ENABLING CARDIOVASCULAR MULTIMODAL, HIGH DIMENSIONAL, INTEGRATIVE ANALYTICS

by

VICTOR ROTH CARDOSO

A thesis submitted to
The University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

Institute of Cardiovascular Sciences
College of Medical and Dental Sciences
University of Birmingham
April 2021

ABSTRACT

While traditionally the understanding of cardiovascular morbidity relied on the acquisition and interpretation of health data, the advances in health technologies has enabled us to collect far larger amount of health data. This thesis explores the application of advanced analytics that utilise powerful mechanisms for integrating health data across different modalities and dimensions into a single and holistic environment to better understand different diseases, with a focus on cardiovascular conditions. Different statistical methodologies are applied across a number of case studies supported by a novel methodology to integrate and simplify data collection. The work culminates in the different dataset modalities explaining different effects on morbidity: blood biomarkers, electrocardiogram recordings, RNA-Seq measurements, and different population effects piece together the understanding of a person morbidity. More specifically, explainable artificial intelligence methods were employed on structured datasets from patients with atrial fibrillation to improve the screening for the disease. Omics datasets, including RNA-sequencing and genotype datasets, were examined and new biomarkers were discovered allowing a better understanding of atrial fibrillation. Electrocardiogram signal data were used to assess the early risk prediction of heart failure, enabling clinicians to use this novel approach to estimate future incidences. Population-level data were applied to the identification of associations and temporal trajectory of diseases to better understand disease dependencies in different clinical cohorts.

ACKNOWLEDGEMENTS

I would like to thank my supervisors Paulus and George.

There are many people, involved directly and indirectly to the development of this work. I would like to thank my parents and siblings: Carla, Reginaldo, Hugo, Beto and Luiza. Friends from home: Victor Matheus, Ricardo, Eduardo, Gabriel, Talles and Rafael.

To my late dog Beethoven, may you always stay in our memories sneakily stealing food.

I would also like to thank all the friends I made in Birmingham: Winnie, thanks again for taking me from the college reception to the ICVS on the first day! Luke and John, thanks for introducing me to a whole new world of chocolate rain. Andreas, thanks for showing me that someone can be a wonderful cook and have a sense of humour, style, and be very helpful in a single package. Laura and Dominic, hope you keep on giving some (py)caret/awk shows all around.

On this journey, I had a family away from home, thank you Albert, Renata, Benjamin, and Eric, for the time we stayed together.

Office friends and colleagues from different offices, thank you! Special thanks to Larissa, Liuwei (Chase), Jasmeet, Jun, Sim, and Vasili!

Big thanks to Letícia, thanks for all your support, all the Catan games we played (and that you almost won).

PROJECT OUTCOMES

Chapter 2

Some work on this chapter led to other explorations, one of which was published with collaborators Animesh Acharjee, Joseph Larkman, Yuanwei Xu, and supervisor Georgios Gkoutos:

Acharjee, A., Larkman J., Xu Y., **Cardoso V. R.**, Gkoutos G. V. (2020). "A random forest based biomarker discovery and power analysis framework for diagnostics research." BMC medical genomics 13(1): 1-14. (1): my involvement was in the data analysis and manuscript preparation.

Chapter 3

This chapter involved work developed with collaborators Winnie Chua, Larissa Fabritz, supervisors Paulus Kirchhof and Georgios Gkoutos, and others. This chapter led to these outcomes:

Abstracts

Chua, W., **Cardoso V. R.**, Purmah Y., Tull S., Neculau G., et al. (2018). "P1184 Blood biomarkers associated with atrial fibrillation in a community-based cohort of patients presenting acutely to hospital." EP Europace 20(suppl_1): i229-i229. (2): my involvement was in the model creation and interpretation of the results.

Chua, W., **Cardoso V.**, Crijns H., Schotten U., Guasch E., et al. (2020). “64 A multiple blood biomarker model for identifying patients with prevalent AF” BMJ Publishing Group Ltd and British Cardiovascular Society. (3): my involvement was in model creation and interpretation of the results.

Publications

Chua, W., Purmah Y., **Cardoso V. R.**, Gkoutos G. V., Tull S. P., et al. (2019). "Data-driven discovery and validation of circulating blood-based biomarkers associated with

prevalent atrial fibrillation." European Heart Journal 40(16): 1268-1276. (4): my involvement was in model creation, interpretation of the results and manuscript preparation.

Chua, W., Law J. P., **Cardoso V. R.**, Purmah Y., Neculau G., et al. (2021). "Quantification of fibroblast growth factor 23 and N-terminal pro-B-type natriuretic peptide to identify patients with atrial fibrillation using a high-throughput platform: A validation study." PLoS medicine 18(2): e1003405. (5): my involvement was in the interpretation and visualisation of the results.

Under-review

Chua, W., **Roth Cardoso V.**, Guasch E., Sinner M. F., Brady P., et al. "A Novel Biomarker Model for Detecting Patients With Atrial Fibrillation: A Development and Validation Study.". Pre-print (6): my involvement was in the model creation, interpretation of the results and manuscript preparation.

Chapter 4

This chapter involved work developed with collaborators Jasmeet Reyat, Winnie Chua, Larissa Fabritz, supervisors Paulus Kirchhof and Georgios Gkoutos, and others. This chapter led to these publications:

Abstract

Hepburn, C., Syeda F., Yu T., Homes A. P., **Roth V. C.**, et al. (2018). "128 Desmosomal instability increases atrial arrhythmia susceptibility after endurance training." Heart 104(Suppl 6): A95-A96. (7): my involvement was in the generation of supporting data and interpretation of the results.

Publication

Reyat, J. S., Chua W., **Cardoso V. R.**, Witten A., Kastner P. M., et al. (2020). "Reduced left atrial cardiomyocyte PITX2 and elevated circulating BMP10 predict atrial fibrillation after ablation." JCI insight 5(16). (8): my involvement was in the generation of support data, interpretation of the results and manuscript preparation.

Chapter 5

This chapter had work developed partially with collaborators Winnie Chua, Larissa Fabritz, supervisors Paulus Kirchhof and Georgios Gkoutos, and others, and in the later part of the chapter with collaborators Simrat Gill, Dipak Kotecha, supervisor Georgios Gkoutos, and others.

Abstracts

Gill, S., Sartini C., Uh H., Ghoreishi N., **Cardoso V.**, et al. (2020). "Accurate detection of atrial fibrillation using a smartphone remains uncertain: a systematic review and meta-analysis." European Heart Journal 41(Supplement_2): ehaa946. 3505. (9): my involvement was in the analysis.

Under-review

Gill, S., Bunting, K., Sartini C., **Cardoso V.**, et al. (2021). “Smartphone detection of atrial fibrillation using photoplethysmography: A systematic review and meta-analysis.” Heart Journal. (10): my involvement was in the analysis.

Roth Cardoso, V., et al. (2021). “Validated neural network in routine clinical practice to identify incident heart failure using digital electrocardiograms (cardAIc-ECG)”. Journal of the American College of Cardiology. (11): I had involvement in all stages of the project.

Chapter 8

The work explored in this chapter led to these publications with collaborators Honghan Wu, Simon Ball, Aneel Bhangu, supervisor Georgios Gkoutos, and others:

Publications

Wu, H., Zhang H., Karwath A., Ibrahim Z., Shi T., Zhang X., et al. (2020). "Ensemble learning for poor prognosis predictions: a case study on SARS-CoV2." Journal of the American Medical Informatics Association. (12): my involvement was in data extraction and preparation and manuscript preparation.

Carr, E., Bendayan R., Bean D., Stammers M., Wang W., et al. (2021). "Evaluation and Improvement of the National Early Warning Score (NEWS2) for COVID-19: a multi-hospital study." BMC medicine 19(1): 1-16. (13): my involvement was in the interpretation of the results and manuscript preparation.

CovidSurg Collaborative. Machine learning risk prediction of mortality for patients undergoing surgery with perioperative SARS-CoV-2: the COVIDSurg Mortality Score. (14): my involvement was in the analysis, interpretation of the results and manuscript preparation.

Software

Cardoso, V. R. (2021). gkoutos-group/postcode, Zenodo. (15). This is the supporting software for the analysis on section 3.5 and chapter 7. It provides the functionality of data integration through mapping and collection from different data sources.

Cardoso, V. R. (2021). gkoutos-group/bbcaf_pipeline, Zenodo. (16). This is the analysis pipeline for the cases studied in sections 3.3, 3.5 and 5.2.

Cardoso, V. R. (2021). gkoutos-group/clustering, Zenodo. (17). This contains the methodology for the analysis performed in section 6.6.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Additional biomarker names listed in Appendix 2.1.

AC – Arrhythmogenic cardiomyopathy

AF – Atrial Fibrillation

ANOVA - ANalysis Of VAriance

ANG2 – Angiopoietin

ASA – American Society of Anaesthesiologists

AutoML – Automated Machine Learning

AUCROC – Area Under the Receiver Operating Characteristic Curve

BBCAF – Birmingham Black Country Atrial Fibrillation Registry

BIC – Bayesian Information Criterion

BMI – Body Mass Index

BMP10 – Bone Morphogenetic Protein 10

BNP – B-type Natriuretic Peptide

CA125 – Cancer Antigen 125

CABG - Coronary Artery Bypass Graft

CATCH-ME – Characterizing Atrial fibrillation by Translating its Causes into Health Modifiers in the Elderly

CNN – Convolutional Neural Networks

CRP – (cardiac) C-Reactive Protein

DNA – DeoxyriboNucleic Acid

DNN – Deep Neural Networks

ECG – Electrocardiogram

EHR – Electronic Health Records

ESM1 – Endothelial Specific Molecule 1

EU – European Union

eQTL – expression Quantitative Trait Loci

ETL – Extract Transform Load

FABP3 – Fatty Acid Binding Protein 3

FEV1 – Forced Expiratory Volume in 1 second

FGF23 – Fibroblast Growth Factor 23

FN – False Negative

FP – False Positive

FVC – Forced Vital Capacity

GDF15 – Growth Differentiation Factor 15

GDPR – General Data Protection Regulation

GWAS – Genome-Wide Association Studies

HCM – Hypertrophic cardiomyopathy

HDR – Health Data Research

HES – Hospital Episode Statistics

HF – Heart Failure

ICU – Intensive Care Unit

IDI – Integrated Discrimination Improvement

IGFBP7 – Insulin-like Growth Factor Binding Protein 7

IL – Interleukin-6

KCH – King's College Hospital

LA – Left Arm (ECG lead); Left Atrial (heart chamber)

LASSO - Least Absolute Shrinkage and Selection Operator

LL – Left Leg (ECG lead)

LSOA – Lower layer super output area

LSTM – Long Short-Term Memory

LVEF - Left Ventricular Ejection Fraction

LVESD – Left Ventricular End Systolic Diameter

MAR – Missing At Random

MCAR – Missing Completely At Random

MICE – Multiple Imputation by Chained Equations

MNAR – Missing Not At Random

MRI – Magnetic Resonance Imaging

mRNA – Messenger RiboNucleic Acid

MSE – Mean Squared Error

MSOA – Middle layer super output area

NHS – National Health Service (UK publicly funded healthcare)

NN – Neural Networks

NRI – Net Reclassification Index

NTproBNP – N-terminal pro-B-type natriuretic peptide

NYHA class – New York Heart Association functional classification of heart failure

PC – Principal Component

PCA – Principal Component Analysis

PEFR – Peak Expiratory Flow Rate

PICS – Birmingham Systems Prescribing Information and Communication System

PPG – Photoplethysmograms

R^2 – Coefficient of determination

RA – Right Arm (ECG lead)

RCRI – Revised Cardiac Risk Index

RDF – Resource Description Framework

RIMARC – Ranking Instances by Maximizing the Area under the ROC Curve

RIN – RNA Integrity Number

RNA – RiboNucleic Acid

RNA-Seq – RNA Sequencing

ROSE – Random OverSampling Examples

RL – Right Leg (ECG lead)

ROC – Receiver Operating Characteristic

SARS-CoV-2 – Severe Acute Respiratory Syndrome CoronaVirus 2

SHAP – SHapley Additive exPlanations

SMOTE – Synthetic Minority Oversampling TEchnique

t-SNE – t-Distributed Stochastic Neighbor Embedding

THIN – The Health Improvement Network

TIA – Transient Ischaemic Attack

TN/TNR – True Negative/True Negative Ratio

TnT – (cardiac) Troponin T

TP/TPR – True Positive/True Positive Ratio

UHB – University Hospitals Birmingham

UMAP – Uniform Manifold Approximation and Projection

UK – United Kingdom

USARE – Usable Summarised Anonymised Re-loaded External data framework

+/- – Heterozygous

CHAPTER 1 INTRODUCTION

Typically, health studies have focused on generating data to explore specific scientific hypotheses. This process of collecting data is expensive and burdensome – it requires designing studies, recruiting participants and collecting data. In the current times, more and more data are made available, and many facets of these datasets are still open to investigation. This work proposed to harness the potential of existing datasets and repurposing them by developing novel analytical approaches to interrogate them – rather than generating them. It aims to better understand clinical outcomes with applications as its driving force.

This thesis is multidisciplinary in nature integrating and developing methods stemming from the fields of data science, informatics, computer science, mathematics as well as bioinformatics and health informatics for application in health sciences, with a focus on cardiovascular diseases. It is structured according to different research work related to the investigation of Atrial Fibrillation (AF) and its related comorbidities. Different datasets were explored, such as the Birmingham Black Country Atrial Fibrillation registry (BBCAF), for the development of methods for AF prediction. Transcriptomic analysis of AF-related RNA Sequencing (RNA-seq) data to the identification of novel biological targets, from mice and human samples. Signals from electrocardiogram (ECG) were applied for patient risk stratification, derived from electronic health records. Moreover, this project included the investigation of different patterns of comorbidities in different health data sources and the systematic integration of electronic healthcare records.

These different approaches, when combined, provide a wide background of knowledge that can be used to assess patient risk, investigate new therapeutic targets, and improve patient stratification.

In the sequence on this chapter, different concepts from information and data to different biological terminology are defined.

1.1 Information systems and data

Information systems are instruments that collect, store and provide data. Their use can be related to the origin of ancient history with the Sumerian Cuneiform script, as the earliest known record of written information. Civilizations could extend their reach with this technology that enabled long-distance communications and facilitated trading (18).

Written information in the past assisted with decision making and its importance in decision making is ever increasing. In this thesis, information that can be stored is considered data, and data that denotes any knowledge, right or wrong, are facts derived from the data perspective. Data have implicit meaning and a universe of discourse. It is considered that a collection of related facts is a database (19).

This definition of databases is not limited to electronic ones. It is possible to consider old document systems in hospitals as databases as well. Most of this information would be stored in paper archives individually for each patient with reports and exams, but others are stored in digital repositories, often limited to a specific type of test, e.g., biomarker concentrations or ECG or imaging databases. Since the early 1990s, there has been a shift towards Electronic Health Records (EHR) systems (20). These EHR systems are a substitute for previous paper systems, providing extensive benefits, such as distributed access, collaborative work and faster access to patients' exam results, diagnosis, notes, prescriptions, scheduling, stock management, rotas and any other possible event in the context of the hospital. The University Hospitals Birmingham NHS Foundation Trust (UHB) introduced an electronic health record in 2008, and currently has been using different software to support its EHR initiatives (21), for example, Birmingham Systems Prescribing Information and Communication System (PICS), a rules-based prescription support system.

EHR systems and their datasets are an example of **real-world data**. This is the data reflecting actual clinical practice. Usually, there are minor limitations on participant selection depending on the inclusion criteria. In the context of UHB, limitations could be for cases such as hospitals that are focused on children and/or women care, or patients that live closer to a point of care. However, these limitations are not as penalising as the trial recruitments based on consultations with particular study

practitioners. In real-world data, participants are not bound to specific periods of registration and follow-up, the research support team, or financial constraints that happen in trial studies. Due to participant mixture in these data, research and development of novel hypotheses based on such datasets are expected to suffer minor deviations when applied to external populations. That is, while a study would have a sampled population, these real-world data covering a bigger population sample will more closely approach the overall population and thus a standard clinical routine scenario, and it is then expected to have a lessened deviation for its created models.

Despite the existing benefits of real-world data, it does not solve all types of problems, particularly when evaluating novel traits and treatments. There are no retrospective data for new drugs, neither information on how new treatments will affect patients. The use of real-world data might also require special compliance with ethical regulations. For example, in the United Kingdom (UK), compliance is required with the National Healthcare System (NHS) and other healthcare regulations, such as the Health Research Authority approval (22). This is due to cases in the past where participants were exposed to great harm and risks. Furthermore, ethical requirements ensure that participants data and privacy are kept confidential on a need-to-know basis. Regulations also provide clear information to participants and a plan of action to researchers in case of any unexpected events.

Other types of data considered are **study data** and **simulated data**. Study data contains data collected from trials, these data are usually used to evaluate novel markers, medications, or treatments not done in practice. Simulated data mimics real behaviour, these are commonly applied for modelling medication behaviour before any biological tests. Furthermore, other techniques evolve the generating of new data for model creation. These two cases are explored in this thesis.

The multitude of information in the hospital is an example of big data. Big data is a collection of facts, stored in databases or otherwise, in a wide range of structured and unstructured formats, typically associated with a great number of data-points and data that may contain information that increased exponentially. These characteristics define big data: variety, volume and velocity, the 3 V’s of big data (23). Furthermore, datasets

contain a multitude of information in them and about them. The former is the data itself, the latter is the metadata, the data about the data.

The metadata contains important information about datasets. Clinical care settings and treatment dates are examples of metadata. A patient could have a different treatment depending on the accessibility to healthcare, and local infrastructure. And, to some extent, the way phenotypes are identified and reported by different clinicians lead to a different pattern of the data, e.g., a clinician, due to training and experience, could be focused on some symptoms/signs than others. And, as a broad rule, dates on which a patient was being treated affect significantly the patient pathway: the availability of treatments, new devices, new guidelines, government policies, and external events change the data patterns, making a dataset population substantially different. The inclusion criteria, i.e., the sets of rules used to select patients in a study, are important metadata factors because they could bias the analysis. To some extent, it is possible to say that a dataset from the UK was formed based on a region and population inclusion criteria, although the genetic background of the population could be diverse. Therefore, all factors in the metadata must be considered, as they provide additional meaning to the dataset.

1.1.1 Data modalities

Data have different dimensions and types. These different dimensions and types, similarly to the universe of discourse, discussed previously, help to define the context of the data. The dimensions provide information on contexts, or subgroups, of different data elements. In a biological context, the dimensions may be a cellular, tissue, organ, system, patient and population context. At the patient level, different data types are considered.

Datatypes can be separated into **structured**, **unstructured** or **semi-structured** data. These are categorised due to different approaches and analyses that can be applied to each of them. Structured data refers to all data that contains a tabular or computable structure that is well defined, and the meaning is objectively measured, with clear and unambiguous values. This is the case for a dataset of diagnosis data such as ICD-10s, as the patient data is either annotated with the ICD-10 or not (24). Other examples are

demographics such as age and sex, given that age is marked in a cell and there is no ambiguity about the definition of the term. Unstructured data types contain graphical or signal information, such as medical imaging datasets or recording data, as in the case of electrocardiograms. These data follow a collection protocol and there might be a structure, such as a defined procedure with an expected repetition or sequence recorded. However, the recording itself does not provide any direct information before being interpreted/analysed by a person or algorithm. Semi-structured data are sets of data that contains information in a seemly structured form, such as text. These data have a grammar defining a structure, but the information cannot be directly used in an automatic analysis.

A type of structured data types, omics-related data is considered a type of data that contains quantified data derived from particular measurements of biological processes. The data derived from an omics process contains a multitude of variables in a similar context. Some examples are RNA-seq data, containing a multitude of information about a sample, such as data from transcriptomes expression; proteomics, or blood assays, cases that contain information about a range of blood measurements in the same context; genomics, with genome-wide genotyping data; radiomics with a large number of features extracted from radiographic medical images; and others.

Similar data can be represented using different modalities, for example, rather than annotated with ICD-10 (24) codes, patient data may be available as free text, such as the case of a clinical letter or note. Data may be transformed between modalities. For example, a patient's genome is formed of a combination of four bases; the genome has a grammar and is an example of semi-structured data. These data can be inspected if its grammar is known, but also, this genetic material could be transformed to other formats, such as processed into a signal (unstructured data), which can then be transformed into a structured datatype, indicating measurements. A final report by the investigator could lead to summarised textual data, back to semi-structured data. Figure 1 illustrates this process.

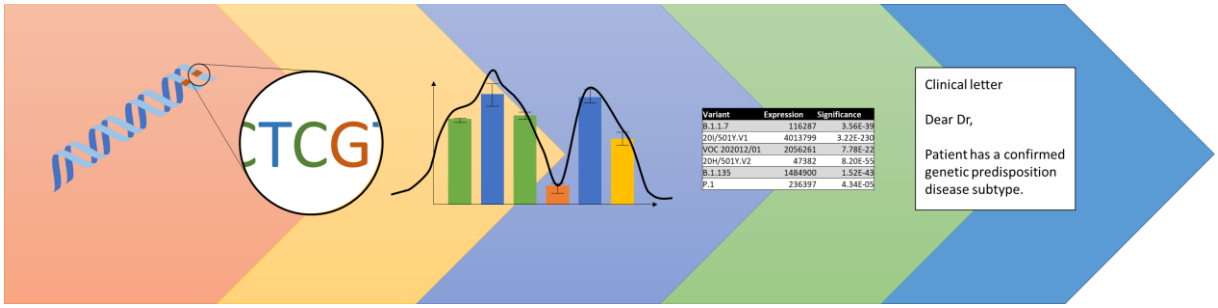


Figure 1: Flow of data through different modalities. From genome (semi-structured data), the zoomed-in sequence of bases is read using an approach that considers the signal (unstructured data) for each read sequencing, marking them as different bases, the sequenced data can be combined into count data (omics, or structured data), which then could be used to generate a clinical output such as a clinical letter (semi-structured data).

These different datasets have different structures and are often stored in different locations of sub-systems, where different parts of the data are stored in different databases, sometimes with different naming conventions and incomplete linkage issues. This is the main challenge of analysing multimodal data. There are also other definitions for data modalities, with some similarities (25).

1.1.2 Data scales

Data modalities refer to different types of view on a data or element. There is also a factor of the resolution of the data. In the following paragraphs is illustrated the different scales of data in a human.

At the deepest resolution, there is information about specific nucleic acids and proteins. One could look at directly quantifiable aspects, e.g., the transcriptome, genome, or biomarkers levels, or structural and functional aspects, e.g., the interaction between proteins.

Cells and tissues are next on the scale, as proteins and other biomolecules compose cells that then compose tissues, cells and tissues behave differently depending on their specialisation. For example, cardiomyocytes, cells that make part of the cardiac muscle, have a differential expression to cells in the liver, or any other tissue. There are also differences in cells and tissues behaviour within the same organ, for instance,

tissues from different heart chambers contain marked differences in their genetic expression.

Next are organs and organ systems, which individually or in combination may function in a disorderly fashion. An unbalance in a system is sometimes a response to another system's abnormal behaviour. This is seen in diseases that are co-morbid with other diseases, and the outcome severity is closely related to another condition. For example, patients with chronic kidney disease are more likely to develop atrial fibrillation and vice-versa.

Individuals in populations are the largest scale, these are information associated with lifetimes, where different lifestyle choices trigger mechanisms hidden in the lowest scale nucleic acids.

When working with data, the population level is the most common unit of operation, with individuals containing depths of information that can be explored on different facets. Populations with a clear separation of outcomes in combination with statistical frameworks provide the basis for the investigation and discovery of new targets and patient stratification.

There is a wide variety of data available, datasets in different modalities and resolutions. This thesis presents the application of approaches to add information through the integration rather than the creation of new data with hypothesis generation and results from separate data modalities and scales. Different methods were applied to integrate and analyse data to advance the knowledge of atrial fibrillation and other cardiac and related comorbidities.

1.2 Heart

The heart has 4 chambers, 2 atria (upper chambers), and 2 ventricles (lower chambers). The heart's function is to pump blood around the body, and it is composed of myocardium, muscle tissues of the heart. The left atrium receives oxygenated blood from the lungs through the pulmonary veins and passes it to the left ventricle, which sends the oxygenated blood under pressure to the body via the aorta. After passing through the body, the deoxygenated blood returns to the right atrium, where it gets

pumped to the right ventricle and back to the lungs. The blood gets oxygenated in the lungs, before going back to the left atrium, restarting the cycle (26). Figure 2 illustrates the blood flow in the heart.

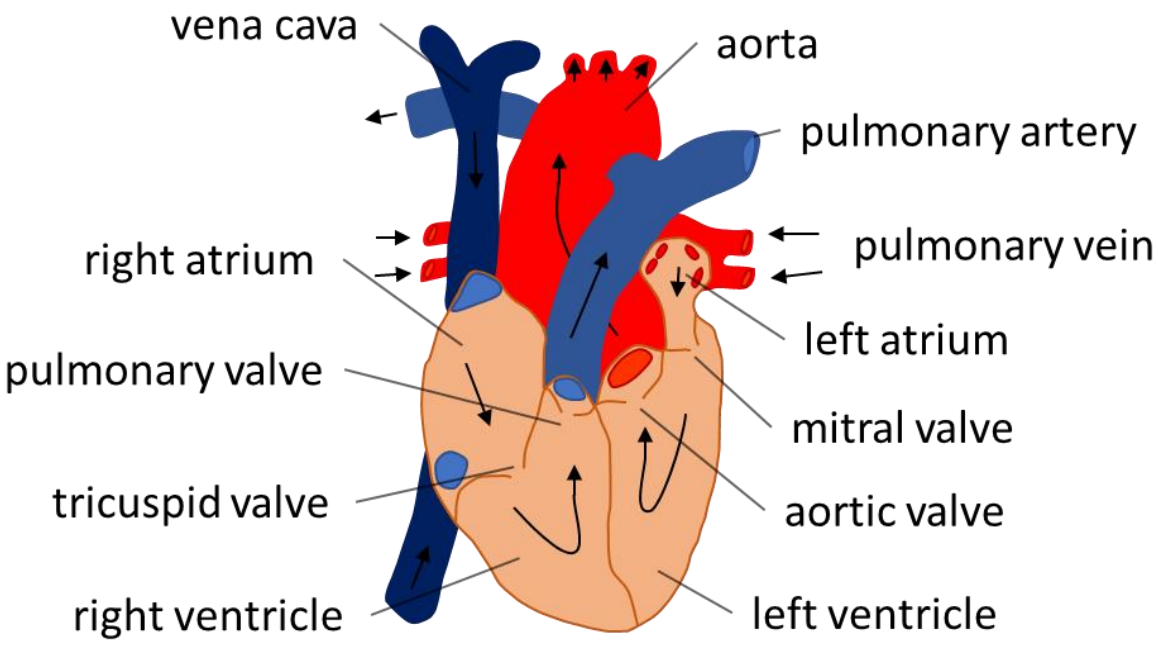


Figure 2: Heart chambers and the flow of blood. Deoxygenated blood arrives from the vena cava to the right atrium. The blood flows to the right ventricle, then to the pulmonary artery and to the lungs, which is then returned oxygenated from the pulmonary vein to the left atrium, then left ventricle before going back to the body by the aorta.

The heart is regulated by the sinoatrial node, which is the natural pacemaker of the heart. Located in the right atrium, the sinoatrial node is also responsible for starting the signalling process. The signal sent by the node (a) contracts the atria, (b) spreads through the heart, and (c) slows down as it passes the atrioventricular region, the region between the atrium and ventricles. In the sequence, the ventricles contract rapidly.

The ejection fraction is a measurement of the heart's capability to eject blood from a heart chamber, it is measured in percentage. A standard measurement comes from the left ventricle, the left ventricular ejection fraction (LVEF) is commonly measured by

an echocardiogram. LVEF's normal range is between 55% and 70%. Reduced values indicate the likeliness of heart failure (27).

1.2.1 Atrial Fibrillation

Atrial Fibrillation (AF) is the most common serious arrhythmia of the heart. It is associated with several health complications and can lead to death. AF is caused when the atria do not contract properly, and instead quivers, because of an irregular activation of the atrium by the sinoatrial node (28).

AF has three different stages: paroxysmal, persistent, and permanent. In the initial form, paroxysmal, the patients have infrequent episodes of AF which are self-terminating. These episodes may become longer and more frequent, and the disease might evolve to persistent AF. A patient with short-standing persistent AF has episodes that last for more than 7 days and may potentially be interrupted with direct cardioversion. The long-standing persistent AF can last for more than 1 year. The final stage, permanent AF, is a type of AF where control of the rhythm is either not intervened and is always persistent, or it is resistant to electrical or pharmacologic cardioversion (29). Figure 3 illustrates the evolution of the disease.

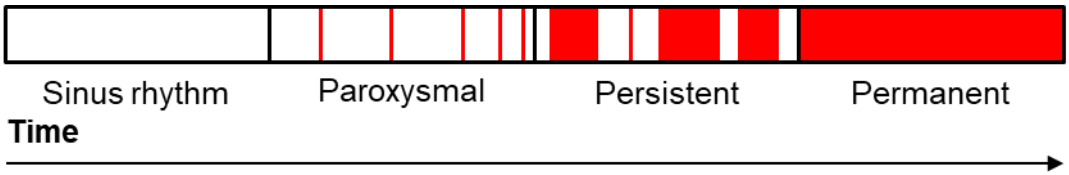


Figure 3: Evolution of atrial fibrillation. Red indicates atrial fibrillation episodes. As the disease progresses over time the atrial fibrillation episodes become longer.

The start and progression of AF are associated with inflammation. Inflammatory pathways contribute to atrial remodelling, and AF leads to inflammation, circling back (30).

AF can be identified by the pulse and then confirmed with an electrocardiogram (ECG) or a Holter ECG. Both devices measure electrical activity in the heart over time. A standard ECG has a recording of 10s and contains 12 leads, which are different signals with different references. Holter ECG is a portable ECG that patients carry for at least

a day and sometimes for as long as 2 weeks at a time. It is used to identify silent AFs, a subtype of atrial fibrillation in which patients go into advanced stages of the disease without being previously identified for AF (29).

Currently, there is no cure for AF, but the treatment of several domains can prevent complications, reduce symptoms, and improve the quality of life in affected patients. In brief, these comprise acute restoration of normal rhythm, treatment of concomitant cardiovascular conditions, anticoagulation to prevent ischemic strokes, rate control, and rhythm control therapy (29). Table 1 describes some treatment approaches for AF.

Table 1: Description of treatments for AF. All treatments aim to improve the quality of life, autonomy and social functioning of AF patients. All but the last treatment also provide the benefit of improved life expectancy. Table based on a figure from (29).

Case	**Treatment**	**Desired outcome**
Acute rate and rhythm control	Beta-blockers, cardioversion	Haemodynamic stability
Manage precipitating factors	Lifestyle changes, treatment of underlying cardiovascular conditions	Cardiovascular risk reduction
Assess stroke risk	Oral anticoagulation	Stroke prevention
Assess heart rate	Rate control therapy	Symptom improvement, preservation of LV function
Assess symptoms	Antiarrhythmic drugs, cardioversion, catheter ablation, AF surgery	Symptom improvement

Some of the conditions commonly co-morbid with AF are hypertension, vascular disease, notably coronary artery disease (CAD), heart failure (HF), diabetes, and stroke (29). AF prevalence is estimated to be 2%, with an increased risk at higher ages; 0.12%-0.16% in patients under 49 years, 3.7%-4.2% in patients aged 60-70 years and between 10% and 17% in patients older than 79 years (31). AF prevalence is expected to be higher given that patients are usually identified only after another serious disease, for example, in the case of stroke 25% of patients are likely to have AF episodes (29).

Of the different comorbidities associated with AF, HF is further studied in this thesis in a predictive model. HF is typically associated with a few symptoms, such as dyspnoea (difficult breathing), fatigue and swelling (32). One way to identify heart failure is through the assessment of left ventricular function using an echocardiogram and its LVEF measurement, a reduced value would indicate heart failure. The LVEF measured from the echocardiogram could be normal and the patient has heart failure, in this case, the patient is categorised as heart failure with preserved ejection fraction. It is also possible to diagnose using blood tests, breathing tests and chest x-ray. HF association with AF is known for more than a century (33). There are many similarities in their risk factors, such as age, hypertension, diabetes, obesity, and other cardiovascular conditions, e.g., valvular, ischaemic and non-ischaemic structural heart disease (34).

Some open research problems addressing AF are the early identification of the disease, the discovery of long-term cures and the improved definition and stratification of patients into subgroups. Although not validated, there is a suggestion in the literature it is that AF could be classified, according to mechanisms, in the following groups: monogenic AF, focally induced AF, postoperative AF, valvular AF, AF in the elderly, polygenic AF and unclassified AF (35).

The main data sources related to AF patients are (a) the clinical history of the patient, phenotypes of preconditions, and associated measurements, (b) electrocardiogram recordings, (c) echocardiogram, images obtained using ultrasound of the frontal surface of the heart, and other imaging modalities of the heart, (d) blood biomarkers such as haemoglobin, creatinine, and cardiovascular biomarkers that may be more specific for AF, and (e) genetic and transcriptomic data, such as RNA-sequencing data and antecedents' information. Their value and the composition of these data is discussed later in this thesis.

Early and correct identification of patients with any condition is crucial to increase patient's quality of life and disease outcomes. There are many models for risk assessment of AF, such as CHA_2DS_2-VASc, which combines different risk-conditions to indicate if a patient that suffers from non-rheumatic AF is likely to have a stroke (36,

37). Also, the New York Heart Association Functional Classification (NYHA class) indicates patients' limitation on physical activity (38).

1.2.2 Electrocardiogram and the PQRST Complex

One of the most common ways to quantify heart activity is by using an Electrocardiogram (ECG). The ECG identifies different sections of the heart's movement by its electric discharge using different leads placed around the chest. The relative position of the leads to the heart influences the signal data that can be obtained. Different leads obtain information focused on specific heart chambers. Wrong positioning of the leads and body movements, such as the abdominal movement during breathing may influence the reading (39). Other factors, such as sex, age, body mass index (BMI) and athlete conditioning alters an ECG signal (40, 41).

There are some variances in the ECG recordings. Usually, a patient that gets an electrocardiogram needs to be in a supine position. The patient rests for a few seconds while 4 limb electrodes, Right Arm (RA), Left Arm (LA), Left Leg (LL) and Right Leg (RL) and 6 chest electrodes are placed in a range around the heart (Figure 4) (42). A standard electrocardiogram recording lasts 10s.

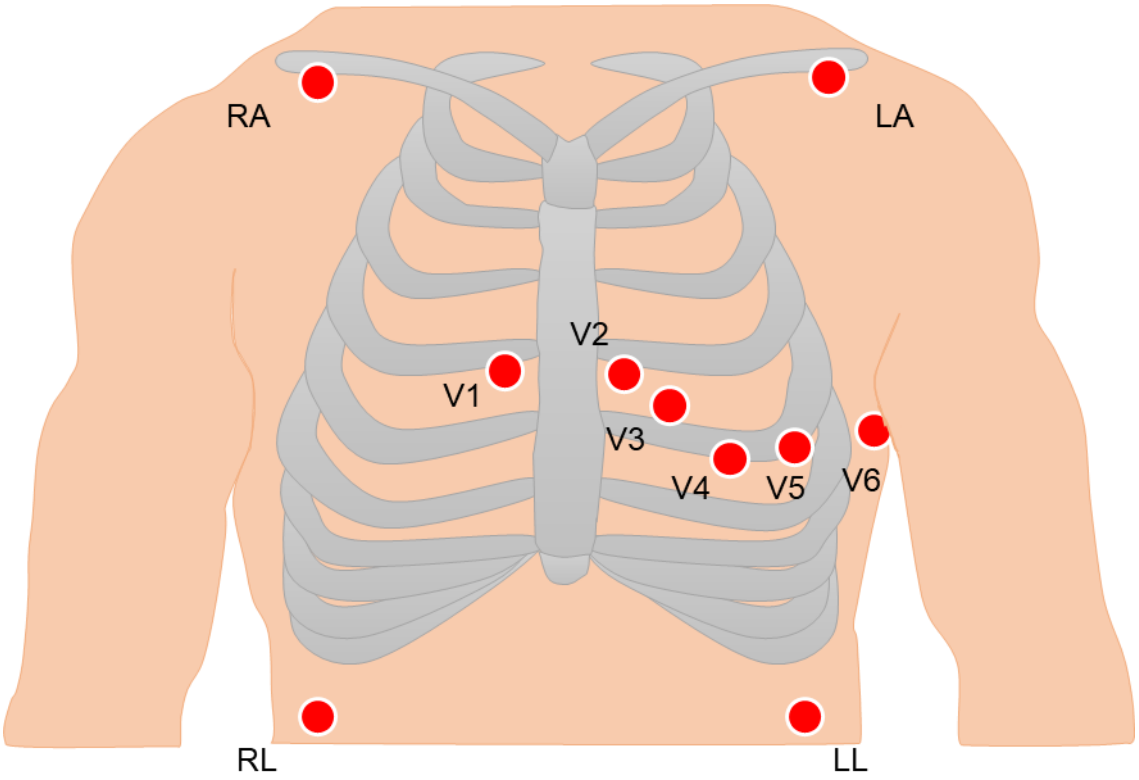


Figure 4: Positioning of the leads in the chest. Depending on the protocol, Right Arm (RA) and Left Arm (LA) leads can be anywhere between the shoulder and the respective elbow. Right Leg (RL) and Left Leg (LL) leads may be placed below the torso and above the respective ankle.

Whilst 10 electrodes are placed, only 9 are directly used to obtain information, as the electrode RL is for grounding. The signals from 9 electrodes are collected and combined as if 12 leads were placed (43). The different calculated leads are:

$$I = LA - RA \quad II = LL - RA \quad III = LL - LA$$

$$Thus, II = I + III$$

$$aVR = \frac{LA + LL}{2} \quad aVL = \frac{RA + LL}{2} \quad aVF = \frac{RA + LA}{2}$$

$$aVR + aVL + aVF = 0$$

Modern ECG devices provide an automatic diagnosis of common conditions that are easily identifiable. There are many nomenclatures for these conditions and approaches to unify them have been proposed (44). It is informally stated by some clinicians that if

the automatic diagnosis does not pick any condition then the patient is very likely to be healthy.

In the first stage of a heartbeat, the *P wave* indicates the start of the beat with atrial depolarisation. After that, there is a time interval with no signal as the blood is moving from the atrium to the ventricles. Following, there is the *QRS Complex* (Q is the lowest next point, R the very high point and S the next lower point): the *Q wave* indicates the depolarisation of the septum, when the ventricles are being activated. The *R wave* indicates the activation of the ventricles (as it has more muscle, it requires higher electricity) and the *S wave* is the electric discharge returning to its neutral position. And the next wave, the *T wave* is the repolarisation of the ventricles before the cycle repeats. There are different intervals classically used in the analysis of ECG signals. Figure 5 illustrates the different parts of a heartbeat and some measurements possible. The *R-R interval*, which denotes the length between two beats, is an important inter-beat variable.

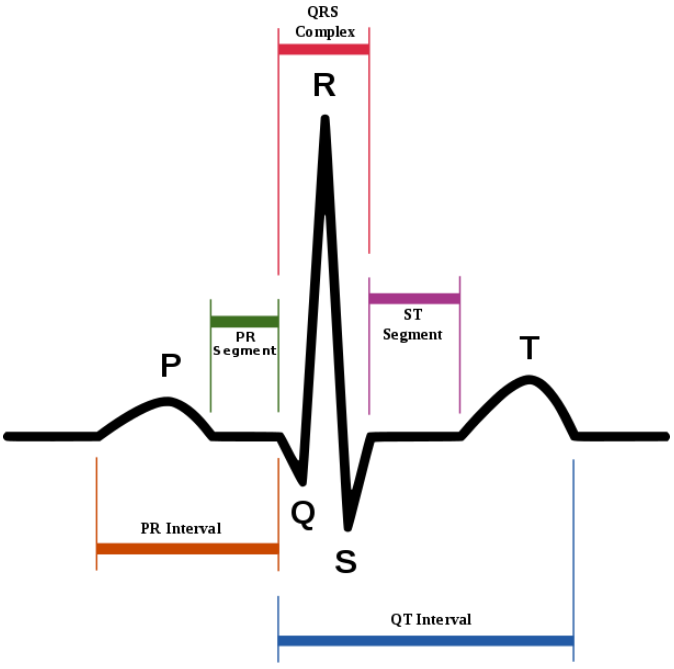


Figure 5: Different stages of a heartbeat. Some intervals are measured to assess normality and diseases.

The combination of the information from the shape of the wave, intervals and amplitudes, and its repetitions can be used to identify diseases. One lead of special

interest to AF is lead II, which indicates the rhythm and can be used to identify diseases associated with it. Patients who suffer from AF tend to have noisier recordings, irregular *R-R intervals* and the *P wave* might be absent. Figure 6 displays a comparison between an atrial fibrillation patient and a healthy patient.

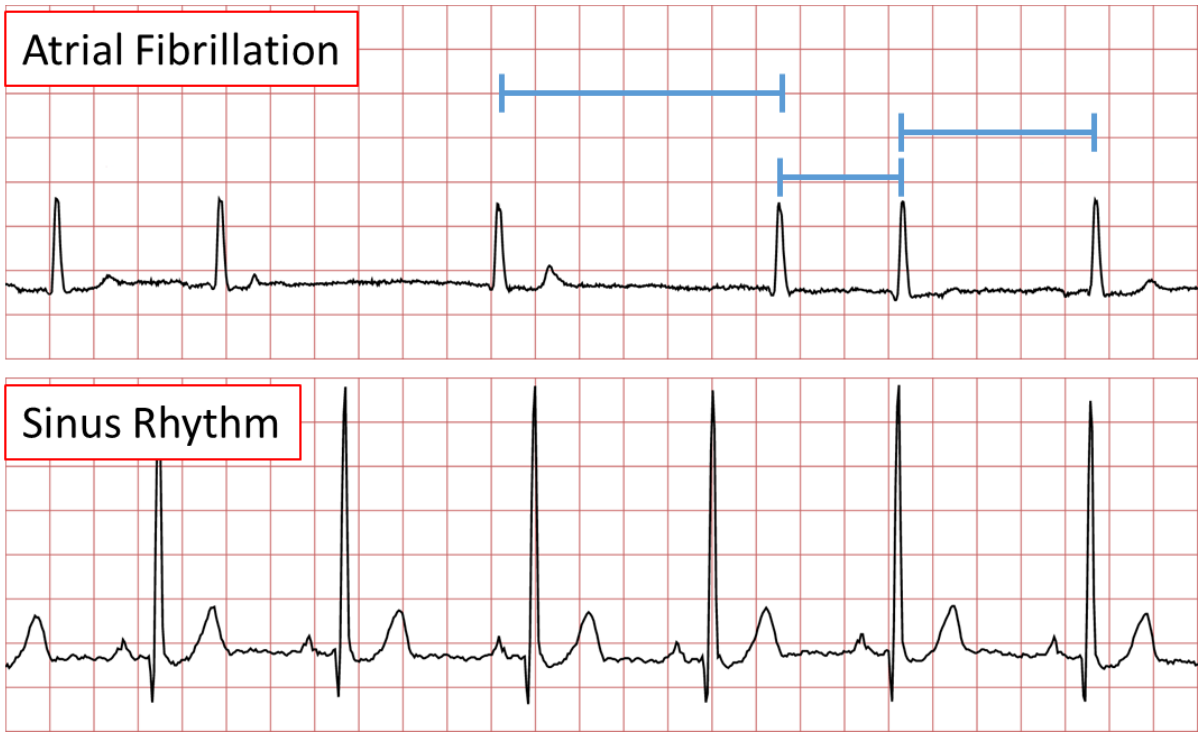


Figure 6: Electrocardiogram comparing an atrial fibrillation patient and a healthy patient. Note the variation in the R-R interval. Signals from MIT-BIH Arrhythmia dataset (45, 46).

In summary, an electrocardiogram offers information about heart activity. It is possible to extract many variables from an ECG, such as P-Q interval, R-R interval, QRS width and Q-T interval. The relationship between these variables can be used in the study of different conditions (47).

There are many methods of storing ECG data in an electronic format (48). ECG recordings handled on this thesis' projects used XML formatting (49). This format contains a header with metadata, such as device information and patient identifier, followed by the different leads in different sections.

1.3 Omics

The genome contains the genetic information for an organism's development, growth, functioning and reproduction. It contains information that is both coded to genes as well as non-coding information that may have regulatory functions. In retro-viruses, the genome information is contained in some ribonucleic acid (RNA), while in other living beings the information is in a deoxyribonucleic acid (DNA), a double helix-shaped molecule formed by two chains of polynucleotide strands (50), which are made of a chain of the nucleotides adenine, cytosine, guanine and thymine in different sequences called chromosomes. In humans, there are 23 pairs of chromosomes which are estimated to have around 20,000 protein-coding genes (51).

The processing of information from genes is enunciated in the central dogma of biology, which describes two processes that explain the flow from the DNA molecule into proteins. The first process is the transcription of DNA into messenger RNA (mRNA), and the second one is the translation of this RNA into proteins (52). These processes are required for the functioning of any living cell and the performance of its biological purposes.

There could be alterations in the functional characteristics in any step. In the first process, a single nucleotide polymorphism (SNP) could affect the whole chain that leads to the expression of a protein. In the second process, the expression of some mRNA could be substantially higher or lower. Furthermore, the manifestation of some proteins may be different in different scenarios.

Omics refer to the wide group of different -omics type of analyses. Genomics is the scientific field that studies the genome. Transcriptomics is the sub-field that studies transcripts, such as the quantification and evaluation of differential expression of the mRNA. Proteomics is the study of interactions, modifications, and location of proteins (53).

1.3.1 Transcriptomics

RNA sequencing (RNA-Seq) is explored in this thesis as an example of omics datasets. This data comes from different biological samples that had their mRNA material

sampled, then these data are processed into a count of transcripts. The different transcripts were then compared for their differential expression. This work aims to understand diseases and their underlying mechanisms (54, 55).

Mice, *Mus musculus*, are used as model organisms because of the similarity to the human genome and the possibility to isolate conditions to be tested, i.e., a knockout mouse, a specimen that does not have one gene – homozygous or heterozygous absence – that is commonly found in a non-genetically modified mouse.

RNA-Seq investigations on this project that used mice involved transgenic mouse heterozygous on the JUP and the PITX2 genes, separately. PITX2 was picked because variants close to it have been associated with AF and is a promising target (56). JUP gene (Plakoglobin) has been associated with arrhythmogenic cardiomyopathy and it is not widely studied (57).

Furthermore, data collected from tissue samples of patients undergoing cardiac intervention surgery were explored using RNA-Seq techniques. As this dataset has a comprehensive view of the patients, and this type of data not being commonly available, this dataset provides a new and unique perspective on cardiac patients.

1.3.2 Genomics

Genomics is explored in this thesis as another example of omics datatype. This datatype was evaluated using Genome-Wide Association Studies (GWAS) analysis (58). This type of analysis involves the use of statistical analysis to identify SNPs of importance to the groups compared.

1.4 Work proposed

As outlined in the introduction, the susceptibility to common cardiovascular rhythm disorders and other cardiovascular diseases is dependent on a multitude of risk factors which can have genetic or environmental origins. Given the technological advances in electronic health record generation, genetic sequencing and protein biomarker quantification, there is a growing wealth of data which, when combined with classical clinical parameters, offers the possibility of improving the care for patients with

underlying cardiac diseases such as AF and HF. As a result, there is a need to better identify patients at increased risk of these cardiac diseases and identify patient subsets who would benefit from selective therapies. The use of advance data analysis methodologies, such as machine learning and artificial intelligence, offers a means to better understand patients with cardiovascular diseases. The best understanding of patients works similarly to what precision medicine targets: with the better understanding of patients, it is possible to personalise their treatment and improve their care (59).

To better understand patients' pattern of morbidity, different aspects of cardiovascular disorders were investigated in this work, separated by their data types and techniques that can be explored: Chapter 2 explains broadly the different methodological tools used. Chapter 3 illustrates the use of structured data, inclusive of the exploration of clinical, biomarkers and socioeconomic data by combining data from the UK Biobank and patients recruited to an AF study. Chapter 4 explores the use of omics data, going deep into the DNA and RNA data with relevance to AF. Chapter 5 explores the use of unstructured data, with ECGs being used as a source of novel morbidity' markers in cardiovascular diseases. Chapter 6 takes a step back and goes into populations, how they group together and explain morbidity. Chapter 7 shows a framework for the collection and processing of data, from the UHB and the UK Biobank, to a streamlined analytics process. Chapter 8 shows the applications of the methods and data described to an urgency scenario, relating to the SARS-CoV-2 pandemic.

CHAPTER 2 DATA DESCRIPTION AND METHODOLOGY

2.1 Introduction

This chapter reviews analysis techniques commonly used in medical studies. These techniques compose the basic toolset used in the analysis described throughout the thesis.

There are 5 parts to this chapter, (a) data pipelines, the description of the analytical framework used, (b) data sources, different datasets that were used in multiple parts of this work, (c) data analysis, basics of statistics, missing values and metrics, (d) artificial intelligence, advanced analytical methods, (e) variable importance, different approaches to identify important predictors in models. The different parts are summarised in Table 2.

Table 2: Summary of the Data Description and Methodology chapter.

Category	Definition	Applications on this thesis
Data pipeline	The end-to-end process to perform data analyses, from planning the experiment to reporting the results achieved.	This methodology guided the analytical process in this thesis. This ensured the results achieved complied with applicable scientific guidelines.
Data sources	Description of the different datasets utilised.	These were used throughout the thesis in multiple projects.
Data analysis	Approaches to analyse the data, basics of statistics, handling unbalanced datasets and metrics	These were used when describing data, creating models and reporting results.
Artificial Intelligence, Machine Learning	Advanced modelling approaches, basics on algorithms such as decision trees and neural networks	In combination with the data analysis section, these methods were used when creating more advanced models, such as ECG models (Chapter 5).
Variable importance	The identification of important predictors is sometimes more important than the performance metric of the overall model.	These were employed when assessing some models results, these were extensively used in the BBCAF analysis (Sections 3.3, 3.4, 3.5, 5.2, 8.3).

2.2 Data pipeline

The process of any data analysis can be separated in different ways. The five main steps are design, collection, transformation, analysis, and data reporting. Depending on the project type, the different stages are iterated multiple times until project closure. This section describes an experimental framework.

Design. The experiment is planned. Questions are raised and possible approaches are hypothesized with or without data available. Requirements raised and complied for data governance, ethics and scale of the experimentation, methods and outcomes expected; for example, in the UK, the Data Protection Act 2018 and the European Union (EU) General Data Protection Regulation (GDPR). The experiment might aim to model risk, stratify patients into known or new sub-groups, identify potential new biomarkers using traditional interpretability models or maximize its performance metric, not limited to these. Some considerations might be about the power of the statistical analysis, whether there are enough numbers for the purposed analysis. In epidemiological studies, it is recommended to have at least 200 cases and 200 controls samples for validation (60).

Collection. The collection of data consists of processing data into an electronically processable format, either by acquisition or transformation. There are three main forms of data collection: (a) extraction of a derived dataset from another electronic source, such as a derived dataset or a web source, (b) compilation of data from other forms of archives, such as typing an exam result or passing the records from written forms into a digitalized system, and (c) forms and questionnaires applied to individuals. In the studies analysed in this thesis, the data are already in a digital format, type (a), and data are combined in different ways for different studies.

Transformation. Not only transformation but also the selection of the data. An electronic processable dataset may not be fit to analysis questions without further preparation of the data. A dataset might need to be combined from different tables, aggregated data might need to be processed, data transformed from a wide to a long format, or vice-versa. Sets of data-points might be further-selected for different analysis using the same or different methods, which is called sensitivity analysis (61).

Analysis. Before executing any analysis, a performance metric has to be selected, it will identify what analytical approach works best, and define a criterion of comparison. The analysis steps can be done as described throughout this chapter. Models can be created in a supervised or unsupervised way: optimizing the separation of classes, or sub-grouping of elements: the metric needs to be compatible with the methodological approach involved.

Data reporting. Data without any visual representation is harder to understand. Data visualisation, as well as performance, metrics, and assumptions can assist the model interpretation and implications of the analysis performed. One might consider the narrative related to a particular dataset and how it is framed (62), key information that is depicted and what is recalled (63), different forms of story and degree of freedom for interpretation (64), design steps (65), the influence of individuals background on the interpretation of results (66), and the influence of spatial patterns on interpretation (67).

Typically, data pipelines are iterative consisted of multiple cycles. Some checklists commonly used to ensure study quality, proper presentation and validation are TRIPOD (68), QUIPS (69), CONSORT (70), Prisma (71) and SPIRIT-AI (72). These checklists support a better analysing and reporting of data studies. Moreover, Banerjee et al. 2021 (73) provide a set of questions for the interpreting and assessment of machine learning studies.

2.3 Data sources

This thesis explores a number of diverse datasets that were selected based on a number of criteria. First, the data needs to be available – although there is a plethora of clinical data, these are typically bound to strict rules of use. One criterion is to be able to contrast **study data** with **real-world data** (section 1.1) – while study data provide an angle on novel targets, the use of real-world data provide an invaluable source of knowledge, with an efficient re-use of data. Another criterion is to be able to explore different modalities and scales of patient data – not all datasets contain the same breadth and depth of information, and these different factors help explain different biological mechanisms (sections 1.1.1 and 1.1.2, and Chapter 6). There were five main datasets used throughout this thesis. The Birmingham Black Country Atrial

Fibrillation registry (BBCAF) is an exemplar of study-data, it contains a good breadth and depth of data, with information about novel biomarkers, electrocardiogram, and omics for some participants. The Characterizing Atrial fibrillation by Translating its Causes into Health Modifiers in the Elderly (CATCH-ME) dataset provides similar data to the BBCAF dataset, the only difference is a wider inclusion scope, i.e., while BBCAF contains participants from the UK, CATCH-ME encapsulates BBCAF and includes participants from different centres in Europe. The UK Biobank contains a large number of participants, with over half a million participants, it contains a depth of information that enables population analysis. University Hospitals Birmingham (UHB) and The Health Improvement Network (THIN) are samples of real-world data, the first with information from secondary and tertiary care, the latter with primary care. The THIN dataset utilises read codes collected from normal practice, and contain a limited number of variables and information for all the participants, however, the information available and number of participants make them an ideal dataset for retrospective analysis on population and morbidity. A summary of these datasets, including their uses is provided in the Table 3.

Table 3: Summary of the different data sources used.

Dataset	Short description	Applications on this thesis
Birmingham Black Country Atrial Fibrillation registry (BBC-AF)	It comes from an AF study performed in the Black Country region. Patients had AF and/or risk factors. The patients had a follow-up and the measurement of multiple biomarkers that were used to assess patient risk.	Analysis of patient risk of developing AF using biomarkers, the influence of socioeconomic factors to predict patient risk, and some investigations using ECGs.
Characterizing Atrial fibrillation by Translating its Causes into Health Modifiers in the Elderly (CATCH-ME)	Dataset formed of the junction of different study datasets, it is formed of the BBCAF dataset and other studies in Europe. It contains a wide range of variables and populations.	This dataset was used to create a model that was validated in this thesis.
UK Biobank	A comprehensive biomedical dataset. It contains over half a million people with a wide range of variables, with comprehensive information about the patient clinical journey.	The main reference for GWAS analyses, population analyses in scale, evaluation of phenotypes, and the data collection framework.
University Hospitals Birmingham (UHB)	Dataset used in the local hospital. It contains real-world data, with variables used in routine care.	This dataset was used for population analyses, creation of patient phenotypes reference, ECGs to predict heart failure study, and the data collection framework.
The Health Improvement Network (THIN)	Primary care dataset containing over ten million data samples from all over England.	This dataset was used for population analysis.

2.3.1 Birmingham Black Country Atrial Fibrillation registry (BBCAF)

This cohort comes from the homonymous study aimed to identify important biomarkers for the prediction of AF. The study in this dataset was approved by the National

Research Ethics Service Committee (BBC-AF Registry, West Midlands, UK, IRAS ID 97753).

BBCAF recruited patients from in-patient or out-patient visits between September 2014 and February 2018 at the Sandwell and West Birmingham NHS Trust (Birmingham, UK). Patients were included if they had atrial fibrillation, or if they had at least 2 CHA_2DS_2-VASc stroke risk factors (74). There were 1600 patients recruited. All patients had a follow-up at 2 years. For patients without diagnosed AF, the patients had a 7-day ECG recording for the assessment of silent AF.

The main dataset contains information about patients such as birth date, sex, ethnicity, weight, height, and postcode. The dataset also includes other information such as recruitment date and risk factors identified during the recruitment visit: advanced age, prior stroke or TIA, arterial hypertension, diabetes mellitus, severe coronary artery disease, stable heart failure, left ventricular hypertrophy and peripheral artery disease. Overall information about AF, including type, frequency, and time since the first occurrence. History of AF treatment, including information if the patient had electrical or chemical cardioversion, catheter ablation or another surgical treatment of AF. Information about a wide range of severe or cardiovascular-related conditions: smoking history, cardiomyopathy, history of syncope, palpitations, unexplained dizzy spells, history of resuscitation, presence of a pacemaker, chronic obstructive lung disease, malignant diseases, hypothyroidism, hyperthyroidism, heart failure, and suspected presence of an acute coronary syndrome. Blood pressure measurements in systolic and diastolic blood pressure. Electrocardiogram recording with measured intervals such as QRS, PQ, QT and QTc, information about other indications from the recording, such as bundle branch block. Echocardiogram recordings, inclusive of measurement of LVEF. Basic blood biochemistry, including haemoglobin, white blood cell count, platelets, alanine aminotransferase (ALAT/GPT), creatinine, glomerular filtration rate, international normalized ratio, prothrombin time, activated partial thromboplastin time. Biomarkers obtained from 1 µL EDTA plasma assessed using Olink Proteomics, (Uppsala, Sweden) cardiovascular panels I and II (75) – each cardiovascular panel assesses 92 protein expression related to different biological processes and disease areas - there are 40 common biomarkers between the different

Olink Cardiovascular panels that were analysed (Appendix 2.1). The dataset also contains the medication history inclusive of inhaled bronchiolitis and steroids, systemic steroids, therapy for chronic lung disease or any other therapy as well as the Montreal Cognitive Assessment for cognitive function (76).

Olink panel cohort. The initial biomarker dataset contains a subset of patients (N = 638). The dataset for analysis with the Olink panels included 7 clinical risk factors: age, sex, hypertension, heart failure, history of stroke or transient ischaemic attack, kidney function and body mass index. The blood biomarkers were measured using Olink cardiovascular panels I and II (75). Sequential participants were separated into a discovery set with 384 patients and the remaining 254 patients into the validation set. These datasets patients had also been assessed using the first and second cardiovascular panel, respectively. The first study describing this dataset is Chua et al. (77). Table 4 describes this dataset.

Table 4: Description of the BBCAF dataset. Categorical variables are reported as n (%), whereas continuous variables are reported as mean (standard deviation) [or median (IQR) for non-parametric distributions]. The independent t-test (or Mann–Whitney U test for non-parametric distributions) and X^2 tests were used to compare continuous and categorical characteristics between patients within the two cohorts. ACEi, angiotensin-converting enzyme inhibitor; BMI, body mass index; CAD, coronary artery disease; eGFR, estimated glomerular filtration rate; NOAC, non-vitamin K antagonist oral anticoagulant; VKA, vitamin K antagonist. [a] Non-parametric distributions. [b] A two-tailed significant difference $P < 0.05$ between patients with and without AF. Table adapted from Chua et al. (77).

	Discovery		Validation	
	No AF (N=215)	**AF (N=169)**	**No AF (N=129)**	**AF (N=125)**
Age (years)	66.0 (57.0–74.0)	73.0 (63.0–79.0)[b]	67.0 (59.1–74.0)	75.0 (67.0–81.5)
Male	130 (60.5)	117 (69.2)	83.0 (64.3)	68.0 (54.4)
Ethnicity				
Caucasian	133.0 (61.9)	142.0 (84.0)[b]	104.0 (80.6)	116.0 (92.8)[b]
Asian	55.0 (25.6)	14.0 (8.3)	13.0 (10.1)	5.0 (4.0)
Afro-Caribbean	25.0 (11.6)	9.0 (5.3)	12.0 (9.3)	4.0 (3.2)
Unknown	2.0 (0.9)	4.0 (2.4)	—	—
BMI (kg/m^2)[a]	28.1 (25.2–32.7)	29.6 (26.0–33.6)	29.1 (25.5–33.4)	28.9 (24.8–32.9)
eGFR ($mL/min/1.73\ m^2$)[a]	72.0 (57.0–87.0)	69.0 (57.5–84.0)	73.0 (58.3–85.0)	64.0 (44.5–79.0)
Diabetes	89.0 (41.4)	37.0 (21.9)[b]	56.0 (43.4)	26.0 (20.8)[b]
Stroke	24.0 (11.2)	21.0 (12.4)	13.0 (10.1)	10.0 (8.0)
CAD	87.0 (40.5)	29.0 (17.2)[b]	78.0 (60.5)	29.0 (23.2)[b]
Hypertension	142.0 (66.0)	104.0 (61.5)	89.0 (69.0)	61.0 (48.8)
Heart failure	31.0 (14.4)	28.0 (16.6)	8.0 (6.2)	12.0 (9.6)
Ejection fraction (%)[a]	60.0 (53.1–67.3)	57.7 (45.0–65.0)[b]	57.0 (45.5–62.5)	55.0 (41.3–61.0)
Admission criteria				
Inpatient	160 (41.6)	97 (25.3)	124 (48.8)	97 (38.2)
Outpatient	55 (14.3)	72 (18.8)	5 (2.0)	28 (11.0)
Concomitant medication				
NOAC	4.0 (1.9)	63.0 (37.3)[b]	1.0 (0.8)	44.0 (35.2)[b]
VKA	5.0 (2.3)	48.0 (28.4)[b]	2.0 (1.6)	41.0 (32.8)[b]
Aspirin	137.0 (63.7)	39.0 (23.1)[b]	98.0 (76.0)	41.0 (32.8)[b]
Antiplatelet agents (clopidogrel, prasugrel, and ticagrelor)	94.0 (43.7)	33.0 (19.5)[b]	82.0 (63.6)	27.0 (21.6)[b]
ACEi	44.0 (20.5)	36.0 (21.3)	58.0 (45.0)	37.0 (29.6)[b]
Angiotensin II receptor blocker	39.0 (18.1)	28.0 (16.6)	22.0 (17.1)	25.0 (20.0)
Beta-blocker	115.0 (53.5)	83.0 (49.1)	85.0 (65.9)	72.0 (57.6)
Diuretic	59.0 (27.4)	66.0 (39.1)[b]	37.0 (28.7)	55.0 (44.0)[b]
Calcium channel antagonist	61.0 (28.4)	42.0 (24.9)	39.0 (30.2)	24.0 (19.2)[b]
Cardiac glycoside	—	33.0 (19.5)[b]	—	28.0 (22.4)[b]
Aldosterone antagonist	13.0 (6.0)	12.0 (7.1)	5.0 (3.9)	10.0 (8.0)
Verapamil/diltiazem	12.0 (5.6)	14.0 (8.3)	5.0 (43.9)	7.0 (5.6)
Antiarrhythmics (amiodarone, dronedarone, flecainide, propafenone, and sotalol)	4.0 (1.9)	17.0 (10.1)[b]	3.0 (2.3)	12.0 (9.6)[b]

ECG cohort. This compiled dataset contains only patients with complete labelling information and signal data. This dataset also contains 638 patients, patients which had their ECG recording in sinus rhythm or during an AF episode, 493 and 145 patients, respectively, while other rhythms were excluded from the analysis. The ECG recording contains 12-lead data at 500Hz over 10 seconds. Patient characteristics are similar to the first consolidated version. The distribution of ECG rhythms is shown in Figure 7.

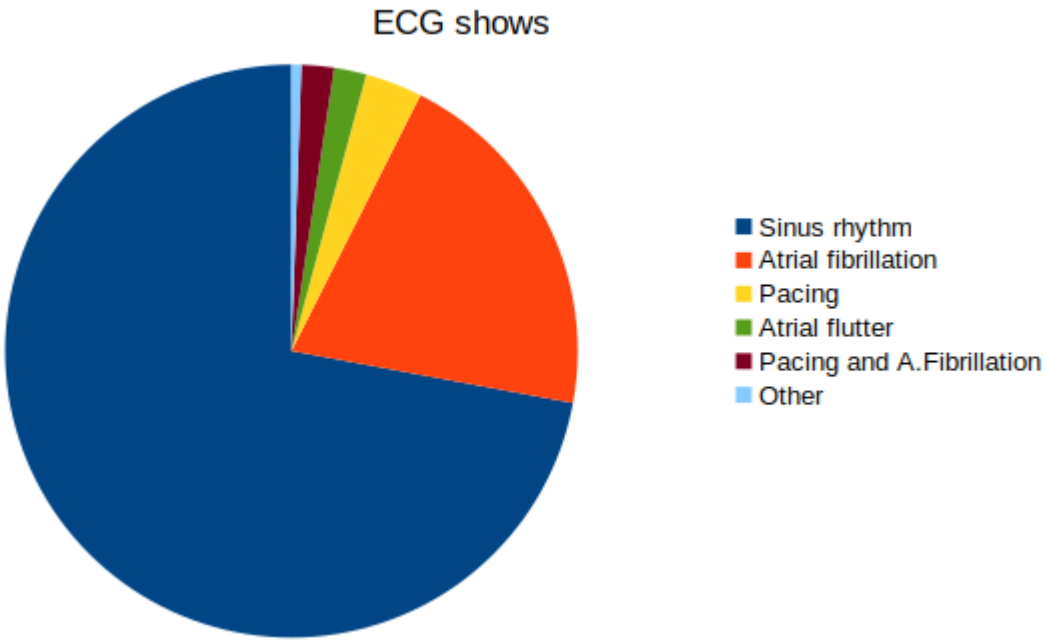


Figure 7: Distribution of ECG rhythm on the BBCAF dataset. Most patients that had a 10s ECG recording had it whilst in sinus rhythm. Some of these patients reported as sinus rhythm have silent AF identified through the holder ECG recording.

Roche cohort. A part of the dataset (N=1485) included 12 blood biomarkers quantified by Roche Diagnostics (Mannheim, Germany) using Elecsys® immunoassays, these are cancer antigen 125 (CA125), growth differentiation factor-15 (GDF15), Interleukin-6 (IL6), N-terminal pro B-type natriuretic peptide (NTproBNP), cardiac Troponin T (TnT), cardiac C-reactive protein (CRP), angiopoietin (ANG2), bone morphogenetic protein 10 (BMP10), endothelial cell-specific molecule-1 (ESM1), fatty acid binding protein 3 (FABP3), fibroblast growth factor 23 (FGF23), and insulin-like growth factor

binding protein 7 (IGFBP7). A description of the discovery dataset is shown in Table 5.

Table 5: Description of the BBCAF Roche biomarkers dataset. Numerical values with a single number in parenthesis indicate mean (standard deviation), other values are non-parametric, and the range is for the 1st and 3rd quartile. Table adapted from the submitted study.

	No AF (N=522)	**AF (N=411)**
Age, years	67 (58-75)	74 (66-80)
Sex, males	309 (59%)	256 (62%)
Ethnicity		
Caucasian	362 (69%)	350 (85%)
Asian	112 (22%)	31 (8%)
Afro-Caribbean	48 (9%)	30 (7%)
BMI, kg/m²	28.7 (25.5-32.5)	29.0 (25.1-33.1)
eGFR, mL/min/1.73 m²	71.7 (26.1)	67.8 (25.9)
Diabetes	238 (46%)	96 (23%)
Stroke/TIA	46 (9%)	38 (9%)
Coronary artery disease	252 (48%)	93 (23%)
Hypertension	333 (64%)	218 (53%)
Heart failure	222 (43%)	219 (53%)
Inpatient admission	469 (90%)	293 (71%)
Biomarkers		
ANG2 (ng/mL)	2.36 (1.73-3.45)	3.64 (2.28-6.14)
BMP10 (ng/mL)	1.95 (1.70-2.32)	2.35 (1.94-2.94)
CRP (mg/L)	4.95 (1.63-18.89)	4.19 (1.57-15.59)
CA125 (per 10 U/mL)	1.23 (0.82-2.01)	1.57 (0.95-3.40)
ESM1 (ng/mL)	2.01 (1.47-2.91)	2.36 (1.78-3.43)
FGF23 (per 100 pg/mL)	1.65 (1.05-2.69)	1.97 (1.35-4.16)
FABP3 (per 10 ng/mL)	3.53 (2.63-5.19)	3.77 (2.82-5.92)
GDF15 (per 100 pg/mL)	18.71 (11.42-31.08)	21.29 (13.41-35.22)
IGFBP7 (ng/mL)	96.23 (82.74-115.30)	110.17 (91.65-140.09)
IL6 (pg/mL)	6.38 (3.31-14.66)	6.49 (3.37-14.69)
NTproBNP (per 100 pg/mL)	4.21 (1.08-14.34)	11.20 (3.51-28.61)
TnT (per 100 pg/mL)	0.30 (0.12-1.09)	0.22 (0.12-0.50)

BBCAF is contained in the broader Characterizing Atrial fibrillation by Translating its Causes into Health Modifiers in the Elderly (CATCH-ME) dataset (78).

2.3.2 Characterizing Atrial fibrillation by Translating its Causes into Health Modifiers in the Elderly (CATCH-ME)

This incorporates data from different research centres and projects in partnership under the European Union's Horizon 2020 research and innovation programme. This

includes international trials, such as the Flecainide short-term versus long-term study, and the local BBCAF study (79) (80) (77).

Due to the different dataset contexts that incorporate this consort study, different types of information under different settings were obtained for the study. The data collected is sparse, there are many columns and many missing variables. All studies collected patient information and some clinical history data. ECG and echocardiograms were partially or fully collected. Some studies collected blood material, whilst others also had tissue samples. The complete dataset contains 14494 records.

The dataset contains basic information about a patient: age, sex, weight, height, ethnicity, original study origin. Diastolic and systolic blood pressure on recruitment. Medical history: type, frequency, and time since the first occurrence of AF; presence of a pacemaker, history of cardioversion, catheter ablation or any surgical procedure and mention of post-operative AF; presence of diagnosed heart failure, diastolic dysfunction, hypertension, stroke, TIA, history of any bleeding, valvular heart disease, diabetes, myocardial infarction, chronic obstructive pulmonary disease, sleep apnoea, chronic kidney disease, hyperthyroidism, hypothyroidism, rheumatic heart disease, pulmonary hypertension, and allergies. History of hospitalisation for any cardiovascular reason. Family history of AF or other cardiovascular conditions. Physical activity, smoking, alcohol consumption, and drug abuse history. ECG measurements, inclusive of intervals and digital recording. Echocardiogram measurements. Laboratory tests, inclusive of international normalized ratio: D-Dimer, Creatinine, Troponin I, Troponin T, glucose, NTproBNP, total cholesterol level, LDL, HDL, and Triglyceride level. Medication history, inclusive but not limited to P2Y12 blockers, beta-blockers, Calcium ion antagonists and aldosterone antagonists.

This dataset also contains some follow-up data: death, AF and other cardiovascular events, medication changes and other laboratory measurements.

2.3.3 UK Biobank

UK Biobank dataset (81). The UK Biobank dataset contains information about 502,489 participants recruited in the United Kingdom aged 40-69 years. The average

recruitment age is 56 years. The study recruited participants between 2007 and 2010. This study collected (and up to 2021 still collects) extensive health-related data about its participants.

The information collected contains baseline questionnaires, assessing sociodemographic, family history, psychosocial, environmental, lifestyle, cognitive function, health status, and food frequency. Physical measurements. Electrocardiogram and magnetic resonance imaging (MRI). Patient primary and secondary care diagnosis and operations dates. Biological measures of blood, urine, and saliva (82). More information about available data is shown on the UK Biobank Data Showcase website (83).

Types of variables used and applied include demographics, ICD-10 diagnoses (84), ICD-9 diagnoses (85), self-reported conditions, medication history, cognitive and laboratory tests, (MRI)-derived values, electrocardiogram recordings and genotyping.

Coded diagnosis in the UK Biobank contains in-patient data from 1997 onward, outpatient data from 2003 onwards and other accident and emergency data from 2008 onwards

CATCH-ME validation cohort. The validation cohort used for the model described in section 3.6 required a few fine tunings. The main variables used were the patient's age, sex, height and weight, blood pressure, history of hypertension, AF, diabetes, tricuspid valve disease, and myocardial infarction, medications such as P2Y2 blockers, beta-blockers and aldosterone antagonists, signs of abnormality in ECG, blood pressure and blood sugar (HBA1c) levels, left atrial volume, left ventricular end-systolic diameter, and diastolic septal wall thickness were included as variables.

Variables that indicate areas and volumes of the heart come from the UK Biobank return dataset 1362, and it is limited with the number of MRI available when it was created. The left ventricular end-systolic diameter is not available, and only the respective volume was available. A transformation between the left ventricular end-systolic volume (LVESV) and the left ventricular end-systolic diameter (LVESD) was done using the Teichholz formula (86):

$$LVESV = \frac{7.0}{(2.4 + LVESD)} LVESD^3$$

The application of the model required complete-cases data (restricted to participants with no missing values). Due to the limitation of cardiovascular measurements from MRI, the number of patients was reduced. In total, there were 27 patients with and 4137 without AF on the validation dataset. A flowchart with the number of patients is shown in Figure 8. Table 6 displays the baseline characteristics and the range of values.

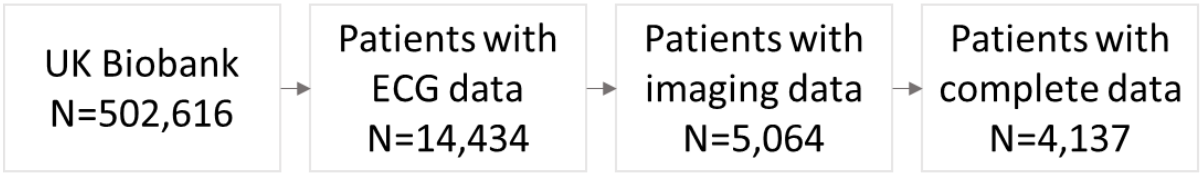


Figure 8: Flowchart of UK Biobank patients.

Table 6: Baseline characteristics in the UK Biobank for the CATCH-ME validation. The samples are separated by patients with (AF) and without (No-AF) atrial fibrillation. ***** (a) Continuous variables with a normal distribution are summarized as mean (standard deviation), (b) Continuous variables which were not normally distributed are summarized as median (IQR), and (c) Categorical variables are reported as the number of cases (%).

Predictor	*****	**Total (n=4137)**	**AF (n=27)**	**No-AF (n=4110)**	**p value**
Age (Years)	b	56 (12)	62 (8)	56 (12)	<0.001
Gender (Female)	c	2181 (52.7)	6 (22.2)	2175 (53.0)	0.003
BMI (kg/m²)	b	26.2 (5.4)	28.3 (4.7)	26.2 (5.4)	0.019
Height (cm)	a	169.4 (4.3)	177 (7.6)	169.3 (9.2)	<0.001
Hypertension	c	1193 (28.8)	16 (59.3)	1177(28.6)	0.001
Diastolic Blood Pressure (mmHg)	a	81.3 (9.9)	82.4 (10.1)	81.3 (9.9)	0.59
Systolic Blood Pressure (mmHg)	a	135.3 (17.7)	140.8 (20.1)	135.3 (17.7)	0.168
HbA1c (mmol/mol)	b	34.7 (4.8)	36 (7)	34.7 (4.8)	0.192
Diabetes	c	188 (4.5)	4 (14.8)	184 (4.5)	0.035
CABG	c	33 (0.8)	2 (7.4)	31(0.8)	0.005
Myocardial Infarction	c	48 (1.2)	2 (7.4)	46 (1.1)	0.032
Tricuspid Valve Disease	c	0 (0)	0 (0)	0(0)	-
Coronary Artery Disease	c	87 (2.1)	2 (7.4)	85 (2.1)	0.21
Left Atrial Volume (mm³)	b	65 (26.35)	77.0 (32.93)	65.0 (26.45)	0.005
Left Ventricular End Systolic Diameter(mm)	b	28.5 (5.83)	29.48 (5.98)	28.50 (5.84)	0.066
ECG Parameters					
Signs of old infarction on ECG	c	394 (9.5)	5 (18.5)	389 (9.5)	0.205
Signs of acute ischemia on ECG	c	167 (4.0)	3 (11.1)	164 (4.0)	0.167
Left Ventricular Hypertrophy	c	151 (3.6)	0 (0)	151 (3.7)	0.617
Medication					
Aldosterone-antagonists	c	8 (0.2)	0 (0)	8 (0.2)	1
Beta-blockers	c	313 (7.6)	12 (44.4)	301 (7.3)	<0.001
P2Y12 blockers	c	64 (1.6)	1 (3.7)	63 (1.5)	0.898

2.3.4 University Hospitals Birmingham (UHB)

University Hospitals Birmingham NHS Foundation Trust (UHB) dataset is an aggregation of different data sources from within the hospital closely associated with the University of Birmingham. UHB contains data from Queen Elizabeth, Good Hope, Heartlands, and Solihull hospitals (87). The use of this data is approved on the Health Research Authority Research Ethics Committee reference 20/PR/0659.

The dataset is formed of different products sub-systems that integrate the operational infrastructure for healthcare service. The two main products datasets utilised are the

Birmingham Systems Prescribing Information and Communication System (PICS) and the Solus Cardiology Solution (21) (88). The local copy of the Hospital Episode Statistics (HES) contains compiled information about in-patients.

PICS contains the most information about a patient. It contains consolidated data regarding the patient's registry: sociodemographic, exams, laboratory tests, diagnoses, symptoms, medications, hospitalisation data, alerts, and actions by the clinical staff. All the data is associated with event-time.

Solus Cardiology contains information from the cardiology side, the data explored is inclusive of echocardiogram reports and ECG recordings. The ECG recordings have 10s recordings, stored in a compacted format such as XML or PDF format (89) (90). Due to technical differences of the PDF dataset, such as limited precision on the stored data and complexity of data extraction, only the XML files were utilised.

HES dataset contains operations and diagnoses information associated with in-patient visits from November 2014 to May 2018. It contains multiple ICD-10s associated with each patient episode. This is inclusive of basic sociodemographic information, episode start and end date, and the main reason for admission.

Both HES and PICS diagnoses are recorded as ICD-10 (84), with some terms from the NHS Classification ICD-10s, available on the NHS Digital website (91). The operation codes are reported as OPCS-4 codes (92). The average age on the system is 54 years.

In this thesis, the UHB dataset with its subparts is the closest example of a real-world dataset. It is composed of data that is routinely collected, without manual filtering and selection from a study perspective.

2.3.5 The Health Improvement Network (THIN)

The Health Improvement Network (THIN) is a database containing information from primary care in the UK. It contains coded information from over 10 million participants in over 500 general practitioner clinics (93).

This dataset provides a good representation of the UK population care journey, it provides a comprehensive description of a patient's use of medical resources, containing information from the administrative side, such as registration and change of clinic, clinical tests, dates of each data collection, and extensive use of read codes to describe the different patient events (94).

2.4 Data analysis

2.4.1 Discovery and validation datasets

After identifying the dataset and collecting the data, it is imperative to validate the results obtained from a dataset into another. This separation will avoid bias on created models. This is a similar rationale to having one study identifying a new factor, and then further studies checking the real behaviour in other data. In this case, the experimental design is already separating the dataset into discovery and validation of the findings, making the results more robust and reliable. It also helps when applying methods that tend to over-fit on the discovery data, such as the case of models with a large number of parameters, such as neural networks.

The dataset is usually separated into two subsets: the training or discovery dataset and the testing dataset (sometimes called validation). The former dataset will be used to create the model. The latter will be used to obtain the final results after selecting the best model.

There are cases where the training process may produce different model variants. To identify the best performing model on the training dataset it is possible to further split the training set into another training and an internal validation set. In no instance, any set or part of it should be combined with the training or internal validation sets. Any overlap of datasets will overestimate the performance of created models.

In scenarios where there are multiple data points per real (biological) sample, it is strongly suggested to keep the biological samples in the same dataset. This will depend on the data type and source, but usually keeping the related data together will avoid creating a model that learns individualised data points. For example, a dataset with multiple ECG recordings for each patient must have a clear separation of patients

into different datasets, avoiding models to learn how to identify specific patients who have or not a condition.

There is no specific rule about how a dataset should be split. In some cases, the dataset might be split depending on the time of record, data source or technology used for the measurements. For example, in a multi-centre study, the data points could be split depending on the centre county. As another example, blood tests that were measured using a specific version of a microarray could be separated into the training and validation sets and the newer version of the assay as the test set. In most scenarios, random sampling without replacement, placing the elements randomly into the different sets, is enough.

It is possible to create models using subsections of the dataset: K-fold cross-validation separates the set into K parts which are then used to created K models using K-1 parts as training data and 1 part as validation, going through all the parts to identify the best performing parameters. Leave-one-out is a variation of the K-fold in which there is one part for each data point. This approach is commonly used in cases where a limited number of data points is available.

In other cases of limited data points, a bootstrapping approach can be used. Bootstrapping utilizes random subsets of the dataset, sampled with replacement (95). These samples are executed with similar methods and the overall results are compiled. Bootstrapping repetitions utilises statistical power to provide model robustness.

In either case of data splitting, it is recommended to keep the same distribution of values, especially the ratio of cases and controls. Furthermore, it is a good idea to compare the performance of the training and validation sets before going to the test set. If the training set gets a much higher performance than the validation set it means the models are over-fitting – the model has not learned but is instead memorizing the data points. If either the validation set is performing considerably better, or both the measurements are unsatisfactory it indicates that the model is under-fitting – it is not learning the model and is rather tending to a mean output. For the first problem, obtaining more samples may correct the problem, whilst for the former problem increasing the number of variables is an option.

2.4.2 Unbalanced datasets

There are some cases where it is not possible to maintain a balance on the whole data or part of it. These are unbalanced datasets, and they do not contain a symmetric distribution of values, for a dataset and contain one predominant value type. For example, a patient dataset has for every 4 healthy patients, 1 patient with AF – this is an example of a dataset with 80% of participants in the largest group, the healthy group, also called the majority class. These datasets tend to be biased towards the majority class. Different reviews describe methods and their applications for handling class imbalance problems (96) (97).

An unbalanced dataset can be corrected through a resampling process, such as up-sampling, down-sampling, and both. These processes will provide a better balancing of the data. This process should only be done on the training set and with care to the interpretation of intermediary results because of the new false distribution, duplicate points or the differentially reduced number of points.

The method Random OverSampling Examples (ROSE) generates new samples in the neighbourhood of elements from the minority class (98, 99). Another approach used is Synthetic Minority Oversampling TEchnique (SMOTE), which generates new samples based on a linear combination of a reference sample and a sample near it (100).

Metrics such as Area Under the Receiver Operating Characteristics Curve (AUCROC) and Precision-Sensitivity curves assist to evaluate the model biases, also provide support on selecting a threshold cut-off to select the model.

2.4.3 Structured datasets datatypes

When progressing into an exploratory and analytical step of data analysis, the use of datatype refers to what type of content a structured data column has. There are two main data types: numerical (continuous) and categorical variables (discrete variables).

Continuous variables are quantitative measurements that have a number indicating its intensity and a unit indicating its meaning and magnitude. For example, a normal glucose level can be assessed as either 5.5 mmol/L or 100 mg/dL. The use of variables

in different units will hinder the interpretation of any analysis. Whilst a variable on a wider scale with more extreme values has a smaller coefficient, a variable with smaller values will have a higher coefficient to compensate. To compensate for unit differences a commonly employed approach is to centre each variable to its mean and to scale all variables to the standard deviation of the population (Equation 2.4.3.1, where X is the variable, μ the mean and σ the standard deviation.).

$$X = \frac{X - \mu}{\sigma}$$

(2.4.3.1)

Another transformation that is commonly applied is the minimum-maximum scaler. It obtains the minimum and maximum value for a variable and transforms these variables into a new range, usually 0 to 1.

Continuous variables can be discretized into different groups, such as transforming the continuous range into normal, abnormally high, or abnormally low values. The 95% confidence interval range may be employed for this task. Quartile intervals or other reference values from the literature are other alternatives.

Categorical variables are either assigned numbers of different groups or names for these groups that indicate a qualitative value that is not expected to have quantifiable steps of increased intensity. For example, in the evolution of AF, a patient goes from a normal rhythm to paroxysmal, persistent, and then permanent AF. The different types of AF cannot be interpreted in any enumerated continuous scale, as the progression into another stage of the disease is either existent or not.

In the opposite case to the discretization transformation, sometimes it is compulsory to use numerical values rather than categorical ones. In this case, new attributes are created, one for each possible category of the original column. For each new attribute, a value of 1 is assigned if the original sample had this new column category whilst a value of 0 or -1 when not (this is usually referred to as one-hot or dummy encoding) (101).

Different variable types imply different types of knowledge. In some cases, it might be easier to think if a value is inside or outside a normality range rather than its real value. Similarly, it may be an option to think of a positive case as the numerical case 1 whilst the negative case as a 0 or -1 value.

The selection of the data type is dependent on the analysis goals, data and algorithms employed. For the case of logistic regression, the target variable must be 0 or 1 (due to the logit function). The independent variables may have any scale, however, if they are on a similar scale they can be more directly compared. In most analysis, the categorical variables must be transformed into numerical features. Machine learning analysis requires continuous variables to be centred and scaled.

The distribution of the data points for each variable is paramount to the different statistical analysis that may be performed. Numerical variables can be in a normal or non-normal distribution. Either the values in a variable are distributed in a normal distribution and statistical tests that depend on this assumption are used, or different non-parametric tests are performed. Categorical variables can have a varying number of possibilities: if there are only two options, the variable is binary, such as the patient has or has not a disease; a variable can have multiple options with very differing proportions. When working with categorical variables the majority class has proportion dominance over the others, while the minority class contains the lowest proportion.

An unbalanced variable has values tending to go into a range or a category of values. For the numerical case, whilst most patients are in a close range, there might be outliers in either extreme of the distribution.

From these different distributions or range of values, many metrics are used to compare groups of samples. For brevity reasons, some of these metrics will only be cited: count (number of samples), minimum value, maximum value, average, standard deviation, median, mode (most frequent value), quartiles, confidence interval, ratios and percentage. These functions are more adequately applied when describing different types of distribution of values, e.g., a continuous variable with a gaussian distribution can be described using the mean and the standard deviation; a categorical

or binary variable may be described using the percentage of samples in one or each of its options.

Depending on the distribution of the variables different reporting protocols are considered. The distribution of a variable can be visually assessed or inspected through a statistical test such as Shapiro-Wilk or Anderson-Darling (102, 103). The default reporting of variables includes mean value and standard deviation for normal-distributed values; in the case of nonparametric distributions, it is usually reported the median value and the 1st and 3rd quartile; for categorical variables, the percentage of each class, or the majority class for a binary variable, is usually shown. These variables will provide some information on the distribution of values. Visualisations, such as scatter plot, bar plot, box plot, violin plot and histograms assist in the understanding of the variable, avoiding biases from only using descriptive values (104).

2.4.4 Transformation

A dataset might contain too many variables to have it analysed directly. Transformations offer ways to compare the importance of factors directly.

Principal Component Analysis (PCA) is a transformation that extracts axes called Principal Components (PCs), ordered by decreasing variance from a dataset of continuous variables (105). Normally, it is used to extract axis from datasets with hundreds to thousands of columns to utilize two axes that explain the most variance – PC1 and PC2 – so that the dataset can be plotted in a 2D representation. This transformation can be particularly useful when dealing with gene expression datasets that normally contain thousands of values. PCA can also be used to find underlying patterns in the dataset, for example, in RNA-Seq to group samples and GWAS to group ancestry groups (58). PCA transformed variables can be mapped back to identify variables that explain the most variance in the dataset.

Other well-known transformations of variables, focused on visualisation are the t-Distributed Stochastic Neighbor Embedding (t-SNE) and the Uniform Manifold Approximation and Projection (UMAP) (106) (107). These approaches utilise Neural

Network to learning representations of the data points, further discussed in section 2.5.5.

2.4.5 Missing values

Missing values are common in any dataset type. Missing values are originated due to three major reasons: (a) Missing Not At Random (MNAR) is the case when the value is missing because of a condition that is not explicitly explained in the variables, e.g. a person did not give the right answer to its financial question because the person was not willing to reveal its financial status; (b) If the value is missing because a datasheet was lost it is Missing Completely At Random (MCAR); (c) In the case that the missingness of data is dependent on other variables it is called Missing At Random (MAR) (108).

There are three main approaches to deal with missing values in datasets. The first approach is to ignore the columns or rows with large amounts of missing values or replace them with an actual value, such as mean/median value or imputing the value using models. A second approach commonly employed is the imputation using Multivariate Imputation by Chained Equations (MICE) which suffices for most requirements (109, 110). A third possibility is to leave the information that the value was missing as another column while imputing the value with one of the approaches so that the information of missing value is kept as in the MAR case (111). In this work, the first and second approaches were used.

2.4.6 Null hypothesis

In statistics, the null hypothesis is the assumption that there is no association between two events (112). Rejecting the null hypothesis indicates that there is an association, and we confirm our alternative hypothesis. This assures that the probability of the null hypothesis being true is low, thus the new theory is considered valid. However, one should pay attention to other errors that may occur: a type I error is the probability of rejecting the null hypothesis when it is true, and a type II error is the probability of accepting a null hypothesis when the alternative hypothesis is true.

One way of evaluating a hypothesis is by comparing two groups. For example, a group of mice with a mutated gene (case), and another group without a mutated gene (control) have their expression levels of multiple genes is compared to validate that a gene is differentially expressed between them – the null hypothesis is that there is no difference in the mutated one. After exploration, it may be identified to be false, the alternative hypothesis is indeed true.

The threshold used for significance may be different ones and has led to many discussions (113). In the medical literature, the probability value (p-value) of 5% is a commonly employed threshold to validate a hypothesis type I error (114).

Tests executed depend on the distribution of the data points. When comparing categorical variables, the Fisher exact test may be performed, and for a large number of samples, a fast Pearson chi-squared test may be used instead, as it provides similar results (115, 116). To evaluate if a distribution is normal, it can be assessed using Shapiro-Wilk, Kolmogorov-Smirnov or Anderson-Darling tests (117). For normal distributions, Student's T-test indicates if there is a difference (118). When there is a nonparametric distribution Mann-Whitney U or Wilcoxon signed-rank test can be performed (119, 120). There are also other tests for specific scenarios: analysis of variance (ANOVA) is a test for comparing 2 or more groups of variables (121). These approaches were used throughout the thesis to describe dataset variables.

2.4.7 Logistic regression

A model has to be created when it is required to understand the behaviour of a dependent variable over some independent variables. A model is the result of data, algorithm, and learning. This section describes the process of creating a model using logistic regression.

To model multiple independent variables, a regression can be used. Regression models employ a mathematical function as follows:

$$f = a_o + a_1 x_1 + \cdots + a_n x_n \quad (2.4.7.1)$$

Where the a_i describe the coefficients of the different variables *i* and the x_i describe the variables. The a_0 coefficient is the constant term or intercept. The coefficients of the model can be optimized using different approaches, such as least squares or gradient descent (122) (123).

To predict the risk score of having a disease we will need to obtain a range of values from 0 to 1 (0% to 100% chance). Given that *f* may assign "infinitely" negative or positive values, it is required to transform these values to a 0 to 1 range. A frequently employed transformation function is the logistic function, which is defined as:

$$Logistic(x) = \frac{e^x}{e^x + 1} = \frac{1}{1 + e^{-x}}$$
(2.4.7.2)

The final probabilistic model with multiple independent variables utilizing the logistic function is then as follows:

$$p(x) = \frac{1}{1 + e^{-(a_o + a_1 x_1 + \cdots + a_n x_n)}}$$
(2.4.7.3)

The model created up to now will be able to give a score to each data point, or patient, given some variables. To evaluate the importance of each variable an odds-ratio is utilized. The odds-ratio indicates the influence that a variable has on the model and its absolute number may be used to indicate the most important variables.

$$OR_i = e^{-a_i}$$
(2.4.7.4)

Formula (1.4) above describes the odds ratio for the i^{th} variable. The odds ratio is normally used with probabilistic testing, such as the 95% confidence interval, to validate that the variable contributes to the model if its mean does not cross the 1 value. If the odds-ratio is 1 it means that the variable has no importance at all, if the value is above 1 it means that the variable leads to the positive case, while below 1 it contributes to the negative case. A significance value is also normally assessed.

The odds ratio is a good metric to evaluate variable importance when creating linear models. There are many other metrics to evaluate model performance and different factors over the variables. Mainly focusing on binary problems (case-control), different metrics are described in the following section. For the scenario of multi-class outcomes, the comparison may be done one primary class against all or all against all.

2.4.8 Confusion matrix

For binary problems, the True Positive (TP or hit) indicates the number of predictions that were correctly predicted as being in the positive class. True Negative (TN or correct rejection) follows as the correct predictions of the negative class (the other condition). False Positive (FP) and False Negative (FN) are incorrect predictions of the positive and the negative class, respectively. A confusion matrix, Table 7, is used to display all these results together, where the columns indicate the actual outcome, and the lines indicate the predicted outcome.

Table 7: Confusion matrix. This table indicates the performance of a model.

		Actual class	
		Positive	**Negative**
Predicted class	**Positive**	TP	FP
	Negative	FN	TN

One of the most common forms of measurements of performance is accuracy. Accuracy measures the number of correct predictions over the total amount of predictions as a fraction (Equation 2.4.8.1).

$$Accuracy = \frac{Correct\ predictions}{Number\ of\ predictions} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.4.8.1)$$

When evaluating the performance of a model it is useful to consider a combination of measurements so to provide a better understanding of the model capabilities. Accuracy indicates the overall performance of a model and could be hiding the fact that the dataset is unbalanced and thus the predictor might not be working as good as

it should – it could be predicting all the instances as the majority class. For example, a dataset with 96% of the majority class could yield an accuracy of 96% whilst it is a model that cannot segregate the data points. To avoid such issues, the accuracy can be seen in conjunction with other measurements, such as sensitivity, specificity, precision, and/or F-measure, explained next.

Sensitivity, also known as True Positive Rate (TPR) or recall in other domains, is the rate of true positive predictions over the total amount of positive examples (Equation 2.4.8.2).

$$Sensitivity = TPR = \frac{TP}{P} = \frac{TP}{TP + FN} \quad (2.4.8.2)$$

Specificity or True Negative Rate (TNR) indicates the relation of correctly predicted negatives over the total amount of negatives (Equation 2.4.8.3).

$$Specificity = TNR = \frac{TN}{N} = \frac{TN}{TN + FP} \quad (2.4.8.3)$$

Precision or Positive Predictive Value (PPV) measures how good are the positively predicted values (Equation 2.4.8.4).

$$Precision = PPV = \frac{TP}{TP + FP} \quad (2.4.8.4)$$

The F-measure is the harmonic mean of sensitivity and specificity (Equation 2.4.8.5), and can also be used to strike the right balance between precision and recall (Equation 2.4.8.6) where β indicates the number of times sensitivity is more important than precision:

$$F_1 = \frac{2 * (Precision * Sensitivity)}{Precision + Sensitivity} \quad (2.4.8.5)$$

$$F_\beta = (1 + \beta^2)\frac{(Precision * Sensitivity)}{\beta^2 * Precision + Sensitivity}$$
(2.4.8.6)

Type I error or False Positive Rate (FPR, Equation 2.4.8.6) indicates the probability of having a false identification, such as the probability of having identified as positive a disease that is not there. Type II error or False Negative Rate (FNR, Equation 2.4.8.7) is the probability of identifying as negative a value that is positive. These two tests are essential in comparative statistics to verify if a statistic is significantly different.

$$Type\ I\ error = FPR = \frac{FP}{P} = \frac{FP}{TN + FP} = 1 - TNR$$
(2.4.8.6)

$$Type\ II\ error = FNR = \frac{FN}{P} = \frac{FN}{TP + FN} = 1 - TPR$$
(2.4.8.7)

The metrics sensitivity, specificity, precision and F-measure are good indications for model performance. These metrics can be directly applied in a model that does not have a score risk, e.g., a rule-based model might say that everyone with stroke and HF will have AF.

However, most models will provide a scoring value for each data point, e.g., a patient has a 70% risk of getting AF. This model can be given different cut-off points as a decision criterium to action, a percentage value that will define if the patient will be treated or not. To assess the overall performance of this model the combined effect of multiple prediction thresholds is used. In these cases, metrics such as the Area Under the Receiver Operating Characteristic Curve are useful.

2.4.9 Area Under the Receiver Operating Characteristic Curve

The Area Under the Receiver Operating Characteristic Curve (AUCROC), or C-statistic, is a measurement that considers not the number of correct predictions but how close to the correct prediction they were. It is calculated using the predicted scores originated from a model and provides a way to obtain different performance points

measured by the sensitivity and specificity originated from the model when using different thresholds (124) (125).

The rational of a threshold is that if the threshold cut-off is 50% and samples are scored higher or equal to 50%, then it means that they are of the positive class. If a sample is lower than 50%, then it is in the negative class. If the threshold is changed to another number, then the resulting model can be made to a more specific scenario. The threshold cut-off can be tailored to predict higher-risk or lower-risk patients. For example, if a predictor is not performing well for the classification of a patient as AF/no-AF, more patients with AF risk can be obtained by lowering the threshold, consequently increasing the sensitivity of the prediction, whilst reducing its specificity. A similar effect can be obtained by fine-tuning the precision.

The AUCROC is calculated with the integral of the sensitivity x specificity curve given different thresholds. One way of calculating this integral is by summing the rectangle area of the finite steps (126). To calculate the AUCROC using the rectangle method, a list of prediction scores for the model in increasing order is needed. The different predicted scores are used as thresholds. Calculate the different values of sensitivity and specificity for all the thresholds. Plot the values in a line plot. The area can be calculated by the sum of the partial points, considering the difference of length and height between the points. Appendix 2.2 exemplifies this process. The AUCROC confidence interval can be calculated and models can be compared using DeLong's algorithm (127) (128).

AUCROC measurement indicates the overall performance of a model. It can be used to compare different risk models, and after selection a threshold cut-off, it can be used with its confusion matrix, a table summarising the predictive performance.

2.4.10 Other metrics

Other metrics are usually used and may provide another form of interpreting the results.

For regression problems, the coefficient of determination (129), or R^2 can be calculated as:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = \frac{SS_{reg}}{SS_{tot}}$$

SS_{res} stands for the residual sum of squares, which represents the accumulated value of discrepancy between the real value and the predicted. SS_{tot} is the total sum of squares, indicating how far from the mean the values are spread, proportional to the variance of the data. *SSreg* is the regression or explained sum of residuals. They are calculated using:

$$SS_{res} + SS_{reg} = SS_{tot}$$

$$SS_{res} = \sum_i (Y_i - \hat{Y}_i)^2$$

$$SS_{reg} = \sum_i (\hat{Y}_i - \bar{y})^2$$

$$SS_{tot} = \sum_i (Y_i - \bar{y})^2$$

$$\bar{y} = \frac{1}{N}\sum_i^N Y_i$$

Where Y_i is the true output value, $\hat{Y}_i$ is the predicted value, *i* represent each instance in the dataset that goes up to its N^{th} term, $\bar{y}$ represents the mean value.

An R^2 of 1 indicates that there is no residual, the model has a perfect prediction on the dataset used. A model that predicts its mean value has *SSres* equals to *SStot*, leading to an R^2 of 0. If *SSres* is smaller than *SStot* it means that the model has some predictive power, despite not being perfect, has the capability of explaining some of the effects on the data, the value will range between 0 and 1. If *SSres* is bigger than *SStot*, the model is predicting worse than a "*mean model*".

The mean squared error (130) can be calculated as:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2$$

Where Y_i is the true output value, $\hat{Y}_i$ is the predicted value, i represent each instance in the dataset that goes up to its N^{th} term.

A value of 0 indicates that the model can precisely identify the samples tested, there is no upper bound on the error.

For classification models, the Brier score (131) has an identical formula to the mean squared error. The boundaries of the Brier score are 0 and 1, due to classification models only being able to predict between 0 and 1, where 0 indicates a model that perfectly predicts the samples.

2.4.11 What measure should be used?

Different metrics can be traced together for performance measurement. Other than AUCROC, the Precision-Sensitivity (Precision-Recall) curve and its area under the curve indicates a model performance under unbalanced scenarios. Whilst AUCROC considered both classes, the Precision-Sensitivity curve focuses on the minority class.

Table 8 describes some reference values of classification performance measurements. The definition of how good the accuracy of a model is depending on the balance of the dataset. The best AUC, in the usual conception, is considered the one farthest from 0.5. However, if a model is corrected to a new dataset and the performance is oscillating from very close to 0 to very close to 1, it is more likely that the model is performing due to chance, and the model must be revised. Similarly, if a model strongly oscillates its performances from one dataset to another, it indicates that the model has a variance problem.

A model with consistent wrong matches, such that the predictions are precisely off, indicate that the model is biased, or have modelled a biased model of the problem. This is the case of a model created under a higher risk population, the model will baseline, its intercept value in logistic regression, as incremented risk.

Table 8: Performance measurements and identification of performance for classification problems. * Although commonly the worst model has an area under the curve of 0.5, the worst possible model will have AUCROC 0 with its parameters overfit to some specific data points.

Performance measure	Range of values	Best	Worst
Accuracy	0-1	1	0
Brier score	0-1	0	1
Sensitivity, specificity, precision	0-1	1	0
AUCROC	0-1	1	0.5*
Precision-sensitivity area under curve	0-1	1	0.5*

2.5 Artificial Intelligence, Machine Learning

Artificial Intelligence (AI) is the use of methods that simulate intelligence to make decisions. AI includes all the methods that use static rules, algorithms that do not mimic learning and all the methods from Machine Learning (ML). ML has methods that simulate intelligence with learning. The learning of an ML task is its training, which results in a model with trained coefficients or functions.

Overall, ML is a group of algorithms that can improve their performance with experience. Mitchell (132) translates that into this question: “Given performance measurement P and task T can this AI perform better with more experience E?”(132). For example, an algorithm is used to perform a task of predicting a condition in the dataset of patients (experience) and using as a performance measurement the ability to correctly separate them, such as a person improving its technique the machine model improves with more samples.

Machine learning encapsulates approaches from classical statistics and computational algorithms. The results are backed by statistical testing and are analysed using computing approaches enabling it to operate in a larger magnitude of data, rendering it possible to obtain more complex solutions to more complex problems.

Many algorithms and approaches exist within the field of machine learning and most of them differ in their fundamental assumptions, bias and applicability to certain

domains. Therefore, it is essential to explore different options and apply different approaches.

2.5.1 Supervised and unsupervised learning

Before the exploration of different algorithms, it is required to define sub-categories of algorithms depending on the aim of the model.

Supervised learning involves algorithms where the target of learning is known. The target is another variable in the dataset, such as a number or a category. The algorithm tries to understand the patterns that lead from the other variables to the target one, e.g., a model could be aiming to identify if the patient has AF given different blood tests, for new patients the model tries predicting the risk of AF.

Unsupervised learning is the approach when there is no specific goal for the algorithm to evaluate a data sample against an output. These algorithms can be used to identify subgroups of variables, e.g., a sub-group of patients with a disease, or to identify patterns, such as correlations between variables.

Semi-supervised approaches utilize a dataset that has partial information numerical or labels (categories/classes) outputs and with many, potentially predominant, unlabelled data points to the creation of a model. This is especially useful to complete a dataset that has limited labelling and can be used as an assistant to complete data that is complete but requires manual interpretation, such as measurements or diagnosis out of images.

2.5.2 Association rule mining

Association rule mining, a type of unsupervised learning, is a method to obtain patterns on the dataset. It identifies associations between the pairings of categorical values, i.e., dataset of all conditions in the hospital EHR can be passed through a rule mining algorithm to identify what conditions go together with others and the inference of these conditions as rules, e.g., AF patients usually had another condition, such as stroke, heart failure and chronic kidney disease.

Apriori algorithm can quickly assess these patterns in a population, and identify linked conditions with more than two terms (133). This approach is especially useful for rare or directly linked conditions, where having a condition is highly conditioned to having another condition.

Apriori algorithm contains two main parts. Initially, frequent itemsets are identified. This is done bottom-up through incremental passages on the dataset counting if the number of elements sets in a dataset is above a support threshold. After each passage, these different sets are combined into bigger item sets, until there is no more combination of bigger item sets or all the combinations are tested. After the identification of the item sets that are above the support threshold, different rules are evaluated if they are above a confidence threshold calculated by the Bayes theorem (134).

Apriori can also be used on continuous values after transforming them using a discretization transformation (135). A discretization transformation is usually done to indicate an abnormality of values, e.g., if a blood test indicates that the glucose level is outside the 95% confidence interval.

Another algorithm for the identification of associations in a dataset is the CM-SPADE algorithm (136) (137). While Apriori identify rules without time constraints, CM-SPADE uses time sequences to identify associations.

2.5.3 Classical and contemporary machine learning

It is possible to separate the algorithms into classical and contemporary machine learning algorithms. Classical machine learning algorithms may be applied, as a baseline, before the use of more advanced methods. These include the static or dynamic creation of logical rules, logistic regression and other algorithms, such as random forest and support vector machines (138) (139). This includes unsupervised and supervised algorithms (132). In the more contemporary aspect, there is increased use of neural networks (140), in particular deep neural networks, such as convolutional and residual networks.

The main difference between deep networks and previous algorithms is their capability of abstracting their features, i.e., inducing models from raw data like images, texts or

sounds – rather than manually extracting and creating the variables from the data. It is also capable of estimating and simulating operations from different types of classical algorithms using neural networks. Neural networks became popular because of their capabilities of abstracting models, performance and the availability of computing power in the form of modern graphics cards, cloud computing and other specialized devices (141).

2.5.4 Decision trees and random forests

One widely used classical machine learning algorithm is decision trees (142). Decision trees have a tree-like structure. The root of the tree is the complete dataset, at each element of the three a decision is made based on a variable, if the variable is above or below a certain value, then the tree branches into other branches. Each other branch has other decision rules that further separate the dataset until a decision is reached. Figure 9 exemplifies a decision tree for the case of identifying patients with AF in a dataset.

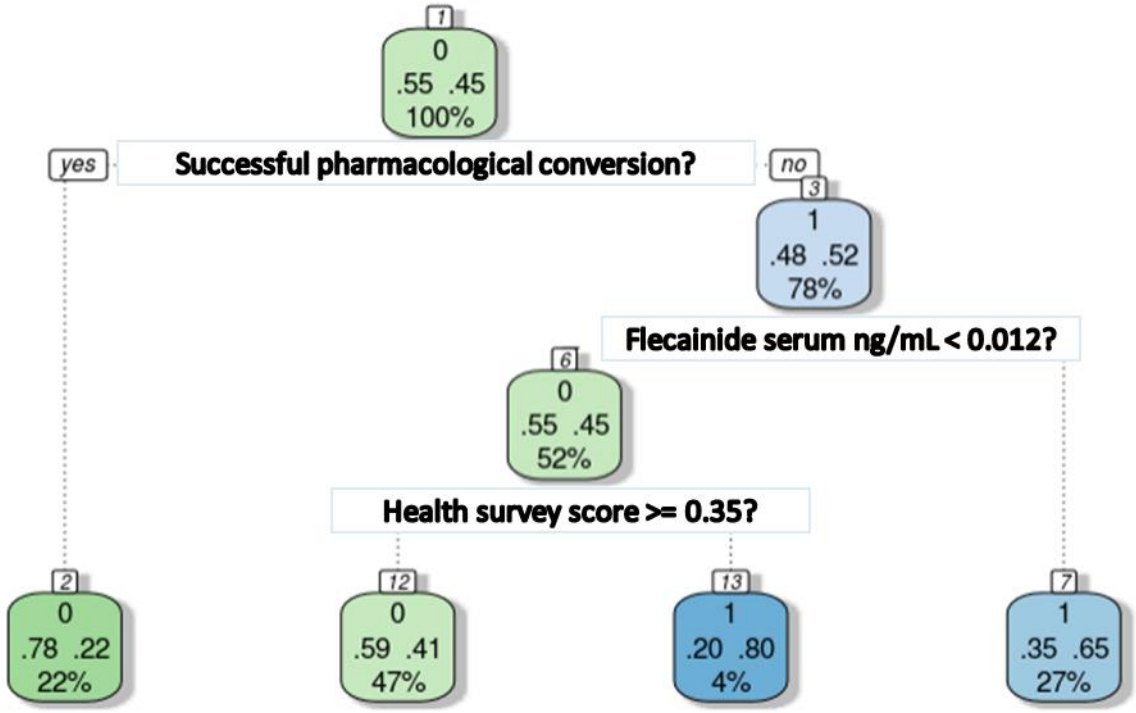


Figure 9: Sample decision tree. The different variable conditional expression under each node indicates the condition required to follow different paths if a condition is fulfilled. For each node, the number on top indicates the majority class in the node, the middle numbers indicate the rate of samples in this node without and with the outcome class, and the bottom percentage indicates the total percentage of elements that reach this node.

One algorithm for constructing decision trees is the ID3 algorithm. When creating a model using decision trees (model fitting), the initial step is to start with the whole dataset in the root node. On each node, the variables are assessed to identify the one that will have the most information gain or least entropy. This variable is selected to divide the branch. The node is identified with its majority class, and it is further divided if there are more variables or data points available.

Random forest is a machine learning that was built upon decision trees (138). This algorithm utilises a bootstrapping approach to sample with replacement sets of the training set with a limited number of columns. Different decision trees are trained for each subset. The final model consists of different decision trees voting for the predicted class or averaging the prediction for numerical outputs.

Whilst a single decision tree might suffer from the influence of noise, random forests can reduce the variance effect due to the bootstrapping, without increasing the bias.

2.5.5 Neural Networks

While some algorithms perform well on some datasets, they might fail on others. This depends on the data patterns and dataset structure. This is reflected in the literature where different algorithms are employed depending on the problem at hand (143). This is usually seen in classical machine learning algorithms, artificial neural networks (NN) are a versatile type of machine learning algorithm (144).

The performance of NN tends to be better, or at least as good as classical machine learning methods. However, special care is required to keep training, validation and testing sets given that NN tend to over-fit due to its increased number of coefficients. The application of NN might not be possible in datasets with a reduced number of samples.

A NN requires the transformation of data points into numerical values centralised around 0, due to the activation function. For example, a categorical variable such as sex will be transformed into 1 for male and -1 for female (or vice-versa), or a transformation that considers the proportion of each sex. Categorical variables with multiple options will be transformed into one-hot encoded variables.

NN are formed through the combination of different building blocks. In the lowest level, there is a neuron, or a node, which executes some operation on part of the data and passes this information to other nodes in different layers. Each layer may contain several nodes that execute a similar function. The passage of information between nodes is through a link connecting them. The way the connections are made between layers and the type of operation executed on the connected data define its functionality.

There is no limit to the format of a NN, there could be many layers that can be arranged in different ways. The only requirement is the presence of at least one input and one output layer. As the network grows wider and deeper the abstraction increases. If the network is very deep, containing from tens to thousands of layers, it is considered a deep neural network (DNN) (145).

The way the nodes and layers are connected influence the results that can be obtained. These different interconnections are called the architecture of the network. A dense network has nodes that are fully interconnected between layers, this abstracts combinations of the different variables (Figure 10). A convolutional neural network (CNN) operates in blocks of filters and value pooling through vectors of signals, such as a 1-D beat signal or a 2-D Magnetic resonance imaging (MRI) figure. The abstractions resulting from these operations lead to a network that *sees* in different scales the figures – different layers of a CNN return more specific information than previous layers and it is capable of interpreting or classifying something from the original signal (146).

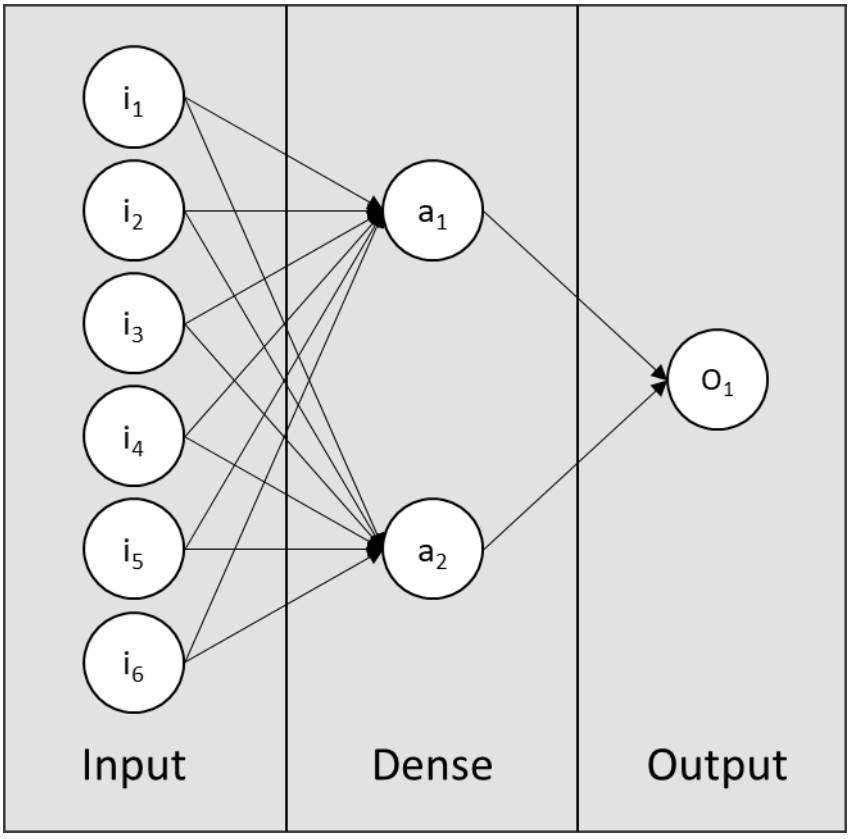


Figure 10: Representation of a neural network with 3 layers. The input layer contains the dataset variables information, the dense layer operates over the input data, and the output layer operates over the dense layer. While each node in the intermediary dense layer can abstract linear functions, the output layer can combine these linear functions into a higher dimensional model.

A NN works like logistic regressions, with an increased number of operations (147). For each node in each layer (after the input layer), an activation function is applied.

This activation function is analogous to a neuron sending a signal as the output of the node. Different functions can be applied, from logit/sigmoid operations to non-linear activations such as Rectified Linear Unit and Mish (148) (149). Activation functions have the following equation:

$$z_1 = \sum_{j \in L-1} (c_{j,1} * i_j) + c_{j,0}$$

$$a_1 = f(z_1)$$

Where z_1 is the accumulator for the first dense node, coefficients $c_{j,1}$ are operated over the previous layer values i_j, added to the bias term $c_{j,0}$, a_1 contains the value after the activation function over the accumulator.

A similar routine is applied to different nodes of the network up to the output layer. The output layer will provide the results of the network for specific input. Outputs for the data points are compiled and assessed against the true value using a loss function. The loss function indicates how wrong the results are, and they exist to emphasize different learning aims, such as minimizing the distribution of categorical values, e.g., cross-entropy loss, or a generic squared error function for broad cases (150).

The result of the loss function, the error of the model, is used to optimise the network. To do this, the amount of error is passed back to the network propagating from the output to the input layer, updating the coefficients. This process is called back-propagation.

$$L2\ Loss = \sum_{i \in samples} \sqrt{(\hat{y}_i - y_i)^2}$$

Where $\hat{y}_i$ is the predicted value, y_i is the true value for i different samples available.

The backpropagation takes steps in the direction where the error is going to be corrected. To correct the error, coefficients need to be changed in the direction that

reduces the error, this direction is the derivate where the error reduces, which is the gradient of the function.

$$c_l = c_l - \alpha \frac{dL}{dc_l}$$

Where c_l are the coefficients for the layer l, α is the learning rate, L is the loss function.

The backpropagation passes the influence of the coefficient from the loss function. If all the coefficients are equal, all the coefficients will be updated equally, and there is no learning. To avoid this issue, one needs to initialise the coefficients of the model with random values, one initializer function that suffices this requirement is the Xavier uniform initializer (151).

There is a multitude of gradient descent algorithms variants, such as Stochastic Gradient Descent, Adaptive Gradient Algorithm and Adam (152) (153) (154). These variants modify the learning process changing the rate and the influence of previous learning iterations. Figure 11 illustrates the process of gradient descent.

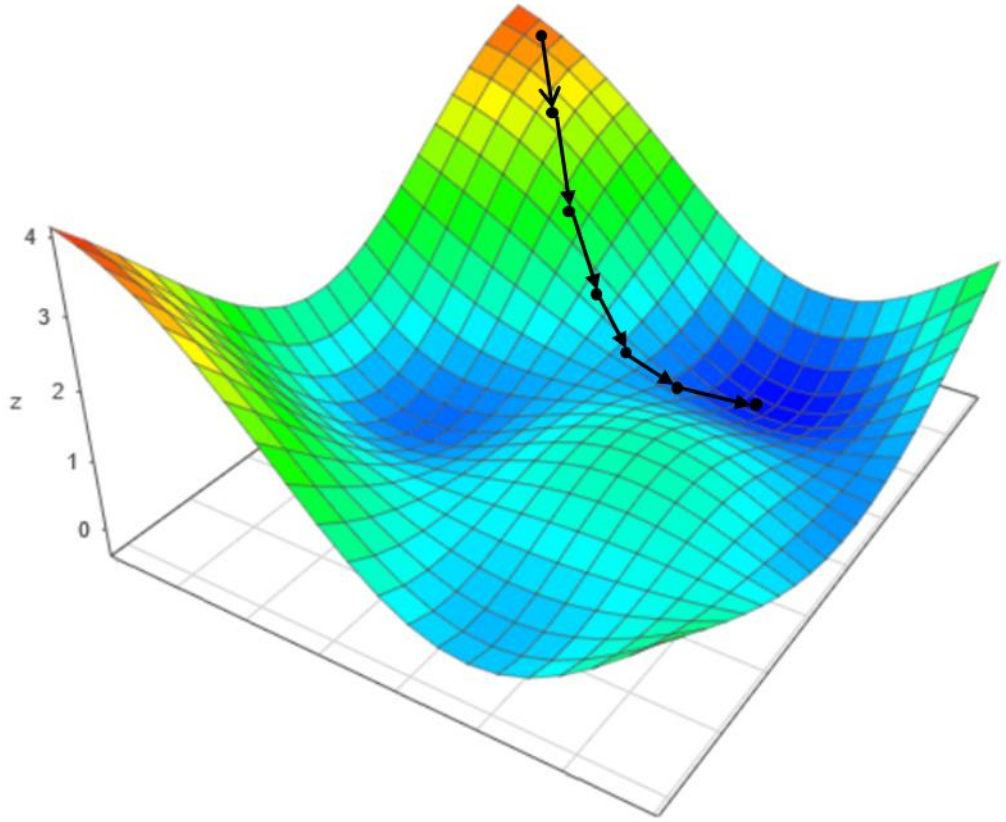


Figure 11: Gradient descent. The z-axis is the error, the coloured surface symbolises different coefficients possible for the model, black points indicate different coefficients and arrows indicate the optimization steps descending to the minimal loss or best model.

Model learning is also affected by the number of samples being seen by the model learning at each iteration. When all the data is trained at once, the coefficients can be updated with all the available information. If the dataset does not fit the memory, the model needs to be trained in batches. These batches can be used to iterate over the whole data updating the coefficients after seeing all the data, also called one epoch, or as mini-batches updating the coefficients after each batch.

Furthermore, the use of a pre-trained neural network model to the creation of another model on the same type of problem quickens the optimization process. This is called pre-training or transfer learning (155).

Automated Machine Learning

The search for the best neural network architecture is a complex problem. In classical machine learning approaches, the hyperparameters are well defined: the number of features in a decision tree is limited, as well as the number of decision trees in a random forest. In NN the hyperparameters are not well defined. The hyperparameters encompass the multitude of layers, their activations, types and dimensions, different links, and combinations of different blocks. This led to the advance of a meta-optimisation approach called Automated Machine Learning. Automated Machine Learning (AutoML) is a subfield in machine learning aiming at identifying hyperparameters for more complex architectures, especially for Neural Architecture Search. A formal language for encoding search spaces and a framework for the evaluation of architectures were part of this process (156) (157). BANANAS is a recent optimization method that at the time of its publication was shown to be the best performing through an evolution approach (158) (159). Despite there being other providers for AutoML, such as cloud services, some approaches reported in the literature are AutoKeras which utilises the well-known Keras framework to implement their solution and it is freely available (160), and Chainer focusing on DNNs (161). He et al. 2019 has a review on AutoML research (162).

Whilst PCA is an approach that linearly transforms the data to a lower-dimensional state, NN enables the transformation using non-linear transformations. It is also possible to reduce the dimensionality of the data, compressing the data using a commonly used architecture called autoencoder (163). This architecture is formed by two parts: an encoder section of the network, where for each layer the network has a reducing number of nodes until an encoded layer, followed by a decoder part, starting from the encoded layer and increasing the number of nodes until the original dimension of the datapoint. The learning is achieved by reconstructing the output to the original input. The encoded layer, with its reduced dimension, forces the network model to learn a reduced dimension representation of the signal, compressed variables with the most significant information of the data. This enables the creation of a network that learns new variables representing higher dimensional data, such as signals or images. One direct application is the use of a reduced version of the data, and thus the important features, without other biasing noises or elements. It is also capable of detecting anomalies where PCA does not detect (164). A different type of autoencoder

is variational autoencoder, which encodes a differential representation of the values, the mean and variance of the data samples. The resulting encoded representation is more continuous than encodings generated in the non-variational approach (165) (166). There is a multitude of other approaches for learning the representation of data samples, Bengio et al. provide a review on different approaches (163).

It is possible to generate samples related to the autoencoder when varying the encoded distribution of the values, with limited results. Generative adversarial networks, another advanced neural network architecture, is capable of better-generating samples (167) (168). These networks are trained with a dual structure: one part tries to generate an image, and the other part discriminates if it is real or generated. This architecture forces the networks to compete with each other leading to high performing results, and in the case of images, generating high-resolution realistic figures (169). This approach can be applied to different types of datasets, from structured to unstructured data, collections of variables, signals (such as audio and medical recordings) and figures could be generated simulating a distribution of samples, which could be normal or abnormal samples (170) (171) (172). Despite not being able to learn all the features of samples as an autoencoder, it opens the possibility of oversampling the dataset with higher fidelity (173) (174).

Long short-term memory

In the case of a continuous stream of information, i.e., a time-based data, long short-term memory (LSTM) layers can be employed. LSTMs are so-called recurrent networks, i.e. the network uses the information of previous time points and predictions combined with the current data point to predict future values (175). An example of an LSTM network applied to the prediction of AF is Yildirim et al. (176), in which a beat-by-beat signal from the PhysioNet was applied through a network for the encoding of the signal and prediction (45).

Other applications of Deep Learning are described by Dargan et al. (177).

2.5.6 Combining models

The data problems are getting more complex with time. Machine learning models handle a vast amount of data and operate on an ever-increasing number of parameters. When a single model is not enough for a prediction it is possible to create a model combining different models, this approach is called an ensemble. One way of creating an ensemble is by giving each sub-model a vote and then picking the most voted results. Alternatively, the ensemble can be done by re-training the output of all models giving each of them a coefficient indicating their weight in the final voting (132). The voting system is employed in the random forests algorithm, where the different decision trees vote towards the best-predicted value (138).

2.6 Variable importance

One could aim at different objectives when applying any analytical framework. In some cases, the aim is to obtain the best model to separate healthy against unhealthy patients. Also, a model can be used to separate data points into novel subgroups of related characteristics. In other cases, a model is applied to obtain important predictors as targets for further investigations. May et al. 2011 reviewed different approaches (178).

Independent of the scenario, a model can be used to assess important variables. The important variables can either be assessed isolated (univariate analysis) or in combination with the other predictors (multivariate analysis).

Investigative models might be focused on the predictors related to the positive case, rather than any variable that assists in separating the data points. Variables related to the target condition are identified for further investigations given their magnitude.

Univariate analysis fit the model using the different variables individually. The model performance metric and the value of the coefficient can be compared between the different predictors. A typical case of variable comparison is using the odds ratio for each variable logistic regression, comparing the magnitude of coefficients over scaled variables. One ML algorithm that directly assesses the importance of independent

variables to a predicted value is the Ranking Instances by Maximizing the Area under the ROC curve (RIMARC) algorithm (179).

Multivariate analysis utilizes the whole set of variables in a single model to evaluate the importance of the variables. To compare the effect between the variables they must be in a similar range of values. However, no effect can be considered without the other co-factors, given that these variables will have this effect due to the combination of the different variables. Depending on the algorithm, intrinsic traits will assist in identifying important variables. One such model for the evaluation of important variables is the elastic net algorithm, which penalises coefficients from logistic regression (180). Each model has peculiarities to the identification of important variables, Kuhn et al. (181) describes different approaches in the caret package.

Independent of the variable importance method, repeated experiments such as K-fold validation or bootstrapping (95) are suggested for increased statistical power of any analysis.

2.6.1 Correlations

It is possible to evaluate the importance of a variable with the degree of relation to the target variable or other variables. For example, it is possible to visually inspect data distributions, indicating predictors of potential. Similarly, correlation measurements between the other variables and outcomes can indicate that a group of variables does not add value to the model.

Commonly used metrics for correlation are the Pearson and Spearman correlation. The first indicates if the variables are linearly related, whilst the latter indicates if they are increasing or decreasing together. The range of correlation goes from -1 to 1, with negative results indicating inverse correlation and positive results indicating a direct correlation between variables.

2.6.2 Wrapper methods: backward, forward, and other searches

These methods involve the wrapping of the model creation into new steps to select the best set of features. This wrapping involves the creation of a model multiple times and assessing the performances measured.

Backward elimination starts the model with all the variables and interactively tests if the model would be better without one of the variables. After removing a variable, the process is repeated until the removal of variables stops improving the model performance.

Forward selection goes in the opposite direction of backward elimination. Different models are created starting from one variable, the best model is selected and a new test for adding a variable is made until there is no further performance gain.

The search can be done in different ways. Exploring exhaustively all the possibilities may be feasible for a small number of variables while applying heuristics or selected groups of variables that intuitively seem to lead to better performance may be necessary for a big number of variables.

The wrapper methods are classical approaches, and they aim to optimize the overall performance score, not necessarily picking the best individual predictors.

2.6.3 Algorithm-dependent importance

There are many algorithms that can be employed to generate models, and each of them address data under a different prism. To assess the ranking of importance, or the magnitude of importance, different methods exist for different algorithms.

To assess the important variables in a linear model, the odds ratio, derived from the model coefficients, indicate what variables are more important, and their absolute value indicate the ranking. In other models, such as LASSO, or another generalised linear model, a similar approach is employed (182).

In recursive partitioning algorithms (183), the reduction in the loss attributed to each variable at each split is calculated and the sum returned. For random forest, the

accuracy on the out-of-bag samples is recorded, then the accuracy is measured after shuffling the predictor variables. The mean difference is calculated for all the trees, then normalised with the standard error (184). Stochastic gradient boosting algorithm utilises a similar algorithm to the one for random forest, but using the entire training dataset, rather than the out-of-bag observations (185).

For the linear support vector machine used in this thesis the importance of variable is estimated through the use of AUC calculated for each predictor (181).

2.6.4 Shapley additive explanations

SHapley Additive exPlanations (SHAP) is a method of evaluating variable importance for more complex models, derived from game-theory approach (186). Although SHAP can be applied for different models, it is a state-of-the-art procedure for the interpretation of neural networks. It can be applied to other models, such as random forests, providing a way of understanding the inside of the machine learning “black box”. It enables that contemporary algorithms with high performance also provide interpretability. SHAP works by calculating the conditional expectation function for the model, providing the measure of additive feature importance for each feature.

2.7 Applications of statistical methods

Different questions can be approached by using each of the methods described above, but moreover, a whole new set of even more complex problems can be address by combining these different methods together.

The essence of quantitative analysis can be solved by the use of statistical methods. This chapter explained principles of the use of different datasets: where different datasets are not available, splitting a dataset enables some verification of created models, and a degree of bias identification (section 2.4.1). On every dataset there are different types of variables available, and the way they are interpreted (either as a number or a category) change their meaning, some metrics and variable transformations are shown (sections 2.4.3). For the visualisation of complex datasets, the whole data need to be transformed: this was traditionally performed using PCA, methods such as t-SNE and UMAP are more commonly performed nowadays to

handle non-linear relations (section 2.4.4). Missing values are common, especially in real-world datasets: different considerations could be made depending on the case, and data imputed (section 2.4.5).

Logistic regression is the first approach shown to model a set of input traits to an output categorical feature (section 2.4.7). This enables the creation of risk models, and the understanding of how much a feature influences the outcome variable. When creating models, it is essential to understand the performance of the predicted output, AUCROC, confusion matrix, and other metrics assist the interpretation and support their use (sections 2.4.9, 2.4.8, and 2.4.10). These metrics are essential to understand if, and how different models behave.

The use of artificial intelligence enables the creation of different types of models: supervised and unsupervised models are exposed, with a main focus on supervised learning, the creation of models that predict an outcome, including logistic regression (sections 2.5.1 and 2.5.3). Decision trees, random forest, and neural networks provide other ways of handling the data and creating models, models that become capable of abstracting non-linear relations (sections 2.5.4 and 2.5.5). Models that can be used isolated, or with other models, approaches such as ensemble enable the use of them together (section 2.5.6).

Independent of the method employed, it is essential to understand the magnitude of the features, towards the outcome or themselves (section 2.6).

A summary of the different methods and how they are used is shown in Table 9.

Table 9: Procedures, objectives and methods applications. A summary of the different methods exposed in the chapter and how they can be used.

Method (section/s)	Objective	Example of use (section)
Data separation (2.4.1)	To enable the verification of model performance and increase the robustness of analyses.	Structured data (chapter 3).
Variable type interpretation (2.4.3)	To support the interpretation of data' features.	Throughout this work.
Visualisation of complex datasets (2.4.4)	To help the visualisation of complex datasets through the reduction of feature numbers.	Data-driven discovery and validation of circulating blood-based biomarkers associated with prevalent atrial fibrillation (3.3)
Missing variables and imputation (2.4.5)	To support the interpretation of missing values rational and the re-insertion of features into missing values.	Covid-19 risk model (8.3)
Logistic regression (2.4.7)[1]	To obtain a model between a set of input features to an output categorical value.	CATCH-ME model validation (3.6)
Performance metrics (2.4.9, 2.4.8 and 2.4.10)	To enable the verification of a statistical model performance.	Throughout this work.
Machine learning supervised models (2.5.1, 2.5.3, 2.5.4, 2.5.5 and 2.5.6).	To obtain a model between a set of input features to an output feature using non-linear abstractions.	Early prediction of heart failure using electrocardiograms (5.4)
Machine learning association rule mining (2.5.2)	To obtain patterns of association between features.	Temporal analysis of data (6.4)
Variable importance (2.6)	To interpret the different features that contribute to a model prediction.	Development and validation of a multiple blood biomarker model (3.4)

[1] Logistic regression is kept separated from other machine learning methods in this table.

CHAPTER 3 STRUCTURED DATA: CLINICAL VIEW OF THE PATIENT

3.1 Introduction

Structured datasets provide a direct link between the data point, patient, and the information fields associated with them. This type of data is defined as having a fixed structure, and the measured values are related to the dictionary of variables, with fixed units or options. Different units identify the scale on which the continuous variables operate, and the magnitude quantifies its effect. Options, despite being also valid when considering continuous variables in a limited measurement resolution, are considered to be the levels in a categorical variable.

All data are susceptible to collection bias, and structured datasets are not exempt from that. Although data dictionaries provide definitions of parameters, interpretation of these definitions that are passed through the body of data generators and curators, varying in time and space, can introduce bias. It is possible to separate these issues into explicit and implicit data bias. It is considered an implicit bias when, for example, patients undertake different treatment pathways when going to specialist care, especially in different countries with different healthcare rules. Studies might consider different definitions for diseases, and different exclusions criteria to what is relevant to the hypothesis or context. These can be made explicit with the definition of terms. Besides, measurement techniques (for example, different assays) can have a varying resolution, although this difference is negligible when considering standard clinical tests. Variables collected may be influenced by the time and day of the week of the measurement, to a minor effect. Approaches such as human phenotype ontology and different Caliber phenotype portal try to systematise these differences (187) (188).

Despite these issues, structured data provides a direct way of analysing datasets. Inconsistencies are treated as noise. The data need to be cleaned and readied for analysis, and the fields are made consistent for analysis. The most common format a structured dataset can be presented is as a table, such as a spreadsheet file or in a SQL database (189). This chapter introduces and discusses the development and

analytical framework and the results of its application across four different cases of structured data related to atrial fibrillation.

Atrial fibrillation (AF) is often identified only after a complication, such as a stroke (190) (191). Most times AF is missed, especially due to paroxysmal presentation in the earlier stages and not seen in standard 10 s electrocardiogram recordings. Furthermore, AF screening is burdensome for patients (192). Early identification of AF can support preventive treatment of associated conditions (193) (74) (194).

Different comorbidities and clinical information are associated with AF and could be used to predict its presence, such as hypertension, ischaemic heart disease, heart failure, prior stroke, diabetes, obesity, and age. The performance when using this information is often limited and requires specialised knowledge.

3.2 BBCAF machine learning pipeline

The cases explored in this chapter consist of data that is structured. Either the information was collected and compiled for a particular study, or the data is available as part of a clinical system formulated in a manner that renders them amenable for structured data analysis.

These data typically contain a set of input variables, such as clinical characteristics, laboratory test results, and other derived variables. A common outcome variable is whether or not the patient has AF – its risk score.

Such datasets can be useful to address aims such as identification of patterns between the variables and subgroups, such as correlations and statistics, creation of supervised/risk models, and the identification of important predictors for these conditions from risk models.

There are different approaches in the literature to the identification of important features, e.g. the RIMARC algorithm is capable of identifying important predictors of persistent AF or death (179) (195).

To facilitate the analysis of structured datasets, a pipeline was built as a methodological approach for the automated development of models and the identification of important biomarkers using different machine learning approaches.

This pipeline was created utilising the R language for statistical computing (196). It is based on the *caret* library (181). The main algorithms utilised are from the libraries *ROSE*, *randomForest*, *e1071*, *glmnet*, *rpart*, *gbm* and *pROC* (197) (184) (198) (182) (183) (185) (199).

This pipeline has three main steps: data preparation, the definition of execution settings and the execution of the pipeline (Figure 12).

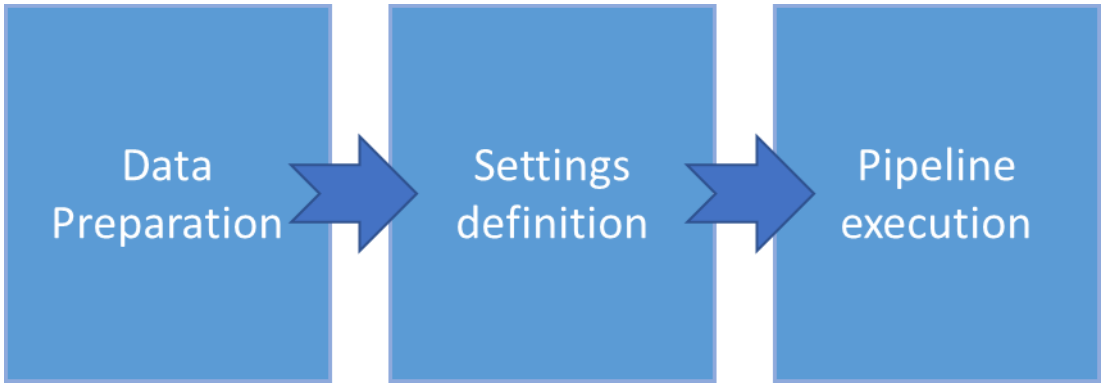


Figure 12: Three steps of the BBCAF machine learning pipeline. Data preparation gets the input data ready for analysis. Settings definition identifies the different options available in the pipeline. Pipeline execution starts the tool and obtains the results.

Data preparation. In this step, the data must be prepared before plugging into the tool. Whenever a dataset is loaded into the system its column definition and types might not be loaded properly. The operator needs to check and correct the column types, recoding categorical columns that might have been assigned as numerical, and transforming cases with missing values e.g., indicated by the numeric value 99. It will also require that the output variable be defined. This step needs to be done manually on a case-by-case basis depending on the dataset.

Settings definition. The tool supports a range of settings with different options: the proportion of split sets, randomised key seed, types of training (cross-validation, repeated cross-validation, bootstrapping or direct train-test), number of folds and

repetition of tests, number of processing cores, criteria for the elimination of rows and columns due to amount of missing values; MICE iterations to impute missing values, selection of algorithms (as support by the caret platform); sampling using up, down or ROSE sampling; if columns that have near-zero variance should be removed; if columns should be centred and scaled; different modalities of feature selection and the number of iterations, grid or random search, feature selection before the algorithms pass or with the algorithm using a wrapper; and other options.

In its default settings, variables are centred and scaled to unit variance – this is done to improve interpretation of model variables, and to improve model training and performance, a step that is for learning algorithms such as support vector machines, where quadratic elements in the algorithm will lead to improper training. Models split the data into training and test sets and explore the training set using cross-validation.

Pipeline execution. Dataset is loaded and executed following the sequence of loading the data, executing a post-loading data preparation as defined by the settings, followed by feature selection and algorithm execution. The results are returned to the operator for evaluation. In the case of a regression model, the performance is measured in R^2 or mean squared error, classification models are evaluated using the AUCROC.

3.3 Data-driven discovery and validation of circulating blood-based biomarkers associated with prevalent atrial fibrillation – Case 1

3.3.1 Introduction

Blood biomarkers, like those measured in clinical practice, provide an alternative and potentially enhance the predictive performance of AF risk models. Numerous candidate biomarkers for the identification of AF have previously been identified, for example, N-terminal pro-B-type natriuretic peptide (NTproBNP) and brain natriuretic peptide (BNP) indicating myocardial stretch, C-reactive protein (CRP) towards inflammation, Galectin 3 for cardiac fibrosis and glomerular filtration rate (GFR) indicating renal function (200) (201) (202) (203) (204). These blood biomarkers are commonly evaluated individually or grouped, and they are often utilised for other cardiovascular diseases, such as coronary artery disease and heart failure (205).

It is hypothesised that there are biomarkers that can better contrast the pattern of risk between patients that have prevalent AF and no AF diagnosis at recruitment.

3.3.2 Data description and analysis

To assess the hypothesis, data at baseline were collected and used to identify differences in patients that have or have not AF. The dataset used for this analysis is the BBCAF cohort containing the assay biomarkers (described in section 2.3.1).

The initial dataset exploration was based on visualisation as well as unsupervised tests through PCA transformation analysis (105). Correction of data points batch-effect was performed using an empirical Bayes method using R package *sva* (206) (207). Missing variables were imputed using the MICE method in R (109, 110).

Statistical tests evaluated the differences in biomarker expression between participants with identified AF against sinus rhythm; comparisons were performed using t-test or Mann-Whitney tests depending on the normality of the data distributions, assessed using Kolmogorov-Smirnov tests (118) (119) (208). Models using logistic regression and the BBCAF machine learning pipeline were created, using a 5-fold cross-validation approach, and a random forest feature selection before executing 5 algorithms representing different machine learning approaches: lasso and elastic-net regularized generalized linear model, support vector machines with linear kernel, random forest, stochastic gradient boosting and recursive partitioning. The performance was measured using AUCROC. Important variables for the logistic regression model were assessed using odds ratio, for the machine learning models their importance was calculated from the scaled importance (more details on variable importance methods are in section 2.6).

3.3.3 Results and discussion

PCA transformation and visualisation indicates that variances in the dataset can be split into two main subgroups, and it shows that there is a bias on the dataset from the different cardiovascular assays utilised (Figure 13). In addition to bias introduced by using 2 different assays, when plotting the variable points one-by-one (Figure 14),

there are also indications of batch effects with strong separation between the different groups of patients within each cardiovascular assay.

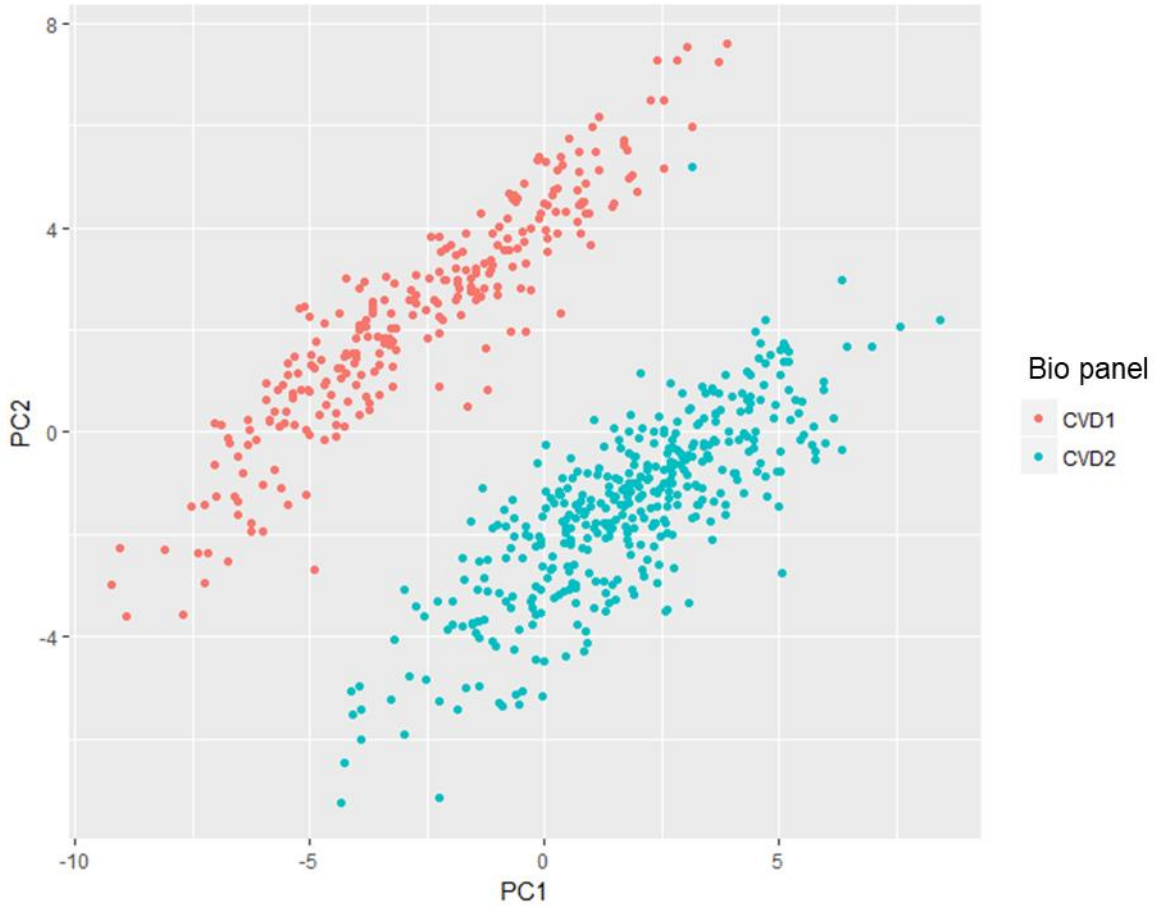


Figure 13: Principal component analysis of the BBCAF dataset. Two subgroups were identified on the data because of batch effects due to differences in the cardiovascular panels utilized. CVD1 stands for Cardiovascular panel I and CVD2 is Cardiovascular panel II.

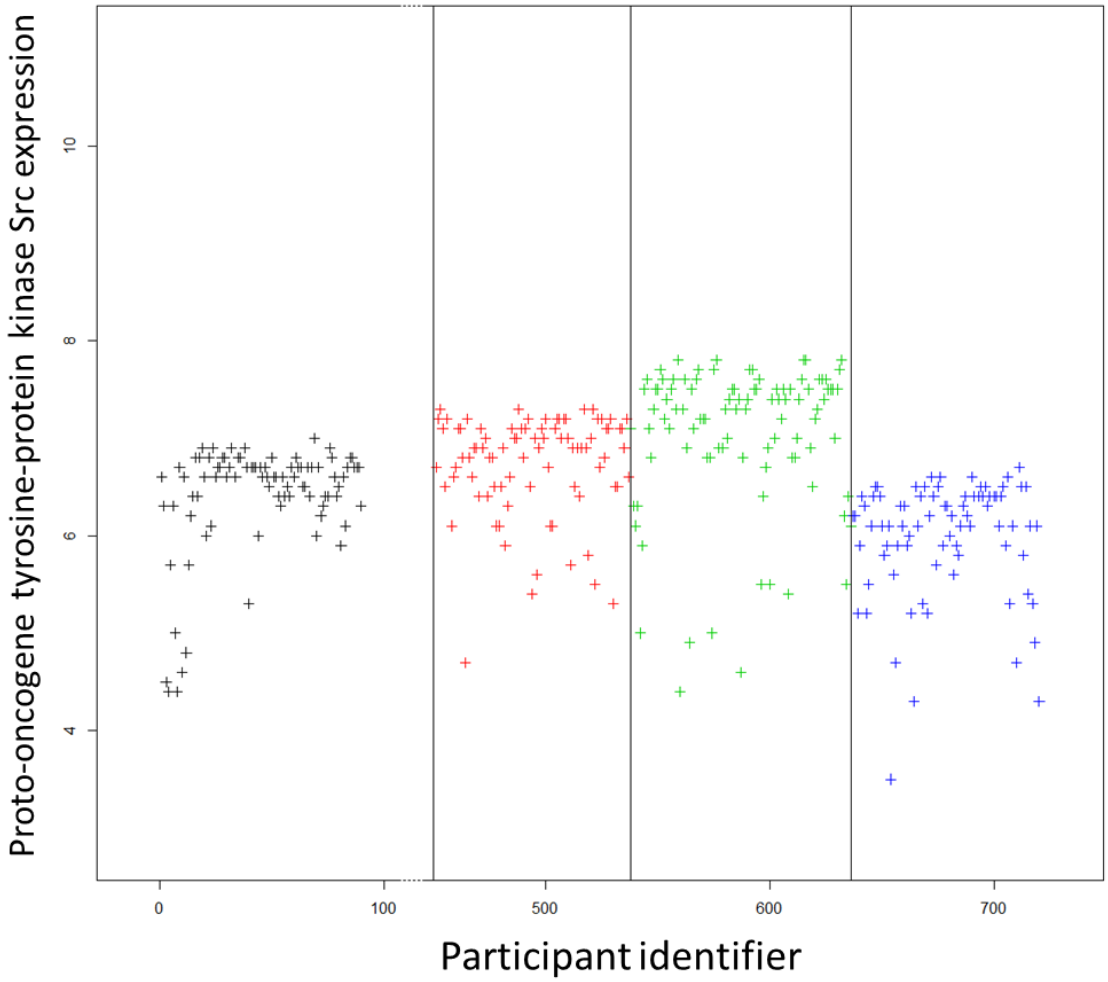


Figure 14: Distribution of the values for Proto-oncogene tyrosine-protein kinase Src before any pre-processing. On the y-axis the estimated measurement of the assay, on the x-axis the patient number in order of recruitment. In black (left) are some samples from the Cardiovascular panel I, in red, green, and blue are different batches of panels for the Cardiovascular panel II.

Issues raised from the batch-effects led to the investigation of correction methods with the assay provider. A correction of batch effect was required and it was done using an empirical Bayes method (206). The method employed fixed the identified batch effects.

Logistic regression indicated increased risk for increased values of different variables (odds ratio): age (1.06 95% CI 1.035 – 1.095), male sex (2.022 1.275-3.564), BMI (1.06 1.021 – 1.115), *BNP* (1.293 1.112 – 1.627) and *FGF23* (1.667 1.363 – 2.344). *TRAIL-R2* was shown to indicate higher risks of AF for lower values (0.242 0.135-0.323).

Machine learning models confirmed features *BNP*, Age and *FGF23* as top predictors for AF risk (Figure 15). The best-identified model in the cross-validation step was the Lasso and elastic-net regularized generalized linear model with a final validation set AUCROC score of 0.697 (95%CI 0.63-0.76).

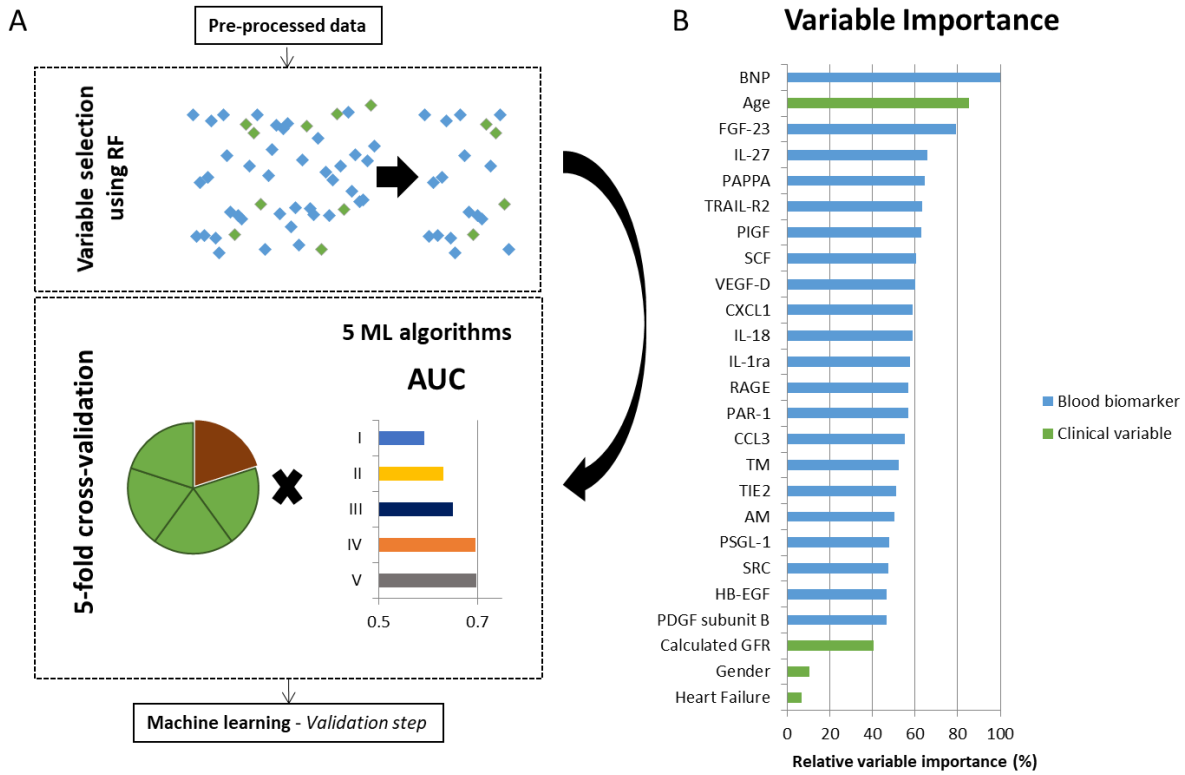


Figure 15: Framework for the BBCAF machine learning pipeline analysis and feature importance results. A indicates the framework and the final scores of models tested, B shows the ranking of variable importance identified in the feature selection stage.

Other machine learning models repeatedly show important variables, such as *FGF-23* showing up within the top predictors for most algorithms, in combinations with age, BNP, and sex (Table 10).

Table 10: Relation of important variables for different models applied in the BBCAF machine learning pipeline. Variables in bold were identified in the logistic regression model.

Ranking	**Lasso and Elastic-Net Regularized Generalized Linear Model**	**Support Vector Machines with Linear Kernel**	**Random Forest**	**Stochastic Gradient Boosting**	**Recursive Partitioning**
1	**Age**	**BNP**	**Age**	PIGF	PSGL-1
2	**TRAIL-R2**	**Sex**	**FGF-23**	**FGF-23**	**FGF-23**
3	RAGE	**FGF-23**	IL-27	SCF	CXC1
4	TM	VEGF-D	**BNP**	PAPPA	TIE2
5	PAR-1	IL-27	PDGF sub-B	VEGF-D	**Age**
6	VEGF-D	RAGE	SRC	IL-27	IL-27
7	IL-1ra	CCL3	**TRAIL-R2**	**Age**	PAPPA
8	PAPPA	ADM	ADM	**TRAIL-R2**	RAGE
9	PSGL-1	SCF	**Sex**	PSGL-1	TM
10	**Sex**	PAPPA	IL-1ra	**BNP**	IL-1ra

Starting from an initial number of 40 biomarkers, repeated analysis and validation using different approaches point to the predominant relevance of *FGF23*, *BNP* and *TRAIL-R2*, the first two indicating an increased risk of AF for increased values, later biomarker

indicating increased risk for lower presence. Furthermore, age and sex predictors were re-identified of valuable importance when creating these models.

Brain natriuretic peptide (BNP) is a peptide discovered in the porcine brain. However, its highest concentrations are found in the heart (209). It is produced by the cardiomyocytes in response to stretch or pressure increase. It is a known marker for cardiovascular risk (209).

TNF-related apoptosis-induced ligand receptor 2 (TRAIL-R2) is associated with the risk of myocardial infarction (210). Univariate analysis did not show a significant difference for this biomarker between AF and sinus patients (P=0.727). The influence on the models is related to other clinical variables.

Fibroblast growth factor 23 (FGF-23) levels are elevated in patients with chronic kidney disease, at it is associated with mortality (211). In the literature, there are studies with a significant difference of *FGF-23* between AF and a non-AF population (212) and studies that show a difference but are not significant (213). *FGF-23* is linked with higher risks of cardiac hypertrophy, which may lead to AF (214).

3.3.4 Limitations

Despite classical techniques and novel machine learning analysis producing similar results, this study has some limitations. Namely, there are potential biases on the patient selection criteria, and imputation of data. Furthermore, external validation of the findings is required, including the exploration of the biomarkers in broad populations, and longitudinal studies with incident development of AF.

Although confounding effects were not explored in this use-case, another study co-authored investigated the effects of comorbidities against some of the biomarkers, those are the cases of heart failure and chronic kidney disease, against NTproBNP and FGF23, which showed that their predictive value for AF remains significant (5).

3.4 Development and validation of a multiple blood biomarker model – Case 2

3.4.1 Introduction

Blood biomarkers measured using research assays provide an estimate of a wide range of parameters that could be used in the assessment of AF risk. After identifying some candidate markers using that technology, it is hypothesized that advanced tests of blood biochemistry, at the same time will reduce the measurement noise, improving the predictive power and also be closer to measurements applied in standard clinical practice biochemistry tests.

3.4.2 Data description and analysis

This study further explored the Birmingham Black Country Atrial Fibrillation registry (described in section 2.3.1). The cohort utilised in this analysis contains the 12 Roche biomarkers.

Dataset for analysis was split into 60% discovery and 40% validation set. Statistical analysis was done creating different risk models using univariate and multivariate logistic regression with backward feature selection, BBCAF machine learning pipeline, and neural networks (NN). Logistic models were adjusted by age, sex, BMI, eGFR, heart failure, stroke/TIA, and hypertension. Important features were identified through logistic regression odds ratio and machine learning models scaled feature importance.

NN models were created using the Keras framework (215). Data were processed using Scikit-learn (216), categorical variables, sex and comorbidities, were transformed using a Min-Max scaler. Continuous variables were centred towards mean and scaled to unit variance using the training set as reference. Models were optimized by assessing a range of hyperparameter (further explored in Appendix 3.1). The final model created contains 2 layers with 256 hidden dense variables with RELU activation (217), and a dropout layer (218). After these, an output layer with sigmoid activation contains the prediction. The model was trained with *adam* optimizer (219), and early stopping with 20 epochs patience. The best model was selected using the best performing model using binary cross-entropy loss. Further ten models were created

using different randomization parameters to verify the importance of variables using the Shapley Additive explanation (SHAP) (220).

3.4.3 Results and discussion

Univariate analysis indicated a significant increase of AF risk for increased levels of *ANG2*, *BMP10*, *FGF23*, *IFGBP7* and *NTproBNP*. High levels of *TnT* were linked to sinus rhythm patients. Other biomarkers were not confirmed significant. The multivariate logistic model performed with an AUCROC of 0.743 (95% CI 0.712-0.775), features selected and positively associated with AF were age, sex, BMI, *ANG2*, *BMP10* and *FGF23*.

After execution of the steps from the BBCAF machine learning pipeline, the best performing model identified is a support vector machine model, which yields in the validation set an AUCROC of 0.733 (95% CI 0.691-0.775), NN model performed with

an AUCROC of 0.784 (0.745-0.822). Figure 16 shows the AUCROC plot for the models.

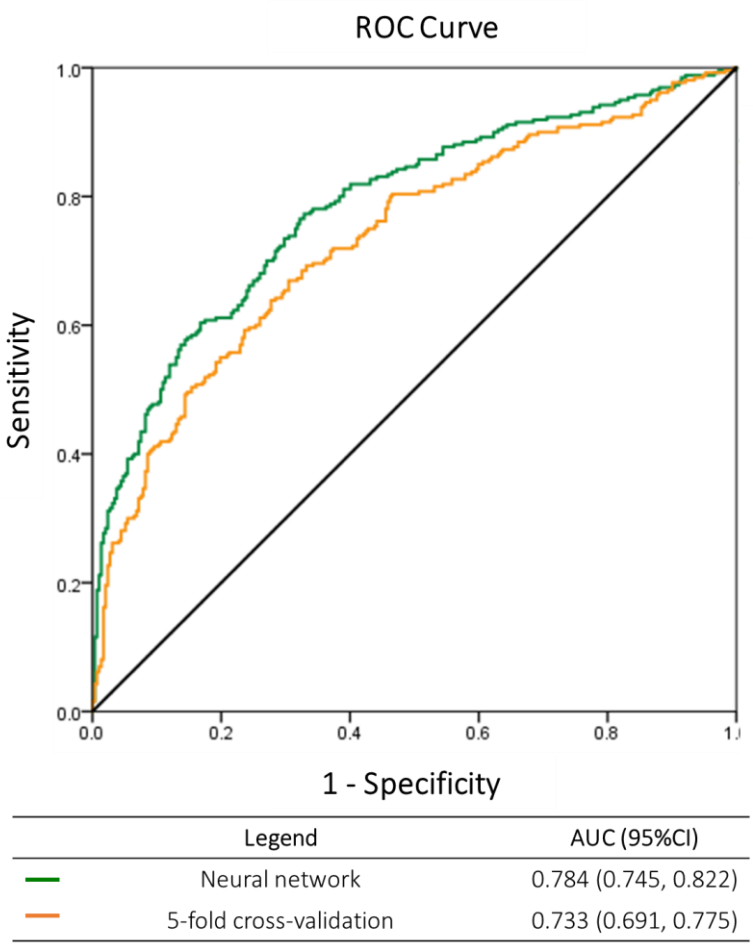


Figure 16: Comparison between the BBCAF machine learning pipeline result (5-fold cross-validation) and the NN performance measure using an AUCROC metric. In the validation set, the BBCAF machine learning pipeline best performing model yielded an AUCROC of 0.733 (95% CI 0.691-0.775), and the NN model scored 0.784 AUCROC (0.745-0.822).

Further to the predicted AUCROC, which provides an overview of the behaviour in the dataset, identifying some cut-offs in the model allows analysing a model that would be used in practice. Cut-offs are representative of the points where a decision will be taken in practice. Table 11 shows different resulting metrics for different hand-picked thresholds. A model that aims to identify most patients as possible, even with lower risk, would pick a low threshold value, such as 0.1, which is expected to collect 95.4% of the positive cases, whilst a threshold of 0.9 will obtain positive patients that are more

likely to be true positives, however in higher numbers of patients than an even higher cut-off. There are different ways of making a decision, such as net benefit, which are left for future improvements in this study (221).

Table 11: Performance metrics for different model thresholds. Changing the threshold value affects the risk level of patients that are obtained and segregate these patients that would be selected in a model in use.

Threshold	Accuracy	PPV	NPV	Sensitivity	Specificity
0.01	47.1%	47.1%	50.0%	99.6%	0.3%
0.1	53.3%	50.2%	79.3%	95.4%	15.8%
0.25	66.7%	60.1%	80.7%	86.9%	48.6%
0.5	71.0%	72.7%	69.9%	61.5%	79.5%
0.75	66.5%	87.9%	61.8%	33.5%	95.9%
0.9	60.9%	92.3%	57.6%	18.5%	98.6%
0.99	54.2%	88.9%	53.6%	3.1%	99.7%

BBCAF machine learning pipeline indicated that the variables *BMP10*, *ANG2*, *TnT*, *FGF23* and age are the most relevant variables for the created models. When evaluating the NN (Figure 17), it is shown that the *ANG2*, *BMP10* and *FGF23* are biomarkers associated with AF risk. Variables such as male, age and BMI are shown again of importance to the risk stratification of AF.

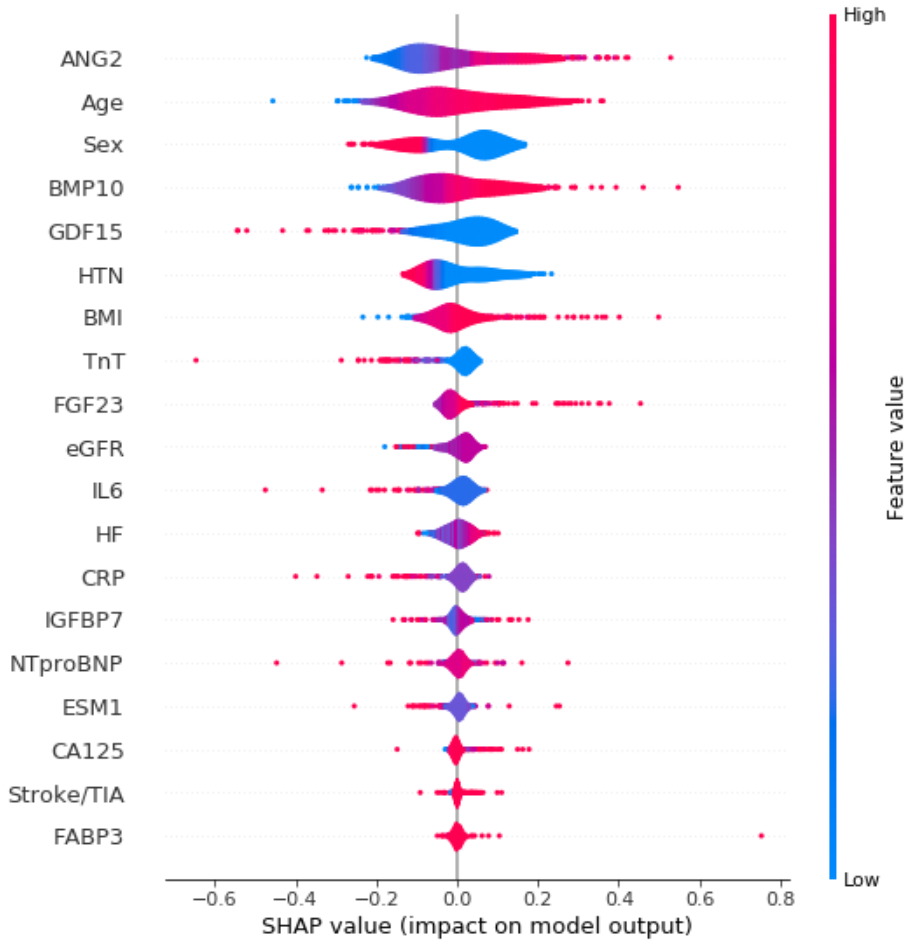


Figure 17: SHapley Additive exPlanations measures the impact of the different variables in the NN model. Increased values of SHAP indicate a positive influence toward predicting AF. Feature values indicate the magnitude of the variable. Increased values of *ANG2*, *BMP10* and *FGF23* are indicators of a higher risk of AF. Being a male, having heart failure or increased BMI are also show as increased risk of AF.

Angiopoietin-2 (ANG2) in increased levels is associated with the presence of oedema in patients with heart failure and worsened outcomes (222). ANG2 has been previously shown a significant difference for AF patients (223). Due to AF pathophysiology being associated with atrial fibrosis, ANG2 may be associated with inflammation in the regions. Further understanding of its role at the cellular level is required to understand its effect on AF.

Bone morphogenetic protein 10 (BMP10) is expressed in cardiomyocytes. The family of BMP proteins are essential to the regulation of cardiovascular structure and function, and it circulates through blood (224).

Different analyses confirm an increased risk of AF for patients who are male, older, and have increased BMI shown in the literature. Biomarkers ANG2, BMP10 and FGF23 are identified as strong candidates for assessment of patient risk. Different models identified (Table 11) can be personalised for practice use, and thresholds can be used to balance the number of participants selected for treatment.

3.5 Socioeconomic factors to atrial fibrillation: a study of the influence of the patient location Birmingham Black Country Atrial Fibrillation dataset – Case 3

3.5.1 Introduction

Patient's lifestyle may be an indication that they have an increased risk of having AF (225) (226). Lifestyle information is not collected in a standardised format on clinical records, sometimes it is collected only as punctual information such as patient's job, residence type, religion, marital status or residence postcode.

A patient residence postcode can be used to infer information from Census data. The Census provides information such as income, employment, education, health, crime, housing, services, and living environment (227). The information contains an aggregation of data from a region, and it is assumed that the aggregated information is very close to the individual level.

In the BBCAF dataset, variables expose different clinical factors, such as age, diseases, medications, and associated risk scores. Furthermore, there are ECG recordings and family disease history. There are limited data about a patient's quality of life: SF-12 and EQ5D variables provide a summarized perspective (228) (229). Information such as participants' professions, and their associated stress levels, education, income, air pollution where they live, accessibility to healthcare, and living conditions are not available. It is hypothesised that exogenous variables derived from the Census can provide further knowledge to improve the understanding of both the patient and its health outcomes to the development of AF. These variables extracted

from the postcode are hypothesised to behave as proxy variables for these data not available.

3.5.2 Data description, integration and analysis

The initial locked data for the BBCAF analysis (described in section 2.3.1) was enriched with information extracted from public databases. Despite the UK census providing information from different aspects about a region, it is not enough and directly usable for some analysis considering only a few aspects of a resident's life. The different datasets utilised are the 2011 UK census (230), UK police statistics dated from March 2016 to February 2019 (231), index of multiple deprivation 2015 (232), and income data (233). All these different datasets are available under the Open Government License (234).

These different datasets provide a wide range of variables explaining different aspects of an average person lifestyle to a region. Census provides information about housing, civil status, dependent children, qualification, economic activity, jobs and occupation, qualification, healthcare, age structure, usual resident population, ethnicity, country of birth, and other information. The census is the most complete source of data for a region, however, due to its frequency and the broad range of variables, it does not provide easy separation of variables that may be used in an analysis. The index of multiple deprivation is commonly applied in the literature for population studies – it is often associated with worse medical outcomes (235). Index of multiple deprivation is more frequently available than the broad census data. Furthermore, crime statistics and income data provide another aspect of information to different small areas.

These different datasets contain a range of variables explaining different aspects of a person's residential region. Data are made available in different scales depending on the type of data. In its primary format, the data is collected to an individual level, aggregated datasets are released with grouped regions of different sizes. For example, the index of multiple deprivation is linked to lower layer super output area, while the basic housing census is released in census output areas.

To preserve the identity of the census participants, postcodes are grouped into output areas, with a minimum of 40 households, recommended more than 125 households. These output areas are further aggregated into different levels, lower layer super output area (LSOA), middle layer super output areas (MSOA), upper layer super output area, and local authority districts. The most commonly used levels for census output are the LSOA and MSOA. A postcode lookup is used to link the postcode to the census output areas (236). Figure 18 illustrates the mapping from a postcode to different outputs areas.

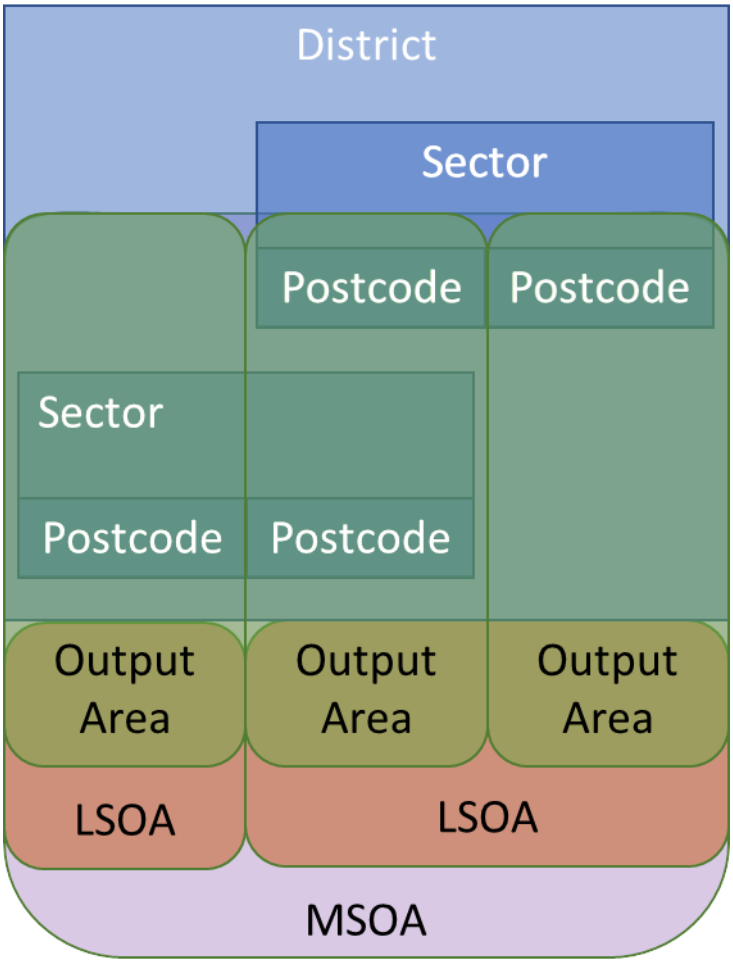


Figure 18: Representation of postcode and output areas. Postcode contains different parts, the outer code indicates the area, district and sub-district; the incode part is formed of the sector and with the unit forms a postcode. Postcodes are grouped into census output areas, further grouped into lower layer super output areas (LSOA) and middle layer super output area (MSOA). There is no clear distinction of postcode districts into an output area level. Larger representation areas are not shown.

BBCAF dataset was linked and complemented of variables from the UK Census 2011 (227), crime reports between March 2016 and February 2019 (231), income estimates for the financial year ending 2018 (233), and index of multiple deprivation for England in 2015 (232). More than 500 new variables were collected from these datasets (237).

Models were created using the BBCAF machine learning pipeline, performance was evaluated using AUCROC and the variable importance.

3.5.3 Results and discussion

The goal of using this linked dataset is to provide a better understanding of general lifestyle qualities that might be affecting the patient health, information that are not routinely collected in a clinical appointment. Figure 19 shows the distribution of patients stratified by conditions and their locations in the country.

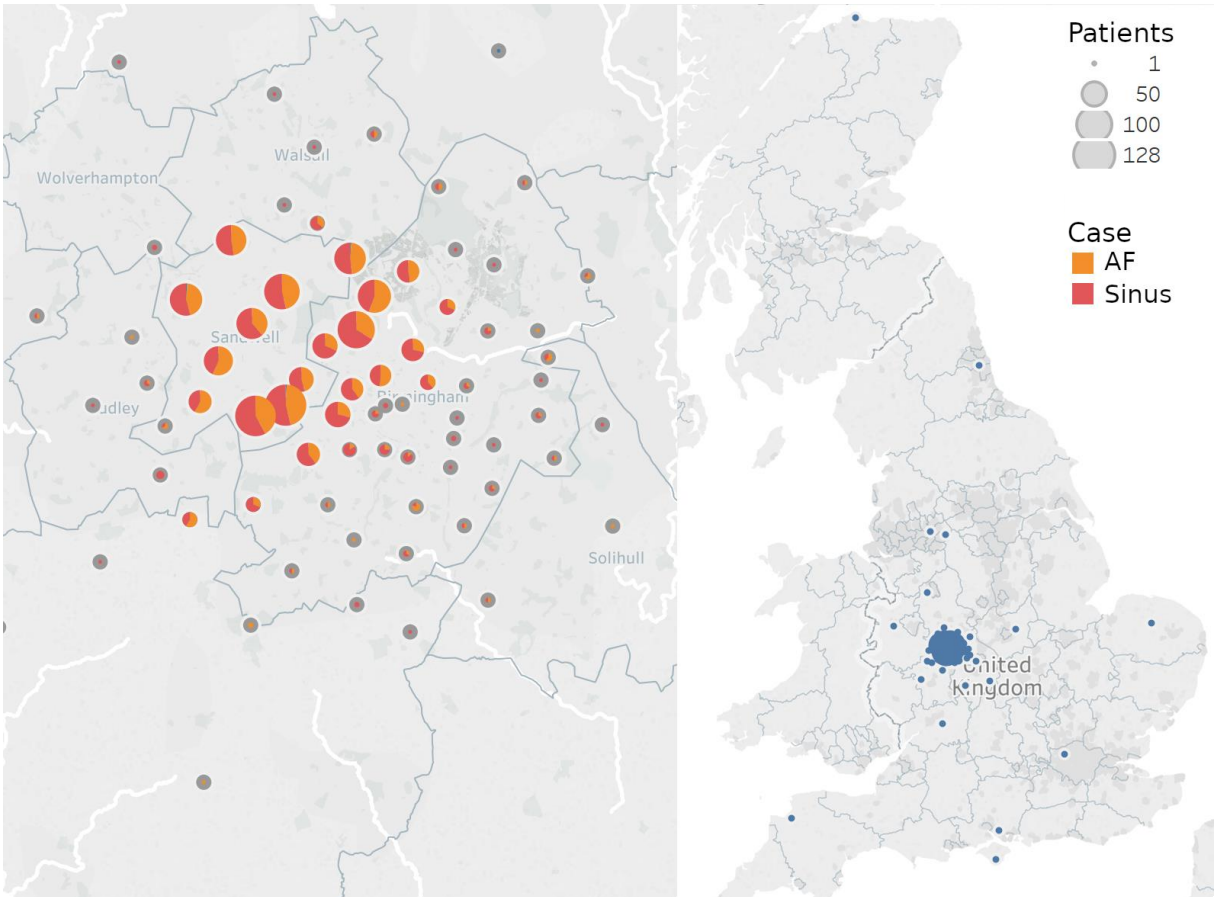


Figure 19: BBCAF patients distributed over the country and concentrated around Birmingham

The best performing model yielded a performance not significantly different from previously created models. Although an increased number of variables was used, and BBCAF machine learning pipeline filtering variables, the only derived variable that showed predictable power was the median age of the participant's region. Other variables were not considered to have enough power to influence the model created in this setting.

The outcomes for this case are: (1) a tool for inclusion of associated data and different data sources, grouped by different levels of geographical (237) which can be applied to other datasets; and (2) different visualizations for the spatial distribution of these patients. Future studies can use the tool for different types of spatial analysis, with the potential to show effects in a more diverse dataset.

3.6 CATCH-ME model validation – Case 4

3.6.1 Introduction

BBCAF is one of the datasets repurposed in the Characterizing Atrial fibrillation by Translating its Causes into Health Modifiers in the Elderly (CATCH-ME) consortium (238). This is a collaborative project between partners in different European countries. The project aimed at integrating and evaluating new insights from 12 AF studies contributed by 6 centres from the United Kingdom, Germany, France, Spain, and the Netherlands.

The CATCH-ME project created different risk models for AF in different scenarios, including models derived from individual patient data across multiple studies. External validation of the created models is paramount to ensure the applicability of the created models into broader practice.

3.6.2 Data description and analysis

A previously created model based on the CATCH-ME dataset (described in section 2.3.1), with parameters age, gender, BMI, height, indication of ECG abnormality, LA volume, LVESD and morbidities hypertension, diabetes, coronary artery disease, tricuspid valvular disease, and medications aldosterone, beta-blockers and P2Y12

inhibitors, was missing an external validation, which was performed across the UK Biobank (described in section 2.3.3).

The equation for the risk model is:

$$
\begin{aligned}
f(x) = {} & -12.27683 + 0.002120737631986 * Age_{CubedDecades} \\
& + -1.01193908416263 * Gender_{Female} + 0.019560817190605 \\
& * Interaction_{AgeGender} + 0.038705190440557 * BMI \\
& + 0.044485448003914 * Height - 0.449706014329128 \\
& * Morbidity_{Hypertension} - 0.67676466318835 * Morbidity_{Diabetes} \\
& - 0.503494449575835 * Morbidity_{CoronaryArteryDisease} \\
& - 0.218267179280336 * ECG_{Abnormal} + 0.295309761910202 \\
& * Morbidity_{TricuspidValvularDisease} - 0.388837016630487 \\
& * Medication_{Aldosterone} + 0.485159907600036 * Medication_{BetaBlockers} \\
& - 1.29272860983173 * Medication_{P2Y12} + 0.02916503904998 \\
& * LA_{volume} + 0.012977946947688 * LVESD
\end{aligned}
$$

Where $Age_{CubedDecades}$ is the patient age divided by 10 and cubed, $Interaction_{AgeGender}$ the result of age in year times gender, $ECG_{Abnormal}$ is if there are signs of infarction, hypertrophy, or ischemia on ECG, LA_{volume} (left atrial) was measured in cm^3 and LVESD (left ventricular end systolic diameter) in mm, morbidity variables are different conditions reported as ICD-10, medications were identified in patient recruitment. Categorical variables were transformed into one or zero values indicating if the condition is present or not.

The analysis was done using Stata v15 (239). The known coefficients for the previously created model were placed into a polynomial formula, and the new dataset was loaded to conform to the term names. Model performance was assessed using AUCROC.

3.6.3 Results and discussion

The total number of patients with complete data collected were 4137, out of these 27 had atrial fibrillation. A summary description is provided in Table 12, the main

differences between the patients with and without AF identified in this cohort are the age, height, presence of hypertension and the prescription of beta-blockers.

Table 12: Summary description for UK Biobank participants identified in the CATCH-ME validation. (a) Continuous variables with a normal distribution are summarized as mean (standard deviation), (b) Continuous variables which were not normally distributed are summarized as median (IQR), (c) Categorical variables are reported as number of cases (%).

Variable	**Total (n=4137)**	**AF (n=27)**	**No AF (n=4110)**	**p-value**
Age, years (b)	56 (12)	62 (8)	56 (12)	<0.001
Gender, female (c)	2181 (52.7)	6 (22.2)	2175 (53.0)	0.003
BMI, kg/m^2 (b)	26.2 (5.4)	28.3 (4.7)	26.2 (5.4)	0.019
Height, cm (a)	169.4 (4.3)	177 (7.6)	169.3 (9.2)	<0.001
Hypertension (c)	1193 (28.8)	16 (59.3)	1177(28.6)	0.001
Diastolic Blood Pressure, mmHg (a)	81.3 (9.9)	82.4 (10.1)	81.3 (9.9)	0.59
Systolic Blood Pressure mmHg (a)	135.3 (17.7)	140.8 (20.1)	135.3 (17.7)	0.168
HbA1c, mmol/mol (b)	34.7 (4.8)	36 (7)	34.7 (4.8)	0.192
Diabetes (c)	188 (4.5)	4 (14.8)	184 (4.5)	0.035
CABG (c)	33 (0.8)	2 (7.4)	31(0.8)	0.005
Myocardial Infarction (c)	48 (1.2)	2 (7.4)	46 (1.1)	0.032
Tricuspid Valve Disease (c)	0 (0)	0 (0)	0(0)	-
Coronary Artery Disease (c)	87 (2.1)	2 (7.4)	85 (2.1)	0.21
Left Atrial Volume, mm3 (b)	65 (26.35)	77.0 (32.93)	65.0 (26.45)	0.005
LVESD, mm (b)	28.5 (5.83)	29.48 (5.98)	28.50 (5.84)	0.066
ECG Parameters				
Signs of old infarction on ECG (c)	394 (9.5)	5 (18.5)	389 (9.5)	0.205
Signs of acute ischemia on ECG (c)	167 (4.0)	3 (11.1)	164 (4.0)	0.167
Left Ventricular Hypertrophy (c)	151 (3.6)	0 (0)	151 (3.7)	0.617
Medication				
Aldosterone-antagonists (c)	8 (0.2)	0 (0)	8 (0.2)	1
Beta-blockers (c)	313 (7.6)	12 (44.4)	301 (7.3)	<0.001
P2Y12_blockers (c)	64 (1.6)	1 (3.7)	63 (1.5)	0.898

The validation model resulted in an AUCROC of 0.71, 95%-CI 0.60-0.81, while the original discovery set had an AUCROC of 0.78, CI 0.76-0.80 (unpublished data).

Despite the number of patients in the validation, it is shown that there is predictive power in this validation, with its predictive scores' confidence interval crossing the confidence interval of the initial model creation. To get a more relevant and conclusive result, a larger number of patients is required, which will be possible when more MRI data is collected by the UK Biobank. Further to the imaging data, it is expected with the added risk of AF by higher age, more patients will have developed AF, and the model can be further explored. This shows the model can be further explored and be used in predictive scenarios.

Figure 20 displays a histogram of AF patients compared to the recruitment date. The UK Biobank targeted to recruit patients between 40-69 years old. The mean age at recruitment in the UK Biobank is circa 56 years. Due to the increased risk with age, the number of patients with AF increases over time as the patients get older. The number of patients with atrial fibrillation diagnosed before recruitment is 6390, and the number of patients with atrial fibrillation after the first year is 7622. As of February 2018, 22160 of the total number of 502,616 patients had atrial fibrillation.

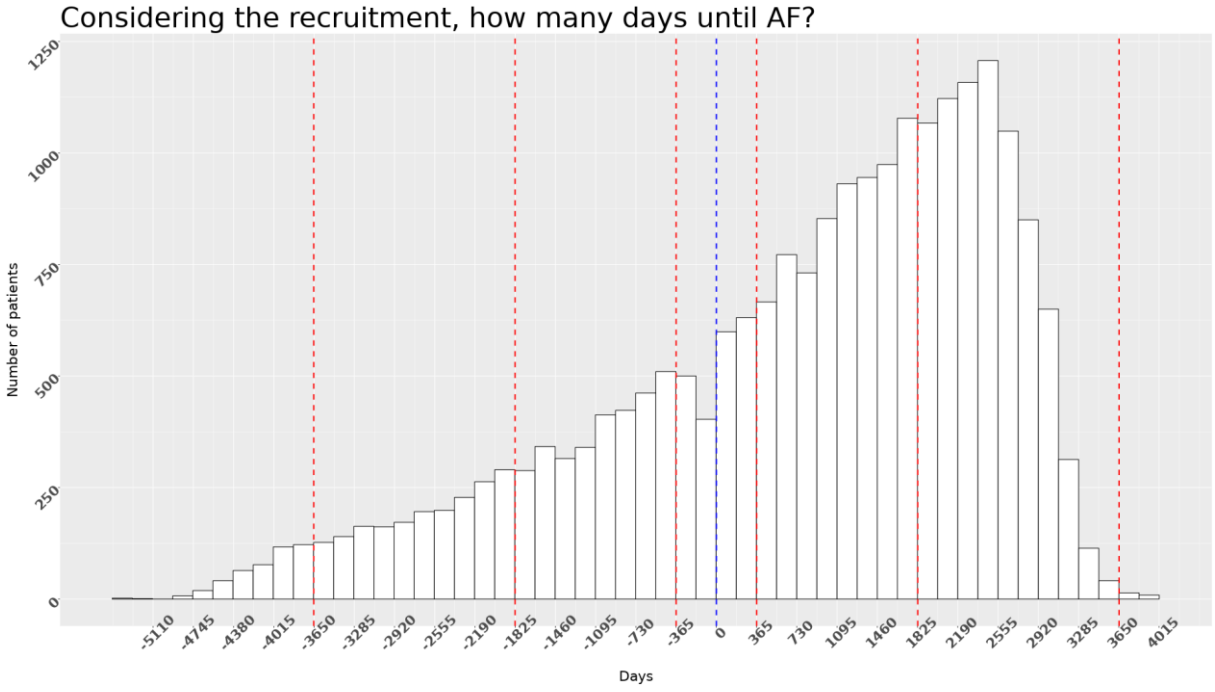


Figure 20: UK Biobank distribution of atrial fibrillation occurrence related to the date of recruitment. The blue line indicates the reference recruitment date, red lines indicate a period of 1, 5 and 10 years before and after recruitment.

3.7 Chapter summary

In this chapter, we explored structured datasets applied to the creation of models evaluating a patient's risk of developing AF.

Previous studies demonstrated the association of AF with other cardiovascular diseases, such as congestive heart failure, valve disease and other conditions such as hypertension, diabetes and age using a follow-up study to evaluate these conditions as independent risk factors (240).

The BBCAF machine learning pipeline allows for a broad search for optimal results in a broad range of datasets tests. It allows for the flexibility of adding a new structured dataset, either following a close approach to methodological approaches used before or changing the settings for different scenarios. The BBCAF machine learning pipeline provided a streamlined structure for the evaluation of different models. Age was repeatedly re-identified in models as one of the major risk factors of AF. Increased BMI and male sex are shown in different cases as relevant markers for atrial fibrillation risk.

Biomarkers BNP, FGF23 and TRAIL-R2 were shown to be of marked importance in the first case explored. Biomarkers ANG2, BMP10, BNP and FGF23 have been shown as relevant for the stratification of AF risk. These biomarkers are targets for improved predictive power and the application of newer models for improved patient stratification. Its applications improve on the availability of different approaches to determine AF patients.

In conclusion, structured datasets are shown as a powerful format for the investigation of different clinical outcomes in the diverse context of variables. Models can be built upon other models, and data can be enhanced for exploration on a wide range of variables in different models.

CHAPTER 4 OMICS: TRANSCRIPTOMICS AND GENOMICS ANALYSIS FOR IMPROVED AF PATIENT-DISEASE STRATIFICATION

4.1 Introduction

Omics datasets, are a type of structured datasets, containing a wide range of information, usually on a limited number of samples (241). There are some exceptional large-scale studies, such as the UK Biobank, which contain a wide range of omics data, such as genotyping, whole-genome, and metabolomics for hundreds of thousands of participants linked to various pathologies.

Data collected from these samples contain several features measured by a variety of biological protocols including array assays and sequencing. Of the types of data available, omics-based biological datasets can exist as proteomics, transcriptomics, metabolomics and genomics. Proteomics study the set of proteins in a cell or organisms, and these can be measured with protein microarrays. Transcriptomics studies the transcripts expressed in a biological material, sourced from a cell, tissue or organ, with these measured with RNA Sequencing (RNA-Seq) techniques, yielding information that has at least fifty thousand transcripts/features in mice, and in the case of the human genome, more than eighty thousand transcripts (242). Genomics indicates variables collected from genome-wide genotyping data, containing data on a much larger scale, e.g. Affymetrix UK Biobank Axiom® array measured circa eight hundred thousand variants and imputed the data to over ninety million single nucleotide polymorphisms (SNP) (81). Metabolomics study metabolites, small molecule substrates that are products or intermediaries of cell metabolism (243). These molecules change all the time, and using this data there is another way of seeing the functioning of the system/body. In the UK Biobank, the metabolomics biomarkers are being collected using nuclear magnetic resonance.

It is important to develop an approach that integrates the various omics with other clinical feature, a multi-omics data integration. This chapter explores analyses of RNA-Seq analysis from mouse and human tissue to further understand gene expression

patterns in AF patients. Followed by Genome-Wide Association Studies (GWAS) investigations in the UK Biobank.

The use of RNA-Seq and GWAS can potentially aid the understanding the pathobiology and pathophysiology of AF. Despite both investigating genetic-related outcomes, they assess different biological patterns.

RNA-Seq explores functional genetics, this technique assesses the expression of biologically active regions that contain proteins and other long/non-coding RNA. It has the potential of identifying how much of a protein, or how much of any mRNA is being produced from the sample material biological function. RNA-Seq allows the exploration of specific tissues, such as evaluating the expression on heart tissues, left and right atria and how much they differ. The use of RNA-Seq also enables the use of transgenic mice to identify signals that surpass species. That is, rather than depending on limited samples collected during open heart surgery, homologous genes behaviour can be explored in different scenarios. This includes the further understand of pathways when partially or fully disabling a specific gene.

On the other side, GWAS measures the genetic aspect of the population, without quantifying the expression. By measuring over hundreds of thousands of SNPs, GWAS is capable of identifying signals that transcend ethnicity and populations. This can be applied to the identification of novel targets associated with a disease, such as the case of atrial fibrillation, where PITX2 was identified to be was strongly associated with atrial fibrillation (264).

In summary, RNA-Seq and GWAS provide views on different scales (section 1.1.2). RNA-Seq enables the investigation of novel targets, with understanding of biological pathways (sections 4.3 and 4.4), while GWAS enables the broader identification of targets associated with genetic predisposition (section 4.6).

4.2 Analytical framework

Before going into different scenarios explored, this section describes the analytical framework.

4.2.1 RNA Sequencing

The procedure for RNA-Seq analysis requires different steps, from *in-vivo* to *in-silica*. The process is formed of experimental design, mice training or patient recruitment, sample collection, sequencing procedures, and other laboratory experiments.

Experimental design plans the experimental setup, hypothesis to protocol, from the number of samples to any process from the start of the experiment to sample collection and analysis to be performed. The hypotheses explored in the different studies are from different literature evidence, described in each subsection. The protocol for the mice training also depends on the hypothesis explored, e.g., mouse sample may be swim-trained or fed a different diet. For reasons of statistical relevance, it is necessary to have at least 3 paired samples, although this depends largely on the analysis protocol (244). Different laboratory experiments were conducted in the samples, and then samples were sequenced.

The sequencing procedure involves the preparation of the library, which goes through the sequencer. The sequencer utilised for the samples is the Illumina NextSeq 500 (245). It outputs digital files with RNA sequences with different markers for each sample. The files can be directly transformed into different *FASTQ* files for each sample (246).

The acquisition of the different biological samples was done by other personnel. The involvement of this project is in the *in-silica* stage. After completing the biological analysis protocol, sample material was sent, and data files were received from collaborators in the Institute of Human Genetics, University of Münster. Sample information was compiled from internal sources and matched with the sample files.

The RNA-Seq analysis involves different steps: quality assessment of the samples, filtering of sequences, transformations of data files, alignment of reads into a reference genome, counting of the aligned reads, and differential expression evaluation.

The first step of quality assessment is operated manually and is usually decided on qualitative analyses. In the sample collection stage (*in-vivo*), metrics such as RNA Integrity Number (RIN – values range from 1 to 10, fully degraded to intact samples,

respectively) indicates the quality of the samples, however, this metric does not suffice as a clear indication of inclusion criteria (247). In the literature, samples with RIN as low as 4 were capable of picking differences between samples (248). On the computational side, there are a few checks that can be performed: FastQC (249) provides a good set of tools to check the samples, if the metrics are very abnormal the sample is removed, other cases the sample is kept to increase the analysis numbers.

After assessing the quality of the samples, the data is aligned. The alignment ratio shows the percentage of reads that are aligned to the reference genome and indicates the quality of the samples. Samples with a low alignment ratio indicate contamination or even a wrong reference genome. After alignment, the transcripts are counted and are processed using differential expression analysis. In the differential expression analysis, the data can be further checked for quality issues, this involves running methods such as PCA to evaluate if the samples are off the norm for the condition type, mislabelled, or if the data is skewed.

Different tools were utilised to accomplish the steps above: data quality was checked utilizing FastQC (249), Trimmomatic was utilized when required to cut lower portion/quality of the reads (250), Samtools was utilised to transform the data as required (251), HTSeq is a counter tool for finalising the quantification of the values (252), alignment was done utilizing hisat2 (253), and processing of transcript counts and differential expression between the groups was done in R utilizing DESeq 2 (196) (254). The mapping from a reference genome to gene symbols was done using BioTools (255).

The reference genomes used on this study for mice and human, respectively, were, *Mus Musculus* (GRCm38.p6) and *Homo sapiens* (GRCh38), both genome assemblies releases of the Genome Reference Consortium (256). Furthermore, two main reference annotations were applied: UCSC and ENSEMBL (257) (258).

4.2.2 RNA Sequencing automation

The different analysis using RNA-Seq data required to perform a set of common operations using different parameters and reference organisms. The main operations,

as described above, are the filtering of reads, alignment to a reference genome, counting of the aligned transcripts and then differential expression comparison using different sets.

To perform these operations in different scenarios and samples, an RNA-Seq pipeline was implemented (259). It contains functions to download and prepare the reference genome. When executing the RNA-Seq pipeline, it will use the reference genome specified, and generate intermediary files for the trimming, alignment and counting steps.

The differential expression step requires the manual setting of sample values and the definition of case-control samples. Output values are the differential expression spreadsheet, volcano plots, and normalized sample values. These generated files contain the results of RNA-Seq analysis, which can be further explored for enrichment analysis (260), or visualisation of samples using PCA.

4.2.3 Genome-Wide Association Studies

Bush & Moore (261) described the GWAS workflow, used as base for this study. SNPs are base-pair changes in the genome. It could be a base-pair that was changed, deleted, or added. These changes may happen either in coding or non-coding sequences of genes. Some of these changes do not necessarily affect the expression of proteins due to genetic code degeneracy. It is considered that very infrequent changes are mutations, whilst common changes (at least 1% of the population) are SNPs.

GWAS data collection involves the application of enriched DNA material into SNP-arrays. SNP-arrays contain probes that attach to different sequences. Then the sequences attached are quantified to the identification of variants.

Association analysis can be performed using Plink (262). It uses a set of statistical tests, Fisher’s exact test and variants, and it has further generalized linear and logistic regression models that allow correcting for interactions.

The number of results when applying these methods is long due to the number of SNPs. Methods such as Benjamini-Hochberg for p-value correction may be used (263).

A standard way of reporting GWAS analysis is through the use of a Manhattan plot. It is a scatter plot, on the x-axis the chromosomes and the y-axis the p-value. There are usually dashed horizontal lines on significant adjusted p-values and auxiliary for extreme significance values.

4.3 Murine RNA-Seq Heterozygous PITX2 – Case 1

4.3.1 Introduction

Initial GWAS explorations in the Icelandic population then validated in European and Chinese populations indicated a strong association between risk of AF and variants on locus 4q25, adjacent to *PITX2* (264). *PITX2* is a homeobox gene (a gene that regulates development in multicellular organisms). During the embryonic stage, *PITX2* regulates left-right asymmetry in the heart and other organs (265). In the human heart, its expression is dominant to the left side, with its isoform *PITX2c* being most highly expressed (56) (266). In mice, reduced *Pitx2* or *Pitx2c* leads to a predisposition to AF without marked structural changes (56).

It is not well understood the underlying behaviour that led to the significance of *PITX2c* in atrial fibrillation. In this case we explored mice samples to explore if there are more patterns to be learned.

4.3.2 Data description and analysis

Biological materials were collected from 12 mice from the MF1 strain at 12 weeks old. 8 male and 4 female, samples paired between wild type and genetically modified heterozygous (+/-) *Pitx2c* knockout (56). These samples weren't challenged in addition to the genetic deletion. Samples were collected for both left and right atria. In total there were 24 data samples.

The analysis was performed using the RNA-Seq framework described in section 4.2.1 using the UCSC GRCm38/mm10 reference genome (257).

4.3.3 Results and discussion

Analysis indicated differentially expressed transcripts between the subgroups of *Pitx2c* +/- and the wildtype mice left atrium. Figure 21 shows the differentiation between the samples.

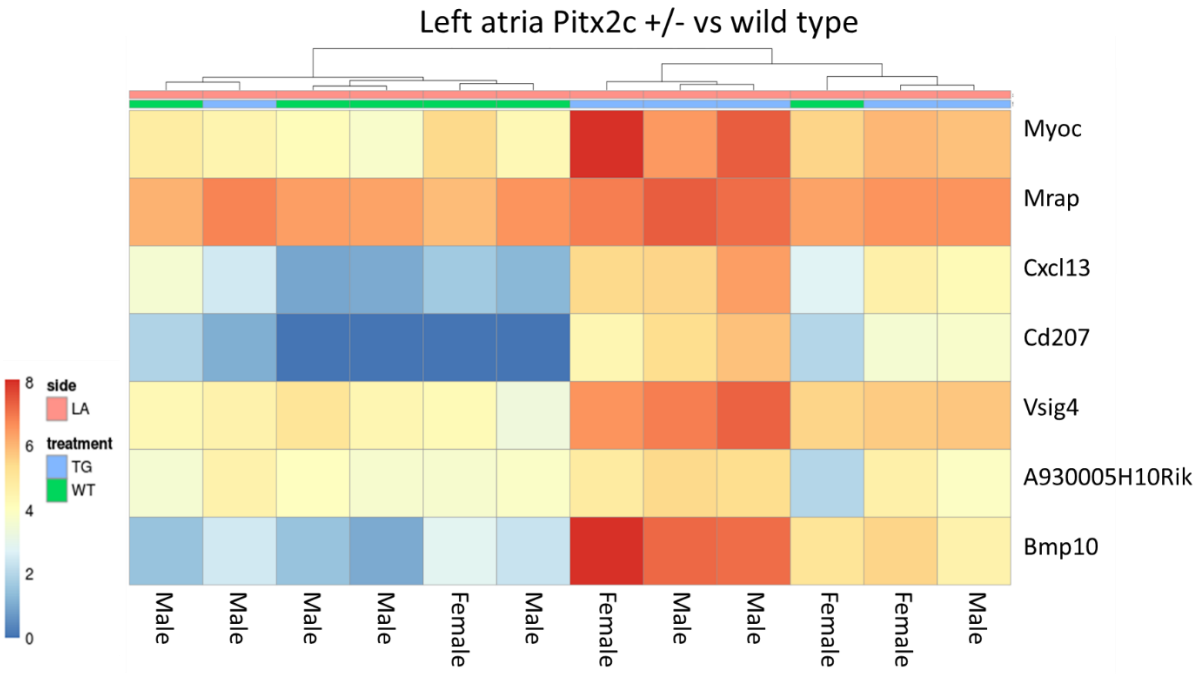


Figure 21: Heatmap comparing the wildtype to the transgenic Pitx2c mice. The transcripts selected are the ones that showed a differential expression between the wild type and the transgenic mice on the left atrium. Blue indicates reduced expression, and red indicates increased expression. The samples were clustered by similarity.

All the samples compared showed a trending profile towards increased expression of some transcripts on the differential expression analysis. Table 13 exposes these differences.

Table 13: Differential expression comparing wild type and Pitx2c +/-. Log2 fold-change indicates how much more expression was identified in the Pitx2c +/- case. P-values were adjusted using the Benjamini-Hochberg approach (263).

Gene	log2 fold-change	p-adjusted
Cd207	4.446945505	9.11E-05
Cxcl13	3.280999901	0.000168927
Bmp10	3.939048513	0.000289024
Myoc	2.02975104	0.010565334
Vsig4	1.720448913	0.010565334
A930005H10Rik	1.37425925	0.013549833
Mrap	0.628376467	0.047386574

Cd207 is a protein-encoding gene. Its upregulated expression was identified in epicardial adipose samples from patients that developed postoperative atrial fibrillation, suggesting a pre-existing inflammatory state of epicardial adipose tissue (267). *Cxcl13* was shown to be involved with the cardiac remodelling in patients with heart failure (268). *Bmp10* is a key marker, as it plays a key role in murine cardiogenesis, with perturbations leading to different heart diseases (269). There is not much information in the literature about *Myoc* and cardiovascular diseases, its homonymous protein is associated with skeleton structural changes and glaucoma (270). *Mrap* gene is associated with glucocorticoid deficiency (271). *Vsig4* is implicated in inflammation (272). *A930005H10Rik* is a long non-coding RNA.

Out of the measured biomarkers, C-X-C motif chemokine 13 (*CXCL13*) and (Bone morphogenetic protein 19 (*BMP10*) proteins are soluble, and thus are biomarkers that are of increased importance for the assessment of atrial fibrillation through blood samples. Reyat et al. 2020 demonstrate a case for *BMP10* (8).

4.4 Murine RNA-Seq Jup +/- – Case 2

4.4.1 Introduction

Arrhythmogenic cardiomyopathy (AC) increases the risk of different life-threatening conditions, inclusive of arrhythmias. It was identified that mutations in genes encoding desmosomes are predominant factors to patients developing AC (273).

Desmosomes are involved in the intercellular junctions of cardiac muscle and they are also involved in embryonic development with alternating adhesive affinity status (274).

One of the first genes identified from the desmosomes causative of AC is *JUP*, which encodes the junction plakoglobin protein. Deletion of *Jup* in mice offered the first model to investigate AC.

This led to the experimentation design with this model organism. The model organism is a heterozygous knockout of the *Jup* gene, due to recessive homozygous leading to embryonic development problems. Samples were paired by sex, then were randomly allocated into endurance swim-training or sedentary lifestyle. The protocol consisted of 8 weeks of swim training. Left and right atrial samples were collected and sequenced.

4.4.2 Data description and analysis

The dataset contains 12 mice from the 129/Sv train (275), 6 mice of each sex, each sex with 3 wildtypes and 3 heterozygous *Jup* knockout, with samples available for left and right atria. In total there were 24 samples. Some mice had AF identified, others not. The average RIN value was 8.65, the lowest value was 7.2. Table 14 summarises the samples.

Table 14: Description of Plakoglobin experiment samples. Transgenic mice are heterozygous knockout of the *JUP* gene. Arrhythmia indicates atrial arrhythmia.

	Left Atrium		**Right Atrium**		
	Female	Male	Female	Male	Total
Transgenic					
No arrhythmia		2		2	4
Arrhythmia	2	2	2	2	8
Wild type					
No arrhythmia	2	4	2	4	12

Data analysis were performed using the RNA-seq framework described above, using ENSEMBL Mus musculus GRCm38.91 reference genome (258). Left atrial samples from wildtype and transgenic mice were compared in subsequent analyses. Differential expression formula corrected for the sequencing batch.

4.4.3 Results and discussion

Swim-training led to more atrial arrhythmias in Jup hearts only. RNA-Seq analysis confirmed that *Jup* is significantly different between the genotype groups. Transcripts that indicate genes Ankyrin repeat domain 2 (*Ankrd2*) and Actin alpha 1 skeletal muscle (*Acta1*) had reduced expression in the transgenic groups, although no significant difference is shown (Table 15).

Table 15: RNA-Seq results for Plakoglobin analysis.

Transcript	**Gene Symbol**	**log2-FoldChange**	**p-value adjusted**
ENSMUSG00000001552	Jup	0.855755706	0.004050806
ENSMUSG00000025172	Ankrd2	-1.793113321	0.436866578
ENSMUSG00000031972	Acta1	-2.319484251	0.783772261

Ankrd2 protein is associated with stress response, it is not essential to cardiovascular development, although there are associations between its absence and conditions such as hypertrophic and dilated cardiomyopathies (276) (277). *Acta1* is also associated with dilated cardiomyopathy (278). The link between AF and hypertrophic cardiomyopathies (HCM) is not fully understood, despite AF being a common complication of HCM (279) (280).

4.5 Human RNA-Seq – Case 3

4.5.1 Introduction

Model organisms can be used to understand the interactions and effects when altering specific genomic regions of those samples. It is an essential tool to explore the dynamics when specific genes are either silenced or enhanced. These alterations provide insights on pathways, facilitating the targeting of specific markers. Knowing pathways and differences associated with a marker can assist the understanding of a disease, however, this does not provide the whole picture.

The human genetic material is a complex structure. Its interactions and dynamics are not completely understood. Ancestry information can provide some perspective on risks inherited to a patient, subtle mutations over generations make the problem even more complicated.

Analyses of patient data provide an understanding of the effects of a patient and its complex dynamics under clinical observation. It is also the analysis that generates new targets for further investigations in model organisms.

Previous studies showed a strong association between AF and variants on loci 4q25, adjacent to *PITX2* (264). PITX2 regulates the asymmetry between the left and right side of multiple organs, inclusive of the heart (265). It is of paramount importance the evaluation of the differential expression between atria, to further understand the differences and any mechanism that might assist in understanding AF. Further to the marked difference between atria, and due to a marked difference in AF between males and females (281), gender differences were also evaluated.

4.5.2 Data description and analysis

Experimental data were collected from consented patients in the CATCH-ME project. Patient data contain a comprehensive clinical view of the patient, included of comorbidity history, cardiovascular operations, and lifestyle including physical activity (the complete description of variables is shown in 3.6.2).

The tissue samples were collected from consented patients undergoing heart surgery, e.g., bypass surgery, with material coming from the left and right atrial appendages. The material was sourced from different institutes in Europe and sequenced in the University of Münster, Germany. Due to this and other factors, the data quality is heterogeneous.

Table 16 indicates the profile of the set of patients that underwent RNA-Sequencing and had data available. Most patients had samples from either the left or the right atrium, 74 and 39, respectively, 17 patients had samples from both sides. The relation between sex and atrium samples available is shown in Table 17.

Table 16: Summary of Human RNA-Seq dataset. Smoker indicates if the patient has a smoking history. Coronary artery disease (CAD), myocardial infarction (MYOC), heart failure (HF), diabetes (DIA), chronic kidney disease (CKD), stroke, and transient ischaemic attack (TIA) indicate the patient morbidity history.

		Heart Rhythm Paroxysmal AF	Permanent AF	Persistent AF	Sinus Rhythm	Missing	Total
Sex	Female	13	1	9	12		35
	Male	22	2	18	52		94
	Missing					1	1
Smoker	Yes	9	2	8	26		45
	No	7	1	9	16		33
	Missing	19		10	22	1	52
CAD	Yes	20	1	13	36		70
	No	15	2	14	26		57
	Missing				2	1	3
MYOC	Yes	7		5	20		32
	No	26	3	21	41		91
	Missing	2		1	3	1	7
HF	Yes	5		12	25		42
	No	30	3	14	35		82
	Missing			1	4	1	6
DIA	Yes	2	1	8	17		28
	No	33	2	19	44		98
	Missing				3	1	4
CKD	Yes	2		1	5		8
	No	29	3	24	57		113
	Missing	4		2	2	1	9
Stroke	Yes	3		4	5		12
	No	32	3	23	58	1	117
	Missing				1		1
TIA	Yes	1		1	4		6
	No	34	3	26	59	1	123
	Missing				1		1
	Total	**35**	**3**	**27**	**64**	**1**	**130**

Samples were sequenced in batches; some samples were replicated in different sequencing batches. These samples are considered technical replicates. The differential expression design formula corrects the batch and the sex of the samples. Participants with missing sex data were ignored.

Differential expression results with adjusted p-values below 0.05 were grouped into more highly expressed transcripts in the left atrium and right atrium. Top 3000 hits were selected to run Metascape gene analysis (282).

Table 17: RNA-Seq data available for the human dataset.

	Right Atrium	Left Atrium	Total
Male	40	68	108
Female	16	22	38
Missing		1	1
Total	56	91	147

These RNA-Seq data were processed using reference genome ENSEMBL Homo sapiens GRC38.93.

4.5.3 Results and discussion

Left and right atria

The overall comparison led to 11,403 significant transcripts, inclusive of PITX2 as the most significantly differentially expressed, more highly expressed in the left atrium. A list of 122 highly differential expression (absolute log2FoldChange above 2) significant transcripts is in Appendix 4.1. A representation of the differential expressed transcripts is shown in Figure 22.

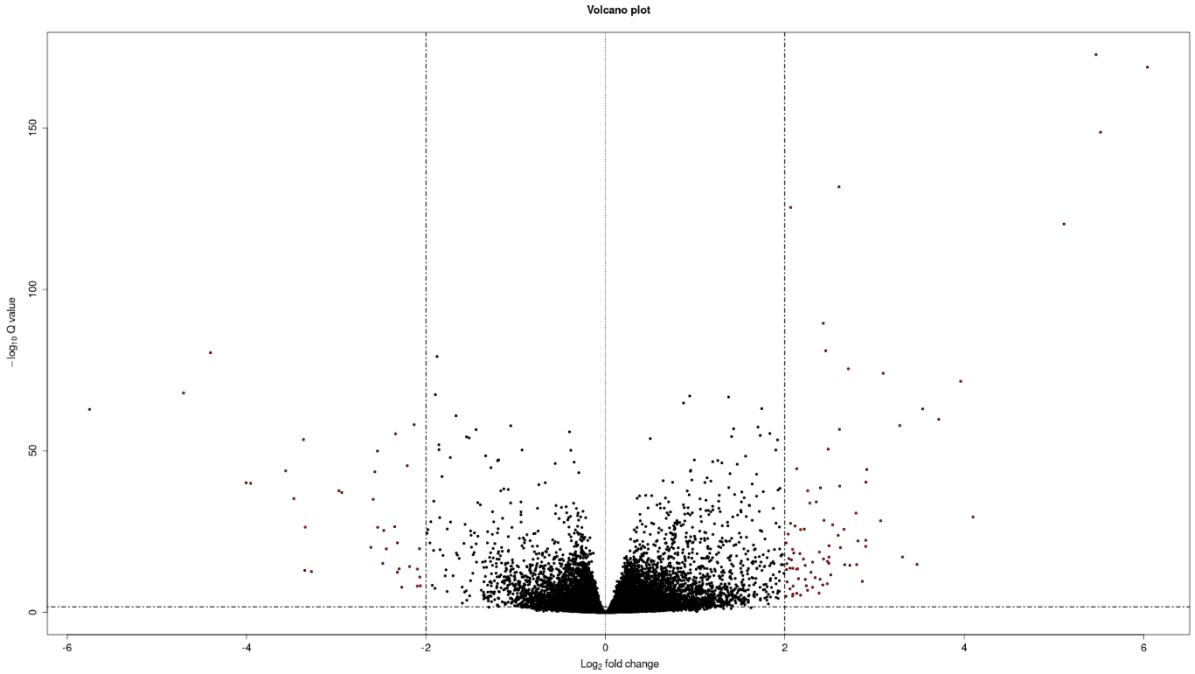


Figure 22: Volcano plot for the Human RNA-Seq LAxRA. The horizontal line indicates significance of values. The vertical lines on -2 and 2 indicate values that have a very significant differential expression.

Gene enrichment analysis on more highly left atrium expressed transcripts indicated a few pathways (shown in Table 18).

Actin cytoskeleton organization pathway is associated with immune pathology (283), and it is associated with changes in motor proteins linked to cardiomyocyte electrophysiology. Extracellular structure organization is linked to changes in cardiomyocytes size. Cell junction organization and response to wounding are implicated from AF.

Table 18: Relevant pathways identified in more highly expressed left atrium transcripts.

GO identifier	Description
GO:0034330	cell junction organization
GO:0009611	response to wounding
GO:0030036	actin cytoskeleton organization
GO:0043062	extracellular structure organization

4.6 Sampled GWAS of Atrial Fibrillation – Case 4

4.6.1 Introduction

For a set of AF patients, it is unclear what their AF origins are. First and foremost, in some datasets, such as the UK Biobank (81), the definition of AF severity is unclear, with most patients defined under the broad WHO ICD-10 term I48 Atrial fibrillation and flutter (24). Furthermore, it is suggested that AF patients could be categorized according to their AF mechanistic origins (35). Although there are signs that might indicate an origin for AF, e.g., clinical history and biochemistry results, there is no clear indication that a patient had AF due to genomic predisposition.

In this analysis, it is assumed that the different mechanistic origins for AF are valid, and that, for a subgroup of patients identified with AF, there must be at least a considerable number of participants with a genomic predisposition.

In a contrast to standard analysis comparing overall sample population, i.e. all cases compared against all controls (284), it is hypothesized that when analysing sub-samples of these participants it is possible to probe sub-groups, which would have a marked genomic difference, and exposing a stronger signal. These signals would then indicate new genomic targets.

4.6.2 Data description and analysis

This analysis used the UK Biobank dataset, inclusive of patient comorbidity history and genotyped data (81). Participants were identified using the Hospital Episode Statistics set, resulting in 21,486 AF cases, and 466,891 controls. Samples were also matched with their age, sex, and ancestry information, and these data were used as correction terms.

Table 19: Samples description for the sampled GWAS analysis on AF participants in the UK Biobank. There were 21,486 AF cases and 466,891 controls (total of 488,377 participants identified).

	Cases		Controls	
Variable	**Number**	**Percentage**	**Number**	**Percentage**
Age Recruitment (Median)	64		57	
Atrial Fibrillation	21486	100.0%	0	0.0%
Coronary Artery Disease	8141	37.9%	33862	7.3%
Chronic Kidney Disease	2287	10.6%	7665	1.6%
Diabetes	3789	17.6%	26365	5.6%
Heart Failure	4120	19.2%	4474	1.0%
Hypertension	13467	62.7%	96214	20.6%
Ischaemic Heart Disease	8141	37.9%	33862	7.3%
Peripheral Vascular Disease	939	4.4%	2861	0.6%
Pulmonary Embolism	760	3.5%	4453	1.0%
Sex (Male)		66.0%		44.0%
Stroke Tia	238	1.1%	7958	1.7%
Valvular heart disease	1866	8.7%	1932	0.4%

In an approach similar to bootstrapping (285), a 10% subset of case and control samples were sampled over each 100 executions. Samples were filtered for similarity to avoid biasing the results (286). GWAS analysis and sample multiple testing correction was performed as described in section 4.2.3.

4.6.3 Results and discussion

The experiments resulted in 1328 SNP variants with p-values adjusted under 0.05. Those variants are in 21 regions on chromosomes 1, 4, 12 and 16, as indicated in Table 20.

Table 20: Significant regions identified on the sampled GWAS analysis on AF participants in the UK Biobank.

rs2723298	rs1375302	rs1448817	rs17042081	rs17042171	rs2200733	rs6843082
rs13141190	rs7193343	rs2106261	rs879324	rs3853445	rs11931959	rs35323363
rs6666258	rs6658392	rs13376333	rs521511	rs883079	rs1895585	rs13124249

All regions identified were reproduced and previously identified in the literature (287).

Some analyses in the literature involve the increased use of bigger datasets and a combination of genomic comparison techniques (288), e.g. standard GWAS, GWAS meta-analysis, gene set enrichment analysis (289), expression quantitative trait loci (eQTL) analysis (290), and phenome-wide association analysis (291). Combining these techniques there is an increase in the analysis power, although limited by the availability of data.

Although a limited number of regions were identified, it is shown that analysis using sub-samples of the population yield targets loci for AF. In future steps, the use of different analysis approaches, or enhanced datasets is suggested for improved results.

4.7 Chapter summary

Since the identification of PITX2 on loci 4q25, the number of variants identified is ever increasing with the wide use of different population data (288). Experiments performed to verify and reinforce some of the results previously obtained, both in the case of GWAS and RNA-seq, the first performed in a well-known dataset and the latter in novel data. The analyses carried out aid towards a better understanding of the different mechanisms that interact with biological targets associated with cardiovascular disease predisposition.

CHAPTER 5 UNSTRUCTURED DATA: ELECTROCARDIOGRAMS

5.1 Introduction

Images carry more meaning than words. And unstructured datasets can carry more information that structured or semi-structured data.

Unstructured datasets contain information about signals that are usually measured in a limited time or space-resolution constraints, e.g., recording of an output signal and images, respectively. These types of data contain embedded information, which can be extracted and expanded into other types of data, especially structured data. They contain information that is not directly interpretable for people without practical knowledge, or to people absent of tools to the identification of features that can then be understood. Some features are repeating patterns in a dataset, abnormality, and artefacts, e.g., time between signal peaks, alterations and inversions of signals, and the presence of recording errors not natural to the signals.

Numerous techniques have been developed to exploit unstructured datasets. These techniques usually target the extraction of features or the direct creation of models for decision-making. In other words, the signals can be plugged into models to support other models with additional variables (unsupervised learning) or to learn the direct use of the information for the prediction of an outcome (supervised learning) (292).

One can think of different ways to handle unstructured datasets. An unstructured dataset can be transformed into a structured one through the use of aggregative functions, such as collecting specific points in intervals, e.g., the first point each day, or the average for each minute of recording. The data can be collected in conjunction with an event, for example, when the patient was under a crisis, or after a certain amount of time as a procedures follow-up. The values can also be collected about values reaching a critical point, such as a peak in heart rate being an important point to the collection of its associated blood biomarkers.

Unstructured datasets can also be directly applied in a model. This is commonly seen in image models. Tags in images can be classified using the VGG-19 model on

ImageNet, a model which is both capable of differentiating outcome classes and creating intermediary variables from an input (293). Further to the identification of outcome classes, models are capable of identifying elements and their boundary boxes on images (294).

In the cardiovascular domain, those techniques are applied to a wide variety of cases (295). There is a wealth of literature on the analysis of electrocardiogram (ECG) signals preceding the use of deep learning, an analysis that use hand-engineered features to explicitly transform the signal into more useful variables. This scenario is exemplified by Lankveld et. al. 2016, in which a range of features is extracted from an ECG recording, and the best features are selected as predictors to the condition explored (296). More modern approaches adopt the direct application of the unstructured data in a deep learning model, for example, extraction of data from ECG recording images (297), identification of cardiovascular movement (298), and the identification of diseases (299).

In this chapter, ECGs are applied and exemplified in the case of unstructured datasets. ECG recordings contain a range of signals from different leads placed in different spots in the chest. The signals are hypothesized to not only contain direct information about different heart chambers and their physical and electrical movement but also underlying information that can provide hints on the understanding of cardiovascular patterns and the evolution of diseases. Two scenarios are investigated: the extraction of features from a study ECG dataset for the identification of atrial fibrillation, and the use of real-world routinely collected ECG data for the early identification of heart failure risk. Further to these two cases, a section explores and expands the optimization of neural networks for the learning of complex models, describing a framework for this process.

5.2 Feature extraction from electrocardiogram-derived images – Case 1

5.2.1 Introduction

An ECG contains information that is easily identifiable by a trained cardiologist. For example, it is possible to identify a patient that suffers from AF using their recording,

the patient will usually have irregular R-R intervals, noisy recording, and/or absent P wave. These visual metrics are directly identifiable by a human and are usually composed of variables formed of linear relationships. These features are well known in the literature, and recording devices normally assess some of these features.

In an approach, closely related to omics analysis described in the previous chapter (radiomics), it is hypothesized that it is possible to automatically extract a wide range of features from an ECG recording, and those variables can provide advanced features for the identification of AF patients.

Similar approaches have been applied to the identification of features from tumour images (300), and ECG recordings (301). It is expected that the extraction of data from multiple beats at once can identify traits associated with inter-beat dynamics and the method is not bounded by the use of special methods for irregular beats (302). Furthermore, the investigation of variables from all leads, rather than just lead II, might provide more information on patient outcomes and insights not related to rhythm. This contrasts to the Rahhal et. al. 2018 (301) study that utilised single-beats on lead II recordings extracted from the MIT PhysioNet dataset (45).

5.2.2 Data and methods

This study utilises the ECG cohort of the BBCAF dataset (section 2.3.1).

The state of the art approach for the prediction of AF using ECG signals, published in early 2019, utilizes convolutional neural networks resulting in an almost perfect stratification of AF patients as well as of patients suffering from several other conditions, exceeding the performance of most clinical practitioners (303). This further justifies the potential of the use of modern machine learning processes for the analysis of ECG signals. The approach proposed differs from this as well as other studies in that, in this use case, electrocardiogram signals are considered as merely another set of variables, that can be used both in combination with other variables as well as independently.

A wavelet transform was applied to the ECG signal and then features were extracted using VGG-19 (from ImageNet) (293). This approach is similar to Rahhal et. al. 2018

(301). The difference is that it employs a larger signal, containing at least 2 complete beats, followed by VGG-19 data transformation. Then features are then collected from the second last dense layer (Figure 23).

While the application of a wavelet transform does not generate variables that can be directly used from the original data, it generates different views on the data structures and patterns. Each wavelet transformation is similar to a single layer of an image (Figure 23 A): a computer image is formed of three channels, red, green and blue, and the output of the proposed models are formed of channels, from the wavelet transformations of Daubechies 4, Coiflet 1 and Biorthogonal (304). To transform the intermediary image-like features to a variable-like format, pre-trained image-capable CNN models were applied, namely, the VGG-19 was used to extract these features (Figure 23 B).

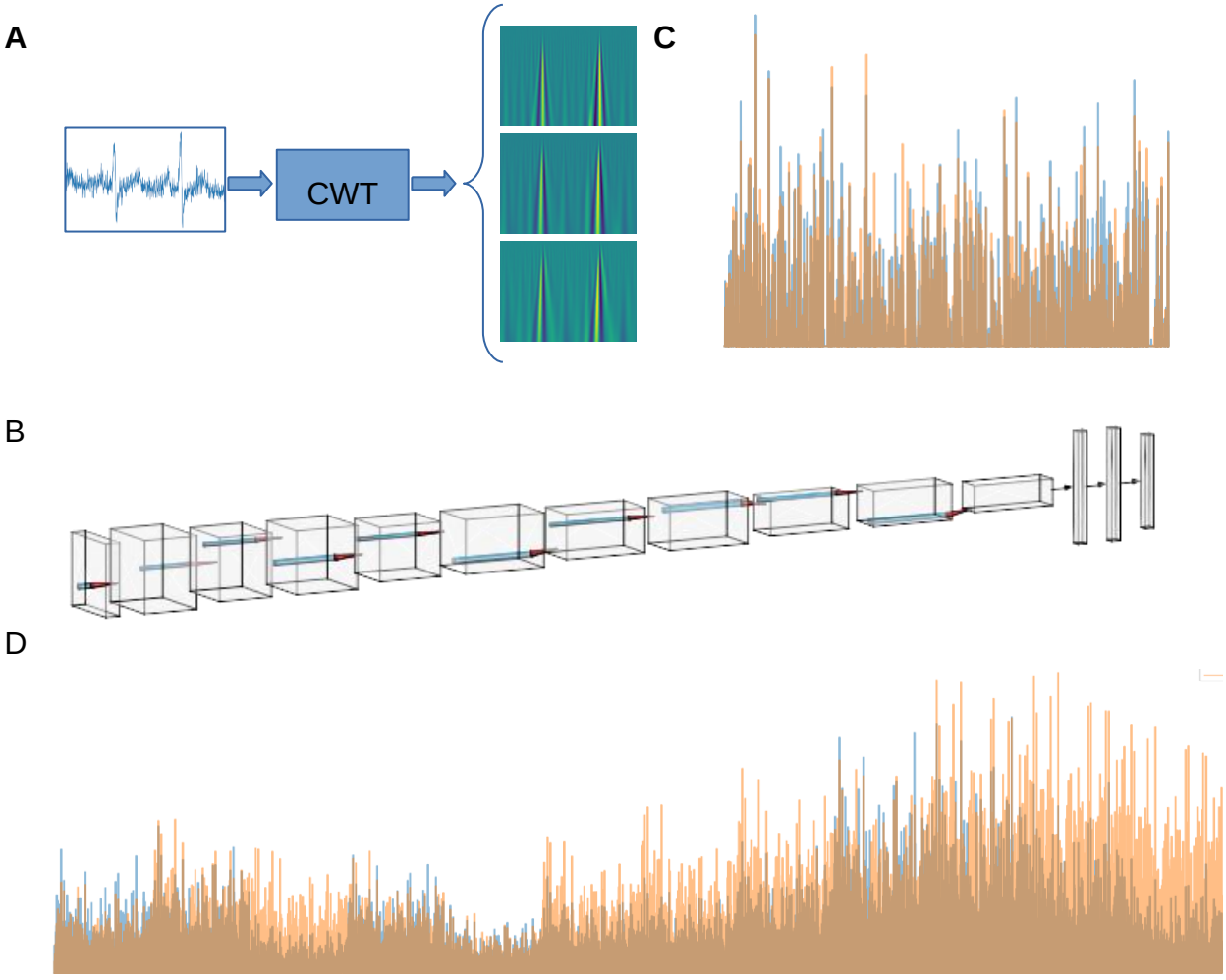


Figure 23: Simplified version of the executed pipeline. **A** is the first step, transforming the signal of each lead using a continuous wavelet transform (CWT). **B** is a summarized version of the neural network applied; in the end the dense layers contain “interpreted features” from the input data. **C** is the set of extracted variables (4096 features) for the AVL lead, in orange variables from a sample sinus patient and in blue variables for a sample atrial fibrillation patient. **D** represents all the features extracted for the 12 leads (12x4096 features). The order of leads shown is AVF, AVL, AVR, I, II, III, V1, V2, V3, V4, V5, V6.

Each patient had a feature vector collected after applying the process, resulting in 49,152 features (12x4,096). These features were treated as a set of variables and applied within the BBCAF machine learning pipeline. Important variables were identified using their relative importance to the created models. Different algorithms were evaluated in an initial step, then these models were further enhanced with the addition of other features from the BBCAF cohort, clinical factors and biomarkers.

5.2.3 Results and Discussion

In the model created with only the ECG-derived variables selected to use only 1,136 variables of the whole set. The best performing model reached an AUCROC performance of 0.78 in the training set, whilst performing with an AUCROC of 0.71 in the test set.

In the model combined with the clinical variables, it is possible to achieve a slightly higher performance value (95% confidence interval 0.5862-0.7876). Despite the similar performance, the created models show new targets from the ECG that can improve prediction.Figure 24 depicts the important variables for the combined model.

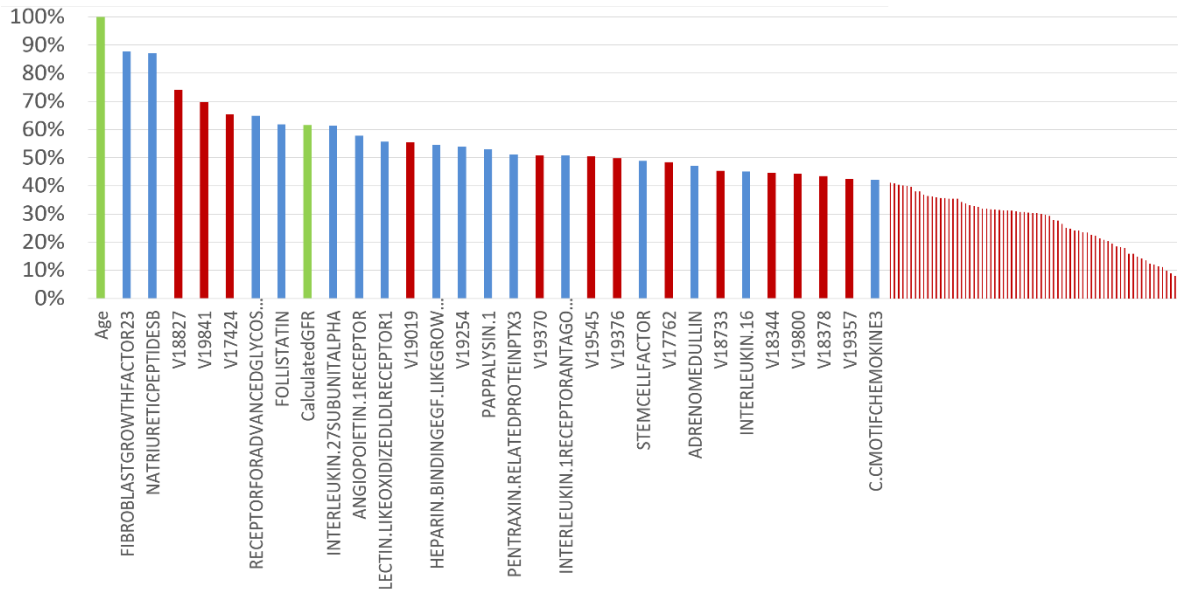


Figure 24: Scaled importance of the different variables of the best performing model using the features extracted from lead II, combined with biomarkers and clinical parameters. In green clinical variables, blue biomarkers and red variables extracted from ECGs.

The resulting performance of those models shows that the models allow a degree of separation between patients. These experiments show the potential of markers derived from a neural network for the assessment of atrial fibrillation risk. Highlighted variables can be treated as biomarkers, and can be handled in a similar way to other omics datasets. These biomarkers could also be considered new targets to be further explored. The use of systematic approaches in higher dimensional datasets could provide insights into ECG mechanisms. Future work should consider investigating the

important features in the recording that lead to a specific variable, and approaches such as SHAP should assist in this process (220).

Related to the application of this approach to this problem, there are other methods for the direct extraction of features. Approaches such as Yildirim et. al. 2018 (305) that a signal is compressed through the use of an autoencoder neural network, and the compressed signal can be decompressed in another device for computation. Despite the capabilities of compression, the compression is not lossless. An application of a similar method to the compression of an ECG signal and its use for the evaluation of patient risk could not converge in experiments realised. Chen et. al. 2018 (306) explored the use of variational autoencoders in the identification of anomalous ECGs.

Also considering the use of NN in ECGs, there are approaches in the literature making use of generative adversarial networks to the synthesis of ECG signals based on other recordings for a patient, performing denoising and selecting features on single beats (307), possibly enabling the creation of a model using a reduced amount of data. Similarly, LSTM approaches are seen in the literature as an alternative to the application of convolutional neural networks (308).

While AF is a condition diagnosed by the explicit indication of abnormalities in an ECG recording (section 1.2.1), it causes pathophysiological changes in the heart, which reinforce these changes seen in recordings. Other heart conditions, despite not being directly diagnosed or predicted using ECGs, may have underlying patterns, that advanced machine learning models could pick. The next sections explore a background knowledge on neural networks, and their application to the identification of heart failure, a disease associated with AF.

5.3 Neural Network optimization

Many aspects affect Neural Networks (NN). There are different types of architectures, layers, functions and settings that have differing effects on the outcome (309).

In section 2.5.5, it was seen that an optimizer iteratively updates the weights/coefficients of the model over epochs and batches. Section 3.4.2 describes a NN optimization approach and the implementation of NN for the BBCAF problem. This

section describes the methodology for the training of advanced neural networks in unstructured datasets, inclusive of different layer types.

The BBCAF model created previously (section 3.4) utilised a range of variables that had no direct structure between them, a dense network sufficed. On the dense network, it was required to select parameters such as the number of nodes and the depth of the network. When working with unstructured data, the network is required to handle the interaction between the different variables, the relationships between the current sampled point and the previous/next ones. This interaction between these signal points is done with convolutional layers, and when applied with other blocks they form building blocks with many different settings.

Further to the use of convolutional and dense layers, the main building blocks for unstructured signal analysis are batch normalization, pooling layers, and addition layers. Batch normalization improves the learning of a NN reducing the covariate shift in the layers, speeding the training whilst reducing generalization error (310). Pooling layers are aggregative functions that combine the output of a previous layer, usually a convolutional layer into min or max operators that will then reduce the dimension of the output (311 347). Addition layers can be used to connect branches of the network that were separated, for example, a network can be separated into two paths, one that goes through some layers and another that bridges directly, then these parts can be merged. When a branch of the network bridges over another branch the formed block is a residual block.

Residual blocks improve models through the abstraction of new features, at the same time using older features on the next steps of the model (312). Residual blocks can be formed in different ways. In experiments realised, the residual block is formed of two paths, starting with an input signal. On a path, the input signal passes through two blocks with convolution and batch normalization, on the other path the signal goes through convolution and max pooling. The values are then combined and output. The size of the kernel and filters can be altered for each of these building blocks, altering how the network will detect small features of a signal. This is usually done from a bigger

to a smaller kernel size, whilst increasing the number of features. Figure 25 illustrates the network architecture used.

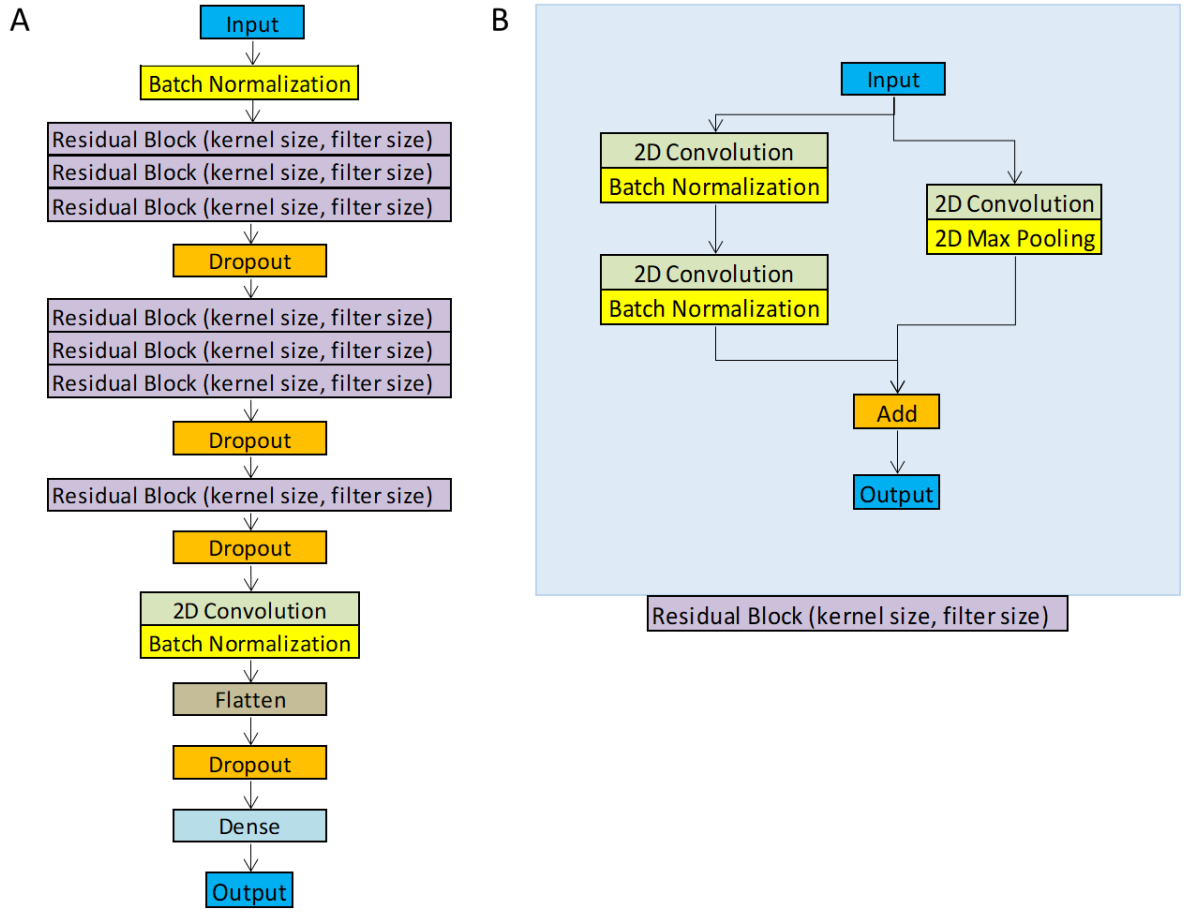


Figure 25: Illustration of the network architecture utilised for the unstructured analysis of electrocardiograms. Activation functions are not shown. Each rectangle indicates a type of layer on the network. (A) shows an overall structure of the network, hyperparameter search tested different numbers of layers in different sections of the network. (B) indicates the structure of a residual block, where there is a path where more computation is performed and another path that bridges data to the other layers.

There is a large number of optimization decisions to make, from the number, order and size of layers; kernel and strides; activation and loss functions, and so on. To obtain a more robust model, different network alternatives are required to be tested.

The framework developed contains three main parts: (1) generation of search settings, composed of different search hyperparameters, (2) model creation with settings generated and (3) compilation of the results with performance assessment.

Generation of search settings goes through alternatives, making sure that the network is compatible with the input signal and output prediction. This involved the selection of input files, such as specific signal and dataset versions, output performance files and model checkpoints for each case number. It also specified the number of epochs, early termination protocols, and whether to execute normal sampling or random sampling of batch samples, all with different settings for layers and number of filters.

Model creation. The creation of models follows a parallel paradigm using the resources available. Slurm is utilized to control the GPU resources, queueing the different search architectures tried (313).

Results compilation. Model creation assessed the samples using binary cross-entropy, weighting the unbalanced state of the dataset. The best performing model as assessed in the validation set had its AUCROC calculated, and confusion matrices were generated for different thresholds over the curve.

5.4 Early prediction of heart failure using electrocardiograms – Case 2

Our analysis of ECG in AF demonstrated that there are patterns that can be explored in a computational fashion. In the literature, the work of Attia et al. 2020 has shown that the ECG itself can be used for the development of models that can predict incident AF (299). The prediction of incident AF is due to the physiological changes that come with the evolution of AF morbidity, i.e., while the patient had no clear ECG of the disease, their heart was already changing. In other cardiovascular morbidities, such as HF, there are also underlying physiological changes in the heart that lead to the evolution of the disease. As such, and as a natural extension to our work and given the association of atrial fibrillation to heart failure, we hypothesised whether standard 12-lead ECG can be used to the early prediction of HF risk. To be able to assess this hypothesis, NN were employed at the UHB dataset. The UHB dataset provides a larger number of ECG recordings, facilitating NN to learn rarer patterns.

5.4.1 Introduction

HF is a complex clinical disease triggered by various causes. It has significant importance due to high treatment costs, rising prevalence and impact on morbidity and mortality. The early identification of factors can help prevent the development of heart failure and reduce costs (32). There are a multitude of models to incident heart failure in the literature considering different predictors for patient risk (314). All models consider age, followed by a substantial number considering blood pressure and sex as well. Some models use the left ventricular hypertrophy score from ECG recordings as well.

5.4.2 Data and methods

This study uses the UHB dataset (section 2.3.4). ECGs were collected between December 2008 and March 2018. The ECGs had either a sampling of 500 Hz or 1 kHz. Clinical information was collected from internal systems up to October 2019. Patient data was collected and matched to the date of ECG recording: age, sex, presence of coronary artery disease, hypertension, diabetes, chronic kidney disease, AF, HF. For a patient with multiple recordings the data is matched up to the date of each recording. Patients were excluded if they had HF diagnosis before the ECG date, when the ECG data was not suitable for analysis, such as invalid lead signal, and in cases where at the end of follow-up the patients had a reported myocarditis, cardiomyopathy or any congenital malformation. Patients were allocated to cases if they had a diagnosis of HF, while patients without a coding were marked as controls (Figure 26).

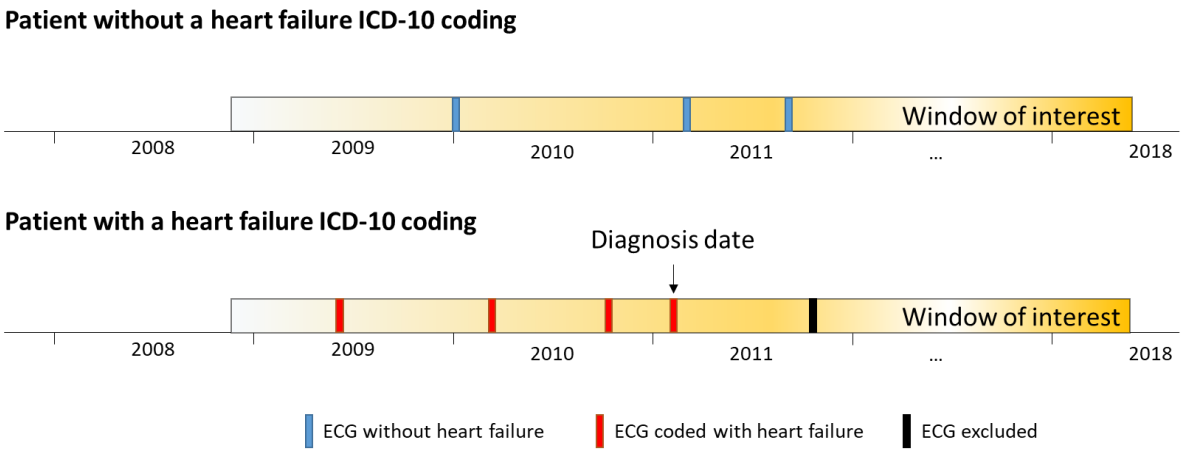


Figure 26: Patient inclusion for the heart failure model using electrocardiograms. On top a patient without a diagnosis of heart failure is considered a control case. On the bottom, a patient with a heart failure coding has all data added as case up to the date of diagnosis, and further ECGs are excluded.

The 500 Hz ECG dataset was used for model training, selection and initial evaluation, with 70%, 15% and 15% splits, respectively. The 1 kHz was down sampled to 500 Hz and used as an external validation set, it was separated into two 50% cohorts. If a patient would be separated into multiple sets it was instead kept on a single set to avoid biasing the models; patients that were both in the 500 Hz and 1 kHz cohort were excluded from the 500 Hz set. After exclusions there were 137,018 valid ECGs (65,565 patients) in the 500 Hz set and 11,886 valid ECGs (9,508 patients) in the 1 kHz set (Figure 27). A description of the population is shown in Table 21.

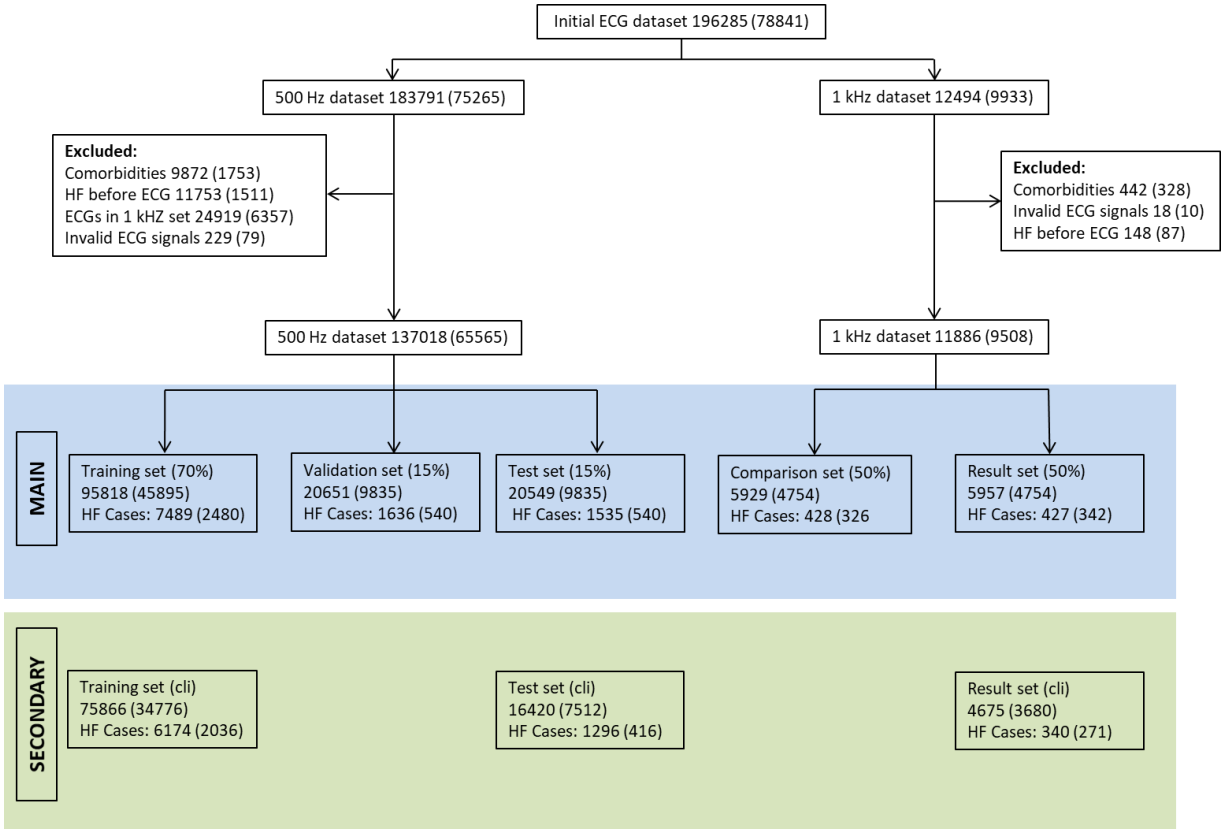


Figure 27: Patient flowchart for the electrocardiogram for the prediction of heart failure case. The numbers indicate the number of ECGs, followed by patients in parenthesis. *cli* indicates the cohort with complete set of clinical variables.

Table 21: Patient description for the cohort used on the electrocardiograms to the prediction of incident heart failure model. The patient information reported in this table comes from the first ECG recording for each patient. [A] is reported as mean (standard deviation) [B] indicates continuous variables, reported as median with inter-quartile range (1st-3rd quartile), [C] are discrete variables reported as the number of participants (percentage). Anderson-Darling test was applied to assess the normality of the continuous variables, and all continuous variables were identified as non-normal distributions.

	Characteristic	500 Hz cohort n = 65,565	Missing	1 kHz cohort n = 9,508	Missing
	ECGs per patient	2.09 (2.31) [A]	-	1.25 (0.62) [A]	-
	Develop HF	3520 (5%) [C]	-	668 (7%) [C]	-
Basic	Age at exam, years	60 (44-73) [B]	4%	61 (45-73) [B]	4%
Basic	Sex, female	30268 (46.2%) [C]	3%	4300 (45%) [C]	3%
Basic	BMI	27.5 (24-31) [B]	28%	28 (24.5-32) [B]	33%
Biochemistry	Sodium (mmol/L)	140 (138-142) [B]	29%	140 (138-142) [B]	23%
Biochemistry	Creatinine (mmol/L)	79 (66-97) [B]	28%	80 (67-96) [B]	33%
Biochemistry	GFR (mL/min/1·73m²)	86 (59-116) [B]	48%	85 (60-114) [B]	50%
Biochemistry	Haemoglobin (g/dL)	13 (12-14) [B]	60%	13 (12-14) [B]	48%
Biochemistry	Potassium (mmol/L)	4 (4-5) [B]	29%	4 (4-5) [B]	33%
Biochemistry	Urea (mmol/L)	5 (4-7) [B]	28%	5 (4-7) [B]	33%
Co-morbidities	Atrial fibrillation	2578 (3.9%) [C]	-	312 (3%) [C]	-
Co-morbidities	Coronary artery disease	4980 (7.6%) [C]	-	690 (7%) [C]	-
Co-morbidities	Chronic kidney disease	1146 (1.8%) [C]	-	113 (1%) [C]	-
Co-morbidities	Diabetes mellitus	4466 (6.8%) [C]	-	575 (6%) [C]	-
Co-morbidities	Hypertension	8945 (13.6%) [C]	-	1072 (11%) [C]	-

Leads that have a linear relation with other leads were ignored. Leads I, II, and V1-6 were used for analysis.

CNN models were created using the Keras framework with Tensorflow backend (315). Secondary models were created using logistic regression on the clinical variables and DNN output. Model performance was assessed using the AUCROC, confidence interval was calculated using the Delong's method (127). Comparisons between these models were performed using the Net Reclassification Index (NRI), a comparison

metric between models with defined thresholds of risk, and Integrated Discrimination Improvement (IDI), a metric that compares two models without defined thresholds, were calculated in R using PredictABEL (316) (317). For the use of NRI, the bins were separated into 0-20%, 20-40%, 40-60%, 60-80%, and 80-100%.

5.4.3 Results and discussion

The main CNN model yields an AUCROC of 0.806 (95% CI 0.797-0.816), where a model with only clinical parameters on the same cohort yields 0.727 AUCROC (0.715-0.740). The categorical NRI is 0.355 (0.306-0.403), numerical NRI 0.495 (0.440-0.549), while the IDI is 0.145 (0.128-0.162). A model that uses both the CNN output and the clinical values parameters yields an AUCROC of 0.824 (0.814-0.834). When evaluating these models against the external dataset, the 1 kHz set, the CNN model yields 0.763 (0.743-0.783), the clinical model performs with AUCROC 0.735 (0.710-0.760). When combining the CNN output into the logistic model, the model yields 0.782 AUCROC (0.761-0.804). The categorical NRI calculated is 0.123 (0.40-0.207), continuous NRI 0.271 (0.174-0.369), and the IDI was 0.063 (0.033-0.093). All measures are significant with p-value < $5e^{-3}$. Figure 28 shows these results.

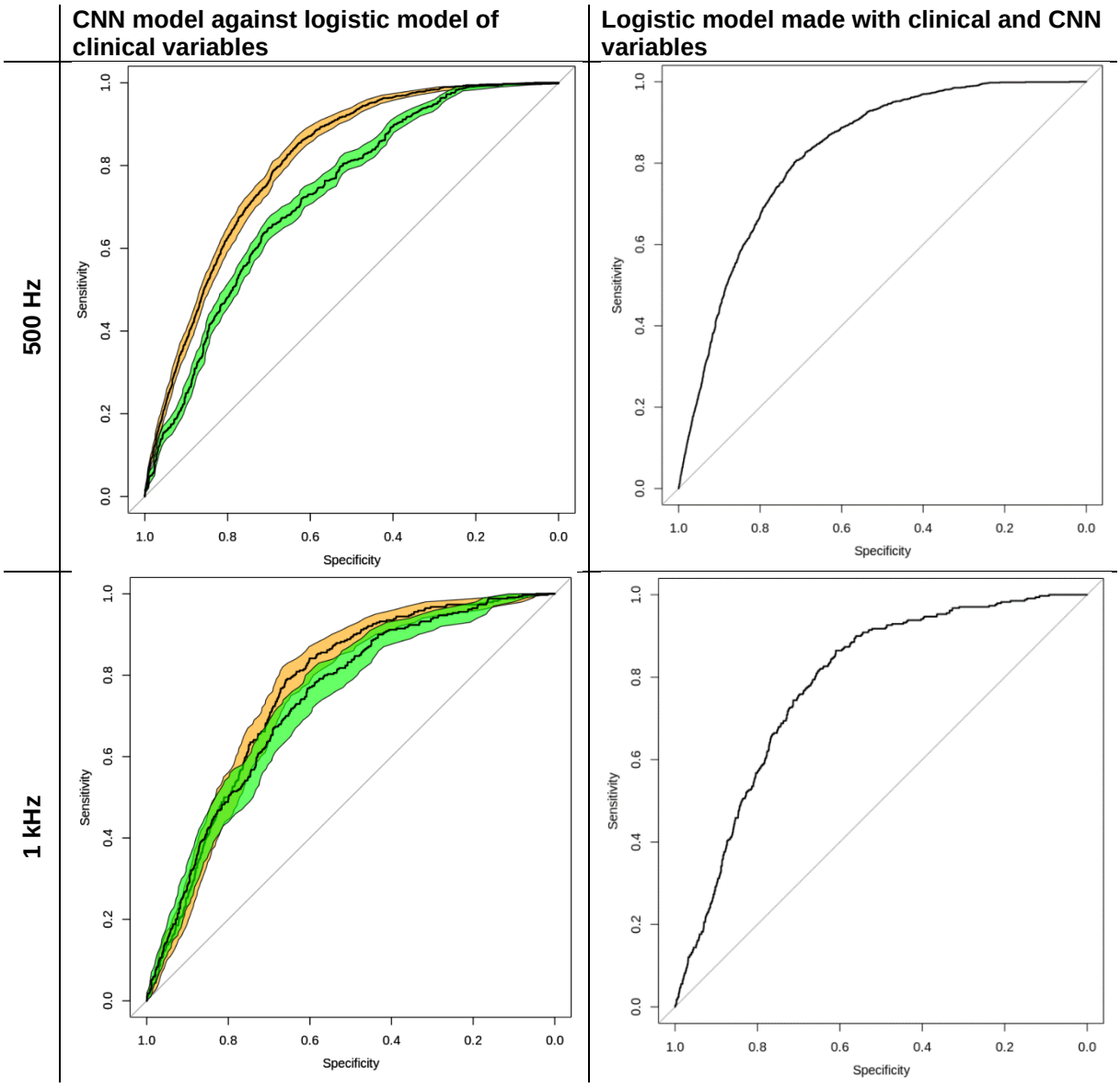


Figure 28: Performance results for the electrocardiogram model to predict incident heart failure. On the left side is a comparison between the CNN and the clinical variables-only models, the shaded part is the 95% CI, orange CNN, green clinical model. On the right is the performance AUC plot for the CNN and clinical variables model. Top and bottom parts indicate the performance of the model on the testing data for 500 Hz and 1 kHz.

We found a model that can be used for the prediction of heart failure. The proposed model contrasts well against a risk model for heart failure based only on clinical information and biochemistry. The proposed model enables the use of ECG data that would not be used to the full extent in heart failure risk assessment.

The developed model can be fine adjusted to the identification of higher risk, or discard patients with lower risk of incident heart failure. More experiments or a clinical trial would be necessary to evaluate with the created models can be effectively used to manage incidence of heart failure.

5.5 Chapter summary

This chapter showed the application of advanced machine learning approaches to the processing and modelling of signal data. While the performance of models using only structured data can provide predictions that support care, advanced solutions using unstructured signals are capable of further assisting the identification of diseases and provide potential to be used in clinical use, both due to their use with routine clinical data, and the novel use of data to improve the clinical risk assessment.

The first case demonstrated that the extraction of (novel) biomarkers from ECG signals can be further explored to aid the identification of patterns that can be used in combination with other variables so as to improve patient/disease stratification. The second case demonstrated a model that can be used to assess patient risk of heart failure. Despite requiring clinical evaluation, this last model shows a great potential to support the diagnosis.

CHAPTER 6 POPULATION DATA

6.1 Introduction

Typically population studies include a set of participants recruited based on a set of rules delineated in the study design, and these rules can sometimes be very limited. In the context of this thesis, these limited rules were exemplified in the BBCAF study (section 3.2), where participants recruited had either AF or other risk factors, and were referred to a hospital setting. These rules can be broader, such as the case of the UK Biobank study, aiming its recruitment to the age range of 40-79 years old (81).

Whilst other chapters explored intrinsic values of different data types, based on small cohorts, this chapter aims to use a limited set of variables and leverage the power of large populations to understand epidemiological features, in terms both of static and longitudinal features. Rather than exploiting patient datasets at an individualised level, patient data are rendered as weighting sample representatives across a big population.

This chapter investigates patient populations behaviors in the context of a real-world datasets (section 1.1). In particular, it examines different patterns across different viewpoints, namely, comorbidities, diseases, as well as the translation between comorbidities and phenotypes, across 4 different use-cases. The approach adopted is generic and can be applied outside the cardiovascular domain and has been used to identify potential relations between non-cardiovascular conditions and cardiovascular specific manifestations.

Moreover, this chapter explores the possibility of better describing a population and of their traits. It introduces clustering, followed by the analysis of prevalence, correlation, and precedence of some diseases. Some of these analyses were performed across different cohorts, using data extracted from a secondary care setting, selected cohorts participating in the UK Biobank, and primary care data collected in THIN.

The first use case explores the admissions of patients into the hospital, and how they could be better separated and treated, through the identification of differences in prevalence and association of comorbidities in difference datasets. In the second use

case, the patterns of comorbidity evolution are investigated from a longitudinal point of view depicting that participants with reported comorbidities are more likely to develop new ones. The third use case explores datasets related to participants that were hospitalised for pulmonary related complications and assess whether they form different subgroup populations. The fourth use case utilises population data to derive potential links between diseases and particular phenotypic manifestations that can be then used to semantically define them (354).

6.2 Clustering of patients

Clustering is a method that aims to identify the best separation of subgroup populations from a given cohort, using a set of features from the dataset. Each subset clusters together participants with more similar features than the overall population. These subsets of participants can then be the target for specific treatments, actions, or further explorations. For example, the identification of comorbidity subgroups associated with different risks can assist the risk assessment of ischaemic heart disease patients (318), in a similar way that the identification of heart failure patient subgroups can assist the practitioner to provide personalised care (319).

Many variations affect clustering: data distribution, shape, how the performance is assessed, and clustering algorithm. Moreover, clustering algorithms may utilise multiple settings ranging from the dissimilarity functions to inherent parameters, also affecting the outcome of the clustering model. A clustering model has strong qualitative features, and many patterns may not be exclusive due to the clustering, but by the interpretation of the analyst.

Despite the variety of clustering approaches, there is no definite approach to the clustering of patients in a comorbidity scenario (320). For this thesis, major clustering methods are considered: k-means, hierarchical clustering and mixture models (321).

k-means clustering is an iterative algorithm that aims to separate the samples into subgroups using some reference points as clusters centres, and a distance metric is evaluated between each sample and the different reference points. k-means clustering has different variations (322). Lloyd’s algorithm starts with randomized points for the

reference points, called centroids, with one centroid for each number of clusters. The algorithm allocates samples to the nearest centroids, and at each iteration of the algorithm, new centroids are created based on the nearest points (323). k-medoids, a variation of k-means, utilises real sample points rather than centroids (324). An implementation of the k-means algorithm is available in R in the package NbClust; this package contains many metrics that can be used both independently or aggregated (325).

After executing clustering algorithms it is necessary to evaluate what is the best number of clusters, a classical method for the evaluation of a good number of clusters is using the elbow method, which identifies the number of clusters where the loss function starts to stabilise (326).

Hierarchical clustering seeks to group the samples from the bottom-up, forming other levels of groups that are re-combined. The algorithm works iteratively. After each iteration, the distance between all the samples, or grouped elements, is re-calculated. The samples, or grouped elements, that are closest are combined. Then, this process is repeated until all the elements are in a single group. It is possible to vary the distance metric between the points as well as how the distance between the groups is calculated. An implementation of this algorithm is available in standard R using function *hclust* (196).

Mixture models premise that there are sub-groups of populations within the overall population, and models probabilistically the participation of a sample in sub-populations. These models aim to identify latent classes that describe underlying phenomena. For continuous features, sub-groups are expected to have some geometric features, with different parameters, such as distribution, volume, shape and orientation. There are many packages in R for these analyses, for continuous variables is *mclust* (327), for categorical variables *poLCA* (328) (329), and for both data types, there is *clustMD* (330). These models can be assessed using the Bayesian information criterion (BIC), which penalises the performance as the number of parameters increases (331).

These approaches require the analyst to decide the number of clusters that will be used. Typically, a range of cluster numbers are tested, for example, 2 to 10, and then a final decision is made based on the available metrics. Stability is considered a good indication of the quality of the modelled clusters (332). It is a good indication if models will be recreated using the specific number of clusters and if minor variations in the dataset and/or model settings affect the model creation. One approach of assessing the stability of a model is through the use of a bootstrapping technique: datasets are sampled with replacement, analyses are performed and the variance of the clusters are evaluated (333). For example, in an approach by Dolnicar & Leisch 2010 (334), bootstrapping was applied using the Calinski-Harabasz index to assess the variance of the clusters (this index is calculated using the variance of clusters centres divided by the sum of within clusters variances, i.e., the separation of the data points divided by the compactness of the values), and the Rand-index (ratio of samples that are in the same sub-group for different models) was collected to measure the concordance between different cluster models (335) (336).

The analyses performed in this chapter utilise, in addition to the model creation, the assessment of variables difference through statistical testing, creation of risk models, survival curves for each subgroup, sub-group prediction modelling, and transformation of data points using PCA to assist understanding.

6.3 Comorbidities associations – Case 1

6.3.1 Introduction

When someone has a condition, they are more likely to have other extenuating conditions. Although many comorbidity associations are known in the literature, some populations have a slightly different comorbidity pattern (337). The understanding of the population comorbidity patterns can assist the clinical decision and facilitate the treatment course selection. In this activity, we are exploring the different disease associations in the UHB dataset and comparing it to the UK Biobank. By performing this analysis, we aim to obtain a better overall picture of different datasets and their patterns of comorbidity.

6.3.2 Data and methods

The UK Biobank and the UHB datasets were used for this analysis (sections 2.3.3 and 2.3.4). For this study, a list of 65 clinical important comorbidities, such as AF, heart failure, stroke, hypertension and diabetes, was employed for this analysis (**Appendix 6.1**). ICD-10 codes were collected from the whole patient record.

Two main statistical analysis were employed. First, the overall prevalence was calculated using the number of patients with a condition divided by the total number of patients. Then, associations between diseases were calculated using the Bayes theorem (134), with its formula described below, where condition B is the pre-existing condition and condition A is the secondary condition, in the form that the resulting probability is "*given that you have condition B what is the probability that you'll have condition A*".

$$P(condition\ A \mid condition\ B) = \frac{P(condition\ A, condition\ B)}{P(condition\ B)}$$

The associations were measured in a pairwise approach for all comorbidities. After the generation of the associations, the differences between the different population' associations were evaluated.

6.3.3 Results and discussion

Figure 29 depicts graphically the overall disease prevalence. Hypertension is a common disease in these populations, with around 25% prevalence, followed by diabetes around 10% depending on the dataset. Some of the conditions can be associated with complications, such as type 2 diabetes mellitus, whilst "at risk of falls" indicate a predisposition from older patients that led to hospitalization. The differences for these cohorts do not surpass 10%.

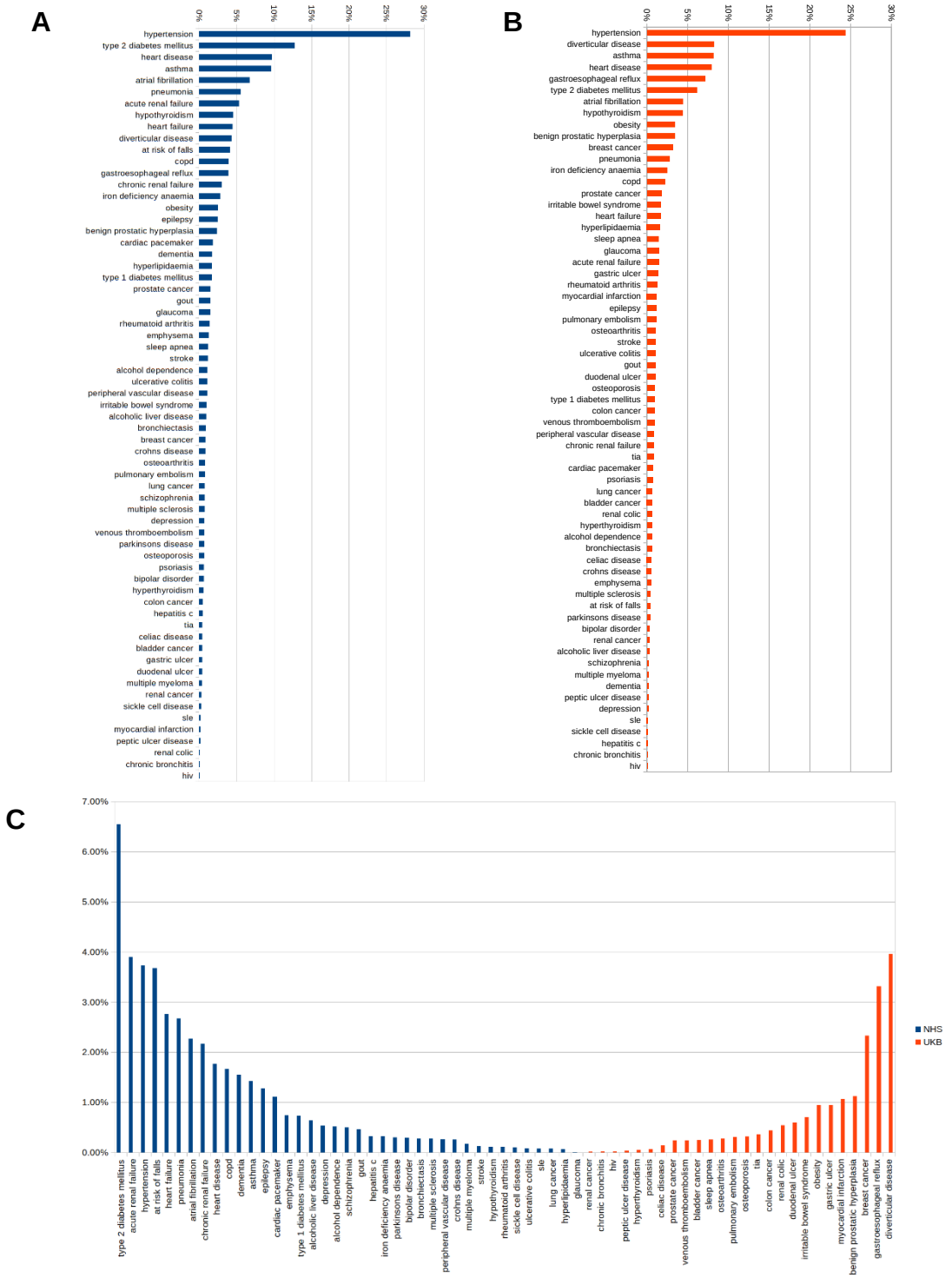


Figure 29: Prevalence of different conditions in the UHB (A) and the UK Biobank (B). Comparison (C) of the prevalence of diseases between the UHB (blue) and the UK Biobank (red) - the vertical axis indicates the percentage of prevalence difference between the sets.

The associations for the different conditions and each dataset are presented in Appendixes 6.2 and 6.3. Appendix 6.4 presents the differences between comorbidities associated with the two datasets.

For example, atrial fibrillation, dementia, osteoporosis and osteoarthritis have increased associations with patients at risk of fall within the UHB dataset. Patients that have more falls with injury are usually older, in a similar way that the conditions cited above are associated with increased age. This shows an underlying difference between the datasets: UHB patients had more *at risk of falls* reported, possibly due to being a more emergency care setting, while the UK Biobank population would have more diseases reported with or without falls.

Mental health conditions, such as depression and bipolar disorder are shown to be more commonly associated with the UK Biobank. These conditions are more commonly reported in primary care settings, where they are usually identified and treated, and when patients have treatments in a hospital, they do not usually need to report these conditions.

Different cohorts may have different patterns of comorbidity, for example, there might be an inherent population predisposition, care setting, and or data processing reasons. These approaches help to pinpoint these differences and provide a starting point for investigating the population differences. We could identify associations in the population, describe the data, and provide a comparative description of the cohorts.

6.4 Temporal analysis of data – Case 2

6.4.1 Introduction

While the prevalence for a population provides insights on the final picture seen in a cohort, as seen in the previous case, sometimes it is important to know the temporal pathogenesis of relative conditions, as it could be the case that a set of patients following a different pathway may have a different set of outcomes.

Other studies have previously investigated patterns of disease evolution, among those, the Danish Disease Trajectory Browser investigated relation pairs for

comorbidities in the Danish population (338), Zemedikun et al. 2018 investigated relations between comorbidities in clusters from the UK Biobank (339), and Vetrano et al. 2020 investigated patterns of comorbidities in clusters in a Swedish population of older adults (340).

In this case study, we hypothesize that there is a time factor that provides new insights into the evolution of diseases and the cohort.

6.4.2 Data description and analysis

This case utilises the UK Biobank dataset (section 2.3.3). A set of 28 important comorbidities is selected for analysis (Appendix 6.5). All patients with comorbidities from the UK Biobank are used, and each disease is identified from the HES records, where a coded diagnosed affirms that the patient had the condition on a specific date, while missing codes are considered that the patient was not diagnosed with it. The analysis is separated into two aspects, a pairwise patterns of comorbidity between a set of important diseases, and a rule mining algorithm is performed over the time-sequence.

Pairwise disease precedence. To evaluate the precedence of diseases, all pairs of the 28 conditions were assessed. For each pair, all patients that had both conditions were selected, and the difference in days between the dates of diagnosis for the pair was calculated. Then, these values were plotted in a histogram.

Time-sequence mining. This analysis uses the *CM-SPADE* algorithm to identify time-based rules for diseases (section 2.5.2) (137). To make this algorithm work, for each patient the different times of diagnosis are compiled. For each different time, the patient is marked with the conditions that happened before or at the time of the event. *CM-SPADE* algorithm goes through the different patients and events building a sequence of common events with their related statistics. Figure 30 illustrates how the data was selected.

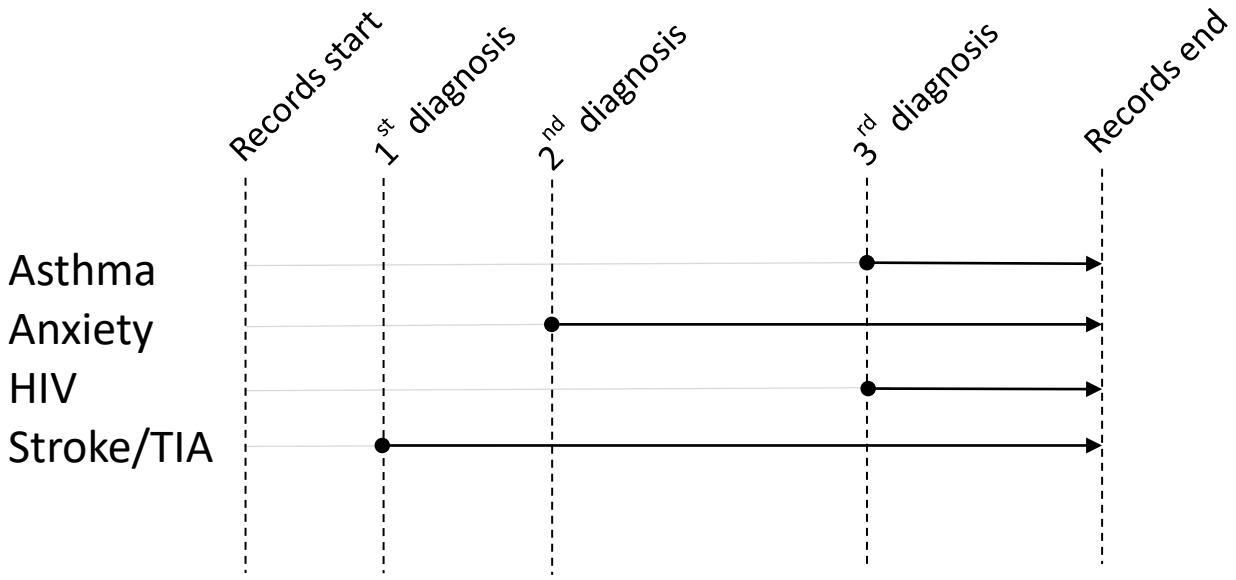


Figure 30: Illustration for the different time points used for time-based analysis. The vertical lines indicate the beginning and end of the records and all the different diagnosis events for a hypothetical patient. A patient might have up to 28 different events recording new diagnoses.

6.4.3 Results and discussion

Pairwise disease precedence. The mean number of patient-disease pairs is 1,415.39 (standard deviation 2,936.01). Ischaemic heart disease and hypertension are part of the pair with the highest number of patients, with 29,810. Results are shown in Figure 31. The V-shape comes from the reduced number of days in the centrally selected bins. After visual inspections, two diseases were identified to have marked contrasts in their histograms, those are asthma and HIV.

Asthma tends to be identified before other diseases; it has a strong genetic predisposition component. Different studies have seen asthma precedence over other diseases, cases where asthma was diagnosed before anxiety (341), arthritis (342), and chronic kidney disease (343). Asthma treatment with or without steroids influences the outcomes of these patients. Patients that had steroids medication have a higher risk of developing chronic kidney disease. Not only the initial presence of asthma and its treatment influences the predisposition to other conditions, but it also leads to lifestyle changes (343).

HIV has a much different pattern than other comorbidities. This is due to it being an infectious condition. It is not an estimated subsequent step of any condition. But it might

proceed with the development of multiple conditions if the patient contracted it early enough or if it was left untreated.

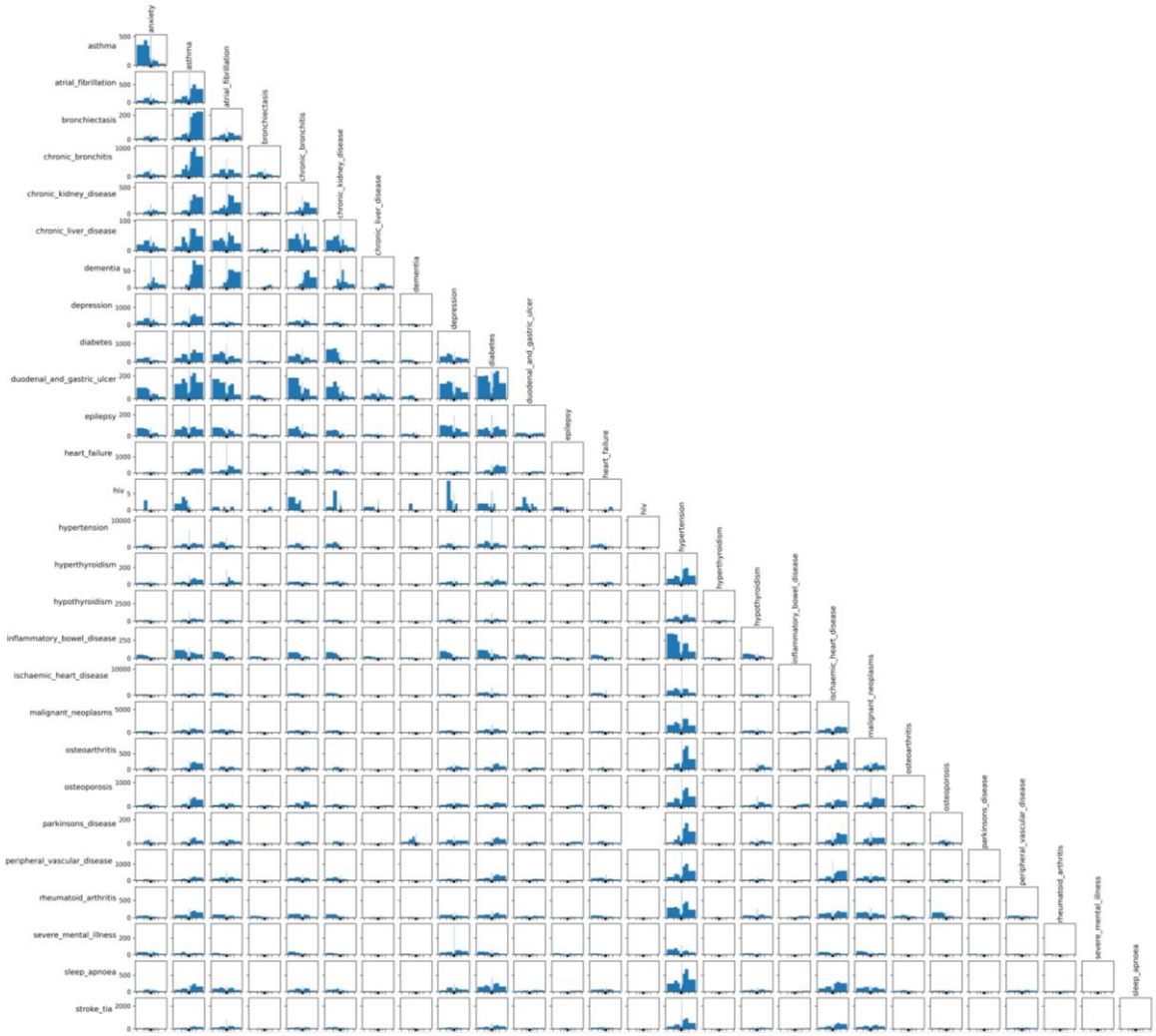


Figure 31: Representation of pairwise disease precedence in the UK Biobank. Histogram of time between disease on the row against the one on the columns. Ticks in the middle indicate a close interval between diseases. Positive (to the right of centre line): column disease before; row disease after. Negative (to the left of centre line): column disease after; row disease before. The V pattern is due to the bins selected. The bins boundaries are: -20 years, -10 years, -5 years, -2 years, -1 year, -6 months, 0, +6 months, +1 year, +2 years, +5 years, +10 years, +20 years. Empty relations indicate there is no data available for the pair. Bin height is scaled to the row of pairwise relations.

Time-sequence mining. Due to the intention of knowing the different pathways between the diseases, only participants with at least 3 comorbidities were selected for analysis. This reduced the UK Biobank dataset to 65,537 participants.

Despite the large number of rules found in the algorithm, the results are exemplified on participants having asthma. In the UK, the prevalence of asthma is 6.5% (344). For patients that had at least 3 reported comorbidities, the prevalence of asthma goes to 26.52%. Out of these patients, 71.43% are reported with asthma twice, falling to 54.64% on thrice (this due to the prevalence of other conditions/how early the patient had asthma). The population that had asthma reported after anxiety is 1.54%, and the population of the dataset that had reported anxiety after asthma is 2.50%. It is also the case that 13.61% of the population have asthma before reported hypertension, whilst 13.41% have it the other way around (hypertension before asthma), and while the former has the confidence of 51.31% of happening, the latter only 17.03%, i.e., patients that have asthma are more likely to have hypertension coded after an asthma diagnosis, rather than the other way around.

Furthermore, there are some associations between the 3 terms identified, such as 2.11% of the patients will have asthma followed by diabetes then hypertension. For those with the two first conditions, 41.30% will have hypertension. Looking at it the other way, 2.42% of the patients will have asthma followed by hypertension and then diabetes, for those participants with the two first conditions, only 17.79% will have diabetes.

Figure 32 illustrates the trajectories explored. Despite similar numbers of patients that have one condition then another, the ratio of those with some conditions that end up developing other conditions is very different, with conditions such as hypertension predominantly happening last.

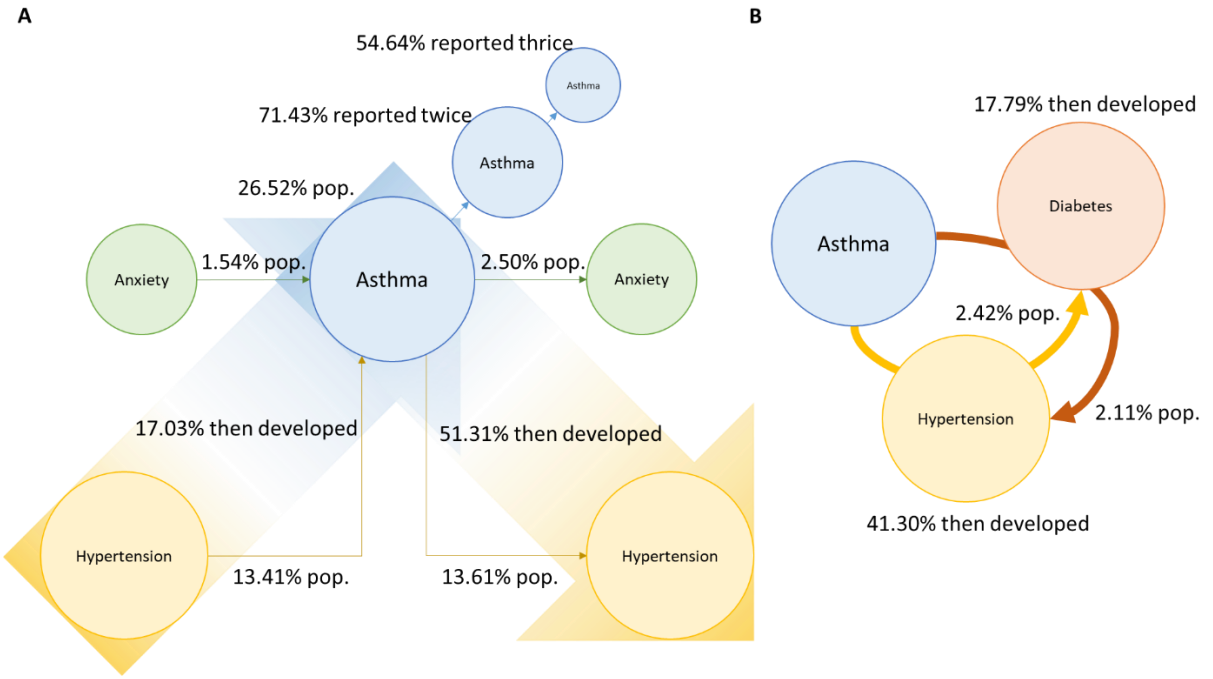


Figure 32: Illustration of patients' trajectory. The different circles indicate different conditions in the dataset, the edges indicate trajectories identified, the numbers indicate the prevalence and incidence of different subgroups. Population (pop) percentages are for patients that had both (A) or all 3 (B) conditions. "Then developed" percentages indicate patients that had the first (A) or the two initial conditions (B) that would then develop the second (A) or third (B) condition.

The patterns of comorbidity are very complex, and there are different ways of assessing them. This study has exposed a way of investigating these medical datasets on a big scale. This way provides an overview of the population and might be done as the first assessment of the population morbidity trajectory.

6.5 Chronic obstructive pulmonary disease patients' stratification – Case 3

6.5.1 Introduction

Chronic obstructive pulmonary disease (COPD) is a common disease, it can be prevented and treated. Patients usually have some symptoms: dyspnoea, cough and sputum production. The development of COPD is propitiated by genetic predisposition, smoking, pollution, and low air quality. Other chronic diseases increase COPD morbidity and mortality (345).

The degree of COPD is determined by the relation of two measurements: the forced expiratory volume in 1s (FEV1), and the forced vital capacity (FVC). These measurements are obtained in a spirometry test, and a value of are $\frac{FEV1}{FVC} < 70\%$ is a common threshold to indicate at least a moderate airflow limitation.

Decreased values of either FEV1 or FVC are associated with AF incidence (346), and the hospitalization of AF patients is higher among those with low FEV1. The Rotterdam study showed a 28% increase in risk of AF for COPD patients (347). Furthermore, COPD is associated with other cardiovascular conditions, with an effect where one conditions worsen the other - patients require a combined treatment to supress the worsening of either conditions (348) (349).

In this case we hypothesized that a more personalised treatment for a COPD patient could improve care. We investigated the presence of subgroups of COPD patients are the time of first admission with diagnosed COPD into the hospital. This study was performed using the UK Biobank data.

6.5.2 Data description and analysis

The UK Biobank dataset was utilised for this analysis. 14,689 participants that had a diagnosis of COPD were included. Participants without FEV1 or FVC data, or a value above 70% were excluded from analysis. In total, 6,195 participants were included on analysis (Figure 33).

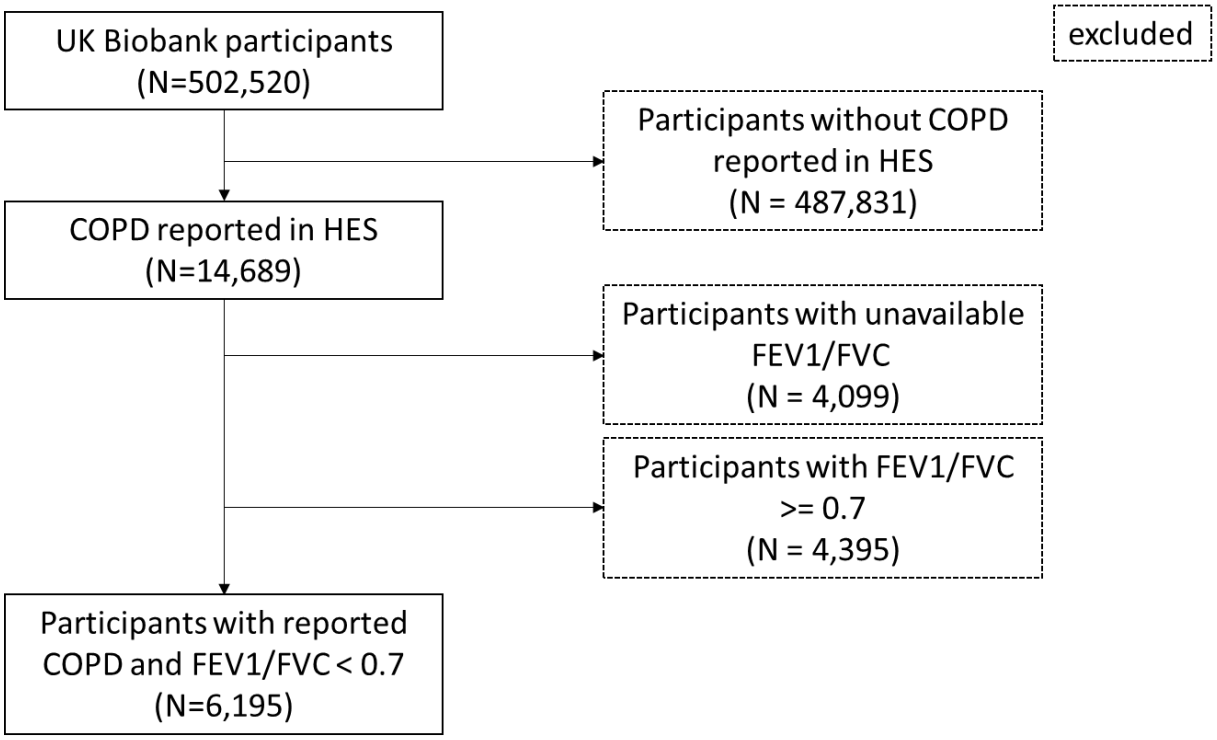


Figure 33: Patient flowchart for the COPD clustering study. Patients were collected from the UK Biobank. Dashed boxes indicate excluded participants.

Patients had extensive data collected: demographic variables: age, sex, smoking status and body mass index; biochemistry results inclusive of eosinophils count and percentage, neutrophil count and percentage, lymphocyte count and percentage, white cell count and percentage, platelet count, haemoglobin, C-reactive protein, estimated glomerular filtration rate (derived using the Modification of Diet in Renal Disease Study formula (350)), urine albumin creatinine ratio, haemoglobin A1c, total cholesterol, LDL cholesterol, HDL cholesterol, triglycerides, and vitamin D concentration; symptoms: shortness of breath walking on ground level, chronic cough, chronic phlegm, wheeze, fatigue, chronic pain, weight change (loss/gain), poor sleep/insomnia, chest pain, anxiety and low mood; comorbidities: ischaemic heart disease, stroke, transient ischaemic attack, peripheral vascular disease, heart failure, hypertension, atrial fibrillation, asthma, bronchiectasis, diabetes, chronic kidney disease, dementia, duodenal and gastric ulcer, epilepsy, HIV, hyperthyroidism, hypothyroidism, inflammatory bowel disease, chronic liver disease, malignant neoplasms, osteoarthritis, osteoporosis, Parkinson's disease, sleep apnoea, rheumatoid arthritis,

depression, and anxiety; other clinical measurements such as systolic and diastolic blood pressure, restring pulse rate, total bone mineral density score, FEV1, FEV1 z-score, FEV1 percent predicted, FVC, FVC z-score, peak expiratory flow rate (PEFR), predicted PEFR, percentage of predicted PEFR, LVEF.

All comorbidities were collected up to the date of COPD diagnosis using UK Biobank HES records. Other variables were collected at the time of patient recruitment.

Clustering analysis was performed in R using the poLCA package (196) (328). The experiment was repeated 100 times with bootstrapping. Comorbidities data was used for analysis and the remaining variables were used for reporting. The best number of clusters was identified using the BIC (331). After the identification of the best number of clusters, models were created using the whole data. Rand-index was used to compare cluster models (335) (section 6.2). Fisher test was performed on the categorical variables, analysis of variance (ANOVA) on the numeric variables.

6.5.3 Results and discussion

After model creation, the best number of classes differed between the models. The best number of clusters is 3, with some bootstraps showing best performance on up to 5 clusters (Figure 34).

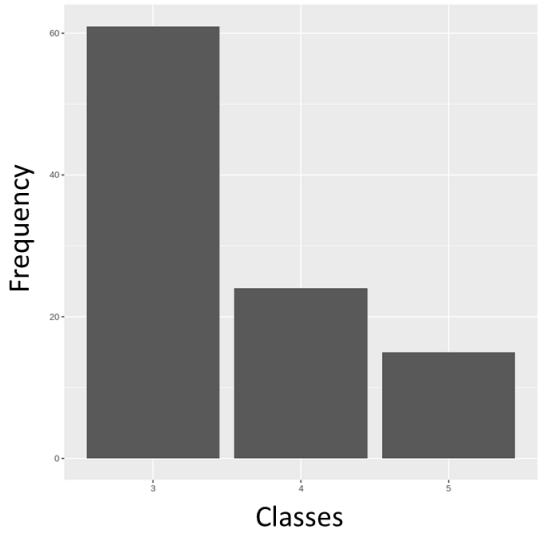


Figure 34: Frequencies of the best number of clusters over the bootstrapping iterations.

The agreement, as measured by the Rand-index, was of at least 72%, and up to 86% between the different models. These indicate that, although there are some variations using a different number of clusters, the participants in a group are commonly grouped together by other models. A final model with 3 clusters was used. Most of the patients were in a same group 1, with 4,081 patients, a smaller group 2 contained 348 patients, whilst a group 3 contained 1,766 patients.

The overall distribution of comorbidities in shown in Table 22. A relation of numerical variables is shown in Table 23, information of number of comorbidities is shown in Table 24. A complete relation of categorical variables is available in Appendix 6.6.

Table 22: Distribution of comorbidities in each COPD clusters. Values are reported as number of participants (percentage in group).

	Comorbidities	**Fisher tests (p-value)**	**Group 1 (4081)**	**Group 2 (348)**	**Group 3 (1766)**
Comorbidities	Anxiety	p < 1e-3	13 (0%)	143 (41%)	56 (3%)
	Asthma	p < 1e-3	850 (21%)	147 (42%)	490 (28%)
	Atrial Fibrillation	p < 1e-3	105 (3%)	26 (7%)	450 (25%)
	Bronchiectasis	0.148425787	115 (3%)	12 (3%)	66 (4%)
	Chronic Kidney Disease	p < 1e-3	28 (1%)	19 (5%)	186 (11%)
	Chronic Liver Disease	p < 1e-3	29 (1%)	28 (8%)	22 (1%)
	Dementia	p < 1e-3	9 (0%)	6 (2%)	20 (1%)
	Depression	p < 1e-3	139 (3%)	199 (57%)	89 (5%)
	Diabetes	p < 1e-3	117 (3%)	23 (7%)	612 (35%)
	Duodenal and Gastric Ulcer	p < 1e-3	144 (4%)	27 (8%)	110 (6%)
	Epilepsy	p < 1e-3	49 (1%)	33 (9%)	46 (3%)
	Heart Failure	p < 1e-3	24 (1%)	6 (2%)	314 (18%)
	HIV	1	3 (0%)	0 (0%)	1 (0%)
	Hypertension	p < 1e-3	985 (24%)	118 (34%)	1554 (88%)
	Hyperthyroidism	p < 1e-3	10 (0%)	30 (9%)	25 (1%)
	Hypothyroidism	p < 1e-3	189 (5%)	75 (22%)	117 (7%)
	Inflammatory Bowel Disease	p < 1e-3	44 (1%)	20 (6%)	42 (2%)
	Ischaemic Heart Disease	p < 1e-3	265 (6%)	46 (13%)	1048 (59%)
	Malignant Neoplasms	p < 1e-3	706 (17%)	96 (28%)	333 (19%)
	Osteoarthritis	p < 1e-3	35 (1%)	29 (8%)	37 (2%)
	Osteoporosis	p < 1e-3	128 (3%)	85 (24%)	53 (3%)
	Parkinson's Disease	p < 1e-3	5 (0%)	6 (2%)	12 (1%)
	Peripheral Vascular Disease	p < 1e-3	69 (2%)	19 (5%)	356 (20%)
	Rheumatoid Arthritis	p < 1e-3	83 (2%)	55 (16%)	62 (4%)
	Severe Mental Illness	p < 1e-3	20 (0%)	17 (5%)	9 (1%)
	Sleep Apnoea	p < 1e-3	38 (1%)	12 (3%)	79 (4%)
	Stroke Tia	p < 1e-3	42 (1%)	17 (5%)	189 (11%)

These groups have different comorbidity patterns. Patients in the first group tend to have a smaller comorbidity history than the others. Group 2 have the most patients that have some mental issues and malignant neoplasms, whilst group 3 contains patients that suffer the most from cardiovascular conditions.

Table 23: Relation of numerical features for each COPD cluster. Values are reported as median (1st-3rd quartile).

Numerical features	ANOVA (p-value)	Group 1 (4081)	Group 2 (348)	Group 3 (1766)
Age (disease)	1.41E-40	65.31 (59.94-69.57)	65.32 (58.88-69.68)	67.97 (63.43-71.37)
BMI	6.25E-63	26.23 (23.34-29.61)	26.16 (23.21-29.46)	28.72 (25.44-32.54)
Weight	4.72E-71	74.4 (64.30-84.93)	72.65 (62.83-84.40)	82.6 (71.62-95.60)
Time until death	1.86E-08	1054 (243.00-2536.00)	612 (122.33-1378.33)	745 (229.83-1616.33)
Time until re-admission	2.99E-13	37 (0.00-518.00)	21.5 (0.00-242.50)	32.5 (0.00-316.08)

Further to the comorbidity separation, group 3 has higher BMI and older patients. Ages of groups 1 and 2 are similar, although ANOVA analysis showed differences between all groups.

Table 24: Information about number of comorbidities for each COPD cluster. Values are reported as number of patients (percentage in group).

Number of comorbidities	Group 1 (4081)	Group 2 (348)	Group 3 (1766)
comorbidities (>=1)	2820 (69%)	348 (100%)	1766 (100%)
comorbidities (>=2)	1146 (28%)	348 (100%)	1766 (100%)
comorbidities (>=3)	259 (6%)	290 (83%)	1353 (77%)
comorbidities (>=4)	18 (0%)	169 (49%)	809 (46%)
comorbidities (>=5)	1 (0%)	81 (23%)	400 (23%)
comorbidities (>=6)	0 (0%)	37 (11%)	178 (10%)
comorbidities (>=7)	0 (0%)	12 (3%)	75 (4%)
comorbidities (>=8)	0 (0%)	6 (2%)	21 (1%)
median (IQR)	1 (200%)	3 (100%)	3 (100%)

Patients in the groups 2 and 3 have more comorbidity history.

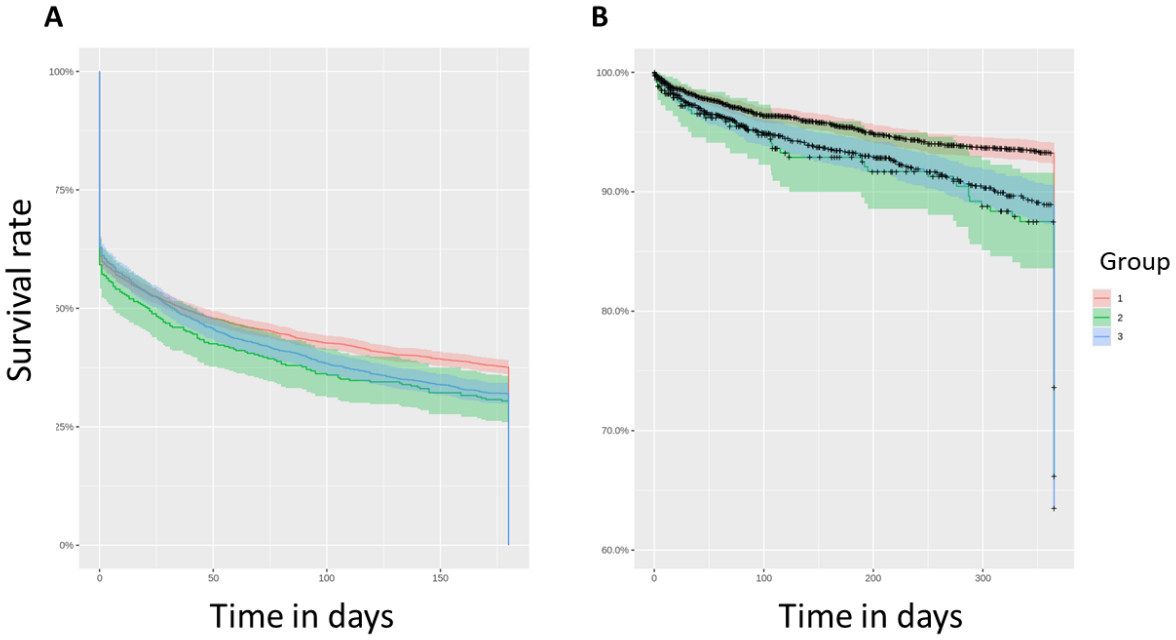


Figure 35: Survival analysis for each COPD cluster. In **A** the readmission curves censored to 180 days, in **B** the mortality curves up to 360 days.

Patients in group 1, with a reduced severity and number of comorbidities tend to take longer to be readmitted and have a lower mortality rate. Patients with mental issues and malignant neoplasms tend to be admitted earlier, and their mortality, although varying is very similar to group 3, with predominantly cardiovascular diseases (Figure 35).

The analysis shows that, for a new patient being admitted with COPD diagnosis, this patient tends to belong to one of 3 distinct groups: a cardiovascular group, where a quarter of the patients have atrial fibrillation; a mental issues/higher risk patient group, where there is not a predominant profile of cardiovascular disease, but the severity risk of the group is high. Group 1 contains lower risk patients, these tend to have a lower number of comorbidities, and are shown to take longer to be readmitted and tend to survive longer. These different groups can then be considered for targeted clinical care.

6.6 Patient phenotypes – Case 4

6.6.1 Introduction

Reference values can simplify the interpretation of laboratory tests, values that are too far from normally seen values indicate that a patient might suffer from a condition and could indicate the need for a clinical investigation. In some cases, the abnormal measurement is an indication of the disease itself, such as the case of haemoglobin, where the measurement of haemoglobin A1c, glycated haemoglobin, which is linked to sugar, when in elevated numbers indicate a higher risk for diabetes, a much higher number indicates that the patient has diabetes.

Despite not always having a cut-off to indicate the definition of diseases, abnormal values can be good indicators of conditions, such as the case of BNP and heart failure, or other cardiovascular conditions.

In this case, we explore the associations between conditions and patient biochemistry results. These associations could be used as a comprehensive relation between these terms, even indicating conditions that are not usually referenced in the literature, and it is hypothesised that by applying this relation it is possible to capture genetic differences.

6.6.2 Data description and analysis

The UHB dataset was utilised for this analysis, the part of the dataset used contains patient biochemistry phenotypes, and comorbidities coded using ICD-10. The UK Biobank dataset is utilised to validate the results, the dataset used contains patient comorbidity history and SNP information.

Each phenotype result available contains a value, date when the exam was performed, and a label indicating if the patient had a measurement that was higher or lower than the defined NHS Biochemistry Reference Ranges (351). For normally distributed values, this normality range is calculated with the 95% confidence interval for a population. The normality range has differences depending on sex, age, and genetic ancestry.

For each patient and each test available, the first phenotype instance was collected up to September 2019, e.g., for a patient that had 3 glucose tests and a single potassium test: only the first glucose test and the single potassium test were collected. Then, for each exam instance, all ICD-10s up to 7 days after the date the phenotype were collected. These criteria assure that any diagnosis due to a blood test is taken into consideration and that a patient is not used more than once for comparison.

For each pair of biomarker and condition, the following analysis was performed: when the number of samples was under 5, Fisher's exact test was performed to obtain the significance, other cases chi-squared test was employed (115, 116). The odds ratio was assessed using Fisher's contingency matrix, between each abnormality term and the normality status of the blood test. The formula is:

$$OR = \frac{N_{condition,abnormal\ test} \times N_{no\ condition,normal\ test}}{N_{condition,normal\ test} \times N_{no\ condition,abnormal\ test}}$$

Cases without 5 patients available for each possibility were ignored. P-values were corrected using the Benjamini-Hochberg method (352) (353), comparisons with an adjusted p-value under 0.05 were kept for investigations.

Significant abnormal terms with OR above 1 were then combined to form a simplified set of abnormal terms: *above*, *not above*, *normal*, *not below*, *below*, and *not normal*. The combined OR was calculated as the distance magnitude (square root of the sum of the squared terms).

After compiling the relation of terms, the list of significant associations was matched against the UK Biobank cohort. For each patient disease, the corresponding phenotype and the abnormality status was added to the patient data. For the different phenotypes, patients were separated into normal and abnormal samples, then evaluated for genomic differences using GWAS.

6.6.3 Results and discussion

There were 140 phenotypes collected, including biomarkers that have minor differences, such as being reported in different units (a full relation is available in

Appendix 6.7). Despite having similar terms, they represent technical variations and represent traits associated with a period. After filtering for significant conditions, there were 134 phenotypes with valid associations and, 67,643 relations were generated on 3,628 unique ICD-10 terms, the mean number of relations per ICD-10 term was 18.64.

In the UK Biobank, 392,296 patients with a least one ICD-10 coded were selected and 11,469 unique ICD-10s, and 134 blood phenotypes. In total, there are 3,449,190 connections between the nodes, a total of 3,396,584 reported ICD-10s, an average of 8.65 ICD-10 per patient, and 52,606 relations between ICD-10s and phenotypes.

The relationships between diseases and phenotypes were created using abnormality information. A disease can be associated with a phenotype that is normal, abnormally low, abnormally high, abnormally high and low, not low, or not high – these cases cover all the options for values that can be normal, higher or lower than normal. It could also be the case that a disease is not associated with any phenotype value range, while a disease can't be associated with all of these intervals at the same time.

GWAS investigations on different phenotypes did not yield novel significant SNPs. However, evaluating glucose phenotype measured by haemoglobin results, some top SNPs were found to be associated with diabetes (Appendix 6.8 list some highly important SNPs). Furthermore, the representation might enable the translation of terms between different dataset types, and, if expanded on other terms and datasets, can provide links to data analysis, as an ontology (354).

Limitations. We did not investigate if associations change when going up the ICD-10 hierarchy. It could be the case that *E10.0 Type 1 diabetes mellitus with coma* indicates one association, *E10 Type 1 diabetes mellitus* another, *E10-E14 Diabetes mellitus* something else, and the overall *Chapter IV Endocrine, nutritional and metabolic diseases*, which contains all the terms, could indicate another phenotype relation. It was expected that the strongest signal would be in the lowest terms.

Many variations could give different insights on the dataset, such as the use of a more comprehensive dataset, such as primary healthcare records that rather than using blood tests would use more described phenotypes from *Read Codes*. There could be

different groups of normality ranges, directly generated from the population used, and the use of a different exam, rather than limiting the analysis to the first clinical presentation.

6.7 Chapter summary

This chapter challenged patient cohorts to identify disease patterns. It showed that datasets on a population scale are powerful to identify patterns between multiple diseases and variables. We demonstrated it through different epidemiological exploratory analysis.

Cases 1 and 2 provide an overall description of a dataset, it is possible to identify conditions that are more highly expressed in a cohort, and the data-driven approaches employed to assist the analyst to understand broad data patterns. These approaches start separating the patients into comorbidity subgroups. Case 3 shows that within the population, there are also sub-populations with distinct subtypes of a disease, a separation that enables a targeted treatment. Case 4 illustrates the generation of novel representations for patient characteristics, with the passage of these representations between different datasets to the increment of patient data. All cases explored the power of datasets at a population level, and applications to a better understanding of the data, improving the knowledge about patients and diseases.

Despite no case specifically focusing on cardiovascular disease, there are marked results that show cardiovascular insights in this scale of population data. In the first case, the prevalence of cardiovascular conditions is exposed with a highlight on the severity of NHS patients, and the differential of cardiovascular morbidity expression (Appendix 6.4). The identification on patients' pathway of hypertension when asthma or diabetes are involved also show the effect of cardiovascular conditions on the temporal development of morbidity (Figure 32). In the third case studied, the use of machine learning techniques to cluster patients shows a clear sub-grouping of cardiovascular patients.

CHAPTER 7 A NOVEL CLINICAL DATA INTEGRATION FRAMEWORK ACROSS MULTIMODAL MULTIDIMENSIONAL DISPARATE RESOURCES

7.1 Introduction

A multitude of tools exists for the storage and handling of data. The storage of data may be in different formats, such as flat files, comma-separated values, spreadsheets, or in more traditional table-structured relational databases systems, such as PostgreSQL and Microsoft SQL (355, 356).

Data scientists frequently operate with datasets available from one or multiple sources, which may range in format, mode, or subject. If there are a multitude of datasets there is an overhead of extracting and concatenating datasets tailored to the available data sources; this process of data integration imposes a challenge for a data scientist who already usually expends the majority of their time processing and cleaning the data.

In the infrastructure of a usual hospital, there are different datasets with either different systems or operated from different locations due to a variety of products handling parts of the operation. None of the different datasets fit into a workstation memory (e.g., in the local UHB data a single database backup uses over 130GB of disk space) and the different datasets may require filters and other corrective procedures before use. Nonetheless, the dataset may contain facts spread over different tables and schemas, requiring the scientist to learn data details, further complicating and delaying actual analysis.

One widely used way of integrating data in the biomedical literature is through the use of ontologies (357). Ontologies are a hierarchy of concepts, providing a common language to the representation of things. Ontologies have a domain vocabulary, e.g. the human phenotype ontology (358). Ontologies are composed of classes, identifiers indicating phenomena, and logical axioms indicating relationships between these classes. A formal definition of knowledge allows ontologies to be used in a computational fashion: the use of the same identifier in different datasets can be used to link the knowledge together, and the relationship between classes can be used to expand on its knowledge. For example, in the human phenotype ontology (358), the

term *HP:0005110* indicates *Atrial fibrillation*, this term is a child of the *Atrial arrhythmia* term, which is a child of the *Supraventricular arrhythmia* term, and so on until the root term of the ontology. Furthermore, the *Atrial fibrillation* term is connected to other hierarchies with other equivalent terms, such as the Medical Subject Headings (359), enabling the use of information in different scopes.

There are many tools for the handling of data passage and concatenation of different datasets. Many solutions are private software, some of them offer cloud solutions to data integration, e.g., Oracle (Austin, USA) and Microsoft Azure (Mountain View, USA). These solutions aim to compile data from different data sources for real-time business analytics. Despite the offerings of private solutions to data analysis, the availability of data integration solutions to research study construction is limited, usually limited to individual packages and solutions that handle some data operations, such as transform and merge datasets.

Analysts work and integrate data using a set of computing language and tools. A recent survey of the Kaggle data science website shows R, Python and SQL as the top used languages (360). SQL (Structured Query Language) is a database language specialised in the handling of large datasets; it offers capabilities for data integration, however, is very limited on the analysis in itself, and is usually used to integrate and pre-format datasets before exporting them for further analysis in other statistical analysis languages (189). Python and R contain different solutions to the integration of data. Whilst in default R, operations to read and merge datasets can be performed, libraries such as *tidyverse* can improve the process, with its comprehensive set of functions to do any type of data preparation and transformation (361). Python has different libraries that are capable of reading and operating upon datasets, interacting with different formats of data, software and frameworks; some well-known packages are *numpy* and *pandas* (362).

Further to standardised packages, other specialised tools are capable of extracting data for analysis. These packages collect data from a format, transform and load into another format, these are *Extract-Transform-Load* (ETL) tools. Bonobo is an ETL package in Python that utilizes a graph architecture for the matching of data, however

does not allow the specification of rules for data extraction and requires a specific implementation for each data source (363). Bubbles is an ETL package in Python for data processing and data quality measurement, it works on top of other data extraction packages, it is capable of data integration however does not allow filtering with rules based on each data-point values (364). Karma utilises an ontology approach to the integration of different data sources, it contains semantic rules from the RDF data model (365) (366). And, in a similar approach, LinkSuite utilises ontologies for the integration of data in different scales (367).

In the healthcare context, there are large amounts of data, followed by a broad range of solutions for data management (25). The systematic handling of healthcare data aims toward a system that can enable personalised medicine and reduced time for the development of trials. In addition to these, the governance, reproducibility and interpretation of models are of increased importance in medicine (368). The green button is an approach that illustrates the re-use of EHR data, where a patient is matched against a patient in a similar scenario (369). Hemingway et al. 2018 explore tendencies for big data analysis in cardiovascular research, with an increased number of trial datasets that can be re-explored for further research questions (370). And on the other side, the use of EHR systems can improve the results from trial datasets, as seen in cerebrocardiovascular death risk assessment, where EHR models perform better than the Framingham dataset (371).

In the NHS there are a few approaches that illustrate the use of data integration: OpenPrescribing.net unifies open England's NHS drug prescription datasets and provide an online tool for analysis (372); schematise a framework for correlation analysis (373). These approaches represent different data in a unified framework, enabling the re-use of data.

There are different tools and approaches to extracting, integrating, and managing healthcare datasets. While they may be able to handle big datasets, they do not handle dependencies between datasets or have a limited performance for bigger datasets. Moreover, some approaches ignore aspects of time boundaries for reproducible data extraction and do not have standardised data. Still, there are requirements to (1)

integrate data in different silos, (2) simplify the data collection pipeline and (3) handle datasets in a generic interface. This chapter describes a framework for the handling of big tabular datasets, and integrator, a study cohort preparation tool that is included of cross-mapping references over different datasets and fast compilation of data with complex relations.

7.2 Integrator – Case 1

7.2.1 Introduction

It was described in the chapter introduction that there are different tools for the extraction and integration of data, such as Bonobo, Karma, and R or Python with their many packages. Still, these packages require multiple steps and internal knowledge of underlying datasets.

A solution to combining datasets is through the transformation of the data into the OpenEHR standards (374), but it requires big transformations to the underlying dataset. There is a balance between changing the whole system for the generation of research datasets, and the use of the normal dataset. Furthermore, there is a barrier to the understanding of the whole system for the construction of datasets for analysis. This tool proposes to become an intermediary, solving the problem without big transformations; it collects the data from different sources and output a dataset ready for analysis (Table 25 compares the different methods).

Table 25: Comparison between different data approaches to study data collection. Despite the differences between the tools, it is exposed as in the different tools found. Knowledge of databases indicate how much of the databases you need to know to efficiently work using the different tools; Modification of databases indicate if the underlying database needs to be changed to use the approach; Data collection toolset indicates how much of the process needs to be implemented to efficiently use the approach; Extraction rules indicate if it is supported, included of time-based rules on time and association with other terms.

Solution	Knowledge of databases	Modification of databases	Data collection toolset	Extraction rules
Standard R/Python packages (361) (362)	Yes	No	Manual	Manual
ETL tools (363) (364)	Yes	Depends	Yes	Some
OpenEHR (374)	No	Yes	Yes	Yes
Integrator (proposed approach)	No	No	Yes	Yes

7.2.2 Data description and analysis

Integrator framework is built in Python on top of the *pandas* package and is available online on GitHub (15). First, the different components of the approach are discussed, the mapping and data extractors, then the main architecture, which unites the framework.

Mapping different contexts. Data may be labelled using different identifiers, which may be equivalent, indicate different scales, mode of presentation, or different data sources altogether. It is very often the case that the data needs to be linked before compiling related data. To do that, two types of reference mappings were implemented:

a mapping from a sub-scale into an upper-scale, such as the case of different levels of identifiers, and mapping from different reference identifiers, such as the case of synonym identifiers, for the same thing, in different datasets (Figure 36).

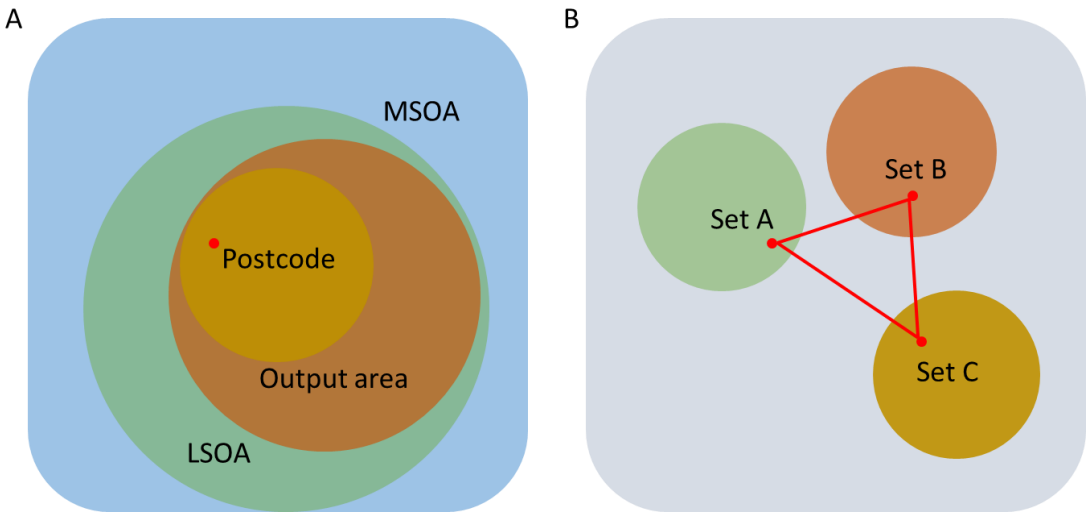


Figure 36: The different data mapping modes. A represents the case of identifiers that are in different scales, such as postcode, where census uses groupings of identifiers. B indicates the case where a data point is in different datasets with a different set of identifiers.

A variable can be associated with others through different mappings. The mapping framework extracts the minimum number of steps using a breadth-first search (375), combining operations that have the same source (Algorithm 1). The complexity of the algorithm is $O(E*I + E^2*O^2)$, where E is the number of mapping extractors, I the number of input variables, and O the number of output variables (target required columns). Despite the apparent complexity of the algorithm, this method reduces the requirement of multiple extractions on the database, i.e., this method is run once, while multiple extractions would be data intensive.

Algorithm 1: Python pseudocode for the matching algorithm. Edges indicate different data source extractors with links between mappings. Path is formed of information from the input variable, the extractor for the link, and the output variables.

```
procedure Ends (list of paths)
        ends = empty list
        for path in list of paths:
                ends.append("last item of path")
        return ends

procedure HasSameEdge (path, other_path)
        for InputVariable, ConnectingEdge in path
                for OtherInputVariable, OtherConnectingEdge in other_path
                        if InputVariable == OtherInputVariable and ConnectingEdge == OtherConnectingEdge
                                return true
        return false

procedure CombinePaths (path, other_path)
        for InputVariable, ConnectingEdge, OutputVariable in path
                for OtherInputVariable, OtherConnectingEdge, OtherOutputVariable in other_path
                        if InputVariable == OtherInputVariable and ConnectingEdge == OtherConnectingEdge
                                NewPath = path
                                NewPath.remove(InputVariable, ConnectingEdge)
                                NewPath.append(InputVariable, ConnectingEdge, [OutputVariable, OtherOutputVariable])
                                return NewPath
        return failure

procedure Mapping (InputVariables, Edges, OutputVariables)
        PathsToExplore = InputVariables
        CompletePaths = empty
        for path in PathsToExplore
                for NewVariable in Edges[path]
                        NewPath = path
                        NewPath.append(Edges[path])
                        if NewVariable in "output variables"
                                if NewVariable not in Ends(CompletePaths)
                                        CompletePaths.add(NewPath)
                        else
                                PathsToExplore.add(NewPath)
        for variable in OutputVariables
                if variable not in Ends(CompletePaths)
                        return failure
        CompiledTargets = empty
        for path in CompletePaths
                for other_path in CompletePaths
                        if HasSameEdge(path, other_path)
                                CompletePaths.remove(path)
                                CompletePaths.remove(other_path)
                                CompletePaths.add(CombinePaths(path, other_path))
        return CompletePaths
```

Data extractors. Extractors are components that return data from a set of identifier variables. These components must be implemented to fit the underlying dataset, and it must return a pandas data frame. The two types of extractors defined in the framework are directly associated variables and time-associated variables. The first type is initialised with the required reference columns for direct matching when executed other data associated with the identifiers of relevance are merged into the original dataset; the latter type requires the specification of the identifier and time reference for each of them, and the data source is matched with the identifier and event time, returning features associated with an event. Figure 37 illustrates these different data sources.

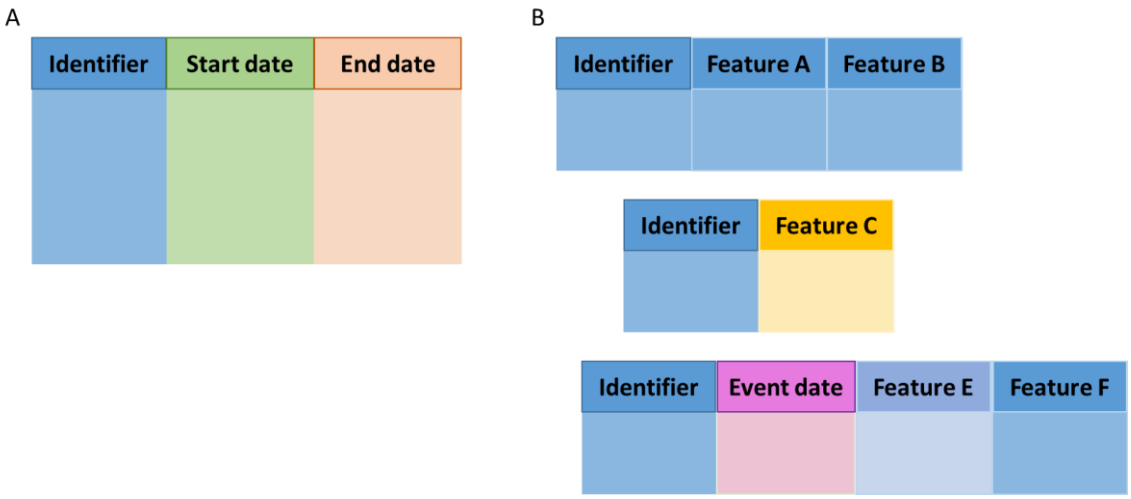


Figure 37: Illustration of the datasets involved in the process. (A) The input dataset, which contains at least one identifier column and may contain other reference mapping columns, such as other identifiers and date variables. (B) Different source datasets that an extractor will query, on the top two columns different features are extracted from different sources, in the bottom dataset a date column provides advanced questions that can be asked.

Abstract implementation for these extractors is available, and for some data problems explored in this thesis, problems described in sections 4.6, 5.4 and 6.5, were implemented to fit the UK Biobank and the UHB data.

Architecture. The tool combines the use of the mapping and data extractor functionalities. Figure 38 illustrates how the tool works.

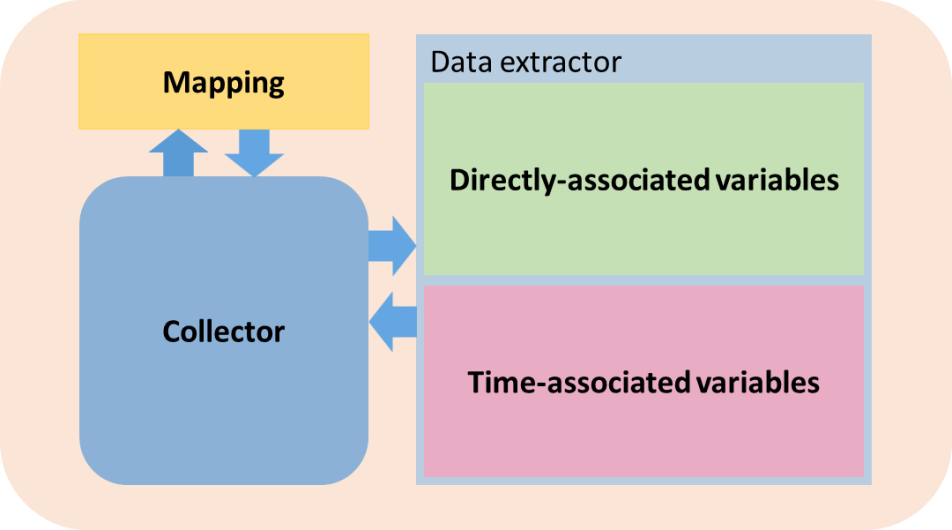


Figure 38: The different components in Integrator. The Collector component interfaces with mapping handles and the data extractors, mapping handler maps one identifier to another type, data extractors handle the collection of data from the database back to Collector.

Collector orchestrates the different parts of the tool, it goes through identifiers needed and set mapping transformations, add these operations as a data extractor, and handlers the data collection and concatenation. Mapping indicates different mappings that could be in a dataset and that may be required to select all the required data. Data extractor is separated into two main types of data that may be collected: data directly mapped via an identifier and data that is associated with a time.

The Collector procedure has three main steps: an initial setup step for configuration, a step for data collection and a final step for data return.

1) **Initial setup**. Check for a valid input dataset format. Make a copy of the data extractors' relation. Verify mappings and build a graph to map reference variables.
2) **Data collection**.
    i. Identify missing reference variables and append to the beginning of the extractors relation if required, using the mapping sources.
    ii. For each chunk in the dataset:
        a. For each data source/extractor in the mapping:
            1. Collect relation of reference variables to mapping.

2. Call data extractor with reference variables.
3. Merge extracted variables in batches.

3) **Return new dataset**. The dataset can be returned in a file that is appended, or as a data-frame ready for analysis.

Applied cases. This tool was applied in different scenarios, explored differently throughout this thesis:

1. Sampled GWAS of Atrial fibrillation – Section 4.6
2. Early prediction of heart failure using electrocardiograms – Section 5.4
3. Chronic obstructive pulmonary disease patients' stratification – Section 6.5

These cases involved the implementation of the solution to the underlying data and the application of the tool to the collection of data. The UK Biobank dataset, despite being static for most of it, had some different files, and refreshes of the data, which were handled with the creation of a special purpose locator for the different fields; these were combined with extractors that handle the unique format of the UK Biobank dataset, that is formed of several fields, instances, and arrays on its main part – ranges of lists – and HES extractors for another set of important data. The UHB dataset is formed mostly of events, and the data had to be pre-located and readied into data extractors; the main requirement being the implementation of connectors that link the application to the anonymised data in SQL format.

7.2.3 Results and discussion

The solution implemented was capable of handling the different data sources available in the UK Biobank and UHB scenarios. In the problem described in Section 4.6, it was required to collect age, sex and comorbidity data for the UK Biobank patients. Initially, all the patient identifiers were collected, and data sources that collect UK Biobank data fields of age and sex were directly matched, and information of comorbidity was extracted for different sets of ICD-10s. The problem described in Section 6.5 has some similarities to the above, as some basic variables were collected, such as age and sex, but in this problem, the comorbidities had to be collected up to the date of COPD admission. The patients were identified when was COPD first reported, and then the comorbidities were time-restricted up to the date of COPD, with the use of a time-associated extractor for comorbidities. The problem in Section 5.4 required that the complete patient history be matched up to the different dates of ECG recordings; this

resembles a lot of the COPD case, just now using the extractors developed for the UHB data, collecting basic clinical data, biochemistry results, and comorbidity information for each patient's ECG.

For the explorations considered, the main benefit was seen in qualitative terms. On the setup of the tool, it was required to understand the database, but then followed by a reduced burden to understand the underlying database structure, the specifications of the dataset, locations, column types, and how the different data sources are linked. This led to savings to the time an analyst would require understanding and collecting from different data sources. When there are minor changes to the type of data an analyst is looking for, such as the alteration in a set of ICD-10 values, time reference of extraction, or the addition of one or another field, the modification to the data collection procedures is trivial. When there are efforts to link another dataset there is still a development task that needs to be fulfilled and underlying knowledge of the dataset is still required.

Underlying knowledge of the database is essential when linking individual data points, e.g., patient identifiers. Disease codes are commonly represented in a defined language, such as ICD-10s, which contain a hierarchy that resembles the use of ontology of integrating knowledge (24), these can be widely used, both for its use on extraction and to link external knowledge over different datasets (a mention of the cross-use of ICD-10s is shown in section 6.6).

While an exclusively ontological approach to data extraction avoids some ETL procedures, e.g. creating a common data model (376), the use of ontologies for data integration requires the definition of a vocabulary, semantic relationships, variable links, and extraction protocol (377) (378). The proposed solution is an intermediary solution, it allows for flexible data extraction, with proved use-cases, without the requirements for an ontological framework or data model.

The approach provides a solution to the problem of collecting and integrating data for analysis. It enables the quick collection of data for different types of retrospective cohort studies, enabling the whole process of analysis in shorter times. As a comparison, the BBCAF study collected data for 2 years, while the collection of

equivalent data using Integrator would take circa 30 minutes, although with only retrospective data.

7.3 USARE Framework – Case 2

7.3.1 Introduction

The analysis of Big Data in healthcare has the potential to generate an enormous amount of value to healthcare, as it is expected that using different data sources can be leveraged to reduce care costs and increase the quality of care (379).

There is a plethora of issues and challenges for the application of Big Data in healthcare settings (380) (381). Some of these challenges are related to organizational, processual, and compliance of the data use – those are the cases of shifting the company to become more data-driven, the insertion of the analytical methods into the daily routine and matters of governance and ethical approvals to the use of the data, respectively. In the context of the UHB, these issues are in the process of ongoing implementation to being solved, with initiatives such as Health Data Research (HDR) UK on the hospital side, and the enveloping of research into clinical practice.

On the technical side, some of the challenges of Big Data are associated with the veracity of the data: it is fragmented, not standardised, and inaccuracy or inconsistency (379). In this section we explore a framework to solve these technical limitations, including how to use the data, and structure it in a way to comply with regulations, reducing the know-how requirements, with a use case in the UHB.

One of the main technical challenges in EHR is having the data in a common format that can be analysed. There are different standards, and for a complete makeover of the data, big transformations are required. This is the case of formats such as Observational Medical Outcomes Partnership Common Data Model (376), where it proposes that different centres and data formats are transformed into a common one, where the data can be commonly treated and analysed.

The EHR systems in the UHB are composed of different software to handle parts of the healthcare routine, for example, PICS handles most information about a patient, from biochemistry results to diagnosis and symptoms (21). Solus Cardiology contains some cardiovascular information, such as ECG records (88), whilst the local copy of HES contains information about in-patients, inclusive of diagnosis and operations. They have a common identifier for participants, and where available, they utilise the same hierarchy for diagnosis, ICD-10s (84).

The different data sources are not directly compatible with each other, as they have dissimilar format and structures. Furthermore, staff do not necessarily know details about the different databases, and how to effectively combine the different sources for analysis.

7.3.2 Proposed framework

The variables required from the UHB dataset are laboratory results, inclusive of biochemistry tests and other physical measurements. Diagnosis and operations, inclusive of ICD-10s and OPCS-4, and ECG recordings. All these data have their associated patient and date of the event.

The main requirements proposed to be solved are:

1. Collection of data sources into a single structure.
2. Increase the veracity of the available data.
3. Reduce barrier to healthcare data analysis.
4. Non-disturbance of clinical EHR systems.
5. Compliance with regulations and policies.

These issues can be solved with the use of the **U**sable **S**ummarised **A**nonymised **R**e-loaded **E**xternal data (USARE) framework proposed. It is paramount that the clinical EHR systems are not affected by any change, due to it, backup copies of the original databases are used, and no live system is used. Moreover, the handling of the dataset is all made within the hospital systems, and protected by the network firewall, with access permissions checks both on the outside and the internal network.

The different backups of data sources were re-loaded into Docker containers (382), these containers maintain isolation of the server used and the re-loaded SQL database (189). To combine the different data sources in the hospital, data integration is needed (383), and a common way of doing it is loading different datasets and transforming them to fit a new format, this is done through an extract transform load (ETL) method (384), data is read, transformed and loaded into a new database.

To facilitate the data transformation, a set of randomised identifiers were created from the list of collected identifiers in the different data sources used. This was used as a key element of the anonymization layer, as with it, it is not possible to trace back the order of the participants. When transformations were done from the original data sources, the newly created sets contained transformed identifiers, and for the different variables required, data were read and transformed into a common way, variables that were not relevant were ignored and not compiled into the new format, this formed a summarised view of the data.

Due to ethical agreements, variables such as age were approximated, and other identifiable data were ignored. Moreover, due to data inconsistencies, some data points had to be processed and cleaned. For example, some laboratory results had invalid values, sometimes with non-numeric values, others with values that were out of boundaries, such as extremely high values or unexpected numbers, for example, BMI values above thousands, and blood pressure lower than 10 mmHg. In some places, there was a non-conformity of formatting, where a dot would be interchanged with a comma on the decimal sign (385). These issues were handled with the creation of transformation rules for the set of imported variables.

The newly created data passed through the different layers up to the final summarised layer, this does not contain any direct link to the original data, in a format that can be read using Integrator, which can be directly input by an analyst. The complete framework is described in Figure 39.

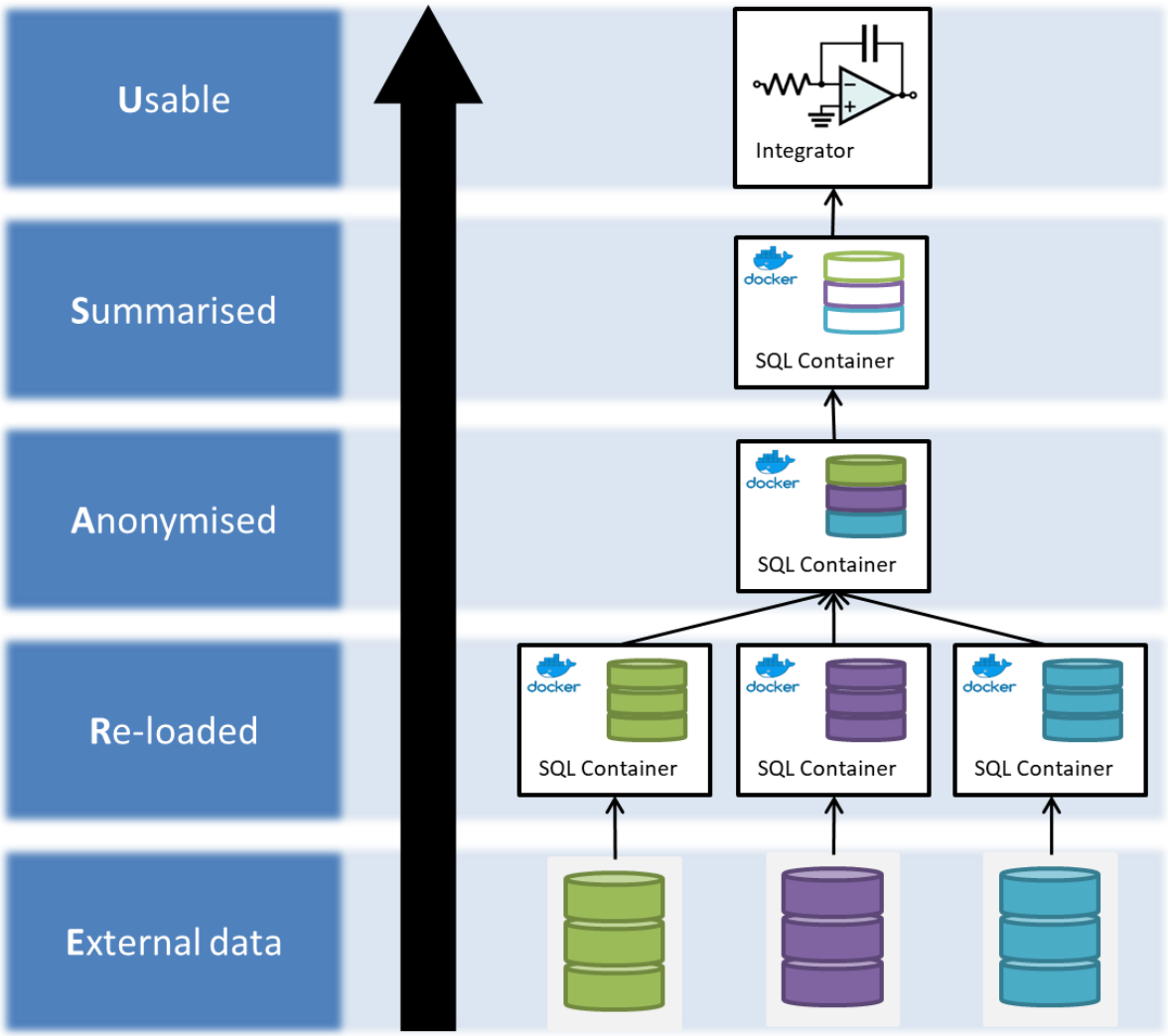


Figure 39: USARE framework for re-usable data. From the bottom-up, external data contains data sources from different databases. The datasets are reloaded into intermediary data storages that are combined into an anonymised dataset. A summarised version of the data is created and used in the analysis.

7.3.3 Results and discussion

The proposed framework complies with the different technical requirements aimed. The framework works with external data, isolated from other operating data, they are re-loaded internally, anonymised, processed to only keep what is needed, summarised and then being available in a usable format to the analyst, it provides an intermediary solution to enable the analysis of healthcare data. It facilitates the demonstration of the benefits of the increased use of big data systems before a bigger transformation is employed. This was seen on models such as the ECG HF models (section 5.4).

In the literature, conceptual models such as Pecoraro et al. 2014 (386) involved the transformation of data on multiple stages, an intermediary stage, called the staging data model, contains similar data to the approach employed. The data in this intermediary state could be further transformed into a data model that other systems use.

7.4 Chapter summary

There is an overall trend in the use of big data approaches for healthcare. There is an expected cost reduction, time to results, and the possibility of the application of new insights into healthcare problems (387). The approaches described in this chapter enable the use of healthcare data for analysis. Approaches such as the Integrator, as part of the USARE framework or not, can be used to collect big data that can be used by an analyst. It reduces the burden of knowing the underlying database structure, compiling and pre-processing the data, with its uniform structure. It was shown to work in scenarios explored in this thesis, such as the collection of comorbidities for analysis, and the modelling of heart failure using electrocardiograms. This systematic integration enables the use of real-world dataset selectively, in opposition to the realisation of research with just controlled study data.

CHAPTER 8 COVID-19

8.1 Introduction

In December of 2019, an increased number of pneumonia cases was associated with a novel coronavirus, deemed to be originated from a seafood market in Wuhan, China (388). The novel coronavirus quickly expanded to several cities in China, and in Europe, the first country to have identified cases was Italy, followed by other European countries and worldwide. Transmissibility, or the R_0 rate of the coronavirus, was identified to be high and adds to the severity of the disease (389). The coronavirus, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), was classified as a pandemic by the World Health Organization and several countries started with measures to counter the spread and damage of the newfound disease, which patients were usually admitted with fever, cough, and shortness of breath; there was commonly a need for intensive care unit (ICU) treatment and ventilators are often necessary (390).

People with SARS-CoV-2 with cardiovascular complications were more likely to be admitted into ICU (391). It has been shown that the virus binds to ACE2, leading to hypertension and heart failure complications (392). Patients with SARS-CoV-2 have increased IL-6 levels, an inflammation marker, showing myocarditis and arrhythmias (393). Patients with AF, especially as a new-onset condition, were associated with higher mortality (394). Furthermore, “long covid” has been associated with cardiac symptoms, such as chest pain and palpitations (395). Despite all these links with arrhythmias and cardiovascular conditions, the link between SARS-CoV-2 and cardiovascular diseases is not well understood.

To understand more about the disease and to improve the care of cardiovascular patients, the multitude of data being generated related to the condition must be explored in a unified fashion similar to the approach depicted in this thesis. Different tools described through the thesis were applied to different scenarios on handling the novel coronavirus projects.

There are three main projects involved: first and foremost, (a) improving the prediction of Covid-19 models, then (b) we assessed the risk of patients undergoing elective and emergency surgery, and (c) we explored a solution to assist the rescheduling of patients that were affected from limited healthcare services during the pandemic.

The surgery risk dataset contains patients which had coronavirus and had to undergo surgery. Usually, the surgery is under emergency and there are both risks of not doing the surgery, risk of surgery with coronavirus and further risks of obtaining coronavirus when hospitalized. These patients have the highest risk on the datasets explored.

The patient rescheduling project investigates the patients which had their routine treatment postponed or cancelled due to modified hospital treatment and strategy during the pandemic. These patients do not have a risk per se of coronavirus, however, their continued specialised treatment can be affected, leading to increased risks of complications. This project aims to assist clinicians in the rescheduling of patients.

8.2 Improving the performance of risk models – Case 1

8.2.1 Introduction

Not long after countries started with lockdown, many researchers shifted their focus onto understanding Covid-19, trying to assess patient risk and looking for solutions to counter the pandemic. In the literature, there was a deluge of risk models, over a hundred mortality risk models created up to the 1st of July 2020, most of them with a high risk of bias (396).

To make the models more robust, it is hypothesized that an ensemble model, a model formed of multiple submodels, will be more generalizable, countering bias that would be found looking at some models individually, and synergise the models. My involvement was supporting the analysis, discussing, and presenting the results.

8.2.2 Data description and analysis

Four datasets were used in the analysis, a first Wuhan cohort (Wuhan01), with 2869 adults with Covid-19 admitted into Wuhan Sixth Hospital or Taikang Tongji Hospital, this dataset contains patients admitted between the 1st and 23rd of February 2020, who

died or were discharged on or before the 29th of March 2020; another Wuhan dataset (Wuhan02) contains 357 adults from Tongji Hospital, data collected between 1st and 4th of March 2020. Two UK datasets were used, a King's College Hospital (KCH) with 1475 adults hospitalised between the 1st of March and 2nd of April 2020, who were followed up to 8th April 2020; a UHB dataset containing 693 adults hospitalised in the Queen's Elizabeth Hospital Birmingham between the 14th of March and 13th of April 2020 and had follow-up data up to the 19th of April 2020. The mortality rates of these datasets are 2.4%, 45.7%, 26.9%, and 19%, respectively.

Table 26: Description of patients in the Covid-19 ICU dataset.

	Wuhan01 Cohort (N=2869)		Wuhan02 Cohort (N=357)		KCH Cohort (N=1475)		UHB Cohort (N=693)	
Variables	**Not Poor Prognosis (N=2738)**	**Poor Prognosis (N=131)**	**Did Not Die (N=194)**	**Died (N=163)**	**Not Poor Prognosis (N=949)**	**Poor Prognosis (N=526)**	**Not Poor Prognosis (N=477)**	**Poor Prognosis (N=216)**
Age, years	60 (49-68)	70 (63-78)	51 (37-62)	69 (62-77)	69 (54-81)	75 (60-86)	72 (57-82)	70 (56-80)
Male, percentage	1389 (50.7)	84 (64.1)	91 (46.9)	118 (72.4)	514 (54.2)	330 (62.7)	254 (53.2)	144 (66.7)
Red cell distribution width, percentage	12.9 (12.3-13.5)	13.0 (12.5-14.0)	12.0 (11.8-12.7)	12.9 (12.3-13.9)	–	–	13.7 (12.7-15.4)	13.9 (13.2-15.1)
Albumin, g/L	38.3 (35.5-40.7)	31.6 (28.7-35.0)	37.5 (34.2-40.2)	30.1 (27.6-33.0)	38.0 (35.0-41.0)	36.0 (33.0-39.0)	31.0 (26.0-35.0)	28.0 (22.0-32.0)
C-reactive protein, mg/L	2.1 (0.8-7.3)	59.9 (14.2-120.0)	19.5 (3.8-49.8)	114.1 (61.9-178.8)	72.5 (28.8-127.9)	112.2 (56.8-216.5)	83.0 (42.0-140.2)	180.0 (102.5-267.0)
Serum blood urea nitrogen, mmol/L	4.3 (3.6-5.4)	6.8 (5.0-11.0)	–	–	–	–	6.3 (4.5-10.4)	8.1 (5.4-13.1)
Lymphocyte count, 10^9/L	1.5 (1.1-1.9)	0.7 (0.5-1.1)	1.1 (0.8-1.5)	0.6 (0.4-0.8)	1.0 (0.7-1.4)	0.9 (0.6-1.4)	0.9 (0.7-1.3)	0.9 (0.6-1.2)
Direct bilirubin, umol/L	3.3 (2.5-4.4)	5.4 (3.5-7.2)	3.5 (2.5-4.7)	6.2 (4.4-9.2)	–	–	10.0 (7.0-14.0)	11.0 (8.0-20.0)
Lactate dehydrogenase, IU/L	174.6 (150.3-210.2)	332.2 (244.9-461.0)	250.0 (202.2-310.5)	567.0 (427.5-762.0)	–	–	316.5 (245.8-411.0)	436.0 (340.0-623.0)
Serum sodium, mmol/L	141.6 (140.0-143.2)	139.8 (137.4-143.4)	139.2 (136.5-141.2)	138.9 (135.8-143.6)	–	–	137.0 (134.0-140.0)	138.0 (135.0-143.0)
Neutrophil count, 10^9/L	3.5 (2.7-4.5)	6.7 (4.8-9.9)	–	–	5.1 (3.7-7.4)	6.6 (4.5-9.4)	4.7 (3.4-6.7)	6.7 (4.8-9.4)
Oxygen saturation, percentage	97.8 (97.0-98.2)	96.6 (94.5-97.7)	–	–	19 (18-20)	23 (20-28)	94.0 (93.0-96.0)	92.0 (88.0-94.0)

Seven models from the literature were chosen from the literature, Ji et al. 2020 (397), Shi et al. 2020 (398), Gong et al. 2020 (399), Lu et al. 2020 (400), Levy et al. 2020 (401), and Yan et al. 2020 (402). These models come from 6 different regions, from two countries, China and the United States. The original datasets where these models were created contain median ages between 44 and 65 years old, with mortality rates varying between 7% and 52%. The models had their parameters collected from published material and reimplemented for analysis. Then, the different models were ensembled using a bagging predictors approach (403).

The bagging approach employed uses a competence assessment framework to assist when grouping the results from the different models, this is done using 3 key elements: (a) similarity of a model with the patient, (b) general competence of the model, where bigger datasets used for model derivation are ranked higher, and (c) data completeness of the patient related to the model employed. Further details are available in Wu et al. 2020 (12).

8.2.3 Results and discussion

For each dataset, a different model performed best. On Wuhan01, Xie et al. 2020 performed (AUCROC of 0.888 95%CI 0.874-0.926), on Wuhan02, Dong et al. 2020 (AUCROC of 0.881 95%CI 0.841-0.913), on KCH and UHB, Levy et al. 2020 performed best (AUCROC 0.658 95%CI 0.629-0.685, and AUCROC 0.660 95%CI 0.617-0.713). No model performed consistently across the different datasets, whilst the ensemble model had consistent discrimination in all cases: 0.914 (95%CI 0.891-0.937), 0.890 (95%CI 0.856-0.921), 0.665 (95%CI 0.640-0.692), and 0.683 (95%CI 0.643-0.723) on Wuhan01, Wuhan02, KCH and UHB, respectively.

These analyses showed that a single model did not perform consistently well on different datasets. While a model was best in a cohort it was often the case that it could not be used to the same performance power in another dataset. This is partially due to patients having different eligibility to hospital care, as in the UK only patients in a more severe state were admitted into hospitals, and the different stages of the pandemic when these patients were admitted. The combination of the models into an ensemble provided a model that performed better, and more consistently on different datasets.

8.3 Surgery risk – Case 2

8.3.1 Introduction

CovidSurg is an initiative under the GlobalSurg project (404). The overall project and different initiatives investigate factors that may affect a surgery outcome: patient clinical history, biochemistry profile, surgery type, urgency modality, patients' follow-up, and other factors, such as the experience of the surgical team, centre and country of operation. At the time of writing, GlobalSurg collaborative is formed by surgeons and anaesthetists from 122 countries.

Due to the pandemic, it was estimated in May 2020 that over 28 million surgeries could be cancelled (405). The assessment of the patient risk of undertaking surgery is extremely important, as for different operations there is always a balance of risk of them taking the procedure and having Covid-19 complications, or the risk of not doing the procedure and risking evolving the condition. Half the patients that had Covid-19 undergoing surgery had postoperative pulmonary complications and were associated with higher mortality risk (406).

We explored the use of CovidSurg data to better understand the mortality risk during the Coronavirus pandemic. This study is approved under clinical audit terms in the UK, and in other countries equivalent ethical approvals were done. My involvement was supervising, discussing, verifying and re-executing the analytical framework, and presentation of results.

8.3.2 Data description and analysis

The data comes from 756 hospitals across 69 countries.

Adult patients were included if they had Covid-19 infection between 7 days before and 30 days after surgery. The dataset is formed of basic patient information: age, sex, haemoglobin, white cell count, C-reactive protein, American Society of Anaesthesiologists (ASA) grade, Revised Cardiac Risk Index (RCRI) score, information about respiratory comorbidities, and smoking status; Covid-19 status: need of preoperative respiratory support, the timing of diagnosis before or after the surgery;

operation factors: surgical speciality, surgery indication, urgency, grade, and type of anaesthesia. The primary outcome was mortality at 30 days after surgery.

Variables with more than 20% missing values were ignored from the analysis. Data were imputed using MICE (109, 110). Different algorithms were applied and used to identify important predictors: generalised linear models, random forest, and elastic net (182) (184) (180). Hyperparameter search was done using a grid search over 10- and 5-fold resampling. Then, feature important was formed merging the ranked important features from both resampling, up to a maximum of 5 features.

The analysis dataset contains 8492 patients, patients had mostly abdominal surgery (40.6%), orthopaedic (33.8%), and head and neck surgery (9.8%). Most surgeries were emergency (80.8%), and 57.3% of all surgeries were for benign diseases. The overall mortality rate was 17.2%. The dataset was separated in the order of events into 6777 derivation and 1715 validation patients. The patients from the derivation cohort had surgery between February 1st and May 31st, 2020, while the validation cohort had surgery between June 1st and July 31st, 2020.

8.3.3 Results and discussion

The best performing model performed with an AUCROC of 0.73 (95%CI of 0.71-0.74) in the derivation set, and 0.80 (0.77-0.83) in the validation set (Figure 40). The important predictors identified were age, ASA grade, RCRI score, and preoperative respiratory support.

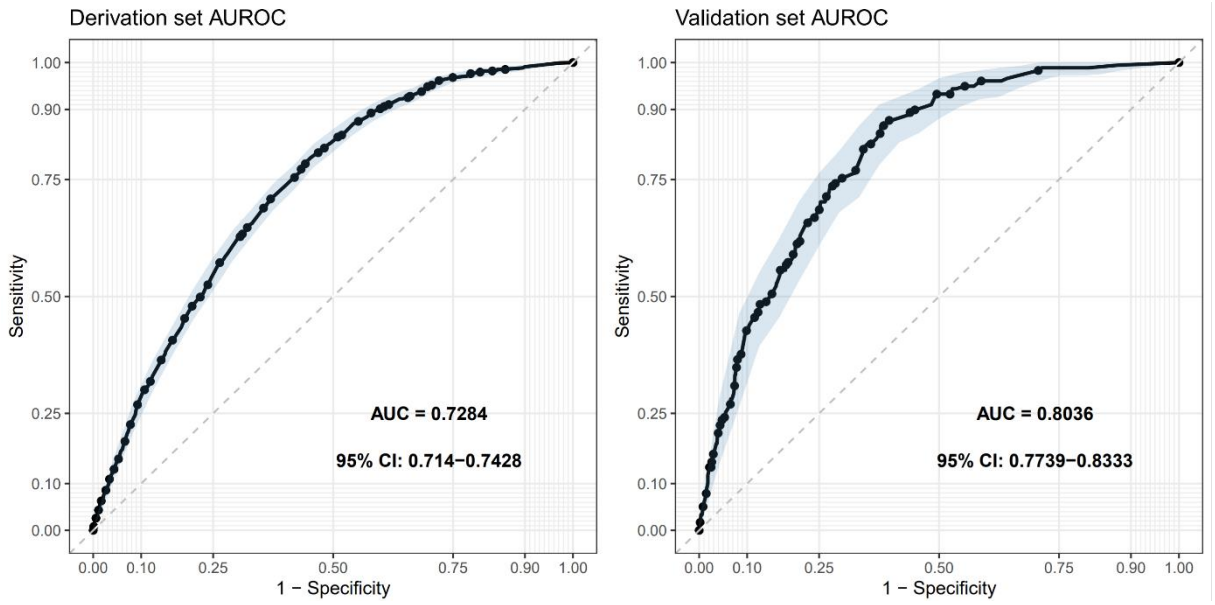


Figure 40: Comparative of AUCROC performance between derivation and validation sets for the CovidSurg case.

The developed model was made available online as a risk identification tool (407).

The created models utilise data from many countries, it is diverse and generalised. It utilises a single number of variables that can be easily obtained and applied in practice. Patient risk assessment is a paramount tool to evaluate if a surgery should be done or not. Despite the situation in some countries starting to becoming more controlled, with vaccinations and a reduced number of Covid-19 cases, it is still essential to understand the risk of operations, especially in low-income countries, where vaccination will still take years to occur (408).

The created models assist the evaluation of patient risk, can assist the identification of patients with higher mortality risk, providing a way to reducing the backlog of surgeries affected by Covid-19.

8.4 Patient rescheduling – Case 3

8.4.1 Introduction

Further to disruptions in operations, in the UHB there were disruptions to appointments as well. Clinicians that would normally be on out-patient clinic work had to support

Covid-19 hospitalisations. The overall healthcare operated into hard strains in the UK; primary care had the routine changed, with some routine procedures cancelled, and dentistry services were temporarily stopped; patients had appointments only in acute situations at the beginning of the pandemic (409) (410). This led to an accumulation of patients that should have had follow-up but could not be slotted for it.

To improve the situation of care in the UHB, we created a tool to identify patients that would be of higher risk and provided this tool as a webpage to clinicians. The main requirements are communication with other hospital systems, identification of comorbidities and other risk factors for the different patients, and the presentation of the results to clinical staff. My involvement in this work was idealising the project, collecting from hospital data sources, setting up the tool, discussing, and testing.

8.4.2 Data description and analysis

For this use case, we utilised the UHB dataset, the main body of UHB data described previously (section 2.3.4) was incremented with scheduling data from patient administration systems and clinical documents stored as EMC Documentum (411).

The tool was created as an internal website, with an application programming interface that could be accessed by other software on the network. The tool was formed of a few key elements: (a) its internal representation of patients, formed of data collected internally, and updated on demand; (b) components to extract data from other sources; and (c) web interface system.

The internal representation of patients. There are key elements required by this tool. For each clinician, there is a set of patients that they are providing specialised care, these were used for authentication and organisation purposes; a clinician was not able to see other colleagues' patients unless explicitly given permission. Patients have their age, sex and BMI collected, also the information if they had any of 65 important comorbidities (Appendix 6.1). A relation of conditions obtained from textual inferences of the patient and their families were also represented as categorical terms. These data were stored on an object format in a MongoDB server that communicates with the webpage (412).

Extraction of data from other sources. Main data elements come from the internal PICS database (21), as it contains the basic clinical information, and records of important comorbidities, these were extracted using SQL queries (189). Diseases information from documents were collected from data files available for the target patients, Komenti performs a Natural language processing over the files and identifies relevant diseases (413). These data are format into the internal representation used for the patients.

Web interface system. The overall product was created as a webpage in Python using the *gunicorn* library (414). The server was integrated into a Docker image so that this system can be run independently of the environment setup of a server (382). The different components were executed on-demand, for example, when a new clinician is registered on the system, or when this clinician needs a refresh of their patients' data. For a patient, the last appointment date was shown, with the list of comorbidities that could be relevant when prioritising their care.

8.4.3 Results and discussion

This tool showed some capabilities of the UHB infrastructure and the group to find solutions to the urgent needs of the hospital. It provided a system to identify patients that could require a priority of care needs.

At the time of writing this tool has been used mainly by a clinician that is our UHB point of contact, but still requires validation from more practitioners to evaluate if there is a benefit to both the speed and the efficiency of care.

8.5 Chapter summary

This chapter explored three initiatives to challenges of clinical care during the Covid-19 pandemic. First and foremost, there were many studies published in the literature to assess patient mortality risk in hospitals. While some models performed well in their derivation cohorts they could not be used more extensively. A model formed of an ensemble of other models performed more consistently over different patient cohorts. The newly formed model and the approach employed can be further combined with other models and used to assess patient's risk, as, at the time of writing, there are still

a large number of hospitalisations, especially in countries that are only now going through the second wave of the disease. The second use case investigated a way to alleviate the situation of delayed surgeries, with the creation of a model that could improve clinical decision of patient risk undergoing surgery. In the last case, we focused on handling the backlog of patient appointments, some patients had higher complication risks and a way to prioritise then was implemented. These different initiatives focused on different aspects of the pandemic and provided different solutions to the challenges seen.

CHAPTER 9 CONCLUSION

Data come in different shapes and sizes, unstructured, semi-structured and structured, from the very fine graded genotype data to population scales. The major aim of this work was to investigate different methodological approaches to different data types to the better understanding and risk assessment of patients disease, with a particular focus on cardiovascular diseases.

9.1 Investigations and outcomes

Structured datasets were explored in chapter 3, different methods were applied to the better identification of patients with AF. Factors of risk from the literature were re-identified: morbidities of risk, age, BMI, and male sex. Novel biomarker predictors were identified, and risk models with these biomarkers were created, showing the potential of these biomarkers to the improvement of patient stratification. The work employed in this chapter led to a few published/under-review outcomes (4) (5) (6).

Omics datasets, examined on chapter 4, were explored to both assess the potential identifying novel AF biomarkers. Due to the fact that locus 4q25, adjacent to gene *PITX2*, showed very significant differences between healthy and AF patients, investigations on their pathophysiological patterns were warranted. The initial case examined the differences on mice that were +/- knockout on *PITX2*. This study showed a few protein-coding transcripts with possible implications to AF. Two of these transcripts are protein-encoding. The proteins CXCL13 and BMP10 were of increased importance due to their solubility and of potential biomarker value (8).

Unstructured datasets, investigated in chapter 5, form a data type that is not commonly explored to the full extent in the literature. Using a combination of signal processing and machine learning, biomarkers were extracted from the different leads of an ECG recording. Novel approaches to the use of ECGs to predict HF were explored in real-world data with promising results. HF is commonly diagnosed using imagining techniques, and the risk is usually assessed using models such as age and NTproBNP. The use of ECG recordings to support HF diagnosis is another tool that can be added to the cardiologist clinical toolset (11).

Population data, chapter 6, were explored to better understand patient cohorts and stratify for potential targeted treatment. While the other chapters studied different data modalities and their application to different use cases within the cardiovascular space, this chapter explored the use of larger scale cohorts, aiming to explore patterns that transcend cardiovascular morbidities – to converge back in cardiovascular patterns and a multitude of other morbidity relations. A multitude of associations and morbidity patterns were identified linking cardiovascular and other diseases patterns.

In chapter 7, an approach to combine datasets was proposed. The proposed framework was employed in a number of cases (sections 4.6, 5.4 and 6.5) to handle the extraction of a complex number of variables and elements from diverse datasets. The first case relates to how complex multivariate datasets can be compiled. The second case exposed a logical framework to enable the use of the proposed solution to real-world, routine healthcare systems. The third case illustrated the collection of data related to the date of patients first diagnosis, where the date of one event was used as reference to the collection of other data.

Chapter 8 focused on the unprecedented pandemics and its effect on cardiovascular patients. Analytical pipelines developed for different applications were repurposed to respond to the real-time need of the pandemic, demonstrating the power and flexibility of these approaches. The approaches employed to support the crisis led to a good extent of results (12) (13) (14).

9.2 Limitations and future work

The work on this thesis focused on a variety of data sources, including clinical trial data and routinely collected secondary datasets, model organisms datasets and biological data related to human participants. The approaches employed can be applied beyond the datasets used in this thesis, however further experimentation is required to validate the results and expand the usability of data methods.

The use of data is a problem of its own. It is not trivial to collect and get data into a shape that can be analysed. Data that were originally collected to answer a scientific research question are often not suitable for interrogation of unplanned analysis. There

are assumptions that need to be made and reported. Such assumptions get even more complicated on real-world dataset, where a curated dataset might not be available, the data might be situated in different silos and in different modalities. This is a line of work that can be further explored. How to make data simpler, and more traceable.

There is an ever-growing number of risk models created in the medical domain. As explored in this thesis, a risk model can be *just a mouse click away*. There is limited knowledge on how they work, how to make the best use of them, and apply them. That is not even considering how to make models properly. These are problems that can be better explored, especially when the care is more integrated with research.

Methodologies employed to some problems can be translated to other problems of equivalent data scales and modalities. For example, the analysis that employed unstructured data, the ECG signals, can be applied on electroencephalography to assess brain function; analysis using biomarkers can be applied to a variety of other biological hypothesis; techniques to the integration of datasets can be further expanded to cover different data scenarios.

Data and models can also be thought as different building blocks, and one is limited without the other. These two elements could be combined, and tools that can handle the different analytical steps could be integrated to quicker evaluation of research questions and enable a more widespread use.

REFERENCES

1. Acharjee A, Larkman J, Xu Y, Cardoso VR, Gkoutos GV. A random forest based biomarker discovery and power analysis framework for diagnostics research. BMC medical genomics. 2020;13(1):1-14.

2. Chua W, Cardoso VR, Purmah Y, Tull S, Neculau G, Gkoutos GV, et al. P1184Blood biomarkers associated with atrial fibrillation in a community-based cohort of patients presenting acutely to hospital. EP Europace. 2018;20(suppl_1):i229-i.

3. Chua W, Cardoso V, Purmah Y, Crijns H, Schotten U, Guasch E, et al. 64 A multiple blood biomarker model for identifying patients with prevalent AF. BMJ Publishing Group Ltd and British Cardiovascular Society; 2020.

4. Chua W, Purmah Y, Cardoso VR, Gkoutos GV, Tull SP, Neculau G, et al. Data-driven discovery and validation of circulating blood-based biomarkers associated with prevalent atrial fibrillation. European heart journal. 2019;40(16):1268-76.

5. Chua W, Law JP, Cardoso VR, Purmah Y, Neculau G, Jawad-Ul-Qamar M, et al. Quantification of fibroblast growth factor 23 and N-terminal pro-B-type natriuretic peptide to identify patients with atrial fibrillation using a high-throughput platform: A validation study. PLoS medicine. 2021;18(2):e1003405.

6. Chua W, Roth Cardoso V, Guasch E, Sinner MF, Brady P, Casadei B, et al. A Novel Biomarker Model for Detecting Patients With Atrial Fibrillation: A Development and Validation Study.

7. Hepburn C, Syeda F, Yu T, Holmes AP, Roth VC, Wright T, et al. 128 Desmosomal instability increases atrial arrhythmia susceptibility after endurance training. Heart. 2018;104(Suppl 6):A95-A6.

8. Reyat JS, Chua W, Cardoso VR, Witten A, Kastner PM, Kabir SN, et al. Reduced left atrial cardiomyocyte PITX2 and elevated circulating BMP10 predict atrial fibrillation after ablation. JCI insight. 2020;5(16).

9. Gill S, Sartini C, Uh H, Ghoreishi N, Cardoso V, Bunting K, et al. Accurate detection of atrial fibrillation using a smartphone remains uncertain: a systematic review and meta-analysis. European Heart Journal. 2020;41(Supplement_2):ehaa946. 3505.

10. Gill S, Bunting K, Sartini C, Roth Cardoso V, Ghoreishi N, Uh HW, et al. Smartphone detection of atrial fibrillation using photoplethysmography: A systematic review and meta-analysis.

11. Roth Cardoso V, Gill S, Bunting K, Tica O, Kolias V, Karwath A, et al. Validated neural network in routine clinical practice to identify incident heart failure using digital electrocardiograms (cardAIc-ECG).

12. Wu H, Zhang H, Karwath A, Ibrahim Z, Shi T, Zhang X, et al. Ensemble learning for poor prognosis predictions: a case study on SARS-CoV2. Journal of the American Medical Informatics Association. 2020.

13. Carr E, Bendayan R, Bean D, Stammers M, Wang W, Zhang H, et al. Evaluation and Improvement of the National Early Warning Score (NEWS2) for COVID-19: a multi-hospital study. BMC medicine. 2021;19(1):1-16.

14. Collaborative C. Machine learning risk prediction of mortality for patients undergoing surgery with perioperative SARS-CoV-2: the COVIDSurg mortality score. The British journal of surgery. 2021.

15. Cardoso VR. gkoutos-group/postcode. 1.0.0 ed: Zenodo; 2021.

16. Cardoso VR. gkoutos-group/bbcaf_pipeline. 1.0.1 ed: Zenodo; 2021.

17. Cardoso VR. gkoutos-group/clustering. v1.0.0 ed: Zenodo; 2021.

18. Mark JJ. Writing 2011 [Available from: https://www.ancient.eu/writing.

19. Elmasri R, Navathe SB. Fundamentals of Database Systems: Pearson; 2015.

20. Evans R. Electronic health records: then, now, and in the future. Yearbook of medical informatics. 2016;25(S 01):S48-S61.

21. Trust NF. Birmingham Systems Prescribing Information and Communications System (PICS). Birmingham2012.

22. Authority HR. NHS Health Research Authority London2021 [Available from: https://www.hra.nhs.uk/.

23. Hurwitz JS, Nugent A, Halper F, Kaufman M. Big data for dummies: John Wiley & Sons; 2013.

24. Organization WH. The ICD-10 classification of mental and behavioural disorders: diagnostic criteria for research: World Health Organization; 1993.

25. Manogaran G, Thota C, Lopez D, Vijayakumar V, Abbas KM, Sundarsekar R. Big data knowledge system in healthcare. Internet of things and big data technologies for next generation healthcare: Springer; 2017. p. 133-57.

26. Standrin S, Borley NR, Collins P, Crossman AR, Gatzoulis MA, Healy JC, et al. Heart and great vessels. Gray's Anatomy: The Anatomical Basis of Clinical Practice: Elsevier Churchill Livingstone; 2008. p. 959-87.

27. Robotham JL, Takata M, Berman M, Harasawa Y. Ejection fraction revisited. Anesthesiology: The Journal of the American Society of Anesthesiologists. 1991;74(1):172-83.

28. Nattel S. New ideas about atrial fibrillation 50 years on. Nature. 2002;415:219-26.

29. Kirchhof P, Benussi S, Kotecha D, Ahlsoon A, Atar D, Casadei B, et al. 2016 ESC Guidelines for the management of atrial fibrillation developed in collaboration with EACTS. European Heart Journal. 2016;37(38):2893-962.

30. Hu Y-F, Chen Y-J, Lin Y-J, Chen S-A. Inflammation and the pathogenesis of atrial fibrillation. Nature Reviews Cardiology. 2015;12(4):230.

31. Zoni-Berisso M, Lercari F, Carazza T, Domenicucci S. Epidemiology of atrial fibrillation: European perspective. 2014.

32. Ponikowski P, Voors AA, Anker SD, Bueno H, Cleland JGF, Coats AJS, et al. 2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: The Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC)Developed with the special contribution of the Heart Failure Association (HFA) of the ESC. European Heart Journal. 2016;37(27):2129-200.

33. Mackenzie J. Diseases of the Heart, 3rd edn. London: Frowde. Hodder and Stoughton. 1914;101:102-3.

34. Anter E, Jessup M, Callans DJ. Atrial fibrillation and heart failure: treatment considerations for a dual epidemic. Circulation. 2009;119(18):2516-25.

35. Fabritz L, Guasch E, Antoniades C, Bardinet I, Benninger G, Betts TR, et al. Defining the major health modifiers causing atrial fibrillation: a roadmap to underpin personalized prevention and treatment. Nature Reviews Cardiology. 2016;13:230-7.

36. Gage BF, Van Walraven C, Pearce L, Hart RG, Koudstaal PJ, Boode B, et al. Selecting patients with atrial fibrillation for anticoagulation: stroke risk stratification in patients taking aspirin. Circulation. 2004;110(16):2287-92.

37. Lip GY, Nieuwlaat R, Pisters R, Lane DA, Crijns HJ. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. Chest. 2010;137(2):263-72.

38. Association NYH. Diseases of the heart and blood vessels: nomenclature and criteria for diagnosis: Little, Brown; 1964.

39. Einthoven W, Fahr G, Waart Ad. On the direction and manifest size of the variations of potential in the human heart and on the influence of the position of the heart on the form of the electrocardiogram. American Heart Journal. 1950;40(2):163-211.

40. Simonson E, BLACKBURN JR H, PUCHNER TC, Eisenberg P, Ribeiro F, Meja M. Sex differences in the electrocardiogram. Circulation. 1960;22(4):598-601.

41. Bessem B, de Bruijn MC, Nieuwland W. Gender differences in the electrocardiogram screening of athletes. Journal of science and medicine in sport. 2017;20(2):213-7.

42. Meek S, Morris F. ABC of clinical electrocardiography: introduction. I—Leads, rate, rhythm, and cardiac axis. BMJ: British Medical Journal. 2002;324(7334):415.

43. Kligfield P, Gettes LS, Bailey JJ, Childers R, Deal BJ, Hancock EW, et al. Recommendations for the standardization and interpretation of the electrocardiogram: part I: the electrocardiogram and its technology a scientific statement from the American Heart Association Electrocardiography and Arrhythmias Committee, Council on Clinical Cardiology; the American College of Cardiology Foundation; and the Heart Rhythm Society endorsed by the International Society for Computerized Electrocardiology. Journal of the American College of Cardiology. 2007;49(10):1109-27.

44. Mason JW, Hancock EW, Gettes LS. Recommendations for the standardization and interpretation of the electrocardiogram: part II: electrocardiography diagnostic

statement list a scientific statement from the American Heart Association Electrocardiography and Arrhythmias Committee, Council on Clinical Cardiology; the American College of Cardiology Foundation; and the Heart Rhythm Society Endorsed by the International Society for Computerized Electrocardiology. Journal of the American College of Cardiology. 2007;49(10):1128-35.

45. Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. circulation. 2000;101(23):e215-e20.

46. Moody G. A new method for detecting atrial fibrillation using RR intervals. Computers in Cardiology. 1983:227-30.

47. Fujiki A, Yoshioka R, Sakabe M, Kusuzaki S. QT/RR relation during atrial fibrillation based on a single beat analysis in 24-h Holter ECG: The role of the second and further preceding RR intervals in QT modification. Journal of Cardiology. 2011;57(3):269-74.

48. Trigo JD, Alesanco Á, Martínez I, García J. A review on digital ECG formats and the relationships between them. IEEE Transactions on Information Technology in Biomedicine. 2011;16(3):432-44.

49. Lu X, Duan H, Zheng H, editors. XML-ECG: An XML-based ECG presentation for data exchanging. 2007 1st International Conference on Bioinformatics and Biomedical Engineering; 2007: IEEE.

50. Watson JD, Crick FH. Molecular structure of nucleic acids. Nature. 1953;171(4356):737-8.

51. Lane L, Bairoch A, Beavis RC, Deutsch EW, Gaudet P, Lundberg E, et al. Metrics for the Human Proteome Project---2013-2014 and Strategies for Finding Missing Proteins. J Proteome Res. 2014;13(1):15-20.

52. Kukurba KR, Montgomery SB. RNA sequencing and analysis. Cold Spring Harbor Protocols. 2015;2015(11):pdb. top084970.

53. Toxicity Co. Genomics, Transcriptomics and Proteomics: Glossary of Terms 2020 [Available from: https://cot.food.gov.uk/committee/committee-on-toxicity/cotmtgs/cotmtsem/cotsem1001/49831.

54. Brown TA. Genome 3: Garland Science Publishing; 2007.

55. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biology. 2016;17(13).

56. Kirchhof P, Kahr PC, Kaese S, Piccini I, Vokshi I, Scheld H-H, et al. PITX2c is expressed in the adult left atrium, and reducing Pitx2c expression promotes atrial fibrillation inducibility and complex changes in gene expression. Circulation: Cardiovascular Genetics. 2011;4(2):123-33.

57. Li J, Swope D, Raess N, Cheng L, Muller E, Radice G. Cardiac tissue-restricted deletion of plakoglobin results in progressive cardiomyopathy and activation of {beta}-catenin signaling. Molecular Cell Biology. 2011;31(6):1134-44.

58. Reed E, Nunez S, Kulp D, Qian J, Reilly MP, Foulkes AS. A guide to genome-wide association analysis and post-analytic interrogation. Statistics in Medicine. 2015;34(28):3769-92.

59. Shah SH, Arnett D, Houser SR, Ginsburg GS, MacRae C, Mital S, et al. Opportunities for the cardiovascular community in the precision medicine initiative. Circulation. 2016;133(2):226-31.

60. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. Journal of clinical epidemiology. 2016;74:167-76.

61. Thabane L, Mbuagbaw L, Zhang S, Samaan Z, Marcucci M, Ye C, et al. A tutorial on sensitivity analyses in clinical trials: the what, why, when and how. BMC medical research methodology. 2013;13(1):92.

62. Hullman J, Diakopoulos N. Visualization rhetoric: Framing effects in narrative visualization. IEEE transactions on visualization and computer graphics. 2011;17(12):2231-40.

63. Borkin MA, Bylinskii Z, Kim NW, Bainbridge CM, Yeh CS, Borkin D, et al. Beyond memorability: Visualization recognition and recall. IEEE transactions on visualization and computer graphics. 2015;22(1):519-28.

64. Segel E, Heer J. Narrative visualization: Telling stories with data. IEEE transactions on visualization and computer graphics. 2010;16(6):1139-48.

65. Pontis S, Babwahsingh M. Improving information design practice: A closer look at conceptual design methods. Information Design Journal. 2016;22(3):249-65.

66. Peck EM, Ayuso SE, El-Etr O, editors. Data is personal: Attitudes and perceptions of data visualization in rural pennsylvania. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems; 2019.

67. Schiewe J. Empirical studies on the visual perception of spatial patterns in choropleth maps. KN-Journal of Cartography and Geographic Information. 2019;69(3):217-28.

68. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. British Journal of Surgery. 2015;102(3):148-58.

69. Hayden JA, van der Windt DA, Cartwright JL, Côté P, Bombardier C. Assessing bias in studies of prognostic factors. Annals of internal medicine. 2013;158(4):280-6.

70. Schulz KF, Altman DG, Moher D, Group C. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. Trials. 2010;11(1):32.

71. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. Journal of clinical epidemiology. 2009;62(10):e1-e34.

72. Rivera SC, Liu X, Chan A-W, Denniston AK, Calvert MJ. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. bmj. 2020;370.

73. Banerjee A, Chen S, Fatemifar G, Zeina M, Lumbers RT, Mielke J, et al. Machine learning for subtype definition and risk prediction in heart failure, acute coronary syndromes and atrial fibrillation: systematic review of validity and clinical utility. BMC medicine. 2021;19(1):1-14.

74. Casselman F, Coca A, De Caterina R, Deftereos S, Dobrev D, Ferro JM, et al. 2016 ESC Guidelines for the management of atrial fibrillation developed in collaboration with EACTS. European Heart Journal. 2016;37:2893-962.

75. Proteomics O. Proseek Multiplex CVD II panel | Olink. Olink.

76. Nasreddine ZS, Phillips NA, Bédirian V, Charbonneau S, Whitehead V, Collin I, et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. Journal of the American Geriatrics Society. 2005;53(4):695-9.

77. Chua W, Purmah Y, Cardoso VR, Gkoutos GV, Tull SP, Neculau G, et al. Data-driven discovery and validation of circulating blood-based biomarkers associated with prevalent atrial fibrillation. European Heart Journal. 2019;40(16):1268-76.

78. Chua W, Easter CL, Guasch E, Sitch A, Casadei B, Crijns HJ, et al. Development and external validation of predictive models for prevalent and recurrent atrial fibrillation: a protocol for the analysis of the CATCH ME combined dataset. BMC cardiovascular disorders. 2019;19(1):1-9.

79. Kinkorova J. Horizon 2020, new EU Framework programme for research and innovation, 2014-2020. Casopis lekaru ceskych. 2014;153(5):254-6.

80. Kirchhof P, Andresen D, Bosch R, Borggrefe M, Meinertz T, Parade U, et al. Short-term versus long-term antiarrhythmic drug treatment after cardioversion of atrial fibrillation (Flec-SL): a prospective, randomised, open-label, blinded endpoint assessment trial. The Lancet. 2012;380(9838):238-46.

81. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS medicine. 2015;12(3).

82. Manolio TA, Weis BK, Cowie CC, Hoover RN, Hudson K, Kramer BS, et al. New models for large prospective studies: is there a better way? American journal of epidemiology. 2012;175(9):859-66.

83. Biobank U. UKB: UKB Showcase 2015 [Available from: https://www.ukbiobank.ac.uk/.

84. Organization WH. International statistical classification of diseases and related health problems: World Health Organization; 2004.

85. Slee VN. The International classification of diseases: ninth revision (ICD-9). American College of Physicians; 1978.

86. Chengode S. Left ventricular global systolic function assessment by echocardiography. Annals of cardiac anaesthesia. 2016;19(Suppl 1):S26.

87. Birmingham UH. University Hospitals Birmingham 2020 [Available from: https://www.uhb.nhs.uk/home.htm.

88. HDClinical. Solus Cardiology. 2020.

89. Bray T, Paoli J, Sperberg-McQueen CM, Maler E, Yergeau F. Extensible markup language (XML) 1.0. W3C recommendation October; 2000.

90. Technical Committee : ISO/TC 171/SC 2 Document file formats Esaaoi. ISO 32000-1:2008 Document management — Portable document format — Part 1: PDF 1.7. 2008.

91. Digital N. NHS ICD-10 5th Edition data files 2016 [Available from: https://digital.nhs.uk/data-and-information.

92. Health NCf, Health Do, Centre SCI. OPCS Classifications of Interventions and Procedures: The Stationery Office; 2006.

93. Blak B, Thompson M, Dattani H, Bourke A. Generalisability of The Health Improvement Network (THIN) database: demographics, chronic disease prevalence and mortality rates. Journal of Innovation in Health Informatics. 2011;19(4):251-5.

94. Benson T. The history of the Read codes: the inaugural James Read Memorial Lecture 2011. Journal of Innovation in Health Informatics. 2011;19(3):173-82.

95. Efron B, Tibshirani RJ. An Introduction to the bootstrap: Champman & Hall/CRC; 1993.

96. Branco P, Torgo L, Ribeiro RP. A survey of predictive modeling on imbalanced domains. ACM Computing Surveys (CSUR). 2016;49(2):1-50.

97. Haixiang G, Yijing L, Shang J, Mingyun G, Yuanyue H, Bing G. Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications. 2017;73:220-39.

98. Menardi G, Torelli N. Training and assessing classification rules with imbalanced data. Data Mining and Knowledge Discovery. 2014;28(1):92-122.

99. Lunardon N, Menardi G, Torelli N. ROSE: A Package for Binary Imbalanced Learning. R journal. 2014;6(1).

100. Lusa L. SMOTE for high-dimensional class-imbalanced data. BMC bioinformatics. 2013;14(1):106.

101. Garavaglia S, Sharma A. A SMART GUIDE TO DUMMY VARIABLES: FOUR APPLICATIONS AND A MACRO2016 February. Available from: https://stats.idre.ucla.edu/wp-content/uploads/2016/02/p046.pdf.

102. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). Biometrika. 1965;52(3/4):591-611.

103. Anderson TW, Darling DA. Asymptotic theory of certain" goodness of fit" criteria based on stochastic processes. The annals of mathematical statistics. 1952:193-212.

104. Anscombe FJ. Graphs in Statistical Analysis. The American Statistician. 1973;27(1):17-21.

105. Hotelling H. Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology. 1933;24(6):417-41.

106. Maaten Lvd, Hinton G. Visualizing data using t-SNE. Journal of machine learning research. 2008;9(Nov):2579-605.

107. McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:180203426. 2018.

108. Donders ART, Heijden GJMGvd, Stijnen T, Moons KGM. Review: A gentle introduction to imputation of missing values. Journal of Clinical Epidemiology. 2006;59(10):1087-91.

109. Rubin D. Inference and Missing Data. Biometrika. 1976(63):581-90.

110. Buuren Sv, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software. 2011;45(3).

111. Witten IH, Frank E, Hall MA. Data Mining: Practical Machine Learning Tools and Techniques. Burlington: Elsevier; 2011.

112. Everitt BS, Skrondal A. Null hypothesis. The Cambridge Dictionary of Statistics. Cambridge: Cambridge University Press; 2010.

113. Sullivan GM, Feinn R. Using effect size—or why the P value is not enough. Journal of graduate medical education. 2012;4(3):279.

114. Rosner B. Hypothesis Testing: One-Sample Inference. Fundamentals of Biostatistics. Boston: Cengage Learning; 2106. p. 211-78.

115. Fisher RA. On the interpretation of χ 2 from contingency tables, and the calculation of P. Journal of the Royal Statistical Society. 1922;85(1):87-94.

116. Pearson K. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science. 1900;50(302):157-75.

117. Razali NM, Wah YB. Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. Journal of statistical modeling and analytics. 2011;2(1):21-33.

118. Kalpić D, Hlupić N, Lovrić M. Student's t-Tests. 2011.

119. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. The annals of mathematical statistics. 1947:50-60.

120. Wilcoxon F. Individual comparisons by ranking methods. Breakthroughs in statistics: Springer; 1992. p. 196-202.

121. Fisher RA. Statistical methods for research workers. Breakthroughs in statistics: Springer; 1992. p. 66-70.

122. Moré JJ. The Levenberg-Marquardt algorithm: implementation and theory. Numerical analysis: Springer; 1978. p. 105-16.

123. Ruder S. An overview of gradient descent optimization algorithms. arXiv preprint arXiv:160904747. 2016.

124. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition. 1997;30(7):1145-59.

125. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143(1):29-36.

126. Stewart J. Calculus: Cengage Learning; 2011.

127. Sun X, Xu W. Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. IEEE Signal Processing Letters. 2014;21(11):1389-93.

128. Demler OV, Pencina MJ, D'Agostino Sr RB. Misuse of DeLong test to compare AUCs for nested models. Statistics in medicine. 2012;31(23):2577-87.

129. Coefficient of Determination. The Concise Encyclopedia of Statistics. New York, NY: Springer New York; 2008. p. 88-91.

130. Mean Squared Error. The Concise Encyclopedia of Statistics. New York, NY: Springer New York; 2008. p. 337-9.

131. Brier GW. Verification of forecasts expressed in terms of probability. Monthly weather review. 1950;78(1):1-3.

132. Mitchell TM. Machine Learning: McGraw-Hill; 1997.

133. Agrawal R, Srikant R, editors. Fast Algorithms for Mining Association Rules in Large Datasets. VLDB '94 Proceeding of the 20th International Conference on Very Large Data Bases; 1994; San Francisco.

134. Jeffreys H. Scientific inference: Cambridge University Press; 1973.

135. Dougherty J, Kohavi R, Sahami M, editors. Supervised and Unsupervised Discretization of Continous Features. Machine Learning; 1995.

136. Fournier-Viger P, Gomariz A, Campos M, Thomas R, editors. Fast vertical mining of sequential patterns using co-occurrence information. Pacific-Asia Conference on Knowledge Discovery and Data Mining; 2014: Springer.

137. Fournier-Viger P, Lin JC-W, Kiran RU, Koh YS, Thomas R. A survey of sequential pattern mining. Data Science and Pattern Recognition. 2017;1(1):54-77.

138. Breiman L. Random forests. Machine learning. 2001;45(1):5-32.

139. Cortes C, Vapnik V. Support-vector networks. Machine learning. 1995;20(3):273-97.

140. Hertz J, Krogh A, Palmer RG. Introduction to the theory of neural computation. Redwood City: Addison-Wesley; 1991.

141. Cloud G. Cloud TPUs - ML accelerators for TensorFlow. Google.

142. Quinlan JR. Induction of decision trees. Machine learning. 1986;1(1):81-106.

143. Caruana R, Niculescu-Mizil A, editors. An empirical comparison of supervised learning algorithms. ICML '06 Proceedings of the 23rd international conference on Machine learning; 2006; Pittsburgh.

144. Schmidhuber J. Deep learning in neural networks: An overview. Neural networks. 2015;61:85-117.

145. Alom MZ, Taha TM, Yakopcic C, Westberg S, Sidike P, Nasrin MS, et al. The history began from alexnet: A comprehensive survey on deep learning approaches. arXiv preprint arXiv:180301164. 2018.

146. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. arXiv. 2015.

147. Parr T, Howard J. The matrix calculus you need for deep learning. arXiv preprint arXiv:180201528. 2018.

148. Nair V, Hinton GE, editors. Rectified linear units improve restricted boltzmann machines. Proceedings of the 27th international conference on machine learning (ICML-10); 2010.

149. Misra D. Mish: A Self Regularized Non-Monotonic Neural Activation Function. arXiv preprint arXiv:190808681. 2019.

150. Janocha K, Czarnecki WM. On loss functions for deep neural networks in classification. arXiv preprint arXiv:170205659. 2017.

151. Glorot X, Bengio Y, editors. Understanding the difficulty of training deep feedforward neural networks. Proceedings of the thirteenth international conference on artificial intelligence and statistics; 2010.

152. Sutskever I, Martens J, Dahl G, Hinton G, editors. On the importance of initialization and momentum in deep learning. International conference on machine learning; 2013.

153. Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization. Journal of machine learning research. 2011;12(Jul):2121-59.

154. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980. 2014.

155. He K, Girshick R, Dollár P, editors. Rethinking imagenet pre-training. Proceedings of the IEEE International Conference on Computer Vision; 2019.

156. Negrinho R, Gormley M, Gordon GJ, Patil D, Le N, Ferreira D, editors. Towards modular and programmable architecture search. Advances in Neural Information Processing Systems; 2019.

157. Sciuto C, Yu K, Jaggi M, Musat C, Salzmann M. Evaluating the search phase of neural architecture search. arXiv preprint arXiv:190208142. 2019.

158. White C, Neiswanger W, Savani Y. BANANAS: Bayesian Optimization with Neural Architectures for Neural Architecture Search. arXiv preprint arXiv:191011858. 2019.

159. Vikhar PA, editor Evolutionary algorithms: A critical review and its future prospects. 2016 International conference on global trends in signal processing, information computing and communication (ICGTSPICC); 2016: IEEE.

160. Jin H, Song Q, Hu X, editors. Auto-keras: An efficient neural architecture search system. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; 2019.

161. Tokui S, Okuta R, Akiba T, Niitani Y, Ogawa T, Saito S, et al., editors. Chainer: A deep learning framework for accelerating the research cycle. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; 2019.

162. He X, Zhao K, Chu X. AutoML: A Survey of the State-of-the-Art. arXiv preprint arXiv:190800709. 2019.

163. Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence. 2013;35(8):1798-828.

164. Sakurada M, Yairi T, editors. Anomaly detection using autoencoders with nonlinear dimensionality reduction. Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis; 2014.

165. Kingma DP, Welling M. An introduction to variational autoencoders. arXiv preprint arXiv:190602691. 2019.

166. Doersch C. Tutorial on variational autoencoders. arXiv preprint arXiv:160605908. 2016.

167. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al., editors. Generative adversarial nets. Advances in neural information processing systems; 2014.

168. Hong Y, Hwang U, Yoo J, Yoon S. How generative adversarial networks and their variants work: An overview. ACM Computing Surveys (CSUR). 2019;52(1):1-43.

169. Karras T, Laine S, Aila T, editors. A style-based generator architecture for generative adversarial networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2019.

170. Donahue C, McAuley J, Puckette M. Adversarial audio synthesis. arXiv preprint arXiv:180204208. 2018.

171. Hartmann KG, Schirrmeister RT, Ball T. EEG-GAN: Generative adversarial networks for electroencephalograhic (EEG) brain signals. arXiv preprint arXiv:180601875. 2018.

172. Shaham TR, Dekel T, Michaeli T, editors. Singan: Learning a generative model from a single natural image. Proceedings of the IEEE International Conference on Computer Vision; 2019.

173. Karras T, Aila T, Laine S, Lehtinen J. Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:171010196. 2017.

174. Bau D, Zhu J-Y, Strobelt H, Zhou B, Tenenbaum JB, Freeman WT, et al. Gan dissection: Visualizing and understanding generative adversarial networks. arXiv preprint arXiv:181110597. 2018.

175. Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Computation. 1997;9(8):1735-80.

176. Yildirim O, Baloglu UB, Tan R-S, Ciaccio EJ, Acharya UR. A new approach for arrhythmia classification using deep coded features and LSTM networks. Computer methods and programs in biomedicine. 2019;176:121-33.

177. Dargan S, Kumar M, Ayyagari MR, Kumar G. A Survey of Deep Learning and Its Applications: A New Paradigm to Machine Learning. Archives of Computational Methods in Engineering. 2019:1-22.

178. May R, Dandy G, Maier H. Review of input variable selection methods for artificial neural networks. Artificial neural networks-methodological advances and biomedical applications. 2011;10:16004.

179. Güvenir HA, Kurtcephe M. Ranking Instances by Maximing the Area under the ROC Curve. IEEE Transactions on Knowledge and Data Engineering. 2013;25(10):2356-66.

180. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology). 2005;67(2):301-20.

181. Kuhn M, With contributions from Wing J, Weston S, Williams A, Keefer C, Engelhardt A, et al. caret: Classification and Regression Training. R package version 6.0-76 ed2017.

182. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software. 2010;33(1):22.

183. Therneau T, Atkinson B, Ripley B. Recursive Partitioning and Regression Trees. R package version 4.1-11. ed2017.

184. Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002;2(3):5.

185. Ridgeway G, others cf. Generalized Boosted Regression Models. R package version 2.1.3 ed2017.

186. Lundberg SM, Lee S-I, editors. A unified approach to interpreting model predictions. Proceedings of the 31st international conference on neural information processing systems; 2017.

187. Robinson PN, Mundlos S. The human phenotype ontology. Clinical genetics. 2010;77(6):525-34.

188. Denaxas S, Gonzalez-Izquierdo A, Direk K, Fitzpatrick NK, Fatemifar G, Banerjee A, et al. UK phenomics platform for developing and validating electronic health record phenotypes: CALIBER. Journal of the American Medical Informatics Association. 2019;26(12):1545-59.

189. Melton J. Database language sql. Handbook on Architectures of Information Systems: Springer; 1998. p. 105-32.

190. Grond M, Jauss M, Hamann G, Stark E, Veltkamp R, Nabavi D, et al. Improved detection of silent atrial fibrillation using 72-hour Holter ECG in patients with ischemic stroke: a prospective multicenter cohort study. Stroke. 2013;44(12):3357-64.

191. Wachter R, Groeschel K, Gelbrich G, Hamann GF, Kermer P, Liman J, et al. Holter-electrocardiogram-monitoring in patients with acute ischaemic stroke (Find-AFRANDOMISED): an open-label randomised controlled trial. The Lancet Neurology. 2017;16(4):282-90.

192. Freedman B, Camm J, Calkins H, Healey JS, Rosenqvist M, Wang J, et al. Screening for atrial fibrillation: a report of the AF-SCREEN International Collaboration. Circulation. 2017;135(19):1851-67.

193. Adderley NJ, Nirantharakumar K, Marshall T. Risk of stroke and transient ischaemic attack in patients with a diagnosis of resolved atrial fibrillation: retrospective cohort studies. bmj. 2018;361.

194. Boriani G, Laroche C, Diemberger I, Fantecchi E, Popescu MI, Rasmussen LH, et al. Asymptomatic atrial fibrillation: clinical correlates, management, and outcomes in the EORP-AF Pilot General Registry. The American journal of medicine. 2015;128(5):509-18. e2.

195. Oto E, Okutucu S, Katircioglu-Öztürk D, Güvenir HA, Karaagaoglu E, Borggrefe M, et al. Predictors of sinus rhythm after electrical cardioversion of atrial fibrillation: results from a data mining project on the Flec-SL trial dataset. Eurospace. 2017;19(6):921-9218.

196. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2017.

197. Lunardon N, Menardi G, Torelli N. ROSE: A package for binary imbalanced learning. R Journal. 2014;6(1):11.

198. Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F. e1071: Misc Functions of the Department of Statistics. R package version 1.6-8 ed. TU Wien: Probability Theory Group (Formerly: E1071); 2017.

199. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011;12:77.

200. Patton KK, Heckbert SR, Alonso A, Bahrami H, Lima JA, Burke G, et al. N-terminal pro-B-type natriuretic peptide as a predictor of incident atrial fibrillation in the Multi-Ethnic Study of Atherosclerosis: the effects of age, sex and ethnicity. Heart. 2013;99(24):1832-6.

201. Sinner MF, Stepas KA, Moser CB, Krijthe BP, Aspelund T, Sotoodehnia N, et al. B-type natriuretic peptide and C-reactive protein in the prediction of atrial fibrillation risk: the CHARGE-AF Consortium of community-based cohort studies. Europace. 2014;16(10):1426-33.

202. Smith JG, Newton-Cheh C, Almgren P, Struck J, Morgenthaler NG, Bergmann A, et al. Assessment of conventional cardiovascular risk factors and multiple biomarkers for the prediction of incident heart failure and atrial fibrillation. Journal of the American College of Cardiology. 2010;56(21):1712-9.

203. Ho JE, Yin X, Levy D, Vasan RS, Magnani JW, Ellinor PT, et al. Galectin 3 and incident atrial fibrillation in the community. American heart journal. 2014;167(5):729-34. e1.

204. Watanabe H, Watanabe T, Sasaki S, Nagai K, Aizawa Y. Close Bidirectional Relationship Between Chronic Kidney Disease and Atrial Fibrillation: The Niigata Preventive Medicine Study. Am Heart Assoc; 2008.

205. Thomas MR, Lip GY. Novel risk markers and risk assessments for cardiovascular disease. Circulation research. 2017;120(1):133-49.

206. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8(1):118-27.

207. Leek JT, Johnson WE, Parker HS, Fertig EJ, Jaffe AE, Storey JD, et al. sva: Surrogate variable analysis. R package version. 2017;3(0):882-3.

208. Kolmogorov-Smirnov A, Kolmogorov A, Kolmogorov M. Sulla determinazione empírica di uma legge di distribuzione. 1933.

209. Federico C. Natriuretic peptide system and cardiovascular disease. Heart views: the official journal of the Gulf Heart Association. 2010;11(1):10.

210. Skau E, Henriksen E, Wagner P, Hedberg P, Siegbahn A, Leppert J. GDF-15 and TRAIL-R2 are powerful predictors of long-term mortality in patients with acute myocardial infarction. European Journal of Preventive Cardiology. 2017;24(15):1576-83.

211. Kendrick J, Cheung AK, Kaufman JS, Greene T, Roberts WL, Smits G, et al. FGF-23 associates with death, cardiovascular events, and initiation of chronic dialysis. Journal of the American Society of Nephrology. 2011;22(10):1913-22.

212. Mathew JS, Sachs MC, Katz R, Patton KK, Heckbert SR, Hoofnagle AN, et al. Fibroblast growth factor-23 and incident atrial fibrillation: the Multi-Ethnic Study of Atherosclerosis (MESA) and the Cardiovascular Health Study (CHS). Circulation. 2014;130(4):298-307.

213. Alonso A, Misialek JR, Eckfeldt JH, Selvin E, Coresh J, Chen LY, et al. Circulating fibroblast growth factor-23 and the incidence of atrial fibrillation: the Atherosclerosis Risk in Communities study. Journal of the American Heart Association. 2014;3(5):e001082.

214. Faul C, Amaral AP, Oskouei B, Hu M-C, Sloan A, Isakova T, et al. FGF23 induces left ventricular hypertrophy. The Journal of clinical investigation. 2011;121(11).

215. Chollet, François, al. e. Keras. 2015.

216. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12:2825-30.

217. Nair V, Hinton GE. Rectified linear units improve restricted boltzmann machines. Proceedings of the 27th International Conference on International Conference on Machine Learning; Haifa, Israel. 3104425: Omnipress; 2010. p. 807-14.

218. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. J Mach Learn Res. 2014;15:1929-58.

219. Kingma DP, Ba J. Adam: A method for stochastic optomization. Conference paper at the 3rd International Conference for Learning Representations, San Diego. 2017.

220. Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. Adv Neur In. 2017;30.

221. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. bmj. 2016;352:i6.

222. Pöss J, Ukena C, Kindermann I, Ehrlich P, Fuernau G, Ewen S, et al. Angiopoietin-2 and outcome in patients with acute decompensated heart failure. Clinical Research in Cardiology. 2015;104(5):380-7.

223. Freestone B, Chong AY, Lim HS, Blann A, Lip GY. Angiogenic factors in atrial fibrillation: a possible role in thrombogenesis? Annals of medicine. 2005;37(5):365-72.

224. Morrell NW, Bloch DB, Ten Dijke P, Goumans M-JT, Hata A, Smith J, et al. Targeting BMP signalling in cardiovascular disease and anaemia. Nature Reviews Cardiology. 2016;13(2):106.

225. O'Neal WT, Qureshi W, Judd SE, Glasser SP, Ghazi L, Pulley L, et al. Perceived stress and atrial fibrillation: the REasons for geographic and racial differences in stroke study. Annals of Behavioral Medicine. 2015;49(6):802-8.

226. Severino P, Mariani MV, Maraone A, Piro A, Ceccacci A, Tarsitani L, et al. Triggers for atrial fibrillation: the role of anxiety. Cardiology Research and Practice. 2019;2019.

227. Statistics OfN. 2011 Census aggregate data. UK Data Service. 2016.

228. Turner-Bowker D, Hogue S. Short form 12 health survey (SF-12). Encyclopedia of Quality of Life and Well-Being Research Edn Dordrecht: Springer Netherlands. 2014:5954-7.

229. Herdman M, Gudex C, Lloyd A, Janssen M, Kind P, Parkin D, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Quality of life research. 2011;20(10):1727-36.

230. Sweet D. Office for National Statistics. National Records of Scotland; Northern Ireland Statistics and Research Agency. Social Trends 41–Health data.'Life expectancy at birth and age 65 and infant and neonatal mortality rate United Kingdom'. 2011.

231. Police.uk. data.police.uk 2020 [Available from: https://data.police.uk/.

232. Statistics OoN. English indices of deprivation 2015: UK Government; 2015 [Available from: https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015.

233. Statistics OoN. Income estimates for small areas, England and Wales: UK Government; 2020 [Available from: https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandworkinghours/datasets/smallareaincomeestimatesformiddlelayersuperoutputareasenglandandwales.

234. Archives TN. Open Government Licence for public sector information 2020 [Available from: http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/.

235. Charlton J, Rudisill C, Bhattarai N, Gulliford M. Impact of deprivation on occurrence, outcomes and health care costs of people with multiple morbidity. Journal of health services research & policy. 2013;18(4):215-23.

236. Statistics OoN. Postcode to Output Area Hierarchy with Classifications (August 2018) Lookup in the UK: UK Government; 2018 [Available from: https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015.

237. Cardoso VR. Postcode/Integrator 2018 [Available from: https://www.github.com/gkoutos-groups/postcode.

238. Chua W, Easter CL, Guasch E, Sitch A, Casadei B, Crijns HJ, et al. Development and external validation of predictive models for prevalent and recurrent atrial fibrillation: a protocol for the analysis of the CATCH ME combined dataset. BMC cardiovascular disorders. 2019;19(1):120.

239. StataCorp L. Stata statistical software: Release 15. 2017.

240. Benjamin E, Levy D, Vaziri S, D'Agostino R, Belanger A, Wolf P. Independent risk factors for atrial fibrillation in a population-based cohort. The Framingham Heart Study. JAMA. 1994:271(11):840-4.

241. Micheel CM, Nass SJ, Omenn GS. Omics-based clinical discovery: Science, technology, and applications. Evolution of Translational Omics: Lessons Learned and the Path Forward: National Academies Press (US); 2012.

242. Breschi A, Gingeras TR, Guigó R. Comparative transcriptomics in human and mouse. Nature Reviews Genetics. 2017;18(7):425.

243. Idle JR, Gonzalez FJ. Metabolomics. Cell metabolism. 2007;6(5):348-51.

244. Slonim DK, Yanai I. Getting Started in Gene Expression Microarray Analysis. PLoS Computational Biology. 2009;5(10).

245. Illumina. Sequencing and array-based solutions for genetic research 2020 [Available from: https://www.illumina.com/.

246. Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 2010;38(6):1767-71.

247. Schroeder A, Mueller O, Stocker S, Salowsky R, Leiber M, Gassmann M, et al. The RIN: an RNA integrity number for assigning integrity values to RNA measurements. BMC molecular biology. 2006;7(1):1-14.

248. Romero IG, Pai AA, Tung J, Gilad Y. RNA-seq: impact of RNA degradation on transcript quantification. BMC biology. 2014;12(1):1-13.

249. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010.

250. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics. 2014.

251. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.

252. Anders S, Pyl PT, Huber W. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31(2):166-9.

253. Kim D, Langmead B, Salzber S. HISAT: a fast spliced aligner with low memory requirements. Nature Methods. 2015;12:357-60.

254. Anders S, Huber W. Differential expression analysis for sequence count data. Nature Precedings. 2010:1-.

255. Saurin A. BioTools: ENSEMBL Gene ID to Gene Symbol Converter 2020 [Available from: https://www.biotools.fr/mouse/ensembl_symbol_converter.

256. Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen H-C, Kitts PA, et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome research. 2017;27(5):849-64.

257. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. UCSC browser. Genome Research. 2002;12:996-1006.

258. Cunningham F, Achuthan P, Akanni W, Allen J, Amode MR, Armean IM, et al. Ensembl 2019. Nucleic acids research. 2019;47(D1):D745-D51.

259. Cardoso VR. rna-seq-scripts 2019 [Available from: https://github.com/gkoutos-group/rna-seq-scripts.

260. Reimand J, Isserlin R, Voisin V, Kucera M, Tannus-Lopes C, Rostamianfar A, et al. Pathway enrichment analysis and visualization of omics data using g: Profiler, GSEA, Cytoscape and EnrichmentMap. Nature protocols. 2019;14(2):482-517.

261. Bush WS, Moore JH. Genome-wide association studies. PLoS Comput Biol. 2012;8(12):e1002822.

262. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American journal of human genetics. 2007;81(3):559-75.

263. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological). 1995;57(1):289-300.

264. Gudbjartsson DF, Arnar DO, Helgadottir A, Gretarsdottir S, Holm H, Sigurdsson A, et al. Variants conferring risk of atrial fibrillation on chromosome 4q25. Nature. 2007;448(7151):353-7.

265. Lu M-F, Pressman C, Dyer R, Johnson RL, Martin JF. Function of Rieger syndrome gene in left–right asymmetry and craniofacial development. Nature. 1999;401(6750):276-8.

266. Kahr PC, Piccini I, Fabritz L, Greber B, Schöler H, Scheld HH, et al. Systematic analysis of gene expression differences between left and right atria in different mouse strains and in human atrial tissue. PloS one. 2011;6(10):e26389.

267. Ninni S, Ortmans S, Seunes C, Coisne A, Marechal X, Modine T, et al. Postoperative atrial fibrillation following cardiac surgery is associated with pre-existing inflammatory dysregulation in epicardial adipose tissue. Insights from transcriptomic analysis. Archives of Cardiovascular Diseases Supplements. 2019;11(1):105-6.

268. Waehre A, Halvorsen B, Yndestad A, Husberg C, Sjaastad I, Nygård Sl, et al. The Homeostatic Chemokine CXCL13 and Its Receptor CXCR5 Are Regulated in Heart Failure and Are Involved in Cardiac Remodelling. Am Heart Assoc; 2009.

269. Chen H, Shi S, Acosta L, Li W, Lu J, Bao S, et al. BMP10 is essential for maintaining cardiac growth during murine cardiogenesis. Development. 2004;131(9):2219-31.

270. Fingert JH, Stone EM, Sheffield VC, Alward WL. Myocilin glaucoma. Survey of ophthalmology. 2002;47(6):547-61.

271. Metherell LA, Chapple JP, Cooray S, David A, Becker C, Rüschendorf F, et al. Mutations in MRAP, encoding a new interacting partner of the ACTH receptor, cause familial glucocorticoid deficiency type 2. Nature genetics. 2005;37(2):166-70.

272. Huang X, Feng Z, Jiang Y, Li J, Xiang Q, Guo S, et al. VSIG4 mediates transcriptional inhibition of Nlrp3 and Il-1β in macrophages. Science advances. 2019;5(1):eaau7426.

273. Austin KM, Trembley MA, Chandler SF, Sanders SP, Saffitz JE, Abrams DJ, et al. Molecular mechanisms of arrhythmogenic cardiomyopathy. Nature Reviews Cardiology. 2019;16(9):519-37.

274. Garrod D, Chidgey M. Desmosome structure, composition and function. Biochimica et Biophysica Acta (BBA)-Biomembranes. 2008;1778(3):572-87.

275. Fabritz L, Hoogendijk MG, Scicluna BP, Van Amersfoorth SC, Fortmueller L, Wolf S, et al. Load-reducing therapy prevents development of arrhythmogenic right ventricular cardiomyopathy in plakoglobin-deficient mice. Journal of the American College of Cardiology. 2011;57(6):740-50.

276. Bang M-L, Gu Y, Dalton ND, Peterson KL, Chien KR, Chen J. The muscle ankyrin repeat proteins CARP, Ankrd2, and DARP are not essential for normal cardiac development and function at basal conditions and in response to pressure overload. PloS one. 2014;9(4):e93638.

277. Jasnic-Savovic J, Nestorovic A, Savic S, Karasek S, Vitulo N, Valle G, et al. Profiling of skeletal muscle Ankrd2 protein in human cardiac tissue and neonatal rat cardiomyocytes. Histochemistry and Cell Biology. 2015;143(6):583-97.

278. Reza N, Garg A, Merrill SL, Chowns JL, Rao S, Owens AT. ACTA1 novel likely pathogenic variant in a family with dilated cardiomyopathy. Circulation: Genomic and Precision Medicine. 2018;11(10):e002243.

279. Marian A, Roberts R. The molecular genetic basis for hypertrophic cardiomyopathy. Journal of molecular and cellular cardiology. 2001;33(4):655-70.

280. Rowin EJ, Hausvater A, Link MS, Abt P, Gionfriddo W, Wang W, et al. Clinical profile and consequences of atrial fibrillation in hypertrophic cardiomyopathy. Circulation. 2017;136(25):2420-36.

281. Andrade JG, Deyell MW, Lee AY, Macle L. Sex differences in atrial fibrillation. Canadian Journal of Cardiology. 2018;34(4):429-36.

282. Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nature communications. 2019;10(1):1-10.

283. C. Wickramarachchi D, Theofilopoulos AN, Kono DH. Immune pathology associated with altered actin cytoskeleton regulation. Autoimmunity. 2010;43(1):64-75.

284. Nielsen JB, Thorolfsdottir RB, Fritsche LG, Zhou W, Skov MW, Graham SE, et al. Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. Nature genetics. 2018;50(9):1234-9.

285. Efron B. Computers and the theory of statistics: thinking the unthinkable. SIAM review. 1979;21(4):460-80.

286. Haworth S, Mitchell R, Corbin L, Wade KH, Dudding T, Budu-Aggrey A, et al. Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis. Nature communications. 2019;10(1):1-9.

287. Weng L-C, Preis SR, Hulme OL, Larson MG, Choi SH, Wang B, et al. Genetic predisposition, clinical risk factor burden, and lifetime risk of atrial fibrillation. Circulation. 2018;137(10):1027-38.

288. Roselli C, Chaffin MD, Weng L-C, Aeschbacher S, Ahlberg G, Albert CM, et al. Multi-ethnic genome-wide association study for atrial fibrillation. Nature genetics. 2018;50(9):1225-33.

289. Shi Jing L, Fathiah Muzaffar Shah F, Saberi Mohamad M, Moorthy K, Deris S, Zakaria Z, et al. A review on bioinformatics enrichment analysis tools towards functional analysis of high throughput gene set data. Current Proteomics. 2015;12(1):14-27.

290. Franke L, Jansen RC. eQTL analysis in humans. Cardiovascular Genomics: Springer; 2009. p. 311-28.

291. Diogo D, Tian C, Franklin CS, Alanne-Kinnunen M, March M, Spencer CC, et al. Phenome-wide association studies across large population cohorts support drug target validation. Nature communications. 2018;9(1):1-13.

292. Bishop CM. Pattern recognition and machine learning: springer; 2006.

293. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556. 2014.

294. Redmon J, Farhadi A. Yolov3: An incremental improvement. arXiv preprint arXiv:180402767. 2018.

295. Krittanawong C, Johnson KW, Rosenson RS, Wang Z, Aydar M, Baber U, et al. Deep learning for cardiovascular medicine: a practical primer. European heart journal. 2019;40(25):2058-73.

296. Lankveld T, Zeemering S, Scherr D, Kuklik P, Hoffmann BA, Willems S, et al. Atrial fibrillation complexity parameters derived from surface ECGs predict procedural outcome and long-term follow-up of stepwise catheter ablation for atrial fibrillation. Circulation: Arrhythmia and Electrophysiology. 2016;9(2):e003354.

297. Hao P, Gao X, Li Z, Zhang J, Wu F, Bai C. Multi-branch fusion network for myocardial infarction screening from 12-lead ECG images. Computer Methods and Programs in Biomedicine. 2020;184:105286.

298. Bello GA, Dawes TJ, Duan J, Biffi C, de Marvao A, Howard LS, et al. Deep-learning cardiac motion analysis for human survival prediction. Nature machine intelligence. 2019;1(2):95-104.

299. Attia ZI, Noseworthy PA, Lopez-Jimenez F, Asirvatham SJ, Deshmukh AJ, Gersh BJ, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. The Lancet. 2019;394(10201):861-7.

300. Huang Y, Liu Z, He L, Chen X, Pan D, Ma Z, et al. Radiomics signature: a potential biomarker for the prediction of disease-free survival in early-stage (I or II) non—small cell lung cancer. Radiology. 2016;281(3):947-57.

301. Al Rahhal MM, Bazi Y, Al Zuair M, Othman E, BenJdira B. Convolutional neural networks for electrocardiogram classification. Journal of Medical and Biological Engineering. 2018;38(6):1014-25.

302. Agrafioti F, Hatzinakos D. ECG biometric analysis in cardiac irregularity conditions. Signal, Image and Video Processing. 2009;3(4):329.

303. Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nature medicine. 2019;25(1):65.

304. Stephane M. A wavelet tour of signal processing. Elsevier; 1999.

305. Yildirim O, San Tan R, Acharya UR. An efficient compression of ECG signals using deep convolutional autoencoders. Cognitive Systems Research. 2018;52:198-211.

306. Chen S, Meng Z, Zhao Q. Electrocardiogram Recognization Based on Variational AutoEncoder. Machine Learning and Biometrics. 2018:71.

307. Golany T, Radinsky K, editors. PGANs: Personalized generative adversarial networks for ECG synthesis to improve patient-specific deep ECG classification. Proceedings of the AAAI Conference on Artificial Intelligence; 2019.

308. Zhu F, Ye F, Fu Y, Liu Q, Shen B. Electrocardiogram generation with a bidirectional LSTM-CNN generative adversarial network. Scientific reports. 2019;9(1):1-11.

309. Smith LN. A disciplined approach to neural network hyper-parameters: Part 1--learning rate, batch size, momentum, and weight decay. arXiv preprint arXiv:180309820. 2018.

310. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:150203167. 2015.

311. Dumoulin V, Visin F. A guide to convolution arithmetic for deep learning. arXiv preprint arXiv:160307285. 2016.

312. He K, Zhang X, Ren S, Sun J, editors. Identity mappings in deep residual networks. European conference on computer vision; 2016: Springer.

313. Yoo AB, Jette MA, Grondona M, editors. Slurm: Simple linux utility for resource management. Workshop on Job Scheduling Strategies for Parallel Processing; 2003: Springer.

314. Sahle BW, Owen AJ, Chin KL, Reid CM. Risk prediction models for incident heart failure: a systematic review of methodology and model performance. Journal of cardiac failure. 2017;23(9):680-7.

315. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al., editors. Tensorflow: A system for large-scale machine learning. 12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16); 2016.

316. Pencina MJ, D'Agostino Sr RB, D'Agostino Jr RB, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Statistics in medicine. 2008;27(2):157-72.

317. Kundu S, Aulchenko YS, van Duijn CM, Janssens ACJ. PredictABEL: an R package for the assessment of risk prediction models. European journal of epidemiology. 2011;26(4):261-4.

318. Crowe F, Zemedikun DT, Okoth K, Adderley NJ, Rudge G, Sheldon M, et al. Comorbidity phenotypes and risk of mortality in patients with ischaemic heart disease in the UK. Heart. 2020;106(11):810-6.

319. Tromp J, Ouwerkerk W, Demissei BG, Anker SD, Cleland JG, Dickstein K, et al. Novel endotypes in heart failure: effects on guideline-directed medical therapy. European heart journal. 2018;39(48):4269-76.

320. Ng SK, Tawiah R, Sawyer M, Scuffham P. Patterns of multimorbid health conditions: a systematic review of analytical methods and comparison analysis. International journal of epidemiology. 2018;47(5):1687-704.

321. Flynt A, Dean N. A survey of popular R packages for cluster analysis. Journal of Educational and Behavioral Statistics. 2016;41(2):205-25.

322. Morissette L, Chartier S. The k-means clustering technique: General considerations and implementation in Mathematica. Tutorials in Quantitative Methods for Psychology. 2013;9(1):15-24.

323. Lloyd S. Least squares quantization in PCM. IEEE transactions on information theory. 1982;28(2):129-37.

324. Park H-S, Jun C-H. A simple and fast algorithm for K-medoids clustering. Expert systems with applications. 2009;36(2):3336-41.

325. Charrad M, Ghazzali N, Boiteau V, Niknafs A, Charrad MM. Package 'nbclust'. Journal of statistical software. 2014;61:1-36.

326. Thorndike RL. Who belongs in the family? Psychometrika. 1953;18(4):267-76.

327. Scrucca L, Fop M, Murphy TB, Raftery AE. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. The R journal. 2016;8(1):289.

328. Linzer DA, Lewis JB. poLCA: An R package for polytomous variable latent class analysis. Journal of statistical software. 2011;42(10):1-29.

329. Haughton D, Legrand P, Woolford S. Review of three latent class cluster analysis packages: Latent Gold, poLCA, and MCLUST. The American Statistician. 2009;63(1):81-91.

330. McParland D, Gormley IC. Model based clustering for mixed data: clustMD. Advances in Data Analysis and Classification. 2016;10(2):155-69.

331. McLachlan GJ, Rathnayake S. On the number of components in a Gaussian mixture model. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2014;4(5):341-55.

332. Von Luxburg U. Clustering stability: an overview: Now Publishers Inc; 2010.

333. Jain AK, Moreau J. Bootstrap technique in cluster analysis. Pattern Recognition. 1987;20(5):547-68.

334. Dolnicar S, Leisch F. Evaluation of structure and reproducibility of cluster solutions using the bootstrap. Marketing Letters. 2010;21(1):83-101.

335. Caliński T, Harabasz J. A dendrite method for cluster analysis. Communications in Statistics-theory and Methods. 1974;3(1):1-27.

336. Rand WM. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical association. 1971;66(336):846-50.

337. Kalgotra P, Sharda R, Croff JM. Examining multimorbidity differences across racial groups: a network analysis of electronic medical records. Scientific reports. 2020;10(1):1-9.

338. Siggaard T, Reguant R, Jørgensen IF, Haue AD, Lademann M, Aguayo-Orozco A, et al. Disease trajectory browser for exploring temporal, population-wide disease progression patterns in 7.2 million Danish patients. Nature communications. 2020;11(1):1-10.

339. Zemedikun DT, Gray LJ, Khunti K, Davies MJ, Dhalwani NN, editors. Patterns of multimorbidity in middle-aged and older adults: an analysis of the UK Biobank data. Mayo Clinic Proceedings; 2018: Elsevier.

340. Vetrano DL, Roso-Llorach A, Fernández S, Guisado-Clavero M, Violán C, Onder G, et al. Twelve-year clinical trajectories of multimorbidity in a population of older adults. Nature communications. 2020;11(1):1-9.

341. Del Giacco SR, Cappai A, Gambula L, Cabras S, Perra S, Manconi PE, et al. The asthma-anxiety connection. Respiratory medicine. 2016;120:44-53.

342. Shen T-C, Lin C-L, Wei C-C, Tu C-Y, Li Y-F. The risk of asthma in rheumatoid arthritis: a population-based cohort study. QJM: An International Journal of Medicine. 2014;107(6):435-42.

343. Huang H-L, Ho S-Y, Li C-H, Chu F-Y, Ciou L-P, Lee H-C, et al. Bronchial asthma is associated with increased risk of chronic kidney disease. BMC pulmonary medicine. 2014;14(1):1-8.

344. Bloom CI, Saglani S, Feary J, Jarvis D, Quint JK. Changing prevalence of current asthma and inhaled corticosteroid treatment in the UK: population-based cohort 2006–2016. European Respiratory Journal. 2019;53(4).

345. Pauwels RA, Buist AS, Calverley PM, Jenkins CR, Hurd SS. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: NHLBI/WHO Global Initiative for Chronic Obstructive Lung Disease (GOLD) Workshop summary. American journal of respiratory and critical care medicine. 2001;163(5):1256-76.

346. Matarese A, Sardu C, Shu J, Santulli G. Why is chronic obstructive pulmonary disease linked to atrial fibrillation? A systematic overview of the underlying mechanisms. International journal of cardiology. 2019;276:149.

347. Grymonprez M, Vakaet V, Kavousi M, Stricker BH, Ikram MA, Heeringa J, et al. Chronic obstructive pulmonary disease and the development of atrial fibrillation. International journal of cardiology. 2019;276:118-24.

348. Han MK, McLaughlin VV, Criner GJ, Martinez FJ. Pulmonary diseases and the heart. Circulation. 2007;116(25):2992-3005.

349. Rabe KF, Hurst JR, Suissa S. Cardiovascular disease and COPD: dangerous liaisons? European Respiratory Review. 2018;27(149).

350. Levey AS, Bosch JP, Lewis JB, Greene T, Rogers N, Roth D. A more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation. Annals of internal medicine. 1999;130(6):461-70.

351. Trust OUHNF. NHS Biochemistry Reference Ranges Document 2019 [Available from: https://www.ouh.nhs.uk/biochemistry/tests/documents/biochemistry-reference-ranges.pdf.

352. Jafari M, Ansari-Pour N. Why, when and how to adjust your P values? Cell Journal (Yakhteh). 2019;20(4):604.

353. Reiner A, Yekutieli D, Benjamini Y. Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics. 2003;19(3):368-75.

354. Hoehndorf R, Dumontier M, Gkoutos GV. Evaluation of research in biomedical ontologies. Brief Bioinform. 2012;14(6):696-712.

355. Group PGD. PostgreSQL. 11 ed2019.

356. Microsoft. SQL Server. California2019.

357. Hoehndorf R, Schofield PN, Gkoutos GV. The role of ontologies in biological and biomedical research: a functional perspective. Brief Bioinform. 2015;16(6):1069-80.

358. Köhler S, Gargano M, Matentzoglu N, Carmody LC, Lewis-Smith D, Vasilevsky NA, et al. The human phenotype ontology in 2021. Nucleic acids research. 2021;49(D1):D1207-D17.

359. Lipscomb CE. Medical subject headings (MeSH). Bulletin of the Medical Library Association. 2000;88(3):265.

360. Kubacka T. A story told through a heatmap 2019 [Available from: https://www.kaggle.com/tkubacka/a-story-told-through-a-heatmap.

361. Wickham H, Averick M, Bryan J, Chang W, McGowan LDA, François R, et al. Welcome to the Tidyverse. Journal of Open Source Software. 2019;4(43):1686.

362. McKinney W. Python for data analysis: Data wrangling with Pandas, NumPy, and IPython: " O'Reilly Media, Inc."; 2012.

363. Dorgueil R. Bonobo. 2019.

364. Urbanek S. Bubbles. 2015.

365. Knoblock CA, Szekely P, Ambite JL, Gupta S, Goel A, Muslea M, et al., editors. Interactively mapping data sources into the semantic web. Proceedings of the First International Conference on Linked Science-Volume 783; 2011: CEUR-WS. org.

366. Brickley D, Guha RV, McBride B. RDF Schema 1.1. W3C recommendation. 2014;25:2004-14.

367. Ceusters W, Smith B, Fielding JM, editors. LinkSuite TM: Formally robust ontology-based data and information integration. International Workshop on Data Integration in the Life Sciences; 2004: Springer.

368. Toh TS, Dondelinger F, Wang D. Looking beyond the hype: Applied AI and machine learning in translational medicine. EBioMedicine. 2019;47:607-15.

369. Longhurst CA, Harrington RA, Shah NH. A 'green button'for using aggregate patient data at the point of care. Health affairs. 2014;33(7):1229-35.

370. Hemingway H, Asselbergs FW, Danesh J, Dobson R, Maniadakis N, Maggioni A, et al. Big data from electronic health records for early and late translational cardiovascular research: challenges and potential. European heart journal. 2018;39(16):1481-95.

371. Kennedy EH, Wiitala WL, Hayward RA, Sussman JB. Improved cardiovascular risk prediction using nonparametric regression and electronic health record data. Medical care. 2013;51(3):251.

372. Bacon S, Goldacre B. Barriers to Working With National Health Service England's Open Data. Journal of Medical Internet Research. 2020;22(1):e15603.

373. Chakraborty P, Farooq F, editors. A Robust Framework for Accelerated Outcome-driven Risk Factor Identification from EHR. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; 2019.

374. Khennou F, Khamlichi YI, Chaoui NEH. Improving the use of big data analytics within electronic health records: a case study based OpenEHR. Procedia Computer Science. 2018;127:60-8.

375. Cormen TH, Leiserson CE, Rivest RL, Stein C. Introduction to algorithms: MIT press; 2009.

376. OHDSI OHDSaI-. OMOP Common Data Model 2021 [Available from: https://www.ohdsi.org/data-standardization/the-common-data-model/.

377. Zhang H, Guo Y, Li Q, George TJ, Shenkman E, Modave F, et al. An ontology-guided semantic data integration framework to support integrative data analysis of cancer survival. BMC medical informatics and decision making. 2018;18(2):129-47.

378. De Giacomo G, Lembo D, Lenzerini M, Poggi A, Rosati R. Using ontologies for semantic data integration. A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years: Springer; 2018. p. 187-202.

379. Mehta N, Pandit A. Concurrence of big data analytics and healthcare: A systematic review. International journal of medical informatics. 2018;114:57-65.

380. Wang Y, Kung L, Wang WYC, Cegielski CG. An integrated big data analytics-enabled transformation model: Application to health care. Information & Management. 2018;55(1):64-79.

381. Mehta N, Pandit A, Kulkarni M. Elements of healthcare Big Data analytics. Big Data Analytics in Healthcare: Springer; 2020. p. 23-43.

382. Merkel D. Docker: lightweight linux containers for consistent development and deployment. Linux journal. 2014;2014(239):2.

383. Lenzerini M, editor Data integration: A theoretical perspective. Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems; 2002.

384. Kimball R, Caserta J. The data warehouse ETL toolkit: John Wiley & Sons; 2004.

385. ISO I. Quantities and Units—Part 1: General. 2009.

386. Pecoraro F, Luzi D, Ricci FL, editors. A Clinical Data Warehouse Architecture based on the Electronic Healthcare Record Infrastructure. HEALTHINF; 2014.

387. Song T-M, Ryu S. Big data analysis framework for healthcare and social sectors in Korea. Healthcare informatics research. 2015;21(1):3-9.

388. Organization WH. Novel coronavirus 2020 [Available from: https://www.who.int/csr/don/12-january-2020-novel-coronavirus-china/en/.

389. Liu Y, Gayle AA, Wilder-Smith A, Rocklöv J. The reproductive number of COVID-19 is higher compared to SARS coronavirus. Journal of travel medicine. 2020.

390. Cascella M, Rajnik M, Cuomo A, Dulebohn SC, Di Napoli R. Features, evaluation and treatment coronavirus (COVID-19). Statpearls [internet]: StatPearls Publishing; 2020.

391. Hessami A, Shamshirian A, Heydari K, Pourali F, Alizadeh-Navaei R, Moosazadeh M, et al. Cardiovascular diseases burden in COVID-19: Systematic review and meta-analysis. The American journal of emergency medicine. 2021;46:382-91.

392. Zheng Y-Y, Ma Y-T, Zhang J-Y, Xie X. COVID-19 and the cardiovascular system. Nature Reviews Cardiology. 2020;17(5):259-60.

393. Siripanthong B, Nazarian S, Muser D, Deo R, Santangeli P, Khanji MY, et al. Recognizing COVID-19–related myocarditis: The possible pathophysiology and proposed guideline for diagnosis and management. Heart rhythm. 2020;17(9):1463-71.

394. Spinoni EG, Mennuni M, Rognoni A, Grisafi L, Colombo C, Lio V, et al. Contribution of atrial fibrillation to in-hospital mortality in patients with COVID-19. Circulation: Arrhythmia and Electrophysiology. 2021;14(2):e009375.

395. Lee CC, Ali K, Connell D, Mordi IR, George J, Lang EM, et al. COVID-19-associated cardiovascular complications. Diseases. 2021;9(3):47.

396. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. bmj. 2020;369.

397. Ji D, Zhang D, Xu J, Chen Z, Yang T, Zhao P, et al. Prediction for progression risk in patients with COVID-19 pneumonia: the CALL score. Clinical Infectious Diseases. 2020;71(6):1393-9.

398. Shi Y, Yu X, Zhao H, Wang H, Zhao R, Sheng J. Host susceptibility to severe COVID-19 and establishment of a host risk score: findings of 487 cases outside Wuhan. Critical care. 2020;24(1):1-4.

399. Gong J, Ou J, Qiu X, Jie Y, Chen Y, Yuan L, et al. A tool for early prediction of severe coronavirus disease 2019 (COVID-19): a multicenter study using the risk nomogram in Wuhan and Guangdong, China. Clinical infectious diseases. 2020;71(15):833-40.

400. Lu J, Hu S, Fan R, Liu Z, Yin X, Wang Q, et al. ACP risk grade: a simple mortality index for patients with confirmed or suspected severe acute respiratory syndrome coronavirus 2 disease (COVID-19) during the early stage of outbreak in Wuhan, China. 2020.

401. Levy TJ, Richardson S, Coppa K, Barnaby DP, McGinn T, Becker LB, et al. Development and validation of a survival calculator for hospitalized patients with COVID-19. medRxiv. 2020.

402. Yan L, Zhang H-T, Goncalves J, Xiao Y, Wang M, Guo Y, et al. An interpretable mortality prediction model for COVID-19 patients. Nature machine intelligence. 2020;2(5):283-8.

403. Breiman L. Bagging predictors. Machine learning. 1996;24(2):123-40.

404. pandemic GgfscdtC. COVIDSurg Collaborative. Br J Surg. 2020.

405. Birmingham Uo. COVID-19 disruption will lead to 28 million surgeries cancelled worldwide Birmingham2020 [Available from: https://www.birmingham.ac.uk/news/latest/2020/05/covid-disruption-28-million-surgeries-cancelled.aspx.

406. Nepogodiev D, Bhangu A, Glasbey JC, Li E, Omar OM, Simoes JF, et al. Mortality and pulmonary complications in patients undergoing surgery with perioperative SARS-CoV-2 infection: an international cohort study. The Lancet. 2020;396(10243):27-38.

407. CovidSurg. CovidSurg Risk Model Birmingham2021 [Available from: https://covidsurgrisk.app/.

408. So AD, Woo J. Reserving coronavirus disease 2019 vaccines for global access: cross sectional analysis. bmj. 2020;371.

409. Foundation TH. COVID-19: Five dimensions of impact 2020 [Available from: https://www.health.org.uk/news-and-comment/blogs/covid-19-five-dimensions-of-impact.

410. Association BD. Live updates: Coronavirus and dentistry London2020 [Available from: https://bda.org/advice/Coronavirus/Pages/latest-updates.aspx.

411. OpenText. OpenText EMC Documentum. 2021.

412. Chodorow K. MongoDB: the definitive guide: powerful and scalable data storage: " O'Reilly Media, Inc."; 2013.

413. Slater LT, Bradlow W, Hoehndorf R, Motti DF, Ball S, Gkoutos G. Komenti: A semantic text mining framework. 2020.

414. Chesneau B. Gunicorn-python wsgi http server for unix. 2017.

415. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological). 1996;58(1):267-88.

416. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:190711692. 2019.

APPENDICES

Appendix 2.1 Biomarkers collected for BBCAF study

The list of biomarkers is compiled from Chua et al. (77).

Biomarker	Abbreviation
Adrenomedullin	ADM
Agouti-related protein	AGRP
Angiopoietin-1 receptor	TIE2
Cathepsin L1	CTSL1
C-C motif chemokine 3	CCL3
CD40 ligand	CD40L
C-X-C motif chemokine 1	CXCL1
Dickkopf-related protein 1	Dkk-1
Fibroblast growth factor 23	FGF-23
Follistatin	FS
Growth hormone	GH
Heat shock 27 kDa protein	HSP 27
Heparin-binding EGF-like growth factor	HB-EGF
Interleukin-1 receptor antagonist protein	IL-1ra
Interleukin-16	IL-16

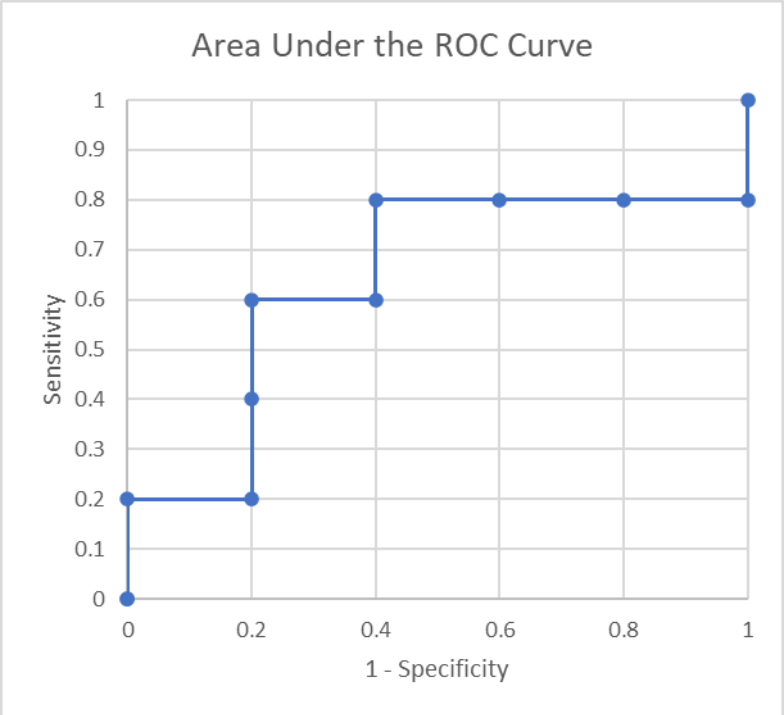
Biomarker	Abbreviation
Interleukin-18	IL-18
Interleukin-27	IL-27
Interleukin-6	IL-6
Lectin-like oxidized LDL receptor	LOX-1
Leptin	LEP
Matrix metalloproteinase-12	MMP-12
Matrix metalloproteinase-7	MMP-7
Melusin	ITGB1BP2
Natriuretic peptides B	BNP
NF-kappa-B essential modulator	NEMO
Pappalysin-1	PAPPA
Pentraxin-related protein PTX3	PTX3
Placenta growth factor	PIGF
Platelet-derived growth factor subunit B	PDGF subunit B
Proteinase-activated receptor 1	PAR-1
Proto-oncogene tyrosine-protein kinase Src	SRC
P-selectin glycoprotein ligand 1	PSGL-1
Receptor for advanced glycosylation end products	RAGE

Biomarker	Abbreviation
Renin	REN
Stem cell factor	SCF
Thrombomodulin	TM
TIM.1	TIM-1
Tissue factor	TF
TNF-related apoptosis-induced ligand receptor 2	TRAIL-R2
Vascular endothelial growth factor D	VEGF-D

Appendix 2.2 Calculating Area Under the Receiver Operating Characteristic Curve

Given a list of samples, the AUCROC can be calculated using the predicted scores. True class indicates the real value of the sample, the prediction indicates the sample risk on the model.

True class	Prediction/threshold	TP	TN	FP	FN	TNR	1-TNR	TPR
	0	5	0	5	0	0	1	1
1	0.1	4	0	5	1	0	1	0.8
0	0.15	4	1	4	1	0.2	0.8	0.8
0	0.18	4	2	3	1	0.4	0.6	0.8
0	0.2	4	3	2	1	0.6	0.4	0.8
1	0.45	3	3	2	2	0.6	0.4	0.6
0	0.46	3	4	1	2	0.8	0.2	0.6
1	0.7	2	4	1	3	0.8	0.2	0.4
1	0.8	1	4	1	4	0.8	0.2	0.2
0	0.81	1	5	0	4	1	0	0.2
1	0.97	0	5	0	5	1	0	0
	1	0	5	0	5	1	0	0



AUCROC plot sample. The blue dots indicate the different thresholds of the model.

$$\sum_{\forall t \; in \; thresholds} TPR_t * (TNR_{t-1} - TNR_t)$$

This equation calculates the AUCROC in steps.

Appendix 3.1 Neural network hyperparameter optimization

There are two main aspects to a model, its parameters, these are configuration variables inherent to the data and model. Hyperparameters are tuneable elements external to the data and model, these are defined by the data practitioner to configure the model creation and evolution. Hyperparameters are elements such as the number of trees and depth of a random forest model (138), the regularization term in the least absolute shrinkage and selection operator (LASSO) model (415).

Due to the influence of these parameters, the same network using different settings, or hyperparameters, can achieve higher performance than a model without hyperparameter optimization. This applies to any application of machine learning, even, for example, RoBERTa neural language model (416).

One can modify the architecture and the number of elements in a neural network without limits. Some patterns work better in some situations, those identified by exploratory works. For example, images operate better with convolutional neural networks of decreasing kernel sizes, and incremented feature maps, rather than having dense layers only. For variables that contain intensity data, and not mapped or connected, such as biomarkers, dense layers suffice in a neural network. There are matters of the number of dense nodes in each layer, regularization which can be done using dropout layer/s, optimizer choice, activation functions and early stopping rules.

To evaluate these different combinations, another dataset is required. The discovery dataset was further split into training and internal validation sets (80%, 20%). The best model from this selection is then returned to have a final comparison against the original validation data.

Early stopping was fixed as 20 epochs (**Figure A2**). Activation function was limited to RELU on hidden layers, and sigmoid in the output node. The optimizer was restricted to *adam*. A different number of layers were evaluated, from 1 to 3 hidden layers, dense layers were tested with 256, 512 and 1024 nodes, dropout ratio tested were 0.05, 0.1 and 0.2. All models were tested 3 times, each with different initialization seeds. In total 81 models were created. **Figure A1** illustrates the training process.

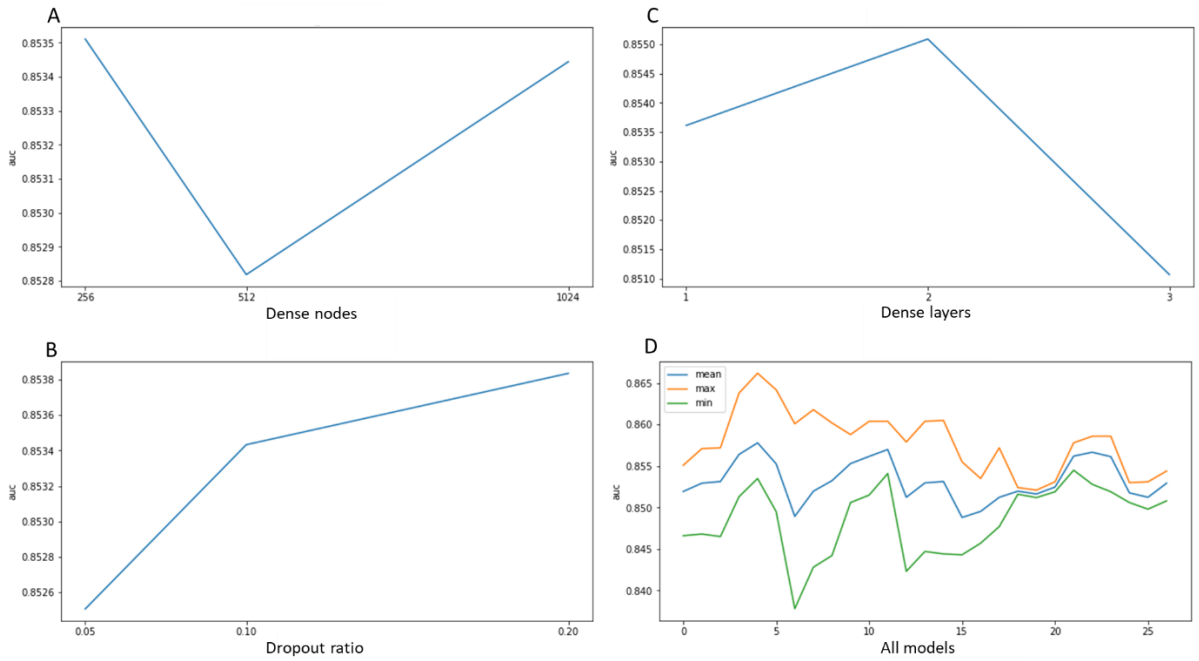


Figure A1: Different hyperparameters evaluated and their performances on the internal validation set. Different graphs indicate the differences in AUC for A dense nodes, B number of dense layers, C dropout ratios, D all the different settings.

There is not much difference between the different settings. The most difference comes from varying the initialization seed of the network. However, as following the criteria of maximizing the AUC value, the network picked contains 2 dense layers with 256 dense nodes each and a dropout ratio of 0.2.

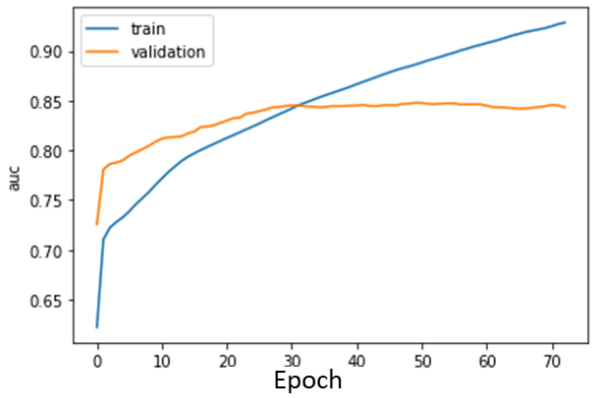


Figure A2: Early stopping criteria. Whilst the performance in the training set is ever-increasing, tending to over-fit, the internal validation performance plateau after some training epochs. A model stopped at this stage prevents worsened generalization performance.

Appendix 4.1 List of significantly expressed genes Human LAxRA

Ensemble ID	GeneSymbol	log2FoldChange	p-val adj
ENSG00000164093	PITX2	5.470003641	2.02E-173
ENSG00000250103	PANCR	6.042799819	1.60E-169
ENSG00000230943	LINC02541	5.521013851	2.15E-149
ENSG00000135439	AGAP2	2.604749826	1.81E-132
ENSG00000143867	OSR1	2.066359897	4.07E-126
ENSG00000250511	AC098798.1	5.113739231	5.52E-121
ENSG00000145242	EPHA5	2.429359655	3.27E-90
ENSG00000112214	FHL5	2.45696871	9.69E-82
ENSG00000105697	HAMP	-4.402966178	4.21E-81
ENSG00000115705	TPO	2.708508129	3.84E-76
ENSG00000250846	EPHA5-AS1	3.098416201	9.30E-75
ENSG00000109991	P2RX3	3.961595981	3.00E-72
ENSG00000239402	CYP4F62P	-4.703220574	1.17E-68
ENSG00000081277	PKP1	3.538850955	1.03E-63
ENSG00000163217	BMP10	-5.750039089	1.34E-63
ENSG00000066405	CLDN18	3.716571513	1.63E-60
ENSG00000162981	FAM84A	-2.132603108	7.37E-59
ENSG00000225670	CADM3-AS1	3.282068525	1.52E-58
ENSG00000142973	CYP4B1	2.611042779	2.09E-57
ENSG00000153976	HS3ST3A1	-2.339592424	4.93E-56
ENSG00000240253	FAR2P3	-3.365040066	3.07E-54
ENSG00000188803	SHISA6	2.484058672	2.82E-51
ENSG00000163395	IGFN1	-2.540693211	1.08E-50
ENSG00000181408	UTS2R	-2.209696091	3.68E-46
ENSG00000169507	SLC38A11	2.132490012	3.37E-45
ENSG00000145423	SFRP2	2.914448347	6.04E-45
ENSG00000249306	LINC01411	-3.564365425	1.44E-44
ENSG00000148942	SLC5A12	-2.569254486	3.24E-44
ENSG00000236081	ELFN1-AS1	2.904542162	4.95E-41
ENSG00000214081	CYP4F30P	-4.006277315	7.42E-41
ENSG00000180053	NKX2-6	-3.954381375	1.10E-40
ENSG00000268297	CLEC4GP1	2.613016075	7.95E-40
ENSG00000183134	PTGDR2	2.398143223	2.69E-39
ENSG00000163017	ACTG2	2.256640038	2.25E-38
ENSG00000228826	AL592494.1	-2.97101456	2.32E-38
ENSG00000196834	POTEI	-2.938744439	7.50E-38
ENSG00000222038	POTEJ	-3.472946834	6.22E-36
ENSG00000196990	FAM163B	-2.588881078	1.00E-35
ENSG00000137077	CCL21	2.350376586	6.20E-35
ENSG00000213088	ACKR1	2.280205941	1.49E-34
ENSG00000132975	GPR12	2.794484518	1.88E-31
ENSG00000185559	DLK1	4.099831016	2.83E-30
ENSG00000248479	AC104137.1	2.436802393	3.44E-29
ENSG00000165970	SLC6A5	3.068726776	4.60E-29
ENSG00000283646	LINC02009	2.064619066	2.92E-28
ENSG00000283221	AC007218.2	2.534789321	7.64E-28
ENSG00000079435	LIPE	2.114638371	1.87E-27
ENSG00000177359	AC024940.2	-2.349692025	2.81E-27
ENSG00000223349	KLF2P3	-3.345967264	4.14E-27
ENSG00000198914	POU3F3	-2.538422478	4.71E-27
ENSG00000124882	EREG	2.217026464	1.61E-26
ENSG00000168447	SCNN1B	2.660074171	1.85E-26
ENSG00000264727	AC005725.1	2.177379808	2.31E-26
ENSG00000272583	AL592494.3	-2.470152537	4.69E-26
ENSG00000100678	SLC8A3	2.038773703	5.51E-25
ENSG00000172061	LRRC15	2.595176211	1.54E-24
ENSG00000130876	SLC7A10	2.903589561	4.98E-23
ENSG00000106178	CCL24	2.815576067	7.59E-23
ENSG00000126733	DACH2	-2.319885416	3.16E-22
ENSG00000224940	PRRT4	2.016005835	3.26E-22
ENSG00000111245	MYL2	2.494377094	2.32E-21
ENSG00000173432	SAA1	2.902763767	3.80E-21
ENSG00000283586	GKN3P	-2.614855865	7.90E-21
ENSG00000105509	HAS1	2.619598633	9.30E-21
ENSG00000128040	SPINK2	-2.07267951	1.84E-20
ENSG00000261319	LINC02152	-2.441425195	2.10E-20
ENSG00000167588	GPD1	2.086906235	3.13E-20
ENSG00000166819	PLIN1	2.38596848	2.24E-19
ENSG00000134962	KLB	2.100708678	4.25E-19
ENSG00000158571	PFKFB1	2.169068151	6.21E-19
ENSG00000146352	CLVS2	3.313764021	7.87E-18
ENSG00000181092	ADIPOQ	2.492721917	8.30E-18
ENSG00000138207	RBP4	2.046499443	2.23E-17
ENSG00000174697	LEP	2.431766243	2.92E-17
ENSG00000253434	LINC02237	2.208700302	3.03E-17
ENSG00000135917	SLC19A3	2.07464291	5.76E-17
ENSG00000149124	GLYAT	2.477564954	1.67E-16
ENSG00000187288	CIDEC	2.311491726	2.95E-16
ENSG00000168333	PPDPFL	-2.482179431	6.28E-16
ENSG00000152785	BMP3	2.492125739	6.63E-16
ENSG00000275385	CCL18	2.023327327	8.23E-16
ENSG00000197632	SERPINB2	3.474779459	1.53E-15
ENSG00000226482	ADIPOQ-AS1	2.802352656	1.75E-15

ENSG00000100739	BDKRB1	2.669288033	1.83E-15
ENSG00000145863	GABRA6	2.726174706	2.70E-15
ENSG00000159387	IRX6	2.228848255	2.90E-15
ENSG00000233639	LINC01158	-2.183526207	6.78E-15
ENSG00000133317	LGALS12	2.057873719	2.01E-14
ENSG00000248869	LINC02511	2.091527476	2.22E-14
ENSG00000184811	TUSC5	2.142827201	3.31E-14
ENSG00000220553	RPL5P19	-2.298695423	3.38E-14
ENSG00000214064	RPL6P5	-2.095533267	4.10E-14
ENSG00000211890	IGHA2	2.133082319	4.43E-14
ENSG00000165478	HEPACAM	2.020484053	5.97E-14
ENSG00000143839	REN	-3.351895738	9.78E-14
ENSG00000118231	CRYGD	-3.277031878	2.43E-13
ENSG00000268416	AC010329.1	-2.319525484	3.98E-13
ENSG00000166104	AC126323.1	2.292393034	4.09E-13
ENSG00000162761	LMX1A	2.512139594	2.54E-12
ENSG00000253369	AC131902.1	-2.069175413	1.21E-11
ENSG00000147647	DPYS	2.339998309	1.76E-11
ENSG00000114771	AADAC	2.153290204	3.80E-11
ENSG00000271239	AC007423.1	2.394412285	5.93E-11

ENSG00000186081	KRT5	2.229028347	6.63E-11
ENSG00000157765	SLC34A2	2.864462833	2.58E-10
ENSG00000128652	HOXD3	2.026460776	4.03E-10
ENSG00000257671	KRT7-AS	2.47517599	1.40E-09
ENSG00000134339	SAA2	2.424098676	3.19E-09
ENSG00000174145	NWD2	2.239504222	5.67E-09
ENSG00000230778	ANKRD63	-2.061815626	6.16E-09
ENSG00000182585	EPGN	2.093758438	7.56E-09
ENSG00000274295	AC131902.3	-2.096248797	7.76E-09
ENSG00000170484	KRT74	-2.269750569	1.66E-08
ENSG00000261618	AC083837.1	2.310090896	1.78E-08
ENSG00000211976	IGHV3-73	2.057242283	5.64E-08
ENSG00000166396	SERPINB7	2.254171541	1.39E-07
ENSG00000205420	KRT6A	2.382528691	1.11E-06
ENSG00000135477	KRT87P	2.132578852	1.27E-06
ENSG00000203783	PRR9	2.090805801	2.49E-06
ENSG00000243264	IGKV2D-29	2.176407322	4.52E-06
ENSG00000258793	AL355102.4	2.087923954	8.56E-06
ENSG00000162891	IL20	2.014121401	1.10E-05

Appendix 6.1 List of 65 important comorbidities

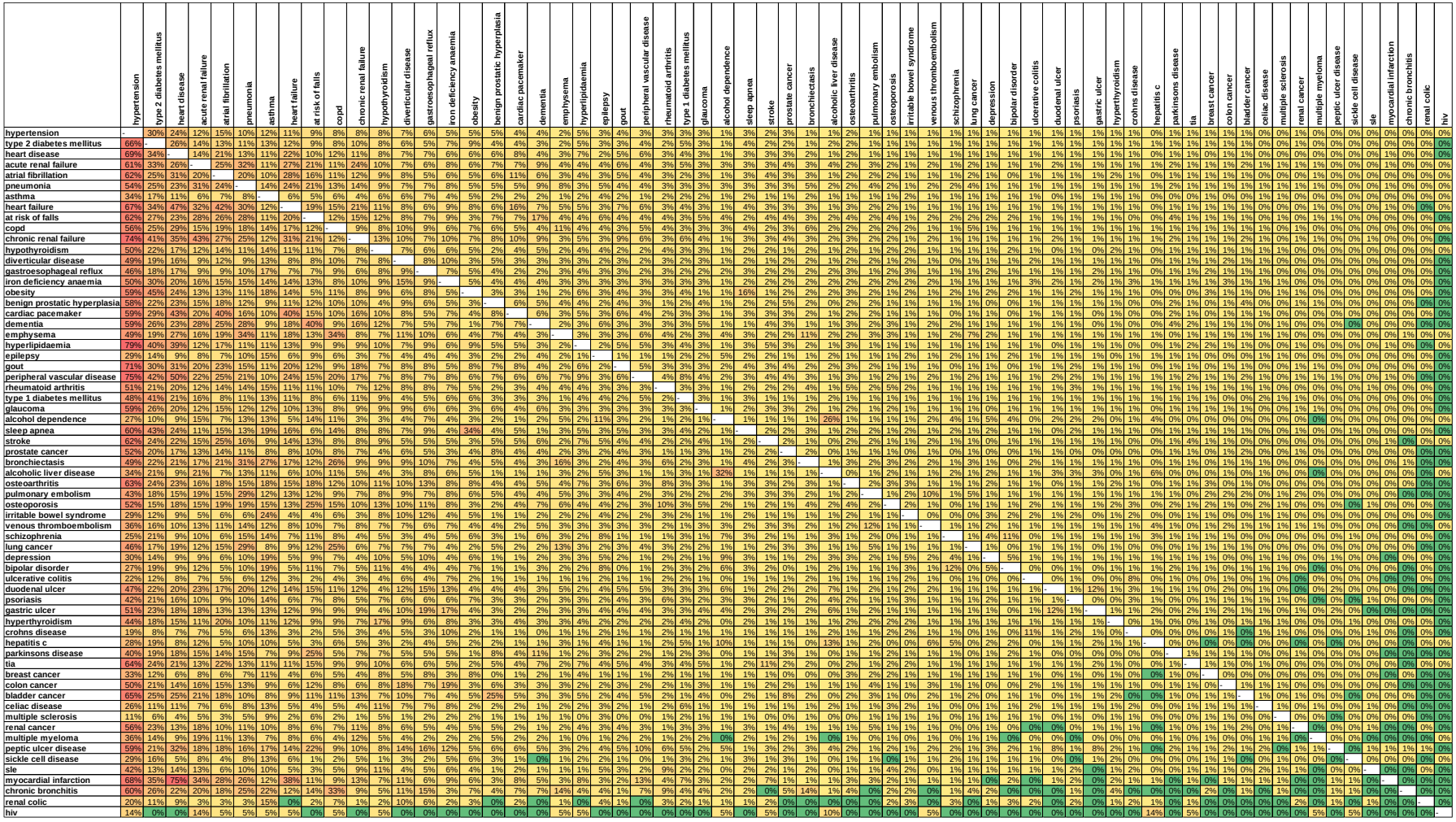
Condition	ICD-10
Acute renal failure	N17.9
Alcohol dependence	F10.2
Alcoholic liver disease	K70 K70.0 K70.1 K70.2 K70.3 K70.4 K70.9
Asthma	J45.9
At risk of falls	R29.6
Atrial fibrillation	I48 I48.0 I48.1 I48.2 I48.3 I48.4 I48.9
Benign prostatic hyperplasia	N40 N40X
Bipolar disorder	F31 F31.0 F31.1 F31.2 F31.3 F31.4 F31.5 F31.6 F31.7 F31.8 F31.9
Bladder cancer	C67 C67.0 C67.1 C67.2 C67.3 C67.4 C67.5 C67.6 C67.7 C67.8 C67.9
Breast cancer	C50 C50.0 C50.1 C50.2 C50.3 C50.4 C50.5 C50.6 C50.8 C50.9
Bronchiectasis	J47 J47X
Cardiac pacemaker	Z95.0
Celiac disease	K90.0
Chronic bronchitis	J42 J42X
Chronic renal failure	N18.9
Colon cancer	C18 C18.0 C18.1 C18.2 C18.3 C18.4 C18.5 C18.6 C18.7 C18.8 C18.9
Chronic obstructive pulmonary disease	J44.9
Crohn's disease	K50 K50.0 K50.1 K50.8 K50.9
Dementia	F03 F03X
Depression	F33.9

Diverticular disease	K57 K57.0 K57.1 K57.2 K57.3 K57.4 K57.5 K57.8 K57.9
Duodenal ulcer	K26 K26.0 K26.1 K26.2 K26.3 K26.4 K26.5 K26.6 K26.7 K26.9
Emphysema	J43.9
Epilepsy	G40 G40.0 G40.1 G40.2 G40.3 G40.4 G40.5 G40.6 G40.7 G40.8 G40.9
Gastric ulcer	K25 K25.0 K25.1 K25.2 K25.3 K25.4 K25.5 K25.6 K25.7 K259
Gastroesophageal reflux	K21 K21.0 K21.9
Glaucoma	H40 H40.0 H40.1 H40.2 H40.3 H40.4 H40.5 H40.6 H40.8 H40.9
Gout	M10.9 M109.0 M109.1 M109.2 M109.3 M109.4 M109.5 M109.6 M109.7 M109.8 M109.9
Heart disease	I25 I25.0 I25.1 I25.2 I25.3 I25.4 I25.5 I25.6 I25.8 I25.9
Heart failure	I50 I50.0 I50.1 I50.9
Hepatitis c	B18.2
Human immunodeficiency virus	B24 B24X
Hyperlipidaemia	E78.5
Hypertension	I10 I10X
Hyperthyroidism	E05 E05.0 E05.1 E05.2 E05.3 E05.4 E05.5 E05.8 E05.9
Hypothyroidism	E03 E03.0 E03.1 E03.2 E03.3 E03.4 E03.5 E03.8 E03.9
Iron-deficiency anaemia	D50 D50.0 D50.1 D50.8 D50.9
Irritable bowel syndrome	K58 K58.0 K58.9
Lung cancer	C34 C34.0 C34.1 C34.2 C34.3 C34.8 C34.9
Multiple myeloma	C90.0
Multiple sclerosis	G35 G35X

Myocardial infarction	I21.9
Obesity	E66 E66.0 E66.1 E66.2 E66.8 E66.9
Osteoarthritis	M15.9
Osteoporosis	M81 M81.0 M81.1 M81.2 M81.3 M81.4 M81.5 M81.6 M81.8 M81.9
Parkinson's disease	G20 G20X
Peptic ulcer disease	K27 K27.0 K27.1 K27.2 K27.3 K27.4 K27.5 K27.6 K27.7 K27.9
Peripheral vascular disease	I73.9
Pneumonia	J18 J18.0 J18.1 J18.2 J18.8 J18.9
Prostate cancer	C61 C61X
Psoriasis	L40 L40.0 L40.1 L40.2 L40.3 L40.4 L40.5 L40.8 L40.9
Pulmonary embolism	I26.9
Renal cancer	C64 C64X
Renal colic	N23 N23X
Rheumatoid arthritis	M06.9 M06.90 M06.91 M06.92 M06.93 M06.94 M06.95 M06.96 M0697 M06.98 M06.99
Schizophrenia	F20 F20.0 F20.1 F20.2 F20.3 F20.4 F20.5 F20.6 F20.8 F20.9
Sickle cell disease	D57 D57.0 D57.1 D57.2 D57.3 D57.8
Systemic lupus erythematosus	M32.9
Sleep apnoea	G47.3
Stroke	I63 I63.0 I63.1 I63.2 I63.3 I63.4 I63.5 I63.6 I63.8 I63.9
Transient ischaemic attack	G45 G45.0 G45.1 G45.2 G45.3 G45.4 G45.8 G45.9
Type 1 diabetes mellitus	E10 E10.0 E10.1 E10.2 E10.3 E10.4 E10.5 E10.6 E10.7 E10.8 E10.9

Type 2 diabetes mellitus	E11 E11.0 E11.1 E11.2 E11.3 E11.4 E11.5 E11.6 E11.7 E11.8 E11.9
Ulcerative colitis	K51 K51.0 K51.2 K51.3 K51.4 K51.5 K51.8 K51.9
Venous thromboembolism	I80.2

Appendix 6.2 Associations of diseases in the UHB

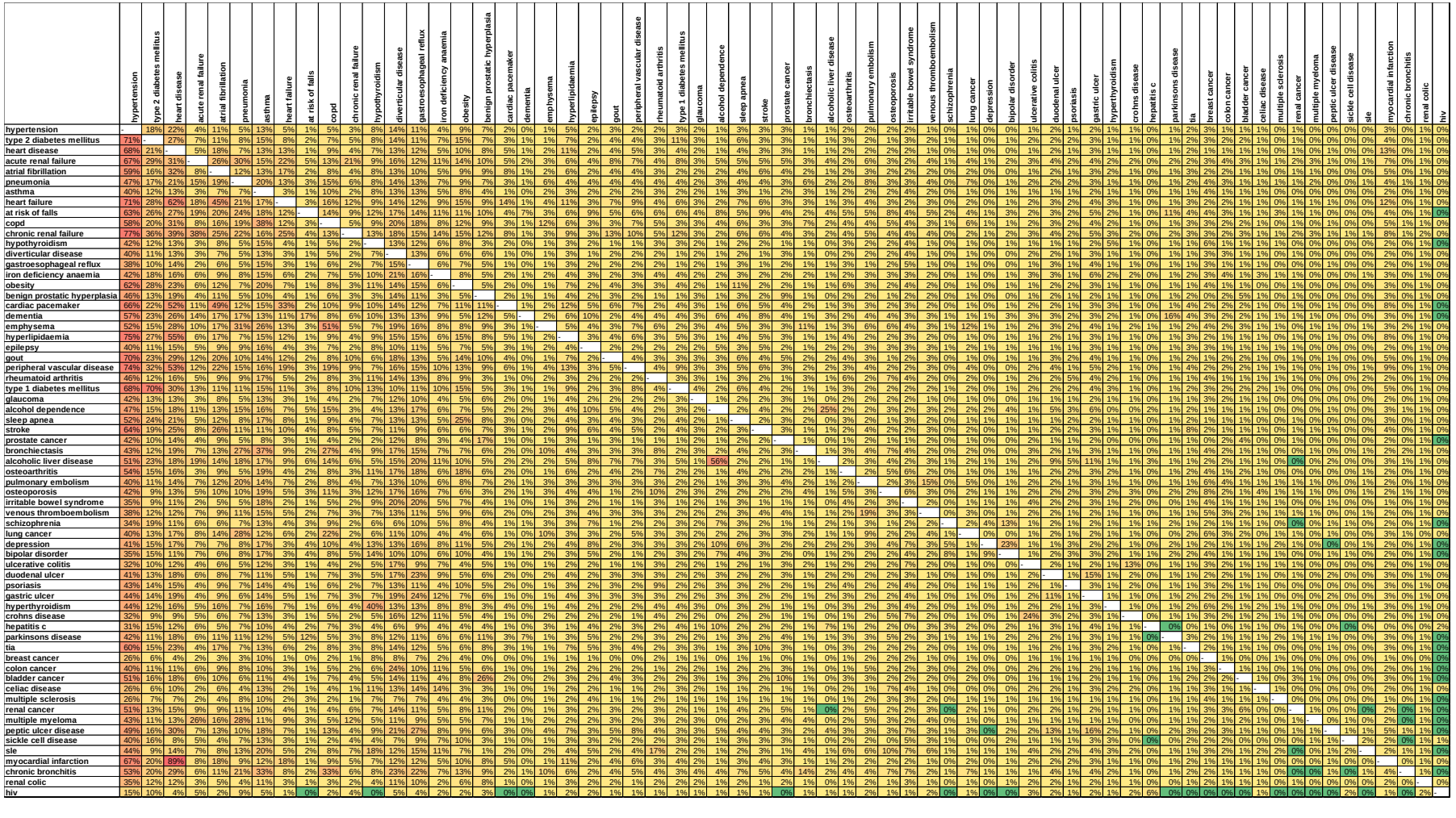


For a row condition, there is a % of patients that also have the column condition. These columns are ordered in descent order by the sum of probabilities.

Note: even though the data in the UHB does not contain all the patient life data there are comorbidities associated with the patient visit. N = 167737.

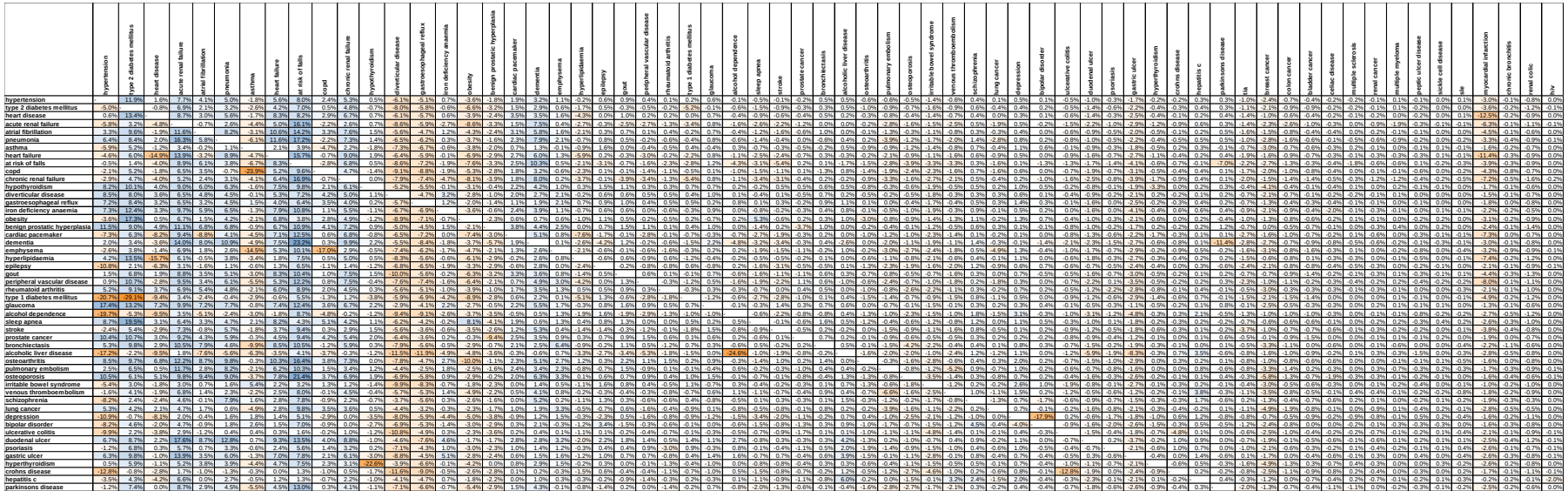
Source: UHB Trust in-patient HES. Minimum episode start date: 2014/11/03. Maximum episode end date: 2018/05/31.

Appendix 6.3 Associations of diseases in the UK Biobank



For a row condition, there is a % of patients that also have the column condition. These values are ordered in descent order by the sum of probabilities from **Appendix 5.2**. Note that the UK Biobank contain data from a patients' "whole life". N = 392296. Source: UK Biobank Main and Secondary ICD-10s.

Appendix 6.4 Comorbidities patterns between the UHB to the UK Biobank



The difference from the UHB to the UK Biobank. A positive value, marked in blue, indicates more association of a row condition to the column condition in the UHB set of patients. Negative values, highlighted in orange, indicate a higher association in the UK Biobank. Note that the UK Biobank contains the whole life records, while the UHB data is from secondary care.

Appendix 6.5 List of 28 important conditions

Condition	ICD-10
Anxiety	F40 F41
Asthma	J45
Atrial fibrillation	I48
Bronchiectasis	J47
Chronic bronchitis	J41 J42 J43 J44
Chronic kidney disease	N18 N19
Chronic liver disease	K70 K71 K72 K73 K74
Dementia	F00 F01 F02 F03
Depression	F32 F33
Diabetes	E10 E11 E12 E13 E14
Duodenal and gastric ulcer	K25 K26 K27 K28
Epilepsy	G40
Heart failure	I50
Human immunodeficiency virus	B20 B21 B22 B23 B24
Hypertension	I10 I11 I12 I13 I15
Hyperthyroidism	E05
Hypothyroidism	E02 E03 E01.8
Inflammatory bowel disease	K50 K51
Ischaemic heart disease	I20 I21 I22 I23 I24 I25
Malignant neoplasms	C00 C01 C02 C03 C04 C05 C06 C07 C08 C09 C10 C11 C12 C13 C14 C15 C16 C17 C18 C19 C20 C21 C22 C23 C24 C25 C26 C30 C31 C32 C33 C34 C37 C38 C39 C40 C41 C43 C44 C45 C46 C47 C48 C49 C50 C51 C52 C53 C54 C55 C56 C57 C58 C60 C61 C62 C63 C64 C65 C66 C67 C68 C69 C70 C71 C72 C73 C74 C75 C76 C77 C78 C79 C80 C81 C82 C83 C84 C85 C86 C88 C90 C91 C92 C93 C94 C95 C96 C97
Osteoarthritis	M15
Osteoporosis	M80 M81 M82
Parkinson's disease	G20
Peripheral vascular disease	I70 I71 I72 I73 I74
Rheumatoid arthritis	M05 M06
Severe mental illness	F20 F31.4 F31.5 F32.3
Sleep apnoea	G47.3
Stroke/TIA	G45 I61 I62 I63 I64

Appendix 6.6 Description of COPD clusters

	Categorical Variables	Fisher tests (p-val)	Group 1 (4081)	Group 2 (348)	Group 3 (1766)
Basic information	Sex (Male)	p < 1e-3	2212 (54%)	132 (38%)	1263 (72%)
	BMI	p < 1e-3			
	Normal		1473 (36%)	129 (37%)	370 (21%)
	Obesity		912 (22%)	77 (22%)	699 (40%)
	Overweight		1581 (39%)	131 (38%)	671 (38%)
	Unclear		17 (0%)	0 (0%)	13 (1%)
	Underweight		98 (2%)	11 (3%)	13 (1%)
	Smoking packs per year	p < 1e-3			
	<20		567 (14%)	58 (17%)	228 (13%)
	>60		509 (12%)	39 (11%)	289 (16%)
	20-40		1217 (30%)	107 (31%)	491 (28%)
	40-60		919 (23%)	77 (22%)	388 (22%)
	Unclear		869 (21%)	67 (19%)	370 (21%)
	Smoking Status	p < 1e-3			
	Mixed		1730 (42%)	184 (53%)	659 (37%)
	Prefer not to answer		34 (1%)	3 (1%)	15 (1%)
	Unclear		438 (11%)	44 (13%)	167 (9%)
	White		1879 (46%)	117 (34%)	925 (52%)
	Blood pressure diastolic	p < 1e-3			
	high		804 (20%)	59 (17%)	401 (23%)
	normal		3043 (75%)	267 (77%)	1272 (72%)
	Unclear		234 (6%)	22 (6%)	93 (5%)
	Blood pressure systolic	p < 1e-3			
	high		2056 (50%)	149 (43%)	1004 (57%)
	normal		1791 (44%)	177 (51%)	669 (38%)
	Unclear		234 (6%)	22 (6%)	93 (5%)
	Pulse rate	0.789605197			
	High		8 (0%)	0 (0%)	3 (0%)
	Normal		223 (5%)	22 (6%)	87 (5%)
	Unclear		3850 (94%)	326 (94%)	1676 (95%)
Comorbidities	Anxiety	p < 1e-3	13 (0%)	143 (41%)	56 (3%)
	Asthma	p < 1e-3	850 (21%)	147 (42%)	490 (28%)
	Atrial Fibrillation	p < 1e-3	105 (3%)	26 (7%)	450 (25%)
	Bronchiectasis	0.148425787	115 (3%)	12 (3%)	66 (4%)
	Chronic Kidney Disease	p < 1e-3	28 (1%)	19 (5%)	186 (11%)
	Chronic Liver Disease	p < 1e-3	29 (1%)	28 (8%)	22 (1%)
	Dementia	p < 1e-3	9 (0%)	6 (2%)	20 (1%)
	Depression	p < 1e-3	139 (3%)	199 (57%)	89 (5%)
	Diabetes	p < 1e-3	117 (3%)	23 (7%)	612 (35%)
	Duodenal and Gastric Ulcer	p < 1e-3	144 (4%)	27 (8%)	110 (6%)
	Epilepsy	p < 1e-3	49 (1%)	33 (9%)	46 (3%)
	Heart Failure	p < 1e-3	24 (1%)	6 (2%)	314 (18%)
	HIV	1	3 (0%)	0 (0%)	1 (0%)
	Hypertension	p < 1e-3	985 (24%)	118 (34%)	1554 (88%)
	Hyperthyroidism	p < 1e-3	10 (0%)	30 (9%)	25 (1%)
	Hypothyroidism	p < 1e-3	189 (5%)	75 (22%)	117 (7%)

	Inflammatory Bowel Disease	p < 1e-3	44 (1%)	20 (6%)	42 (2%)
	Ischaemic Heart Disease	p < 1e-3	265 (6%)	46 (13%)	1048 (59%)
	Malignant Neoplasms	p < 1e-3	706 (17%)	96 (28%)	333 (19%)
	Osteoarthritis	p < 1e-3	35 (1%)	29 (8%)	37 (2%)
	Osteoporosis	p < 1e-3	128 (3%)	85 (24%)	53 (3%)
	Parkinson's Disease	p < 1e-3	5 (0%)	6 (2%)	12 (1%)
	Peripheral Vascular Disease	p < 1e-3	69 (2%)	19 (5%)	356 (20%)
	Rheumatoid Arthritis	p < 1e-3	83 (2%)	55 (16%)	62 (4%)
	Severe Mental Illness	p < 1e-3	20 (0%)	17 (5%)	9 (1%)
	Sleep Apnoea	p < 1e-3	38 (1%)	12 (3%)	79 (4%)
	Stroke/TIA	p < 1e-3	42 (1%)	17 (5%)	189 (11%)
Biochemistry tests	Albumin to Creatinine Ratio	p < 1e-3			
	High		289 (7%)	20 (6%)	238 (13%)
	Intermediate		1247 (31%)	114 (33%)	638 (36%)
	Normal		9 (0%)	0 (0%)	4 (0%)
	Unclear		2536 (62%)	214 (61%)	886 (50%)
	eGFR	p < 1e-3			
	Low		132 (3%)	13 (4%)	167 (9%)
	Normal		3709 (91%)	314 (90%)	1491 (84%)
	Unclear		240 (6%)	21 (6%)	108 (6%)
	Cholesterol HDL	p < 1e-3			
	High		3222 (79%)	265 (76%)	1196 (68%)
	Normal		345 (8%)	32 (9%)	326 (18%)
	Unclear		514 (13%)	51 (15%)	244 (14%)
	Cholesterol LDL	p < 1e-3			
	High		2680 (66%)	206 (59%)	730 (41%)
	Normal		1155 (28%)	119 (34%)	927 (52%)
	Unclear		246 (6%)	23 (7%)	109 (6%)
	Cholesterol Total	p < 1e-3			
	High		2711 (66%)	220 (63%)	709 (40%)
	Normal		1128 (28%)	107 (31%)	951 (54%)
	Unclear		242 (6%)	21 (6%)	106 (6%)
	CRP	p < 1e-3			
	High		897 (22%)	72 (21%)	504 (29%)
	Normal		2932 (72%)	255 (73%)	1152 (65%)
	Unclear		252 (6%)	21 (6%)	110 (6%)
	Eosinophil percent	0.043478261			
	High		1308 (32%)	114 (33%)	558 (32%)
	Moderate/high		972 (24%)	76 (22%)	447 (25%)
	Normal		1629 (40%)	130 (37%)	691 (39%)
	Unclear		172 (4%)	28 (8%)	70 (4%)
	Haemoglobin count	p < 1e-3			
	High		53 (1%)	3 (1%)	14 (1%)
	Low		212 (5%)	27 (8%)	221 (13%)
	Normal		3650 (89%)	291 (84%)	1466 (83%)
	Unclear		166 (4%)	27 (8%)	65 (4%)
	HBA1c	p < 1e-3			
	High		94 (2%)	12 (3%)	286 (16%)

	Normal		3749 (92%)	302 (87%)	1363 (77%)
	Unclear		238 (6%)	34 (10%)	117 (7%)
	Lymphocytes percent	p < 1e-3			
	High		55 (1%)	11 (3%)	19 (1%)
	Normal		3854 (94%)	309 (89%)	1677 (95%)
	Unclear		172 (4%)	28 (8%)	70 (4%)
	Neutrophils percent	0.002998501			
	High		293 (7%)	37 (11%)	142 (8%)
	Normal		3616 (89%)	283 (81%)	1554 (88%)
	Unclear		172 (4%)	28 (8%)	70 (4%)
	Platelets count	0.002998501			
	High		138 (3%)	6 (2%)	41 (2%)
	Normal		3777 (93%)	315 (91%)	1660 (94%)
	Unclear		166 (4%)	27 (8%)	65 (4%)
	Triglycerides	p < 1e-3			
	High		1173 (29%)	102 (29%)	655 (37%)
	Normal		2666 (65%)	225 (65%)	1002 (57%)
	Unclear		242 (6%)	21 (6%)	109 (6%)
	Vitamin D	0.23838081			
	Low		2221 (54%)	197 (57%)	1016 (58%)
	Normal		1437 (35%)	118 (34%)	587 (33%)
	Unclear		423 (10%)	33 (9%)	163 (9%)
	White cell count	0.005997001			
	High		326 (8%)	32 (9%)	170 (10%)
	Normal		3589 (88%)	289 (83%)	1531 (87%)
	Unclear		166 (4%)	27 (8%)	65 (4%)
Symptoms	Anxiety	p < 1e-3			
	No		452 (11%)	15 (4%)	155 (9%)
	Unclear		3433 (84%)	311 (89%)	1543 (87%)
	Yes		196 (5%)	22 (6%)	68 (4%)
	Chest Pain	p < 1e-3			
	No		2541 (62%)	193 (55%)	982 (56%)
	Unclear		79 (2%)	9 (3%)	35 (2%)
	Yes		1461 (36%)	146 (42%)	749 (42%)
	Chronic Cough	0.00149925			
	No		227 (6%)	14 (4%)	63 (4%)
	Unclear		3608 (88%)	321 (92%)	1617 (92%)
	Yes		246 (6%)	13 (4%)	86 (5%)
	Chronic Pain	p < 1e-3			
	No		19 (0%)	0 (0%)	4 (0%)
	Unclear		3935 (96%)	319 (92%)	1689 (96%)
	Yes		127 (3%)	29 (8%)	73 (4%)
	Chronic Phlegm	0.001999			
	No		275 (7%)	16 (5%)	83 (5%)
	Unclear		3608 (88%)	321 (92%)	1617 (92%)
	Yes		198 (5%)	11 (3%)	66 (4%)
	Fatigue	p < 1e-3			
	No		256 (6%)	6 (2%)	62 (4%)
	Unclear		3429 (84%)	311 (89%)	1543 (87%)

	Yes		396 (10%)	31 (9%)	161 (9%)
	Low mood	p < 1e-3			
	No		2755 (68%)	152 (44%)	1179 (67%)
	Unclear		236 (6%)	14 (4%)	129 (7%)
	Yes		1090 (27%)	182 (52%)	458 (26%)
	Poor sleep	0.117441279			
	No		771 (19%)	49 (14%)	341 (19%)
	Unclear		11 (0%)	1 (0%)	2 (0%)
	Yes		3299 (81%)	298 (86%)	1423 (81%)
	Shortness of breath	p < 1e-3			
	No		687 (17%)	42 (12%)	235 (13%)
	Unclear		2799 (69%)	265 (76%)	1236 (70%)
	Yes		595 (15%)	41 (12%)	295 (17%)
	Weight Change	p < 1e-3			
	Gained weight		1263 (31%)	114 (33%)	581 (33%)
	Lost weight		676 (17%)	93 (27%)	333 (19%)
	No		2031 (50%)	127 (36%)	818 (46%)
	Unclear		111 (3%)	14 (4%)	34 (2%)
	Wheeze	0.035482259			
	No		1031 (25%)	95 (27%)	514 (29%)
	Unclear		113 (3%)	12 (3%)	50 (3%)
	Yes		2937 (72%)	241 (69%)	1202 (68%)
Respiratory tests	FEV1/FVC z-score	p < 1e-3			
	low		0 (0%)	0 (0%)	1 (0%)
	normal		3692 (90%)	299 (86%)	1539 (87%)
	Unclear		389 (10%)	49 (14%)	226 (13%)
	FEV1 % predict	p < 1e-3			
	mild		383 (9%)	46 (13%)	159 (9%)
	moderate		1287 (32%)	107 (31%)	557 (32%)
	severe		443 (11%)	31 (9%)	147 (8%)
	Unclear		1903 (47%)	162 (47%)	893 (51%)
	very severe		65 (2%)	2 (1%)	10 (1%)
	FEV1 z-score	p < 1e-3			
	Normal		3692 (90%)	299 (86%)	1540 (87%)
	Unclear		389 (10%)	49 (14%)	226 (13%)
	FVC z-score	p < 1e-3			
	Low		16 (0%)	2 (1%)	4 (0%)
	Normal		3676 (90%)	297 (85%)	1536 (87%)
	Unclear		389 (10%)	49 (14%)	226 (13%)
	Peak expiratory flow rate	p < 1e-3			
	Low		3033 (74%)	277 (80%)	1143 (65%)
	Normal		1031 (25%)	71 (20%)	610 (35%)
	Unclear		17 (0%)	0 (0%)	13 (1%)

Appendix 6.7 Phenotypes collected

Phenotype
Urine 5-HIAA Excrn
A1AT
Angiotensin
Albumin
AlkP (until 21 Nov 11)
AlkP
ALT
AST
Antithrombin
B12 (Diaverum)
B12
Basos
Bilirubin
CD 16/56+ NK cells (x10^9/L)
Immunochemical C1 esterase inhibitor
Complement C3
Complement C4
Calcium
Caeruloplasmin
Corrected Ca
CD 19+ cells (x10^9/L)
CD3+ cells (x10^9/L)
CD4+ cells (x10^9/L)
CD8+ cells (x10^9/L)
Creatine Kinase
LA Normalised Ratio
Creatinine Clearance
Creatinine
CSF Glucose
CSF Protein
D4 Androstenedione
Diab Clinic Creat
DHA Sulphate
DRVVT Screen Ratio
dsDNA abs ELISA
Eosinophils
ESR
Factor XI
Factor VIII clotting
Iron
Ferritin
Fibrinogen
Folate
Fructosamine
Follicle-stimulating hormone
Free T3
Free T4
FVIIIrag
CD3+ T cells (cells/mm3)
CD4+ T cells (cells/mm3)
CD8+ T cells (cells/mm3)
Glom. Filt Rate
Gamma GT
Globulin
Glucose
Finger-stick Glucose
Hydrogen ion conc.
Haptoglobin
Hb (g/dL)
Haemoglobin A1c
Haemoglobin A2
Haemoglobin F
HCT
Hb (g/L)
IgA
IGF-1 (Old test method)
IGF-1
IgG
IgG1
IgG2
IgG3
IgM
HbA1c-IFCC
INR
K
Kappa/Lambda ratio
Lupus Anti Screen
Lactate Dehyd'nase (until 11 Mar 2012)
Lactate Dehyd'nase
Luteinizing hormone
Lymphocytes
MCH
MCHC (g/dL)
MCHC (g/L)
MCV
MG
Monocytes
Neutrophils
Ammonia
Noradrenaline (Excn)
17-OH-Progesterone
Osmolality
Protein C function
pCO2
Platelets
pO2
Phosphate
Prolactin
Free Protein S
ParaThyroid Hormone (Immulite)
Parathyroid Hormone (ng/L)
Parathyroid Hormone (pmol/L)
PTT Ratio
Plasma viscosity
CO Hb
Ionised Calcium PoC
Blood Cl
Blood Glucose
HCT on Blood Gas analyser
Hb
Blood K
Lactate
Met Hb
Blood Na
PoC SO2
RBC
RDW
Reticulocytes
Serum creatinine
Serum free kappa
Serum free lambda
Sex H'mone Bind Glob
Total Carbon Dioxide
Testosterone
Testosterone (mass spec)
Total Protein
Transferrin
Trig
TSH
Thrombin time (ratio)
Thrombin time (seconds)
Urine Cal (Excn.)
Urine creatinine
Urine Creat (Excn.)
Kappa (g/l)
Ur
Urate
Urine Urea (Excn.)
vWF Collagen Binding Activity
WBC

Appendix 6.8 SNPs associated with high glucose

CHR	SNP	BP	A1	TEST	N MISS	OR	STAT	P
16	rs111285796	20361087	TTCTGCCAGA	ADD	170340	0.8632	-4.785	1.71E-06
16	rs4293393	20364588	G	ADD	170393	0.8642	-4.753	2.00E-06
16	rs13333226	20365654	G	ADD	170410	0.8655	-4.712	2.46E-06
17	rs8065820	1957893	A	ADD	170257	1.192	4.708	2.50E-06
16	rs12917707	20367690	T	ADD	170318	0.8667	-4.631	3.65E-06
8	rs591346	9818065	C	ADD	170036	1.11	4.595	4.32E-06
8	rs62500966	52992325	A	ADD	170487	1.276	4.587	4.50E-06
11	rs900145	13293905	C	ADD	170327	0.8904	-4.553	5.29E-06
22	rs28582261	43883747	G	ADD	170102	0.9024	-4.503	6.70E-06
11	rs118018586	80839429	T	ADD	170245	1.401	4.427	9.56E-06
18	rs41378153	47516411	T	ADD	170431	0.7349	-4.414	1.02E-05
9	rs117311531	115859898	G	ADD	170322	0.6994	-4.401	1.08E-05
6	rs118034128	94935813	C	ADD	170129	1.324	4.388	1.14E-05
3	rs9844829	139979092	T	ADD	169909	0.8731	-4.371	1.24E-05
20	rs4346455	10879336	C	ADD	170225	1.104	4.366	1.26E-05
18	rs10503084	61804949	A	ADD	170209	1.153	4.356	1.32E-05
19	rs3108586	37230491	T	ADD	170410	0.9008	-4.349	1.37E-05
9	rs4740912	7791515	G	ADD	168380	0.9047	-4.315	1.60E-05
12	rs11183609	47167837	C	ADD	170123	1.141	4.286	1.82E-05
6	rs59856354	104277260	A	ADD	170345	0.7916	-4.281	1.86E-05
20	rs75637135	6803897	T	ADD	169965	1.257	4.281	1.86E-05
16	rs9924441	9047198	T	ADD	169946	0.8925	-4.228	2.35E-05
6	rs72988541	104241528	T	ADD	169863	0.7916	-4.227	2.36E-05
8	rs656319	9814411	A	ADD	170267	1.1	4.205	2.61E-05
7	rs117076099	93956047	C	ADD	170110	1.398	4.17	3.05E-05
15	rs79287552	57592248	T	ADD	170285	1.181	4.15	3.32E-05
4	rs13114161	116378597	G	ADD	169521	0.7573	-4.138	3.50E-05
11	rs10832021	13324530	G	ADD	169948	0.8994	-4.13	3.62E-05
7	rs73227485	93516460	A	ADD	169727	1.195	4.13	3.63E-05
7	rs58919541	156297224	A	ADD	170377	0.7332	-4.125	3.71E-05
1	rs75129533	247531139	A	ADD	170139	1.238	4.124	3.72E-05
4	rs75138976	19225971	A	ADD	170125	0.8494	-4.123	3.74E-05
1	rs74850688	63506318	T	ADD	169737	1.227	4.102	4.10E-05
17	rs35394823	1943880	G	ADD	170446	1.163	4.098	4.17E-05
3	rs1596152	3749436	G	ADD	169461	1.099	4.09	4.32E-05
10	rs4897807	133490980	T	ADD	169548	0.9103	-4.078	4.55E-05
21	rs2234694	33038865	C	ADD	170373	1.228	4.077	4.56E-05
21	rs2833507	33163612	A	ADD	170049	1.141	4.077	4.57E-05
9	rs4740935	8340155	A	ADD	169960	1.1	4.07	4.70E-05

Relation of SNPs with p-value < $5e^{-5}$ identified when comparing patients with high glucose against the ones without high glucose phenotype reported in the UK Biobank.