence (Austin et al. 2021). The most common approach for dealing with this problem in clinical research is to omit participants with missing values. However this can have a large impact on analysis, potentially leading to inappropriate conclusions (J. G. Ibrahim, Chu, and Chen 2012; Stavseth, Clausen, and Røislien 2019; Jakobsen et al. 2017). This form of analysis is known as complete case analysis (CCA).

Single imputation is another frequently used method, where missing values are replaced using a particular rule. For example, the mean value for the variable, or to use the last observation. Single imputation assumptions are often unrealistic and biased, and therefore should be avoided in most instances.

Multiple imputation (MI) is a commonly used method for tackling missing data, building on single imputation. Multiple reasonable values for a particular variable are imputed for every subject without a data point for that variable in MI, with the outcome variable included in the imputation (Moons et al. 2006). As a result, multiple complete data sets are created. The

results of statistical analyses carried out across each of these complete datasets are combined, and the pooled results assessed (Austin et al. 2021).

However, MI is computationally intensive, and will always be at best an approximation of the true value. The process is often inaccurate where there are high proportions of missing values, large numbers of variables, or too few observations (Sterne et al. 2009).

1.4.4.1 Multiple imputation using multivariate imputation by chained equations

Multiple imputation using multivariate imputation by chained equations (MICE) is a popular MI approach, the steps of which are summarised in Table 1.4.

***Table 1.4**: Multivariate imputation by chained equations (MICE) algorithm for multiple imputation.*

1. For each of the k variables missing data, select an imputation model, e.g. regression.

2. Initially replace the missing values at random by selecting from the observed values for that variable. Alternatively, use another method, such as the mean of the present values to replace missing values. Correlations are reduced but imputations can now take account of all available data.

3. Remove the placeholder values for one variable with missing data:
    a. Model the observed values using the other variables using the method selected in 1).
    b. Randomly perturb the estimated regression coefficients in order to reflect the uncertainty in imputed values
    c. Use the model fitted in (a) to with perturbed coefficients to predict the missing values

4. Execute step 3 for each variable missing data.

5. Cycle through steps 3 and 4 (forming one cycle of the imputation process which creates one imputed data set) your chosen number of times (5 - 20 cycles suggested).

6. Create M imputed data sets by repeating steps 2-5 M times and updating the imputations each time.

1.5 MicroRNAs

MicroRNAs (miRNAs) are small, non-coding RNA molecules that play an important role in gene expression regulation and affect a range of biological processes (O'Brien et al. 2018). The first miRNA was discovered in 1993 by the Ambros group (Lee, Feinbaum, and Ambros 1993). The majority of miRNAs have individual promoters and are transcribed by RNA polymerase II or III into primary miRNAs. These primary miRNAs are then processed into precursor miRNAs and finally into mature miRNAs. Most miRNAs prompt translational repression and degradation by interacting with the 3' untranslated region (3' UTR) of their

target mRNAs, though miRNA interactions with other regions have also been reported (Xu et al. 2014). MicroRNAs may also trigger translation or regulate transcription.

1.5.1 Measuring miRNAs

There are 3 main assays for in depth miRNA expression analysis: quantitative polymerase chain reaction (qPCR), RNA-seq using next generation sequencing (NGS) technology and microarrays. The high-throughput method of qPCR involves the amplification of deoxyribose nucleic acid (DNA) by polymerase chain reaction, which is monitored in real time. The PCR method uses an enzyme to amplify a short section of template DNA in cycles. In each cycle, the number of DNA sections are doubled, leading to exponential amplification of these targets. PCR is highly sensitive, rapid technique, and the quantitative nature allows for the measurement of precise values (either relative or absolute) of amplified DNA in samples.

RNA-seq is another high-throughput method. The workflow can be generalised into five steps. Firstly, the RNA is extracted, and then undergoes reverse transcription into complementary DNA (cDNA). Next, this cDNA is fragmented, and adapters ligated to each end. These adapters include functional elements which allow for sequencing, for example, elements which allow for clonal amplification of the fragments. Amplification is next, as well as size selection and quality control. Finally sequencing can occur by analysing the cDNA library with NGS, resulting in short sequences corresponding with the original fragments. RNA-seq can identify transcripts from organisms without a determined genomic sequence. RNA-seq experiments can detect low background signals, as the cDNA sequences can be mapped to specific regions, allowing for the removal of experimental noise, and are easily quantified.

Microchips designed to study the expression levels of multiple different genes concurrently are called microarrays. The key principle underpinning microarrays is based on the binding of complementary sequences. Messenger-RNA is isolated and converted into cDNA, a more stable form of RNA. Restriction endonucleases then cut the DNA molecules into smaller pieces. These fragments are then labelled with fluorescent dyes; Cy3 (green) and Cy5 (red). The labelled cDNA is loaded onto the microarray, where thousands of single-stranded DAN samples corresponding to a single gene are arranged in a grid. Where the fluorescence binds to the complementary base pair in the sample spot, the gene can be seen to be active. DNA fragments which do not bind to the probes are washed away. Scanning the microarray with a laser allows the fluorescently labelled cDNA to ‘light up’, and the gene is identified. The intensity of the light signals seen with the laser are then used to quantify the amount of original mRNA. Microarrays are rapidly being replaced by sequencing technologies, as the data is only ever displayed as values relative to other signals detected on the array. However, there are still benefits to the method. Microarrays are a relatively inexpensive and robust way of looking for differentially expressed genes across the transcriptome.

A drawback to utilising microarrays (or qPCR) for quantifying miRNA expression levels stems from the short length of mature miRNAs. There are also high levels of sequence homology between miRNAs, and many miRNAs have large numbers of isoforms. These factors combine to create difficulties in primer or probe design, as well as hybridisation.

A comparison of different sequencing types for miRNAs in 2014 determined that the type of sequencing required should depend on the particular requirements of the experiment

(Mestdagh et al. 2014). However, RNA-seq is generally regarded as the superior method compared to microarrays for several reasons. Firstly, it is more quantifiable. Microarray values are only relative to other signals on the microarray, whereas RNA-seq data is quantifiable. RNA-seq is also more sensitive to high and low transcription levels, which microarrays can struggle to detect accurately. Secondly, the RNA-seq allows for mapping of cDNA to specific targeted regions on the genome, removing some experimental noise. Finally, RNA-seq can detect transcripts from previously un-sequenced organisms, unlike hybridisation methods which require species specific probes.

With any profiling method, there are systematic disparities and biases initiated during the experimental process. Within NGS, the preferred profiling method in most cases, sources of bias could be introduced from a range of sources, including RNA sample quality, contamination with RNA during library preparation and reverse transcription. Additionally, total read counts vary depending on the miRNA library used. Taking these systematic variations into account is therefore important, and normalisation is a crucial step before abundances of miRNAs may be compared. The goal of this normalisation is to distinguish between true biological signal and random noise. Normalisation methods can largely be grouped into 2 categories; ones applying linear scaling, and a second category of methods which do not apply linear scaling.

1.5.2 Identifying targets and pathways

Predicting the interactions between miRNA and mRNA targets is challenging because each miRNA can regulate from one to a large number of mRNAs, and each mRNA is targeted by multiple miRNAs. The availability and quantity of both mRNA and the miRNAs targeting them may also contribute to which genes are regulated. Measuring the changes in mRNA levels after over- or under- expressing a miRNA may intuitively appear to be a straightforward way to identify miRNA targets, however there are several drawbacks to this approach. For example, there may be indirect signals reflected in downstream genes of the original miRNA target. Additionally, the experimental set-up may not be representative of the workings within an organism. Finally, the miRNA may restrict the efficiency of translation, which would not necessarily be mirrored in mRNA levels.

The regulatory role of the miRNA varies between cell types, ie. some mRNAs may respond differently to miRNA regulation depending on the cell type (O'Brien et al. 2018). This problem is typically tackled through the prediction of targets and followed up with experimental validation of these interactions. There is a lack of experimental evidence to identify miRNA targets, which has driven an increase in computational algorithms aiming to add to these repertories. As miRNAs are short, and only require part of their sequence to be complementary to their target, computational location of targets is extremely challenging. Adding to this, there is a lack of understanding of the procedures which direct the targeting process for miRNAs (Or, Ben Or, and Veksler-Lublinsky 2021). Nonetheless, bioinformatic tools have been developed to try and predict these interactions, including the most popular algorithms, TargetScan, miRanda and DIANA microT.

1.5.3 miRNAs as biomarkers

MicroRNAs have the potential to be ideal biomarker candidates. MiRNAs can be found in a range of different biological fluids, including saliva, breast milk, urine and blood - as well as blood derivatives such as plasma and serum. They are also highly specific to their originating cell or tissue type, and have been shown to vary according to disease progression. Additionally, the technologies for detecting miRNAs are widely accessible. The development of new assays for miRNAs are also faster and cheaper than comparatively producing new antibodies for protein biomarkers (Condrat et al. 2020)

The first use of miRNAs as biomarkers was in the field of cancer in 2008 (Lawrie et al. 2008), and literature examining their use as potential biomarkers has rapidly expanded since then, across a range of different diseases. Abnormal miRNA expression has been associated with a number of diseases in humans (Peng and Croce 2016; Paul et al. 2018). However, miRNAs have not yet made the leap from research to clinical use. This problem is not unique to miRNAs; thousands of papers have been written suggesting biomarkers for a range of diseases, however, only a handful of biomarkers with clinical application have been successfully endorsed for clinical practice (Drucker and Krapfenbauer 2013).

There are a range of reasons the conversion from biomarker discovery to clinical utility has been challenging. Firstly, in rare diseases, amassing a number of patients large enough that a study aiming to uncover biomarkers has enough power can be difficult. Secondly, miRNAs are expressed in different amounts in different tissues and organs, with many miRNAs displaying tissue specific, or even cell specific expression profiles (Precazzini et al. 2021). In complex diseases, such as pulmonary hypertension, it is unreasonable to expect a single biomarker for stratification to be identified, at least in part because complex diseases often affect multiple biological systems. There are of course exceptions, such as the anti-cancer drug trastuzumab (Herceptin®), which can only be dispensed if the singular pharmacogenomic biomarker HER2/neu receptor is overexpressed. However, these singular biomarkers are only achievable under special circumstances (Fröhlich et al. 2018). Multi-biomarker signatures originating from complex high-throughput data are an alternative to the single biomarkers, which allow for a more complete overview of the diseases under investigation. This is the main area where ML can help in uncovering these relationships.

1.5.4 miRNAs in PAH

The first team to report dysregulation of miRs in developing PAH was (Caruso et al. 2010). The team found miR-21 and let-7a down regulated in serum from patients with IPAH. MiR-22, miR-30 and let-7f downregulated; miR-322 and miR-451 upregulated. This was followed by a study by (Courboulin et al. 2011a) which found seven miRs significantly abnormally expressed in patients with PAH compared with controls (miR-204, miR-450a, miR-145, miR-302b, miR-27b, miR-367, and miR-138). There have since been a range of studies looking at miR levels in different cell and tissue types, such as blood plasma and pulmonary artery smooth muscle cells (Rhodes et al. 2013; Schlosser, White, and Stewart 2013; Courboulin et al. 2011b; F. Li et al. 2017; Brock et al. 2009). MicroRNAs in PAH have been reviewed in (Alex M. K. Rothman, Chico, and Lawrie 2014), and more recently been reviewed in (Santos-Ferreira et al. 2020), where they highlighted four miRNAs of importance in PAH (miR-29, miR-124, miR-140, and miR-204).

The detected numbers of miRNAs related to PH pathobiology has increased in recent years, but the additional definition of shared activity of miRNA across diseases may be useful for forming molecular links underlying potentially surprising disease associations with PH. Additionally, a grouping of convergent miRNAs as well as their downstream genes may be more effective than a single miRNA target at improving, preventing or regressing the overall manifestations of PH, an area in which machine learning might be able to help.

1.6 Aims and Objectives

I hypothesised that machine learning could enhance our understanding of pulmonary hypertension by identifying novel miRNAs involved in the disease process and by uncovering more complex relationships. I had 3 aims to aid in the investigation of this:

1. Explore some supervised machine learning methods to classify a small test cohort of patients with PAH and disease and healthy controls using miRNAs.

2. Explore this signature in a much larger cohort of patients, expanding these analyses to include signatures for PH and CTEPH

3. Investigate the application of unsupervised learning in the larger cohort

Chapter 2: A diagnostic miRNA signature for pulmonary arterial hypertension using a consensus machine learning approach

As part of my PhD thesis, I am including work from my published paper 'A diagnostic miRNA signature for pulmonary arterial hypertension using a consensus machine learning approach', which was published in *EbioMedicine*, June 2021, DOI: https://doi.org/10.1016/j.ebiom.2021.103444

Multiple reports exist on the expression and / or function of individual miRNAs in PAH, and reports of miRNA signatures in other disease but when we searched PubMed database using the terms [("Pulmonary Arterial Hypertension" OR "PAH") AND ("machine learning" OR "ensemble learning") AND ("microRNA" OR "miRNA" OR "miR")] for articles before February 20th 2021 0 results were returned. We hypothesised applying machine learning to microRNAs in PAH may provide novel insights. This was the largest microRNA profiling of PAH patients with 64 treatment naïve patients (sampled at the time of diagnosis), and 43 disease and healthy controls at the time. It is also the first machine learning assessment of microRNAs for PAH.

I produced all figures in this paper, with the exceptions of Figure 2.11 (produced by Cai Davies) and Figure 2.10 (Figure 6) which was produced by Dr Josephine Pickworth. Dr Pickworth carried out the qPCR, as well as writing the methods and supplementary information on the analysis. James Iremonger drafted the sections on plasma preparation and RNA isolation, as well as microarray profiling and preprocessing. I drafted all other sections of the paper, after which feedback was provided by all other authors. I was not involved in the collection or processing of samples.

I have also expanded this chapter to include work done by Cai Davies, a Genomic medicine MSc student I co-supervised. With his project, we aimed to build on the validation section by investigating the added value of miRNA targets in classifying patients with PAH in RNA seq. Tables 2.6, 2.12 and 2.13 are modified from his dissertation, I produced all other tables.

2.1 Introduction

Pulmonary arterial hypertension (PAH) is a rare but progressive cardiopulmonary disease which can be sub-categorised into seven sub-groups: Idiopathic PAH (IPAH), heritable PAH (HPAH), drug and toxin induced, PAH associated with other associated diseases, PAH long term responders to calcium channel blockers, PAH with overt features of venous/capillary involvement, and persistent PH of the newborn (Simonneau et al. 2019) (Table 1.1).

Often insidious at onset, PAH is usually rapidly progressive and patients frequently experience significant delays between initial symptom onset, diagnosis (right heart catheter) and

treatment, with little improvement to these delays over that past 20 years (Kiely, Lawrie, and Humbert 2019; Brown et al. 2011). Screening for PAH in connective tissue diseases (CTDs), including systemic sclerosis (SSc) where up to 10-15% of patients develop PAH has been shown to be beneficial (Hachulla et al. 2005) with several screening tools now available (reviewed in (Kiely, Lawrie, and Humbert 2019) recommended (Khanna et al. 2013)). Screening for other forms of PAH is required, and the identification of blood-based biomarkers may help identify patients at risk earlier and reveal drivers of disease (Bauer et al. 2020; Kiely, Lawrie, and Humbert 2019). Current clinically used blood-based biomarkers are limited to markers of cardiac stress e.g. N-terminal pro B-type Natriuretic Peptide (NT-proBNP) that gives little insight into early disease, or the molecular drivers of disease.

MicroRNAs (miRNA) are small, non-coding RNA molecules found in tissues, blood and plasma. They have been shown to be dysregulated in PAH, and contribute to the disease process in animal models (Anwar et al. 2016; Rameh and Kossaify 2016; Miao, Chang, and Zhang 2018). Blood based miRNA biomarkers can be collected without the need for invasive tissue biopsy, and are present in plasma and serum in a stable form. However, with as many as 2300 miRNAs regulating biological processes (Alles et al. 2019), identifying those relevant for diagnosis of PAH can be computationally challenging.

Machine learning as a field has progressively improved our ability to find relevant features in large and high-dimensional data sets collected from genomic studies (Toh, Dondelinger, and Wang 2019). Supervised machine learning methods have been used successfully to develop classifiers for disease diagnosis, as well as to identify potential disease biomarkers (Hira and Gillies 2015). Specifically in PAH we have previously utilised machine learning approaches to study molecular drivers of, and biomarkers for PAH (Kiely et al. 2019; Rhodes, Wharton, et al. 2017; Rhodes, Ghataorhe, et al. 2017; Bauer et al. 2020). In this study, we identify miRNA biomarkers associated with PAH selected using a consensus of four different supervised machine learning feature selection techniques. We assess the potential of miRNAs as a diagnostic tool by creating binary predictive classification models and assessing the accuracy of these models. Further insight into the role of miRNAs in the pathogenesis PAH and potential candidates for therapeutic intervention is revealed through the analysis of miRNA target genes and pathways in human lung and whole blood transcriptomes.

2.2 Methods

2.2.1 Cohort overview and sample collection

We collected 83 unique plasma samples from sequentially consented patients with suspected pulmonary hypertension and controls, obtained according to the Declaration of Helsinki, with local research ethics committee approval and informed written consent from all subjects from the Sheffield Teaching Hospitals Observational study into Pulmonary Hypertension, Cardiovascular and Lung disease Biobank (STH-Obs, UK REC 18/YH/0441). Patient samples were obtained from the diagnostic right heart catheter and were PAH-treatment naïve. From the 83 samples, 18 patients with SSc-associated PAH (SSc-PAH) and 10 SSc patients without PH (SSc-without PH) were incorporated into the PAH patient groups and controls respectively. All patients with SSc were of the limited cutaneous subtype. The rest of the Sheffield samples

were comprised of 34 IPAH patients and 21 healthy controls. An additional 24 patient and healthy control samples were obtained from the Imperial College London Pulmonary Hypertension sample collection (UK REC 17/LO/0563) and included in the study to remove a single centre bias. All samples were collected between 2007 and 2013, then stored in plasma at -80^{o}C until the miRNA extraction. The cohort comprising all available samples meeting these criteria at the time of miRNA extraction, was randomly assigned to training (two-thirds) and validation (one-third) sets, matched for age, sex and WHO functional class, with demographics seen in Table 2.1, and missing data for patients with PAH found in Table 2.2. The training set was used to build models, which were evaluated in the validation set to minimise overfitting bias.

***Table 2.1**: Basic demographics for a cohort of healthy controls (HC) and patients with PAH from Sheffield and Imperial, profiled for miRNA expression. Patients with systemic sclerosis (SSc) were included in both the HC and PAH classification sets. Not all metrics were available for all patients. Continuous variables described as mean (standard deviation). For missing values, see Table 2.2.*

	Training Set		**Validation Set**	
	HC + SSc without PAH	**IPAH + SSc-PAH**	**HC + SSc without PAH**	**IPAH + SSc-PAH**
No. Sheffield samples	14 + 7	23 + 11	7 + 3	11 + 7
No. Imperial Samples	8 + 0	8 + 0	4 + 0	4 + 0
Total sample no.	**29**	**42**	**14**	**22**
Mean age at sampling in years (years)	54.1 (14.5)	56.5 (14.3)	51.6 (11.7)	57.4 (15.3)
Female (%)	12 + 6 (58.1%)	18 + 6 (57.1%)	7 + 3 (71.4%)	8 + 6 (63.6%)
Alive 5 years follow up (%)	28 (97%)	28 (65%)	14 (100%)	9 (43%)
WHO Functional class (I, II , III, IV)	-	(0,6,33,3)	-	(0,3,17,2)
Patients on immunomodulatory agent at sampling	4	2	2	2
Mean Pulmonary Arterial Pressure (mm Hg)	-	54.9 (15.6)	-	49.4 (13.7)
Pulmonary vascular resistance (dynes)	-	870 (488)	-	753 (448)
6 minute walk distance: Imperial only (m)	-	202 (158)	-	378 (59)
ISWD: Sheffield only (m)	-	214 (169)	-	248 (246)
Cardiac Output (L/min)	-	4.8 (1.4)	-	5.0 (2.0)
Mean pulmonary arterial wedge pressure (mm Hg)	-	10.4 (3.8)	-	10.8 (3.2)

***Table 2.2**: Missing values for PAH patient's data. 6MWD Imperial only, ISWD Sheffield only*

Parameter	No missing training data (%)	No missing validation data (%)	Total no missing (%)
Mean Pulmonary Arterial Pressure	2 (5)	3 (14)	5 (8)
Pulmonary vascular resistance	6 (14)	4 (19)	10 (16)
6-minute walk distance	2 (25)	2 (50)	4 (33)
ISWD	7 (20)	2 (12)	9 (17)
Cardiac Output	3 (7)	4 (19)	7 (11)
Mean capillary wedge pressure	5 (12)	5 (24)	10 (16)

2.2.1.1 Plasma preparation and RNA isolation

Total RNA was isolated from 1 ml of Citrate plasma using the Norgen total RNA slurry format extraction kit (Norgen Biotek Corp. Canada). RNA was concentrated using the RNA Clean and Concentrate-5 kit (Zymo Research Corp, U.S.A).

2.2.1.2 Microarray profiling and preprocessing

Agilent single colour miRNA arrays miRbase v.19 (Agilent Technologies, UK), which can detect up to 2006 human miRNAs, were performed on purified and concentrated plasma RNA in 2015. Raw microarray signals were normalised using the quantile method within the robust mean array (RMA) method from the R package AgiMicrorna (v.2.14.0) (López-Romero 2011), correcting for the background signal. MiRNAs were then filtered, keeping only those expressed in at least 10% of arrays, leaving 393 miRNAs. Expression levels were log2 transformed and all subsequent calculations were performed on this value. Independent filtering increases detection power in high-throughput experiments. Additionally, several of the feature selection methods utilised below cannot account for multicollinearity. As such, we undertook two subsequent filtration steps to reduce the starting number of miRNAs. MiRNAs were filtered down to 179 by those which have been qPCR confirmed to exist by Exiqon, and therefore, we can assume they can be accurately quantified by the Agilent array. We further eliminated features with high mean absolute correlation, using a correlation matrix method. For each feature, the mean absolute correlation based on pairwise correlations was calculated. If a pair-wise correlation was > 0.7, the feature with the greater mean absolute correlation was removed, using the caret package (v6.0-86) in R. Where two miRNAs are highly correlated both with each other and disease status, and both are kept in the model, there is a danger that both may be considered insignificant, potentially missing an important signal. We carried forward our downstream analysis with 42 miRNAs after filtering. The workflow is described in Figure 2.1.

We used PCA and t-SNE analysis to visually explore the data. PCA analysis was carried out using prcomp in R without scaling the data, and a t-SNE analysis was run using the Rtsne package (version 0.15).

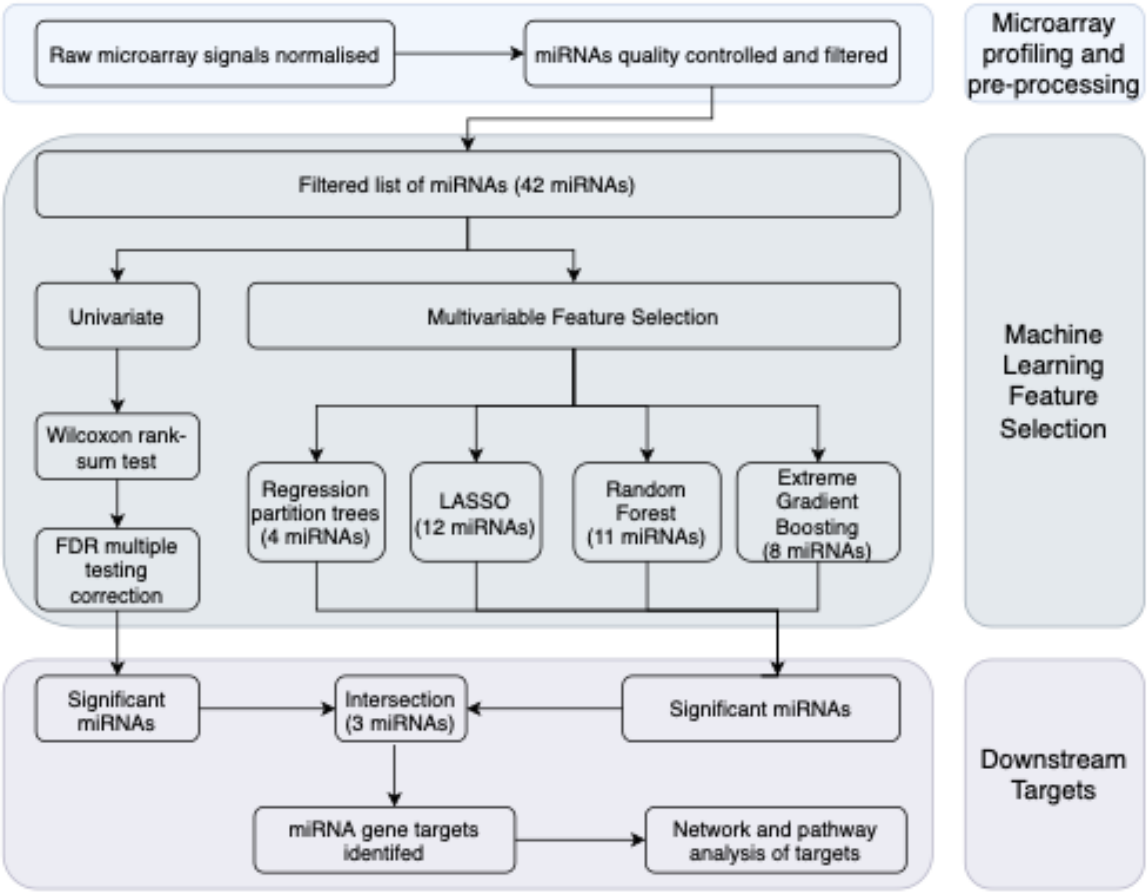


***Figure 2.1**: Machine learning methodology for the identification of miRNAs which may play a role in PAH, and the assessment of their target genes.*

2.2.2 Statistical Analysis

2.2.2.1 Multivariable microRNA selection and model building

All statistical analyses were carried out using R (v4.0.0) (R Core Team 2013). We used both a multivariable and univariable approach to selecting miRNAs. In the multivariable approach, we used four separate feature selection methods simultaneously to identify candidate biomarkers, with the intersection amongst the methods considered the significant miRNAs. In each instance, parameters were tuned using 10-fold cross-validation (repeated 10 times) on the training set. For each of the feature selection methods, we subsequently used a supervised machine learning approach for binary classification to create predictive classification models, based on features selected from the prospective cohort study. For further details on the parameters used, see the code available on github at https://github.com/niamherrington/microarray-miRNA. The guidelines of the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) statement were followed.

Random Forest using Boruta

Boruta is a feature selection random forest wrapper algorithm designed to identify all relevant variables in a classification framework (Kursa and Rudnicki 2010). We performed 300 iterations of the random forest normalised permutation importance function to obtain attribute importance, using default settings within Boruta package (v7.0.0) in R, including the confidence level of 0.01. After the 300 runs were complete, miRNAs still not confidently classified as important variables were rejected along with the miRNAs rejected by the algorithm. This process was then repeated 100 times, with miRNAs selected on at least 10 occasions were carried forward.

We then combined the microRNAs selected by Boruta into a random forest model using the randomForest package (v.4.6-14) (Liaw and Wiener 2002). We selected a random forest model as they are generally robust to overfitting, and capable of learning non-linear relationships. However, the results may not be easily interpretable. The caret package was used to identify 1000 trees as being optimal among the 100, 250, 500, 750, 1000, 1250 and 1500 trees tested. The number of variables available for splitting at each tree node was optimised next, with 1 variable per tree node the best out of a range from 1 to 4. A probability threshold of > 0.5 was used to determine whether a subject was a PAH patient or no PH.

Regression partition tree

Classification trees were calculated using Rpart (v4.1-15) (Therneau and Atkinson 2018) and caret in R. A major advantage of rpart is the interpretable output, that can be displayed graphically. However, a disadvantage is that the trees tend to have a lower predictive accuracy, due to the fact the trees are less robust. The trees were used by the greedy feature selection algorithm, recursive binary splitting to return ordered features, from the root of the tree down.

The fit of the model was controlled by setting the minimum number of observations that must exist in a node for a split to be attempted to four, and the minimum number of observations in any terminal node set to two. The trees were split by minimising the Gini index at each split. This was then cross-validated using 10-fold, repeated cross-validation. We considered a variable selected if it was present in the final tree. A probability threshold of > 0.5 was used to determine whether a subject was a PAH patient or no PH.

LASSO

Least absolute shrinkage and selection operator (LASSO) on binomial logistic regression using the glmnet package in R (v4.0) (Friedman, Hastie, and Tibshirani 2010) was used to select relevant miRNAs, by eliminating parameters with a coefficient of 0. One of the advantages to using a LASSO method is that coefficients are shrunk and removed, reducing variance without substantially increasing the bias (Fonti and Belitser 2017). Additionally, LASSO models allow for effectively interpretable output. However, a drawback to LASSO is a lack of flexibility to fully capture non-linear relationships. We chose the regularisation parameter, λ, using 10-fold cross-validation with binomial deviance as the criterion. From the cross validations, the value of λ with the minimum binomial deviance (λ-min = 0.0502) was selected and used to refit the model. A probability threshold of > 0.5 was used to determine whether a subject was a PAH patient or no PH. To ensure the models were not driven by age and sex, we also attempted to classify patients using these characteristics in a LASSO model.

XGBoost

The final model we used to fit miRNA features to disease diagnosis was the gradient boosting method, using the XGBoost package in R (v1.0.0.1) (T. Chen and Guestrin 2016). We trialled XGBoost as it has been used very effectively in a range of classification problems, consistently winning machine learning competitions on Kaggle, as well as providing insights into biological data sets. However, with many hyperparameters to tune, computational time is longer than some of the other methods, additionally, the results can be difficult to interpret. XGBoost is an extreme gradient boosting method which ranks the features from most to least important. To decide on the regularisation parameter settings, we used a grid search over a range of values, using 10-fold repeated cross-validation on the training set, selecting the optimal values for the final model (Table 2.3). The optimisation ranges were selected by expanding grid searches previously used by other teams on RNAseq data (Y. Li et al. 2017). The ability to fine-tune these parameters in XGBoost means the model is more robust to overfitting. Features contributing to more than a 5% improvement in accuracy to their branches were selected as 'important'. A probability threshold of > 0.5 was used to determine whether a subject was a PAH patient or no PH. Once features had been selected, the model was retrained over the same parameter range, using just selected miRNAs.

Table 2.3*: Parameters used to optimise an XGBoost classifier for PAH using miRNAs. a: the range of each parameter tuned, b: the optimal parameter for the initial xgboost model, c: the final parameter value used for an xgboost model trained on a reduced number of miRNAs.*

Parameter	**Available Range**	**Optimisation range**[a]	**Initial value**[b]	**Optimal value**[c]
No of trees	1 - ∞	100 - 10 000	4300	200
Learning rate	0 - 1	0.01, 0.025, 0.05, 0.1, 0.2, 0.3	0.025	0.025
Maximum tree depth	0 - ∞	1, 2, 3, 4, 5, 6	1	1
gamma	0 - ∞	0, 0.05, 0.1, 0.5, 0.7, 0.9, 1	0.05	1
Minimum child weight	0 - ∞	1,2,3,4	2	1
Subsample rate (row sampling)	0 - 1	0.5, 0.75, 1.0	0.5	0.5
% feature used in each boost (column sampling)	0 - 1	0.4, 0.6, 0.8, 1.0	0.4	0.4

Ensemble

An ensemble of predictions from the above classifiers were generated by averaging the predicted probabilities from each individual supervised machine learning approach, and then using a threshold of > 0.5 to call subjects with PAH.

Comparison with NT-proBNP
All patients, and healthy controls from Sheffield had routine clinical measurements of NT-proBNP. This information was used to compare the accuracy of the miRNA models with NT-proBNP as a classifier by retraining each of the models with NT-proBNP as an additional variable. The performance of standalone NT-proBNP for the cohort was also measured.

Multivariable classifier performance assessment
We also used a leave-one-out cross validation approach (LOOCV) to compare miRNAs selected when the entire dataset was used. All methods above were attempted across the whole dataset, using a LOOCV approach instead of repeated cross validations. AUCs were calculated using the average of the cross validations across the whole dataset, rather than using training and validation sets.

Classification without SSc
Finally, we repeated the above machine learning methods to classify patients with IPAH or healthy controls, using the same training and validation sets described above, without patients with SSc.

2.2.2.2 Univariable analysis

Using a Shapiro-Wilk test (Shapiro and Wilk 1965) for the selected miRNAs, a normality assumption for the majority of miRNAs is violated. As a result, for each miRNA, we performed a non-parametric Wilcoxon rank-sum test, comparing expression levels between patients with PAH and the no PH group, to find a single p-value for each miRNA. These p-values were then adjusted using the Benjamini Hochberg multiple testing correction to control the false discovery rate (FDR) with a cutoff of 0.05. We calculated the discriminatory power of each individual miRNA, using the training set to find an optimal cutpoint by simultaneously maximising sensitivity and specificity, then calculating the accuracy using the validation set. We examined survival using the Kaplan-Meier method for each selected miRNA and calculated the p-value for a log-rank test. All participants were followed up for five years after the sample date, or date of death, with no participants lost to follow up. Cox proportional hazard tests were done using the survival package (v2.44-1.1)

2.2.2.3 Classification performance of multivariable models

To compare classifiers, we looked at how accurately each classifier categorised each patient in the validation set. We also looked at the performance of each feature selection method, by comparing them using the following evaluation metrics, where TP represents true positive, FN represents false negative, TN represents true negative, and FP represents false positive.

- Sensitivity = TP / (TP + FN)
- Specificity = TN / (TN + FP)
- Positive predictive value = TP / (TP + FP)
- Negative predicted value = TN / (TN + FN)
- Correct classification rate = (TP + TN) / (TP + TN + FP + FN)
- Area under the receiver operator characteristic (ROC) curve (AUC); the confidence interval calculated using the method by Delong et al (Elizabeth R. DeLong, DeLong, and Clarke-Pearson 1988).

2.2.3 Pathway Analysis

Gene targets were inferred using DIANA v5.0 microT-CDS (Paraskevopoulou et al. 2013) for the miRNAs which appeared in all four features selection methods, with the threshold for target prediction set to the default of 0.7. We then carried out a network analysis using WebGestalt (Liao et al. 2019) and Cytoscape (v3.7.1) (Shannon et al. 2003). Pathway genes were downloaded from KEGG (Kanehisa and Goto 2000).

2.2.4 External Validation in Whole Blood RNA seq

RNA sequencing was performed on whole-blood samples from 359 patients with PAH, and 72 controls, as previously described (Rhodes et al. 2020). 28 of the Sheffield samples, and two Imperial healthy controls were also included in the miRNA cohort, so we excluded these to ensure the validation set was independent. We split the cohort into the same training and validation groups, and then used XGBoost to classify patients using the gene targets identified using similar optimisation ranges as above. As this dataset is unbalanced due to a comparatively small number of healthy controls, we incorporated a weighting parameter; number of PAH cases / number of controls. The final parameters selected can be seen in Table 2.4. The threshold value was calculated using Youden's Index. We compared this model to 3 additional models formed from randomly selecting 548 genes from all available genes in the whole-blood set and training an XGBoost classifier on each of these sets. In order to select these genes, the 548 gene targets were removed, then the 'sample' function was used 3 separate times to select 548 genes each time.

We also assessed the added value of the selected genes by comparing the Akaike Information Criterion (AIC) of logistic regression models created by using the top 15 genes selected from the gene targets to a random logistic regression model from 15 randomly selected genes.

Table 2.4*: Parameters used to optimise an XGBoost classifier for PAH using mRNAs. a: the range of each parameter tuned, b: the optimal parameter for each model*

Parameter	**Available Range**	**Optimisation range**[a]	**miRNA gene target model** [b]	**Random model 1**[b]	**Random model 2**[b]	**Random model 3**[b]
No of trees	1 - ∞	100 - 10 000	550	800	5150	200
Learning rate	0 - 1	0.01, 0.025, 0.05, 0.1, 0.2, 0.3	0.05	0.05	0.025	0.05
Maximum tree depth	0 - ∞	1, 2, 3, 4, 5, 6	3	2	3	4
gamma	0 - ∞	0, 0.05, 0.1, 0.5, 0.7, 0.9, 1	0.05	0.5	0	0.05
Minimum child weight	0 - ∞	0.2, 0.5, 1, 2	0.2	2	0.2	0.2

Subsample rate (row sampling)	0 - 1	0.5, 0.75, 1.0	0.5	1	0.5	1
% feature used in each boost (column sampling)	0 - 1	0.4, 0.6, 0.8, 1.0	0.4	0.6	0.4	0.4

2.2.5 External Validation in published lung tissue microarray studies

Two publicly available datasets profiling lung tissue from patients with PAH were used to validate the gene target lists. In GEO accession GSE15197 (Rajkumar et al. 2010), differential expression was measured in 13 normal lung tissue samples compared to 18 lung tissue samples with PAH. We excluded seven samples where patients had PH secondary to idiopathic pulmonary fibrosis (IPF). The original study found 13,899 genes differentially expressed between patients with PAH and healthy controls. GEO accession GSE53408 (Zhao et al. 2014) compared 12 samples of lung tissue from patients with PAH to 11 healthy lung tissue samples. Basic characteristics of the two cohorts are described in Table 2.5.

Table 2.5*: Characteristics of 2 GEO datasets, GSE15197 and GSE53408. *Information missing for four patients and three controls*

GSE15197	***n***	**Age, yr**	**Sex (M/F)**	**PVRI, Wood units**	**MPAP, mmHg**
GSE15197 PAH	18	44 ± 10	7/11	20 ± 9	55 ± 7
GSE15197 Normal controls	13	60 ± 11	5/8		
GSE53408 PAH*	8	40 ± 12	3/5		56 ± 9
GSE53408 Normal controls*	8	47 ± 15	4/4		

The GEOR2 interface was used to import data into R using Biobase (v2.42.0) and GEOquery (v2.50.5). The limma package (v3.38.3) used for differential expression analysis with a log2 transform. Gene targets were extracted and FDR corrected (<0.05) using the Benjamini Hochberg correction.

2.2.6 qPCR validation of gene targets

Pulmonary artery smooth muscle cells (PASMCs) purchased from commercial suppliers (Lonza catalogue # CC-2581) taken from healthy donors and PASMCs isolated from four separate IPAH patients (donated from Prof. N Morrell of Cambridge University) as previously described (Pickworth et al. 2017), were grown in culture before being quiesced (0.2% foetal Calf Serum) for 48 hours, and lysed for the isolation of RNA using Trizol. Direct-zol RNA mini-prep kits (Zymo research R2050), and Zymospin column were used to extract RNA as per manufacturer's instructions. RNA (n=3 for each condition) was reverse transcribed to cDNA using RNA to cDNA kit (Applied Biosystems 4387406). Eight genes were selected for

quantitative-PCR (qPCR) and TaqMan probes for FER (Hs00245497_m1), UCR3 (Hs00419575_m1), MTUS1 (Hs00368183_m1), API5 (Hs00362482_m1), PELI1 (Hs00900505_m1), HGF (Hs00300159_m1), GLMN (Hs00369634_m1), PARP8 (Hs01065404_m1) were purchased from Thermo Fisher and run in duplicate. Human ATP5B Hs00969569_m1 was used as control. Relative quantity was calculated using the ΔΔCt method. Analysis was performed using GraphPad Prism v 8.2.

2.2.7 Added value of miRNA gene targets for classification (Cai Davies)

There are 505 genes with differential RNA expression in patients with PAH compared with controls in the validation cohort described above (Rhodes et al. 2020). To investigate the potential added value of identifying miRNA targets, two further XGBoost models were explored, with the final parameters shown in Table 2.6. The first examined the utility of these 505 genes to classify patients (model 1), and the second ran in parallel, examining performance of a classifier built with both these 505 genes, and the gene targets present in the RNA seq set (model 2).

***Table 2.6**: Final model best parameters from two XGBoost models in classifying PAH from controls*

Parameter	**Available Range**	**Model 1 final value**	**Model 2 Final value**
No of trees	1 - ∞	5650	8450
Learning rate	0 - 1	0.01	0.01
Maximum tree depth	0 - ∞	3	3
gamma	0 - ∞	0.9	0.9
Minimum child weight	0 - ∞	2	2
Subsample rate (row sampling)	0 - 1	0.5	0.5
% feature used in each boost (column sampling)	0 - 1	1	0.6

These models were then examined to look for the top 20 genes with the highest importance to the XGBoost models, as measured by the percentage gain, where gain is a measure of the contribution of the feature to the model relative to each feature's contribution for each tree in the model. A higher value for one feature when compared to another infers it is of a higher importance for generating a prediction.

2.3 Results

We profiled the miRNAs from 64 patients with PAH and 43 combined SSc-without PH and healthy controls (no PH). Initial t-Distributed Stochastic Neighbour Embedding (t-SNE) and principal component analysis (PCA, Figure 2.2) showed some separation between groups. Since several of the feature selection methods utilised later cannot account for multicollinearity, we undertook two filtration steps to reduce the starting number of miRNAs. Initially the miRNAs were filtered, removing those failing quality control, and miRNAs highly correlated to each other, to leave 42 miRNAs (Figure 2.3). Next, we selected the miRNAs most predictive of PAH vs no PH using four different supervised machine learning methods.

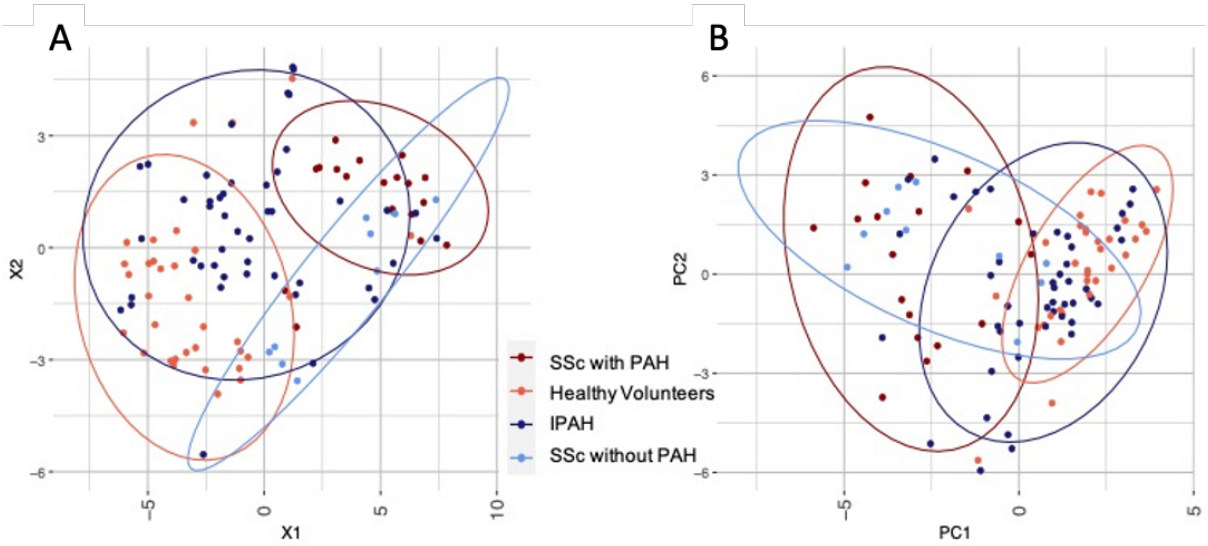


***Figure 2.2**: A) t-SNE plot of subjects in both the training and validation sets. B) PCA plot of subjects in both the training and validation sets, showing the first two principal components. PC1 (27.1% of variance), and PC2 (21.4% of variance).*

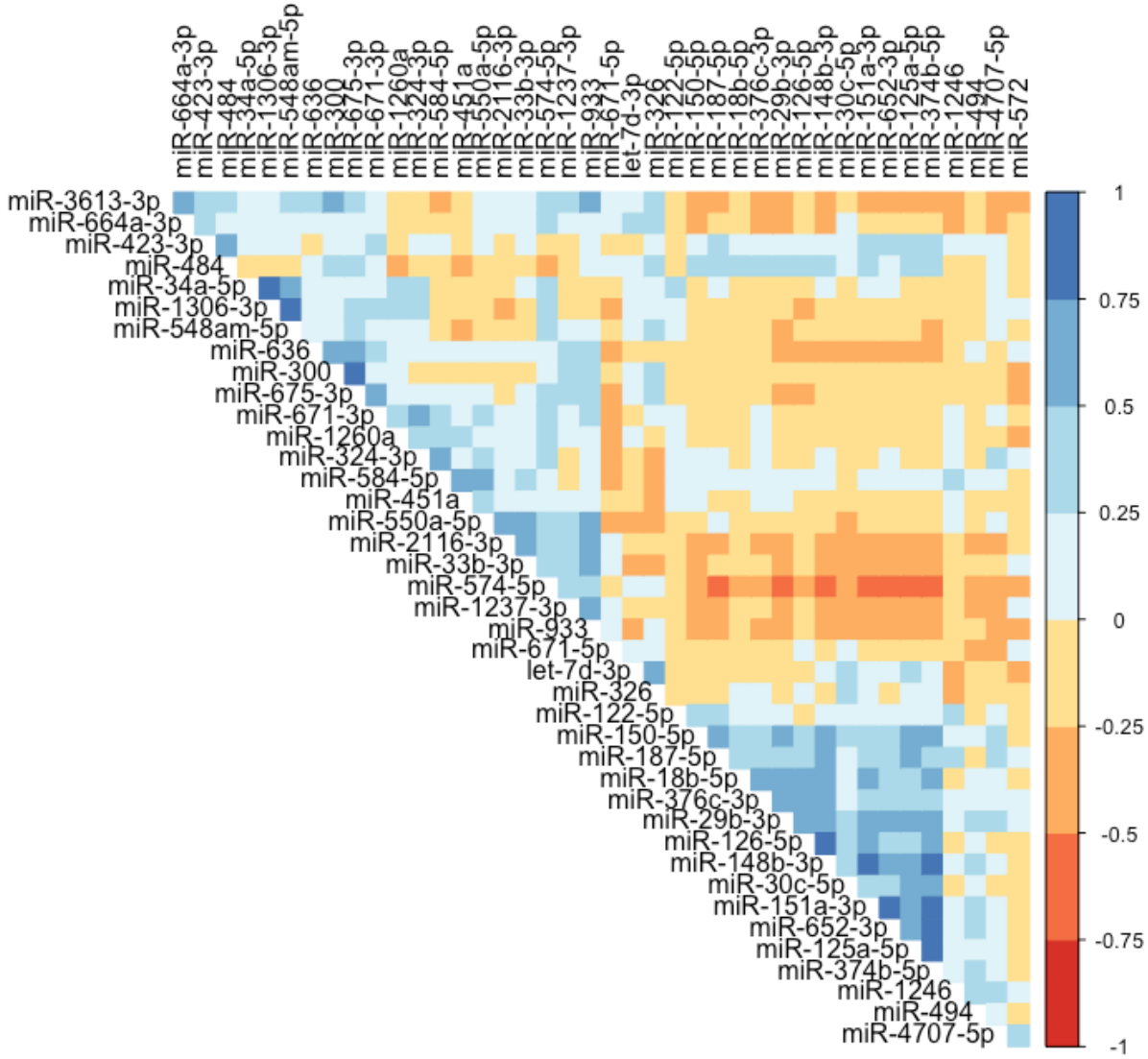


***Figure 2.3**: Correlation plot of the miRNAs remaining after filtering out those with high correlation (Spearman's > 0.7)*

2.3.1 miRNAs selected using supervised machine learning approaches

The disease diagnosis (PAH vs no PH) of 72 individuals was described as a function of the 42 miRNAs using four different machine learning methods. Feature selection was used to determine the miRNAs most relevant to the diagnosis. Four different machine learning techniques were used to select miRNAs and model PAH diagnosis; Boruta (an embedded random forest method), LASSO, regression partition trees, and XGBoost (an extreme gradient boosting method). The features subsets selected by each method were all different, though there were overlapping miRNAs in all (Figure 2.4). Two miRNAs were selected by all four methods; miR-636 and miR-187-5p. These 2 miRNAs were the most consistently selected when different discovery sets were utilised; a training and validation set approach, leave-one-out cross validated approach, and a training and validation set approach without patients with SSc (Figure 2.5).

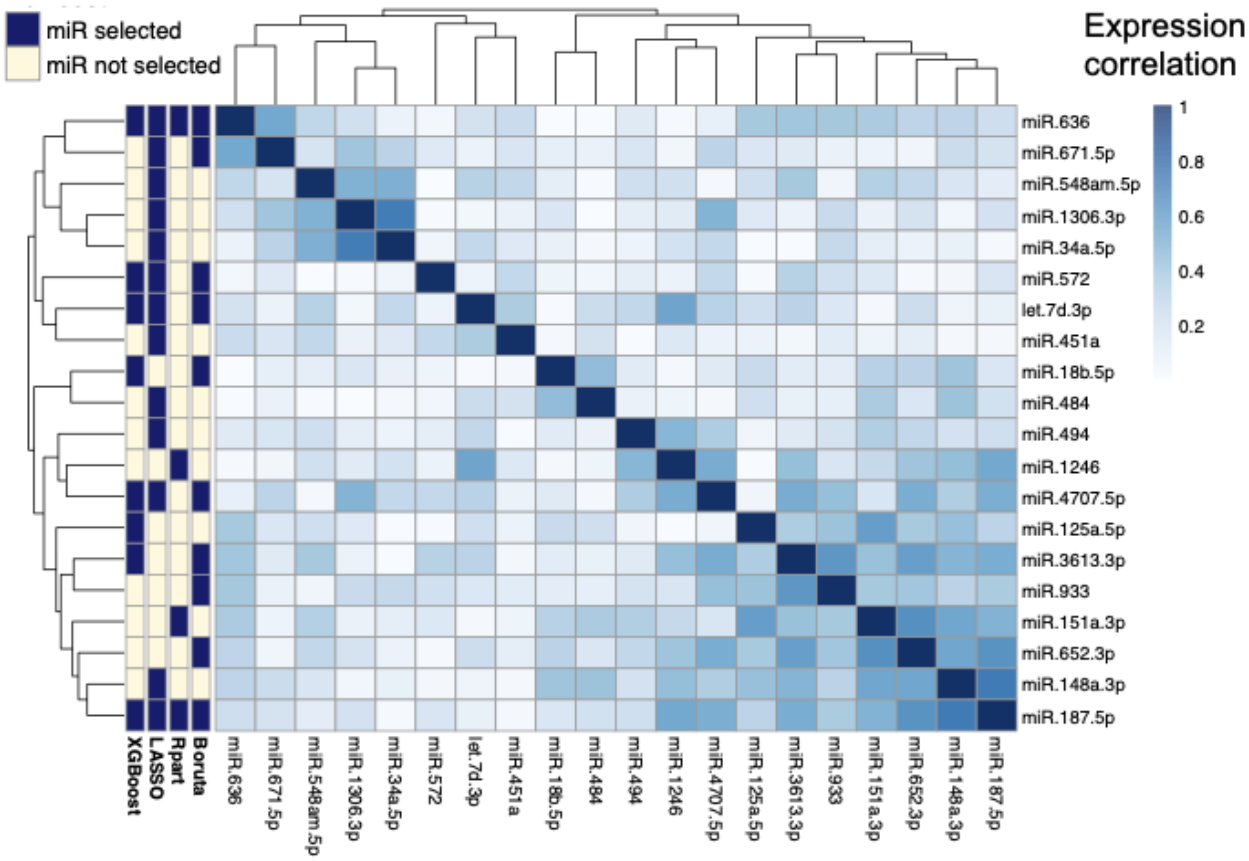


***Figure 2.4**: Absolute Expression correlation (Spearman) matrix between miRNAs selected by machine learning methods (side-bar). Dendrogram orders miRNAs by hierarchical clustering. XGBoost: Extreme gradient boosting method. Rpart: a regression partition tree method. Boruta: a random forest wrapper method for feature selection.*

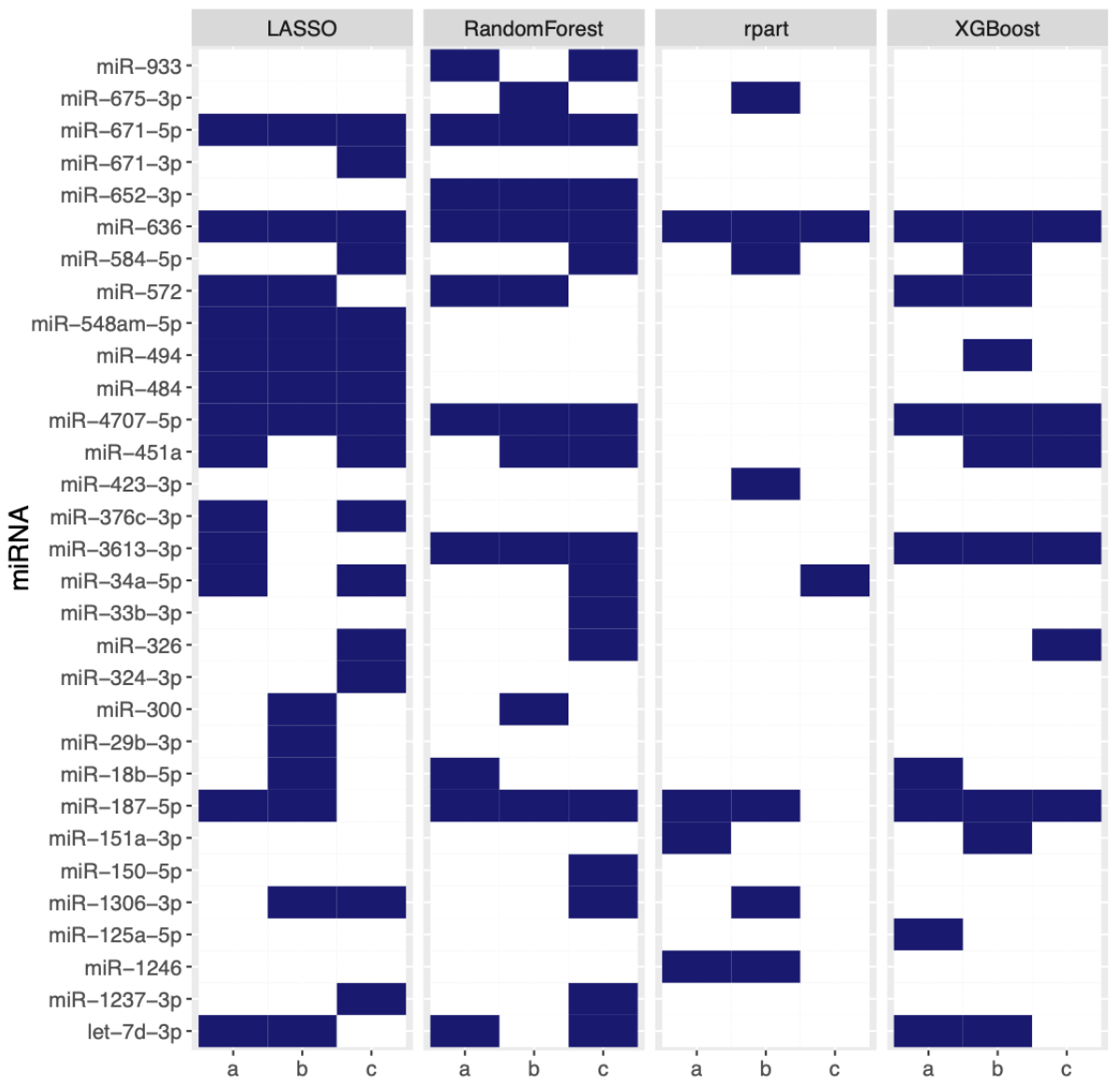


Figure 2.5*: Heatmap of selected miRNAs using four different supervised machine learning approaches across three different discovery sets: a) a training and validation cross validation approach for IPAH and PAH-SSc vs healthy controls and PH-without SSc; b) leave-one-out cross validation approach across the whole dataset for IPAH and PAH-SSc vs healthy controls and PH-without SSc. c) training and validation cross validation approach for patients with IPAH and healthy controls. Blue: miRNA was selected, white: miRNA was not selected*

2.3.2 Performance of PAH classification using miRNAs

To compare the performance of each feature selection method, we looked at how each model performed as a classifier on the validation set. The classification of each subject by each model can be seen in Table 2.7. Boruta random forest had the highest overall accuracy, with 30 out of 35 subjects in the validation set correctly identified.

***Table 2.7**: Model Classifications on the validation set for four different methods; regression partition trees (Rpart), LASSO, random forest wrapper (Boruta), extreme gradient boosting (XGBoost) and an ensemble prediction*

Patient ID	Diagnosis	Random Forest prediction	Rpart prediction	LASSO prediction	XGBoost prediction	Ensemble prediction
1	Healthy control	PAH	PAH	PAH	Control	PAH
2	Healthy control	Control	Control	Control	Control	Control
3	Healthy control	Control	Control	Control	Control	Control
4	Healthy control	Control	Control	Control	Control	Control
5	Healthy control	Control	Control	Control	Control	Control
6	Healthy control	Control	Control	Control	Control	Control
7	Healthy control	Control	Control	Control	Control	Control
8	Healthy control	Control	Control	PAH	Control	Control
9	Healthy control	Control	PAH	Control	Control	PAH
10	Healthy control	PAH	PAH	PAH	PAH	PAH
11	Healthy control	Control	Control	Control	PAH	Control
12	SSc-without PAH	PAH	PAH	PAH	PAH	PAH
13	SSc-without PAH	PAH	PAH	PAH	PAH	PAH
14	SSc-without PAH	Control	Control	Control	Control	Control
15	SSc-PAH	Control	PAH	PAH	Control	PAH
16	SSc-PAH	PAH	PAH	PAH	PAH	PAH
17	SSc-PAH	PAH	PAH	PAH	PAH	PAH
18	SSc-PAH	PAH	PAH	Control	PAH	PAH
19	SSc-PAH	PAH	PAH	PAH	PAH	PAH
20	SSc-PAH	PAH	PAH	PAH	PAH	PAH
21	SSc-PAH	PAH	PAH	PAH	PAH	PAH

Patient ID	Diagnosis	Random Forest prediction	Rpart prediction	LASSO prediction	XGBoost prediction	Ensemble prediction
22	IPAH	PAH	PAH	PAH	PAH	PAH
23	IPAH	PAH	PAH	PAH	PAH	PAH
24	IPAH	PAH	PAH	PAH	PAH	PAH
25	IPAH	PAH	PAH	PAH	PAH	PAH
26	IPAH	PAH	PAH	PAH	PAH	PAH
27	IPAH	PAH	PAH	PAH	PAH	PAH
28	IPAH	PAH	PAH	PAH	PAH	PAH
29	IPAH	PAH	Control	PAH	PAH	PAH
30	IPAH	PAH	PAH	PAH	PAH	PAH
31	IPAH	PAH	PAH	PAH	PAH	PAH
32	IPAH	PAH	PAH	Control	PAH	PAH
33	IPAH	PAH	Control	Control	PAH	Control
34	IPAH	PAH	PAH	PAH	PAH	PAH
35	IPAH	Control	PAH	Control	Control	Control
36	IPAH	Control	PAH	Control	PAH	PAH

The performance of each feature selection method on the validation set was also variable (Table 2.8). The cross validated performance for the training set can be seen in Table 2.9. The Random Forest model had the highest AUC (0.84), but the XGBoost model had a higher accuracy (0.83). The LASSO model had the poorest performance, with an accuracy of 0.72. The number of miRNAs selected by each method also differed, with LASSO selecting the most (13 miRNAs), and the Rpart model behaving more stringently by selecting just four miRNAs. The AUCs for models trained using a leave-one-out cross-validation approach showed similar results (Figure 2.6).

Table 2.8*: Model performance of four classifiers on the validation set; a random forest wrapper method (Boruta), regression partition trees (Rpart), LASSO, and extreme gradient boosting (XGBoost).*

	Random forest	**Rpart**	**LASSO**	**XGBoost**	**Ensemble**
miRNAs selected by model, n	10	4	13	8	20
Sensitivity (95% CI)	0.86 (0.65-0.97)	0.91 (0.71-0.99)	0.77 (0.55-0.92)	0.91 (0.71-0.99)	0.91 (0.71-0.99)
Specificity (95% CI)	0.71(0.42-0.92)	0.64 (0.35-0.87)	0.64 (0.35-0.87)	0.71 (0.42-0.92)	0.64 (0.35-0.87)
Positive predictive value (95% CI)	0.83 (0.61-0.95)	0.80 (0.59-0.93)	0.77 (0.55-0.92)	0.83 (0.63-0.95)	0.80 (0.59-0.93)
Negative predictive value (95% CI)	0.77 (0.46-0.95)	0.82 (0.48-0.92)	0.64 (0.35-0.86)	0.83 (0.52-0.98)	0.82 (0.48-0.92)
Correct classification rate (95% CI)	0.81 (0.64-0.92)	0.81 (0.64-0.92)	0.72 (0.55-0.86)	0.83 (0.67-0.94)	0.81 (0.64-0.92)
AUC (95% CI)	0.84 (0.69-1)	0.79 (0.63-0.95)	0.79 (0.63-0.94)	0.82 (0.66-0.99)	0.85 (0.70-1)

Table 2.9*: Mean 10 fold cross-validated performance on the training set regression partition trees (Rpart), a random forest wrapper method (boruta), LASSO, and extreme gradient boosting (XGBoost).*

	Random forest	**Rpart**	**LASSO**	**XGBoost**
Sensitivity	0.72	0.50	0.65	0.75
Specificity	0.93	0.64	0.83	0.88
Positive predictive value	0.91	0.53	0.79	0.85
Negative predictive value	0.85	0.66	0.79	0.86
Correct classification rate	0.85	0.58	0.76	0.83

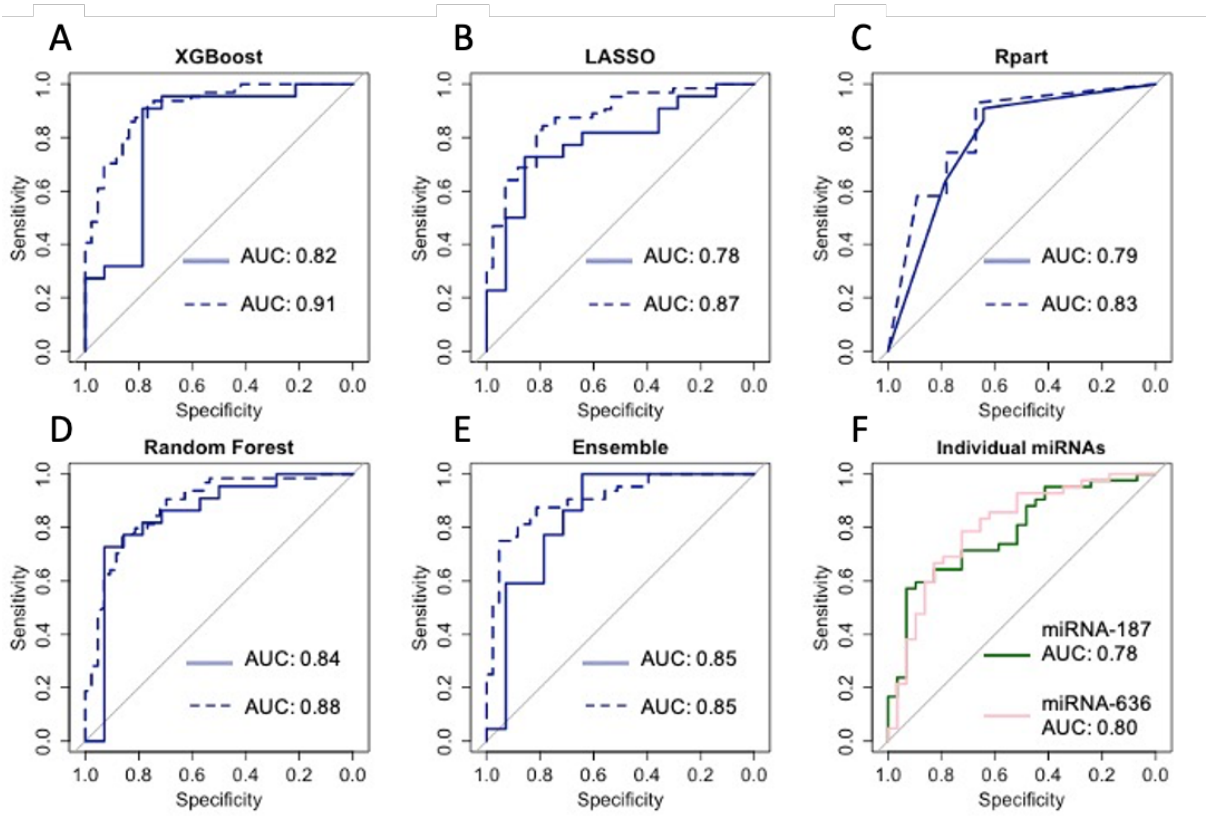


***Figure 2.6**: AUC for miRNA classifiers trained using LOOCV approach, and training / validation approach. Solid lines indicate ROC for the validation set (n = 35), where the model was trained on a separate set. Dashed lines indicate miRNA models trained using a leave-one-out cross validation approach across the whole data set. (A) extreme gradient boosting (XGBoost) utilising eight miRNAs; (B) LASSO utilising 13 miRNAs; (C) regression partition trees (Rpart) utilising four miRNAs; (D) a random forest wrapper method (Boruta) utilising 10 miRNAs; (E) Ensemble approach utilising 20 miRNAs; (F) Average cross validated ROC for miRNA-187-5p and miRNA-636 on the training set.*

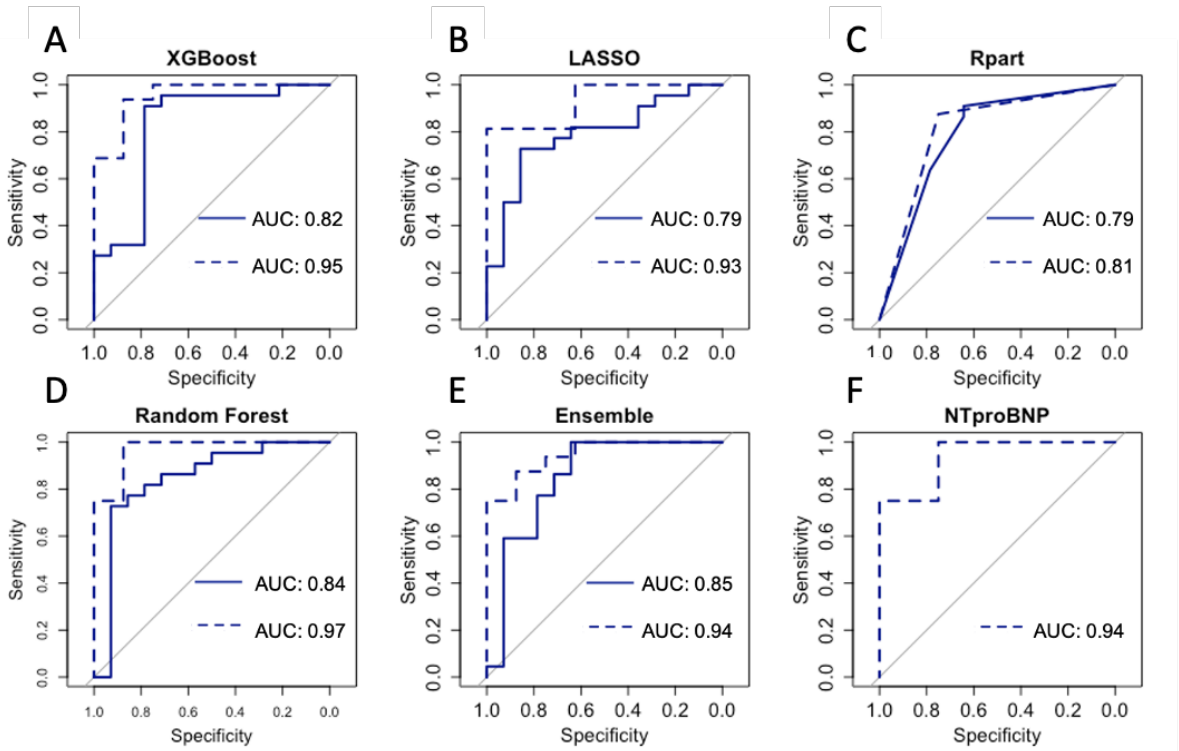


***Figure 2.7**: Solid lines indicate miRNA models, dashed lines indicate miRNA model + NT-proBNP. ROC curves for all four machine learning classifiers on the validation set, and NT-proBNP. (A) extreme gradient boosting (XGBoost) utilising eight miRNAs; (B) LASSO utilising 13 miRNAs; (C) regression partition trees (Rpart) utilising four miRNAs; (D) a random forest wrapper method (boruta) utilising 10 miRNAs; (E) Ensemble approach utilising 20 miRNAs; (F) NT-proBNP alone.*

As multivariable methods are known to select different candidate biomarkers, often with equal accuracy (He and Yu 2010), we focused on the overlapping miRNAs selected by the four different machine learning methods. From the 20 miRNAs selected across all four methods, seven miRNAs are found in more than one model, of these, two were selected by every model; miR-636 and miR-187-5p (Figure 2.5).

For a subset of patients from Sheffield, NT-proBNP levels were assayed at routine clinical appointments. We then used these to compare the models' performances when NT-proBNP levels were included (Figure 2.7). Although the best performing miRNA model (Random Forest) did not perform significantly different to the NT-proBNP classifier alone (miRNA AUC 95% CI = 0.69 - 1 vs NT-proBNP AUC 95% CI = 0.84 - 1), all miRNA models with NT-proBNP saw an improved performance with AUCs (Figure 2.7). Random forest increased from 0.84 to 0.97, rpart from 0.79 to 0.81, LASSO increased from 0.78 to 0.93, and the XGBoost model increased from 0.82 to 0.95, though not significantly larger according to the DeLong test. A clear association of miRNAs with PAH diagnosis may warrant future investigation of specific miRNAs for therapeutic intervention.

2.3.3. Importance of individual miRNAs in PAH classification

To check whether individual miRNAs selected could be used for classification, a univariable analysis was carried out on their expression values (Table 2.10). For each miRNA, the

expression levels of patients and controls were compared using a wilcoxon signed-rank test, then controlled for multiple testing using the Benjamini Hochberg correction (Benjamini and Hochberg 1995) at 0.05. The mean centred expression values for miRNAs selected by at least two feature selection methods can be seen in Figure 2.8a. Ten of the miRNAs identified in the feature selection methods had an adjusted p- value <0.05. We also looked at the univariate discriminatory power of each miRNA individually. MiR-187-5p had an accuracy of 0.78 on the validation set, whereas miR-636 had an accuracy of 0.69. To assess the potential impact of individual miRNAs on disease progression, we also looked at the survival difference in patients when stratifying them based on the median fitted risk of different miRNAs. However, no miRNA had a significant cox proportional hazard p-value (Table 2.11).

***Table 2.10**: Minimum, mean and maximum expression values for 43 miRNAs remaining when correlating miRNAs have been filtered out for the validation set, grouped by patients with pulmonary arterial hypertension, and healthy and disease controls. BH adjusted p-values for wilcoxon-signed rank tests.*

miRNA	PAH patients			Healthy and disease controls			Adjusted p-value
	Min	Mean	Max	Min	Mean	Max	
let-7d-3p	2.726	3.442	7.135	2.752	3.132	4.537	0.5237
miR-122-5p	2.618	3.064	8.362	2.597	2.888	3.971	0.5307
miR-1237-3p	2.674	3.472	4.685	3.106	3.779	4.533	0.0289
miR-1246	2.595	3.909	6.123	2.649	4.067	13.529	0.1827
miR-125a-5p	2.608	2.876	5.046	2.626	2.714	3.657	0.0428
miR-126-5p	2.654	2.956	3.686	2.713	2.911	3.354	0.7980
miR-1260a	5.628	7.489	9.274	5.949	7.532	8.466	0.9493
miR-1306-3p	2.585	2.976	4.329	2.581	2.806	3.545	0.0499
miR-148a-3p	2.581	2.829	3.771	2.612	2.691	2.998	0.3808
miR-148b-3p	2.589	2.782	3.579	2.585	2.69	3.143	0.1134
miR-150-5p	2.62	3.325	7.367	2.646	2.972	5.987	0.1335
miR-151a-3p	2.577	2.806	4.406	2.593	2.657	3.215	0.0237
miR-187-5p	2.473	2.822	5.34	2.453	2.558	3.024	<0.0001
miR-18b-5p	2.576	2.635	2.787	2.586	2.631	2.684	0.1522
miR-2116-3p	2.758	3.121	4.053	2.852	3.255	4.597	0.1522
miR-29b-3p	2.639	2.812	3.285	2.638	2.773	2.944	0.9493

miRNA	PAH patients			Healthy and disease controls			Adjusted p-value
	Min	Mean	Max	Min	Mean	Max	
miR-300	2.65	2.896	3.557	2.75	3.157	5.633	0.0057
miR-30c-5p	2.611	2.952	3.691	2.647	2.873	3.634	0.3808
miR-324-3p	2.931	3.712	4.997	2.816	3.517	4.56	0.1285
miR-326	2.624	2.809	3.209	2.654	2.781	3.353	0.4509
miR-33b-3p	2.728	3.169	4.454	2.898	3.324	4.471	0.1388
miR-34a-5p	2.682	2.892	3.478	2.649	2.798	3.087	0.1826
miR-3613-3p	2.882	3.881	8.708	3.455	4.394	6.831	0.0003
miR-374b-5p	2.545	2.769	3.496	2.6	2.701	3.565	0.1852
miR-376c-3p	2.524	2.707	3.811	2.546	2.611	2.762	0.2802
miR-423-3p	2.52	2.622	3.06	2.531	2.601	2.782	0.1285
miR-451a	6.925	10.525	14.237	7.365	11.645	14.218	0.0237
miR-4707-5p	2.499	2.862	3.648	2.482	2.638	2.836	0.0105
miR-484	2.707	3.117	3.821	2.765	3.067	3.849	0.1243
miR-494	2.642	3.961	8.154	2.637	4.426	8.353	0.4509
miR-548am-5p	2.688	2.997	4.141	2.701	2.908	3.563	0.9493
miR-550a-5p	2.567	2.804	3.454	2.66	2.861	3.473	0.5840
miR-572	2.541	4.645	7.474	2.62	4.203	5.936	0.0105
miR-574-5p	3.761	7.854	11.055	6.992	8.464	10.159	0.0946
miR-584-5p	2.606	3.487	4.924	2.66	3.839	5.977	0.8776
miR-636	2.721	3.269	4.954	2.737	3.95	5.509	<0.0001
miR-652-3p	2.498	2.627	3.596	2.506	2.543	2.737	0.0105
miR-664a-3p	2.759	3.283	5.719	2.869	3.294	3.693	0.1285
miR-671-3p	2.607	2.684	2.85	2.593	2.692	3.01	0.9493
miR-671-5p	2.492	4.682	9.917	2.538	3.699	9.023	0.0027

miR-675-3p	2.642	3.1	4.601	2.722	3.417	6.002	0.1243
miR-933	2.601	2.983	3.67	2.796	3.143	3.8	0.0104

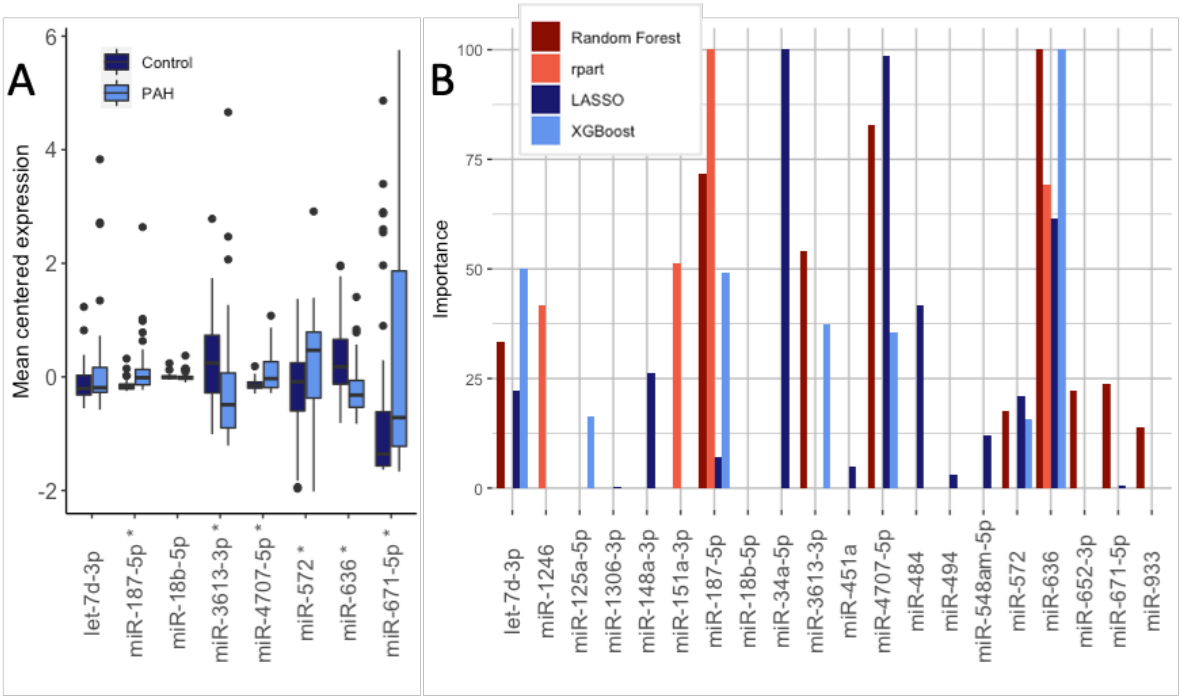


***Figure 2.8**: (A) Comparison of mean centred expression values for both training and validation groups (n = 107) of miRNAs for patients with pulmonary arterial hypertension (PAH) and no PH controls (Control) selected by two or more feature selection methods. * MicroRNAs with a significant difference between groups (adjusted p-value for Wilcoxon rank-sum test < 0.05). (B) Variable importance scores for the miRNAs selected by the feature selection methods, scaled between 0 - 100 per method.*

***Table 2.11**: Cox proportional hazard for miRNAs selected by a feature selection method*

microRNA	Regression beta coefficient	Wald statistic	P-value for miRNA	HR (95% CI for HR)
hsa-let-7d-3p	-0.0691	0.05	0.816	0.933 (0.521-1.67)
hsa-miR-1306-3p	0.755	3.17	0.0749	2.13 (0.927-4.88)
hsa-miR-148a-3p	-0.4	0.21	0.647	0.671 (0.121-3.71)
hsa-miR-187-5p	-0.37	0.39	0.535	0.691 (0.215-2.22)
hsa-miR-34a-5p	1.3	2.2	0.138	3.67 (0.657-20.5)
hsa-miR-451a	-0.0776	0.35	0.556	0.925 (0.715-1.2)
hsa-miR-4707-5p	0.257	0.21	0.646	1.29 (0.432-3.87)
hsa-miR-484	-1.53	2.16	0.142	0.216 (0.028-1.67)
hsa-miR-494	0.0832	0.26	0.608	1.09 (0.791-1.49)
hsa-miR-548am-5p	0.397	0.4	0.525	1.49 (0.437-5.06)
hsa-miR-572	0.167	0.57	0.451	1.18 (0.766-1.82)
hsa-miR-636	0.218	0.27	0.6	1.24 (0.551-2.81)
hsa-miR-671-5p	-0.0273	0.08	0.782	0.973 (0.802-1.18)
hsa-miR-18b-5p	-0.0098	0	0.998	0.99 (0.000646-1520)
hsa-miR-3613-3p	-0.365	1.67	0.196	0.694 (0.399-1.21)
hsa-miR-652-3p	-0.431	0.21	0.65	0.65 (0.101-4.18)
hsa-miR-933	-1.48	1.97	0.16	0.228 (0.029-1.79)
hsa-miR-151a-3p	-0.493	0.57	0.45	0.611 (0.17-2.2)
hsa-miR-1246	0.266	1.53	0.216	1.31 (0.856-1.99)
hsa-miR-125a-5p	-0.25	0.21	0.65	0.778 (0.264-2.3)

2.3.4 PAH classification performs similarly well using miRNA targets

Two miRNAs were identified by all four feature selection methods: miR-187-5p and miR-636. These miRNA were also ranked highest in a variable importance analysis (Figure 4B). In order to investigate the novel role these miRNAs play in PAH, we predicted their target genes. The two miRNAs had 20 predicted gene targets in common (*VAMP7, LMO3, DGKH, YTHDF3, DNAL1, PPP2R2A, ZDHHC15, UBN2, CDKN1B, FAM63B, PARP15, SOCS5, ZNF844, HECTD2, RIMS3, ZNF720, FCHO2, CBX5, PALM2, GABRB2*), with 630 targets in total.

Feature selection methods can be unstable when there are few samples for training. To counter this we verified the selected miRNAs gene targets in a previously published whole blood RNA seq data set (Rhodes et al. 2020), as well as two independent expression studies (Rajkumar et al. 2010; Zhao et al. 2014) .

The whole blood RNA seq data set contained 54 independent healthy controls and 347 PAH patients. Utilising the miRNA target gene set in this RNA seq data set (of which 548 target

genes were present), an XGBoost model was used to classify PAH from non-PH, using a cut-off of 0.841. We used XGBoost as a classifier, as the XGBoost model had the highest correct classification rate for the miRNA set. This produced a model with 0.86 AUC (95% CI 0.78-0.94), and an accuracy of 0.89 for the validation set. This classification model also allowed us to rank the genes contributing the most to the model. The top 15 gene targets are shown in Figure 2.9.

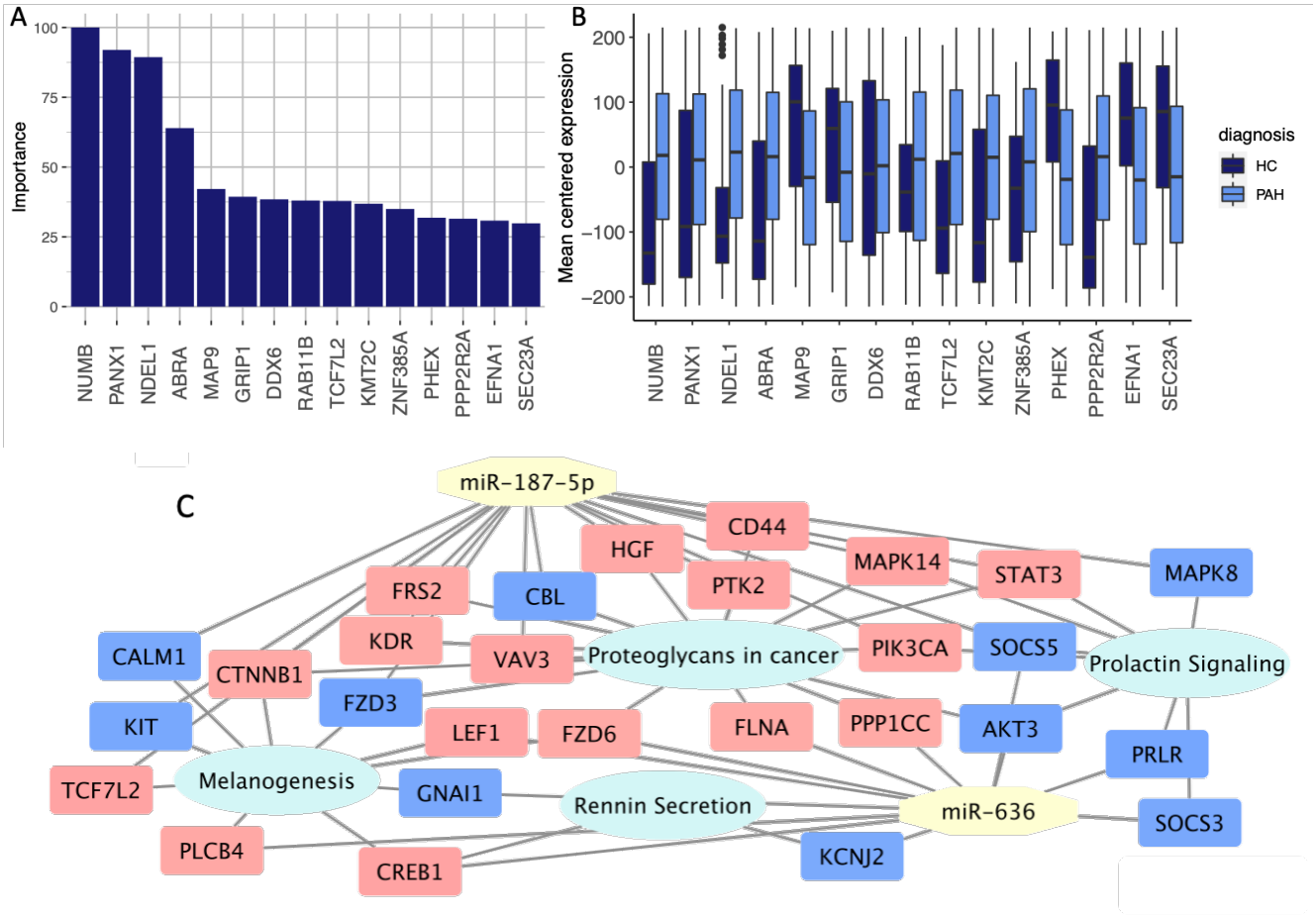


***Figure 2.9**: (A) Top 15 genes ranked with the highest importance in classifying patients in an RNAseq dataset (n = 401), scaled between 0 and 100. (B) Mean centred gene expression for top 15 genes (C) Significantly enriched KEGG pathways of the gene targets from miR-636 and miR-187-5p present in the validation RNA seq dataset. Down regulated genes in pink, up-regulated in blue.*

Three randomly selected groups of gene panels produced similar results, with AUCs of 0.83, 0.92 and 0.81. However, large numbers of genes in the 3 random models had a high correlation coefficient (> 0.7) with the top 15 genes driving the original XGBoost model (121, 269, and 254 genes) suggesting that a large volume of information was shared between models.

A model derived from genes with a correlation coefficient < 0.7 was also derived, with an AUC of 0.80 (95% CI 0.70-0.90). However, here again, connections can be drawn to the original model, with several of the top genes driving the model included in pathways enriched in the original model, such as prolactin signalling and rennin secretion.

We also developed 2 logistic regression models, the first using the top 15 genes which contributed the most to the XGBoost model of gene targets, and the second using 15 randomly selected genes which had a correlation coefficient < 0.7 to these 15 genes. The AIC from the first model was 538, compared to an AIC of 198 for the second model.

From the list of 630 target genes, 592 were found in at least one lung tissue dataset. GSE15197 contained 587 of the gene targets, with 281 found to be differentially expressed (adjusted p-value <0.05). All133 predicted gene targets that were profiled in GSE53408 were differentially expressed. Narrowing this down, 61 genes were differentially expressed in the same direction in both datasets. Basic characteristics of the two cohorts are described in Table 2.5. A pathway analysis of all 630 gene targets showed four enriched KEGG pathways: proteoglycans in cancer, rennin secretion, melanogenesis, and prolactin signalling pathway (Figure 2.9C). Widening the network to include miRNAs selected by at least two feature selection methods showed that of these miRNAs, miR-3613, miR-671 and miR-18b-5p also targeted genes from all of these pathways, with miR-572 targeting genes in the proteoglycans in cancer pathway.

From the pathways identified and putative links to PAH pathogenesis, seven gene targets (*FER, GLMN, PARP8, MTUS1, HGF, PELI1* and *UBR3)* were selected for qPCR validation based on putative links to PAH pathogenesis using four control human pulmonary artery smooth muscle cells (PASMC) and four with IPAH (Pickworth et al. 2017). Two genes in particular, *MTUS1* and *UBR3* showed a significant increase in expression in patient derived PASMCs compared to independent control cells (Figure 2.10). There were no significant differences in expression for the other genes.

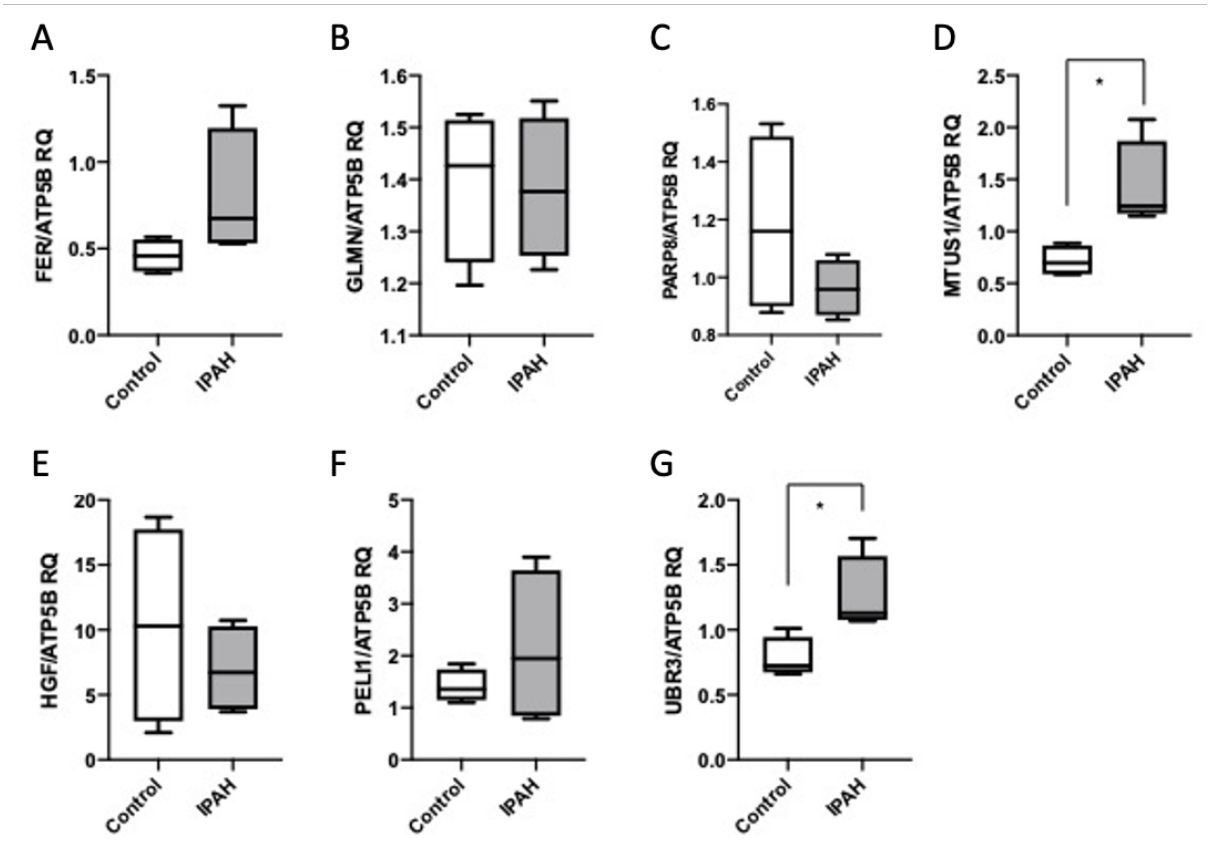


***Figure 2.10**: qPCR RQ relative quantification box plots for (A) FER, (B) GLMN, (C) PARP8, (D) MTUS1, (E) HGF, (F) PELI1, (G) UBR3*

2.3.5 Added value of miRNA targets (Cai Davies)

1053 genes were filtered from the whole blood RNA seq data set described above, with a starting set of well detected genes (described as genes with two or more reads in at least 95% of control or patient samples, n = 25966). These genes were used to generate two models in XGBoost. Model 1, contained 505 differentially expressed genes (described in (Rhodes et al. 2020). Model 2 contained genes from Model 1, and with the addition of the miRNA target genes (n = 548) created a larger feature set of 1053 genes. Both models had high AUCs (0.95) and a high sensitivity (Table 2.12).

***Table 2.12**: Performance of two XGBoost models classifying PAH from healthy controls in RNAseq in the validation set.*

	Sensitivity	**Specificity**	**AUC (95% CI)**
Model 1	0.98	0.63	0.95 (0.91-0.99)
Model 2	0.99	0.54	0.95 (0.90-0.99)

These models were then examined to look for genes with the highest importance to the XGBoost models, as measured by the percentage gain. The bar plot shows 144 genes with a percentage gain above 5% (Figure 2.11). The top 20 genes for each model were compared

against each other, with 12 of the same genes with the highest importance appearing in both models (Table 2.13).

***Figure 2.11**: Genes with a percentage gain >5% (n = 144) in Model 2 XGBoost classification model. Genes ordered by descending percentage gain. Plots in blue are RNA signatures and plots in red are microRNA gene targets. Created in GraphPad Prism by Cai Davies.*

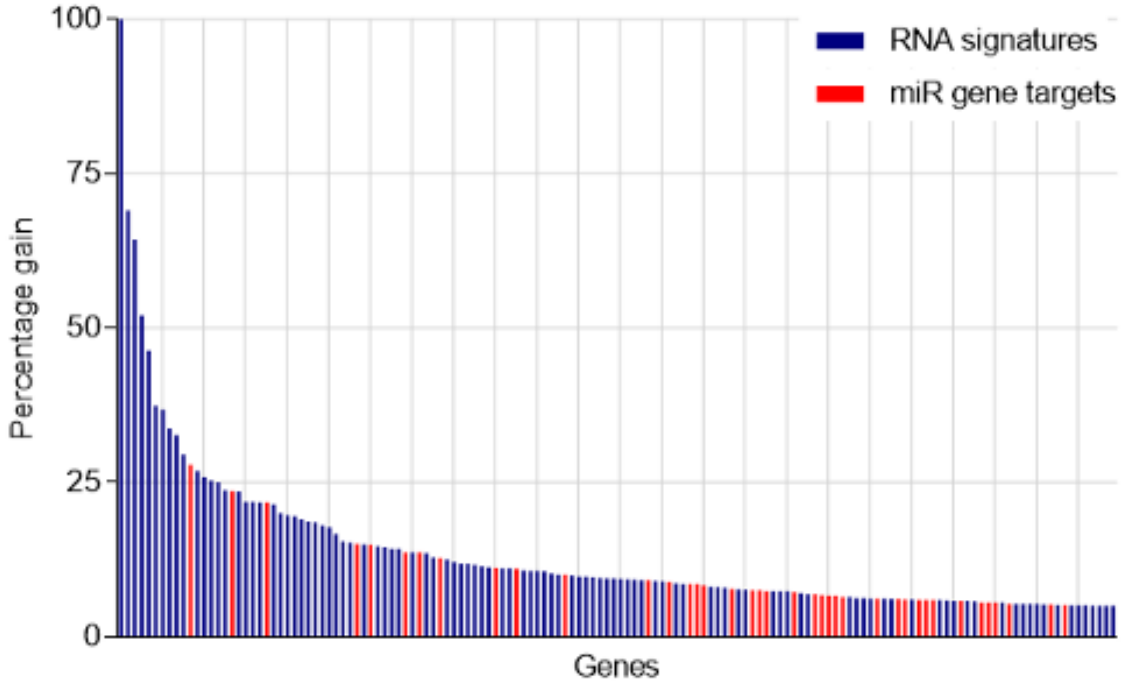


Table 2.13: *Genes that appeared within the top 20 importance of both XGBoost models. Their percentage gain (importance) is compared with genes showing an increase in gain highlighted in green with an increase in gain are termed genes of interest.*

Gene	Model one gain (%)	Model two gain (%)	Difference of gain (%)
MMP28	34.84	68.99	0.49
CPT1A	23.67	46.39	0.49
XKRX	41.82	64.30	0.35
ZNF763	22.62	33.67	0.33
RALA	17.42	25.36	0.31
NRG1	37.12	52.12	0.29
HLTF	25.51	32.67	0.22
KLF10	29.79	36.77	0.19
ANKRD34A	25.83	29.55	0.13
AC009299.4	100.00	100.00	0.00
FAM132B	28.38	23.74	-0.20
AC018890.6	26.41	21.85	-0.21

2.4 Discussion

There is increasing evidence that changes in miRNA expression levels are associated with progression of PAH. Here, we used miRNA expression profiles and a consensus machine learning approach to identify two consistently prioritised miRNAs with high accuracy at identifying PAH from no PH controls, as candidates for further investigation. We subsequently identified putative miRNA gene targets and integrated public lung tissue RNA datasets to validate differential regulation of key miRNA targeted genes, again identifying candidates for further investigation. An extreme gradient boosting method of classifying patients based on the putative gene targets in an overlapping cohort had a similar AUC, providing further

validation. This data suggests that combining different approaches for selecting miRNAs can reveal diagnostic biomarkers and insights into regulators of disease.

Of the supervised machine learning approaches we tested, we found that a random forest approach identified patients with PAH with the highest sensitivity, although an XGBoost approach had a similarly high AUC. Adding NT-proBNP to the random forest model resulted in a model with a higher classification accuracy compared to NT-proBNP alone. This shows NT-proBNP and miRNAs may provide complementary phenotypic information and therefore both should be incorporated in future prospective validation analyses.

It is important to consider whether the features selected at each point are true biomarkers or false positives. Machine learning provides an unbiased approach to predicting patient status, but also the potential to identify previously unknown interactions and identify novel biological features (Lopez-Rincon et al. 2019; Neumann et al. 2016). Our approach of investigating the biomarkers identified through multiple feature selection techniques increases confidence in the generation of reproducible biomarker panels, and reduces the number of miRNAs for potential clinical investigation. The selected miRNAs ranked highly in terms of variable importance (Figure 2.8B).

Both miRNAs selected have previously been linked to PAH. MiR-187 has previously been identified as significantly upregulated in endoarterial biopsy samples in a porcine model (A. Rothman et al. 2017), and in human lung tissue (W. Chen and Li 2017), in concordance with our findings. However, one study on cardiac tissue from the sugen5416 plus hypoxia rat model found miR-187-5p to be significantly downregulated (Joshi et al. 2016). MiR-636 has been reported to correlate with maximum change in pulmonary vascular resistance (PVR) in a small study on a paediatric PAH population (Kheyfets et al. 2017). The above literature reports support the evidence that miR-187-5p and miR-636, identified here as candidate biomarkers may be associated with disease progression of PAH providing validation that our machine learning approach identified miRNA biomarkers of relevance. Several other miRNAs identified as having a high importance score by the feature selection methods have also previously been seen in PAH, for example MiR-4707-5p has been identified as a potential target for PH (Jin et al. 2020). Additionally, miR-34 has been seen to have decreased expression in PAH (Alexander M. K. Rothman et al. 2016; K.-H. Chen et al. 2018), and let-7d, which has been identified as a potential biomarker for the presence and severity of PH in patients with SSc (Izumiya et al. 2015). Similarly, the target genes driving the classification in an independent RNAseq dataset, *TCF7L2,* which ranked highest in importance has previously been seen to be differentially expressed in the lung tissue of IPAH patients (Saygin et al. 2020) as well as in the cardiac muscle tissue in a rat model (Hołda et al. 2020). Some of these target genes also showed weak to moderate correlation with available clinical features, such as lung function forced vital capacity (Table 2.14).

***Table 2.14**: Spearman's correlation coefficients for gene targets with sample demographics. Highest correlation coefficient reported.*

Clinical Variables	Gene	Absolute max correlation coefficient
Demographics		
Age at Sampling	*EFNA1*	-0.3044634
Survival		
Survival (years since sampling).	*BRWD1*	-0.2952542
Died during Follow up	*KAT2A*	-0.2616826
HLA-DPA1/DPB1 SNP (rs28568300)	*MSI2*	-0.228825
Lung Function		
Forced Vital Capacity (FVC, L)	*KDM6A*	-0.4217336
Forced Expiratory Volume in 1 second (FEV, L)	*KDM6A*	-0.3891385
Carbon monoxide transfer (KCO %Pred)	*NET1*	0.2068199
Right Heart Catheter		
Mean Right Atrial Pressure (mm Hg)	*HDGF*	0.2827278
Pulmonary Arterial Wedge Pressure	*NDRG4*	-0.2265114
Pulmonary Vascular Resistance	*BTBD3*	-0.2398163
Cardiac Index	*BTBD3*	0.2597571
Exercise Capacity		
6 minute walk distance (6MWD,m)	*HDGF*	-0.2896662
Incremental Shuttle Walk Distance (ISWD,m)	*HDGF*	-0.2885012

Our main aim in this study was to investigate the relationship between miRNAs and clinical classifications, not to develop a diagnostic tool. ML methods can capture more complex, non-linear relationships, where a straightforward univariable analysis cannot. A limitation to this study is the relatively small sample size used to both generate and validate the miRNAs as classifiers. This may have resulted in some model overfitting and therefore a possible overestimation of effect size. In order to mitigate this, we validated the gene targets in separate published datasets, and used qPCR to validate potentially interesting genes. The target gene data contained a far larger number of variables, with 548 genes for each of the 401 subjects, necessitating our use of ML in this dataset. As a result, future studies based on larger

retrospective and prospective clinical cohorts are warranted, and currently underway (ClinicalTrials.gov NCT04193046) to corroborate the utility of these, and potentially other miRNAs as classifiers and biomarkers. In such a small cohort, there was a danger the models could have been driven by factors such as age and sex, but classification using only these factors yielded an accuracy of 0.57 in the validation set. We also noted that the AUC confidence intervals for males and females on the training and validation sets overlapped. Additionally, both SSc and PAH, as individual diseases can be heterogeneous (Launay et al. 2017). As such within our cohorts of mixed IPAH and SSc-PAH there are likely to be variations between patients, and equally, our control group included 10 disease controls and 33 healthy controls. We also attempted a leave one out cross validation approach across the whole dataset, which resulted in similar miRNAs being selected (Figure 2.5). These mixed groups likely reduce the risk of overfitting to a specific patient phenotype, and increase the chance that this analysis could be replicated in other PAH cohorts.

The two candidate miRNAs selected from the microarray study have not been further quantified by PCR. However, correlations between miRNA microarray expression and PCR have been shown to have very high correlation coefficients (Pradervand et al. 2009). Consequently, further validation of the two miRNAs identified in a larger, independent cohort are necessary before a clinical application can be considered.

In summary, our approach using four machine learning feature selection algorithms on miRNA data from microarrays identified a two miR-signature for PAH from patient plasma. These circulating miRNAs, and their target genes may provide a novel PAH signature, reveal novel disease mechanisms and highlight future putative drug targets.

Chapter 3: Diagnostic miRNA signatures for treatable forms of pulmonary hypertension highlight challenges with the current clinical classification

Following the results of successfully classifying patients with PAH from controls using a small panel of miRNAs from microarray profiling, I hypothesised that this classification approach could scale up and I examined a large cohort of patients with all types of PH, along with disease controls using a large panel of miRNAs profiled by qPCR. I became involved in a miRNA biomarker profiling study with Janssen pharmaceuticals, in collaboration with MiRXES, University of Sheffield, The Royal Papworth Hospital (Cambridge), and Imperial College London. This study aimed to identify different microRNA signatures or biomarkers associated with the different diseases profiled. It is the largest retrospective study of this kind laying the groundwork for the prospective clinical trial CIPHER (ClinicalTrials.gov Identifier: NCT04193046) by identifying signatures that can predict disease types. Chapters 3 and 4 form a manuscript under preparation for submission to the European Heart Journal.

3.1 Introduction

The heterogenous cardiopulmonary condition pulmonary hypertension, as defined by an at rest mPAP > 20mmHg. PH is associated with reduced life expectancy. Patients often present with generic symptoms such as shortness of breath, and diagnosis along with referral to a specialist is often delayed. Since 2004, the World Health Organisation (WHO) has categorised PH into five subgroups, these remain the current clinical classifications (PAH, PH due to left heart disease (PH-LHD), PH due to lung disease (PH-lung), CTEPH and miscellaneous PH) (Simonneau et al. 2019) (Table 1.1). Although the diagnosis requires a right heart catheterisation for confirmation, patients may be triaged with echocardiography and BNP or NT-proBNP. An echocardiography requires specialist interpretation to offer an estimate of mPAP. BNP plasma concentration levels offer a potential alternative. BNP or the prohormone form (NT-proBNP) are the only circulating biomarkers adopted by ERS/ESC guidelines for route clinical use. However, both BNP and NT-proBNP are measures of cardiac stress, and neither differentiate between different underlying causes. As biomarkers, BNP and NT-proBNP are not perfect, with limited sensitivity, even in heart failure (Shah 2022).

MicroRNAs have shown promise as biomarkers in a variety of diseases. MicroRNAs are well preserved in blood plasma, and both more stable and more easily measured than some alternatives, such as proteins (Pritchard, Cheng, and Tewari 2012). In Chapter 2, microarrays were used to profile the miRNAs. Microarrays were one of the first approaches used to analyse miRNAs in large numbers, with the advantage of being less expensive than some other methods, whilst allowing for relatively large numbers to be measured in parallel. However, the quantification of microarrays is restricted to a linear range, and the specificity can be low for miRNAs with closely related sequences (Pritchard, Cheng, and Tewari 2012), and they are sometimes unable to detect low levels of miRNAs. Here, MiRXES's qPCR miRNA assay

technology was used to profile the miRNAs. qRT-PCR is known as the ‘gold standard’ for nucleic acid quantification, with high sensitivity and specificity (Pritchard, Cheng, and Tewari 2012). MicroRNA levels change with disease and may offer an alternative or be additive to BNP as a blood test to diagnose and risk stratify patients. Since miRNAs have a more diverse cellular origin than BNP, we hypothesised that the distribution of circulating miRNAs across the different presentations of PH would inform molecular endotypes in a PH cohort.

In the UK PAH diagnosis is made at specialist PH referral centres, where an important road-block to rapid diagnosis is the identification of patients at highest-risk of PH. As such, the primary objective of this study was to use the existing UK cohorts to identify circulating miRNA as biomarkers using MiRXES’s qPCR miRNA assay technology, and investigate the potential for developing miRNA signatures to identify PH and PAH from DC. Identifying patients with PAH or CTEPH from other forms of PH would be of clinical use, so this study also aimed to identify miRNA signatures to separate patients with PAH from other forms of PH, and patients with CTEPH from other forms of PH.

3.2 Methods

3.2.1 Sample collection

This comprised 1150 patients with PH and 334 disease controls as summarised in Table 3.1. Patients were recruited from 3 UK national PH referral centres at the Hammersmith Hospital (Imperial), Royal Hallamshire Hospital (Sheffield) and Royal Papworth Hospital (Cambridge) as summarised in Table 3.1. All cases were diagnosed between 2008 and 2019 using contemporaneous diagnostic guidelines (Galiè et al. 2015). All samples were obtained following informed consent to one of three cohorts: the Imperial College Prospective Study of Patients with Pulmonary Vascular Disease cohort (PPVD, UK REC Ref 17/LO/0563), the Sheffield Teaching Hospitals observational study of pulmonary hypertension, cardiovascular and other respiratory diseases (STH-ObS, UK REC Ref 18/YH/0441) or Papworth cohort. All samples were collected as per local standard operating procedures and stored at -80oC until assayed. All cases/samples were pre-processed into training, interim and validation datasets to balance age, sex, PH classification and recruitment site. The validation samples were analysed separately.

***Table 3.1**: Demographics for the training, interim and validation cohorts. Normally distributed variables reported as mean (standard deviation), not normally distributed variables reported as median [IQR]. Categorical variables reported as number (% from reported total of column).*

Clinical Variable	Missing	Training	Interim	Validation	All
n		952	185	347	1484
Sex: Female	2 (0.1%)	578 (60.7%)	124 (67.0%)	212 (61.4%)	914 (61.7%)
Age (years)	5 (0.3%)	64.0 [21.8]	64.0 [23.0]	65.5 [20.0]	64.0 [21.0]
Body Mass Index	106 (7.1%)	27.8 [8.6]	27.4 [8.9]	28.0 [8.3]	27.8 [8.6]
Blood pressure - diastolic (mm Hg)	162 (10.9%)	74 [14]	74 [16]	76 [16]	75 [15]
Blood pressure - systolic (mm Hg)	0 (0%)	129 [28]	128 [30]	130 [29]	129 [29]
Pulmonary Vascular Resistance (dynes)	403 (27.2%)	480 [640]	590 [560]	410 [549]	480 [621]
Mean pulmonary artery pressure (mm Hg)	233 (15.7%)	42.0 [24.0]	42.5 [20.8]	37.0 [26.0]	41.0 [25.0]
Cardiac Output (L/min)	295 (19.8%)	4.1 [2.1]	4.1 [2.1]	4.3 [2.0]	4.1 [2.1]
Cardiac Index	421 (28.4%)	2.2 [1.0]	2.2 [1.1]	2.2 [1.1]	2.2 [1.0]
Functional Class	41 (2.8%)				
I		34 (3.7%)	6 (3.3%)	13 (3.9%)	53 (3.7%)
II		185 (20.0%)	35 (19.3%)	68 (20.4%)	288 (20.0%)
III		632 (68.1%)	123 (68.0%	230 (68.9%)	985 (68.3%)
IV		77 (8.3%)	17 (9.4%)	23 (6.9%)	117 (8.1%)
Plasma NT-proBNP (log2 pg/ml)	0 (0%)	9.4 [4.0]	9.4 [4.0]	9.0 [3.4]	9.2 [3.8]
PH Treatment naive	0 (0%)	730 (76.7%)	131 (70.8%)	295 (85.0%)	1156 (77.9%)

Clinical Variable	Missing	Training	Interim	Validation	All
Site	0 (0%)				
Cambridge		295 (31.0%)	39 (21.1%)	104 (30.0%)	438 (28.3%)
Imperial		440 (46.2%)	95 (51.4%)	141 (40.6%)	676 (43.6%)
Sheffield		217 (22.8%)	51 (27.6%)	102 (29.4%)	435 (28.1%)
Diagnosis	0 (0%)				
CTED		53 (5.56%)	11 (5.95%)	17 (4.90%)	81 (5.46%)
Symptomatic but no PH		154 (16.2%)	20 (10.8%)	79 (22.8%)	253 (17.0%)
Idiopathic PAH		188 (19.7%)	41 (22.1%)	51 (15.0%)	280 (18.9%)
Heritable PAH		8 (0.840%)	4 (2.16%)	6 (1.73%)	18 (1.21%)
Drug and toxin induced PAH		4 (0.420%)	0 (0%)	1 (0.288%)	5 (0.337%)
Associated PAH		115 (12.1%)	34 (18.4%)	39 (11.2%)	188 (12.7%)
PH-LHD		119 (12.5%)	18 (9.73%)	41 (11.8%)	178 (12.0%)
PH-Lung		72 (7.56%)	11 (5.95%)	34 (9.80%)	117 (7.88%)
CTEPH		215 (22.6%)	43 (23.2%)	74 (21.3%)	332 (22.3%)
Misc PH		24 (2.62%)	3 (1.62%)	5 (1.44%)	32 (2.16%)

3.2.2 Quantification of serum NT-proBNP and miRNAs (Performed by MiRXES)

Total RNA was extracted from 200 µl of serum or plasma using the Maxwell® RSC miRNA Plasma and Serum Kit (Promega, Madison, USA) as per the manufacturer's recommendations with two modifications: (a) three proprietary spike-in controls (with 20 nucleotide unique RNA sequences) signifying low, medium and high RNA levels (MiRXES, Singapore) were added to

lysis buffer C prior to sample RNA isolation. These controls are used to track the effectiveness of RNA isolation and normalise for technical variations; (b) bacteriophage MS2 RNA (Roche, Basel, Switzerland) was added at 0.4ng per sample isolation to improve RNA isolation yield. For biomarker discovery, a highly-controlled RT-qPCR workflow was used to quantify the expression of miRNA in each sample. Isolated RNA was reverse transcribed using miRNA-specific reverse transcription (RT) primers according to manufacturer's instructions (ID3EAL Customised Individual miRNA RT Primer, MiRXES) on QuantStudio™ 5 Real-Time PCR System (Applied Biosystems, Foster City, CA, USA).

3.2.3 Pre-processing of miRNA expression data (Performed by MiRXES & Chris Rhodes)

We measured expression in 590 miRNAs for the discovery and interim cohorts and 359 detectable miRNAs were measured in the validation cohort. Data from 326 miRNAs that were detected in no less than 90% of samples in both cohorts were combined. Missing values were imputed separately in the combined discovery and interim, and validation sets , by replacing missing values with miRNA mean – four standard deviations. miRNA data were further global normalised (A novel and universal method for microRNA RT-qPCR data normalisation). The pre-processing was carried out by MiRXES. Samples from Cambridge showed higher total miRNA counts than the other centres. To correct for this batch effect, total miRNA counts were modelled with a LASSO model composed of 11 miRNAs. A linear regression using this model was then used to adjust the counts, retaining the mean miRNA levels.

3.2.4 Classification of patients into PH subtypes

We attempted seven different classifications, firstly PH vs DC, and PAH vs DC. Then each of the PH subtypes vs the other PH groups (Table 1.1); PAH vs other PH types, PH-LDH vs other PH types, PH-lung vs other PH types, CTEPH vs other PH types. Finally, we looked at PAH vs CTEPH, the two treatable subtypes of PH. For each classification, the dataset was initially split into predetermined training, interim and validation subsets. All statistical analyses were carried out using R (v4.0.3). Following the approach used in Chapter 2, we used four different machine learning feature selection methods, each fed forward to binary classifiers. For each method, parameters were tuned using 10-fold cross validation (repeated 10 times). Weights were also added to each classifier. Each patient was weighted as follows:

Control in each comparison: $1/(Number\ of\ controls) * 0.5$

Targeted group in each comparison: $1/(Number\ of\ targets) * 0.5$

Each classifier was then assessed for AUC using the interim set. The models with the best cross-validated AUC on the discovery set, and highest AUC on the interim set were refined, using the combined discovery and interim sets, with the best performing method (assessed as the highest mean cross-validated AUC on the combined sets) was selected as the final model. Final performance was then assessed on the validation set. Overview of classifier training and testing is described in Figure 3.1.

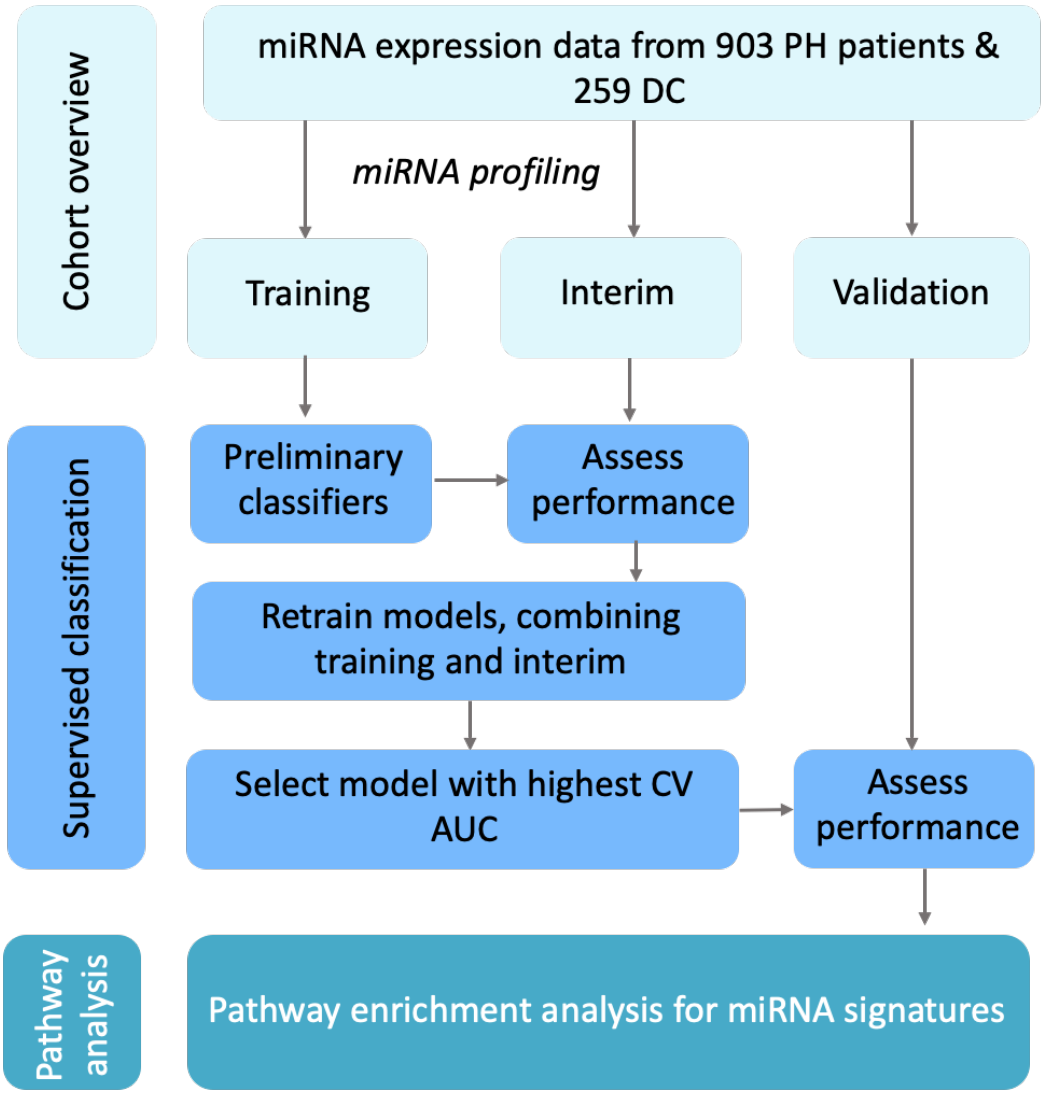


***Figure 3.1**: Method overview. Data set split was into predetermined training, interim and validation sections. Four machine learning methods used to classify patients in the training set, then refined using the combined training and interim. Final performance assessed on validation cohort before pathway enrichment for miRNAs in signatures.*

3.2.4.1 Boruta and Random Forest

Boruta, a feature selection wrapper algorithm based on random forest was utilised to detect potentially appropriate features within the classification framework. We implemented 300 repetitions with default settings of the normalised permutation importance function using random forest to get variable importance, within the Boruta package (v.7.0.0). Once the 300 runs were completed, miRNAs not classified as important by the algorithm were rejected. This implementation was carried out 100 times, selecting the miRNAs present in at least 99 of the 100 repetitions. These miRNAs were utilised to create a random forest model (randomForest package v4.6-14). The caret package (v. 6.0-88) was used to identify the ideal number of trees, from 1000, 1500, 2000 or 2500. The number of variables available for splitting at each tree node was optimised next, across the range 1 to 15 (Table 3.2). Once a model was created, if the original number of variables selected was greater than 10, the model was re-tuned by removing the variables with the lowest contributed importance to the model.

***Table 3.2**: Final Random forest model parameters for each comparison selected from cross validated folds on the training and interim sets. *Random forest not the final model for this comparison*

Comparison	Optimal number of trees	Number of variables available for splitting at each tree node
PH vs DC*	1000	1
PAH vs DC*	1000	9
PAH vs other PH*	1000	1
PH-LHD vs other PH	1000	5
PH-lung vs other PH	1000	3
CTEPH vs other PH*	1000	2
PAH vs CTEPH*	1000	1

3.2.4.2 Recursive partition trees

Classification trees were computed using Rpart (v4.1-15)(Therneau and Atkinson 2018) and caret in R. Trees were generated from the root of the tree downwards, using a greedy feature selection algorithm and recursive binary splitting to return features in order. The tree construction was controlled by setting the minimum number of observations in a terminal node to 5 and the minimum number of observations in a node for a split to occur set to 4. The Gini index was minimised to create each split. A variable was deemed to be selected if it appeared in the final model.

3.2.4.3 LASSO

Least absolute shrinkage and selection operator (LASSO) using the glmnet package in R (v4.1-2) was used to choose pertinent miRNAs by eliminating variables with a coefficient shrunk to 0. The regularisation parameter lambda, (λ), was chosen using binomial deviance, using repeated 10-fold cross-validations. The value of λ with minimum binomial deviance was selected and used to fit the final model.

3.2.4.4 XGBoost

Finally, we looked at XGBoost (v1.4.1.1), a gradient boosting method which ranks features in order of importance. The parameters were tuned over previously described optimisation

ranges for miRNAs (Errington et al. 2021). Once models had been trained, the variables were ranked in terms of their importance. The top nine miRNAs were then selected, and the model re-trained over the same parameter range using only the nine selected miRNAs. The optimal values for these parameters for each model can be seen in Table 3.3 below.

***Table 3.3**: Final XGBoost model parameters for each comparison. *XGBoost not the final model for this comparison*

Comparison	No of trees	Max tree depth	Learning rate	Gamma	% feature used in each boost (column sampling)	Min child weight	Subsample rate (row sampling)
PH vs DC	150	6	0.05	0.1	1	1	1
PAH vs DC	1500	2	0.01	0	1	2	1
PAH vs other PH	3950	1	0.025	0	1	1	1
PH-LHD vs other PH*	1200	8	0.05	0.7	0.8	1	1
PH-lung vs other PH*	2400	10	0.01	0	1	1	0.75
CTEPH vs other PH	200	3	0.1	0	1	5	1
PAH vs CTEPH	2550	1	0.01	0.05	0.6	2	0.75

3.2.4.5 NT-proBNP

We also investigated the performance of NT-proBNP as a standalone variable for classification. The glm function and caret (v. 6.0-88) package in R were used to build logistic regression models for each comparison, using log 2 values for NT-proBNP, setting the family to binomial. Weights were added to each patient as above.

3.2.5 Pathway Enrichment Analysis

An over representation analysis was carried out for each miRNA panel using the GeneTrail 3.2 miRNomics platform. The gene pathway resources miRTarBase, REACTOME and WIKIPATHWAYS from the miRPathDB 2.0 collection were selected as the databases on which the enrichment analysis was performed. A significance level of 0.05 was used for FDR adjusted results (Benjamini and Yetutieli 2001).

3.2.6 Validation in an external cohort

A separate validation dataset cohort of patients from the Brigham and Women's Hospital (BWH) in the US had miRNA plasma levels measured using the same MiRXES platform (Table 3.4). The cohort contained 158 disease controls and 55 patients with PAH. We compared these to 118 samples from Sheffield who had also had plasma miRNA levels measured (76 patients with PH, 22 disease controls and 20 healthy controls). Correlation between serum and plasma levels for matched samples were examined using a spearman's correlation. Due to poor correlation between serum and plasma in the matched samples from Sheffield, we fitted logistic regression (LR) models using the same miRNAs selected by the XGBoost models from our discovery cohort (training and interim) on mean centred data. Only the distribution of PAH and DC patients were comparable in the discovery cohort and this external cohort, so we assessed the models on differentiating PAH from DC using the same set of miRNAs.

Table 3.4: Patient demographics for the Brigham and Women's Hospital cohort. Normally distributed variables are reported as mean (standard deviation), and non-normally distributed variables are reported as median [IQR]. Categorical variables reported as number (% from reported total of column). ERA, endothelin receptor antagonists; PDE5, phosphodiesterase-5 inhibitors.

Clinical Variable	**Missing**	**Disease Controls**	**PAH**
N		157	54
Sex: Female	0 (0%)	111 (70.7%)	37 (68.5%)
Age (years)	0 (0%)	56.0 [21.0]	67.5 [14.8]
Body mass index (kg/m2)	5 (2.4%)	25.0 [7.2]	31.3 [8.3]
Systemic blood pressure - diastolic (mmHg)	0 (0%)	70.0 [16.0]	70.0 [18.0]
Systemic blood pressure - systolic (mmHg)	0 (0%)	124 [23.0]	137 [22.8]
Mean pulmonary artery pressure (mmHg)	0 (0%)	15.0 [4.0]	33.0 [14.5]
Pulmonary Arterial Wedge Pressure (mmHg)	156 (72.9%)	8 [2.75]	12.0 [7.50]
Cardiac output (L/min)	0 (0%)	5.08 [1.28]	4.93 [1.79]

Cardiac index (L/min/m2)	0 (0%)	2.78 [0.62]	2.54 [0.73]
Pulmonary vascular resistance (dynes.s.cm-5)	0 (0%)	94.0 [49.0]	328 [126]
WHO functional class	20 (9.5%)		
I		38 (27.3%)	2 (3.8%)
II		76 (54.7%)	23 (44.2%)
III		24 (17.3%)	23 (44.2%)
IV		1 (0.7%)	4 (7.7%)
NT-proBNP	124 (58.8%)	105 [228]	256 [810]
PH treatment-naive	0 (0%)	154 (98.1%)	39 (72.2%)
ERA		1 (0.6%)	3 (5.5%)
Prostanoid		0 (0%)	4 (7.4%)
PDE5		2 (1.3%)	14 (25.9%)
Other PH drug		0 (0%)	1 (1.9%)

3.2.7 Code availability

All methods and parameters used are described in the git repository available at https://github.com/niamherrington/MiRXES-miRNA.

3.3 Results

3.3.1 Model performance

For most comparisons, an XGBoost model had the highest cross-validated AUC across the training and interim sets. Distinguishing PH-LHD and PH-lung from other forms of PH proved to be an exception, with the random forest models performing best.

The performance, measured by area-under ROC (AUC), of miRNAs and NT-proBNP in distinguishing PH from disease controls (DC), which includes all symptomatic patients (including CTED) in which PH was excluded by cardiac catheterization, was first derived from training and interim samples, then a final check made by evaluating the AUC in the validation data set (Table 3.5, Figure 3.2).

There was no significant difference in the overall performance between NT-proBNP and the miRNA signature in distinguishing patients with PH or PAH from symptomatic disease controls (All PH vs DC, PAH vs DC, Table). However, the miRNAs showed superiority in distinguishing between treatable subtypes of PH from other types. CTEPH vs other PH and PAH vs other PAH had AUCs of 0.69 and 0.70 respectively (compared to 0.51 and 0.58 for NT-proBNP). Additionally, the miRNAs also performed better than NT-proBNP at distinguishing PAH from CTEPH (AUCs of 0.77 compared to 0.55 for NT-proBNP). No performance was analysed on samples from patients with PH-miscellaneous due to the small sample size (n = 32 training, interim and validation combined). ROC curves are shown in Figure 3.2.

Table 3.5: *Performance of miRNA signatures in training and validation datasets. Mean cross validated AUCs on discovery and interim datasets, and AUC on validation set for the best performing models trained on miRNAs and NT-proBNP across five clinically defined classes (Pulmonary Hypertension groups 1-4 and Disease Control). *P value for DeLong test of miRNA and NT-proBNP models on the validation set.*

Comparison	**miRNA AUC mean CV (sd)**	**NT-proBNP AUC mean CV (sd)**	**Validation miRNA AUC (95%CI)**	**Validation NT-proBNP AUC (95% CI)**	**DeLong test p value***	**Machine Learning Model**
All PH vs DC	0.75 (0.05)	0.78 (0.05)	0.70 (0.64-0.76)	0.78 (0.73-0.84)	0.00379	XGBoost
PAH vs DC	0.82 (0.04)	0.76 (0.06)	0.73 (0.66 – 0.80)	0.79 (0.72 – 0.85)	0.193	XGBoost
PAH vs other PH	0.71 (0.06)	0.58 (0.05)	0.69 (0.62-0.75)	0.51 (0.44-0.59)	1.33e-04	XGBoost
PH-LHD vs other PH	0.70 (0.06)	0.56 (0.06)	0.67 (0.60 – 0.74)	0.57 (0.49 - 0.64)	0.0482	RF
PH-lung vs other PH	0.69 (0.06)	0.54 (0.05)	0.65 (0.57 - 0.72)	0.53 (0.45 - 0.61)	0.0522	RF
CTEPH vs other PH	0.70 (0.06)	0.54 (0.04)	0.70 (0.63-0.78)	0.58 (0.50-0.66)	0.0141	XGBoost
PAH vs CTEPH	0.74 (0.06)	0.55 (0.05)	0.77 (0.70 - 0.84)	0.55 (0.46-0.64)	2.13e-07	XGBoost

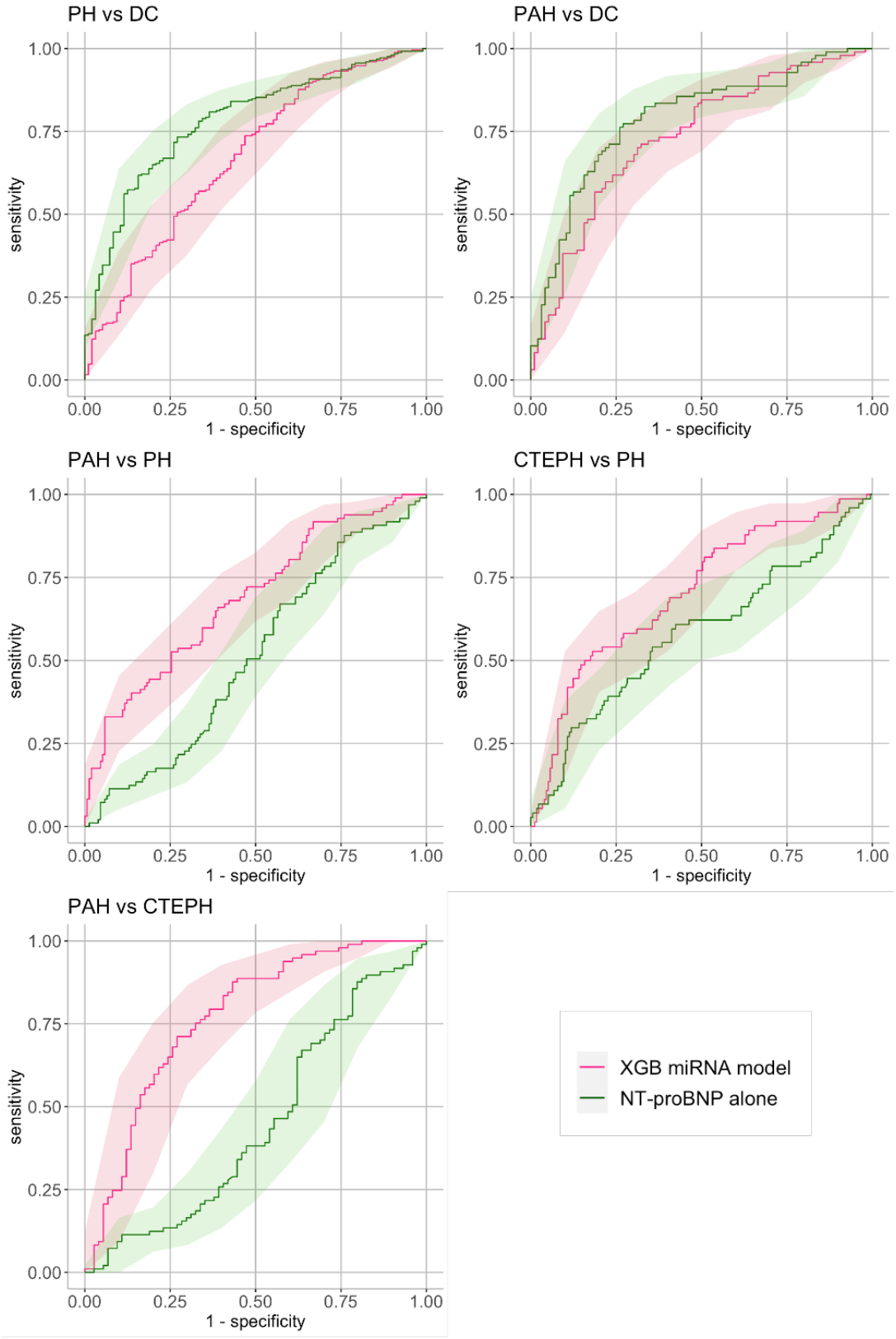


***Figure 3.2**: Performance of miRNA signatures in validation data set. Graphs show receiver operator characteristic curves for the performance of miRAN signatures and NT-proBNP for selected comparisons. Pink lines are XGBoost models from miRNAs, green shows a logistic regression for NT-proBNP. Shaded areas show 95% confidence intervals*

Within each classifier, there appear to be groups of patients for which the miRNA signatures are not able to accurately classify patients. We reasoned there may be a degree of overlapping pathology within these patients, and examined the occurrence of the PAH miRNA signature within other PH groups as an example. Using XGBoost, the PAH miRNA signature, detected in 61% of patents clinically classified as PAH, is found in 5% of PH-LVD and 12% of PH-lung and 10% of CTEPH patients (Figure 3.3). The other 3 machine learning models gave similar results (Figure 3.4)

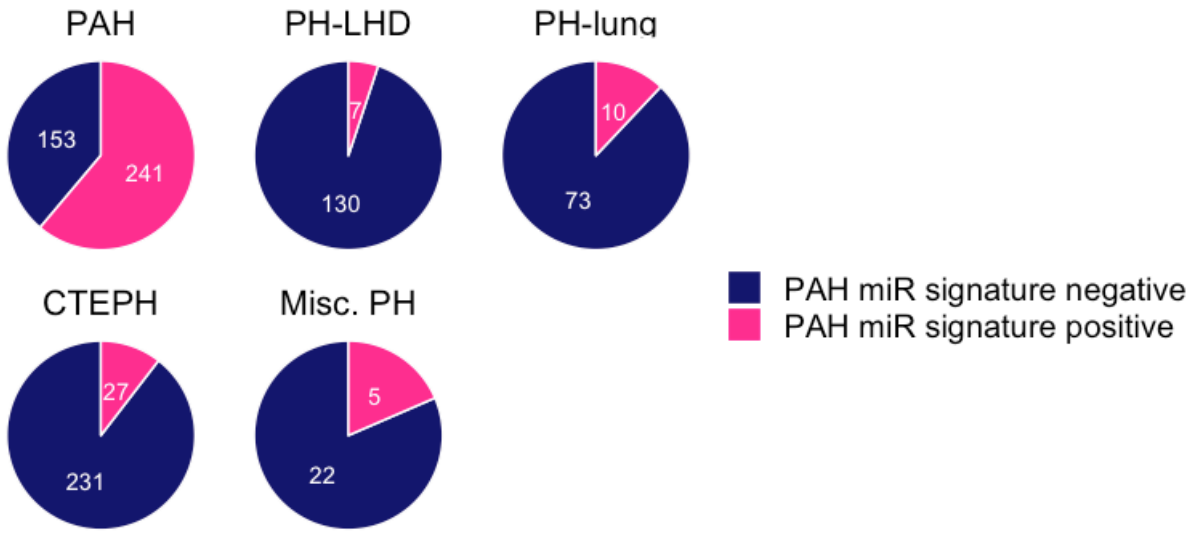


***Figure 3.3**: Percentage of each Dana Point Classification group identified by the PAH miRNA signature. Each pie chart demonstrates the breakdown of the number of patients incorrectly classified with the XGBoost derived PAH miR signature within each clinical PH classification group, defined as the majority vote across all training and interim CV folds. In the PAH cohort, the pink colour indicates the patient was correctly identified, and the blue represents the patients who were incorrectly classified. For PH-LHD, PH-lung, CTEPH and miscellaneous PH (Misc. PH), the pink represents the patients incorrectly classified as PAH by the miRNA signature, and the blue indicates the miRNA signature correctly identified the patient did not have PAH. Numbers represent patients within each classification group identified by PAH miRNA signature, or not.*

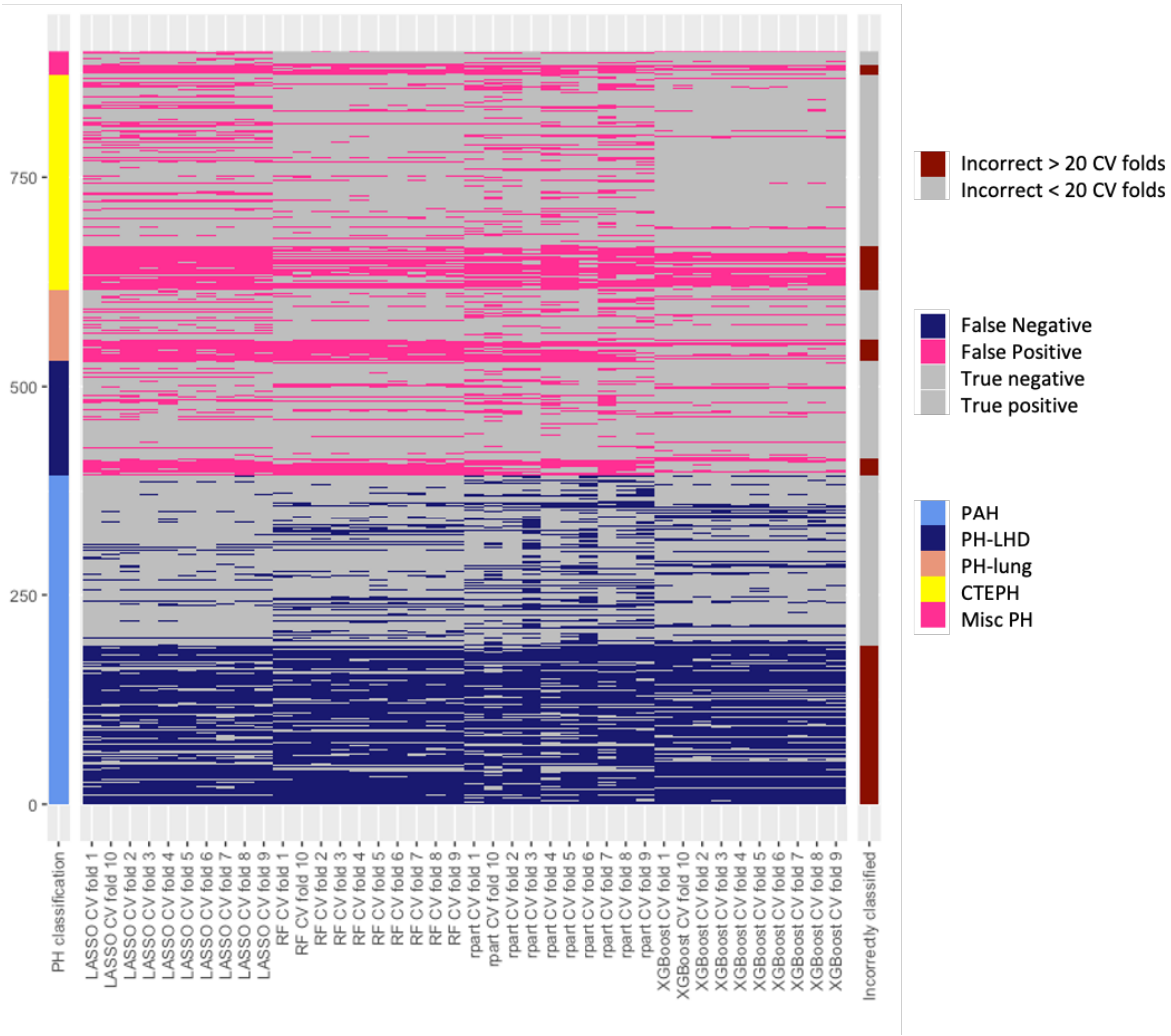


***Figure 3.4**: Predictions for each ML classifier examined over the training and interim sets for detecting PAH from other forms of PH. Each column represents a cross-validated fold, and each row a patient. In the main body of the figure, where the model has made a correct prediction, the corresponding square is grey, and then blue or pink for false negative or false positive predictions respectively. On the far right hand side, the patients incorrectly classified in more than half (>20) the cross validated folds across all four models are highlighted. RF: Random Forest, rpart: recursive partitioning, LASSO: least absolute shrinkage and selection operator. CV: cross validation*

3.3.2 Model miRNAs

The miRNAs driving these models were investigated next. No miRNA appeared in each classification model, and each classification contained miRNAs unique to that signature, with miRNAs repeating across signatures contributing varying degrees of importance (Figure 3.5, Figure 3.6). The signals that separate PH and PAH from disease controls share six miRNAs (miR-RNA-151a-5p, miR-210-3p, miR-30a-5p, miR-193b-3p, miR-126-3p and miR-10b-5p).Hsa-miR-34a-5p and hsa-miR-135a-5p appeared in most models distinguishing between subtypes of PH. Additionally, miR-34a-5p had the highest importance in the model distinguishing between CTEPH and other forms of PH, and the model looking at PAH versus CTEPH. The relative abundance of miR-34a-5p features in discriminating PH-LHD, PH-lung and CTEPH from PH, and CTEPH from PAH is of interest, with miR-34a previously reported

to be dysregulated in PH (Rothman et al. 2016) and play a role in regulating mitochondrial function (Chen et al. 2018). Similarly, we have previously reported changes in expression of miR-150 (Rhodes et al. 2013) within an IPAH population, and miR-150 formed part of the panel of miRNA that distinguishes PAH from other forms of PH (Figure 3.5).

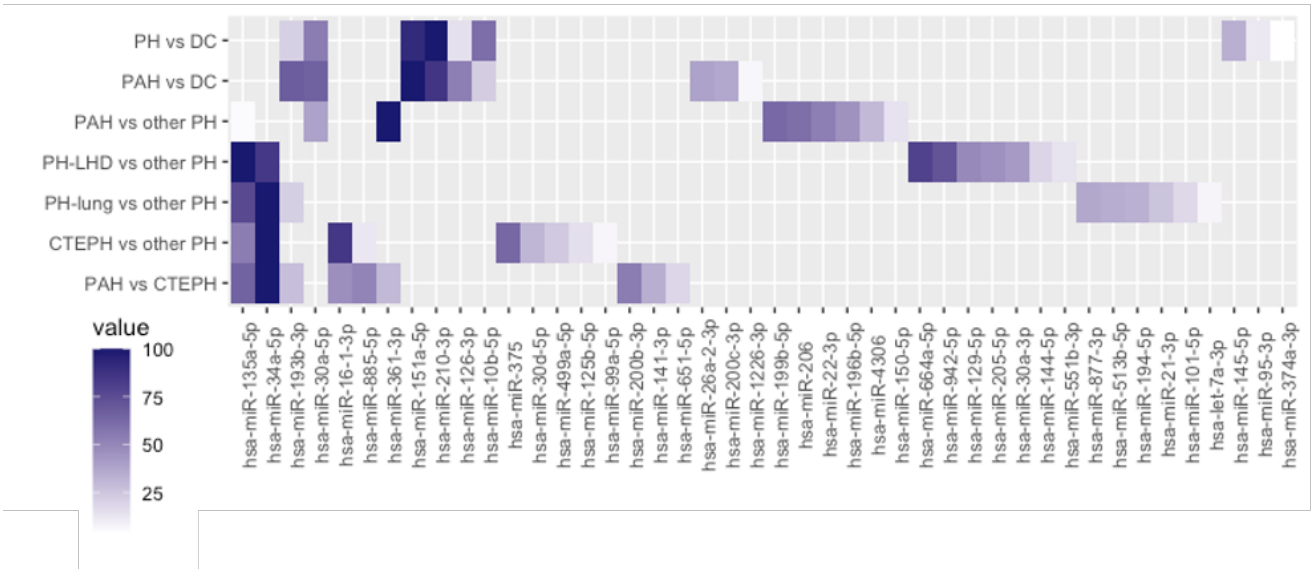


Figure 3.5*: MicroRNA signatures and importance. A heatmap showing the variable importance of each miRNA in differentiating clinical classes. Darker values indicate a higher importance score for that miRNA, scaled to between 1-100.*

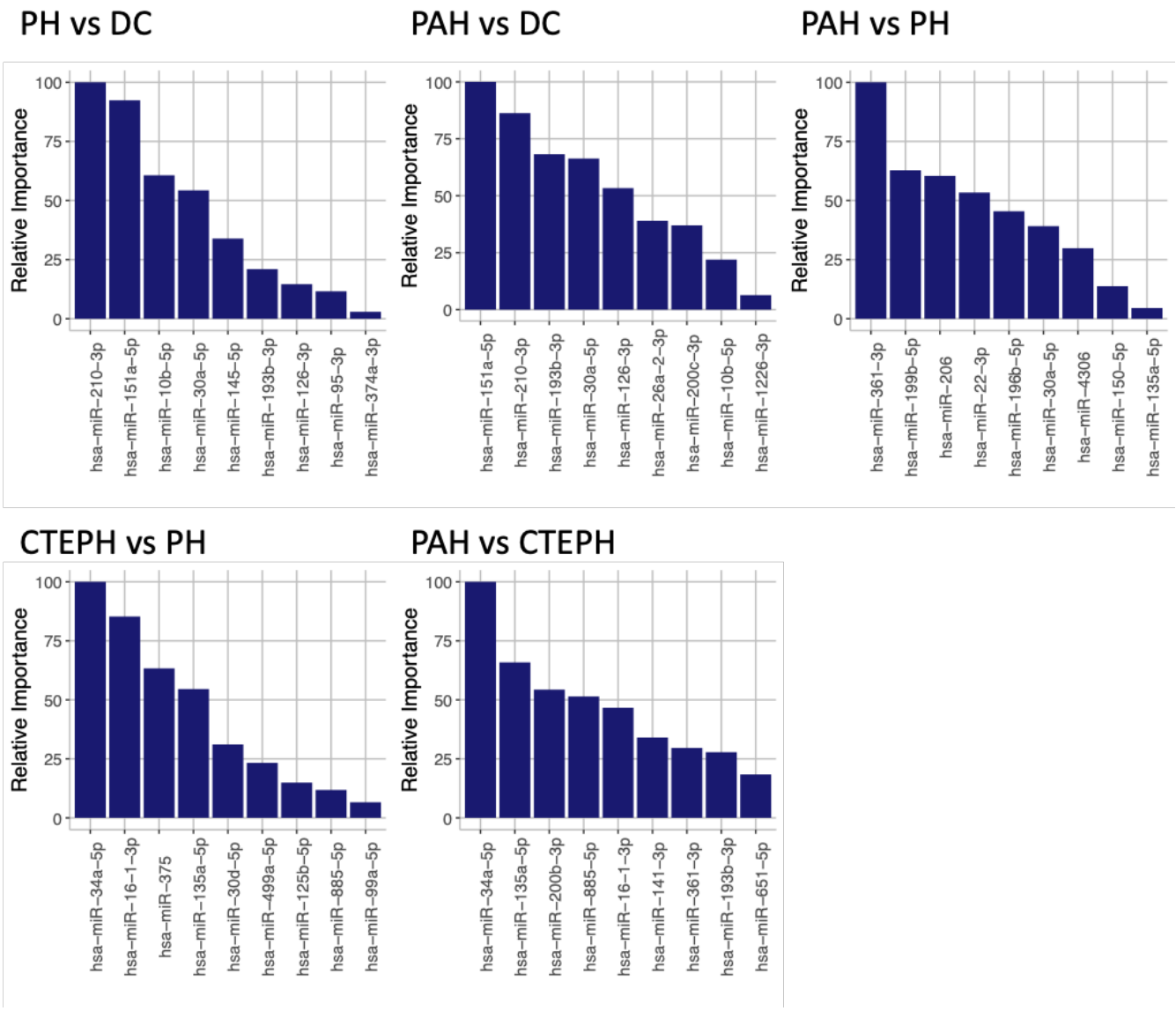
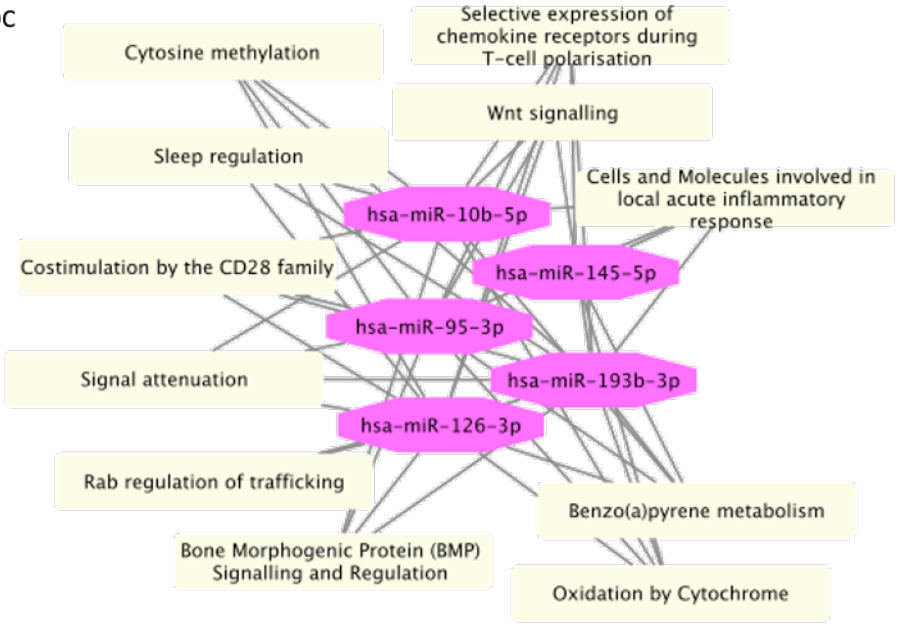


***Figure 3.6**: MicroRNA variable importance plots for XGBoost models*

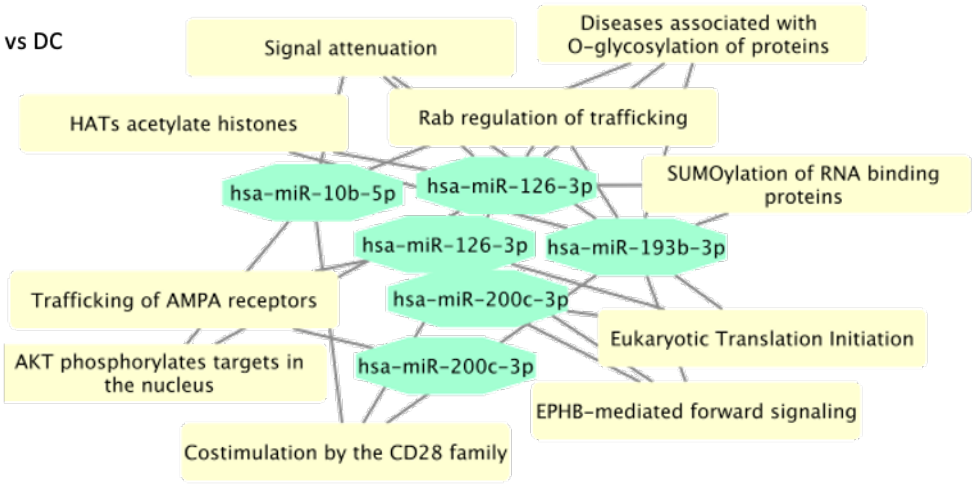
3.3.3 Enriched Pathways

Following miRNA selection, we assessed the enriched pathways (FDR adjusted p-value < 0.05) in each comparison based on the derived signatures (Figure 3.7). Five pathways were seen in more than one comparison. Three pathways were seen in both the PAH vs DC and PH vs DC comparisons; signal attenuation, costimulation by the CD28 family, and rab regulation of trafficking. Each comparison also saw pathways uniquely enriched within that comparison. For example, the PAH vs PH signature saw an enrichment in the FGFR signalling, not seen in the other comparisons. The involvement of FGFR signalling in PAH has been seen in multiple other studies (Zheng et al. 2015).

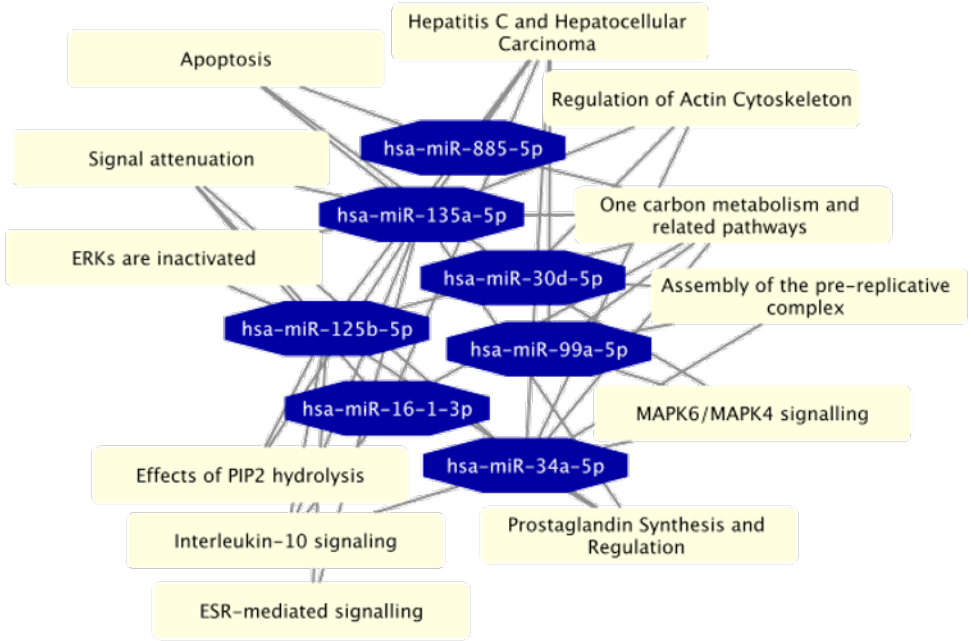
A) PH vs DC



B) PAH vs DC



C) CTEPH vs PH



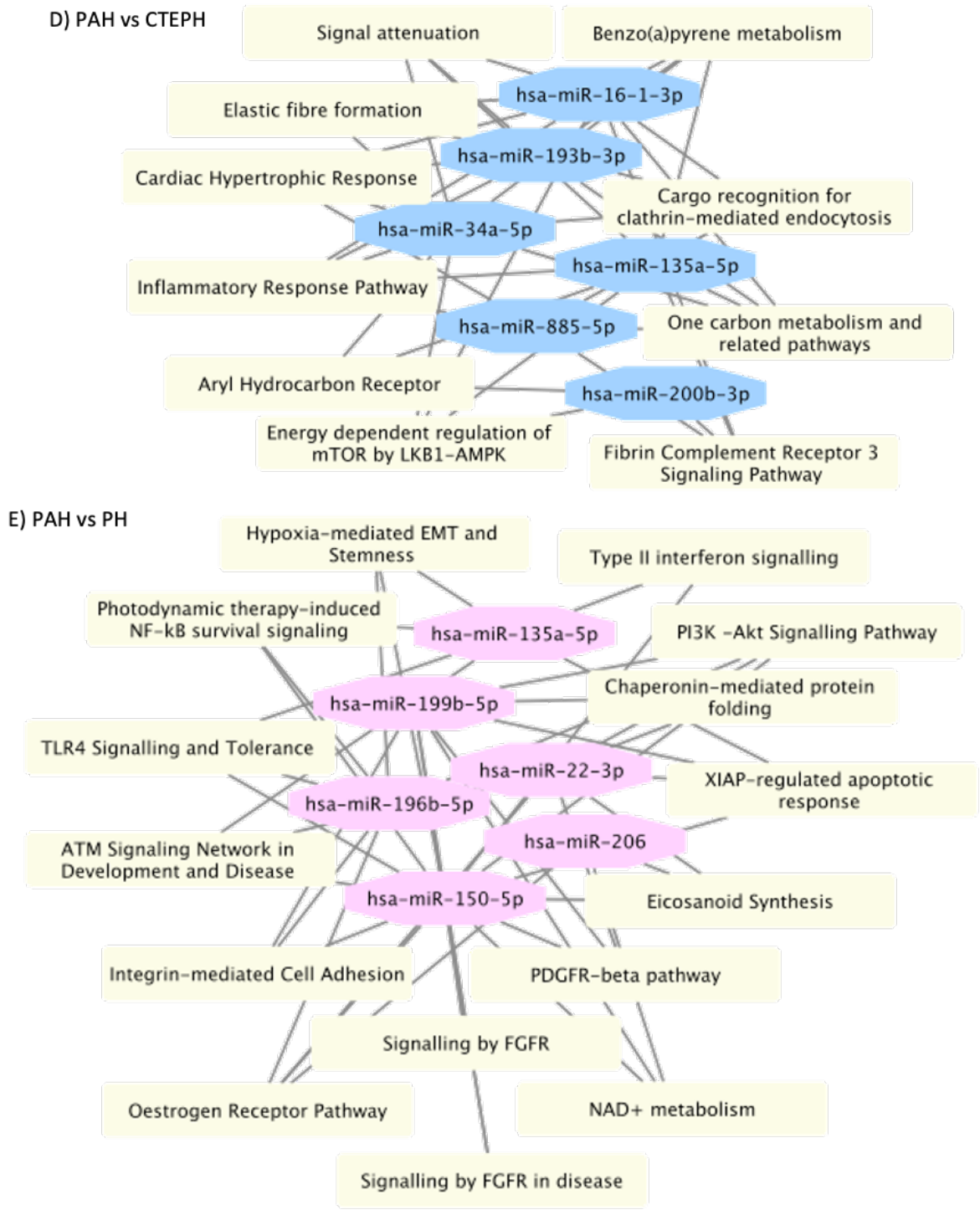


***Figure 3.7**: Top 10 (ranked by q-value) enriched pathways in miRNA signatures for five comparisons. A) PH vs DC, B), PAH vs DC, C) CTEPH vs PH, D) PAH vs CTEPH, E) PAH vs PH*

3.3.4 Validation of the PAH vs DC signature in an external cohort

Owing the poor correlation coefficients seen between serum and plasma in matched samples from Sheffield (Table 3.6), LR models were developed from mean centred miRNA data (Figure 3.8), using the miRNAs selected by and XGBoost model to compare performance in a separate cohort of patients from the Brigham and Women's Hospital in the United States who had plasma samples taken using the same MiREXES platform. The largest difference in median levels for a miRNA between the PAH and DC groups was miR-126-3p in the US cohort (0.394 **Δ**Ct and -0.519 **Δ**Ct for PAH and DC respectively). The smallest median difference between PAH and DC groups was miR-1226-3p in the UK validation cohort (0.0410 and 0.571 respectively).

Table 3.6*: Spearman correlation coefficient between plasma and serum for miRNAs in the PAH vs DC signature in matched patients from Sheffield.*

miRNA	**Correlation coefficient**	**P-value**
hsa-miR-210-3p	0.375	0.00244
hsa-miR-151a-5p	-0.0153	0.904
hsa-miR-193b-3p	0.802	<2.2e-16
hsa-miR-30a-5p	0.804	1.32e-15
hsa-miR-126-3p	0.161	0.203
hsa-26a-2-3p	0.0948	0.474
hsa-miR-200c-3p	0.183	0.149
hsa-miR-10b-5p	0.869	<2.2e-16
hsa-miR-1226-3p	0.163	0.206

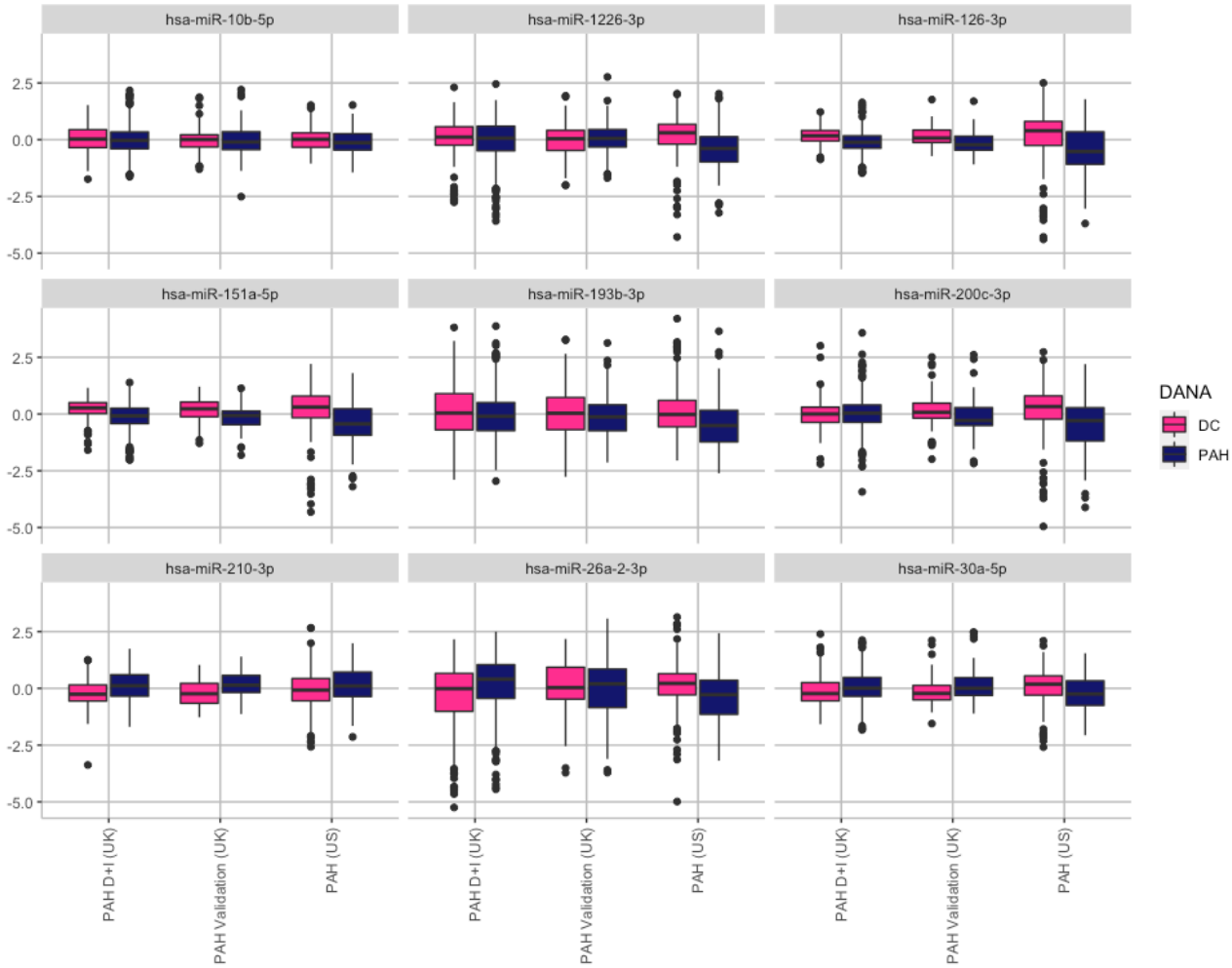


***Figure 3.8**: Mean centred expression profile of miRNAs forming a signature to classify patients with PAH and DC across two separate cohorts of patients.*

Although the coefficients for the miRNAs were not the same between the models (Figure 3.9), similar AIC (20) and mean cross validated AUC (0.80) were seen between models (Table 3.6), suggesting that the same combination of miRNAs can be used to differentiate PAH from DC in both cohorts. A wilcoxon rank sum test of the patient scores showed significant differences in the predicted scores for patients with PAH and DC in both cohorts (p-values 2.2e-16 and 7.22e-11 for UK and US cohorts respectively). The individual patients' scores for the Logistic Regression models for classifying PAH from disease controls (DC) are shown in Figure 3.9B.

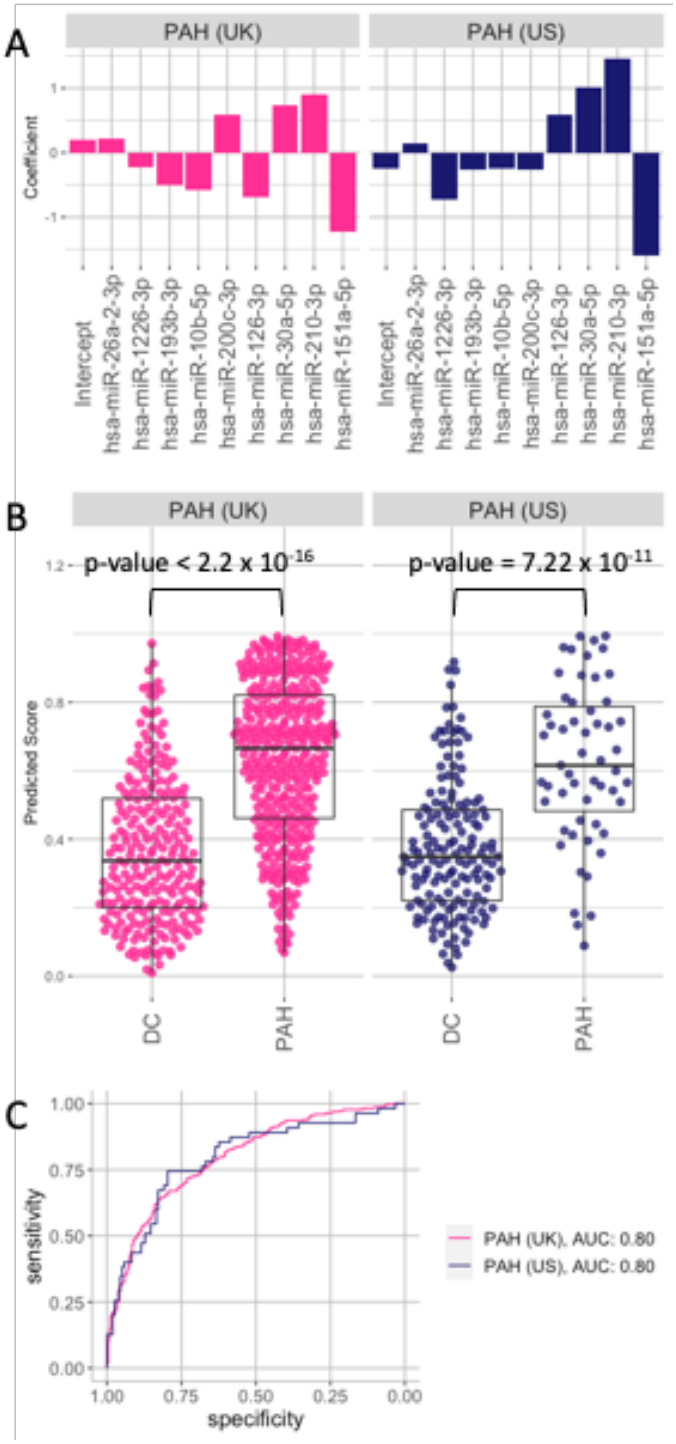


Figure 3.9*: Validation of PAH miRNA signature in both UK and US cohorts using the same miRNAs as in the diagnostic signature. (A) Coefficients for each miRNA in the logistic regression model used to classify PAH vs Disease Control (DC) in each cohort. (B) Predicted score of each patient being PAH from the logistic regression models. P values for Wilcoxon rank sum test between PAH patients and DC (C) ROC for both models*

3.4 Discussion

Targeting an early diagnosis of PH remains a key clinical aim. In this project, we derived a panel of circulating miRNAs from a serum screen from an unbiased screen of over 600 miRNAs, with 1484 patients presenting at three expert UK clinics; the most comprehensive screen to date. We compared the performance of this model with NT-proBNP in detecting PH from disease controls. We identified small panels of miRNAs which distinguish two treatable subtypes of PH from a population of patients at risk of the condition; PAH and CTEPH. These panels could potentially be used independently or combined with NT-proBNP within the patient investigation pathway. This provides the prospect of a point of care test earlier in the diagnosis pathway, identifying prospective patients for treatment at an earlier date.

Changes in levels in circulating miRNAs are well recognised. As such, we expected that the miRNA panel would out-perform the single marker NT-proBNP. However, NT-proBNP did not appear to perform significantly better in the distinction of PH from other breathless controls. The added value of the miRNAs appears to be in identifying the treatable subgroups (PAH and CTEPH) from within the PH group, suggesting miRNAs may hold more disease specific aetiological information, rather than the more generic cardiac stress marker NT-proBNP. Previous studies have mostly focused on comparing smaller cohorts of patients with healthy controls, limiting the utility of these findings. Here, we have compared patients across the spectrum of PH with other patients of similar presentation, resulting in miRNA signatures with a greater potential value.

A key limitation to this study was the validation in a separate cohort, due to the discrepancies between miRNA concentrations in plasma and serum. Other commentators have noted results may be significantly different depending on if serum or plasma is used (Saliminejad, Khorshid, and Ghaffari 2019). We hope to address this in the upcoming CIPHER trial (ClinicalTrials.gov Identifier: NCT04193046). Additionally, this study did not include miR-636 or miR-187-5p, the two miRNAs identified in the previous chapter. Hsa-miR-34a was also identified as a potential classifier for PAH in the previous chapter, building on other studies. Hsa-miR-34a-5p was one of two miRNAs appearing in most models classifying between subtypes of PH, along with hsa-135a-5p. In a study looking at PAH and disease controls (Alexander M. K. Rothman et al. 2016), both hsa-miR-135a-5p and hsa-miR-34a-5p were identified as dysregulated. A later study looking at PAH and healthy controls (K.-H. Chen et al. 2018) also identified hsa-miR-34a-3p as dysregulated. Interestingly, hsa-miR-34a-5p was identified in signatures for PH subtypes versus other PH for all subtypes except PAH versus other PH, though this is perhaps explained by the different make-up of the cohorts

The six miRNAs in both the signatures for PAH versus DC and PH versus DC (miR-RNA-151a-5p, miR-210-3p, miR-30a-5p, miR-193b-3p, miR-126-3p and miR-10b-5p) have all previously been identified as dysregulated in PAH or cardiovascular disease, lending confidence to our hypothesis that key miRNAs are being identified. The down regulation of hsa-miR-126 has been shown to be associated with right ventricular failure (Potus et al. 2015). Mir-151-5p identified in hypertension-induced cardiovascular disease (Amirlatifi et al. 2022). MiR-193 and miR-210 were noted as down regulated in rats with PAH compared with controls (Xiao et al. 2017). Previous studies have also identified the miR-30 family as decreased in cardiovascular disease, with involvement in vascular remodelling (Zhang et al. 2019). Finally,

miR-10, a miRNA associated with inflammation has also been identified in discriminating patients who will reject heart transplantation (Duong Van Huyen et al. 2014) .

In the majority of cases, the XGBoost models achieved the highest AUCs. XGboost has been effectively deployed in a range of clinical settings, for example in predicting mortality in critically ill influenza patients (Hu et al. 2020), and chronic kidney disease diagnosis (Ogunleye and Wang 2020). However, XGBoost models have several drawbacks for translation to clinical use. Primarily, the XGBoost models are 'black box' models; they are difficult for a clinician to unpick and understand. To try and add to the interpretability, we have included the feature importance scores, which quantitatively represent the miRNA's contribution to the model, and can be visualised (Figure 3.6). A future direction to explore in more depth the contribution of miRNAs to the models could look at Shapley additive explanations, as used to aid clinicians looking at mortality in influenza patients (Hu et al. 2020) and to predict acute myocardial infarction (L. Ibrahim et al. 2020).

The miRNA panel derived to distinguish PAH from PH was also noted in some patients clinically ascribed to the other diagnostic groups (Figure 3.3). An overlap of vascular histology in patients with PAH and CTEPH has been documented extensively (Moser and Bloor 1993). Vascular remodelling has more recently been noted in the lungs of patients with both PH-LHD (Fayyaz et al. 2018) and PH-lung (Bunel et al. 2019). We suggest that the presence of the PAH miRNA panel in other clinical groups could signal a common pathology. The potentially shared pathology, particularly the incidence of pre- and post- capillary PH might have been a limiting factor in the identification of these miRNA signatures. This led to the suggestion that an examination of the distribution of miRNAs, agnostic to clinical classifications using unsupervised learning might help inform the clinical presentation of patients from a mixed cohort of PAH, PH-LHD and PH-lung and is investigated in the next chapter.

Chapter 4: Clustering in PH

4.1 Introduction

Pulmonary hypertension (PH) is a rare but often fatal disease, acknowledged to be heterogeneous with a wide array of pathobiological mechanisms, phenotypes and aetiologies. Current classification is based on clinical presentation and hemodynamics, which does not allow for individualised treatment for distinct patient phenotypes. PH can be categorised into five subgroups; PAH, PH due to left heart disease (PH-LHD), PH due to lung disease (PH-lung), chronic thromboembolic disease (CTEPH) and miscellaneous PH (Simonneau et al. 2019) (Table 1.1). However, more than five phenotypes exist amongst patients with PH, and unravelling the heterogeneity within each PH category is essential for the advancement of treatment and individualised care.

Making a correct sub-diagnosis of PH is problematic as PH may be multifactorial, symptoms are non-specific and diagnostic tests can be problematic to unravel. Clinical classification based mainly on haemodynamics can be challenging, and may impede identification of treatment responders and new therapy development. PH-LHD and PH-lung can be occasionally diagnosed and treated as PAH (Barnett and Selby 2015). Treatments which target the vasculature in the lungs in patients with PH-LHD have all had negative results in multicentric clinical trials (Fernández et al. 2019). The distinction between PAH from PH-LHD and PH-lung is a clinical challenge (Figure 3.4). Some commentators have even suggested a pathology continuum between PAH and PH-LHD, embracing ‘atypical’ PAH (Opitz et al. 2016).

After demonstrating miRNAs repeated across panels of signatures discriminating between PH subtypes in the previous chapter, we reasoned that these shared miRNAs may reflect overlapping pathology. Previous studies have successfully used unsupervised clustering on RNA-seq. For example, a group used unsupervised hierarchical clustering on ovarian low-grade serous carcinomas from different locations to determine that fallopian tubes are likely to be the cellular source of low-grade serous carcinomas (Qiu et al. 2017). A different study used integrative non-negative matrix factorisation to cluster RNA-seq and methylation datasets to identify heterogenous subtypes of Pancreatic Ductal Adenocarcinomas (Roy, Singh, and Gupta 2021).

A recent study looked at transcriptomic heterogeneity in PAH (Kariotis et al. 2021). We hypothesised that circulating miRNA may also inform molecular endotypes in a PH cohort. We reasoned that the potential shared pathology might be a restrictive factor in the efficacy of miRNA signatures for the current clinical classifications, and subsequently restrict the potential of miRNAs to uncover underlying molecular drivers. As a result, we took an unsupervised clustering approach to patients from PAH, PH-LHD and PH-lung to examine how the miRNA expression levels, agnostic to clinical classification may group patients. Misc PH was excluded from the analysis owing to a small number of patients. Additionally, we decided to remove patients with CTEPH. Despite shared vascular pathology between PAH and CTEPH patients, with imaging, the diagnosis of CTEPH in the clinic is relatively straightforward, compared with the challenge present by confidently differentiating patients with PAH, PH-LHD

and PH-lung based on clinical measurements. We also examined the heterogeneity within each clinical classification group.

4.1.1 Aims

1. To use unbiased partitioning of patients into distinct clusters within each WHO clinical classification group
2. Cluster patients within the combined cohort of PAH, PH-LHD and PH-lung patients.
3. To examine the survival properties, miRNA, and clinical features which may help distinguish these clusters.
4. To look at enriched pathways within each cluster.

4.2 Methods

The pre-processing steps and cohort overview can be found in Chapter 3, and the workflow for this chapter in Figure 4.1.

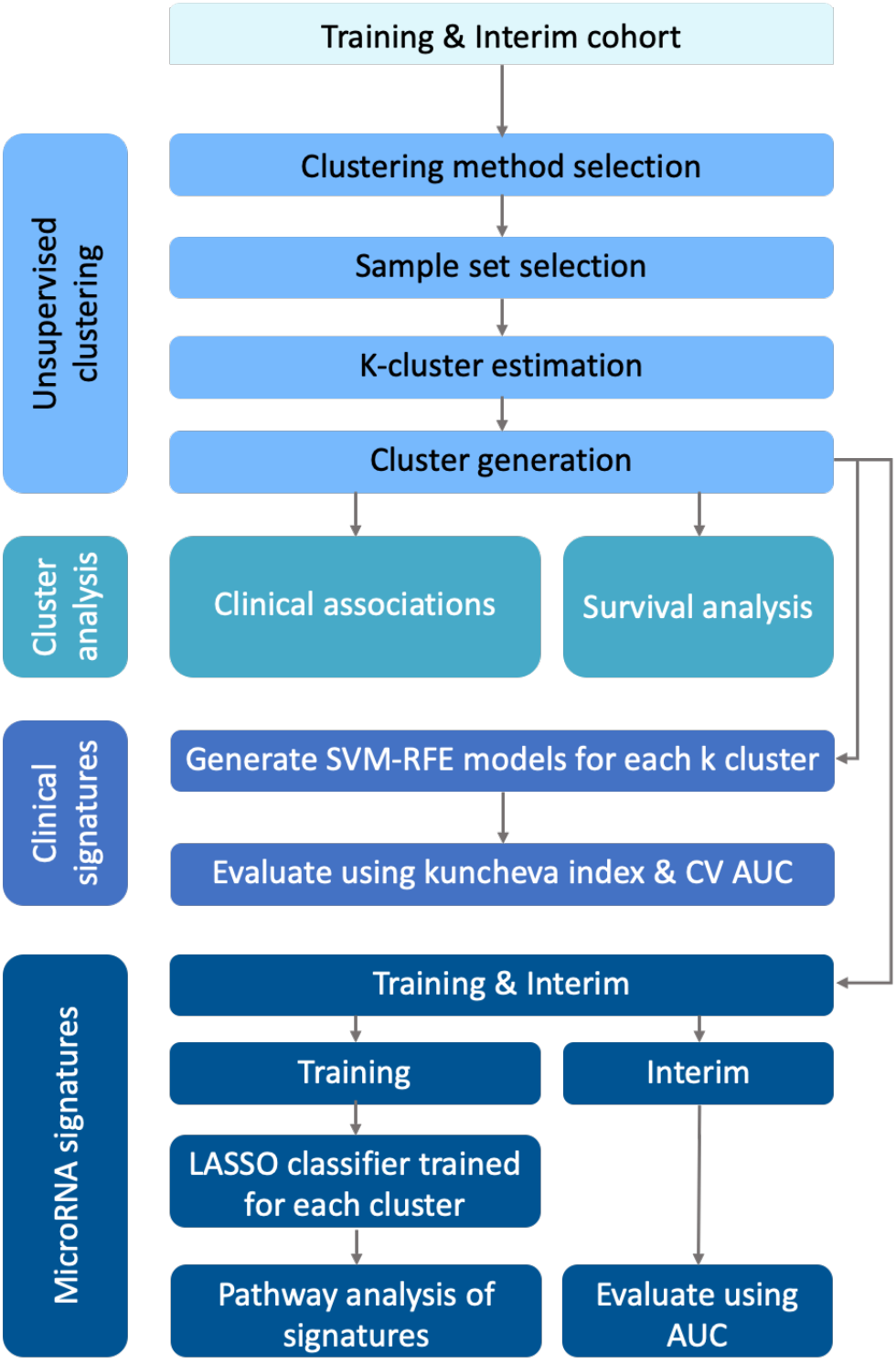


***Figure 4.1**: Method overview. The training and interim cohorts were combined before undergoing unsupervised clustering. Once the clusters had been generated, clinical associations between clusters and survival differences were analysed before the generation of miRNA signatures and clinical signatures for each cluster.*

4.2.1 Unsupervised classification

We examined the utility of unsupervised clustering on both the individual WHO classification groups for PH, and groups PAH, PH-LHD and PH-lung combined, utilising the same pipeline for each grouping.

There are 4 key steps to the unsupervised clustering pipeline primarily developed by Sokratis Kariotis (Kariotis et al. 2021), and executed by him, which can be broadly described as follows:

1. Clustering method selection
2. Sample set selection
3. k - estimation
4. Cluster generation

4.2.1.1 Clustering method selection

The first step aims to determine which clustering method best suits the data from three methods (k-means, hierarchical and spectral clustering).

The consistency of which the different methods partition the data is used as the metric to decide which algorithm is most suitable for the data set. This is a measure of robustness. The clustering method was run multiple times with variations on the parameters used. If the runs agreed on the data partitioning this is a sign data points have not been randomly assigned but are revealing patterns in the underlying data structure. The intra-agreement was calculated using the average of the adjusted Rand index (package fossil v0.3.7).

4.2.1.2 Sample set selection and k estimation

All three of the unsupervised clustering algorithms used are affected by high dimensional data, both in an increase of computational cost and loss of performance from data noise. As a result, the miRNAs were ranked in descending order of variance (stats v3.6.0 R package). Multiple subsets were extracted, starting with the 50 most variable miRNAs and increasing in size by 50 each time. Multiple runs of spectral clustering were computed (for k = 2,3,4,5,6), and the miRNA subset with the highest stability used as measured by a bootstrap approach (package fpc v2.2-3).

An ensemble learning majority voting method was used to determine the most suitable number of clusters (k). 14 machine learning indices were calculated for each k within the range [2,10], and then averaged. The averages were compared with the ideal value for each index, and used to determine which k was most suitable, according to that index.

4.2.2 Cluster analysis

4.2.2.1 Survival

Survival analysis was carried out (R packages survival v 3.2-13 and survminer v0.4.9) to identify survival differences between clusters. Patients having undergone transplantation were excluded from analysis, and no patients were lost to follow-up. Kaplan-Meier curves were plotted for each cluster. Cox models and hazard ratios were subsequently fitted to the clusters, adjusting for age and sex. Time was measured in days from sample to event date, or census

date. All deaths were recorded as an event, with patients undergoing transplant removed from the survival analysis. Survival was examined at 10 years, and cox models were also fitted for functional class and REVEAL risk group.

4.2.2.2 Clinical associations

Frequency tables were created for sex, functional class, site, diagnosis, and comorbidities within each cluster. Pairwise comparisons were made between the clusters in R, using Fisher's exact test or chi squared test where appropriate, using a Benjamini Hochberg correction for multiple test correction.

For the continuous variables, normality was assessed visually using box-plots and QQ-plots, followed by a shapiro-wilk test for each cluster for each continuous variable. Where all clusters were assessed to be normally distributed, an anova test was used. Where a group was not-normally distributed, a Kruskall-Wallis test was used instead. Where these tests resulted in significant p-values, they were followed up with appropriate post-hoc tests. P-values were adjusted using the Benjamini-Hochberg method. The variance of each cluster was also assessed. Where there were large differences in variance between groups, the data were log-transformed.

4.2.2.3 Cluster signatures

After splitting the cohort into training and interim sets as described in Chapter 3, miRNA signatures for each cluster were identified using LASSO classification models (glmnet v4.1-2). Models were fitted for each cluster versus all the others to create six binary models. The regularisation parameter, λ, was selected for each model using 10-fold cross-validation across the training set to select the value of λ with the minimum binomial deviance in each instance. Signature performance was then evaluated using the interim set. Pathway enrichment analysis was performed using the method as described in Chapter 3 for each cluster using the miRNA signatures.

4.2.2.4 Missingness assessment and imputation

In order to generate clinical signatures for the clusters in PAH, PH-LHD and PH-lung combined, missing clinical variables were imputed. Prior to imputation, the data were assessed both for missingness and patterns within the missingness within the training and interim sets. The clinical variables with the highest percentage of missing data were ISWT (75%), predicted transfer factor for carbon monoxide (TLco predicted, 49%) & 6-minute walk distance (36%). Missing values were high for 6-minute walk distance and ISWT as only one of these was available per site. Sheffield patients had ISWT recorded, and whereas patients from Cambridge and Imperial had 6-minute walk tests. Missing percentages for variables can be found in Table 3.1. Data was not imputed for ISWT, TLco predicted, or 6-minute walk distance due to the high percentage of missing data. No participants were missing data for disease classification, sample site, treatment naive status, sex, NT-proBNP or uric acid.

The MICE (Multivariate Imputation by Chained Equations) method was used to impute data in R (mice R package, v3.13.0). TLco predicted had nearly 50% missing data so was excluded from both the imputation and subsequent clinical signature. Ethnicity was also excluded due to the low granularity of available data. The data were also assessed on missingness patterns. 56 patients with no available right heart catheter information were removed as these could not be accurately imputed. Based on recommendations in the MI literature (White, Royston, and

Wood 2011), all the variables from the analysis model were included in the imputation model to ensure relationships between the variables of interest were retained, as well as auxiliary variables, such as the miRNAs. The number of imputations was set to 60, as ~45% of patients had no missing variables once TLco, ethnicity, ISWT and 6-minute walk distance were removed. The number of iterations was set to 50. The convergence of the algorithm was checked along with the means and standard deviations of imputed values.

4.2.2.5 Cluster clinical signatures (Emmanuel Jammeh)

Emmanuel Jammeh derived clinical signatures from the available imputed clinical data using support vector machines (SVM) as the estimator in a method previously described (Kariotis et al. 2021). Briefly, The SVM model was combined with recursive feature elimination (RFE), a method used to remove unrelated and superfluous features, retaining the most informative features. The combined SVM-RFE (Guyon et al. 2002) was then used to identify clinical signatures for each cluster. The robustness of the signatures were then evaluated using the Kuncheva index (Kuncheva 2007), and the AUC was used for classification performance evaluation. The contributions of each feature to the final model were measured using the absolute value of the SVM coefficients.

For each cluster signature, k subsamples were taken from the dataset using random sampling without replacement, such that each subsample contained slightly different samples, and fewer samples than the original dataset.

Each of the k subsamples was then split into b bootstrap samples to minimise the effect of variations within the feature selection. SVM-RFE was carried out on all bootstrap samples to generate b feature rankings and b candidate signatures of variable sizes. Ten-fold cross-validation with 10 repetitions was used to calculate the classification performance of each signature.

4.2.3 Code availability

All methods and parameters for methods I have used are described in the git repository available at https://github.com/niamherrington/MiRXES-miRNA.

4.3 Results

4.3.1 Clustering within each clinical classification group

Addressing the first aim within the chapter we examined the clustering within each clinical classification group and identified two clusters as the optimal number within each (Table 4.1), however the small number of patients with Misc PH led to the exclusion of this group from further analysis.

***Table 4.1**: Clustering analysis within each clinical classification group on the discovery and interim sets combined. * Optimal miRNAs are defined as the number of miRNAs used for clustering analysis, the most stable number for that comparison. ** 14 machine learning indices votes for k clusters. Table produced by Sokratis Kariotis*

PH Group	patients	Clustering method	Optimal miRNAs *	K votes **
PAH	394	spectral	55	**K2: 7** **K7: 3** **K3: 1** **K8: 1** **K6: 2**
PH-LHD	137	spectral	185	**K2: 7** **K5: 3** **K3: 2** **K8 :2**
PH-lung	84	spectral	70	**K2: 6** **K6: 1** **K3: 3** **K7: 4**
CTEPH	258	spectral	75	**K2: 7** **K7: 2** **K3: 1** **K8: 4**
Misc PH	27	spectral	55	**K2: 7** **K6: 2** **K3: 1** **K7: 3** **K5: 1**

We compared overall survival between clusters within each WHO group, and found no significant differences within classification groups over all time (Figure 4.2). However, at the 5 year mark, Cluster B in PH-LHD had significantly worse survival (p-value 0.013).

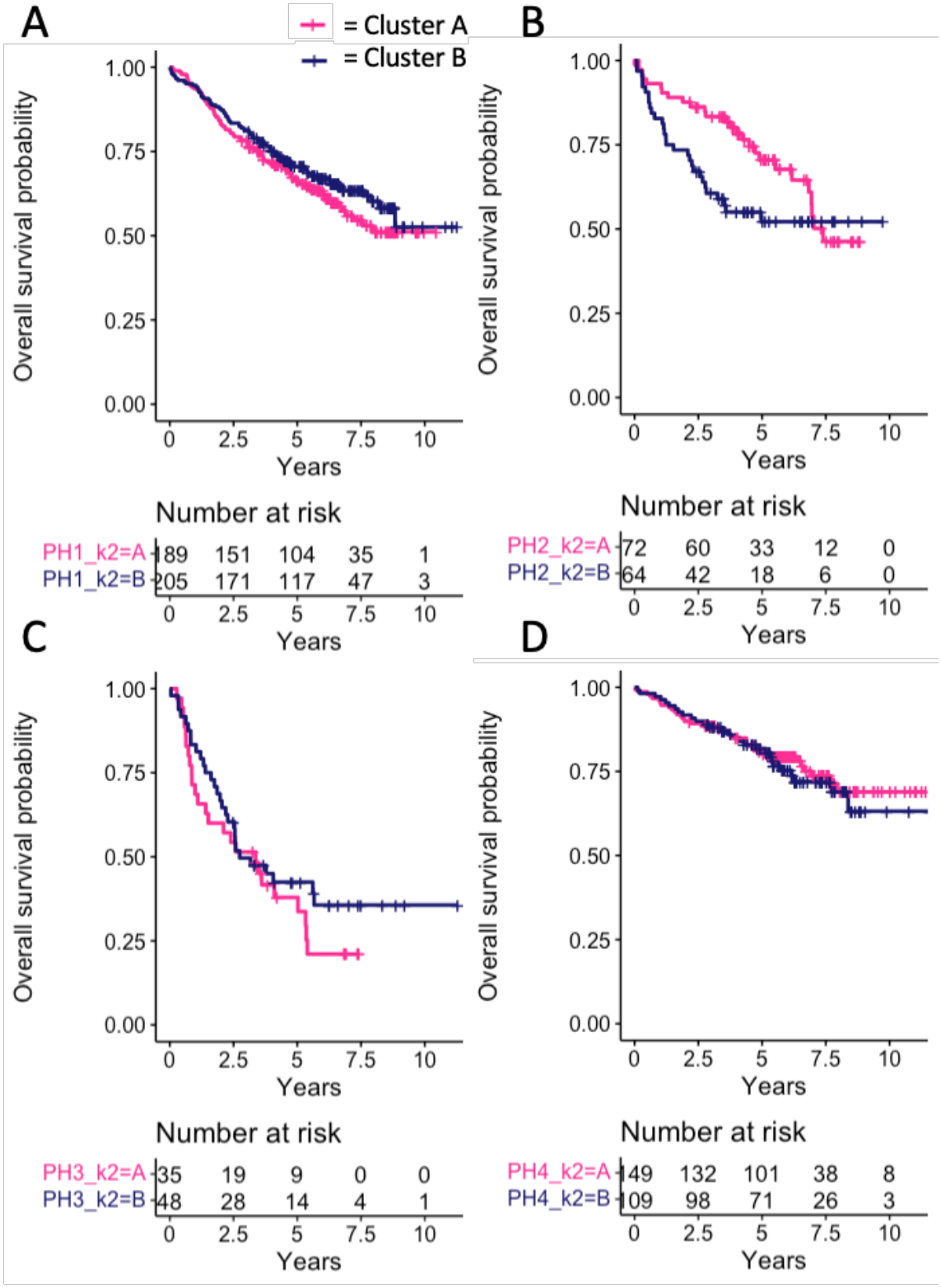


***Figure 4.2**: Kaplan-Meier curves showing survival profiles for two clusters within each WHO clinical classification group. (A) PAH (B) PH-LHD (C) PH-lung (D) CTEPH*

After comparing survival between clusters, we checked for enrichment of clinical variables within the clusters (Table 4.2). Only one variable was significantly different; eGFR (estimated glomerular filtration rate, adjusted p-value 0.0350) showed a significant difference between

clusters A and B in CTEPH patients (Table 4.2). Despite this, some heterogeneity was seen in the CTEPH patients, for example creatinine levels were higher in cluster A (adjusted p-value 0.0828, median values 98 and 87 for clusters A and B respectively). In the PAH clusters, some differences were observed in predicted forced vital capacity (FVCP, mean values 89 and 94 for clusters A and B respectively) and predicted TLCO (median values 50.0 and 57.7 for clusters A and B respectively).Several demographic variables did not show significant differences between clusters. For instance sex had adjusted p-values of 0.540, 0.873, 1.00, and 0.930 across PAH, PH-LHD, PH-lung and CTEPH respectively, and BMI had adjusted p-values of 0.777, 0.888, 0.643, and 0.925 respectively.

***Table 4.2**: Adjusted p-values from enrichment tests for clinical parameters between two clusters within each WHO group (Chi-squared or fisher exact test for categorical variables, and t-test or wilcoxon for continuous variables). * p-value < 0.05*

Clinical Parameter	**PAH**	**PH-LHD**	**PH-lung**	**CTEPH**
6MWD	0.540	0.593	1.0	0.925
Age	0.273	0.888	1.00	0.832
BMI	0.777	0.888	0.643	0.925
BPDIA	0.932	0.824	0.643	0.994
BPSYS	1.00	0.873	1.00	0.925
Cardiac Index	0.540	0.801	1.00	0.561
Cardiac Output	0.942	0.593	1.00	0.780
Comorbidity: Atrial Fibrillation	0.273	0.593	0.918	0.780
Comorbidity: COPD	0.965	0.873	0.643	0.780
Comorbidity: CTD	0.965	0.888	1.00	1.00
Comorbidity: Diabetes	0.486	0.593	0.500	0.336
Comorbidity: Ischaemic heart disease	0.932	0.888	1.00	0.925
Comorbidity: Scleroderma	1.00	0.593	1.00	0.832
Comorbidity: Sleep Apnea	0.965	0.593	0.485	0.336
Comorbidity: Thyroid Disease	0.960	0.593	1.00	1.00
Creatinine	0.932	0.873	1.00	0.0828

Clinical Parameter	PAH	PH-LHD	PH-lung	CTEPH
dPAP	0.960	0.693	1.00	0.780
eGFR	0.932	0.886	1.00	0.0350 *
Functional Class	1.00	0.705	1.00	0.780
Predicted forced vital capacity	0.273	0.888	1.00	0.780
ISWT	0.960	0.787	0.746	0.897
mPAP	0.932	0.593	1.00	0.780
mRAP	1.00	0.888	1.00	0.994
NT-proBNP	0.932	0.693	1.00	0.780
PAWP	0.540	0.888	0.500	0.925
Platelet Count	0.540	0.991	0.411	0.930
PVR	0.808	0.200	1. 00	0.925
REVEAL risk group	0.960	0.969	0.715	0.127
Sex	0.540	0.873	1.00	0.930
Site	0.273	0.593	0.567	0.0828
sPAP	0.540	0.593	1.00	0.897
SvO2	0.614	0.200	1.00	0.930
TLCO predicted	0.273	0.873	1.00	0.925
Treatment	0.540	1.00	0.500	0.130
Uric Acid	0.540	0.991	1.00	0.897

4.3.2 Clustering of all patients with PAH, PH-LHD and PH-lung

After examining the heterogeneity within each of the current clinical classification groups, we looked at combining PAH, PH-LHD and PH-lung, and the heterogeneity within the combined groups. We reasoned that since we earlier demonstrated signatures for both PAH and CTEPH

(from other forms of PH), but that the PAH signature was also seen in other classes of PH (Figure 3.3) potentially due to shared pathology, particularly the incidence of pre- and post-capillary PH, that the miRNA clustering approach might identify common endophenotypes of PH across these three classification groups. We chose to exclude miscellaneous PH due to the small patient numbers. Spectral clustering was again selected as the preferred clustering method, with an optimal miRNA subset selection of 50 miRNAs, and six clusters selected as the optimal number (Figure 4.3)

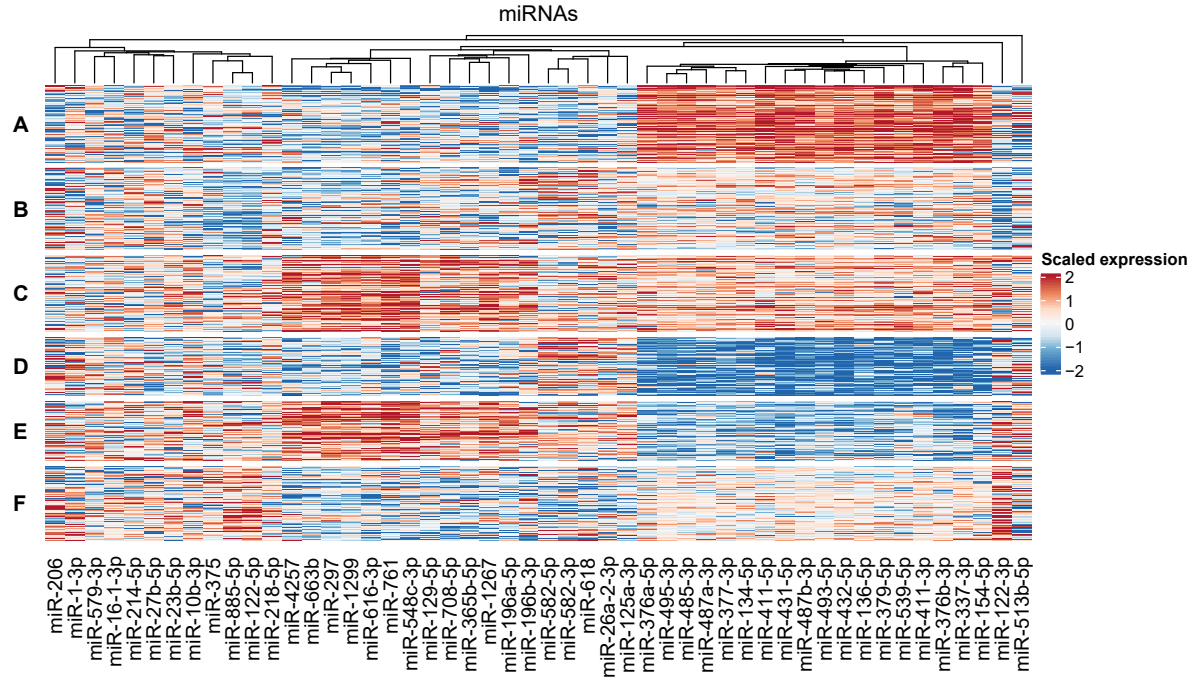


***Figure 4.3**: Unsupervised clustering of patients across PAH, PH-LHD and PH-lung using their miRNA profiles. Heatmap shows z-score scaled miRNA expression. Figure produced by Sokratis Kariotis.*

The traditional WHO clinical classification groups were spread between all six clusters, with no cluster containing a single classification group (Figure 4.4). Cluster E had the highest percentage of PAH patients (72.1%), cluster A had the highest percentage of PH-LHD patients (23.7%), and cluster D had the highest percentage of PH-lung patients (15.5%). No cluster contained a single classification.

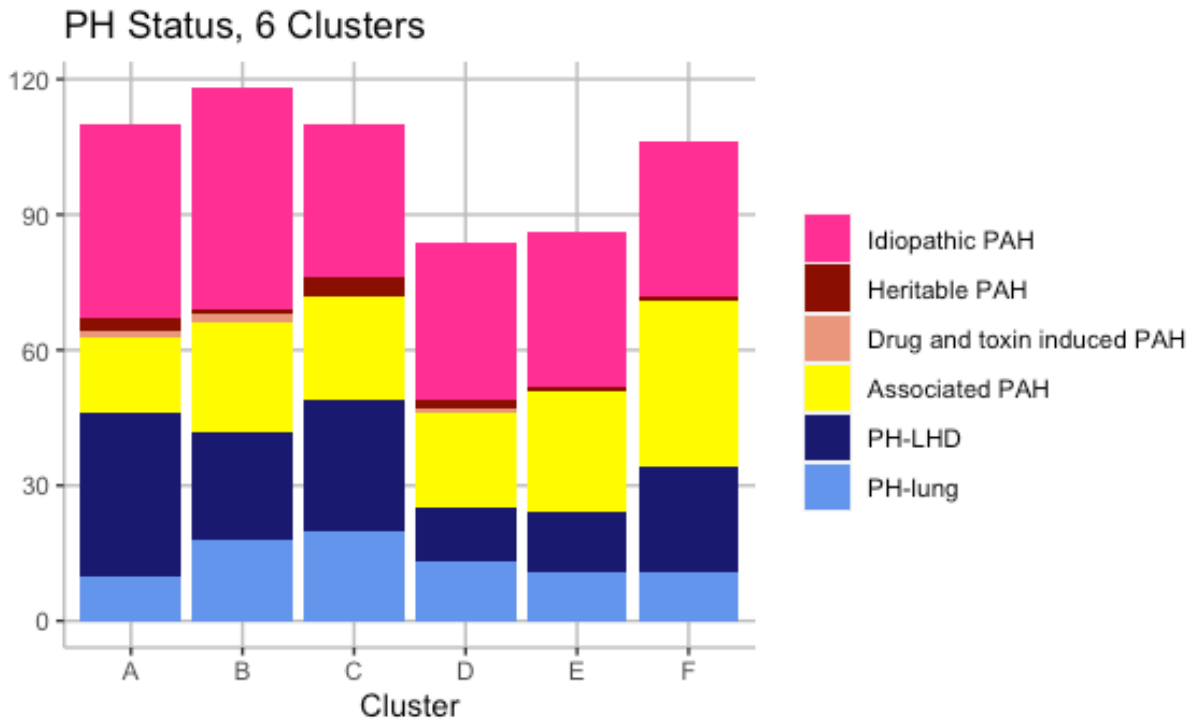


***Figure 4.4**: WHO clinical classification breakdown into six clusters. Cluster E had the highest percentage of PAH patients, cluster A had the highest percentage of PH-LHD patients, and cluster D had the highest percentage of PH-lung patients. No cluster contained a single classification.*

4.3.2.1 Clinical outcomes

By the end of the observation period, 12 patients had undergone transplants and were excluded from further analysis. A further 257 patients had died (41.9%). Kaplan-Meier curves were constructed and log-rank tests performed to compare survival distributions between clusters (Figure 4.5). We also undertook Cox regression to test for any statistically significant survival differences between the clusters. As expected, a higher percentage of deaths occurred amongst men (107 of 209, 51.2%) compared to women (150 of 405, 37.0%), as well as a higher percentage of deaths occurring above the median age (52.5%), compared with the below median age bracket (32.1%). As such, we repeated the survival analysis using a multivariate Cox regression which included the patients' age at sampling and sex (Table 4.3). Taking cluster A as the baseline, clusters C and F showed significantly worse survival with hazard ratios of 1.61 and 1.56 respectively.

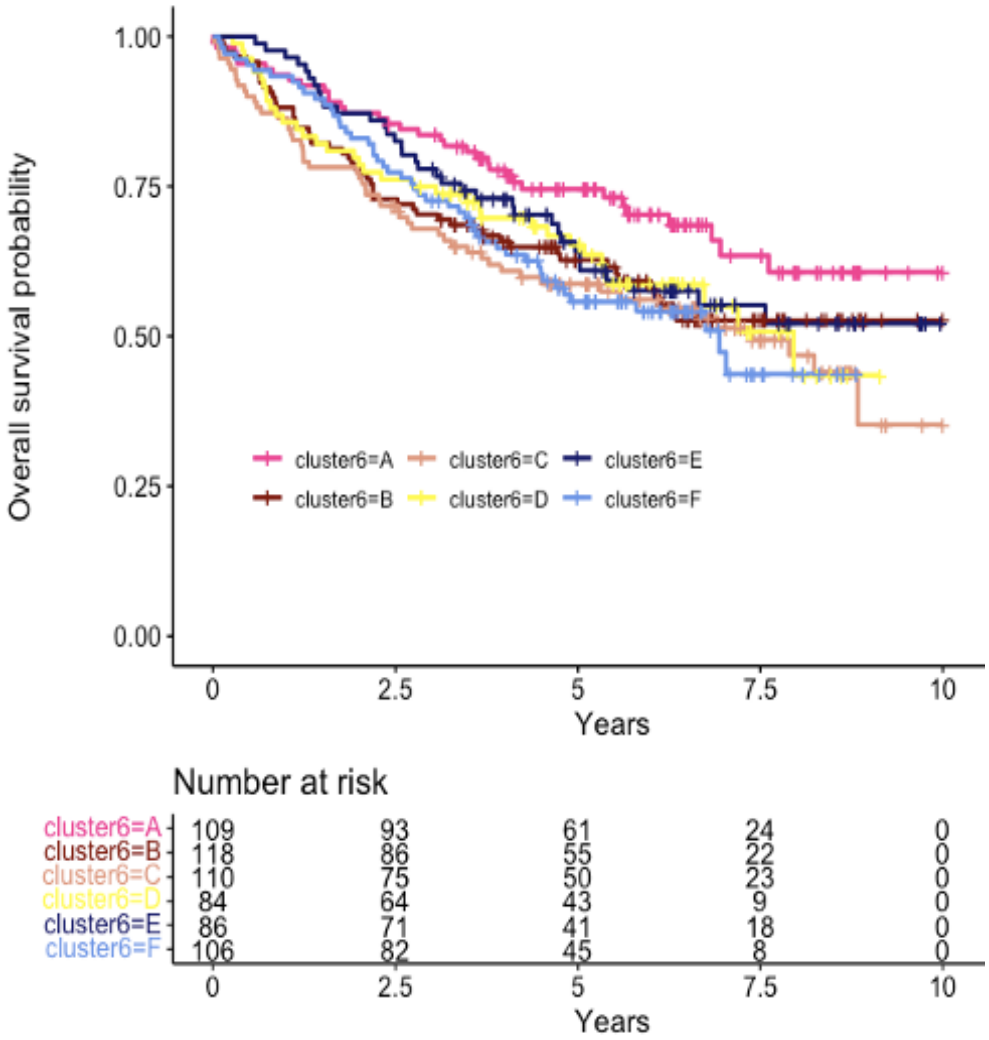


***Figure 4.5**: Kaplan-Meier curves for each of the six clusters. Cluster C and F showed significantly worse survival.*

***Table 4.3**: Cox proportional hazard ratios for sex, age, and cluster at 10 years (95% Confidence Interval). HR = Hazard Ratio. * p-value < 0.05*

Characteristic	**10 year HR**	**p-value**
sex		
F	—	—
M	1.45 (1.13 - 1.86)	3.56e-3 *
age	1.04 (1.03 - 1.05)	2.93e-13*
cluster		
A	—	—
B	1.48 (0.96, 2.28)	0.0789
C	1.61 (1.05, 2.49)	0.0302 *
D	1.50 (0.93, 2.40)	0.0935
E	1.24(0.78, 2.00)	0.364
F	1.56 (1.01, 2.42)	0.0461*

For completeness, we also double checked survival curves when patients were stratified by REVEAL risk group and functional class (Figure 4.6), which stratified mostly as expected. Taking the high risk REVEAL group as baseline, there was no significant difference for the intermediate risk group (p-value 0.238), however the low risk group had significantly better survival (p-value 9.06e-10).

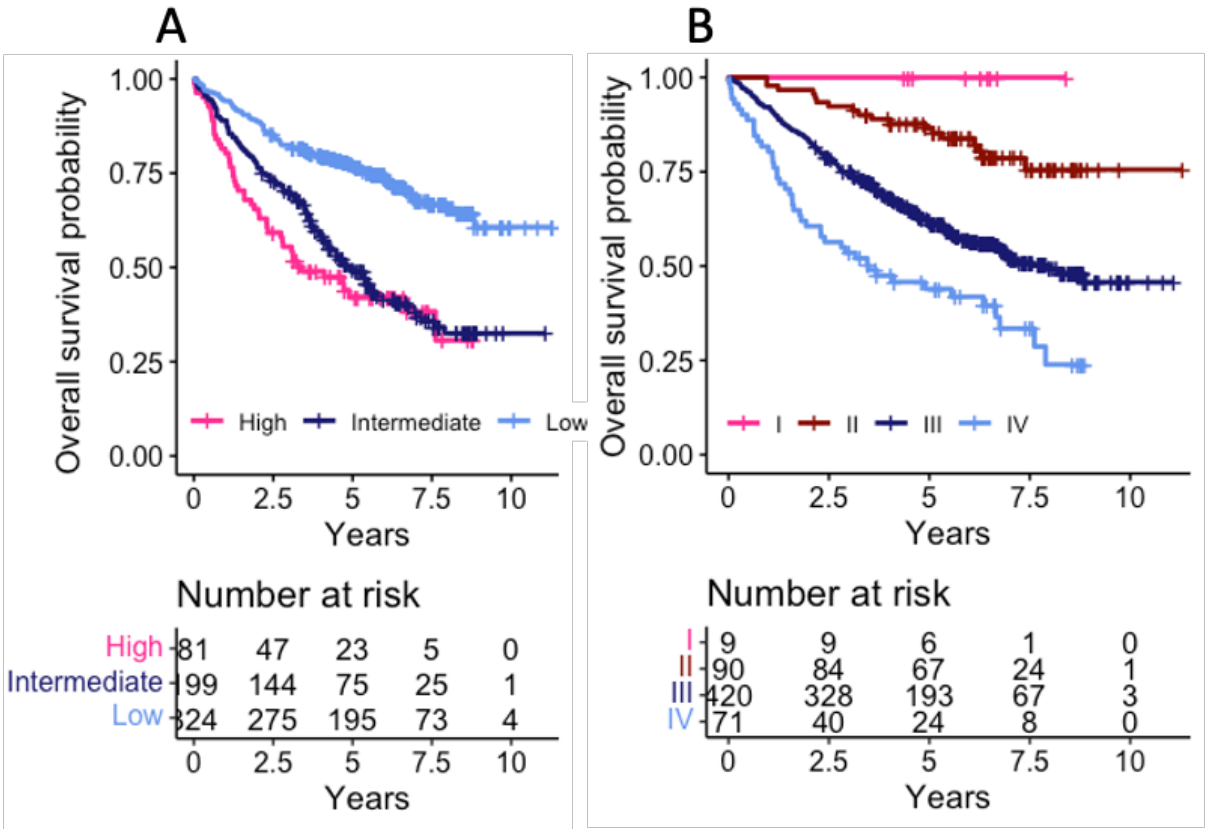


***Figure 4.6**: Survival stratified by A) REVEAL risk group B) Functional class for PAH, PH-LHD and PH-lung*

Two risk stratification variables, REVEAL group and WHO functional class were also examined alongside survival between clusters. Although the REVEAL risk groups did not show significant differences between clusters, the two clusters with poor survival, C and F had the two lowest percentage compositions of low risk group patients, 47.3% and 45.9 respectively% (Figure 4.7). Similarly, WHO functional class again did not show significant differences between groups; however, Cluster A had the lowest percentage of functional class IV patients (9.2%). Clusters C and E had the highest percentage of functional class IV patients (14.4% and 16.7% respectively) (Figure 4.8).

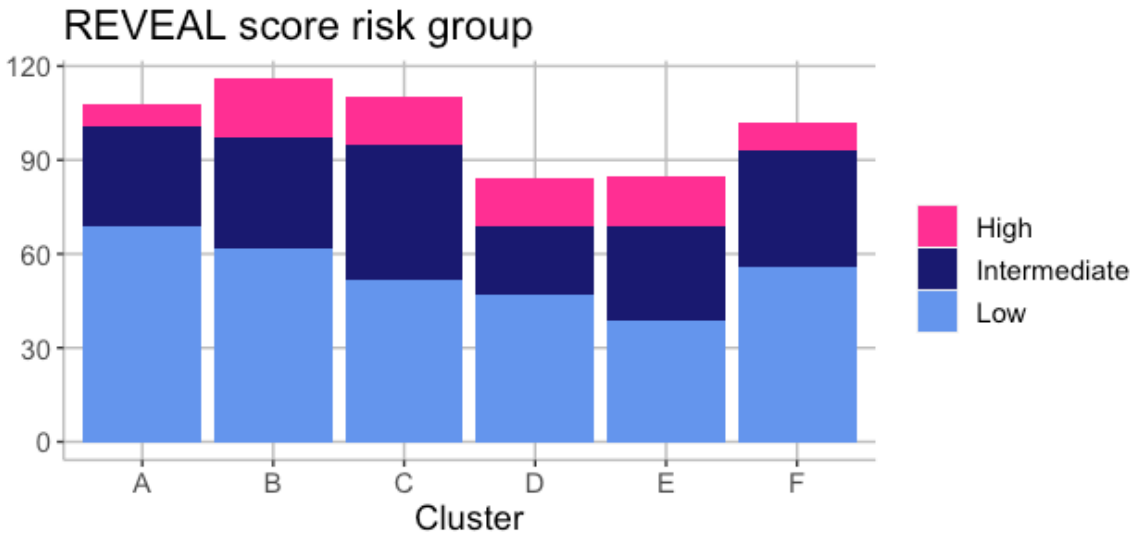


***Figure 4.7**: REVEAL group breakdown by cluster. Cluster A had the best survival outcomes, with the lowest percentage of high risk patients (6.5%) and the largest percentage of low risk patients (63.9). Clusters C and E had the worst survival outcomes, with the smallest percentage of low risk patients (47.3% and 45.9% respectively).*

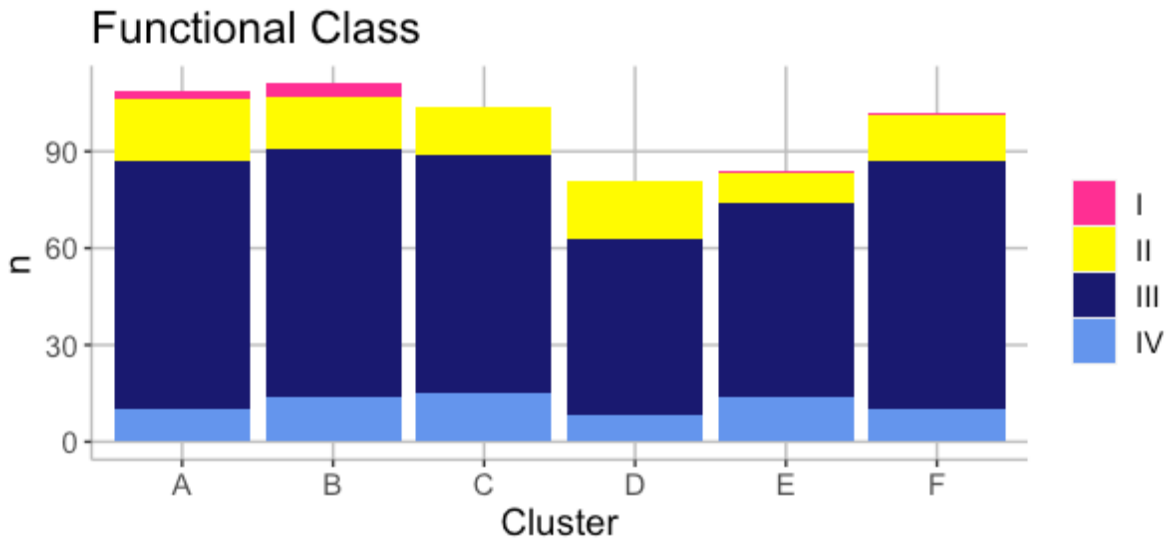


***Figure 4.8**: WHO functional class breakdown by cluster. Cluster A had the best survival outcomes, and the lowest percentage of functional class 4 patients (9.2%). Clusters C and E had the worst survival outcomes and the highest percentage of functional class 4 patients (14.4% and 16.7% respectively).*

4.3.2.2 Biomarker Associations

Clinical measures as recorded for each patient at the time of sampling were assessed for statistically significant differences between clusters (Table 4.4 and Table 4.5). There were no significant differences between the clusters for demographic variables such as age, BMI and sex (adjusted p-values of 0.931, 0.146 and 0.591 respectively). None of the comorbidities noted were present in significantly different proportions between clusters. However, several of the clusters showed significantly different levels of important clinical variables (Figure 4.9).

NT-proBNP had a significantly higher median value (2066) in cluster C compared to all the other clusters, with the exception of Cluster E which also had poor survival (adjusted p-value 4.97e-04). Cluster C also had significantly higher mRAP values compared to all other clusters (adjusted p-value 1.18e-03). Cluster A, with the best survival, had significantly lower mPAP and dPAP than clusters B, C, and E (adjusted p-values 1.89e-03 and 0.0215 respectively).

***Table 4.4**: P-values for clinical variable association to cluster groups for discovery & interim sets. Post-hoc tests were carried out for continuous variables with a p-value < 0.05.*

Clinical Parameter	**P-value**	**Adjusted P-value**	**Significant post-hoc tests**
Treatment	4.54e-07	1.59e-05	
NT-proBNP	2.84e-05	4.97e-04	A/C, B/C, C/D, C/F, D/E
mRAP	1.01e-04	1.18e-03	A/C, B/C, C/D, C/E, C/F
sPAP	1.36e-04	1.19e-03	A/B, A/C, A/D, A/E, E/F
SvO2	1.83e-04	1.28e-03	A/C, A/E, B/C, B/E, C/D, C/F, D/E, E/F
mPAP	3.24e-04	1.89e-03	A/B, A/C, A/E, E/F
Site	5.00e-04	2.50e-03	
PVR	4.99e-03	0.0215	A/C, A/E, E/F
dPAP	5.52e-03	0.0215	A/B, A/C, A/E
Creatinine	0.0116	0.0406	C/D
Uric Acid	0.0374	0.111	C/D
6MWD	0.0379	0.111	A/E, D/E, E/F
eGFR	0.0455	0.123	C/D
BMI	0.0582	0.146	
REVEAL risk group	0.0685	0.160	
Comorbidity: Atrial Fibrillation	0.115	0.231	
Comorbidity: CTD	0.108	0.231	
Comorbidity: Scleroderma	0.119	0.231	
Comorbidity: Diabetes	0.153	0.283	
TLCO predicted	0.169	0.296	

Clinical Parameter	P-value	Adjusted P-value	Significant post-hoc tests
Cardiac Index	0.206	0.344	
Comorbidity: COPD	0.225	0.358	
Comorbidity: Ischaemic heart disease	0.344	0.524	
Comorbidity: Thyroid Disease	0.374	0.545	
Predicted forced vital capacity	0.394	0.552	
Sex	0.473	0.591	
PAWP	0.452	0.591	
Cardiac Output	0.463	0.591	
Functional Class	0.496	0.599	
ISWT	0.570	0.665	
BPSYS	0.625	0.685	
Platelet Count	0.627	0.685	
Age	0.931	0.931	
BPDIA	0.925	0.931	
Comorbidity: Sleep Apnea	0.916	0.931	

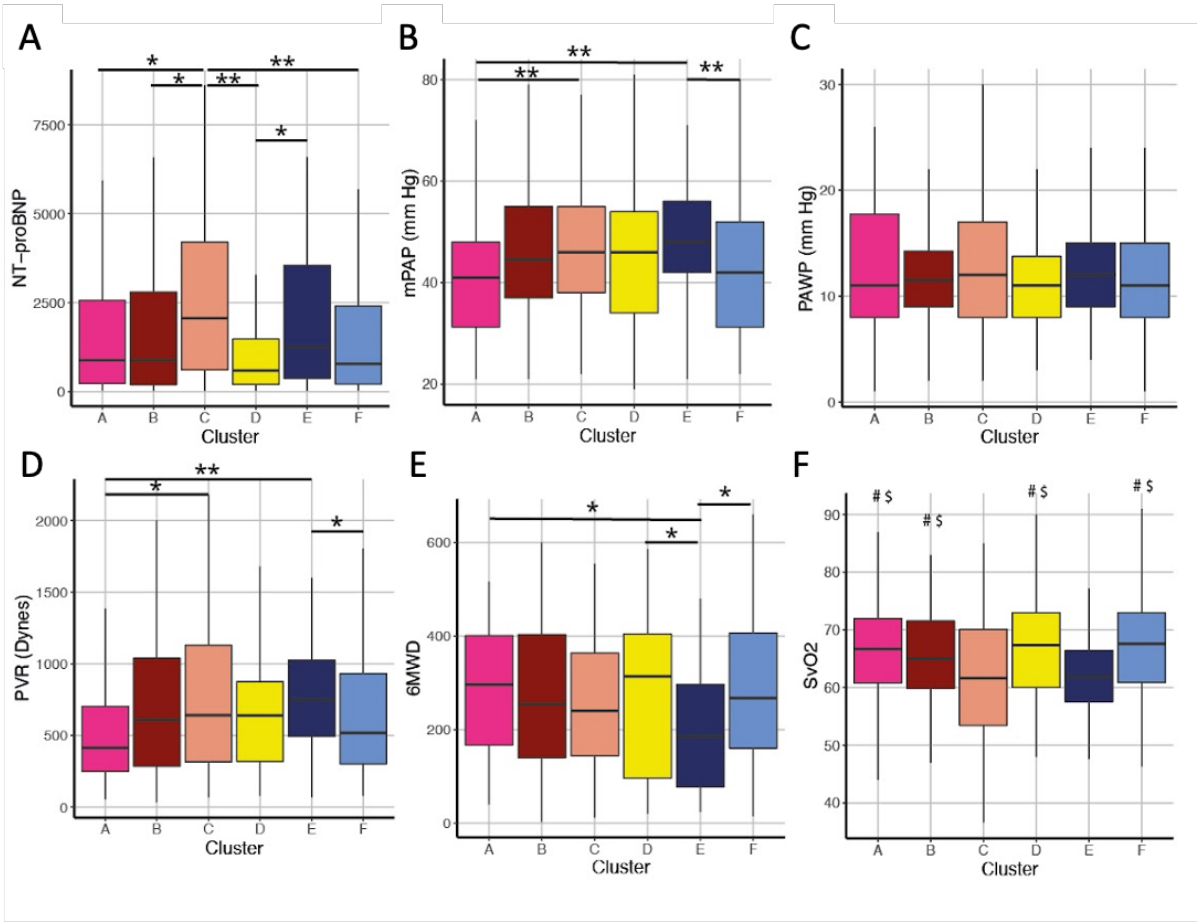


***Figure 4.9**: Clinical characteristics of the six miRNA clusters. Box and Whisker plots show patients' quantitative traits with outliers removed for A) Serum NT-proBNP levels, B) mean PAP, C) PAWP, D) PVR, E) six-minute walk distance, and F) SvO2 within each miRNA Cluster. * p<0.05, ** p<0.01 on specific comparisons. For H) $ p<0.05 compared to cluster C, and # p<0.05 compared to cluster E following Kruskal-Wallis chi-squared followed by Benjamini-Hochberg post-hoc analysis.*

***Table 4.5**: Main clinical characteristics for the 6 clusters across the training & interim cohorts at the time of sampling. Values described as mean (standard deviation) for normally distributed variables, or median [IQR] for variables which are not normally distributed. Categorical variables described as number (percentage of available data).*

Clinical Variable	***A***	***B***	***C***	***D***	***E***	***F***	***All patients***
n	*110 (17.9%)*	*118 (19.2%)*	*110 (17.9%)*	*84 (13.7%)*	*86 (14.0%)*	*106 (17.2%)*	*614*
PH classification							
PH1.1	*43 (39.1%)*	*49 (41.5%)*	*34 (30.9%)*	*35 (41.7%)*	*34 (39.5%)*	*34 (32.1%)*	*229 (37.3%)*
PH1.2	*3 (2.7%)*	*1 (0.9%)*	*4 (3.6%)*	*2 (2.4%)*	*1 (1.2%)*	*1 (0.9%)*	*12 (2.0%)*
PH1.3	*1 (0.9%)*	*2 (1.7%)*	*0 (0%)*	*1 (1.2%)*	*0 (0%)*	*0 (0%)*	*4 (0.1%)*
PH1.4	*17 (15.5%)*	*24 (20.3%)*	*23 (20.9%)*	*21 (25.0%*	*27 (31.4%)*	*37 (34.9%)*	*149 (24.3%)*
PH2	*36 (32.7%)*	*24 (20.3%)*	*29 (26.4%)*	*12 (14.3%)*	*13 (15.1%)*	*23 (21.7%)*	*137 (22.3%)*
PH3	*10 (9.1%)*	*18 (15.3%)*	*20 (18.2%)*	*13 (15.5%)*	*11 (12.8%)*	*11 (10.4%)*	*83 (13.5%)*
Sex: Female	*80 (72.7%)*	*79 (66.9%)*	*67 (60.9%)*	*54 (64.3%)*	*53 (61.6%)*	*72 (67.9%)*	*405 (66.0%)*
Functional Class							
1	*3 (2.8%)*	*4 (3.6%)*	*0 (0%)*	*0 (0%)*	*1 (1.2%)*	*1 (1.0%)*	*9 (1.5%)*
2	*19 (17.4%)*	*16 (14.4%)*	*15 (14.4%)*	*18 (22.2%)*	*9 (10.7%)*	*14 (13.7%)*	*91 (15.4%)*
3	*77 (70.6%)*	*77 (69.4%)*	*74 (71.2%)*	*55 (67.9%)*	*60 (71.4%)*	*77 (71.4%)*	*420 (71.1%)*
4	*10 (9.2%)*	*14 (12.6%)*	*15 (14.4%)*	*8 (9.9%)*	*14 (16.7%)*	*10 (16.7%)*	*71 (12.0%)*
REVEAL risk group							
High	*7 (6.5%)*	*19 (16.4%)*	*15 (13.6%)*	*15 (17.9%)*	*16 (18.8%)*	*9 (8.8%)*	*81 (13.3%)*
Intermediate	*32 (29.6%)*	*35 (30.2%)*	*43 (39.1%)*	*22 (26.2%)*	*30 (35.3%)*	*37 (36.3%)*	*199 (32.9%)*

Low	*69 (63.9%)*	*62 (53.4%)*	*52 (47.3%)*	*47 (56.0%)*	*39 (45.9%)*	*56 (54.9%)*	*325 (53.7%)*
Body Mass Index	*27.2 [8.8]*	*27.0 [7.0]*	*27.5 [11.0]*	*27.7 [9.2]*	*29.1 [7.8]*	*25.1 [8.3]*	*27.2 [9.0]*
Age (years)	*62.3 [21.0]*	*64.1 [23.0]*	*66.0 [15.8]*	*64.2 [23.7]*	*65.9 [19.8]*	*65.0 [21.3]*	*65.0 [21.1]*
FVCP	*89.6 (22.2)*	*92.5 (24.6)*	*85.3 (22.4)*	*86.9 (21.7)*	*85.8 (23.5)*	*86.8 (22.1)*	*88.0 (22.9)*
NT-proBNP (log2 pg/ml)	*9.79 [3.48]*	*9.79 [3.82]*	*11.0 [2.79]*	*9.21 [2.87]*	*10.3 [3.25]*	*9.61 [3.50]*	*9.95 [3.49]*
Uric Acid	*7.0 [3.75]*	*7.0 [4.0]*	*8.0 [3.0]*	*7.0 [3.0]*	*7.0 [3.0]*	*7.0 [3.75]*	*7.0 [3.0]*
6 minute walking distance (m)	*296 [233]*	*255 [263]*	*240 [220]*	*314 [309]*	*186 [218]*	*268 [246]*	*265 [249]*
ISWT	*220 [168]*	*160 [265]*	*200 [208]*	*150 [210]*	*150 [200]*	*150 [190]*	*170 [210]*
Blood pressure - diastolic (mm Hg)	*75.0 [16.0]*	*74.0 [18.0]*	*73.5 [15.3]*	*72.0 [11.0]*	*74.5 [14.0]*	*73.0 [18.0]*	*74.0 [16.0]*
Blood pressure - systolic (mm Hg)	*126 [29.5]*	*127 [31.5]*	*128 [27.5]*	*127 [35.8]*	*125 [30.0]*	*132 [32.0]*	*127 [31.0]*
Pulmonary Vascular Resistance (dynes)	*414 [455]*	*610 [754]*	*642 [813]*	*640 [813]*	*751 [531]*	*519 [631]*	*589 [640]*
Mean pulmonary artery pressure (mm Hg)	*41.0 [16.8]*	*44.5 [18.0]*	*46.0 [17.0]*	*46.0 [20.0]*	*48.0 [14.0]*	*42.0 [20.8]*	*44.0 [19.0]*
sPAP (mm Hg)	*64.5 [32.3]*	*73.0 [34.0]*	*74.0 [29.5]*	*74.0 [36.5]*	*78.0 [29.0]*	*68.0 [33.3]*	*72.0 [32.0]*
dPAP (mm Hg)	*24.0 [10.8]*	*30.0 [13.0]*	*30.0 [13.5]*	*28.0 [14.0]*	*29.0 [11.0]*	*26.0 [16.3]*	*27.0 [14.0]*
Mean right atrial pressure (mm Hg)	*9.0 [6.5]*	*10.0 [8.0]*	*13.0 [8.0]*	*9.0 [6.0]*	*10 [6.0]*	*9.0 [7.5]*	*10.0 [8.0]*
SvO2 (%)	*66.7 [11.2]*	*65.0 [11.7]*	*61.7 [16.6]*	*67.4 [13.0]*	*61.9 [8.95]*	*67.6 [12.1]*	*65.0 [12.6]*
TLco predicted	*58.0 [30.0]*	*54.0 [30.7]*	*48.0 [23.6]*	*51.0 [29.5]*	*47.0 [27.3]*	*55.0 [29.5]*	*53.0 [29.0]*
Pulmonary	*11.0*	*11.5*	*12.0*	*11.0*	*12.0*	*11.0*	*11.0*

arterial wedge pressure	*[9.75]*	*[5.25]*	*[9.0]*	*[5.75]*	*[6.0]*	*[7.0]*	*[7.0]*
Platelet Count	*231 [103]*	*216 [95.8]*	*210 [97.3]*	*205 [88.0]*	*221 [113]*	*221 [107]*	*218 [101]*
eGFR	*69.0 [30.0]*	*70.5 [37.3]*	*55.7 [36.5]*	*72.0 [29.0]*	*66.3 [35.5]*	*69.0 [31.0]*	*68.0 [33.2]*
Cardiac Output (L/min)	*4.05 [2.02]*	*4.20 [1.94]*	*3.73 [2.12]*	*3.93 [1.91]*	*3.90 [2.0]*	*4.21 [2.34]*	*3.97 [2.14]*
Cardiac Index (L/min/m2)	*2.40 [1.08]*	*2.40 [1.04]*	*2.10 [1.15]*	*2.28 [1.11]*	*2.04 [1.04]*	*2.38 [1.24]*	*2.26 [1.15]*
Creatinine	*85.0 [30.0]*	*83.0 [40.0]*	*95.5 [55.5]*	*80.0 [30.0]*	*85.0 [42.0]*	*82.0 [35.8]*	*84.0 [39.3]*
Comorbidity							
COPD	*11 (10.0%)*	*16 (13.6%)*	*21 (19.1%)*	*6 (7.1%)*	*10 (11.6%)*	*12 (11.3%)*	*76 (12.4%)*
Sleep Apnoea	*6 (5.5%)*	*3 (2.5%)*	*5 (4.5%)*	*4 (4.8%)*	*4 (4.7%)*	*5 (4.7%)*	*27 (4.4%)*
Atrial Fibrillation	*23 (20.9%)*	*16 (13.6%)*	*29 (26.4%)*	*11 (13.1%)*	*18 (20.9%)*	*18 (17.0%)*	*115 (18.7%)*
Connective Tissue Disease (excluding SSc)	*13 (11.8%)*	*6 (5.1%)*	*10 (9.1%)*	*5 (6.0%)*	*14 (16.3%)*	*9 (8.5%)*	*57 (10.9%)*
Scleroderma	*10 (9.1%)*	*12 (10.2%)*	*12 (10.9%)*	*9 (10.7%)*	*14 (16.3%)*	*22 (20.8%)*	*79 (12.9%)*
Type 2 diabetes mellitus	*13 (11.8%)*	*16 (13.6%)*	*9 (8.2%)*	*18 (21.4%)*	*14 (16.3%)*	*14 (13.2%)*	*84 (13.7%)*
Thyroid disease	*4 (3.6%)*	*1 (0.8%)*	*3 (2.7%)*	*3 (3.6%)*	*0 (0%)*	*3 (2.8%)*	*14 (2.3%)*
Ischaemic heart disease	*2*	*3 (2.5%)*	*5 (4.5%)*	*1 (1.2%)*	*6 (7.0%)*	*4 (3.8%)*	*21 (3.4%)*
Treatment							
Treatment naive	*38 (34.5%)*	*45 (38.1%)*	*20 (18.2%)*	*47 (56.0%)*	*19 (22.1%)*	*32 (30.2%)*	*201 (32.7%)*
ERA	*26 (23.6%)*	*32 (27.1%)*	*10 (9.1%)*	*24 (28.6%)*	*11 (12.8%)*	*24 (22.6%)*	*127 (20.7%)*
Prostanoid	*9 (8.2%)*	*8 (6.8%)*	*5 (4.5%)*	*3 (3.6%)*	*3 (3.5%)*	*7 (6.6%)*	*35 (5.7%)*

Other PH drug	*4 (3.6%)*	*3 (2.5%)*	*0 (0%)*	*1 (1.2%)*	*1 (1.2%)*	*1 (0.9%)*	*10 (1.6%)*

4.3.2.3 MicroRNA cluster signatures

LASSO models were fitted to the clusters to identify some key miRNAs for each cluster. After training the models on the training data set, the performance was measured in the interim dataset (Table 4.6). The signature for Cluster B showed a lower performance (AUC 0.592 in the interim set), but the other five clusters all had high performance AUCs (0.87-0.97 in the training set, and 0.80-0.96 in the interim set).

Table 4.6*: Performance for six miRNA LASSO models classifying distinct clusters showing mean cross validated (CV) AUC on the training set and an AUC for the interim set.*

Cluster	**CV mean AUC (sd)**	**Interim AUC (95% CI)**
Cluster A	0.92 (0.034)	0.91 (0.82 - 0.99)
Cluster B	0.69 (0.084)	0.59 (0.45 - 0.74)
Cluster C	0.90 (0.048)	0.88 (0.80 - 0.95)
Cluster D	0.97 (0.020)	0.94 (0.89 - 0.98)
Cluster E	0.94 (0.031)	0.96 (0.93 - 1.0)
Cluster F	0.87 (0.046)	0.80 (0.72 - 0.88)

The combination of individual miRNA was unique to each cluster. However several signatures selected the same miRNAs, albeit with different coefficient values, in many cases in opposite directions (Figure 4.10). Each cluster was defined by between 7-12 miRNAs, with 38 miRNAs selected in total. Cluster A had the largest absolute miRNA coefficient value, hsa-miR-4257 (-0.805). Hsa-miR-761 was selected as a marker for all clusters except ClusterA. Hsa-miR-761 inhibits mitochondrial fission and commentators have suggested that modulation of their levels may help tackle apoptosis and myocardial infarction (Long et al. 2013). The cluster with the highest coefficient for miR-761 was cluster E, a cluster with poor survival outcomes. The expression levels in the miRNA signatures across the clusters can be seen in Figure 4.11.

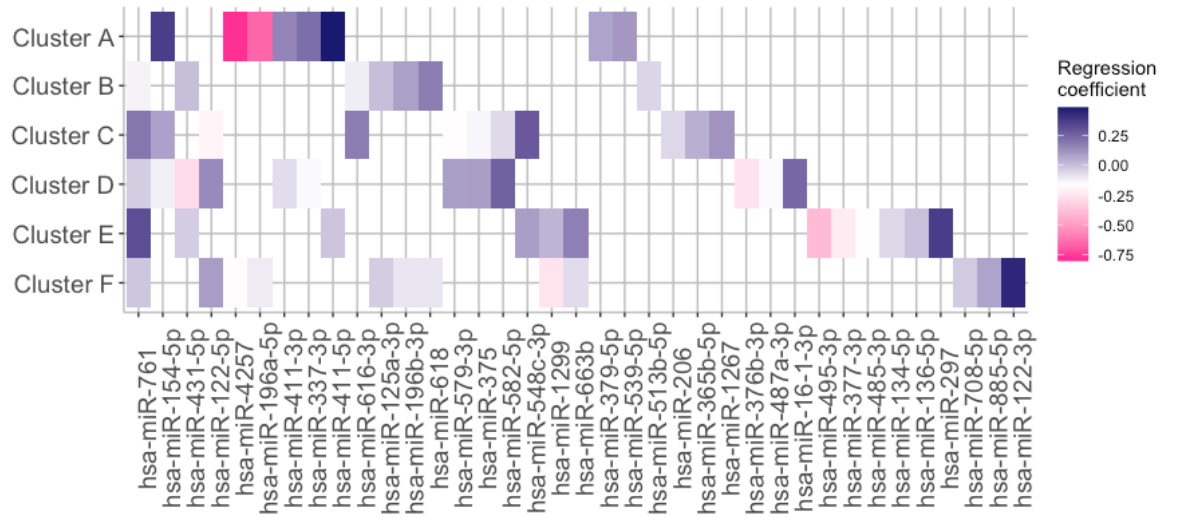
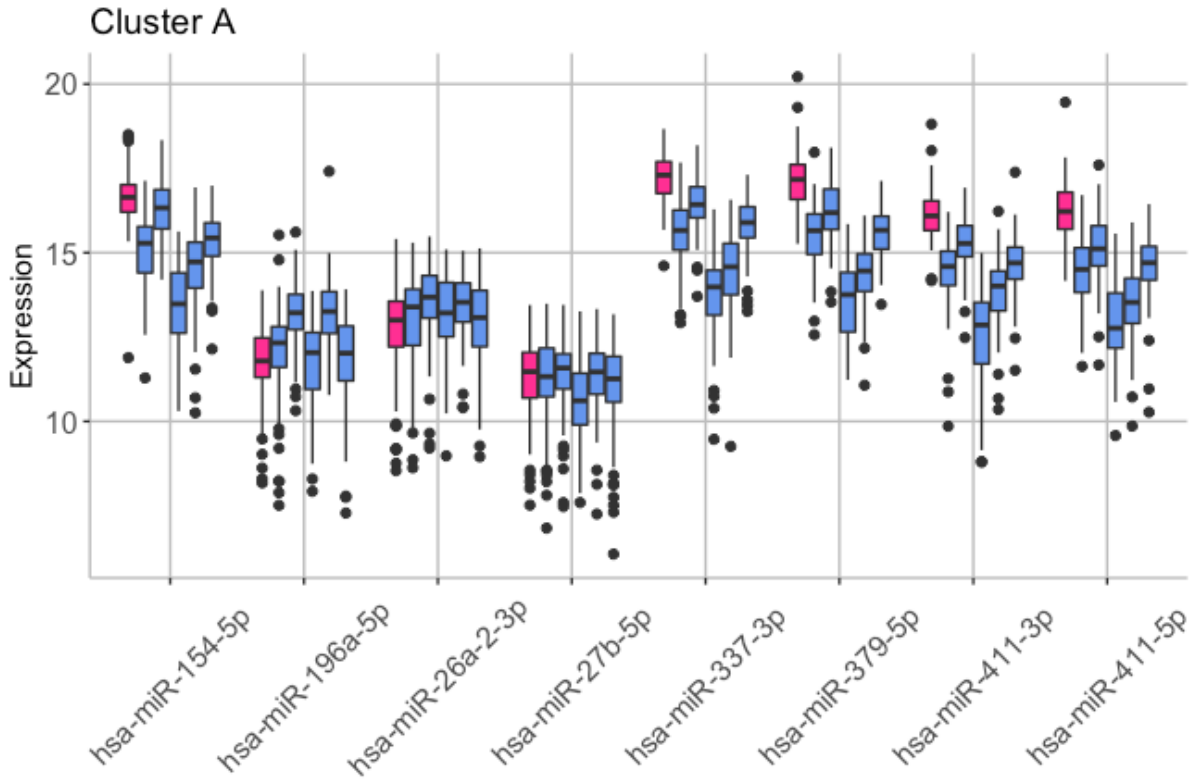
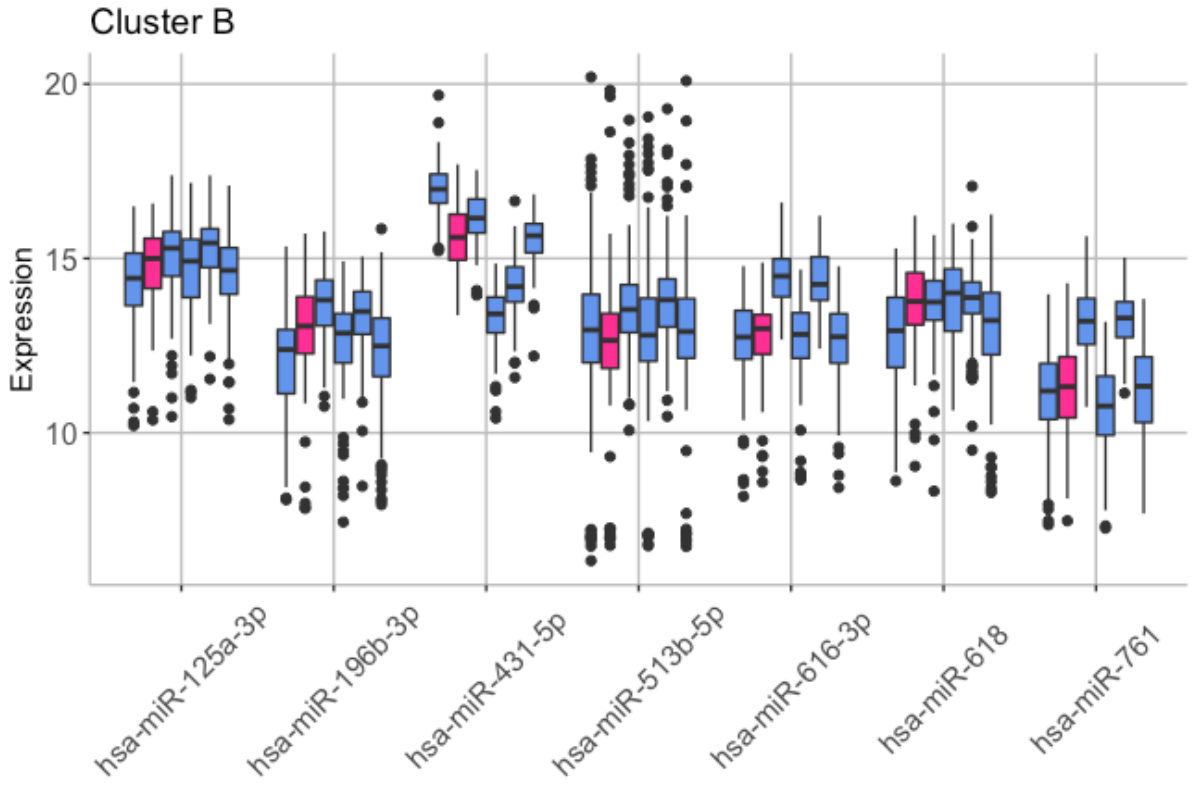
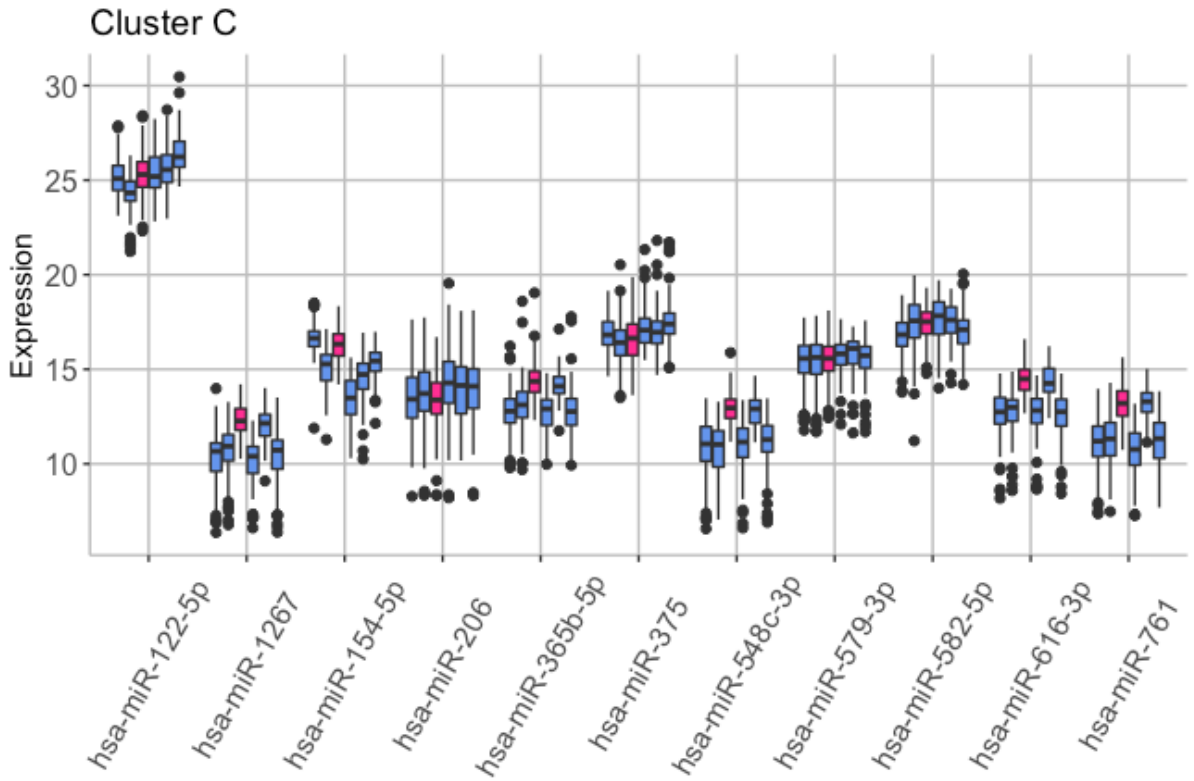
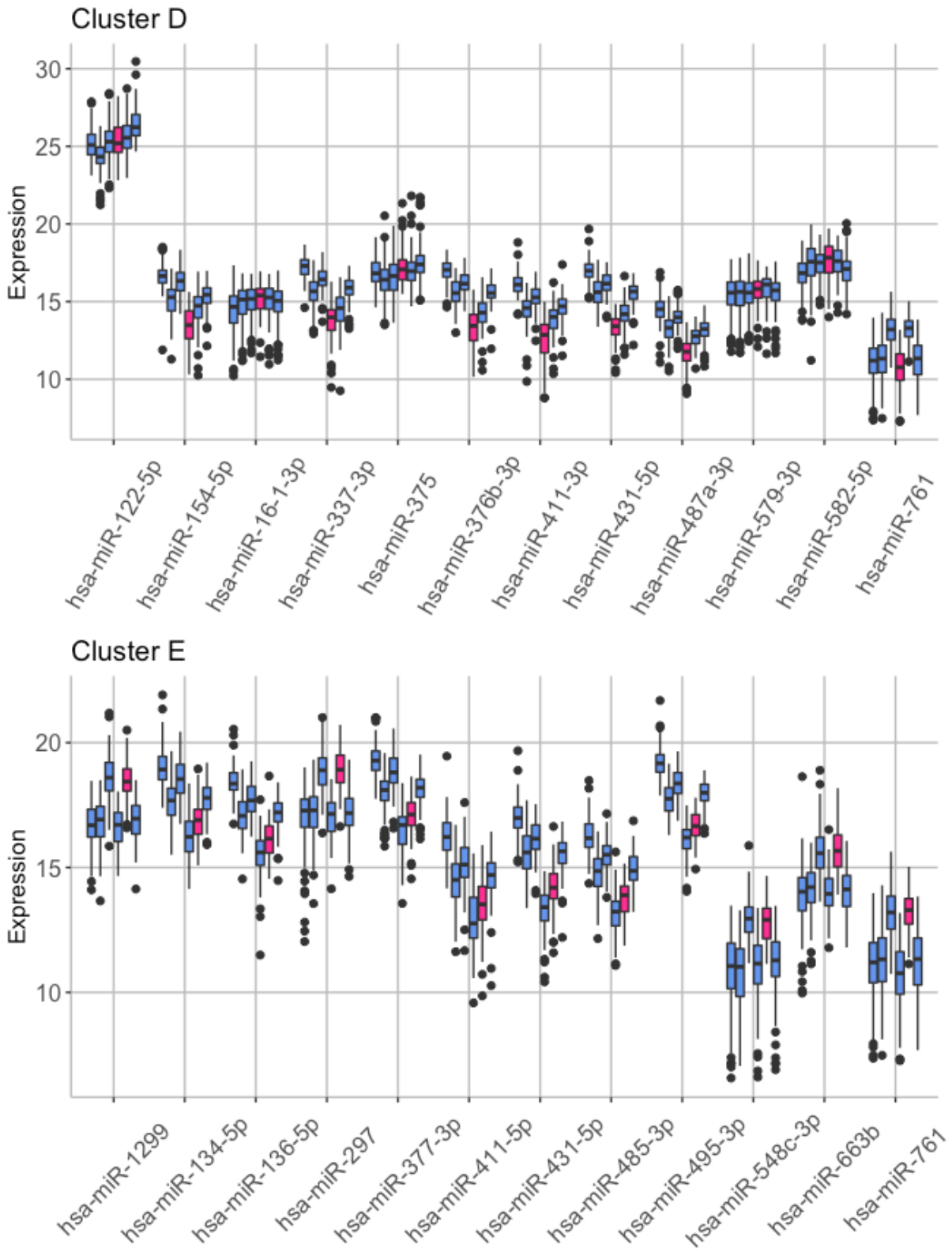


***Figure 4.10**: miRNA coefficients for individual LASSO models describing six clusters*




Cluster B
Expression
20
15
10
hsa-miR-125a-3p
hsa-miR-196b-3p
hsa-miR-431-5p
hsa-miR-513b-5p
hsa-miR-616-3p
hsa-miR-618
hsa-miR-761


Cluster C
Expression
30
25
20
15
10
hsa-miR-122-5p
hsa-miR-1267
hsa-miR-154-5p
hsa-miR-206
hsa-miR-365b-5p
hsa-miR-375
hsa-miR-548c-3p
hsa-miR-579-3p
hsa-miR-582-5p
hsa-miR-616-3p
hsa-miR-761


Cluster D
Expression
30
25
20
15
10
hsa-miR-122-5p
hsa-miR-154-5p
hsa-miR-16-1-3p
hsa-miR-337-3p
hsa-miR-375
hsa-miR-376b-3p
hsa-miR-411-3p
hsa-miR-431-5p
hsa-miR-487a-3p
hsa-miR-579-3p
hsa-miR-582-5p
hsa-miR-761
Cluster E
Expression
20
15
10
hsa-miR-1299
hsa-miR-134-5p
hsa-miR-136-5p
hsa-miR-297
hsa-miR-377-3p
hsa-miR-411-5p
hsa-miR-431-5p
hsa-miR-485-3p
hsa-miR-495-3p
hsa-miR-548c-3p
hsa-miR-663b
hsa-miR-761

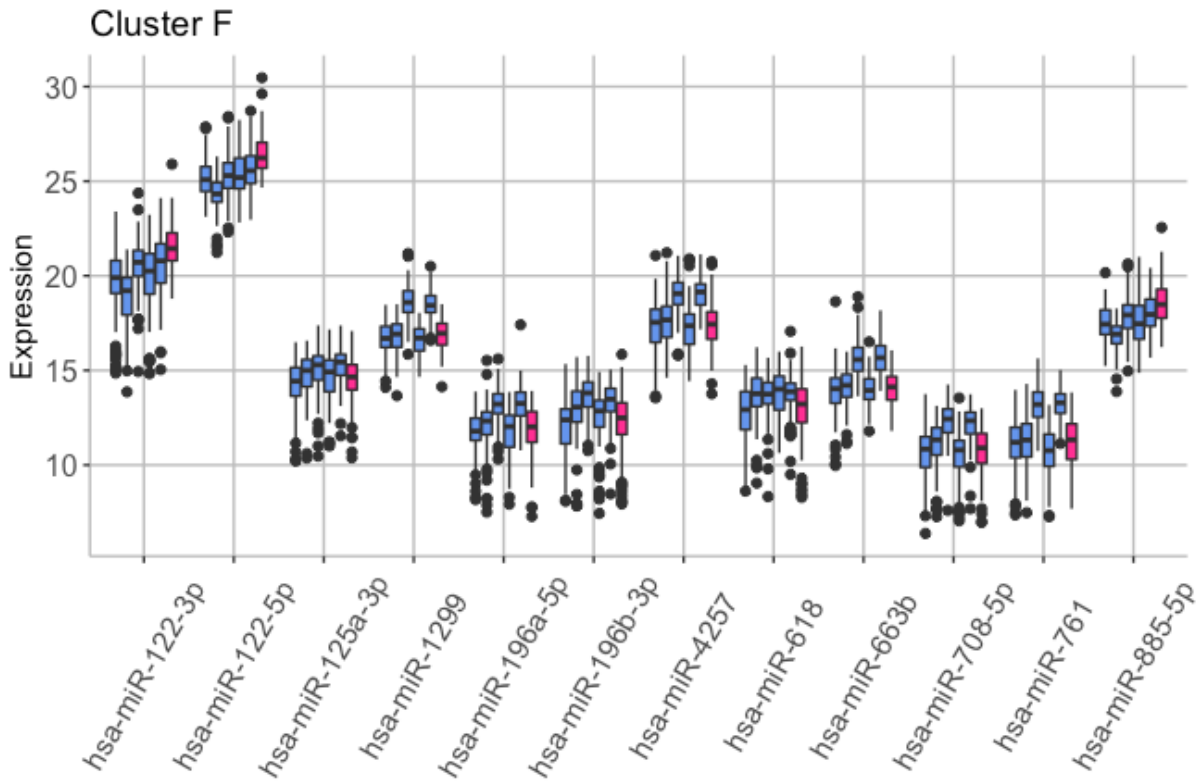


***Figure 4.11**: Expression levels of miRNA signatures to distinguish between clusters A-F in the training and interim sets.*

4.3.2.4 Clinical signatures for Clusters

Clinical signatures were derived for each cluster next by Emmanuel Jammeh, using imputed values where data was unavailable. Signatures were derived for a range of signatures sizes. Again, Cluster B showed the lowest AUC (Figure 4.12). NT-proBNP features in only one of the clinical signatures, cluster C. Although there didn't appear to be significant differences in functional class between clusters in a univariable analysis, functional class was a driving factor in the signature for Cluster A, along with creatinine and eGFR. In cluster C, with significantly worse survival outcomes, the driving factors were mRAP, mPAP and PAH.

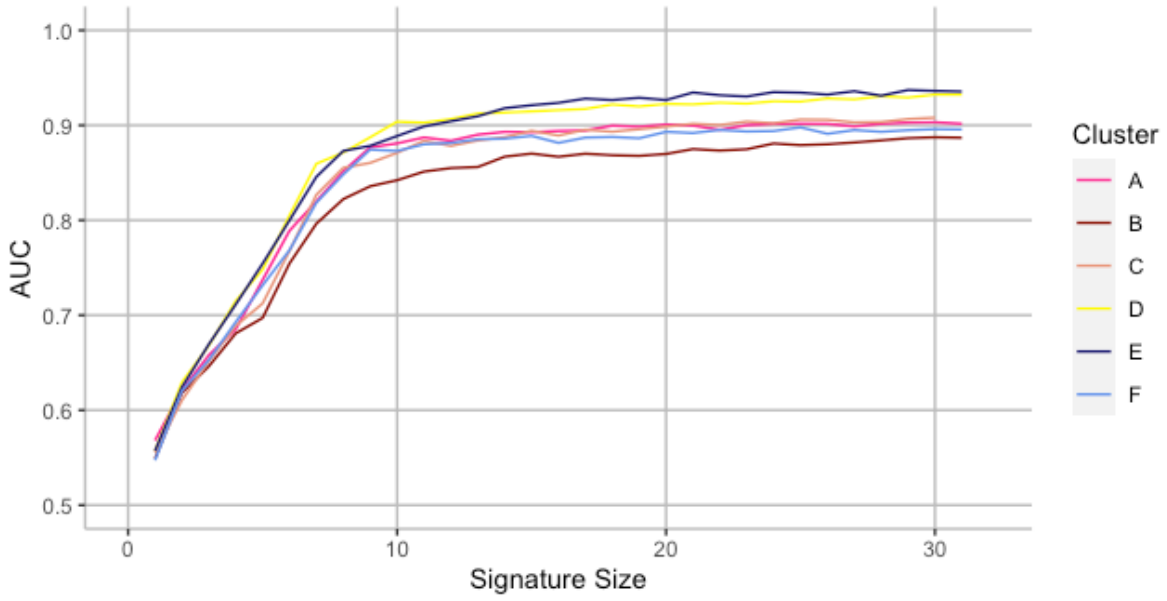


***Figure 4.12**: Classification performance for signatures of different sizes for each cluster. The results suggest that clinical signatures comprising 15 features may be sufficient to adequately describe each cluster with acceptable classification performance with AUC of 0.8. Figure generated from data provided by Emmanuel Jammeh*

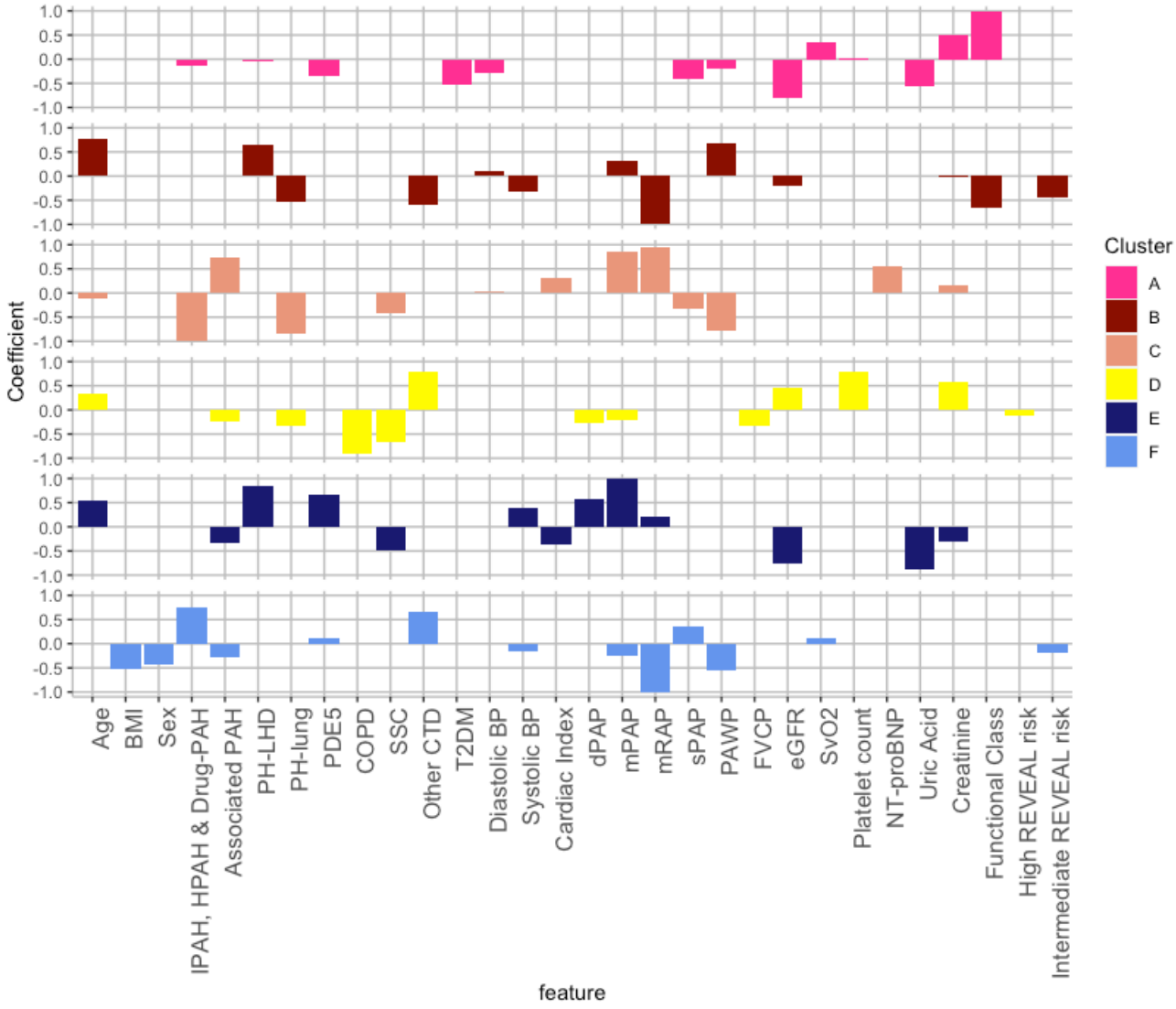


***Figure 4.13**: Selected clinical signatures and the coefficients of each feature in each subgroup. Figure generated from data provided by Emmanuel Jammeh*

4.3.2.5 Pathways

The miRNAs in the signatures for each cluster were linked to potential target genes and then enriched for pathways. The analysis highlighted distinct pathways for each miRNA cluster signature (Figure 4.14). Several of these have been previously found to be associated with PH, such as leptin insulin overlap (Cluster B), focal adhesion (Cluster E) and interleukin families (Clusters C, D and E).

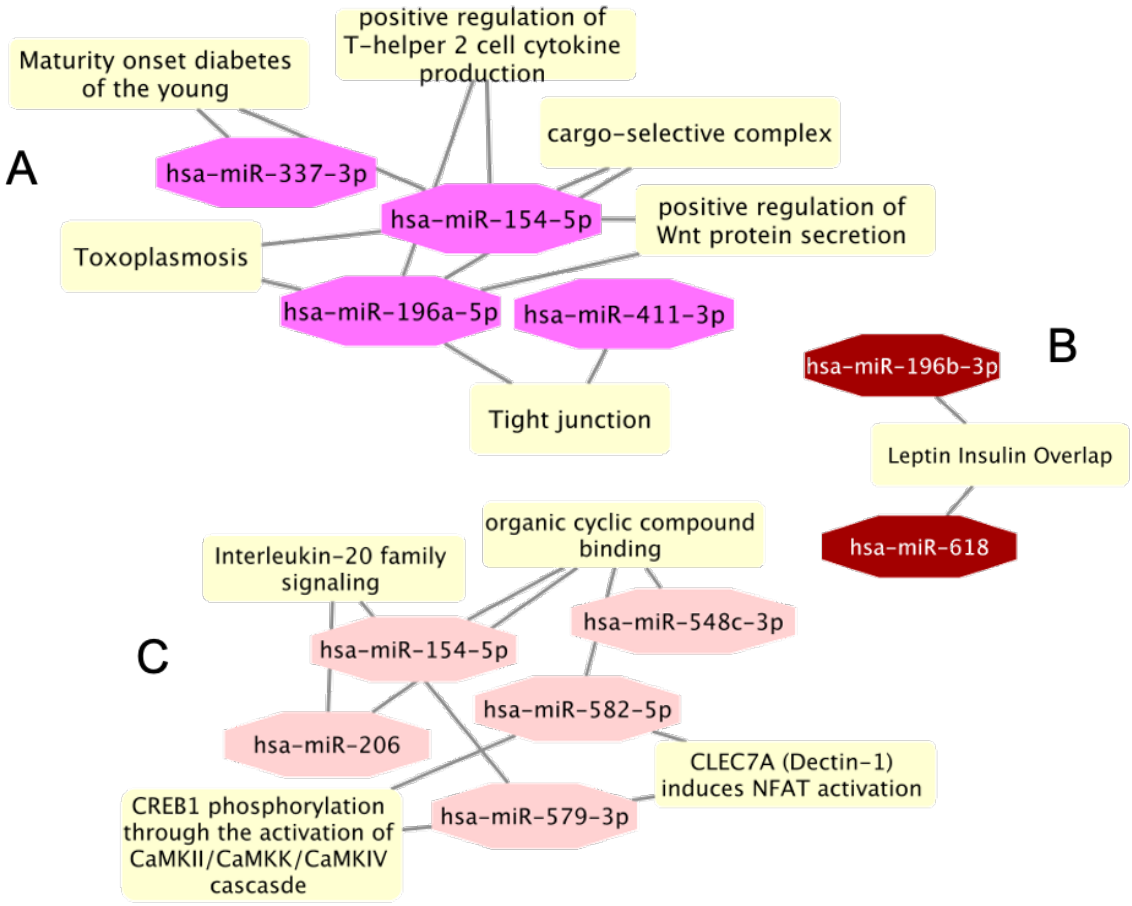


***Figure 4.14**: Enriched pathways for miRNA signatures for (A) Cluster A, (B) Cluster B and (C) Cluster C*

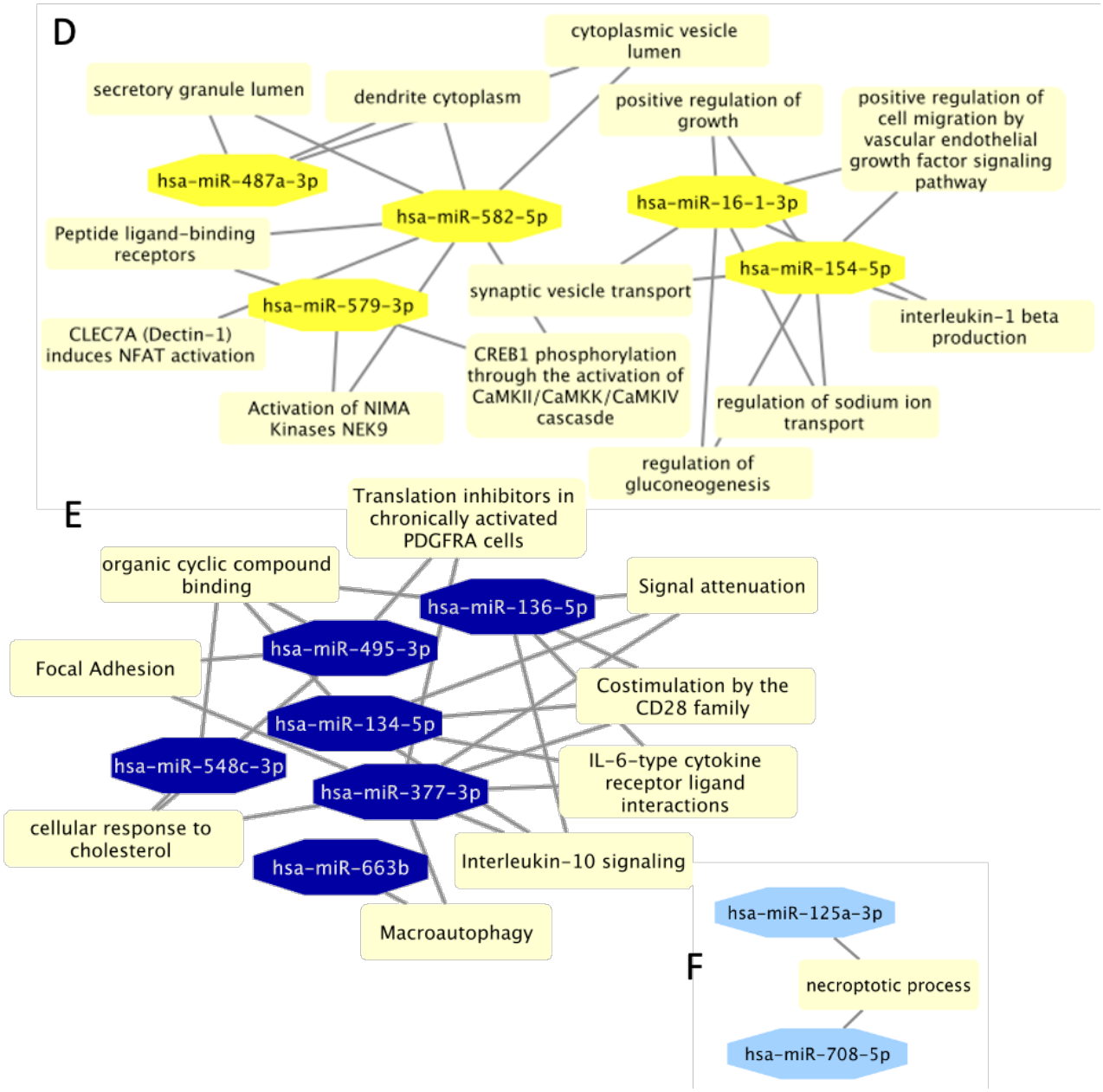


Figure 4.14 *cont: Enriched pathways for miRNA signatures for (D) Cluster D, (E) Cluster E, and (F) Cluster F*

4.4 Discussion

We began our analysis by looking at the heterogeneity within each clinical classification, with two clusters identified as the optimal number of clusters for each clinical classification group. Although only 1 clinical variable was seen to be significantly different between clusters (eGFR in CTEPH patients), there was a significant difference in survival seen at 5 years between clusters A and B in patients with PH-LHD. This is a potential avenue for future research, delving into further differences between the clusters which could lead to these survival differences. For example, the distribution of vasodilator responders within the clusters, or evidence of some pre-capillary involvement.

Diagnostic challenges for patients with PAH, PH-LHD or PH-lung remain, which undermines the ability to create clear diagnostic signatures. An unsupervised machine learning approach identified six molecular clusters from the miRNAs assigned to the clinically defined PAH, PH-LHD and PH-lung clinical groups. All three clinical groups were represented across the clusters. A closer examination of clinical features showed that a range of variables showed significant differences between clusters; including the haemodynamic variables (sPAP, mRAP, mPAP, PVR & dPAP), as well as lung function (SvO2) and biochemical variables (NT-proBNP). Most notably Cluster A was associated with the best survival, low mPAP, low PVR but not the lowest NT-proBNP, while Clusters C and F were associated with the worst survival but were associated with different clinical feature profiles. Cluster C was defined by high NT-proBNP, mPAP and PVR while Cluster F was defined by significantly lower NT-pro BNP. The characteristics highlight the challenges of using single biomarkers (e.g. NT-proBNP) or clinical features (PVR) to attribute risk and molecular mechanisms.

All the clusters except for E also had some overlap in miRNAs with the signatures for discriminating PH clinical classification groups seen in Chapter 3. Clusters C and D selected miR-375 in their signatures, a miRNA found in the CTEPH vs other PH signature. Cluster D and cluster F also selected miR-16-1-3p, a miRNA found in both the CTEPH vs other PH and PAH vs CTEPH signatures. MiR-513b-5p, part of cluster B's signature, was also found in the PH-lung vs other PH groups signature. Finally, the miRNA signature for cluster A contained miR-26a-2-3p, a miRNA found in the PAH vs DC signature. Although CTEPH patients were excluded from this clustering analysis, the shared miRNAs in the signatures suggest there is an element of shared molecular pathology between CTEPH and subgroups of other types of PH, which could be explored in the future by expanding the range of patients clustered.

Several pathways of interest were found to be enriched within clusters. For example, a range of cytokines have been associated with pulmonary hypertension as there is evidence in animal models suggesting inflammation may contribute to the development of pulmonary hypertension, especially PAH (Groth et al. 2014).

Leptin insulin overlap was shown to be enriched in cluster B. Leptin signalling has been shown to be involved in a range of different cardiac pathologies, as well as the proliferation of pulmonary arterial smooth muscle cells by activating extracellular signal-regulated kinase (ERK), signal transducer and activator of transcription 3 (STAT3), and Akt pathways (Chai et al. 2015). The association of leptin-insulin signalling within cluster B may offer a way of selecting patients for further studies in this pathway, recognising that insulin resistance may also be present as co-morbidity of PH-LHD and PH-lung.

Focal adhesion, a pathway enriched in cluster E, involves the regulation of cell migration. Suppressing PASMC migration by inhibiting focal adhesion kinase has been shown to inhibit the progression of PAH, and has been highlighted as a potential therapeutic target (Paulin et al. 2014). Likewise, another enriched pathway in cluster E, IL-6 remains of interest in PAH (Toshner and Rothman 2020), despite a recent report that IL-6 antagonists in an unselected PAH population showed no benefit (Toshner et al. 2022), targeting a subgroup of PH based on their miRNA signature remains an option.

Although appearing distinct, there is considerable overlap between supervised and unsupervised machine learning. In the previous chapter, we used machine learning models

to attempt to solve a classification problem. The unsupervised approach demonstrated that the traditional clinical classifications are not fully representative of the underlying molecular heterogeneity.

Looking forward, another consideration to make is whether miRNAs are the right medium to examine heterogeneity within PH. When the methodology for unsupervised clustering was applied to an RNAseq cohort of patients with IPAH and HPAH, six distinct clusters were uncovered (Kariotis et al. 2021), with significant differences in survival and clinical parameters. In this cohort, an examination of PAH patients alone denoted two clusters as the optimal number. Although the make-up of patients differed slightly (with APAH patients included here), it may be that RNAseq is better placed to provide more insights. Alternatively, proteomics could be considered.

Chapter 5: Conclusion

This thesis documents the use of miRNAs in classifying patients with Pulmonary Hypertension in both a pilot study and a large cohort. I explore the use of four supervised machine learning methods to create miRNA diagnostic signatures before examining the outputs from an unsupervised approach.

In Chapter 2 I used a consensus of multiple machine learning approaches to identify two miRNAs that were able to distinguish PAH from both disease and healthy controls. The study was the largest microRNA profiling of PAH patients with 64 treatment naïve patients, and 43 disease and healthy controls at the time. It was also the first machine learning assessment of microRNAs for PAH.

The miRNAs identified (miR-636 and miR-187-5p) were not quantified by qPCR as part of the study, and these were unavailable to sequence in the study in Chapter 3. Therefore, future work could look to investigate and validate these miRNAs.

Chapter 3 built on the methods developed in Chapter 2, examining a much larger cohort of patients, with 1150 patients with PH and 334 disease controls, examining circulating levels of 326 miRNAs. I used machine learning methods to derive panels of 9 miRNAs to detect PH and PAH from disease controls, as well as signatures to detect subtypes of PH from the larger PH cohort. This time, no healthy controls were included. The miRNA panels performed favourably to the current clinical standard NT-proBNP at discriminating between subtypes of PH.

We noted a superior performance of miRNAs compared to NT-proBNP in discriminating between subtypes of PH. However, the AUCs of PH vs DC and PAH vs DC (0.78 and 0.79 respectively) may be too low for clinical application. One area in which these signatures could add clinical value could be to integrate a circulating miRNA signature with the NT-proBNP cutoffs into the workup of patients with suspected PH, to help differentiate the treatable subtypes PAH and CTEPH. As the study was not set up to look at patients above and below NT-proBNP thresholds, the patient numbers were too small to investigate signatures specific to these groups, however this could be a future direction of investigation.

In Chapter 4 I explored the application of unsupervised machine learning to miRNAs to a subset of the cohort of patients examined in Chapter 3, containing PAH, PH-LHD, and PH-lung. This unsupervised approach identified six distinct molecular clusters that displayed differences in survival, haemodynamics NT-proBNP and 6-minute walking distance, as well as distinct molecular pathways. These circulating miRNAS may offer greater insight into the heterogeneity of PH than clinical phenotyping alone. Each of the molecular subtypes had unique pathways, some of which have previously been identified as potential therapeutic targets. As such, investigating the longitudinal data from patients in the different clusters to investigate treatment responses could also be a future direction.

Although not explored in this thesis, as part of the Janssen study, patients also had metabolomic data profiled. This provides another avenue of investigation for the clusters

described in Chapter 4 by performing a differential analysis of the metabolites between clusters. Alternatively, investigating whether metabolomic signatures can be derived for the clusters could provide a method to pull out metabolomic differences between clusters.

Despite many technological advances, the transferral of genetic biomarkers from the computational identification stage to the clinic has been slow. Transferring the biomarkers discovered into the clinical domain provides a challenge. The number of biomarkers with approved clinical usage compared to those declared in research papers shows that the majority of biomarker candidates have been discarded as possibilities or have yet to reach the clinic (Deyati et al. 2013).

There are several challenges involved in expediting this process. For example, the integration of multidisciplinary teams which contribute to clinical translation of personalised genomic medicine across a range of disciplines such as bioinformatics, epidemiology and omics is hampered by the compartmentalisation of research. An related issue may occur at the evaluation stage of publication, where most reviewers do not have expertise in every discipline included. Once a solution has reached the clinic, the implementation of a new technology or test may necessitate extra training for the healthcare professionals involved in the roll out, as well as new technologies.

This thesis provides a proof of principle that molecular classification of PH may be achieved. Looking forward, the CIPHER clinical trial has recently finished recruiting, where the aim will be to prospectively validate some of the miRNA signatures developed. The initial aim is to investigate whether these signatures may be used as an early diagnostic signature, and hold the potential to validate the cluster signatures.

References

Abbas, Mostafa, and Yasser El-Manzalawy. 2020. “Machine Learning Based Refined Differential Gene Expression Analysis of Pediatric Sepsis.” *BMC Medical Genomics* 13 (1): 122.

Alkhateeb, Abedalrhman, Iman Rezaeian, Siva Singireddy, Dora Cavallo-Medved, Lisa A. Porter, and Luis Rueda. 2019. “Transcriptomics Signature from Next-Generation Sequencing Data Reveals New Transcriptomic Biomarkers Related to Prostate Cancer.” *Cancer Informatics* 18 (March): 1176935119835522.

Alles, Julia, Tobias Fehlmann, Ulrike Fischer, Christina Backes, Valentina Galata, Marie Minet, Martin Hart, et al. 2019. “An Estimate of the Total Number of True Human miRNAs.” *Nucleic Acids Research* 47 (7): 3353–64.

Amirlatifi, Shahrzad, Mahboubeh Pazoki, Mehran Amrovani, Shima Ghezelbash, and Fatemeh Dastyar. 2022. “Evaluation of Genes and Molecular Pathways Involved in the Development of Cardiovascular Disease in Preeclampsia Patients: Biological System and Bioinformatics Analysis Approach.” https://doi.org/10.21203/rs.3.rs-1408549/v1.

Anwar, Anjum, Gregoire Ruffenach, Aman Mahajan, Mansoureh Eghbali, and Soban Umar. 2016. “Novel Biomarkers for Pulmonary Arterial Hypertension.” *Respiratory Research* 17 (1). https://doi.org/10.1186/s12931-016-0396-6.

Armstrong, Iain, Catherine Billings, David G. Kiely, Janelle Yorke, Carl Harries, Shaun Clayton, and Wendy Gin-Sing. 2019. “The Patient Experience of Pulmonary Hypertension: A Large Cross-Sectional Study of UK Patients.” *BMC Pulmonary Medicine* 19 (1): 67.

Austin, Peter C., Ian R. White, Douglas S. Lee, and Stef van Buuren. 2021. “Missing Data in Clinical Research: A Tutorial on Multiple Imputation.” *The Canadian Journal of Cardiology* 37 (9): 1322–31.

Barnett, Christopher F., and Van N. Selby. 2015. “Overview of WHO Group 2 Pulmonary Hypertension Due to Left Heart Disease.” *Advances in Pulmonary Hypertension*. https://doi.org/10.21693/1933-088x-14.2.70.

Bauer, Yasmina, Simon de Bernard, Peter Hickey, Karri Ballard, Jeremy Cruz, Peter Cornelisse, Harbajan Chadha-Boreham, et al. 2020. “Identifying Early Pulmonary Arterial Hypertension Biomarkers in Systemic Sclerosis: Machine Learning on Proteomics from the DETECT Cohort.” *The European Respiratory Journal: Official Journal of the European Society for Clinical Respiratory Physiology*, December. https://doi.org/10.1183/13993003.02591-2020.

Bazan, Isabel S., and Wassim H. Fares. 2015. “Pulmonary Hypertension: Diagnostic and Therapeutic Challenges.” *Therapeutics and Clinical Risk Management* 11 (August): 1221–33.

Benjamini, Yoav, and Yosef Hochberg. 1995. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society: Series B (Methodological)*. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Benza, Raymond L., Mardi Gomberg-Maitland, C. Greg Elliott, Harrison W. Farber, Aimee J. Foreman, Adaani E. Frost, Michael D. McGoon, et al. 2019. “Predicting Survival in Patients With Pulmonary Arterial Hypertension: The REVEAL Risk Score Calculator 2.0 and Comparison With ESC/ERS-Based Risk Assessment Strategies.” *Chest* 156 (2): 323–37.

Brock, Matthias, Michelle Trenkmann, Renate E. Gay, Beat A. Michel, Steffen Gay, Manuel Fischler, Silvia Ulrich, Rudolf Speich, and Lars C. Huber. 2009. “Interleukin-6 Modulates the Expression of the Bone Morphogenic Protein Receptor Type II through a Novel STAT3-microRNA Cluster 17/92 Pathway.” *Circulation Research* 104 (10): 1184–91.

Brown, Lynette M., Hubert Chen, Scott Halpern, Darren Taichman, Michael D. McGoon, Harrison W. Farber, Adaani E. Frost, et al. 2011. “Delay in Recognition of Pulmonary Arterial Hypertension: Factors Identified from the REVEAL Registry.” *Chest* 140 (1): 19–26.

Brunner-La Rocca, Hans-Peter, and Sandra Sanders-van Wijk. 2019. “Natriuretic Peptides in Chronic Heart Failure.” *Cardiac Failure Review* 5 (1): 44–49.

Bunel, Vincent, Alice Guyard, Gaëlle Dauriat, Claire Danel, David Montani, Clément Gauvain, Gabriel Thabut, et al. 2019. “Pulmonary Arterial Histologic Lesions in Patients With COPD With Severe Pulmonary Hypertension.” *Chest* 156 (1): 33–44.

Cai, Jie, Jiawei Luo, Shulin Wang, and Sheng Yang. 2018. “Feature Selection in Machine Learning: A New Perspective.” *Neurocomputing*. https://doi.org/10.1016/j.neucom.2017.11.077.

Caruso, Paola, Margaret R. MacLean, Raya Khanin, John McClure, Elaine Soon, Mark Southgate, Robert A. MacDonald, et al. 2010. “Dynamic Changes in Lung microRNA Profiles during the Development of Pulmonary Hypertension due to Chronic Hypoxia and Monocrotaline.” *Arteriosclerosis, Thrombosis, and Vascular Biology* 30 (4): 716–23.

Chai, Sanbao, Wang Wang, Jie Liu, Huan Guo, Zhifei Zhang, Chen Wang, and Jun Wang. 2015. “Leptin Knockout Attenuates Hypoxia-Induced Pulmonary Arterial Hypertension by Inhibiting Proliferation of Pulmonary Arterial Smooth Muscle Cells.” *Translational Research: The Journal of Laboratory and Clinical Medicine* 166 (6): 772–82.

Chan, Hei-Nga, Di Xu, See-Lok Ho, Man Shing Wong, and Hung-Wing Li. 2017. “Ultra-Sensitive Detection of Protein Biomarkers for Diagnosis of Alzheimer’s Disease.” *Chemical Science*. https://doi.org/10.1039/c6sc05615f.

Chen, Kuang-Hueih, Asish Dasgupta, Jianhui Lin, François Potus, Sébastien Bonnet, James Iremonger, Jennifer Fu, et al. 2018. “Epigenetic Dysregulation of the Dynamin-Related Protein 1 Binding Partners MiD49 and MiD51 Increases Mitotic Mitochondrial Fission and Promotes Pulmonary Arterial Hypertension: Mechanistic and Therapeutic Implications.” *Circulation* 138 (3): 287–304.

Chen, Tianqi, and Carlos Guestrin. 2016. “XGBoost: A Scalable Tree Boosting System.” In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’16*, 785–94. New York, New York, USA: ACM Press.

Chen, Weidan, and Shoujun Li. 2017. “Circulating microRNA as a Novel Biomarker for Pulmonary Arterial Hypertension Due to Congenital Heart Disease.” *Pediatric Cardiology*. https://doi.org/10.1007/s00246-016-1487-3.

Chudova, Darya, Jonathan I. Wilde, Eric T. Wang, Hui Wang, Nusrat Rabbee, Camila M. Egidio, Jessica Reynolds, et al. 2010. “Molecular Classification of Thyroid Nodules Using High-Dimensionality Genomic Data.” *The Journal of Clinical Endocrinology and Metabolism* 95 (12): 5296–5304.

Coghlan, J. Gerry, Christopher P. Denton, Ekkehard Grünig, Diana Bonderman, Oliver Distler, Dinesh Khanna, Ulf Müller-Ladner, et al. 2014. “Evidence-Based Detection of Pulmonary Arterial Hypertension in Systemic Sclerosis: The DETECT Study.” *Annals of the Rheumatic Diseases* 73 (7): 1340–49.

Condrat, Carmen Elena, Dana Claudia Thompson, Madalina Gabriela Barbu, Oana Larisa Bugnar, Andreea Boboc, Dragos Cretoiu, Nicolae Suciu, Sanda Maria Cretoiu, and Silviu Cristian Voinea. 2020. “miRNAs as Biomarkers in Disease: Latest Findings Regarding Their Role in Diagnosis and Prognosis.” *Cells* 9 (2). https://doi.org/10.3390/cells9020276.

Courboulin, Audrey, Roxane Paulin, Nellie J. Giguère, Nehmé Saksouk, Tanya Perreault, Jolyane Meloche, Eric R. Paquet, et al. 2011a. “Role for miR-204 in Human Pulmonary Arterial Hypertension.” *Journal of Experimental Medicine*. https://doi.org/10.1084/jem.20101812.

———. 2011b. “Role for miR-204 in Human Pulmonary Arterial Hypertension.” *The Journal of Experimental Medicine* 208 (3): 535–48.

DeLong, Elizabeth R., David M. DeLong, and Daniel L. Clarke-Pearson. 1988. “Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach.” *Biometrics*. https://doi.org/10.2307/2531595.

DeLong, E. R., D. M. DeLong, and D. L. Clarke-Pearson. 1988. “Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach.” *Biometrics* 44 (3): 837–45.

Drucker, Elisabeth, and Kurt Krapfenbauer. 2013. “Pitfalls and Limitations in Translation from Biomarker Discovery to Clinical Utility in Predictive and Personalised Medicine.” *The EPMA Journal* 4 (1): 7.

Duong Van Huyen, Jean-Paul, Marion Tible, Arnaud Gay, Romain Guillemain, Olivier Aubert, Shaida Varnous, Franck Iserin, et al. 2014. “MicroRNAs as Non-Invasive Biomarkers of Heart Transplant Rejection.” *European Heart Journal* 35 (45): 3194–3202.

Errington, Niamh, James Iremonger, Josephine A. Pickworth, Sokratis Kariotis, Christopher J. Rhodes, Alexander Mk Rothman, Robin Condliffe, et al. 2021. “A Diagnostic miRNA Signature for Pulmonary Arterial Hypertension Using a Consensus Machine Learning Approach.” *EBioMedicine* 69 (July): 103444.

Fayyaz, Ahmed U., William D. Edwards, Joseph J. Maleszewski, Ewa A. Konik, Hilary M. DuBrock, Barry A. Borlaug, Robert P. Frantz, Sarah M. Jenkins, and Margaret M. Redfield. 2018. “Global Pulmonary Vascular Remodeling in Pulmonary Hypertension Associated With Heart Failure and Preserved or Reduced Ejection Fraction.” *Circulation*. https://doi.org/10.1161/circulationaha.117.031608.

FDA-NIH Biomarker Working Group. 2016. *BEST (Biomarkers, EndpointS, and Other Tools) Resource*. Silver Spring (MD): Food and Drug Administration (US).

Fernández, Ana I., Raquel Yotti, Ana González-Mansilla, Teresa Mombiela, Enrique Gutiérrez-Ibanes, Candelas Pérez Del Villar, Paula Navas-Tejedor, et al. 2019. “The Biological Bases of Group 2 Pulmonary Hypertension.” *International Journal of Molecular Sciences* 20 (23). https://doi.org/10.3390/ijms20235884.

Fonti, Valeria, and Eduard Belitser. 2017. “Feature Selection Using Lasso.” *VU Amsterdam Research Paper in Business Analytics*. https://beta.vu.nl/nl/Images/werkstuk-

fonti_tcm235-836234.pdf.
Friedman, Jerome, Trevor Hastie, and Rob Tibshirani. 2010. “Regularization Paths for Generalized Linear Models via Coordinate Descent.” *Journal of Statistical Software* 33 (1): 1–22.
Fröhlich, Holger, Rudi Balling, Niko Beerenwinkel, Oliver Kohlbacher, Santosh Kumar, Thomas Lengauer, Marloes H. Maathuis, et al. 2018. “From Hype to Reality: Data Science Enabling Personalized Medicine.” *BMC Medicine* 16 (1): 150.
Galiè, Nazzareno, Marc Humbert, Jean-Luc Vachiery, Simon Gibbs, Irene Lang, Adam Torbicki, Gérald Simonneau, et al. 2015. “2015 ESC/ERS Guidelines for the Diagnosis and Treatment of Pulmonary Hypertension: The Joint Task Force for the Diagnosis and Treatment of Pulmonary Hypertension of the European Society of Cardiology (ESC) and the European Respiratory Society (ERS): Endorsed by: Association for European Paediatric and Congenital Cardiology (AEPC), International Society for Heart and Lung Transplantation (ISHLT).” *European Heart Journal* 37 (1): 67–119.
Groth, Alexandra, Bart Vrugt, Matthias Brock, Rudolf Speich, Silvia Ulrich, and Lars C. Huber. 2014. “Inflammatory Cytokines in Pulmonary Hypertension.” *Respiratory Research*. https://doi.org/10.1186/1465-9921-15-47.
Guyon, Isabelle, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. 2002. “Gene Selection for Cancer Classification Using Support Vector Machines.” *Machine Learning* 46 (1): 389–422.
Hachulla, Eric, Virginie Gressin, Loïc Guillevin, Patrick Carpentier, Elisabeth Diot, Jean Sibilia, André Kahan, et al. 2005. “Early Detection of Pulmonary Arterial Hypertension in Systemic Sclerosis: A French Nationwide Prospective Multicenter Study.” *Arthritis & Rheumatism*. https://doi.org/10.1002/art.21433.
Hewes, Jenny L., Ji Young Lee, Karen A. Fagan, and Natalie N. Bauer. 2020. “The Changing Face of Pulmonary Hypertension Diagnosis: A Historical Perspective on the Influence of Diagnostics and Biomarkers.” *Pulmonary Circulation* 10 (1): 2045894019892801.
He, Zengyou, and Weichuan Yu. 2010. “Stable Feature Selection for Biomarker Discovery.” *Computational Biology and Chemistry* 34 (4): 215–25.
Hira, Zena M., and Duncan F. Gillies. 2015. “A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data.” *Advances in Bioinformatics* 2015 (June): 198363.
Hoeper, Marius M. 2015. “Pharmacological Therapy for Patients with Chronic Thromboembolic Pulmonary Hypertension.” *European Respiratory Review: An Official Journal of the European Respiratory Society* 24 (136): 272–82.
Hołda, Mateusz K., Aneta Stachowicz, Maciej Suski, Dorota Wojtysiak, Natalia Sowińska, Zbigniew Arent, Natalia Palka, Piotr Podolec, and Grzegorz Kopeć. 2020. “Myocardial Proteomic Profile in Pulmonary Arterial Hypertension.” *Scientific Reports*. https://doi.org/10.1038/s41598-020-71264-8.
Hu, Chien-An, Chia-Ming Chen, Yen-Chun Fang, Shinn-Jye Liang, Hao-Chien Wang, Wen-Feng Fang, Chau-Chyun Sheu, et al. 2020. “Using a Machine Learning Approach to Predict Mortality in Critically Ill Influenza Patients: A Cross-Sectional Retrospective Multicentre Study in Taiwan.” *BMJ Open* 10 (2): e033898.
Ibrahim, Joseph G., Haitao Chu, and Ming-Hui Chen. 2012. “Missing Data in Clinical Studies: Issues and Methods.” *Journal of Clinical Oncology*. https://doi.org/10.1200/jco.2011.38.7589.
Ibrahim, Lujain, Munib Mesinovic, Kai-Wen Yang, and Mohamad A. Eid. 2020. “Explainable Prediction of Acute Myocardial Infarction Using Machine Learning and Shapley Values.” *IEEE Access*. https://doi.org/10.1109/access.2020.3040166.
Izumiya, Yasuhiro, Masatoshi Jinnn, Yuichi Kimura, Zhongzhi Wang, Yoshiro Onoue, Shinsuke Hanatani, Satoshi Araki, Hironobu Ihn, and Hisao Ogawa. 2015. “Expression of Let-7 Family microRNAs in Skin Correlates Negatively with Severity of Pulmonary Hypertension in Patients with Systemic Scleroderma.” *International Journal of Cardiology. Heart & Vasculature* 8 (September): 98–102.

Jakobsen, Janus Christian, Christian Gluud, Jørn Wetterslev, and Per Winkel. 2017. “When and How Should Multiple Imputation Be Used for Handling Missing Data in Randomised Clinical Trials – a Practical Guide with Flowcharts.” *BMC Medical Research Methodology*. https://doi.org/10.1186/s12874-017-0442-1.

Jin, Qi, Zhihui Zhao, Qing Zhao, Xue Yu, Lu Yan, Yi Zhang, Qin Luo, and Zhihong Liu. 2020. “Long Noncoding RNAs: Emerging Roles in Pulmonary Hypertension.” *Heart Failure Reviews* 25 (5): 795–815.

Joshi, Sachindra Raj, Vidhi Dhagia, Salina Gairhe, John G. Edwards, Ivan F. McMurtry, and Sachin A. Gupte. 2016. “MicroRNA-140 Is Elevated and Mitofusin-1 Is Downregulated in the Right Ventricle of the Sugen5416/hypoxia/normoxia Model of Pulmonary Arterial Hypertension.” *American Journal of Physiology-Heart and Circulatory Physiology*. https://doi.org/10.1152/ajpheart.00264.2016.

Kanehisa, M., and S. Goto. 2000. “KEGG: Kyoto Encyclopedia of Genes and Genomes.” *Nucleic Acids Research* 28 (1): 27–30.

Kariotis, Sokratis, Emmanuel Jammeh, Emilia M. Swietlik, Josephine A. Pickworth, Christopher J. Rhodes, Pablo Otero, John Wharton, et al. 2021. “Biological Heterogeneity in Idiopathic Pulmonary Arterial Hypertension Identified through Unsupervised Transcriptomic Profiling of Whole Blood.” *Nature Communications* 12 (1): 7104.

Khanna, Dinesh, Heather Gladue, Richard Channick, Lorinda Chung, Oliver Distler, Daniel E. Furst, Eric Hachulla, et al. 2013. “Recommendations for Screening and Detection of Connective Tissue Disease-Associated Pulmonary Arterial Hypertension.” *Arthritis and Rheumatism* 65 (12): 3194–3201.

Kheyfets, Vitaly O., Carmen C. Sucharov, Uyen Truong, Jamie Dunning, Kendall Hunter, Dunbar Ivy, Shelley Miyamoto, and Robin Shandas. 2017. “Circulating miRNAs in Pediatric Pulmonary Hypertension Show Promise as Biomarkers of Vascular Function.” *Oxidative Medicine and Cellular Longevity* 2017 (July): 4957147.

Kiely, David G., Orla Doyle, Edmund Drage, Harvey Jenner, Valentina Salvatelli, Flora A. Daniels, John Rigg, et al. 2019. “Utilising Artificial Intelligence to Determine Patients at Risk of a Rare Disease: Idiopathic Pulmonary Arterial Hypertension.” *Pulmonary Circulation* 9 (4): 2045894019890549.

Kiely, David G., Allan Lawrie, and Marc Humbert. 2019. “Screening Strategies for Pulmonary Arterial Hypertension.” *European Heart Journal Supplements: Journal of the European Society of Cardiology* 21 (Suppl K): K9–20.

Kuncheva, Ludmila I. 2007. “A Stability Index for Feature Selection.” In *Proceedings of the 25th Conference on Proceedings of the 25th IASTED International Multi-Conference: Artificial Intelligence and Applications*, 390–95. AIAP’07. USA: ACTA Press.

Kursa, Miron B., and Witold R. Rudnicki. 2010. “Feature Selection with the Boruta Package.” *Journal of Statistical Software*. https://doi.org/10.18637/jss.v036.i11.

Launay, David, Vincent Sobanski, Eric Hachulla, and Marc Humbert. 2017. “Pulmonary Hypertension in Systemic Sclerosis: Different Phenotypes.” *European Respiratory Review: An Official Journal of the European Respiratory Society* 26 (145). https://doi.org/10.1183/16000617.0056-2017.

Lawrie, Charles H., Shira Gal, Heather M. Dunlop, Beena Pushkaran, Amanda P. Liggins, Karen Pulford, Alison H. Banham, et al. 2008. “Detection of Elevated Levels of Tumour-Associated microRNAs in Serum of Patients with Diffuse Large B-Cell Lymphoma.” *British Journal of Haematology* 141 (5): 672–75.

Lee, Rosalind C., Rhonda L. Feinbaum, and Victor Ambros. 1993. “The C. Elegans Heterochronic Gene Lin-4 Encodes Small RNAs with Antisense Complementarity to Lin-14.” *Cell*. https://doi.org/10.1016/0092-8674(93)90529-y.

Lewis, Robert A., Iain Armstrong, Carmel Bergbaum, Melanie J. Brewis, John Cannon, Athanasios Charalampopoulos, A. Colin Church, et al. 2021. “EmPHasis-10 Health-Related Quality of Life Score Predicts Outcomes in Patients with Idiopathic and Connective Tissue Disease-Associated Pulmonary Arterial Hypertension: Results from a UK Multicentre Study.” *The European Respiratory Journal: Official Journal of the*

European Society for Clinical Respiratory Physiology 57 (2). https://doi.org/10.1183/13993003.00124-2020.

Liao, Yuxing, Jing Wang, Eric J. Jaehnig, Zhiao Shi, and Bing Zhang. 2019. “WebGestalt 2019: Gene Set Analysis Toolkit with Revamped UIs and APIs.” *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkz401.

Liaw, Andy, and Matthew Wiener. 2002. “Classification and Regression by randomForest.” *R News* 2 (3): 18–22.

Li, Fangwei, Wenhua Shi, Yixin Wan, Qingting Wang, Wei Feng, Xin Yan, Jian Wang, Limin Chai, Qianqian Zhang, and Manxiang Li. 2017. “Prediction of Target Genes for miR-140-5p in Pulmonary Arterial Hypertension Using Bioinformatics Methods.” *FEBS Open Bio* 7 (12): 1880–90.

Little, Roderick J. A., and Donald B. Rubin. 2019. *Statistical Analysis with Missing Data*. Wiley.

Li, Yuanyuan, Kai Kang, Juno M. Krahn, Nicole Croutwater, Kevin Lee, David M. Umbach, and Leping Li. 2017. “A Comprehensive Genomic Pan-Cancer Classification Using The Cancer Genome Atlas Gene Expression Data.” *BMC Genomics* 18 (1): 508.

Long, Bo, Kun Wang, Na Li, Iram Murtaza, Jing-Ying Xiao, Yuan-Yuan Fan, Cui-Yun Liu, Wen-Hui Li, Zheng Cheng, and Peifeng Li. 2013. “miR-761 Regulates the Mitochondrial Network by Targeting Mitochondrial Fission Factor.” *Free Radical Biology and Medicine*. https://doi.org/10.1016/j.freeradbiomed.2013.07.009.

Lopez-Rincon, Alejandro, Marlet Martinez-Archundia, Gustavo U. Martinez-Ruiz, Alexander Schoenhuth, and Alberto Tonda. 2019. “Automatic Discovery of 100-miRNA Signature for Cancer Classification Using Ensemble Feature Selection.” *BMC Bioinformatics* 20 (1): 480.

López-Romero, Pedro. 2011. “Pre-Processing and Differential Expression Analysis of Agilent microRNA Arrays Using the AgiMicroRna Bioconductor Library.” *BMC Genomics*. https://doi.org/10.1186/1471-2164-12-64.

Lungu, Angela, Andrew J. Swift, David Capener, David Kiely, Rod Hose, and Jim M. Wild. 2016. “Diagnosis of Pulmonary Hypertension from Magnetic Resonance Imaging-Based Computational Models and Decision Tree Analysis.” *Pulmonary Circulation* 6 (2): 181–90.

Mao, Chengyi, Xiaoxi Zeng, Chao Zhang, Yushang Yang, Xin Xiao, Siyuan Luan, Yonggang Zhang, and Yong Yuan. 2021. “Mechanisms of Pharmaceutical Therapy and Drug Resistance in Esophageal Cancer.” *Frontiers in Cell and Developmental Biology* 9 (February): 612451.

Mestdagh, Pieter, Nicole Hartmann, Lukas Baeriswyl, Ditte Andreasen, Nathalie Bernard, Caifu Chen, David Cheo, et al. 2014. “Evaluation of Quantitative miRNA Expression Platforms in the microRNA Quality Control (miRQC) Study.” *Nature Methods* 11 (8): 809–15.

Miao, Chenggui, Jun Chang, and Guoxue Zhang. 2018. “Recent Research Progress of microRNAs in Hypertension Pathogenesis, with a Focus on the Roles of miRNAs in Pulmonary Arterial Hypertension.” *Molecular Biology Reports*, October. https://doi.org/10.1007/s11033-018-4335-0.

Mohri, Mehryar, Afshin Rostamizadeh, and Ameet Talwalkar. 2012. *Foundations of Machine Learning*. MIT Press.

Moons, Karel G. M., Rogier A. R. T. Donders, Theo Stijnen, and Frank E. Harrell Jr. 2006. “Using the Outcome for Imputation of Missing Predictor Values Was Preferred.” *Journal of Clinical Epidemiology* 59 (10): 1092–1101.

Moser, K. M., and C. M. Bloor. 1993. “Pulmonary Vascular Lesions Occurring in Patients with Chronic Major Vessel Thromboembolic Pulmonary Hypertension.” *Chest* 103 (3): 685–92.

Murdoch, Travis B., and Allan S. Detsky. 2013. “The Inevitable Application of Big Data to Health Care.” *JAMA: The Journal of the American Medical Association* 309 (13): 1351–52.

Neumann, Ursula, Mona Riemenschneider, Jan-Peter Sowa, Theodor Baars, Julia Kälsch,

Ali Canbay, and Dominik Heider. 2016. “Compensation of Feature Selection Biases Accompanied with Improved Predictive Performance for Binary Classification by Using a Novel Ensemble Feature Selection Approach.” *BioData Mining* 9 (November): 36.

O’Brien, Jacob, Heyam Hayder, Yara Zayed, and Chun Peng. 2018. “Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation.” *Frontiers in Endocrinology* 9 (August): 402.

Ogunleye, Adeola, and Qing-Guo Wang. 2020. “XGBoost Model for Chronic Kidney Disease Diagnosis.” *IEEE/ACM Transactions on Computational Biology and Bioinformatics / IEEE, ACM* 17 (6): 2131–40.

Opitz, Christian F., Marius M. Hoeper, J. Simon R. Gibbs, Harald Kaemmerer, Joanna Pepke-Zaba, J. Gerry Coghlan, Laura Scelsi, et al. 2016. “Pre-Capillary, Combined, and Post-Capillary Pulmonary Hypertension: A Pathophysiological Continuum.” *Journal of the American College of Cardiology* 68 (4): 368–78.

Or, Gilad Ben, Gilad Ben Or, and Isana Veksler-Lublinsky. 2021. “Comprehensive Machine-Learning-Based Analysis of microRNA–target Interactions Reveals Variable Transferability of Interaction Rules across Species.” *BMC Bioinformatics*. https://doi.org/10.1186/s12859-021-04164-x.

Paraskevopoulou, Maria D., Georgios Georgakilas, Nikos Kostoulas, Ioannis S. Vlachos, Thanasis Vergoulis, Martin Reczko, Christos Filippidis, Theodore Dalamagas, and A. G. Hatzigeorgiou. 2013. “DIANA-microT Web Server v5.0: Service Integration into miRNA Functional Analysis Workflows.” *Nucleic Acids Research* 41 (Web Server issue): W169–73.

Parker, Joel S., Michael Mullins, Maggie C. U. Cheang, Samuel Leung, David Voduc, Tammi Vickery, Sherri Davies, et al. 2009. “Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes.” *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 27 (8): 1160–67.

Paulin, R., J. Meloche, A. Courboulin, C. Lambert, A. Haromy, A. Courchesne, P. Bonnet, S. Provencher, E. D. Michelakis, and S. Bonnet. 2014. “Targeting Cell Motility in Pulmonary Arterial Hypertension.” *European Respiratory Journal*. https://doi.org/10.1183/09031936.00181312.

Paul, Prosenjit, Anindya Chakraborty, Debasree Sarkar, Malobika Langthasa, Musfhia Rahman, Minakshi Bari, R. K. Sanamacha Singha, Arup Kumar Malakar, and Supriyo Chakraborty. 2018. “Interplay between miRNAs and Human Diseases.” *Journal of Cellular Physiology*. https://doi.org/10.1002/jcp.25854.

Peng, Yong, and Carlo M. Croce. 2016. “The Role of MicroRNAs in Human Cancer.” *Signal Transduction and Targeted Therapy*. https://doi.org/10.1038/sigtrans.2015.4.

Penso, Marco, Mauro Pepi, Laura Fusini, Manuela Muratori, Claudia Cefalù, Valentina Mantegazza, Paola Gripari, et al. 2021. “Predicting Long-Term Mortality in TAVI Patients Using Machine Learning Techniques.” *Journal of Cardiovascular Development and Disease* 8 (4). https://doi.org/10.3390/jcdd8040044.

Pickworth, Josephine, Alexander Rothman, James Iremonger, Helen Casbolt, Kay Hopkinson, Peter M. Hickey, Santhi Gladson, et al. 2017. “Differential IL-1 Signaling Induced by BMPR2 Deficiency Drives Pulmonary Vascular Remodeling.” *Pulmonary Circulation* 7 (4): 768–76.

Potus, François, Grégoire Ruffenach, Abdellaziz Dahou, Christophe Thebault, Sandra Breuils-Bonnet, Ève Tremblay, Valérie Nadeau, et al. 2015. “Downregulation of MicroRNA-126 Contributes to the Failing Right Ventricle in Pulmonary Arterial Hypertension.” *Circulation*. https://doi.org/10.1161/circulationaha.115.016382.

Pradervand, Sylvain, Johann Weber, Jérôme Thomas, Manuel Bueno, Pratyaksha Wirapati, Karine Lefort, G. Paolo Dotto, and Keith Harshman. 2009. “Impact of Normalization on miRNA Microarray Expression Profiling.” *RNA* 15 (3): 493–501.

Precazzini, Francesca, Simone Detassis, Andrea Selenito Imperatori, Michela Alessandra Denti, and Paola Campomenosi. 2021. “Measurements Methods for the Development of MicroRNA-Based Tests for Cancer Diagnosis.” *International Journal of Molecular Sciences* 22 (3). https://doi.org/10.3390/ijms22031176.

Pritchard, Colin C., Heather H. Cheng, and Muneesh Tewari. 2012. “MicroRNA Profiling: Approaches and Considerations.” *Nature Reviews. Genetics* 13 (5): 358–69.
Qiu, Chunping, Nan Lu, Xiao Wang, Qing Zhang, Cunzhong Yuan, Shi Yan, Samina Dongol, et al. 2017. “Gene Expression Profiles of Ovarian Low-Grade Serous Carcinoma Resemble Those of Fallopian Tube Epithelium.” *Gynecologic Oncology* 147 (3): 634–41.
Raghunathan, Trivellore E. 2004. “What Do We Do with Missing Data? Some Options for Analysis of Incomplete Data.” *Annual Review of Public Health* 25: 99–117.
Rajkumar, Revathi, Kazuhisa Konishi, Thomas J. Richards, David C. Ishizawar, Andrew C. Wiechert, Naftali Kaminski, and Ferhaan Ahmad. 2010. “Genomewide RNA Expression Profiling in Lung Identifies Distinct Signatures in Idiopathic Pulmonary Arterial Hypertension and Secondary Pulmonary Hypertension.” *American Journal of Physiology. Heart and Circulatory Physiology* 298 (4): H1235–48.
Rameh, Vanessa, and Antoine Kossaify. 2016. “Role of Biomarkers in the Diagnosis, Risk Assessment, and Management of Pulmonary Hypertension.” *Biomarker Insights* 11 (June): 85–89.
R Core Team. 2013. “R Core Team.” R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. 2013. URL http://www.R-project.org/.
Rhodes, Christopher J., Pavandeep Ghataorhe, John Wharton, Kevin C. Rue-Albrecht, Charaka Hadinnapola, Geoffrey Watson, Marta Bleda, et al. 2017. “Plasma Metabolomics Implicates Modified Transfer RNAs and Altered Bioenergetics in the Outcomes of Pulmonary Arterial Hypertension.” *Circulation* 135 (5): 460–75.
Rhodes, Christopher J., Pablo Otero-Núñez, John Wharton, Emilia M. Swietlik, Sokratis Kariotis, Lars Harbaum, Mark J. Dunning, et al. 2020. “Whole-Blood RNA Profiles Associated with Pulmonary Arterial Hypertension and Clinical Outcome.” *American Journal of Respiratory and Critical Care Medicine* 202 (4): 586–94.
Rhodes, Christopher J., John Wharton, Reinier A. Boon, Tino Roexe, Hilda Tsang, Beata Wojciak-Stothard, Anob Chakrabarti, et al. 2013. “Reduced microRNA-150 Is Associated with Poor Survival in Pulmonary Arterial Hypertension.” *American Journal of Respiratory and Critical Care Medicine* 187 (3): 294–302.
Rhodes, Christopher J., John Wharton, Pavandeep Ghataorhe, Geoffrey Watson, Barbara Girerd, Luke S. Howard, J. Simon R. Gibbs, et al. 2017. “Plasma Proteome Analysis in Patients with Pulmonary Arterial Hypertension: An Observational Cohort Study.” *The Lancet. Respiratory Medicine* 5 (9): 717–26.
Rosenkranz, Stephan, and Ioana R. Preston. 2015. “Right Heart Catheterisation: Best Practice and Pitfalls in Pulmonary Hypertension.” *European Respiratory Review: An Official Journal of the European Respiratory Society* 24 (138): 642–52.
Rothman, Abraham, Humberto Restrepo, Valeri Sarukhanov, William N. Evans, Robert G. Wiencek Jr, Roy Williams, Nicole Hamburger, Kylie Anderson, Jasmine Balsara, and David Mann. 2017. “Assessment of microRNA and Gene Dysregulation in Pulmonary Hypertension by Endoarterial Biopsy.” *Pulmonary Circulation* 7 (2): 455–64.
Rothman, Alexander M. K., Nadine D. Arnold, Josephine A. Pickworth, James Iremonger, Loredana Ciuclan, Robert M. H. Allen, Sabine Guth-Gundel, et al. 2016. “MicroRNA-140-5p and SMURF1 Regulate Pulmonary Arterial Hypertension.” *The Journal of Clinical Investigation* 126 (7): 2495–2508.
Rothman, Alex M. K., Timothy J. A. Chico, and Allan Lawrie. 2014. “MicroRNA in Pulmonary Vascular Disease.” *Progress in Molecular Biology and Translational Science* 124: 43–63.
Roy, Shikha, Amar Pratap Singh, and Dinesh Gupta. 2021. “Unsupervised Subtyping and Methylation Landscape of Pancreatic Ductal Adenocarcinoma.” *Heliyon* 7 (1): e06000.
Saliminejad, Kioomars, Hamid Reza Khorram Khorshid, and Seyed Hamidollah Ghaffari. 2019. “Why Have microRNA Biomarkers Not Been Translated from Bench to Clinic?” *Future Oncology*. https://doi.org/10.2217/fon-2018-0812.
Santos-Ferreira, Cátia A., Mónica T. Abreu, Carla I. Marques, Lino M. Gonçalves, Rui Baptista, and Henrique M. Girão. 2020. “Micro-RNA Analysis in Pulmonary Arterial

Hypertension: Current Knowledge and Challenges." *JACC. Basic to Translational Science* 5 (11): 1149–62.
Saygin, Didem, Tracy Tabib, Humberto E. T. Bittar, Eleanor Valenzi, John Sembrat, Stephen Y. Chan, Mauricio Rojas, and Robert Lafyatis. 2020. "Transcriptional Profiling of Lung Cell Populations in Idiopathic Pulmonary Arterial Hypertension." *Pulmonary Circulation* 10 (1). https://doi.org/10.1177/2045894020908782.
Schafer, Joseph L., and John W. Graham. 2002. "Missing Data: Our View of the State of the Art." *Psychological Methods* 7 (2): 147–77.
Schlosser, Kenny, R. James White, and Duncan J. Stewart. 2013. "miR-26a Linked to Pulmonary Hypertension by Global Assessment of Circulating Extracellular microRNAs." *American Journal of Respiratory and Critical Care Medicine* 188 (12): 1472–75.
Shah, Sanjiv J. 2022. "BNP: Biomarker Not Perfect in Heart Failure with Preserved Ejection Fraction." *European Heart Journal*. https://doi.org/10.1093/eurheartj/ehac121.
Shannon, Paul, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. 2003. "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks." *Genome Research* 13 (11): 2498–2504.
Shapiro, S. S., and M. B. Wilk. 1965. "An Analysis of Variance Test for Normality (complete Samples)." *Biometrika*. https://doi.org/10.1093/biomet/52.3-4.591.
Simonneau, Gérald, David Montani, David S. Celermajer, Christopher P. Denton, Michael A. Gatzoulis, Michael Krowka, Paul G. Williams, and Rogerio Souza. 2019. "Haemodynamic Definitions and Updated Clinical Classification of Pulmonary Hypertension." *The European Respiratory Journal: Official Journal of the European Society for Clinical Respiratory Physiology* 53 (1). https://doi.org/10.1183/13993003.01913-2018.
Singh, Sally J., Milo A. Puhan, Vasileios Andrianopoulos, Nidia A. Hernandes, Katy E. Mitchell, Catherine J. Hill, Annemarie L. Lee, et al. 2014. "An Official Systematic Review of the European Respiratory Society/American Thoracic Society: Measurement Properties of Field Walking Tests in Chronic Respiratory Disease." *The European Respiratory Journal: Official Journal of the European Society for Clinical Respiratory Physiology* 44 (6): 1447–78.
Starling, Randall C., Michael Pham, Hannah Valantine, Leslie Miller, Howard Eisen, E. Rene Rodriguez, David O. Taylor, et al. 2006. "Molecular Testing in the Management of Cardiac Transplant Recipients: Initial Clinical Experience." *The Journal of Heart and Lung Transplantation: The Official Publication of the International Society for Heart Transplantation* 25 (12): 1389–95.
Stavseth, Marianne Riksheim, Thomas Clausen, and Jo Røislien. 2019. "How Handling Missing Data May Impact Conclusions: A Comparison of Six Different Imputation Methods for Categorical Questionnaire Data." *SAGE Open Medicine* 7 (January): 2050312118822912.
Sterne, Jonathan A. C., Ian R. White, John B. Carlin, Michael Spratt, Patrick Royston, Michael G. Kenward, Angela M. Wood, and James R. Carpenter. 2009. "Multiple Imputation for Missing Data in Epidemiological and Clinical Research: Potential and Pitfalls." *BMJ* 338 (June): b2393.
Tekkeşin, Ahmet İlker. 2019. "Artificial Intelligence in Healthcare: Past, Present and Future." *Anatolian Journal of Cardiology* 22 (Suppl 2): 8–9.
Therneau, Terry, and Beth Atkinson. 2018. "Rpart: Recursive Partitioning and Regression Trees." https://CRAN.R-project.org/package=rpart.
Toh, Tzen S., Frank Dondelinger, and Dennis Wang. 2019. "Looking beyond the Hype: Applied AI and Machine Learning in Translational Medicine." *EBioMedicine* 47 (September): 607–15.
Toshner, Mark, Colin Church, Lars Harbaum, Christopher Rhodes, Sofia S. Villar Moreschi, James Liley, Rowena Jones, et al. 2022. "Mendelian Randomisation and Experimental Medicine Approaches to Interleukin-6 as a Drug Target in Pulmonary Arterial Hypertension." *The European Respiratory Journal: Official Journal of the European*

Society for Clinical Respiratory Physiology 59 (3). https://doi.org/10.1183/13993003.02463-2020.

Toshner, Mark, and Alex Rothman. 2020. “IL-6 in Pulmonary Hypertension: Why Novel Is Not Always Best.” *The European Respiratory Journal: Official Journal of the European Society for Clinical Respiratory Physiology*. https://doi.org/10.1183/13993003.00314-2020.

Weber, Michael, and Christian Hamm. 2006. “Role of B-Type Natriuretic Peptide (BNP) and NT-proBNP in Clinical Routine.” *Heart* 92 (6): 843–49.

White, Ian R., Patrick Royston, and Angela M. Wood. 2011. “Multiple Imputation Using Chained Equations: Issues and Guidance for Practice.” *Statistics in Medicine* 30 (4): 377–99.

Wilkens, Heinrike, Stavros Konstantinides, Irene M. Lang, Alexander C. Bunck, Mario Gerges, Felix Gerhardt, Aleksandar Grgic, et al. 2018. “Chronic Thromboembolic Pulmonary Hypertension (CTEPH): Updated Recommendations from the Cologne Consensus Conference 2018.” *International Journal of Cardiology* 272S (December): 69–78.

Williams, Mark H., Clive E. Handler, Raza Akram, Colette J. Smith, Clare Das, Joanna Smee, Devaki Nair, Christopher P. Denton, Carol M. Black, and John G. Coghlan. 2006. “Role of N-Terminal Brain Natriuretic Peptide (N-TproBNP) in Scleroderma-Associated Pulmonary Arterial Hypertension.” *European Heart Journal* 27 (12): 1485–94.

Wu, Jiande, and Chindo Hicks. 2021. “Breast Cancer Type Classification Using Machine Learning.” *Journal of Personalized Medicine* 11 (2). https://doi.org/10.3390/jpm11020061.

Xiao, Tingting, Lijian Xie, Min Huang, and Jie Shen. 2017. “Differential Expression of microRNA in the Lungs of Rats with Pulmonary Arterial Hypertension.” *Molecular Medicine Reports* 15 (2): 591–96.

Xu, Wenlong, Anthony San Lucas, Zixing Wang, and Yin Liu. 2014. “Identifying microRNA Targets in Different Gene Regions.” *BMC Bioinformatics* 15 Suppl 7 (May): S4.

Zhang, Xiaonan, Shaoyang Dong, Qiujin Jia, Ao Zhang, Yanyang Li, Yaping Zhu, Shichao Lv, and Junping Zhang. 2019. “The microRNA in Ventricular Remodeling: The miR-30 Family.” *Bioscience Reports* 39 (8). https://doi.org/10.1042/BSR20190788.

Zhao, Yidan, Jenny Peng, Catherine Lu, Michael Hsin, Marco Mura, Licun Wu, Lei Chu, et al. 2014. “Metabolomic Heterogeneity of Pulmonary Arterial Hypertension.” *PloS One* 9 (2): e88727.

Zheng, Yaguo, Hong Ma, Enci Hu, Zhiwei Huang, Xiaoling Cheng, and Changming Xiong. 2015. “Inhibition of FGFR Signaling With PD173074 Ameliorates Monocrotaline-Induced Pulmonary Arterial Hypertension and Rescues BMPR-II Expression.” *Journal of Cardiovascular Pharmacology* 66 (5): 504–14.