Cardiovascular disease in Type 2 Diabetes Mellitus: A Precision Medicine approach applying Artificial Intelligence for Heart Failure and Mortality Prediction

Narinder Kaur

MSc, BSc (Hons)

Submitted in fulfilment of the requirements for the

Degree of Doctor of Philosophy

School of Cardiovascular & Metabolic Health

College of Medical, Veterinary and Life Sciences

University of Glasgow

January 2025

Abstract

Cardiovascular diseases (CVDs) are the leading cause of morbidity and mortality worldwide, despite substantial advances in diagnosis and treatment. People who suffer from cardiovascular disease often have multiple risk factors and other chronic conditions. Additionally, medical events may be strongly influenced by socioeconomic status. Patient information can be obtained from electronic medical records (EMRs) that, unlike data from clinical trials and registries, provide a broad range of patient characteristics representative of the general population. EMRs covering a population of ~1.1 million people in Greater Glasgow & Clyde (GG&C) Health Board NHS over 50 years (the age at which the incidence and prevalence of disease affecting older people increase rapidly) were used. Information such as demographics, laboratory tests, primary-care prescriptions, hospitalisations and mortality was retrieved. Several steps were required to ensure that the extracted information was appropriate for analysis and transformed for investigations beyond traditional statistics. Accordingly, data on patients with type-2 diabetes mellitus (T2DM) were obtained to examine their health trajectories, including, incident heart failure and death. Novel risk prediction models were built to help understand the development of heart failure (HF) in patients with T2DM. The models were developed using random survival forest (RSF) methodology. This research highlights the limitations of traditional regression models and demonstrates the improvement of risk prediction with RSF methods, which outperformed traditional approaches in both discrimination and calibration. State-of-the-art machine learning interpretation was applied to discover key contributing factors to the development of heart failure and to all-cause mortality. External validation was applied by acquiring EMRs from Hong Kong, Special Administrative Region (SAR) China. The inclusion of two diverse populations found little evidence of ethnicity-related differences in risk factors. GG&C key risk factors for incident HF were loop diuretics, atrial fibrillation (AF), history of coronary artery disease (CAD), older age, lower levels of estimated glomerular filtration rate (eGFR), haemoglobin and serum albumin. Similarly, for Hong Kong, key risk factors were use of loop diuretics, insulin, lower serum albumin, haemoglobin, lymphocyte counts and eGFR. The model based on Hong Kong data showed slightly better performance compared to the Glasgow cohort for incident heart failure (C-index 0.88 and 0.87) and all-cause mortality (0.85 and 0.83).

In both cohorts' older women were more likely to be prescribed loop diuretics. Whether loop diuretics are just a marker of undiagnosed heart failure or whether they accelerate the progression of cardiovascular and renal disease is uncertain. Another key similarity was that patients had prevalent chronic kidney disease (CKD) events in the prescribed loop diuretics groups. Treatment with loop diuretics was strongly associated with all-cause mortality in GG&C and Hong Kong. (GG&C: adjusted hazard ratio: 2.93, (95% CI: 2.821 to 3.04); Hong Kong: adjusted hazard ratio: 1.75 (95% CI: 1.72 to 1.77). Only a minority of patients prescribed loop diuretics had a diagnosis of heart failure, end-stage renal disease or resistant hypertension. Finally, further investigation of social deprivation in GG&C underlined that 41% patients with T2DM were in the most deprived socioeconomic quintile and that they had a 36% higher rate for all-cause mortality compared to those who were least deprived (adjusted HR: 1.36, 95% CI 1.24–1.50, $p < 0.005$).

Contents

List of Tables

List of Figures

Acknowledgements

This PhD thesis is a testimony of Christ. God witnessed me Graduate before this journey began. From the very beginning God directed me into this multidisciplinary research field with the invaluable support from supervisory team. My primary supervisor, Professor John Cleland, provided great expertise in heart failure; his wisdom and support will always be cherished. Professor John made time to attend most of my presentations at conferences. Dr Fani Deligianni encouraged me to aim higher. Her experience in computing science gave me the insights to improve my approach to analyses. Dr Pierpaolo Pellicori, who substantially helped navigate through the medical statistics and reminded me to enjoy this PhD journey. A special thanks to my inspirational colleagues Dr Antonio Iaconelli, Dr Jocelyn Friday and Dr Yola Jones.

From the bottom of my heart I want thank my grandparents, especially both my grandfather's Gurbachan S. Loha and Harbajan S. Parker for being the reason why I have a passion for research to improve the healthcare for people with type 2 diabetes mellitus with or at great risk of cardiovascular disease. Most importantly, words cannot explain how much I appreciate my Dad Dahmaindar Loha, Mum Raj Loha, brother Ricco Loha and sister-in-law Reshma Loha for being the rocks supporting these years of studying, researching and travelling. Their love allowed me to prosper. Finally, I am blessed to say thank you to my fiancé, James Kumar who came at the right time and prayed for my final PhD year.

Declaration

I, Narinder Kaur, declare that the work presented in this PhD thesis is my own and have not been submitted for any other degree at the University of Glasgow or any other institution. Where the work of others has been used, it is acknowledged and appropriately referenced.

Abbreviations & Acronyms

95% CI 95% confidence interval.

ACEi angiotensin-converting enzyme inhibitor.

AF atrial fibrillation.

ALT alanine aminotransferase.

ARB angiotensin receptor blocker.

AST aspartate aminotransferase

ARNi angiotensin receptor-neprilysin inhibitor.

BHF British Heart Foundation

BMI body mass index.

BNF British National Formulary.

CAD coronary artery disease.

CCB calcium channel blocker.

CHI Community Health Index.

CKD chronic kidney disease.

CKD-EPI Chronic Kidney Disease Epidemiology Collaboration.

COD cause of death.

COPD chronic obstructive pulmonary disease.

CPH Cox proportional hazards.

CSV comma-separated value.

CVD cardiovascular disease

DOB date of birth.

DOD date of death.

DPP-4i Dipeptidyl peptidase-4 inhibitor.

eGFR estimated glomerular filtration rate.

EMRs electronic medical records.

EPR electronic patient records.

ESC European Society of Cardiology.

GLP-1RAs Glucagon-like peptide-1 receptor agonists

GP general practitioner.

GP LES General Practice Local Enhanced Services.

HbA1c glycated haemoglobin.

HF heart failure.

HFpEF heart failure with preserved ejection fraction.

HR hazard ratio.

ICD-9 International Classification of Diseases, 9th Revision.

ICD-10 International Classification of Diseases, 10th Revision.

KM Kaplan-Meier survival estimator.

LD loop diuretic.

LES Local Enhanced Services.

MI myocardial infarction.

MNAR missing not at random.

MRA mineralocorticoid receptor antagonists.

NHS National Health Service.

NICE National Institute for Health and Care Excellence

GG&C Greater Glasgow & Clyde Health Board.

PAD peripheral arterial disease.

PIS Prescribing Information System.

RCT Randomised controlled trials.

RSF Random Survival Forest.

SCI Diabetes Scottish Care Information-Diabetes Collaboration.

SCI Store Scottish Care Information Store.

SGLT2i sodium-glucose co-transporter-2 inhibitor.

SHAP SHapley Additive exPlanations.

SIMD Scottish Index of Multiple Deprivation.

T2DM type-2 diabetes mellitus.

UK United Kingdom.

WHO World Health Organisation.

WoS West of Scotland.

Chapter 1 Introduction

This chapter summarises the research background, problem area and research objectives.

1.1 Background

1.1.1 Introduction to Cardiovascular Disease

Cardiovascular disease is common in middle and older-age adults and the leading cause of death globally, taking an estimated 17.9 million lives each year (Di Cesare et al., 2024). The most common heart condition in Scotland is coronary artery disease (CAD) (Health Intelligence Team, 2024), Glasgow has one of the highest levels of CVD in Western Europe, which has been attributed to high rates of smoking, unhealthy diet, obesity and poor lifestyle choices, which are all strongly associated with lower educational attainment and socioeconomic deprivation. Patients with type 2 diabetes mellitus (T2DM) are at increased risk of developing atherosclerosis, either because they are more likely to develop high blood pressure (hypertension), high levels of blood fat (both cholesterol and triglycerides) and kidney dysfunction or because of the effects of dysglycaemia (high blood sugar) itself.

There are many types of CVD, with atherosclerosis being the most common. Other CVD include diseases of the heart valves, irregularity of heart rhythm (e.g. atrial fibrillation or conduction system block), infiltration of the heart muscle making it stiff (e.g. with amyloid), thickening of the heart muscle due to high blood pressure or problems with heart muscle fibres themselves, which is called cardiomyopathy. There are several different types of cardiomyopathies. Dilated cardiomyopathy means that the contraction of the heart is weak, leading the left ventricle (the heart's main pumping chamber) to enlarge and dilate (John Hopkins Medicine, 2021). Hypertrophic cardiomyopathies cause the heart muscle to become thick and stiff, which impairs its pumping action. The prevalence of AF is predicted to double over the next 30 years due to changing demographics and the rise of lifestyle risk factors (Jones et al., 2020). AF is associated with underlying CVD with increased risk of death, stroke and heart failure (Michaud and Stevenson, 2021). Many patients will have more than one type of CVD and T2DM may contribute to or complicate them all.

Heart failure is a final common pathway for many different CVD, including hypertension, diabetes, atherosclerosis, kidney disease and atrial fibrillation (many patients with heart failure will have all of these). Heart failure is common, often accompanied by exertional breathlessness that can be debilitating, associated with high rates of hospitalisation and with a poor prognosis (Groenewegen et al., 2020). The diagnosis is often made only very late in the course of the disease. Sadly, many patients with heart failure die before the diagnosis is made.

1.1.2 Clinical Presentation and Pathophysiology

The clinical presentation of cardiovascular disease varies widely. **Table 1** shows the Symptoms and Signs.

Table 1 Cardiovascular Disease Symptoms & Signs

Cardiovascular Disease Symptoms/Signs	Description
Chest Pain/Discomfort	Substernal pressure, tightness, or discomfort; may radiate to jaw, shoulders, arms, or upper back. Common in acute coronary syndrome (ACS) and angina pectoris.
Dizziness and Lightheadedness	Results from arrhythmias, heart failure, or transient ischemic attacks (TIAs).
Syncope (Fainting)	Sudden loss of consciousness; occurs with serious arrhythmias or severe aortic stenosis.
Sweating (Diaphoresis)	Excessive sweating, often described as "cold sweat"; accompanies ACS and other acute cardiovascular events.
Nausea and Gastrointestinal Symptoms	Nausea, vomiting, and abdominal pain; associated with ACS and heart failure, sometimes mimicking gastrointestinal disorders.
Anxiety and Depression	Higher rates in patients with persistent angina or heart failure; influences perception and reporting of physical symptoms.
Cyanosis	Bluish discoloration of skin and mucous membranes. Indicates poor oxygenation, often seen in severe heart failure or congenital heart disease.
Palpitations	Sensation of irregular or rapid heartbeat; common in arrhythmias such as atrial fibrillation.
Heart Failure Symptoms/Signs	
Shortness of Breath (Dyspnoea)	Occurs at rest or with exertion; seen in heart failure, ACS and valvular heart disease. Often accompanied by fatigue.
Fatigue and Weakness	Notable in heart failure; considered atypical in ACS, especially in women and the elderly.
Oedema (Swelling)	Swelling in legs, ankles, and feet; indicative of fluid retention due to poor cardiac function, often seen in heart failure.
Jugular Venous Distention (JVD)	Visible bulging of the jugular veins. Indicates increased central venous pressure, commonly seen in heart failure.
Tachycardia/Bradycardia	Can indicate various cardiac conditions, including arrhythmias and heart block.

Adapted from Corrine Y Jurgens, Christopher S Lee et al. Circulation, 146, 2022 & Mendis et al., "Global Atlas on Cardiovascular Disease Prevention and Control," World Health Organization (WHO), 2011.

A symptom is a manifestation of disease evident to the patient while a sign is a manifestation of disease that a health professional perceives. Signs are considered more objective, although eliciting them depends on the skills, experience and opinion of the health professional. Signs usually develop much later in the course of a disease compared to symptoms; waiting for signs to appear can delay diagnosis by days, months or even years. Symptoms are considered more subjective but they are the 'hard' reality that the patient experiences.

Symptoms will often be the reason why a person seeks medical advice and assistance. Despite the critical importance of acknowledging both symptoms and signs, they are often overlooked or underestimated by both patients and clinicians. Patients may dismiss symptoms as insignificant, hope they will resolve or attribute them to ageing, especially when symptoms are mild or non-specific. For example, fatigue or light headedness might not prompt immediate medical attention, even though they can indicate serious underlying CVD.

Clinicians may miss subtle signs or fail to understand the significance of certain symptoms, particularly when they do not fit the classic presentation of a condition. Moreover, many patients have trouble in providing a verbal description of their symptoms and there is often great variation amongst individuals and even from the same individual over time or depending on who they are speaking with. This can lead to miscommunication and misunderstanding and complicate the diagnostic process.

Atherosclerosis, which develops when cholesterol builds up in the arteries causing narrowing, that restricts blood flow (Fonarow, 2007), is the most important and common CVD. Atherosclerotic disease of the coronary arteries can cause angina, myocardial infarction (heart attack), sudden death or heart failure. Atherosclerosis of the arteries to the brain can cause a stroke, leading to permanent neurological damage. Stroke is a major cause of disability and mortality worldwide ('Cerebrovascular Disease - AANS', 2024) . Atherosclerosis of the aorta can lead to bulging of the wall (aneurysms), which can rupture leading to catastrophic internal bleeding which is often fatal. Atherosclerosis of the arteries to the legs can cause muscles cramps (Sanchis-Gomar et al., 2016) during exercise (intermittent claudication) and may jeopardise the viability of the limb, requiring amputation (Li et al., 2020). Atherosclerosis is the primary underlying cause of CAD and stroke. It is characterised by the accumulation of lipid-rich plaques within the arterial wall over years or decades (**Figure 1**).

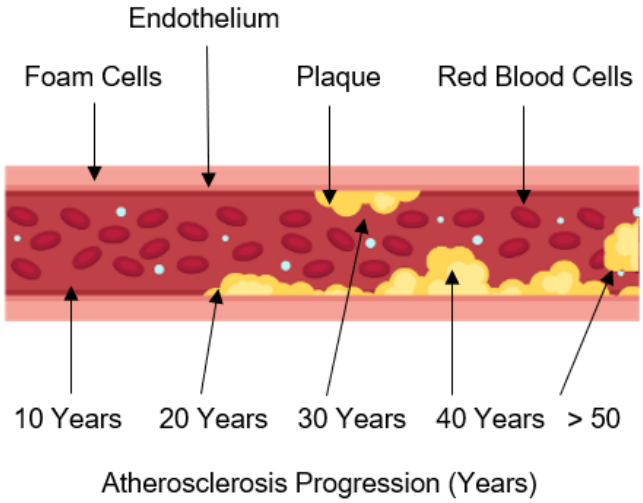


Figure 1 Atherosclerosis Progression

The process involves:

- Endothelial Dysfunction: The inner lining of the arteries (endothelium) becomes damaged due to factors like hypertension, smoking and high cholesterol. This dysfunction makes the endothelium more permeable to lipids and other substances.
- Lipid Accumulation and Oxidation: Low-density lipoproteins (LDL) penetrate the damaged endothelium and accumulate in the arterial wall. These lipids become oxidised, which attracts immune cells like macrophages.
- Formation of Foam Cells: Macrophages engulf oxidised LDL, transforming into foam cells. These foam cells accumulate to form fatty streaks, an early sign of atherosclerosis.
- Plaque Formation: Smooth muscle cells move to the inner layer of the artery, multiply, and produce proteins that form a fibrous cap over the lipid core. This plaque can bulge into the artery, narrowing it and reducing blood flow.
- Plaques can cause narrowing of the lumen of the artery restricting blood flow. If this occurs in the coronary arteries, it may cause angina during exercise. If it occurs in the leg arteries, it may cause muscle cramps during exercise.
- Plaque haemorrhage: as atherosclerosis develops new, fragile capillaries grow in from the vessels that surround the outside wall of the artery (vasa vasora). A similar process can cause eye problems in people with diabetes (diabetic retinopathy). If the bleeds are small, this attracts macrophages (white blood cells) that come to clean up the mess – but

the fat in the red cell membranes may be oxidised and turn the macrophages into foam cells. This may be an important mechanism of plaque growth. Larger bleeds will cause the plaque to rupture.

- Plaque Rupture and Thrombosis: Plaques can become unstable and rupture either because of bleeding, a large amount of 'lipid gruel' and/or stress to the vessel wall. Ulceration through the lining of the vessel exposes the inside of the plaque to the bloodstream, triggering the formation of a thrombus (blood clot), which can block the artery, leading to a heart attack (myocardial infarction) or stroke.

Inflammation plays a central role in endothelial dysfunction and the initiation and progression of atherosclerosis.

Atherosclerosis is a consequence of genetics, unhealthy lifestyle choices, co-morbid conditions (like hypertension, diabetes, obesity and hyperlipidaemia), environmental factors, cytokine activation (i.e. inflammation) and increasing age.

Some people have single gene defects (for instance familial hypercholesterolaemia) that may cause severe atherosclerosis affecting people even in their teenage years. Many people have lots of small defects in many genes (polygenic risk) that cumulatively increase the risk of atherosclerosis. Genetic propensity to hypertension and T2DM will also increase risk. Unhealthy lifestyle choices include smoking (especially tobacco), diets rich in saturated, oxidised fat (e.g. clarified butter, cooked fat-rich foods), processed foods (that usually have a high salt content), sedentary behaviours and excessive alcohol consumption. Environmental pollution, especially particulate matter from smoking, fires or exhaust fumes is highly atherogenic (and carcinogenic). In low-income countries, cooking is often done on an open fire inside the house with no conventional chimney, leading to high levels of air pollution. Obesity may play a central role in atherosclerosis through its association with hypertension, diabetes, diet and sedentary lifestyle. Obesity might also be responsible for low-grade chronic inflammation (Verma et al., 2024). Table 2 provides an overview of the major risk factors associated with cardiovascular disease (NICE, 2023; Health Intelligence Team, 2024a; Timmis et al., 2022; Mills et al., 2020; Segerer and Seeger, 2018; Moran et al., 2022; NICE Guidelines, 2014), which can be classed as either modifiable or non-modifiable.

Table 2 Major Risk Factors of Cardiovascular Disease

Risk Factor	Causes
Hypertension (High Blood Pressure)	Genetic predisposition, high salt intake, obesity, physical inactivity and excessive alcohol consumption
Dyslipidaemia (High Cholesterol)	Diet high in saturated and trans fats, Genetic factors (e.g., familial hypercholesterolemia), obesity and physical inactivity
Smoking	Tobacco use (active and passive exposure) , nicotine addiction, environmental and social factors
Diabetes Mellitus	Diet high in saturated and trans fats, genetic factors (e.g., familial hypercholesterolemia), obesity and physical inactivity
Obesity	High-calorie diet, lack of physical activity, genetic factors and metabolic disorders
Physical Inactivity	Sedentary lifestyle, lack of regular exercise and occupation-related inactivity
Poor Diet	High intake of saturated fats, trans fats and cholesterol, high consumption of processed foods and sugary beverages, low intake of fruits and vegetables
Excessive Alcohol Consumption	Heavy drinking habits, social and cultural influences and genetic predisposition
Age	Natural aging process leading to arterial stiffening and reduced cardiovascular function
Family History of CVD	Genetic predisposition and shared lifestyle factors within families
Sex	Male (higher risk at a younger age) and postmenopausal status in women (increased risk)
Stress	Chronic psychological stress, work-related stress and social and economic pressures
Chronic Kidney Disease	Diabetes, hypertension and polycystic kidney disease
Sleep Apnoea	Obesity, anatomical factors (e.g., enlarged tonsils), alcohol consumption and Smoking
Inflammation	Chronic inflammatory conditions (e.g., rheumatoid arthritis, lupus), infections and obesity

Adapted from British Heart Foundation. (2020). Cardiovascular Disease Statistics 2020, European Society of Cardiology. (2019). Cardiovascular Disease Statistics 2019, NICE Guidelines. (2019). Cardiovascular disease: risk assessment and reduction, including lipid modification & Benjamin, E. J., et al. (2019). Heart disease and stroke statistics—2019 update: a report from the American Heart Association. Circulation.

Diet, physical activity, obesity and smoking, can be altered through lifestyle changes and medical interventions. Non-modifiable risk factors include age and sex. Although the genome is relatively unmodifiable, the consequences of a genetic disposition to a high cholesterol or blood pressure may be readily modified by following a healthy lifestyle and by modern therapy. The interplay between these factors determines an individual's overall risk of developing CVD.

Socioeconomic status (SES) also plays an important role in the natural history of CVD (Psaltopoulou et al., 2017). SES itself is not conventionally classified as a risk factor for CVD but is an important determinant of many, including smoking, obesity and lifestyle (Schultz et al., 2018). Socioeconomic deprivation is also associated with lower educational attainment and poorer access to healthcare services (DD et al., 1997). Addressing socioeconomic disparities is essential to reduce the burden of CVD and improve health outcomes at both individual and population levels (Foster et al., 2018).

Ethnic disparities in CVD risk and health outcomes also exist (George et al., 2017), which may be attributed to a complex interplay of genetic, cultural and socioeconomic factors. South Asians, have a higher risk of CVD compared to those of European ancestry (Razieh et al., 2022). Scotland's population is mostly White (87%) (Scotland Census, 2022) and therefore analyses based on national data may conceal important differences in ethnic minority populations.

The cardiovascular disease continuum, illustrated in Figure 2, starts with a cluster of cardiovascular risk factors (Dzau et al., 2006). Once these risk factors occur, there are early functional changes in the vascular system (vascular ageing), activation of neuro-endocrine and inflammatory pathways, eventually resulting in end-organ damage and either sudden death due to an arrhythmia or vascular occlusion or the development of problems such heart failure, renal failure, disabling stroke or dementia.

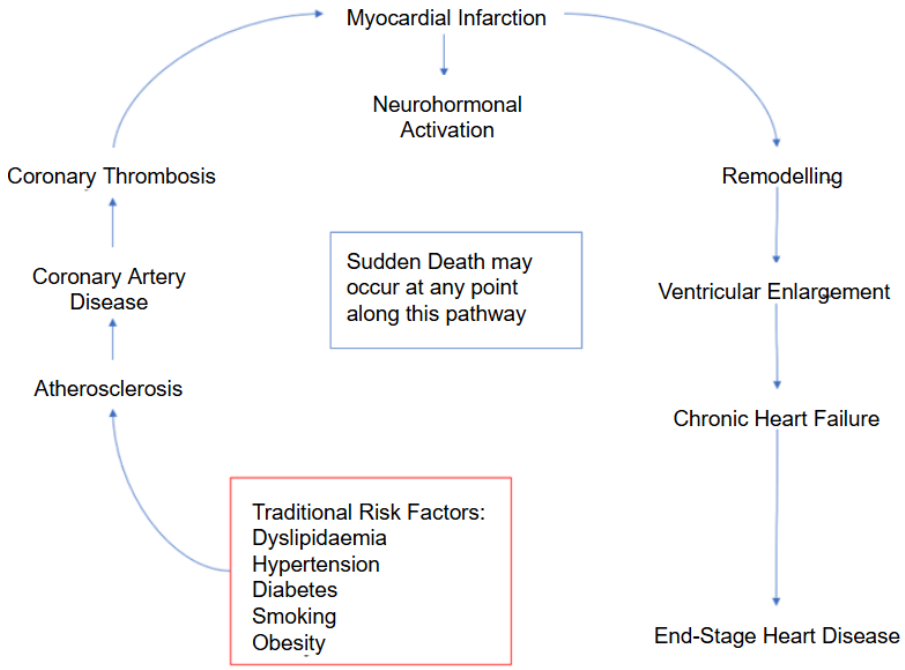


Figure 2 Cardiovascular Disease Continuum. Adapted from Dzau VJ Antman EM, Black HR, et al. The cardiovascular disease continuum validated: Clinical evidence of improved patient outcomes: Part I: Pathophysiology and clinical trial evidence (risk factors through stable coronary artery disease). Circulation. 2006; ;114(25):2850-2870. doi:10.1161/CIRCULATIONAHA.106.655688

1.1.3 Multimorbidity

Multimorbidity, defined as the coexistence of two or more chronic conditions (Hassaine et al., 2020), is increasingly common in patients with CVD (Kraemer, 1995). Multimorbidity makes patient management much more complex. Patients may struggle to cope with a greater burden of ill health, often exacerbated by depression. Clinicians may struggle because they lack the expertise to deal with one or more of the patients' problems. Treatments may struggle because they are contraindicated or ineffective if, for instance the patient has poor kidney function, or because treatments may adversely interact with one another. Hypertension, diabetes, obesity and CKD often co-exist, making treatment time consuming and complex (Guthrie et al., 2012). Among the various comorbid conditions that contribute to the development and progression CVD, T2DM stands out as a common, important and potentially modifiable risk factor (Ormazabal et al., 2018).

1.1.4 Type 2 Diabetes Mellitus and Heart Failure

T2DM is the most common form of diabetes worldwide (Einarson et al., 2018), accounting for approximately 90% of all cases in the UK. T2DM is characterised by insulin resistance. Endogenous insulin levels are usually high but not high enough to normalise blood glucose. Insulin resistance is associated with a cluster of metabolic abnormalities collectively known as metabolic syndrome and an increased risk of CVD which can be attributed both to shared risk factors, such as obesity, sedentary lifestyle and factors associated with and possibly consequence of T2DM, such as hypertension and dyslipidaemia. Insulin resistance and chronic hyperglycaemia, hallmarks of T2DM, contribute to endothelial dysfunction, plaque formation and atherosclerosis (Sattar and Gill, 2014; Ormazabal et al., 2018). Inflammation and oxidative stress also play important roles in the pathogenesis of both T2DM and atherosclerosis. Patients with T2DM are at increased risk of developing heart failure (HF), a clinical syndrome of cardiac dysfunction leading to congestion (JGF et al., 2021), meaning either an excess volume of blood (blood is mostly water) in the pulmonary (lung) or systemic venous systems (haemodynamic congestion, like water being held back by a dam) or excess water in the tissues, leading to pulmonary or peripheral oedema (Clark, 2022). The pathophysiology underlying the development of HF in the context of T2D is complex (Anker et al., 2023) but will involve:

- High insulin levels cause the kidney to retain water and salt (congestion), although this may be counteracted by high levels of glucose in the urine, which increases urine production and salt excretion. Indeed, a recent treatment for diabetes that increases glucose excretion in the urine appears to be a highly effective treatment for heart failure (Zannad et al., 2020).
- Persistently increased glucose levels may cause widespread vascular damage.
    - Increased arterial stiffness will increase the load on the heart
    - damage to the small vessels of the kidney may cause protein leakage and water and salt retention leading to hypertension and (initially subclinical) congestion.
- Hypertension puts an additional load on the heart and can damage the kidney.
- Atherosclerosis can choke the blood supply to the heart and kidneys and may lead to a heart attack, damaging the heart muscle and reducing the heart's ability to pump blood.
- Inflammatory pathways may be activated for multiple reasons (Libby et al., 2002).

- Classification of HF is based on which side of the heart is affected: left-sided and right-sided HF, shown in **Figure 3**.

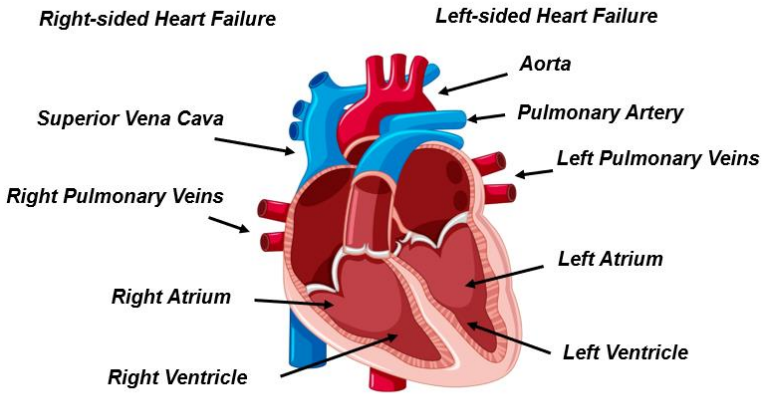


Figure 3 Anatomy of the Heart

Left-sided heart failure is the most common form of heart failure: the left ventricle is the heart's main pumping chamber. HF can be further classified as with preserved ejection fraction (HFpEF) or reduced ejection fraction (HFrEF). Ejection fraction (EF) is a key measure of heart function, representing the percentage of blood the left ventricle pumps out with each heartbeat (Savarese et al., 2022). In HFpEF, the left ventricle maintains a normal ejection fraction but cannot relax properly during diastole. This impairs the heart's ability to fill with blood between beats, causing a backup of blood into the pulmonary veins (pulmonary congestion), leading to symptoms such as shortness of breath, only during exercise if congestion is mild but even at rest when it becomes severe (Ponikowski et al., 2016). HFrEF occurs when the left ventricle cannot contract effectively, which also causes pulmonary congestion.

The most common cause of right-sided heart failure is left-sided failure. The right ventricle is responsible for pumping blood to the lungs for oxygenation. When the right ventricle fails, blood backs up into the systemic veins (systemic venous congestion), leading to peripheral oedema and enlargement of the liver (Clark, 2022; McDonagh et al., 2023). Cardiac dysfunction and congestion usually progress silently over many months or years until some triggering event (such as the onset of atrial fibrillation, an infection, a heart attack or a large meal full of salt) causes the symptoms and signs of heart failure to become apparent.

Even then, the diagnosis is often missed or the patient is put on diuretic treatment because of swollen ankles without considering the reason (Lawson et al., 2021). Loop diuretics are supportive for manging congestion and fluid overload. However, their use can sometimes mask an underlying heart failure diagnosis, delaying appropriate disease-modifying treatment. While effective for symptom relief, they do not directly improve cardiovascular outcomes or alter disease progression (Cuthbert et al., 2024).

1.1.5 General Treatment and Management

Effective management of CVD involves a combination of lifestyle modifications, pharmacotherapy and constant monitoring and audit of progress towards targets as recommended by clinical practice guidelines. Here we discuss key components of cardiovascular treatment and management.

Clinical guidelines are established to assist clinicians in decision-making for the appropriate care and treatment for patients. The guidelines are evidence-based, using systematic reviews, clinical trials and other medical literature. The World Health Organisation (WHO) promotes global awareness of cardiovascular disease (Hill-Briggs et al., 2021). WHO guidelines focus on population-level interventions, such as promoting healthy diets, physical activity and tobacco cessation. Countries have standard guidelines followed by clinicians. The European Society of Cardiology (ESC) updates clinical practice guidelines regularly primarily aimed at cardiologists but also now including patients on its committees (Foundation, 2020). The British Heart Foundation (BHF) is the biggest funder of research into heart and circulatory diseases in Europe. The National Institute for Health and Care Excellence (NICE) in the UK produces evidence-based guidelines for the management of both CVD and diabetes. These guidelines were formally encouraged and developed in the 1990s with committees comprising health professionals and patients (Rawlins, 1999; Ryan et al., 1996). However, clinical guidelines usually reflect average treatment effects rather than the specific needs of an individual patient. Guidelines are typically developed from the perspective of a single disease area. This approach does not sufficiently address the complex needs of patients with multiple chronic conditions. The approach may overlook specific patient factors such as genetic predispositions, lifestyle factors and individual responses to treatments i.e. different ethnicities.

As a result, there is growing interest in using ML to support personalised decision-making, by identifying high-risk subgroups or predicting individual treatment responses across diverse populations.

Treatments are a cornerstone of both primary and secondary prevention of CVD in patients with diabetes. Primary prevention, in addition to treatments aiming to improve glucose control, often includes the use of antihypertensive agents, statins, and antiplatelet therapy to manage CVD risk factors.

Secondary prevention may involve more intensive therapies such as beta-blockers, ACE inhibitors, and mineralocorticoid receptor agonists as well as intensification of glucose control with agents such as SGLT2 inhibitors and GLP-1 receptor agonists, which have been shown to reduce cardiovascular events in diabetic patients (Zannad et al., 2020). ML methods can enhance this treatment framework by helping to stratify patients based on predicted treatment benefit or likelihood of adverse outcomes, thereby refining therapeutic decisions beyond the "one-size-fits-all" recommendations in current clinical guidelines.

Despite the valuable guidance provided by clinical guidelines, their population-based recommendations highlight the need for more individualised approaches. ML-driven models hold promise in bridging this gap, enabling data-driven, patient-specific care strategies in both prevention and treatment pathways.

The management of CVD is mostly initiated when a patient experiences an adverse cardiovascular event or during hospitalisation. 80% of heart failure diagnoses are made in hospital, despite 40% of patients showing symptoms that warrant earlier assessment. Patients diagnosed with T2DM should also be assessed for CVD as most will have other risk factors such as hypertension and dyslipidaemia (Moran et al., 2022).

Laboratory tests are important for identifying some risk factors and to monitor if treatment is working. For patients with diabetes, HbA1c levels are monitored to assess long-term glycaemic control. Lipid profiles are measured to identify and manage dyslipidaemia. Kidney function tests, including serum creatinine and estimated glomerular filtration rate (eGFR), are crucial in

detecting and monitoring the progression of diabetic nephropathy. Inflammatory markers such as C-reactive protein (CRP) are also associated with an increased risk of cardiovascular events.

1.1.6 Advances in Cardiovascular Disease Prevention

Over the last decade, risk scores have been developed, such as Quantified Risk (QRISK3), Systematic Coronary Risk Evaluation (SCORE2) (Collaboration et al., 2021) and Framingham Risk Score (FRS). Identification of high risk may trigger early intervention to delay or prevent the onset of overt clinical disease.

QRISK3 is a predictive algorithm developed by clinicians and academics in the UK in order to estimate the 10-year risk of developing CVD. It considers various risk factors, including age, sex, ethnicity, smoking status, diabetes, family history of CVD, blood pressure, cholesterol levels, body mass index (BMI), deprivation and comorbidities. QRISK3 is frequently updated and might be widely used in primary care settings. However, it requires manual input of data by clinicians. Ethnic groups other than White are not well-represented.

The accuracy of QRISK3 relies on the quality and completeness of data entered. Incomplete or inaccurate data will lead to incorrect risk estimates.

SCORE2 is a CVD risk prediction algorithm developed by the European Society of Cardiology. It estimates the 10-year risk of both fatal and non-fatal cardiovascular events in European populations. SCORE2 improved on the original SCORE model by including a wider range of age groups and adjusting for contemporary risk profiles. Key factors included age, sex, smoking status, systolic blood pressure, total cholesterol, HDL cholesterol. However, SCORE2 is of limited use outside of Europe. Some key factors such as CKD are not included. The FRS estimates the risk of cardiovascular events but is based on data from a predominantly white American population, which may limit its applicability.

These risk prediction algorithms are based on survival analyses, which are used to predict the time until an event occurs. Many studies utilise the Cox proportional Hazards (CPH) model to calculate probability of risk. CPH is essentially a regression model which investigates the survival time of patients. The CPH model struggles with high-dimensional data where the number of covariates is large relative to the number of events(Jiang et al., 2024). Recently,

machine learning (ML) techniques have overcome many limitations of the standard Cox proportional hazards model by handling, high-dimensional data, non-linear relationships, time-varying covariates and complex patient data more effectively. ML also provides higher predictive accuracy and can be made interpretable through advanced methods.

ML-based survival models, such as random survival forests, gradient boosting survival models, and deep learning-based methods (e.g., DeepSurv), have shown promise in identifying patients at high risk of cardiovascular outcomes using routinely collected EMR data. These models can output longitudinal patterns in laboratory values, medication histories and prior comorbidities, providing a more dynamic and personalised risk estimation than traditional tools.

For patients with T2DM, EMRs provide an especially rich source of information. Predictive ML models have been used to identify patients at risk of developing heart failure, chronic kidney disease progression, or major adverse cardiovascular events. EMR-based models allow for the inclusion of real-world, time-updated clinical parameters such as HbA1c trajectories, blood pressure variability, renal function changes, and medication adjustments—features that traditional static models often overlook.

Moreover, the integration of these ML tools into clinical decision support systems is increasingly feasible through EMR platforms. By offering real-time, patient-specific risk estimates at the point of care, these systems can inform preventive strategies, enhance guideline adherence, and support shared decision-making between clinicians and patients.

However, the interpretation and application of these insights require the contextual knowledge and judgment of medical experts. This is known as collaborative intelligence, which represents the integration between computational decision-making and clinical expertise, harnessing the strengths of both to improve healthcare outcomes. For CVD prevention and management, this approach integrates sophisticated algorithms, such as machine learning models, with the understanding and experience of healthcare professionals.

Conclusion

Cardiovascular disease is an important global health challenge but diagnosis, even of some of its most severe forms, is often missed until it is too late. Few patients have a single uncomplicated cardiovascular condition. Problems, including T2DM, commonly conspire to drive the development and progression of CVD. Despite advances in medical research, there is a paucity of individualised risk assessment tools for CVD that have been shown to work for ethnically diverse populations across continents. New approaches to modelling applied to large epidemiologically representative datasets from very different cultures and geographies might identify readily available data that produces generalisable results.

1.2 Problem Statement

CVD is a leading cause of morbidity and mortality, globally. Despite advances in clinical guidelines and treatment, there are important gaps in early detection, individualised care and management, particularly for patients with multimorbidity, those who are socioeconomically deprived or, in a European context, those from ethnic minorities. This research seeks to address the following key issues:

Early Detection and Diagnosis: The variability in symptoms and signs of CVD, coupled with the subjective nature of symptom reporting, leads to delays in diagnosis and treatment. This is particularly true for atypical presentations, which are often overlooked by both patients and clinicians.

Management of Multimorbidity: The coexistence of multiple chronic conditions, particularly T2DM and heart failure, complicates clinical management. Current guidelines often focus on single diseases and may not adequately address the complex needs of patients with multimorbidity.

Socioeconomic and Ethnic Disparities: There may be substantial disparities in CVD outcomes across different socioeconomic and ethnic groups, which current risk prediction tools and clinical guidelines may not adequately reflect..

Standardised vs. Individualised Care: Clinical guidelines are primarily based on population averages and may not account for individual patient factors such as genetics, lifestyle and specific demographic markers. This can lead to suboptimal treatment strategies for diverse patient populations.

1.3 Project Aim

The aim of this research project is to assess cardiovascular morbidity and mortality in T2DM populations by applying novel artificial intelligence methods for early risk detection.

This research will help clinicians to identify which patients are at greater risk of incident heart failure and/or death. Highlighting such risks to clinicians and patients might improve appropriate investigation and treatment, thereby improving patients' wellbeing and prognosis, particularly amongst patients with the greatest socioeconomic disadvantage.

1.4 Research Objectives

The following objectives of expected achievement from the research project are:

1. Investigate the use of electronic medical records in T2DM populations and extract clinically relevant patient characteristics. Develop an appropriate high-level EMRs modelling process plan (**Chapter 3.4 Data Preparation**).
2. Apply statistical analysis to confirm clinical variable associations through correlations, survival analysis methods and multivariable analyses.
3. Leverage advanced machine learning methods for risk prediction and artificial intelligence interpretation.
4. Develop a support system for clinicians and patients to improve understanding of the relative importance of risk factors and their evolution over time, overall and for an individual patient.
5. Perform external validation in two distinct T2DM populations, accounting for ethnicity and diverse patient profiles.

1.5 Thesis Structure

This thesis is structured to provide an introduction and background of the problem, a review of the current literature, description of available datasets, before showing analyses and then a final discussion and conclusion. The outline of each chapter follows:

- Chapter 1 gives a general introduction to CVD, clinical presentation and pathophysiology, multimorbidity with the focus on T2DM and heart failure. It then summarises the general treatment, management of cardiovascular disease and advances in CVD prevention.
- Chapter 2 provides a brief literature review of T2DM as a CVD risk factor, the need for risk stratification, an overview of current risk prediction models, limitations, emerging approaches and inclusion of at-risk populations i.e. socioeconomically deprived people.
- Chapter 3 illustrates the EMRs from two diverse populations, it also explains how the datasets were extracted and prepared for analysis (Objective 1).
- Chapter 4 is the first of analysis chapters and predicts incident heart failure in the first T2DM population: Glasgow, West of Scotland. Survival analysis machine-learning is introduced and interpretation (Objectives 2 and 3) .
- Chapter 5 carries out external validation for incident heart failure risk prediction using the second T2DM population: Hong Kong. Casual inference is applied and the interface of the support system for clinicians is presented addressing Objective 4.
- Chapter 6 investigates socioeconomic groups in Glasgow, West of Scotland to predict mortality, highlighting the need for risk stratification.
- Chapter 7 analyses the two T2DM populations, focusing on the treatment of loop diuretics strongly linked to prognosis.
- Chapter 8 summarises key findings and provides a discussion of results, strengths, limitations, future work and conclusions.

Chapter 2 Literature Review

This chapter examines the development and application of risk prediction models aimed at identifying T2DM patients at elevated cardiovascular risk. Additionally, the review addresses the importance of including factors such as social deprivation, multimorbidity and population diversity in model design. These considerations highlight the need for tailored prediction tools that better reflect the complex, multifactorial nature of cardiovascular risk in T2DM patients and support precision medicine approaches in clinical practice.

2.1 Taxonomy of Literature Review

Taxonomy in **Figure 4** presents the structure of the areas explores in the literature review.

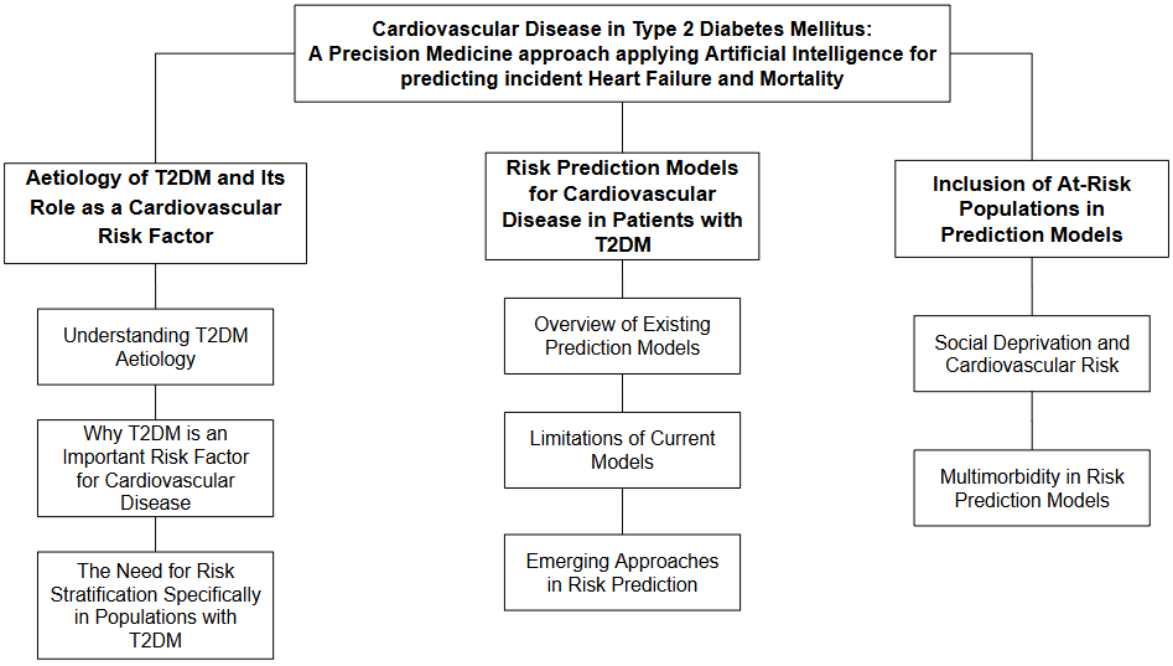


Figure 4 Taxonomy of Literature Review

2.2 Aetiology of T2DM and Its Role as a Cardiovascular Risk Factor

2.2.1 Understanding T2DM Aetiology

Insulin is a hormone secreted by islet-cells in the pancreatic gland. Insulin is a key regulator of cell uptake of glucose and consequently blood sugar levels. Type 2 diabetes mellitus (T2DM) is caused by resistance of cells (especially muscle and fat) to the effects of insulin, Typically, insulin levels are raised but the increase in insulin is insufficient to compensate for the insulin resistance leading to intermittent or persistent hyperglycaemia (high levels of blood sugar). Eventually, the islet-cells may become exhausted, leading to low insulin levels, more often encountered in Type 1 diabetes mellitus. T2DM is due to a combination of genetic, environmental and lifestyle factors but strongly associated with obesity and ageing (Kolb and Martin, 2017).

Genetics play an important role in the development of T2DM. Some populations, especially those of South Asian, African, and Hispanic descent, have a higher risk of developing T2DM (Suzuki et al., 2024), but often face barriers to accessing timely and effective care, resulting in high rates of poorly managed diabetes. Environmental and lifestyle factors also contribute to the development of T2DM. Sedentary behaviour, stress, poor diet and obesity are important risk factors that can lead to insulin resistance and impaired glucose metabolism (Zheng et al., 2018). Socioeconomic deprivation is also often associated with poorer lifestyle choices and access to healthcare (Kyrou et al., 2020), complicating disease management and prevention strategies (Gonzalez et al., 2018). Patients with T2DM face an increased risk of adverse cardiovascular events, highlighting the need for multifactorial interventions that are beyond blood glucose management. Addressing these factors through public health and community interventions is essential for reducing the morbidity and mortality burden of T2DM.

2.2.2 Why T2DM is an Important Risk Factor for Cardiovascular Disease

Individuals with T2DM have a higher risk of developing cardiovascular disease (Ahmad et al., 2024) not only because of the deleterious effects of persistently elevated blood glucose and insulin and insulin resistance, but also due to associated central obesity, high blood pressure and increased cholesterol and triglyceride levels.

These metabolic disorders contribute to endothelial dysfunction (Tziomalos et al., 2010) (narrowing of arteries), inflammation and atherosclerosis (Libby et al., 2002) (buildup of fats, cholesterol and other substances) which exacerbating CVD progression. Over time, these processes accelerate the narrowing and hardening of arteries, which increases the likelihood of coronary artery disease, myocardial infarction, stroke, peripheral artery disease and heart failure.

The United Kingdom Prospective Diabetes Study (UKPDS) and the ADVANCE (Action in Diabetes and Vascular Disease) trials demonstrated that T2DM patients are at a greater risk of cardiovascular events compared to non-diabetic individuals (Chalmers, 2005). T2DM and hypertension are also associated with a decline in kidney function (diabetic nephropathy), which is closely related to the development of cardiovascular complications, which may increase the risk of heart failure (Marx et al., 2023). This interrelationship highlights the importance of early intervention and management in patients with T2DM to reduce the risks associated with cardiovascular disease. Even in patients with well-controlled blood glucose, residual cardiovascular risk remains due to metabolic disturbances (Guan et al., 2024). Blood glucose control alone may be insufficient to prevent the complications of T2DM, emphasising the importance of a multi-faceted approach that also targets blood pressure, lipid levels, and lifestyle factors to reduce cardiovascular risk. Managing T2DM effectively requires careful monitoring and control of risk factors.

2.2.3 The Need for Risk Stratification Specifically in Populations with T2DM

General risk prediction models, such as the SCORE2 model (collaboration et al., 2021), were developed to estimate 10-year CVD risk across European populations. However, this does not account for the unique risk profile of individuals with T2DM, who often have a higher burden of cardiovascular and metabolic risk factors. Effective risk stratification allows healthcare providers to identify people most vulnerable to adverse events and implement tailored, evidence-based interventions that address their unique risk profiles. This is important in diverse populations, where genetic, lifestyle and socioeconomic factors further modify the risk profile. Harnessing information from multiple risk factors may enhance an individual person, precision approach to cardiovascular prevention in T2DM. However, including multiple clinical characteristics creates complex and high-dimensional datasets.

Applying machine learning to EMRs may identify patterns and relationships that traditional analyses may overlook, enhancing risk stratification. This might enable timely, individualised interventions that better align with each patient's unique profile, rather than managing average risk for all patients.

2.3 Risk Prediction Models in T2DM Patients

2.3.1 Overview of Existing Prediction Models

There are a variety of risk prediction models for patients with T2DM with the focus on heart failure or all-cause mortality as outcomes. **Table 3** shows models based on landmark randomised clinical trials. Typically, models based on clinical trial data rely upon traditional regression modelling and multivariable analysis. However, only patients who fulfil the inclusion/exclusion criteria for the trial and who are approached by investigators and are willing to give consent are included in such analyses. These highly selected patients may not be representative of the general population with T2DM. Moreover, only information of interest to investigators is collected, which will rarely include social deprivation scores.

The RECODE study (Basu et al., 2017) used the Action to Control Cardiovascular Risk in Diabetes (ACCORD) multinational trial dataset from United States and Canada to predict all-cause and CVD mortality for patients with T2DM, with moderate to strong prediction performance. Following this, Italian investigators (Copetti et al., 2021) also used the ACCORD trial data to create a risk score called ENFORCE, but the generalisability of the model is limited because few participants were from ethnicities other than White. The ACCORD dataset was also used to predict incident of heart failure in the WATCH-DM analysis (Segar et al., 2019), which improved on the standard cox proportional hazards method, by applying machine-learning adapted random survival forest for a risk score model, which resulted in better discrimination and calibration. However, these models are all limited by the fact that they are based on clinical trial datasets.

Table 4 shows risk prediction models for all-cause mortality and heart failure in patients with T2DM using data from EMR from the UK, New Zealand, Hong Kong and Singapore.

The UK-based QResearch and CPRD study focussed specifically on predicting incident heart failure in patients with T2DM (Hippisley-Cox and Coupland, 2015), using separate equations for men and women over a period of 10-years. This analysis uniquely includes social deprivation, family history and HDL/cholesterol ratio as predictors. Despite high-quality calibration (R^2 values of 41.2% in women, 38.7% in men) and ROC statistics close to 0.78, this study has limited international generalisability because it cannot account for ethnic or geographical diversity outside the UK.

The PREDICT-1° Diabetes study from New Zealand focuses on cardiovascular disease (CVD) risk prediction in T2DM patients over a five-year period (Pylypchuk et al., 2021). This study's moderate C-index suggests there is room for considerable improvement, possibly through the inclusion of additional biomarkers (Wells et al., 2017). The model based on the New Zealand dataset may also have limited applicability to other regions or populations, especially those with different healthcare systems or lifestyle factors affecting CVD risk. The absence of ethnic diversity in the UK study and the regional focus of the PREDICT-1° study suggests that the models may not be easily transferable to other countries with different demographic and clinical characteristics.

To address the limitations of patient diversity a study focusing on Asian populations: Hong Kong and Singapore targeted (Quan et al., 2019) similar outcomes (mortality, cerebrovascular disease, ischemic heart disease) over a five-year risk prediction period. With an improved C-index of 0.778 for mortality and lower values for CVD outcomes, this study emphasises the influence of demographic and clinical predictors across Asian populations. The large sample size provides robust findings. Additionally, another Hong Kong study on all-cause mortality in T2DM patients employed multiple predictive models (Lee et al., 2021), including Cox Proportional Hazards, random survival forests, and DeepSurv. The model's performance metrics (C-index of 0.75 for mortality and 0.79 for cardiovascular mortality) indicate good discriminatory power. The study provides advanced metrics such as the variability of HbA1c and fasting blood glucose (FBG), which are rarely considered in other models. However, neither of these analyses were validated amongst other ethnicities.

Future research could enhance these models by integrating broader socioeconomic, ethnic, racial and regional variability across multiple populations to improve generalisability. Additionally, integrating machine learning models alongside traditional regression approaches, as seen in the Hong Kong study, may offer improved predictive performance through nonlinear interactions.

Table 3 Analyses using Data from Landmark Randomised Trials for Risk Prediction in Patients with T2DM

Studies	**RECODe United Sates** (Basu et al., 2017)	**ENFORCE Italy** (Copetti et al., 2021)	**WATCH-DM United States** (Segar et al., 2019)	**ALTITUDE** (Malachias et al., 2020)
N=Population(s)	ACCORD (n=9,635), DPPOS (n=1,018), Look AHEAD (n=4,760)	Gargano Mortality Study (n=1,019), Foggia Mortality Study (n=1,045), ACCORD (n = 3,150)	ACCORD (n = 8,756 *ALLHAT n* = 12,063)	N= 5,509 (with complete data) from 36 countries and randomised the trial
Variables	Age, HbA1c, BMI, smoking status, duration of diabetes, eGFR	age, antihypertensive and insulin therapy, body mass index (BMI), diastolic blood pressure (DBP), low-density lipoprotein (LDL) cholesterol, triglyceride high-density lipoprotein cholesterol (HDL-C) and albumin/creatinine ratio (ACR) levels	Blood pressure, HbA1c, lipid levels, BMI, eGFR, comorbidities	Age, sex, diabetes status, albuminuria, eGFR, history of CVD, potassium, serum creatinine.
Methods	Cox Proportional Hazards	Cox Proportional Hazards	Random Forest Survival	Cox Proportional Hazards Model(s) with NT-proBNP and without.
Outcomes(s)	All-cause mortality CV mortality ESKD	All-cause mortality	Incident Heart Failure	All-cause mortality and CV composite (CC) outcome (CVD death, resuscitated cardiac arrest, nonfatal myocardial, infarction, stroke, or heart failure hospitalisation).
Metrics	C-index: 0.75 0.79 0.73	survival C-index was 0.81 (95%CI: 0.72–0.89) and Validation Cohort: 0.78 (95%CI: 0.68–0.87)	(C-index 0.77 [95% CI 0.75–0.80] Validation ALLHAT :C-index 0.74 [95% CI 0.72–0.76])	NT-proBNP alone: Death: 0.745 CC outcome: 0.723, Base model: Death: 0.744, CC outcome: 0.731, Base model + NT-proBNP:

					Death: 0.779 CC outcome: 0.763

Table 4 Analyses using observational data from electronic medical records for Risk Prediction in Patients with T2DM

Studies	**QResearch / CPRD** (Hippisley-Cox and Coupland, 2015)	**PREDICT-1° Diabetes New Zealand** (Pylypchuk et al., 2021)	**Hong Kong / Singapore study** (Quan et al., 2019)	**Hong Kong study** (Lee et al., 2021)	**Swedish Cohort Study**(Sattar et al., 2023)
N=Population(s)	437,806 (Derivation cohort), 137,028 (QResearch validation cohort), 197,905 (CPRD validation cohort)	N=46,652	N=678,750 (Hong Kong) and N=386,425 (Singapore)	N= 273,678	Individuals with T2DM: 679,072 from the Swedish National Diabetes Register. Matched Controls: 2,643,800 individuals without diabetes
Variables	Age, BMI, systolic blood pressure, cholesterol/HDL ratio, HbA1c, material deprivation, ethnicity, smoking, diabetes duration, type of diabetes, atrial fibrillation, cardiovascular disease, chronic renal disease, family history of premature coronary heart disease	18 predictors, including diabetes-related measures (e.g., renal function), demographics, medications	age, duration of diabetes, gender, smoking status, body mass index, systolic and diastolic blood pressure, HbA1c, low-density lipoprotein-cholesterol, pre-existing conditions (atrial fibrillation and CKD)	Age, sex, baseline comorbidities, anaemia, mean values of neutrophil-to-lymphocyte ratio, high-density lipoprotein-cholesterol, total cholesterol, triglyceride, HbA1c and fasting blood glucose (FBG), measures of variability of both HbA1c and FBG.	Age, sex, Glycated haemoglobin (HbA1c), Systolic blood pressure, Estimated glomerular filtration rate (eGFR), Lipids & Body mass index (BMI) and other comorbidities
Methods	Cox proportional hazards models; separate equations for men and women	Cox proportional hazards, 5-year, risk prediction	Cox proportional hazards models, five-year risk prediction	Cox Proportional Hazards, random survival forests, DeepSurv	Cox proportional hazards
Outcomes(s)	Incident heart failure diagnosis	CVD risk	Mortality, cerebrovascular	All-cause mortality	Coronary artery disease, Acute myocardial

			disease, ischemic heart disease among Chinese people with T2DM		infarction, Cerebrovascular disease and Heart failure
Metrics	Calibration: R^2 (41.2% in women, 38.7% in men); D statistic (1.71 in women, 1.63 in men); ROC statistic (0.78 in women, 0.77 in men)	C-index = 0.73,	C-index for mortality: 0.778, cerebrovascular disease: 0.695, ischemic heart disease: 0.644	C-index: 0.75 for mortality, 0.79 for CV mortality, 0.73 for ESKD	Hazard Ratios (HRs): HbA1c: Most important for atherosclerotic events. BMI: Explained >30% of HF risk. HF risk in T2D (all risk factors at target): HR = 1.50 (95% CI: 1.35–1.67)

2.3.2 Limitations of Current Models

Most prediction models for patients with T2DM bring challenges in generalisability and interpretability, restricting their clinical adoption, especially when attempting to apply them to individual patients.

Most studies, including those reviewed, use datasets specific to certain regions, ethnicities, or healthcare systems, which restricts model performance (Boyd et al., 2023). Risk models should be validated in multiple, diverse cohorts and regions to ensure they maintain accuracy and reliability across different populations. Recalibrating models in new cohorts and training them on multinational datasets should improve generalisability. However, without external validation, models are only reliable in patient-groups closely resembling their original dataset (Steyerberg et al., 2018).

Another important limitation is in interpreting risk prediction model results. CPH, provide limited interpretability. Patient data involves complex interactions between various risk factors. These limitations can lead to oversimplified models that do not accurately reflect the complexities of patient health. The cox model struggles with time-dependent predictors, residual confounding and interactions between variables, which limit its accuracy in capturing complex relationships in medical research. The CPH assumes the relationship between the predictors and the outcome (hazard) is constant over time. If this assumption is violated (when the effect of a variable changes over time) the model's results may be invalid or misleading.

However, interpretability also becomes more challenging when advanced machine learning methods are applied. When these models capture complex, nonlinear interactions, they lack clear mechanisms to explain how each predictor contributes to risk.

Moreover, interpretability is critical for clinical application, as it helps healthcare providers understand why a patient is at risk and what factors are modifiable (Holzinger et al., 2019). Clinicians are more likely to use models if they can understand the factors driving predictions (Tonekaboni et al., 2019).

Enhancing interpretability of AI models is crucial to improving trust and transparency in healthcare applications. Overall, addressing the limitations of generalisability and interpretability requires the use of diverse datasets, diligent external validation and applying explainable AI methods for model risk prediction.

The application of machine learning to CVD and T2DM risk prediction has grown considerably over the past decade. While literature focused specifically on incident CVD in T2DM populations is limited, many studies have successfully applied ML models to predict outcomes in either condition separately or in populations with overlapping risk factors. Traditional models like CPH remain the benchmark in clinical settings due to their interpretability; however, their assumptions of proportional hazards and linearity may not hold in complex, multimorbid populations.

RSF, an ensemble tree-based extension of the random forest algorithm for time-to-event data, have shown robustness in handling high-dimensional EMR data, non-linear relationships and interactions among variables without requiring variable selection. Other ML approaches such as Gradient Boosting Survival Models (e.g., CoxBoost or XGBoost with survival adaptations) and neural network ensembles have also been explored. Many of these algorithms offer higher discriminative performance compared to traditional models, particularly in large EMR datasets.

In choosing the appropriate ML method for this thesis, models were evaluated for their ability to handle high-dimensional EMR data, accommodate censoring and provide interpretable outputs. RSF was prioritised due to their established use in health informatics literature and their capacity to capture complex risk patterns in multimorbid populations. The tree-based approach is able to capture full patient overview and output the most contributing risk factors.

2.3.3 Emerging Approaches in Risk Prediction

Recently acknowledged interpretable artificial intelligence techniques, such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016) provide insights into individual patient predictions. SHAP explains model predictions by calculating the contribution of each feature to the final prediction using Shapley values from cooperative game theory. LIME generates interpretable explanations for individual predictions by building a simplified, interpretable model around the local area of the prediction, showing how each feature influences that specific prediction. These techniques introduce individualised risk prediction including the reasoning behind contributing risk factors. However, such interpretability methods are not yet routinely applied in survival analysis models commonly used in clinical studies, although they hold potential for enhancing transparency and trust in predictive models.

Currently there are very few risk calculators available for patients with T2DM. Those available require manual input of clinical data and lack supporting tools for interpretation. Most risk calculators are focused on CVD (Kengne, 2013; Committee, 2022). For example, the Diabetes Lifetime Perspective model (DIAL2) (Østergaard et al., 2023), was developed to provide individualised CVD risk for patients with T2DM.

It estimates life-years gained without CVD events based on competing-risk Cox models and validated data from multiple large T2DM European (only) cohorts. However, relying on manual input for such calculations is not practical, as it introduces human error and is inefficient, limiting the usability and scalability of these tools in real-world clinical settings.

There is a growing interest in applying casual inference techniques to clinical studies, aiming to move beyond associations. These methods, such as propensity score matching, purpose to establish causal relationships, interpret response to treatment and reducing confounding in observational datasets. For example, causal models allow for the identification of patient-specific factors that may influence the response to treatment, enabling more tailored and effective care (Lundberg et al., 2017). However, despite their popularity in academic research, these techniques have yet to be widely adopted in real-world clinical practice.

Accurate causal inference requires high-quality data, which may not always be available in real-world clinical settings (Hernán and Robins, 2016). Causal models can be powerful in controlled studies, however their applicability to diverse, real-world populations is sometimes uncertain. However, unless used with great care, they have the potential to introduce systematic bias.

For instance, low blood pressure is a powerful determinant of prognosis for patients with heart failure. A treatment called digoxin increases blood pressure. If people taking digoxin are matched to people not taking digoxin and matching includes blood pressure, then sicker patients taking digoxin will be matched to patients with a more favourable prognosis (higher intrinsic blood pressure) and analysis of outcomes may suggest that digoxin is harmful. Matching of patients before they received digoxin would be required to avoid this problem. Therefore, it is important to assess the applicability of causality in studies.

2.4 Inclusion of At-Risk Populations in Prediction Models

2.4.1 Social Deprivation and Cardiovascular Risk

Social deprivation has been linked to diseases and increased mortality rates (Wright et al., 2019). Physical, mental and social health are all important aspects of loss of well-being and disease (Chandola and Conibere, 2015). Lower income levels are associated with poorer access to education and educational attainment, poorer diet (Rosengren et al., 2019), higher rates of smoking and alcohol consumption, more sedentary behaviour, poorer access to quality healthcare and less successful self-care. This leads in turn to poorer health and increases in morbidity and mortality. However, health risk factors related to socioeconomic status may be difficult to obtain or inaccurate (pack-years of smoking, alcohol consumption). Inclusion of social deprivation may compensate for the lack of or inaccuracy of data from other sources as well as being an important risk factor in its own right, helping to ensure that all aspects of an individual's health—both clinical and social—are considered in the model.

However, factors related to socioeconomic status may at times be difficult to obtain due to limited data access associated with patient characteristics. Inclusion of social deprivation and its external factors is essential for predictive models in healthcare.

Understanding whether an individual is socially deprived, alongside clinical factors, provides a more defined view of their health risks and can improve the accuracy of prediction models for at-risk populations. These models can help ensure that all aspects of an individual's health are considered in the decision-making process.

A large-scale study (Deepali Nagar et al., 2021) in England and Wales using over 3.7 million EMRs developed the cox model with the inclusion of ethnicity and socioeconomic status. The study illustrated a major disparity in diabetes risk across ethnic groups, with Bangladeshi men and women showing the highest adjusted hazard ratios compared to white populations. The inclusion of social deprivation allowed the algorithm to capture the compounding effects of socioeconomic disadvantage on diabetes risk. Ethnicity is also a key inclusion for risk prediction models. Ethnicity is known to improve risk classification in CVD disparities (van Apeldoorn et al., 2024a). One study highlights the predictive capacity of five ML models for CVD events and ethnicity groups in a cohort of 145,600 diabetes patients in New Zealand (Nghiem et al., 2024).

The Gradient Boosting decision tree model performed the best for predicting CVD. Key predictors varied by model and ethnic group, with factors like age, area deprivation and prior hospitalisations being important across groups. Moreover, this inclusion creates fairness in risk assessment for patients with T2DM. Addressing health inequalities, ethnically and socioeconomically diverse populations is important.

Some research shows ML models, while effective in risk prediction, are bias according to the population data. For example, models trained on majority populations can underperform for minority groups, reinforcing systemic health disparities. This refers to Algorithmic bias: differences in the predictive power of models when applied to different subgroups of a population (Dang et al., 2024). Therefore, efforts to ensure algorithmic fairness requires the integration of socioeconomic status such as deprivation scores or stratifying models by demographic subgroups. Addressing fairness in algorithm design is not only an ethical imperative but also ensures better health outcomes for underserved populations.

2.4.2 Multimorbidity in Risk Prediction Models

Multimorbidity, the presence of multiple chronic conditions in a single patient, affects the accuracy and effectiveness of risk prediction models (Rahimi et al., 2018). Patients with multimorbidity often face complex interactions between diseases, making it challenging to predict their outcomes based on single-condition models. Including multiple chronic conditions into risk prediction models enhances their ability to reflect the true complexity of patient health, enabling more accurate assessments and personalised interventions. Precision in diabetes care considers all attributes to help understand individual patient profiles for CVD risk prediction. Traditional risk scores provide population-level insights but lack the ability to personalise risk predictions effectively. These models typically rely on limited number of clinical and demographic factors. This may not fully capture the complexity of contributing factors to disease progression in T2DM. ML has emerged as a transformative approach, with the integration of diverse and high-dimensional datasets such as EMRs. With the use of interpretable and adaptive ML techniques, identification of complex interactions and key predictors are useful for producing clinically relevant risk prediction algorithms. These may be further deployed to evidence-based support tools for T2DM populations considering various outcomes such as heart failure and mortality.

2.5 Conclusion

Several risk prediction models for people with T2DM based on multinational randomised trials exist, but these are of limited value for clinical practice because they include only patients selected and invited to participate and who agree to do so. Many of the studies were population-specific and there was no advanced interpretability used. Traditional cox proportional hazards models have provided a foundation for understanding population-level risks. However, they are constrained by their reliance of on linear assumptions and limited inclusion of complex variables. This research builds on these limitations by applying advanced ML models and interpretation to diverse populations. Implementing risk prediction models specifically for T2DM populations, overcoming these limitations is essential. Prediction models based on EMRs will reflect more closely the population served but the data may lack the structure, granularity or completeness found in trials and registries. Ultimately, a robust model should work similarly well when applied to multiple datasets, whether they be trials, registries or EMR.

Chapter 3 Descriptive Analytics of Type 2 Diabetes Across Two Populations Using Electronic Medical Records

3.1 Introduction

As data becomes increasingly ubiquitous in our digital world, it is now present in every aspect of our lives. Electronic medical records (EMRs) have become the foundation of modern healthcare practice and systems, providing a digital solution that ensures patient information is accessible, integrated and actionable across various settings. Acquisition of large amounts of patient information in EMRs, exponential growth in computing power and the application of machine learning has the potential to transform medical research and patient care. In healthcare, the growing volume and complexity of data has changed how medical information is recorded, stored, used and interpreted. Data are now obtained that tracks a patient's history, records the basis for diagnoses, monitors the response to treatment and records outcomes. Clinicians can identify adverse health trends, recognise modifiable risks for timely intervention and monitor early indicators of success or failure following treatment. This also encourages the rise of precision medicine, using electronic patient records (EPRs) to tailor medical care according to the unique characteristics of each patient. However, this wealth of healthcare data also presents challenges. Ensuring the accuracy, consistency, reliability and security of EPRs is important for many reasons, including ethical and legal. Logical processes and data validation are essential to confirm that conclusions drawn from such data are accurate and reliable. Therefore, standardised methodologies are crucial. The medical field has already begun to adopt standard practices for data handling, but these vary across institutions and regions. These standards ensure consistency and interoperability across different datasets, making large-scale analysis possible. This chapter outlines the process of obtaining and extracting large quantities of data from multiple sources in a reliable and ethical manner. By following established best practices and implementing a standardised methodology, this chapter also aims to contribute to the growing body of knowledge on how EMRs are leveraged for clinical research, including advances in precision medicine. This chapter provides a detailed exploration of the baseline characteristics and clinical profiles of individuals with T2DM using EMRs from two distinct populations: Scotland and Hong Kong. These insights are essential for tailoring predictive models and ensuring their robustness across diverse populations.

3.2 Electronic Medical Records (EMRs)

The development of EMRs represents a fundamental moment in the history of healthcare, a shift driven by both technological innovation and the need for greater efficiency in patient care. The first steps towards digital record-keeping in healthcare were taken in the 1960s, primarily in the United States, where early systems focused on automating hospital administration and billing processes. In the 1970s and 1980s, pioneering academic medical centres, such as those at Harvard (Barnett et al., 1979) and the Mayo Clinic (Ellsworth et al., 2016), began experimenting with computer-stored medical records (Mcdonald and Tierney, 1988), which set the basis for more advanced EMR systems. These early EMR prototypes were primarily used in hospital settings, with a focus on improving efficiency and reducing errors in patient management. By the 1990s there was an adoption of EMRs across the globe. Technological advances enabled greater interoperability between systems. By the end of the decade, countries begun implementing national strategies to promote the use of EMRs within their healthcare systems.

Scotland adopted a regionally focused approach by developing the Scottish Care Information (SCI) programme, which began in the 1990s. This initiative integrated patient records across hospitals, general practices and other healthcare settings, culminating in the creation of the Emergency Care Summary (ECS), ensuring that key patient information was available across Scotland's healthcare providers to improve continuity of care and patient safety. Some data are available nationally, including prescriptions, social deprivation, hospitalisations and deaths, but other data, such as blood and pathology tests, are only available at a regional level. Although regional data could be aggregated to provide national level data, there are many administrative barriers, and it has rarely been attempted or successful. Data are linked and made available in a trusted research environment (TRE), which is called a SafeHaven in Scotland. In Hong Kong, the Hospital Authority (HA) initiated the development of its Clinical Management System (CMS), also in the 1990s, which laid the foundation for a territory-wide electronic health record (EHR) infrastructure. These records are anonymised: patients are assigned anonymous reference keys, ensuring individual identities cannot be linked. Sensitive information such as names, identification number and addresses are removed or not included in the system. By the 2000s, Hong Kong had advanced its system to provide access to patient information across its public hospitals and clinics, contributing to better care coordination.

3.2.1 Using EMRs for Research and Public Health

Initially, EMRs were designed to improve the efficiency and accuracy of patient tracking, providing healthcare professionals with an overview of patient profiles. Nowadays, EMRs are crucial for public health and epidemiological research. They provide invaluable insights for prevention, treatment monitoring and long-term healthcare planning for populations. Observational studies are carried out to draw inferences from the study population which represents the general population. By capturing vast amounts of patient data over time, reflecting routine clinical care, researchers are able to conduct large-scale studies across diverse populations, improving the generalisability of findings.

3.2.2 Data Integrity in EMRs

Data integrity is a critical component when working with sensitive patient records, especially in healthcare research. Trust in the EMRs, both in terms of data collection and analysis, is essential to ensure that the findings are valid and reproducible. Managing EMRs requires patience, diligence and commitment to obtaining honest and reliable results. Without these elements, the conclusions drawn for clinical research are at risk of being flawed or misleading.

EMRs preprocessing greatly relies upon the logical processing. Each stage (data extraction, transformation and analysis) requires integrity and appropriate handling of patient data. Errors at any stage can lead to compromised results, making it vital to possess not only technical skills but also acknowledgement of clinical contexts. For example, reading clinical guidelines for diabetes diagnosis was important to collect the right data from EMRs. By maintaining these principles, the research delivers clinically relevant results while safeguarding patient rights.

It is also important to ensure data are representative and reflect the real-world context of the general patient populations. This included addressing issues such as missing data and conducting thorough data validation to certify consistency and completeness. Data integrity in this thesis required a simultaneous approach that combined technical expertise and ethical responsibility.

3.2.3 Data Privacy Acts & Regulations

The General Data Protection Regulation (GDPR) in the EU, the UK Data Protection Act 2018 and the Hong Kong Personal Data (Privacy) Ordinance (PDPO) enforce strict regulations on how personal data are collected, stored and shared. These laws are designed to protect patient confidentiality by enforcing safeguards like anonymisation and encryption to prevent unauthorised access to patient identifiable information.

3.3 Data Sources

3.3.1 Glasgow, West of Scotland

In Scotland, everyone has free access healthcare through the NHS, including free prescriptions. People receive a unique community health index (CHI) identifier linked to all healthcare contacts and deaths. This identifier remains constant across various sets of data that might be acquired about that individual. These are changed to study IDs for research purposes in the Safe Havens. Safe Havens are secure environments that have been broadly used to support access to EMRs for research while protecting patient identity and privacy. Safe Havens provide secure access to linked, de-identified EMRs (Lee et al., 2021). Deidentification transforms information that could identify an individual in a data set with a study identifier (ID) for that individual. However, unlike anonymisation, deidentification can be reversed back with secure access. This enables large-scale population-level studies. The research ID is the same across data sources, enabling data linkage within multiple data sets. With this advantage, it was important to acquire the necessary patient attributes. Selected datasets were examined, queried and visualised for deeper understanding. Due to the voluminous collection of the greater Glasgow & Clyde population data, several csv files were linked together to form a structured and useable dataset. **Table 5** presents the clinical data extracted from the available csv files which represent patient information. For the West of Scotland Safehaven in Greater Glasgow & Clyde (GG&C), access was granted by the local privacy advisory committee (project code GSH/20/CE/004) **(Appendix A1)**.

Table 5 Safe Haven Data Availability

Safe Haven Data Sources	Description	Data Extracted	Limitations
Demographics	Basic patient information	Safe Haven ID, age, sex & Scottish multiple index deprivation score in 2016	Standard demographic details
Scottish Care Information-Diabetes Collaboration (SCI Diabetes)	National database for diabetes care and management	Safe Haven ID, Date of diabetes diagnosis, Description of diabetes, Diabetes Codes & Ethnicity	No specific limitations
Scottish Care Information Store (SCI Store)	Repository for laboratory results (Biochemistry and haematology) and clinical data from Scottish health records	Safe Haven ID, Glucose & HbA1c, Haematology - haemoglobin, white cell count, neutrophils, lymphocytes. Renal function - sodium, urea and creatinine, uric acid, eGFR. Liver function - bilirubin, AST, ALT, alkaline phosphatase. Lipids – total and HDL, VLDL, LDL cholesterol and triglycerides. Urine – microalbuminuria	Limited to biochemistry and haematology disciplines
General Practice Local Enhanced Services (GP LES)	Data on locally commissioned services provided by general practices (Primary Care)	Safe Haven ID, Body Mass Index, smoker, weight & hypertension	Height, Weight, BMI are limited. Systolic & Diastolic blood pressure values not available
Prescribing Information System (PIS)	Data on prescriptions issued to patients	Safe Haven ID, BNF Chapter Codes, BNF Drug codes, Dispatched Date & Prescribed Date	No specific limitations
SMR01 (Hospitalisations)	Hospital admissions, including diagnoses, operations, and procedures	Safe Haven ID, International Classification Disease codes, 10th edition.	Covers inpatient and day case records
Deaths	Information on patient mortality, including date and cause of death	Safe Haven ID, Date of Death & cause of death.	No specific limitations

3.3.2 Hong Kong, SAR China

The second data source for external validation was the Clinical Data Analysis and Reporting System (CDARS) from Hong Kong, SAR China. CDARS is a territory-wide EMRs system managed by the Hospital Authority of Hong Kong, which provides an integrated platform for capturing patient-level data across public healthcare institutions. Since 1995, the system makes clinical data available for research and audit purposes (Gao et al., 2021). An important feature of CDARS is its emphasis on patient privacy and data security. To protect patient identities, the system anonymises personal information by using a unique Reference Key for each patient. This Reference Key is generated from identifiable data but is securely encrypted to prevent reverse identification, ensuring that individual patient records remain anonymous during research or analysis. This encryption process allows researchers to work with patient data while adhering to stringent privacy regulations. CDARS links different types of clinical data, including inpatient and outpatient records, demographic information. laboratory results, diagnostic imaging, prescription records and death records. These data sources are unified in a single platform, allowing for the longitudinal tracking of patient histories across different healthcare services and episodes of care.

The system uses a common patient identifier to ensure that all records from multiple clinical encounters are consistently linked to the correct individual, this enables research into patient outcomes, disease progression and treatment efficacy.

Data are typically provided in CSV file format, facilitating the integration of information from multiple sources. CDARS has been widely used in previous research studies (Lee et al., 2021; Tse et al., 2024; J. Zhou, Lee, Lakhani, et al., 2022; J. Zhou, Lee, Liu, et al., 2022; Lee et al., 2022; Kwok et al., 2024) for conditions such as chronic obstructive pulmonary disease (COPD), diabetes, cardiovascular diseases and cancer. CDARS provides a powerful tool for large-scale healthcare research and facilitates comparisons with other EMR systems globally.

3.3.3 Routinely collected EMRs

Both Glasgow and Hong Kong's EMRs are routinely collected from healthcare systems, capturing real-time data across patient visits and hospitalisations, including laboratory tests and prescriptions. Unlike retrospective data, which looks back on past information, or clinical trial data, which include only patients who fit the inclusion/exclusion criteria and are willing to give informed consent, routinely collected EMRs reflect diverse healthcare trajectories. EMRs provide dynamic and diverse patient data that can be used to identify at-risk populations, including ethnic minorities, facilitate early detection of diseases. These EMRs provide broad, longitudinal insights into patient populations, including ethnic minorities. Routinely collected EMRs allow for early disease detection, population-wide health monitoring and scalability (increasing the scope and size of data collection).

3.4 Data Preparation

Data preparation is a critical component of any research involving data. This becomes a challenge when handling complex and extensive EMRs. Preprocessing raw EMRs into a format suitable for analysis can be challenging due to the vast amount of unstructured information. Large parts of medical records are in written text form and are tedious to use directly without appropriate data processing. It is important to adopt a structured and iterative approach to understanding the stages of data processing. In this research, a high-level methodology of EMRs modelling process illustrated in **Figure 5** is applied.

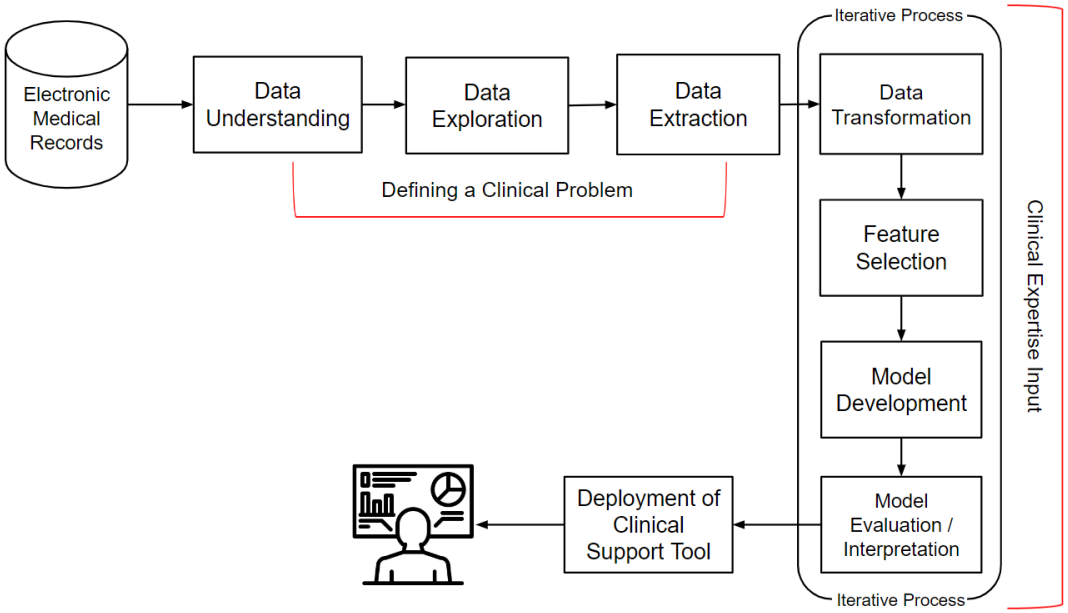


Figure 5 EMR Modelling Process

This methodology is derived from standard cross-industry process for data mining. Data mining involves methods at the intersection of database systems, statistics and machine learning and database systems (Asri et al., 2020). It is the process of transforming data to discover patterns and useful information in large datasets. Data understanding is key to computing clinically relevant results. The first stage of the process was to define the T2DM cohorts. It was important to select and filter the right clinical characteristics for this research. Data exploration was key to discovering the available patient information from multiple datasets in Safe Haven and CDARS. This was followed by extracting the necessary patient information. Data transformation consisted of data grouping, categorising and linking the required patient characteristics by a unique patient identifier.

This prepared the data for feature selection, statistical analysis, machine learning modelling and interpretation. Clinical expertise input was part of each iteration, introducing the integration of collaborative intelligence.

3.4.1 Statistical & Machine Learning Software

Statistical and machine learning analyses were performed using Python. This is a versatile and multipurpose programming language that is widely available and enables robust processing of complex data. Pythons Jupyter Notebook was used, because data are shown at each step as you carry out preprocessing, rather than creating large, automated scripts. This functionality is important when handling sensitive patient data because it allows transparent data preprocessing. **Table 6** presents the python libraries imported to conduct data preparation.

Table 6 Python Libraries

Python Libraries	Use	Version
Data Preprocessing (Cleaning, filtering, searching, selecting and grouping)		
Pandas	Data manipulation and analysis	2.0.3
NumPy	Numerical computing and array operations	1.25.0
SciPy	Scientific computing and technical computing functions	1.11.2
Data Visualisation (Descriptive statistics and results)		
Matplotlib	Visualising graphs and results	3.8.0
Seaborn	Statistical data visualisation based on Matplotlib	0.12.2
NetworkX		
Survival Analysis (Medical statistics and machine learning survival models)		
Lifelines	Survival analysis methods (e.g., Kaplan-Meier estimator, Cox proportional hazards)	0.28.0
scikit-survival	Advanced survival analysis, modeling (e.g., Random Survival Forest) and evaluation metrics (C-index score)	0.16.0
XGBoost	Gradient boosting framework for supervised learning tasks, including survival analysis	1.7.5
Machine Learning (General)		
scikit-learn	General machine learning library for classification, regression, and clustering	1.3.0
Interpretation (Explainability of model predictions and causal inference)		
SHAP	Explaining model predictions using Shapley values	0.41.0
EconML	Causal inference library for estimating treatment effects	0.11.1

3.5 Glasgow Clinical Definitions

3.5.1 Defining Diabetes Mellitus

Engagement with clinical experts defined clinical diagnoses and events. The Scottish Care Information: the SCI-Diabetes registry is a fully integrated shared EMR to support treatment of patients with diabetes in Scotland. Registration occurs automatically when a patient is assigned a diagnostic code "[10]" for diabetes in primary or secondary care. The registry is estimated to capture >99% of all patients with diabetes nationally. This registry was merged with the overall population demographics of those who had a record of diabetes, filtered using SafeHaven ID. Once merged, several patients had repeated rows, **Table 7** presents an example of this.

Table 7 Diabetes Diagnosis with repeated Rows

ID	ETHNICITY_DESC	DATE	ITEMVALUECODE	VALUEDESCRIPTION
11	White - Scottish	2019-12-23	GEN-NYN-01	Yes
11	White - Scottish	2014-05-12	DD-DMT-51B	Impaired Fasting Glucose
11	White - Scottish	2017-11-10	DD-DMT-02	Type 2 Diabetes Mellitus

Each patient in the registry had a record of "Yes", suggesting a diagnosis of diabetes. At a later date, the type of diabetes may have been recorded. In this case, the patients were noted to have an elevated fasting glucose in 2014 and a diagnosis of T2DM in 2017. However, text describing diagnosis were sometimes confusing or blank, reducing my confidence in their reliability. It was important to learn what the diagnostic codes represented as these were consistent. ITEMVALUECODE refers to a coding system for the diabetes registry. Further data transformation was carried to ensure each patient had the most relevant date of diagnosis and diabetes type.

Table 8 portrays the final percentage of patients in different diabetes description categories. There were several other diagnostic codes in numerical format, which did not indicate type of diabetes. These were excluded from the dataset as this analysis focussed only on T2DM, which is the most common type and strongly associated with older age and obesity. T2DM is due to resistance of tissues to the effects on insulin; these patients typically have insulin levels although eventually the pancreatic islet cells that produce insulin may become exhausted so that these patients may eventually require treatment with insulin rather than medicines to improve insulin resistance.

Type-1 diabetes mellitus is the classical but rarer form that often affects children and young adults and is due to failure of the islet cells to produce insulin. These patients are usually treated only with insulin.

Table 8 Percentage of patients in Diabetes Diagnosis Descriptions

Diabetes Description Categories in Glasgow Cohort	
Category	**Percentage of patients**
Type-2 Diabetes Mellitus	71%
Impaired Glucose Metabolism and Other Not Known	10%
Impaired Glucose Intolerance	5%
Impaired Fasting Glucose	4%
Not Known/Unknown	6%
Diabetes Type Not Defined	2%
Type-1 Diabetes Mellitus	2%

To reduce complexity, patients were grouped with the help of clinical experts into five categories using the diagnostic codes shown in **Figure 6**. An example of this is "impaired fasting glucose", "impaired glucose tolerance" and "impaired glucose metabolism and other not known" were defined as at "Risk of Type 2 Diabetes". **See Appendix A2: Diagnostic Descriptions**. As expected, most patients were classified as T2DM. T1DM was excluded, as advised by clinical experts, because it is an autoimmune condition leading to a collapse in insulin production rather than being driven by modifiable lifestyle factors.

```
Type 2             70.789479
Risk of Type 2     11.916023
Other Types         9.248285
Not Defined         6.281431
Type 1              1.764783
Name: Diabetes_type, dtype: float64
```

Figure 6 Diabetes Type Categories

3.5.2 Ethnicity

Ethnicity was also grouped in the SCI-diabetes dataset. Ethnicity data was mapped to six categories shown in **Table 9** (White, Asian, Chinese, Black, Mixed, Other) to standardise variations in reporting and enable clearer analysis. Most Glaswegians are of "White" ethnicity.

Table 9 Ethnicity Groups

SCI-Diabetes: Original Ethnicity Group	Ethnicity Category (Mapping)
White – Scottish, White – British, White – English, White – Polish, White – Welsh, White – Northern Irish, White – Irish, Any other white ethnic group	White
Pakistani, Pakistani Scottish or Pakistani British, Indian, Indian Scottish or Indian British, Bangladeshi, Bangladeshi Scottish or Bangladeshi British, Arab	Asian
Chinese, Chinese Scottish or Chinese British	Chinese
African, African Scottish or African British, Caribbean, Caribbean Scottish or Caribbean British, Black, Black Scottish or Black British	Black
Any mixed or multiple ethnic groups	Mixed
Other – Asian, Asian Scottish or Asian British, Other – Other ethnic group, Other – African, Caribbean or Black	Other

3.5.3 Scottish Index of Multiple Deprivation

The Scottish Index of Multiple Deprivation (SIMD) is a measure of deprivation among 6,976 data zones levels (The Scottish Government, 2017). SIMD observes the extent to which an area is deprived across seven domains: income, employment, education, health, access to services, crime and housing. If an area is identified as 'deprived', this may reflect low incomes but can also mean fewer resources or opportunities. Each patient is given a score 1-5 known as quintiles. This score is ranked from the most deprived to the most affluent: 1-5.

Figure 7 presents patients with diabetes in the five quintiles. 41% are in the highest deprivation quintile 1. Only 15% of those with T2DM were in quintile with least deprivation (quintile 5) in Glasgow.

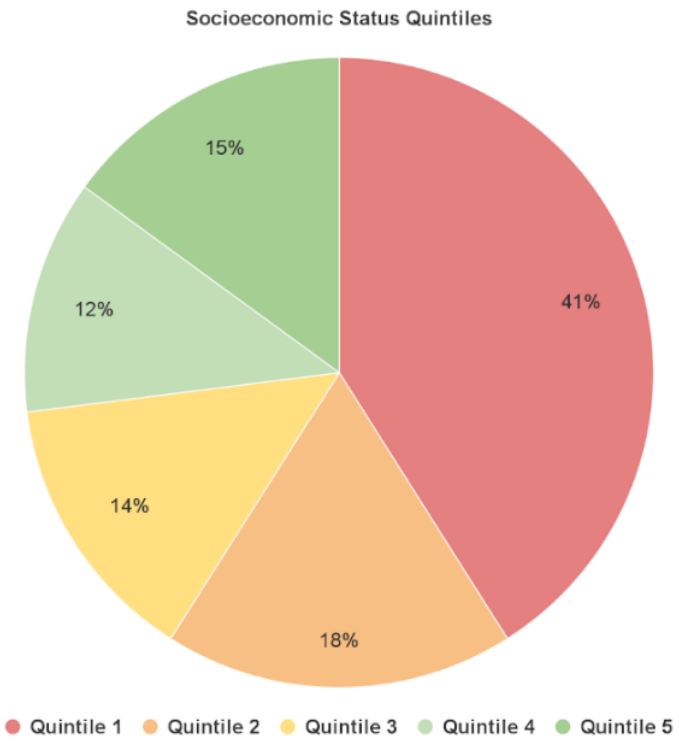


Figure 7 Socioeconomic Status Quintiles in Patients with Diabetes

3.5.4 General Practice Local Enhanced Services (Primary Care)

In the Glasgow population cohort, the following conditions in **Table 10** were extracted from the General practice (GP) – local enhanced services (LES) dataset. These GP practices provide additional care for specific conditions, including coronary heart disease, diabetes, stroke and COPD. Practices can subscribe to different LES services without covering all. For NHS Greater Glasgow & Clyde, 82.8% of practices contributed data, covering a broad range of services. Each GP LES entry includes the safehavenID, event date, a Read code for the service (READCODE), a description and flags indicating whether the entry relates to prescriptions (IsPrescription) or numerical values (IsValue). Additionally, values recorded for services are captured (Value1 and Value2). The coverage of GP LES records extends up to the end of 2018. Extracting these records was labour intensive.

For smoking (**Figure 8**), categories were condensed into "Current Smoker" and "Non-smoker". BMI was stored as calculated value or according to recommended categories as advised by the NHS in Scotland(NHS BMI, n.d.). There was also missing records of BMI for 30% for the Glasgow cohort. BMI record keeping is not important in primary care or secondary care settings unless it is for a certain condition. The Scottish diabetes group (Scottish Diabetes Group, 2021) reported that 1 in 3 people with diabetes in Scotland did not have a BMI recorded in 2019. The missingness of BMI was further investigated in this research. Weight was recorded in kilograms. High blood pressure was grouped from various descriptions presented in **Figure 9**. Essential hypertension means a high blood pressure with no known specific cause such as an endocrine tumour or renal artery stenosis.

Table 10 Extracted Primary Care Conditions

Primary Care Condition	READ Code	Description
Smoking Status	137..00	Current smoker, types of current smoker and non-smoker
Body Mass Index (BMI)	22K..00	Numerical values of BMI using NHS calculator
Weight	22A..00	Weight recorded in kilograms
Hypertension	G20..00, G20z	Essential Hypertension

```
non_smoker['Description'].value_counts()
```

```
Current non-smoker       1158
Current non-smoker 12       3
Current non-smoker 10       3
Name: Description, dtype: int64
```

```
smoker['Description'].value_counts()
```

```
Current smoker                            38050
Current smoker, 0 per day                   581
Current smoker, 10 per day                  371
Current smoker, 20 per day                  362
Current smoker, 15 per day                  178
                                           ...
Current smoker  15 per day                    1
Current smoker, 30 cigarettes per day         1
Very heavy smoker - 40+cigs/d                 1
Current smoker, 45 per day                    1
Current smoker# 18 per day                    1
Name: Description, Length: 95, dtype: int64
```

Figure 8 Smoking Descriptions

```
bp['Description'].value_counts()
```

```
Essential hypertension                    51721
Essential hypertension NOS                 4287
High blood pressure                         334
Hypertension NOS                            168
HYPERTENSION                                 94
Essential hypertension.                      50
hypertension                                 44
BLOOD PRESSURE BORDERLINE                    18
BLOOD PRESSURE RAISED                        10
HYPERTENSION ESSENTIAL                       10
BLOOD PRESSURE HIGH                           7
HIGH BLOOD PRESSURE                           6
Query Essential hypertension                  5
HYPERTENSION ON TREATMENT                     3
H/O: hypertension                             2
New Hypertensive                              2
Hypertension Quality Indicators v1.2          1
Hypertension Quality Indicators               1
Essential hypertension (01/07/2015)           1
Essential hypertension (12/11/2015)           1
Essential hypertension (07/12/2017)           1
Essential hypertension  (07/02/2014)          1
Essential hypertension  (02/05/2017)          1
Essential hypertension.noted                  1
Essential hypertension (08/12/2004)           1
HYPERTENSION LABILE INTERMITTENT              1
Name: Description, dtype: int64
```

Figure 9 High Blood Pressure Descriptions

3.5.5 Prescribing Information System (Medications)

The Prescribing Information System (PIS) in Scotland supports pharmacoepidemiology and pharmacovigilance by providing prescription data for >5.3 million NHS residents (Alvarez-Madrazo et al., 2016). PIS has tracked reimbursed prescriptions with detailed individual prescribing and dispensing records since 2009, linked to the Community Health Index (CHI) number described earlier.

British National Formulary (BNF) codes are used to classify medications. BNF is a pharmaceutical reference book with sections and chapters classifying medications. Each medicine prescribed is assigned a BNF code, which provides details about the medicine's classification, dosage and formulation. For this research, treatments for diabetes (BNF Chapter 6) and cardiovascular disease (BNF Chapter 2) were extracted (**Appendix A3**).

Each prescription record in the PIS dataset includes a prescribing date (PRESC_DATE) and a dispensing date (DISP_DATE). However, two important observations were made. Firstly, for some prescriptions, the same medication and patient were recorded with the same prescribing date but varying dispensing dates, suggesting repeat prescriptions, especially after 2013. Secondly, the dispensing date often falls on the last day of the month, likely reflecting when pharmacies were applied for reimbursement or were reimbursed rather than the actual date of dispensing. To manage this, PRESC_DATE was used for initial prescriptions and DISP_DATE was used to calculate spacing for repeat prescriptions. An example of the coding structure is presented in **Figure 10** for a glucose lowering treatment for diabetes called dapagliflozin. Prescriptions were classified based on their active chemicals. This classification helped with organising medicines by their active components rather than brand names or formulations, especially when some medicines are prescribed as combination tables (for instance angiotensin receptor blockers and thiazide diuretics in one tablet).

BNF Code for Dapagliflozin							
Chapter	Section	Paragrpah	Sub-paragraph	Chemical Substance	Product	Strength and formulation	Generic equivalent
06	**01**	**02**	**3**	**0**	**A**	**G**	*
Endocrine System	Drugs used in diabetes	Oral antidiabetic drugs	DPP-4 Inhibitors	Dapagliflizin	Dapagliflizin	Dapagliflozin 10mg	

Figure 10 BNF Code Breakdown

3.5.6 Scottish Care Information Store (Laboratory Tests)

The Scottish Care Information (SCI) Store is an electronic data repository used by NHS Scotland, designed to store and integrate patient data at a regional health board level. It accepts a wide variety of clinical reports, including biochemistry, hematology, pathology, microbiology and radiology results, as well as other laboratory test types. The SCI Store facilitates the sharing of laboratory and diagnostic information for healthcare providers within the region. **Table 11** shows the extracted laboratory tests with their unique READ Codes and their primary purpose in this research.

Table 11 Extracted Laboratory Tests

Laboratory Test Category	Marker	READ Code	Purpose/Indication
Renal Function	Serum Creatinine	44J3.	Assesses kidney filtration and function.
	eGFR (estimated Glomerular filtration rate)	451E.	Evaluates overall kidney function.
	U Albumin-to-creatinine ratio	46TC.	Early marker of kidney damage, especially in diabetes.
Inflammation	C-reactive Protein	44CS.	Indicates inflammation or infection.
Blood Chemistry	Haemoglobin	423..	Evaluates the oxygen-carrying capacity of red blood cells.
	Glucose	44g..	Measures current blood sugar levels; essential in diabetes diagnosis.
	Serum Albumin	44M4.	Marker of liver function and nutritional status.
	Potassium	44I4.	Electrolyte balance critical for heart and nerve function.
Diabetes Marker	HbA1c	42W5.	Long-term measure of blood sugar control.
Lipid Profile	Total Cholesterol	44P..	Assess cardiovascular risk and lipid metabolism.
	HDL (High-Density Lipoprotein)	441F.	"good" cholesterol, helps remove excess cholesterol from the bloodstream, reducing cardiovascular risk
	LDL (Low-Density Lipoprotein)	44PI.	"Bad" cholesterol; high levels can lead to plaque buildup in arteries.
	VLDL (Very Low-Density Lipoprotein)	VLDLC	Carries triglycerides; high levels linked to atherosclerosis.
	Triglycerides	44Q..	Type of fat; high levels increase risk of heart disease and indicate metabolic issues like diabetes.
Immune Response	Lymphocytes	42M..	Assesses immune response and infection status.
	Neutrophils	42J..	High levels indicate infection or inflammation
Liver Function	AST (Aspartate Aminotransferase)	44HB.	Indicates liver damage or disease.
	ALT (Alanine Aminotransferase)	44G3.	Elevated levels suggest liver injury or inflammation.
	Alkaline phosphatase	44F..	Assesses liver and bone health.
	Bilirubin	44EC.	High levels indicate liver dysfunction or hemolysis.

Data for laboratory tests required extra pre-processing because there were four large files which required integration. For example, for each patient row a column was created for different clinical measurements. **Table 12** presents an example of the original data format. Test types were identified using the CLINICALCODEVALUE field.

Table 12 Unstructured Lab test Dataset

SAMPLEDATE	SAMPLETIME	TISSUETYPE	CLINICALCODEVALUE	CLINICALCODEDESCRIPTION	QUANTITYVALUE	QUANTITYUNIT
2010-01-07	09:00:00	B	44EC.	Total Bilirubin	22.0	umol/L
2010-01-06	14:03:00	B	44I7.	Serum bicarbonate	29.0	mmol/l
2010-01-06	14:30:00	S	44F..	Alkaline Phosphatase	85.0	U/L
2010-01-06	14:59:00	B	44I5.	Serum sodium	132.0	mmol/l
2010-01-08	09:00:00	B	44F..	Alkaline Phosphatase	47.0	U/L

A function was created to identify blood tests taken at time 0 (diabetes diagnosis date) or within the next 6 months, ensuring the closest test date after a diagnosis of T2DM was extracted. Each clinical value was further restructured into separate columns rather than repeated rows, shown in **Table 13**. Rather than impute missing values, clinical experts advised using blood tests in the year prior to diagnosis; usually these were done within a few weeks prior to the diagnosis of T2DM. The "NaN" refers to patients with a missing clinical measurement at a certain time point.

Table 13 Reshaping Lab Tests

Glucose	HbA1c	Total_Cholesterol	LDL	HDL	VLDL	Triglycerides
NaN	53.0	3.8	1.50	2.4	0.7	1.3
11.1	47.0	3.6	1.70	3.6	0.8	1.4
6.8	50.0	4.4	NaN	2.8	NaN	4.4
11.8	38.0	3.2	0.00	3.2	NaN	1.4
12.6	NaN	3.9	1.42	3.9	1.5	3.3

For renal function, the estimated glomerular filtration rate (eGFR), was calculated from serum creatinine levels applying the IDMS-traceable Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) formula(Levey and Stevens, 2010). CKD-EPI was preferred over the Modification of Diet in Renal Disease (MDRD) equation as recommended by recent guidelines (Stevens et al., 2010; Griffiths et al., 2023).

The CKD-EPI formula is:

$$eGFR = 141 \times min\left(\frac{SCr}{k}, 1\right)^{\alpha} \times max\left(\frac{SCr}{k}, 1\right)^{-1.209} \times 0.993^{Age} \times A \times B$$

where:

- $\boldsymbol{SCr}$**:** Serum creatinine (mg/dL)
- $\boldsymbol{k}$**:** 0.7 for women and 0.9 for men
- $\boldsymbol{\alpha}$**:** -0329 for women and -0.411 for men
- **Age:** Patient's age at the time of test
- **A:** 1.159 if female, 1 otherwise
- **B:** 1.159 if African American, 1 otherwise

However, this research assumed no African American participants in the study and serum creatinine values were standardised and converted as required. Moreover, recent studies suggest that eGFR should not be adjusted for ethnicity (Diao et al., 2022). The National Institute for Health and Care Excellence (NICE) removed the recommendation for using an ethnicity adjustment for people of black African or black Caribbean ethnicity (Griffiths et al., 2023).

3.5.7 Mortality

Scottish mortality records are maintained by National Records of Scotland (NRS)(of Scotland, 2019). Each record (row) represents a death and includes the following details:

- SafeHaven ID: Used to link with other tables.
- Date of death.
- Location of death.
- Cause of death: Encoded using internation disease classification (ICD) codes, with up to ten causes listed per person in order of priority (primary, secondary, tertiary, etc.).

The duration between date of diabetes diagnosis and date of death was calculated as shown in **Appendix A4**.

3.6 Hong Kong Clinical Definitions

3.6.1 Socioeconomic Deprivation

The Comprehensive Social Security Assistance (CSSA) scheme in Hong Kong is a government-provided financial aid program designed to provide a safety net for individuals and families who face financial hardship. It aims to meet the basic living needs of recipients by providing cash assistance for essential expenses, such as food, housing and medical care. Eligibility is determined by means testing, which assesses the applicant's income and assets to ensure they fall below a specified threshold.

In the context of health studies, CSSA status has been used as an indicator of deprivation or socioeconomic disadvantage. Individuals receiving CSSA are considered to be experiencing financial hardship, which may correlate with various health outcomes, including higher mortality rates due to a greater accumulation of risk factors, more pre-existing health problems and poorer access to healthcare. However, CSSA status had large amounts of missing data, there was also a delay in assigning CSSA status shown in **Appendix C3.** It was unreliable to use as a measure of socioeconomic deprivation.

3.6.2 Missingness of Body Mass Index and Smoker

Data on BMI was missing in 90% of Hong Kong EMRs (Tsoi et al., 2020). BMI may not be recorded at clinical visits unless relevant to the patient's specific medical condition or treatment plan. The CDARS system captures data from in-patient, out-patient and A&E settings, but the focus is primarily on diagnoses and treatment. There were also very few records reporting smoking in the CDARS system. Smoking status may be underreported unless it is directly relevant to a patient's diagnosis or treatment. Smoking data often relies on self-reporting, which is not routinely updated. Hong Kong has seen a marked decrease in smoking rates compared to many other countries (Smokefree HK, 2024), possibly due to strong public health campaigns and high tobacco taxes. Only 10.2% of adults reported that they smoked in 2021 (Socrates Y WU1, 2021). For this research, the incomplete reporting of smoking status in CDARS may lead to missing data, potentially biasing health outcome studies where smoking is a relevant factor. Therefore, smoking status was not included in analysis for Hong Kong.

3.6.3 Prescriptions

In Hong Kong, the electronic prescribing system (EPS) is a core element of the public healthcare infrastructure, governed by the Hospital Authority (Hong Kong Authority, 2015). Prescriptions are issued electronically, ensuring medications are routinely updated across the public healthcare network. Within CDARS there is longitudinal record of prescribed medications, which greatly supported this research. A Medicine Formulary system is in place to standardise the medications available across public hospitals and clinics (Department of Health, 2020). All prescribed medications are associated with standardised medicine codes, which streamline data entry and reporting in CDARS. However, in the private sector, prescriptions are still predominantly paper-based, meaning data from private clinics and hospitals may not be fully integrated into CDARS. This creates gaps in patient medication histories within the database, especially for those seeking care outside the public system. Despite this limitation, CDARS remains a powerful resource for analysing prescription trends in public healthcare. A custom script was developed to compile multiple CSV files into a merged dataset, facilitating the analysis of long-term loop diuretic usage across the T2DM cohort. **Table 14** presents a snapshot of loop diuretic records extracted from the EPS integrated with CDARS. This approach allowed investigation of patients prescribed loop diuretics before and after (incident) a diagnosis of T2DM.

Table 14 Loop diuretics Records Extracted

Reference_Key	Dispensing Date (yyyy-mm-dd)	Prescription Start Date	Prescription End Date	Drug Item Code	Drug Name	Route	Drug Strength	Dosage	Dosage Unit	Dispensing Duration	Base Unit	Action Status
0	27/02/2009	27/02/2009	28/02/2009	FRUS03	FRUSEMIDE	INJECTION	10MG/ML 2ML	40.0	NAN	EVERY TWENTY-FOUR HOURS	AMP	Issued
1	28/02/2009	28/02/2009	29/02/2009	FRUS01	FRUSEMIDE	ORAL	40MG	1.0	TABLET(S)	DAILY	TAB	Issued
2	28/02/2009	28/02/2009	02/03/2009	FRUS01	FRUSEMIDE	ORAL	40MG	1.0	TABLET(S)	DAILY	TAB	Issued

Each row represents a distinct prescription event, identified by a unique Reference_Key. Key fields include:

- Dispensing Date and Prescription Start/End Dates: These columns show when the medication was dispensed and the specific period during which the medication was prescribed.

- Medicine Item Code and Medicine Name: These fields identify the medication, in this case, Frusemide (a common loop diuretic), with a specific code (e.g., FRUS03 and FRUS01).
- Route: Indicates the method of administration (e.g., INJECTION or ORAL).
- Medicine Strength and Dosage: Specify the potency (e.g., 10MG/ML 2ML) and amount (e.g., 40.0) per administration.
- Dosage Unit and Dispensing Duration: Describe the unit form (e.g., TABLET(S)) and the frequency or interval of administration (e.g., DAILY or EVERY TWENTY-FOUR HOURS).
- Base Unit and Action Status: Define packaging or measurement details (e.g., AMP, TAB) and the status of the prescription (e.g., Issued).

Additionally, under Hong Kong law, all pharmaceutical products must be registered with the Pharmacy and Poisons Board (PPB) ('PHARMACY & POISONS BOARD OF HONG KONG - Pharmacy and Poisons Ordinance', 2024). This registration checks that medications meet strict standards of safety, efficacy and quality before being available to the public. A pharmaceutical product is defined as a substance, or combination of substances, used for treating, preventing, or diagnosing diseases and it must conform to specific regulatory requirements.

3.6.4 Laboratory Tests

In Hong Kong, LOINC (Logical Observation Identifiers Names and Codes) codes are internationally used in clinical laboratory systems to standardise the reporting of laboratory test results. Developed by the Regenstreif Institute (Mok et al., 2013), LOINC provides a universal code system that enables interoperability across different healthcare providers and EMRs. This standardised coding system provides clarity and accuracy in reporting laboratory values, especially critical for population health studies. LOINC codes integrated in CDARS easily categorise and analyse laboratory biomarkers, such as liver enzymes, kidney function tests, inflammatory markers, lipid profiles and blood glucose levels. The adoption of LOINC codes in Hong Kong aligns the region with international standards. **Table 15** presents the extracted laboratory tests for this research including LOINIC code and primary purpose.

Table 15 Extracted Laboratory Tests using LOINIC Codes

Laboratory Test Category	Marker	LOINC	Purpose/Indication
Renal Function	Serum Creatinine	2160-0	Assesses kidney filtration and function.
	eGFR (estimated Glomerular filtration rate)	33914-3	Evaluates overall kidney function.
	U Albumin-to-creatinine ratio	9318-7	Early marker of kidney damage, especially in diabetes.
Inflammation	C-reactive Protein	1988-5	Indicates inflammation or infection.
Blood Chemistry	Haemoglobin	718-7	Evaluates the oxygen-carrying capacity of red blood cells.
	Serum Albumin	1751-7	Marker of liver function and nutritional status.
	Potassium	77142-8	Electrolyte balance critical for heart and nerve function.
	Lymphocytes	731-0	Assesses immune response and infection status.
	Neutrophils	751-8	High levels indicate infection or inflammation
Diabetes Marker	HbA1c	4548-4	Long-term measure of blood sugar control.
Lipid Profile	Total Cholesterol	2093-3	Assess cardiovascular risk and lipid metabolism.
	HDL (High-Density Lipoprotein)	2085-9	"good" cholesterol, helps remove excess cholesterol from the bloodstream, reducing cardiovascular risk
	LDL (Low-Density Lipoprotein)	2089-1	"Bad" cholesterol; high levels can lead to plaque buildup in arteries.
	VLDL (Very Low-Density Lipoprotein)	N/A	Carries triglycerides; high levels linked to atherosclerosis.
	Triglycerides	2571-8	Type of fat; high levels increase risk of heart disease and indicate metabolic issues like diabetes.
Liver Function	AST (Aspartate Aminotransferase)	2571-8	Indicates liver damage or disease.
	ALT (Alanine Aminotransferase)	1742-6	Elevated levels suggest liver injury or inflammation.
	Alkaline phosphatase	6768-6	Assesses liver and bone health.
	Bilirubin	1975-2	High levels indicate liver dysfunction or hemolysis.

Laboratory Test Imputation

Many times, patient records are incomplete or missing due to patient dropout, inconsistent testing schedules, human error and policy (Certain healthcare policies might not require or prioritise the collection of some data points in specific settings or patient groups). Clinical studies may exclude relevant patients or skew available data, misrepresenting true clinical relationships and leading to unreliable conclusions (van Smeden et al., 2021). Imputation reduces these biases by estimating probable values for missing entries (Di et al., 2022), allowing the records to be analysed as if it were complete. Methods like single imputation (e.g., mean substitution) or more complex approaches, such as multiple imputation (Austin et al., 2021), may fill gaps while preserving the variability and structure of the data. Initially, K-Nearest Neighbours (KNN) imputation was applied to handle missing laboratory test values in the dataset, but it significantly distorted results. KNN imputation works by averaging values from the "nearest" cases based on similar patient characteristics. This can create issues in clinical datasets, where patients often display high variability due to age, comorbidities, or lifestyle factors. KNN may inaccurately assume similarity between patients with different health backgrounds, leading to biased imputations. Therefore, the decision was to not impute and use a previous laboratory measurement for each patient.

3.6.5 Mortality

In Hong Kong, mortality records are systematically collected and maintained by the Census and Statistics Department (C&SD) in collaboration with the Department of Health. The data for mortality records come primarily from death certificates, which are required by law to be registered with the Hong Kong Immigration Department within 24 hours of a death. The records provide detailed information, including the date, location and certified cause(s) of death, which are using the ICD codes. These records are anonymised and stored within the CDARS system.

3.7 Glasgow and Hong Kong Hospitalisations

In Glasgow hospitalisations are coded using ICD-10 (10th revision) and in Hong Kong using an earlier version, ICD-9 (9th revision). These are standardised diagnostic coding systems developed by the World Health Organization (WHO). **Figure 11** provides a historical overview of the development of ICD codes. The process began with the International List of Causes of Death (ILCD) in 1893, created by the International Statistics Institute. Over time, the list evolved with multiple revisions to expand and refine the classification of diseases. With ICD-6 in 1948, WHO officially adopted the system as the "International Classification of Diseases." The coding structure continued to improve, with major expansions in ICD-9 and ICD-10, which added more detailed categories to support expanding healthcare needs. The latest version, ICD-11, was introduced in 2022, featuring a restructured coding scheme for greater adaptability and more efficient clustering of related conditions. This evolution highlights the ICD's aim to provide a robust framework adaptable to the changing landscape of global health. **Table 15** shows the extracted ICD codes used in both Glasgow and Hong Kong. Stroke events were further classified as shown in **Table 17** (Kokotailo and Hill, 2005). The evolution of these ICD codes demonstrates an internally grounded coding system.

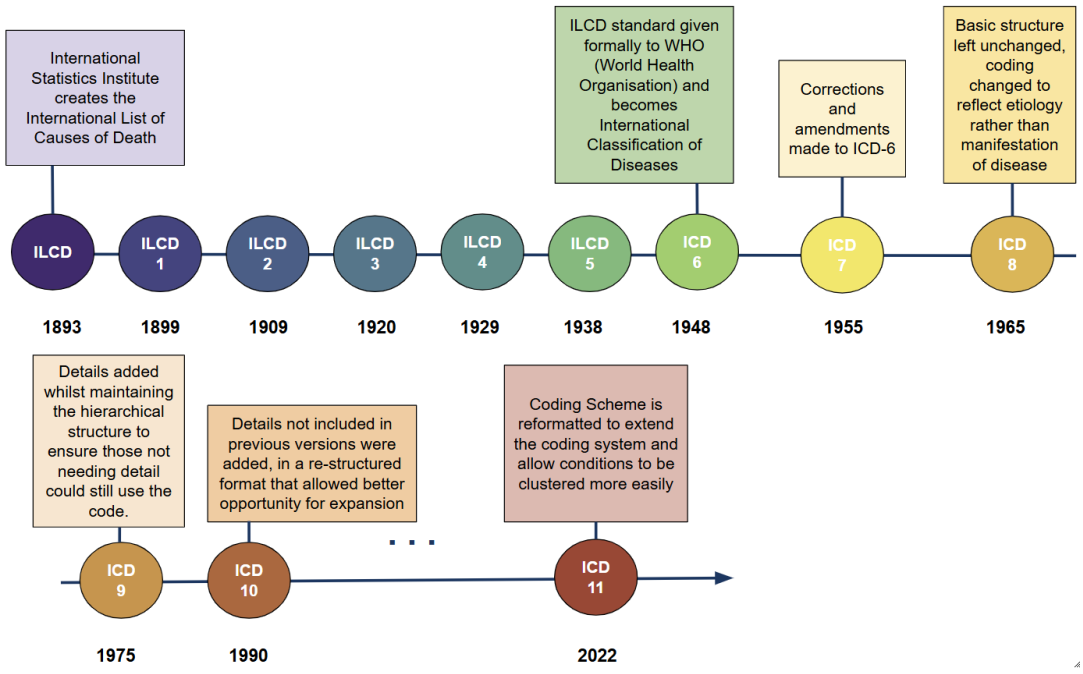


Figure 11 Historical Overview of ICD Codes

Table 17 Extracted ICD Hospitalisation Codes for Glasgow and Hong Kong

Hospitalisations defined by International Disease Classification (ICD) Codes	
Clinical Condition	**Codes & Description**
Angina	**ICD-9**: 413.x (Angina pectoris, with subtypes for unstable and variant angina) **ICD-10**: I20.x (Angina pectoris, including unstable and unspecified types)
Atrial Fibrillation (AF)	**ICD-9**: 427.31 **ICD-10**: I48.x (ranges from unspecified to paroxysmal or persistent atrial fibrillation)
Chronic Obstructive Pulmonary Disease (COPD)	**ICD-9**: 496 **ICD-10**: J44.x (where x can specify chronic bronchitis or emphysema)
Coronary Artery Disease (CAD)	**ICD-9**: 414.0x (Atherosclerosis of coronary arteries) **ICD-10**: I25.x (Chronic ischemic heart disease with specific codes for atherosclerosis types)
Chronic Kidney Disease (CKD)	**ICD-9**: 585.x (where "x" designates CKD stage) **ICD-10**: N18.x (where "x" designates CKD stage from 1 to 5, or end-stage renal disease as N18.6)
Heart Failure (HF)	**ICD-9**: 428.x (Heart failure with specific types like congestive, unspecified) **ICD-10**: I50.x (ranges from unspecified to specific types like left ventricular or congestive heart failure)
Hyperkalemia	**ICD-9:** 276.7 - Abnormally high blood potassium levels. **ICD-10:** E87.5 - Hyperkalemia, elevated potassium levels in the blood.
Hypertension	**ICD-9**: 401.x (essential hypertension) **ICD-10**: I10 (for primary hypertension), and I11.x, I12.x, etc., when accompanied by heart or kidney disease
Myocardial Infarction (MI)	**ICD-9**: 410.x (Acute myocardial infarction, with specific digits indicating episode) **ICD-10**: I21.x (for current MI, including specific artery affected) and I22.x (for subsequent MI)
Peripheral Artery Disease (PAD)	**ICD-9**: 443.9 (Peripheral vascular disease, unspecified) **ICD-10**: I73.9 (Peripheral vascular disease, unspecified)
Type 2 Diabetes Mellitus (T2D)	**ICD-9**: 250.x0, 250.x2 (where "x" denotes type and the code indicates with or without complications) **ICD-10**: E11.x (ranges from unspecified to with specific complications like kidney disease or retinopathy)

Table 16 Stroke Categories in ICD-9 & ICD-10 Codes

Stroke Categories defined by ICD Codes			
Condition	**ICD-9**	**ICD-10**	**Description**
Stroke (Ischemic)	434.x1	I63.x	Occlusion of cerebral arteries
	433.x	I64.x	Stroke, not specified as hemorrhage or infarction
	362.3	H34.1	Retinal vascular occlusion (not specified)
Stroke (Hemorrhagic)	431.x	I60.x	Subarachnoid hemorrhage
	432.x	I61.x	Intracerebral hemorrhage
	436	I67. 81	Acute, but ill-defined cerebrovascular disease
Transient Ischemic Attack (TIA)	435.x	G45.x	Transient cerebral ischemia and related syndromes

3.7.1 Prevalent & Incident Heart Failure

Heart failure (HF) is a clinical syndrome rather than a single pathological diagnosis with a robust definition (Theresa A McDonagh et al., 2021). Symptoms include breathlessness, ankle swelling and fatigue (Groenewegen et al., 2020). Diagnostic codes for heart failure in Glasgow and Hong Kong, derived from hospitalisation events, were identified using the international classification of disease, tenth revision (ICD-10) system (**Table 18**). ICD-10 diagnostic codes for various types of heart failure were grouped together to produce a data column indicating if a patient already had (prevalent) heart failure prior to a diagnosis of T2DM or subsequently developed (incident) heart failure.

Table 18 ICD 10 Codes for Heart Failure

ICD 10 Codes for Heart Failure	
Diagnostic Code	**Description**
I500	Congestive Heart Failure
I5009	Congestive Heart Failure – no information on ejection fraction
I5099	Unspecified Heart Failure - no information on Ejection Fraction
I5091	Heart Failure, unspecified – Preserved Ejection Fraction
I5000	Congestive Heart Failure – Reduced Ejection Fraction
I110	Hypertensive Heart Disease with (Congestive) Heart Failure
I130	Hypertensive Heart and Renal Disease with (Congestive) Heart Failure
I132	Hypertensive heart and chronic kidney disease with heart failure and with stage 5 chronic kidney disease, or end stage renal disease
I139	Hypertensive Heart and Renal Disease, Unspecified

A diagnostic code for heart failure in any diagnostic position, not just the primary position, in a hospitalisation record was accepted (**Appendix A5**). To ensure the heart failure events were incident, patients with *prevalent heart failure* were excluded by applying a 5-year look-back period prior to the diabetes diagnosis.

3.8 Selecting a Study Period

In EMRs, there is not a specified study period or screening stages of patients. Setting the time period for each patient was a sensitive iterative process. Traditionally, most clinical studies investigate data sets from randomised clinical trials. A set period is of dates and expected outcomes is provided. However, working with longitudinal population data required additional steps. In some cases, patients were misdiagnosed. For example, patients were diagnosed with diabetes after their date of death. Most misdiagnosed dates began in the year of the pandemic. To overcome complications, time 0 began at the beginning of a first diabetes diagnosis. This was anytime between the study period presented in **Figure 12**. Between both dates, events were captured: demographics, prescriptions, laboratory tests and hospitalisations. A function was implemented to calculate the number of days between these important clinical characteristics.

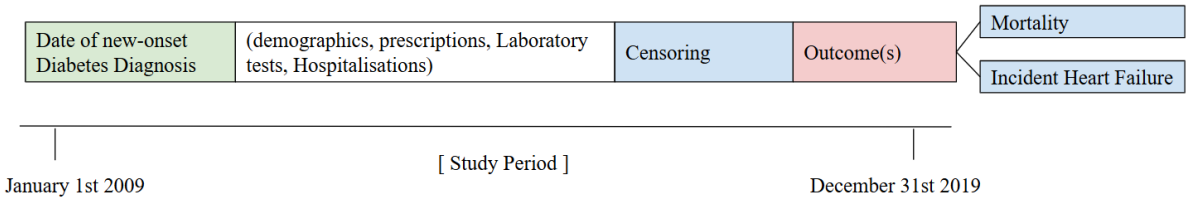


Figure 12 Study Period Representation

3.9 Results

Table 19 shows the baseline characteristics of the T2DM Glasgow, West of Scotland and Hong Kong, China SAR Populations. The Glasgow and Hong Kong cohort has similar age and sex distribution, although there are more percentage of women in Hong Kong. Ethnicity differs, with the Hong Kong cohort predominantly Chinese (92%) compared to Glasgow's population of 85% white ethnicity. Patients in Glasgow have a higher prevalence of chronic kidney disease (5% vs. 1%) and atrial fibrillation (13% vs. 3%) (both $p<0.001$). Differences in some biomarkers show lower total cholesterol and triglycerides but higher neutrophil counts in Glasgow compared to Hong Kong ($p<0.001$ for all). Alanine transaminase (ALT) and aspartate transaminase (AST) levels are notably higher in Hong Kong ($p<0.001$). Metformin use is significantly higher in Hong Kong (68% vs. 32%; $p<0.001$), while insulin and thiazides are more frequently prescribed in Glasgow ($p<0.001$ for both).

Table 19 Baseline Characteristics

Demographics at Baseline:	**Glasgow N = 46,031**	**Hong Kong N= 273,876**
Age (years)	64 (57 – 72)	65 (56 – 75)
Sex		
Men	24,664 (54%)	132,040 (48%)
Women	21,367 (46%)	141,836 (52%)
Ethnicity		
Chinese	N/A	251,966 (92%)
White	39, 290 (85%)	N/A
Other	6,741 (15%)	21,910 (8%)
***Body Mass Index (BMI)**	26 (26 – 31)	25 (23 – 26)
Smoker (Yes)	9,416 (20%)	N/A
Death Record	11,727 (25%)	91,155 (33%)
Socioeconomic Status (SIMD)		64,380 (24%) Receiving comprehensive social security assistance (CSSA)
Quintile 1 – Most Deprived	18,517 (41%)	
Quintile 2	8,355 (18%)	
Quintile 3	6,360 (14%)	
Quintile 4	5,643 (12%)	
Quintile 5 – Least Deprived	7,156 (16%)	
Hypertension (yes)	18,999 (41%)	64,246 (23%)
Chronic Kidney Disease (yes)	2,549 (6%)	3,381 (1%)
Hyperkalaemia (yes)	2,336 (5%)	N/A
Atrial Fibrillation (yes)	6,083 (13%)	7,772 (3%)
COPD (yes)	4,395 (8%)	818 (0.3%)

Coronary Heart Disease (yes)		7,106 (16%)	26,423 (10%)
Myocardial Infarction (yes)		4,545 (11%)	N/A
Peripheral Artery Disease (yes)		1,650 (4%)	346 (0.1%)
Stroke/TIA (yes)		4,010 (9%)	8,986 (3%)
Heart Failure (yes)		4,675 (10%)	11,189 (4%)
Anaemia		2, 302 (5%)	19,425 (6%)
Haemoglobin A1C (mmoL)		55 (46– 61)	56 (51– 63)
Haemoglobin (g/L)	**Men** **Women**	138 (134 – 151) 134 (123 – 139)	131 (132 – 139) 129 (122 – 136)
Lymphocyte count (x10^9/L)		2.0 (2.0 – 2.3)	1.9 (1.7 – 2.4)
Neutrophil count (x10^9/L)		5.0 (3.8-5.6)	5.3 (4.4 – 7.1)
Total Cholesterol (mmol)		4.1 (3.7-5.0)	4.7 (4.3 – 5.2)
Triglycerides (mmol)		1.5 (1.0 – 2.2)	1.5 (1.1 – 1.9)
Serum Albumin (g/L)		37 (35-39)	40 (38 – 42)
eGFR (mL/min/1.73m2)		54 (44 - 61)	64 (53 – 77)
Potassium (mmol)		4.3 (4.0 – 4.6)	4.2 (4.0 – 4.4)
Alanine Transaminase – ALT (U/L)		22 (16-30)	23 (17 – 30)
Aspartate Transaminase – AST (U/L)		20 (16-26)	25 (21 – 39)
Alkaline Phosphatase (U/L)		89 (72-105)	75 (65 – 87)
Bilirubin (µmol/L)		10 (7 - 13)	10.3 (9.2 – 12.8)
Metformin (yes)		14,545 (32%)	185,881 (68%)
Sulphonylureas (yes)		10,204 (22%)	173,525 (63%)
DPP4i (yes)		5,033 (11%)	325 (0.1%)
GLP1-receptor antagonists (yes)		2,139 (5%)	17
Insulin (with Glucose-Lowering Agent)		2,801 (6%)	29,697 (11%)
Statins (yes)		23,802 (52%)	61,401 (22%)
Beta Blockers (yes)		10,243 (22%)	92,309 (34%)
ACEi or ARBS (yes)		20,549 (60%)	121,786 (44%)
Calcium Channel Blockers		4,309 (9%)	109,225 (40%)
Thiazides (yes)		12,021 (26%)	52,096 (19%)
Loop Diuretics		11,403 (25%)	60,152 (22%)

3.10 Discussion

Routinely collected EMRs provide sufficient amount of data for assessing patient health trajectories and overall population investigation. This enhances the ability to study large, diverse cohorts. Several research studies have highlighted the usability of EMRs. Glasgow and Hong Kong capture the key clinical characteristics for disease progression in patients with T2DM. The integration of these datasets highlights the advantages of EMRs for cross-population comparison and generalisability. It represents the universal patterns of disease burden. Despite the similarities in age and sex, the Glasgow cohort consisted of a more ethnically diverse population, while the Hong Kong cohort was predominantly Chinese, reflecting the unique demographics of these regions.

Even though, populations were from different regions, the use of standard codes ensured consistency and comparability in EMRs. Glasgow relied upon READ Codes for laboratory results whereas Hong Kong used LOINC. Although Glasgow and Hong Kong represent distinct healthcare settings, the use of standard coding systems like ICD-9 or ICD-10 ensures consistency and comparability across their EMRs. Both regions leverage national mortality registries to track death outcomes. Scotland's National Records of Scotland and Hong Kong's Death Registry systematically use ICD codes to standardise mortality data, facilitating robust epidemiological comparisons.

Prescription data, however, reflects regional differences in digital infrastructure. In Scotland, medications are managed through the British National Formulary (BNF) prescribing information system, where all prescriptions are digitally recorded, enabling comprehensive tracking of medication history. In contrast, Hong Kong employs an electronic prescribing system (e-Prescription) under the Hospital Authority, which stores medication histories. However, this system is not universal across private sectors or smaller healthcare providers. As a result, prescription records in Hong Kong may lack the completeness achieved in the UK system, where nearly all prescriptions, both in primary and secondary care, are digitised and linked to patient records. These differences highlight how standardised systems enhance comparability but also show regional contrasts in the completeness of prescription data, influencing medication trend analyses.

Furthermore, both cohort measurements majority laboratory test was consistent for example, hbA1c a threshold of ≥48 mmol/mol is used for diagnosing diabetes in NHS guidelines, NICE guidelines and alignment of WHO standards. However, some clinical biomarkers had different thresholds. The threshold for CKD diagnosis is also eGFR <60 mL/min/1.73 m^2, following international standards. However, in some studies, a higher threshold for the diagnosis of early-stage CKD may be used in Hong Kong due to local population characteristics and clinical practices.

Socioeconomic status (SES) is important for understanding health disparities. The Glasgow cohort benefited from rich, detailed records of socioeconomic status, which was collected using the Scottish Index of Multiple Deprivation. This enabled further investigation into how deprivation influences health outcomes, a risk factor that has been shown to influence CVD progression in the UK (Nagar et al., 2021). In contrast, the Hong Kong cohort faced challenges with missing or delayed SES information due to issues in the government's Comprehensive Social Security Assistance assignment system. The delay in assigning CSSA status, which determines financial aid eligibility, hindered the inclusion of socioeconomic factors in the analysis, limiting insights into how SES interacts with health outcomes in Hong Kong. This discrepancy highlights the importance of timely and SES data collection for accurate health risk modelling and intervention planning.

Moreover, patients in Glasgow had a higher mortality, where UK studies often report increased mortality in patients with T2DM (Lin et al., 2024). In contrast, Hong Kong demonstrated relatively lower mortality, which aligns with findings which found that the mortality rates for diabetes-related complications in Hong Kong were generally lower than in Western populations (Wan et al., 2023). Hong Kong carries out screening every 1 to 3 years for risk assessments (Disease Branch et al., 2021) in T2DM patients whereas Glasgow follows stricter annual monitoring protocols using national guidelines (Moran et al., 2022). However, it is important to consider that differences in healthcare systems, preventive measures and access to care likely contribute to these variations.

3.11 Conclusion

The importance of transparency in processing and transforming EMRs is crucial. Clearly defined data sources, methods and justifying each step provides precision to the research results. Using EMRs from two distinct populations, GG&C in the United Kingdom and Hong Kong SAR, China—allows comparison of predictive models for incident heart failure and for mortality for these two very different populations. Substantial data transformation was required to prepare patient cohort for statistical analysis and further machine learning implementation. If these models perform similarly well, it suggests that they might be generalisable to many other geographies, cultures and ethnicities. This chapter demonstrates the value of integrating EMRs from distinct populations for understanding disease trajectories and improving predictive modelling in T2DM. By accounting for differences in data collection methods and leveraging patient data, this work contributes to advancing precision medicine and informing global health policy.

Chapter 4 "Predicting Incident Heart Failure in Patients with Type 2 Diabetes Mellitus: A Machine Learning Approach".

Abstract

Introduction: People with Type 2 diabetes (T2DM) are at increased risk of developing Heart Failure (HF). Using general electronic medical records (EMRs), we applied a machine learning (ML) approach to identify variables that predict incident HF in patients with T2DM (Narinder Kaur et al., 2023).

Methods: National Health Service Scotland EMRs were linked with the Scottish Care Information - Diabetes Registry (SCI-Diabetes), which includes demographic data, blood test results, prescriptions, comorbidities from primary and secondary care diagnostic codes and deaths. Incident HF was defined by the International Classification of Diseases, 10th Revision (ICD-10) codes for hospitalisations, with a look-back period of 5 years to exclude prevalent cases. We developed the random survival forest model: a non-parametric decision tree, which supports time-to-event data, to predict incident HF. We used Cox proportional hazards models to investigate associations between the prescription of loop diuretics and the risk of new-onset heart failure. We applied a state-of-the-art ML explainability method called Shapely Additive Explanations which interprets the direction of association for each contributing risk factor determining a patient's risk score.

Results: Of 29,868 patients with T2DM age ⩾50 years, 8,120 (27%) had coronary artery disease (CAD) at the time of enrolment, and 976 (3%) received a new diagnosis of HF between 2009-19. Key predictors of incident HF were use of loop diuretics, history of atherosclerosis events (myocardial infarction and angina), atrial fibrillation, lower estimated glomerular filtration rate (eGFR) and older age. Individuals prescribed loop diuretics had a 5-fold higher risk of incident HF than those who were not (HR: adjusted for age and sex 5.89 [95% CI 5.27 – 6.58 (<0.005)]). People with greater socioeconomic deprivation were also at greater risk of developing HF. The model c-statistic score was 0.87 and the brier score was 0.02 (low values indicate greater accuracy) for predicting incident HF.

Conclusion: A ML model using readily available EMR data accurately predicts incident heart failure in patients with T2DM.

Patients prescribed loop diuretics are at much greater risk of receiving a diagnosis of heart failure, although this might reflect, at least in part, patients with a previously missed diagnosis of heart failure.

4.1 Introduction

People with T2DM are at increased risk of developing heart failure (HF) (Rosano et al., 2017). The pathophysiological mechanisms linking T2DM and HF are complex, involving metabolic, hemodynamic and inflammatory processes. Identifying individuals with T2DM who are at heightened risk of developing HF is crucial for early intervention and management yet remains a challenge in clinical practice. Machine learning (ML) has the advantage of computing several patient characteristics on a time-to-event basis, compared to conventional statistics. Previous clinical trials have shown insights into the relationships between diabetes and HF but suffer from limitations in population representativeness, as trial participants are typically younger and have fewer comorbidities with under-representation of ethnic minorities and those with the most deprived and most affluent socioeconomic status. Recently, Segar et al developed the WATCH-DM model *(**W**eight [BMI], Age, **H**yper**T**ension, **C**reatinine, **H**DL-C, **D**iabetes **M**ellitus control [fasting plasma glucose])* machine learning risk score model using the variables denoted by the acronym selected from a total of 147 variables. WATCH-DM was applied to a clinical trial dataset called ACCORD (Segar et al., 2019). However, the model is based on patients who were required to fulfil the inclusion and exclusion criteria for a trial. Also, those invited to participate were selected by clinical experts and because most of those invited declined to participate, the results may not be generalisable beyond the trial population. Many people with severe chronic conditions are excluded from clinical trials. The incidence of HF may also vary widely by race and socioeconomic status. Risk score models for predicting incident HF in patients with T2DM should be developed and tested in representative populations but ensuring diverse characteristics if the intention is for them to be generalised.

Survival analysis models, including the Cox proportional hazards model, have long been the standard for time-to-event data (Razaghizad et al., 2022) but face challenges with assumptions like proportional hazards and handling correlated variables. Alternatively, survival-based ML models, handle high-dimensional data and uncover non-linear relationships, potentially improving predictive accuracy.

These models can be paired with interpretability tools, such as SHapley Additive exPlanations (SHAP), to ensure clinicians can trust and act on the predictions. Interpretability is important in clinical practice to foster healthcare professionals trust in the predictive models, enabling them to evaluate and potentially improve on the model and to act more effectively on modifiable risk factors.

4.2 Aim

To identify key predictors of incident HF in a population of patients with T2DM managed by NHS Greater Glasgow & Clyde (GG&C) using a machine learning approach.

4.3 Data Sources

The Glasgow SafeHaven dataset linked with SCI-diabetes (**see Chapter 3, section 3.5.6**) was used for this Chapter.

4.4 Study Patient Information

We obtained National Health Service (NHS) electronic medical records (EMR) of routinely collected health data for GG&C for people aged $\geqslant 50$ years with an incident diagnosis of T2DM between 1st of January 2009 to 31st December 2019. Data were linked with the SCI-diabetes (see section 5.3) dataset using the Glasgow SafeHaven trusted research environment. The prevalence and incidence of HF was identified using ICD-10 codes in any hospitalisation position. Loop diuretic prescriptions were defined by the British National formulary (BNF) classification. The decision was made to focus on repeat prescriptions of oral loop diuretics, specifically those prescribed for periods exceeding 90 days, as the relevance of occasional, isolated prescriptions is uncertain.

Patient characteristics include demographic details such as age, sex, socioeconomic status and ethnicity. Prevalent comorbidities (chronic kidney disease (CKD), chronic obstructive pulmonary disease (COPD), atherosclerotic heart disease, hyperkalaemia, peripheral artery disease, stroke, myocardial infarction, atrial fibrillation, and angina) were defined by the International Classification of Diseases, 10th Revision (ICD-10) codes.

The following routinely collected blood tests were captured as close as possible to the time of T2DM, with a window of ±6 months: glucose, haemoglobin A1c, haemoglobin, total cholesterol, HDL-C, LDL-C, VLDL, triglycerides, serum albumin, serum creatinine, estimated glomerular filtration rate (eGFR), urine albumin-to-creatinine ratio, potassium, lymphocytes, neutrophils, AST, ALT, alkaline phosphatase and bilirubin. Treatments for diabetes, repeated loop diuretics (3 or more consecutive prescriptions), cardiovascular and lipid-lowering medications were extracted ±180 days of T2DM diagnosis.

Figure 13 shows the consort diagram for this analysis. After excluding duplicate entries , 47,396 unique patients had a diagnostic label of diabetes. The first record of T2DM was used as the incidence date, except for 4,947 (9%) patients who had a test showing impaired glucose tolerance or raised fasting plasma glucose and subsequently received treatment for diabetes without a formal diagnosis of T2DM; for these patients the date of entry into the database was used as the incidence date. Of these, n=1,365 were excluded because they either had Type-1 diabetes mellitus or were only prescribed insulin (unusual for incident T2DM). Those with prevalent heart failure (extracting from ICD codes in patients with a history of heart failure: before diabetes diagnosis date) (n = 627) were excluded, with a look-back period of 5 years. Many patients had missing values for BMI, which appeared informative. This is dealt with further in chapter 7. Between 1st January 2009 and 31st December 2019, out of 29,868 patients included, 965 (3%) developed HF.

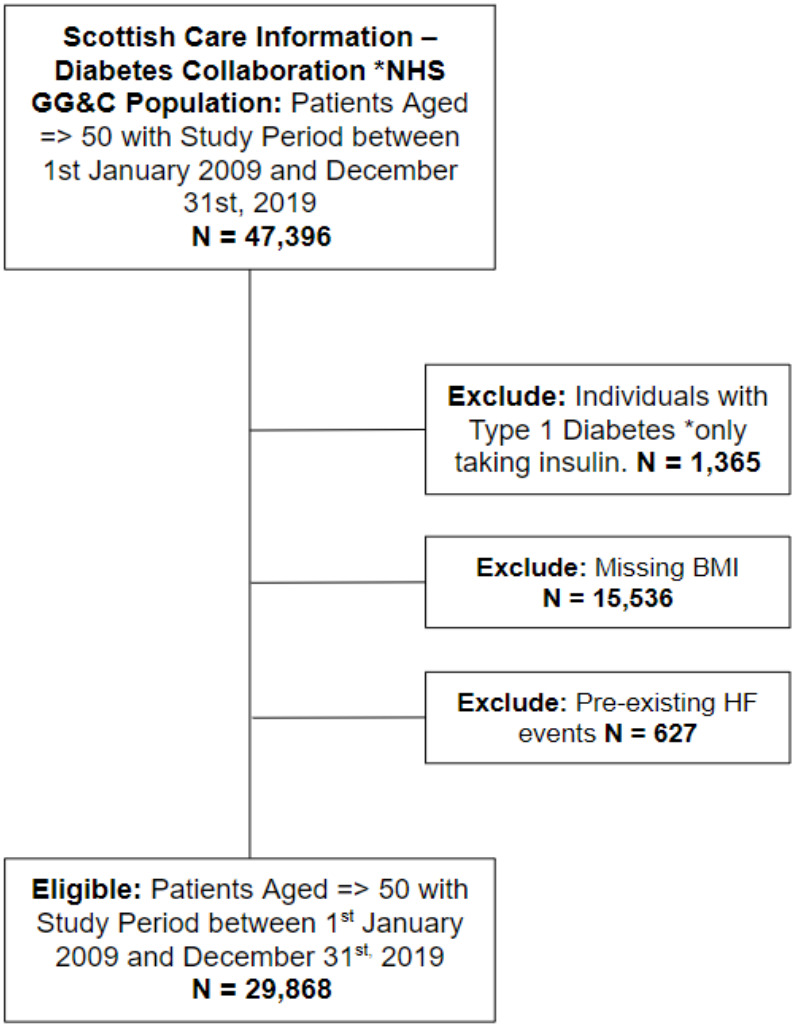


Figure 13 Consort Diagram of T2DM & Incident HF

4.5 Methods

4.5.1 Laboratory Tests & Correlation Analysis

Correlation analysis was key to uncovering linear and non-linear relationships, especially amongst laboratory tests presented in section 4.6.1 (lipid profile, liver function, kidney function, and haematology). Correlations were carried out to reduce highly correlated variables. Correlation methods to assess the relationships between variables were used: Spearman Rank, Pearson's, and Phi coefficients.

Spearman is a non-parametric measure of the strength and direction of association that exists between two variables measured on at least an ordinal scale. Pearson's rho measures the strength and direction of the relationship between two variables. When one variable changes, the other variable changes in the same direction. It assumes both variables are continuous. Phi (or mean square contingency coefficient) denoted by φ is a measure of association between two naturally dichotomy variables (Baak et al., 2019). This refers to non-linear relationships, measuring variables that are completely opposite. Since linear regression assumes a linear relationship between the input and output variables, it fails to fit complex datasets accurately. The application of Phi correlations due to their ability to capture non-linear relationships was further investigated for incident HF. This analysis identified clusters of closely related tests, simplifying the selection process by eliminating redundant variables. Throughout the process, expert cardiologist input was integral to verifying the sense of computational relationships.

Figure 14 outlines an example of the phi correlation for five relationships. For example, triglyceride concentrations are consistently associated with serum creatinine concentrations, due to decreased kidney function. Also, there were high correlations between BMI and a diagnosis of hypertension. Correlations were useful to discover interesting relationships amongst blood tests. For further data modelling, correlations identified strong associations, although correlation tests do not necessarily mean causal relationships amongst variables. In terms of working with high-dimensional data advanced feature selection methods were applied to validate the features influencing survival probability.

Serum Albumin	1.00	0.74	0.76	0.63	0.45
C-reactive Protein	0.74	1.00	0.93	0.85	0.88
Serum Creatinine	0.76	0.93	1.00	0.90	0.95
Haematocrit	0.63	0.85	0.90	1.00	0.73
Triglycerides	0.45	0.88	0.95	0.73	1.00
	Serum Albumin	C-reactive Protein	Serum Creatinine	Haematocrit	Triglycerides

Figure 14 Phi Correlation Example

4.5.2 Survival Analysis

Survival analysis investigates the time-to-events. It focuses on estimating and interpreting the time until an event, such as incident heart failure (Srujana et al., 2022), occurs, considering the possibility of censoring due to lack of long-term, or loss to, follow-up. Typically, standard machine learning models such as linear regression to do not account for time to event data (Prinja et al., 2010).

Key Concepts in Survival Analysis:

1. **Survival Time:**
    - The primary outcome of interest, often denoted as T, represents the time from a defined starting point (e.g., diagnosis, treatment initiation) to the occurrence of the event of interest (e.g., death, relapse).
2. **Event:**
    - The specific outcome or occurrence being studied (e.g., death, disease recurrence, machine failure).
3. **Right-Censoring:**
    - A unique aspect of survival data where the event of interest has not occurred for some subjects during the study period. These subjects are referred to as "censored." Types of censoring include right-censoring (most common), left-censoring, and interval-censoring.
4. **Survival Function:**
    - The survival function S(t) represents the probability that the event of interest has not occurred by time t (Clark et al., 2003). It is defined as:

$$S(t) = P(T > t)$$

5. **Hazard Function:**
    - The hazard function λ(t) represents the instantaneous rate of occurrence of the event at time t, given that the subject has survived up to time t. It is defined as:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{p(t) \leq T < t + \Delta t | (T \geq t)}{\Delta t}$$

Where:

- $\lambda(t)$ is the instantaneous hazard rate at time t. (Risk of heart failure)
- $p(t) \leq T < t + \Delta t | (T \geq t)$ is the conditional probability that the event occurs in the small-time interval, given that the individual has survived up to time t.
- Δt is the small-time interval
- $\lim_{\Delta t \to 0}$ The limit ensures that the hazard function describes risk at a specific moment.

6. **Cox Proportional Hazards Model:**

   - A semi-parametric model that assesses the effect of multiple covariates on the hazard rate. The model assumes that the covariates have a multiplicative effect on the hazard function.
   - The hazard function in the Cox model is given by:

$$\lambda(t|X) = \lambda_0(t)\, exp\left(\beta, X_1 + \beta_2 X_2 + \cdots + \beta_p X_\rho\right)$$

where $\lambda_0(t)$ is the baseline hazard and β represents the coefficients for the covariates X.

4.5.3 Kaplan Meier

Kaplan Meier plots were applied to investigate patients developing HF when prescribed a loop diuretic or not. Kaplan-Meier analysis is a non-parametric statistical method used to estimate the survival function from time-to-event data (Bewick et al., 2004). The Kaplan-Meier survival curve estimates the probability that a patient will survive (i.e., not experience heart failure) beyond a certain time t. Mathematically, the survival function $S(t)$ is defined as:

$$S(t) = \frac{Number\ of\ patients\ surviving\ beyond\ time\ t}{Total\ Number\ of\ patients\ at\ the\ start\ of\ time\ t}$$

The Kaplan-Meier estimator is calculated step-by-step at each time point where an event occurs. The time is divided into intervals based on when heart failure events occur. At each event time t_l, the survival probability is calculated as:

$$s(t_i) = s(t_i - 1) \times \left(1 - \frac{di}{ni}\right)$$

where:

- di is the number of events (incident heart failure) at time t_i.
- ni is the number of patients at risk just before time t_i those who have not yet had the event or been censored.
- $s(t_i - 1)$ is the survival probability up to the previous event time.

The survival curve is a step function that decreases with each event, reflecting the reduction in the proportion of patients who have not yet experienced the event. Finally, the log-rank test is used to compare the survival distributions between the groups (i.e. prescribed loop diuretics vs. not prescribed). It is based on comparing the observed and expected number of events in each group across the entire study period. Additionally, the cox proportional hazards model investigated associations between the prescription of loop diuretics and the risk of incident heart failure.

4.5.4 Penalized Cox Regression

The standard Cox proportional hazards model fails to estimate the coefficients of several features in an analysis because internally it tries to invert a matrix that becomes non-singular due to correlations among features. To overcome this, the Elastic Net Cox regression model, based on the Cox proportional hazard assumption is used. It performs automatic variable selection and regularisation using ridge and lasso regression (Lai et al., 2013). Ridge reduces the impact of features that are not important in predicting the outcome. Lasso improves upon the ridge method. It eliminates many features and reduces overfitting. Moreover, lasso also has constraints in its principles for survival analysis for high-dimensional data. If there is a group of variables among which the pairwise correlations are very high, then the lasso tends to arbitrarily select only one variable from the group. However, this is not reliable for discovering the key prognostic factors for the drivers of disease progression.

Figure 15 presents an example of Elastic Net feature selection of the baseline model. The x-axis represents all variables with reduced coefficients. The coefficient value signifies how much the mean of the dependent variable changes given a one-unit shift in the independent variable while holding other variables in the model constant. The alpha is a parameter that determines how much "weight" is given to both L1 (lasso) & L2 (ridge) penalties. Alpha is an arbitrary hyper-parameter that controls the amount of shrinkage. All coefficients are shrunk almost to zero. When alpha values in the Y-axis are decreased, the coefficients value increases. Higher or lower coefficients mean they have effects on target variables. For example, age has a large coefficient for a wide range of alpha. Whereas triglycerides start to dominate with a small alpha. However, discovering prognostic factors using the elastic net is model specific (Suchting et al., 2017). There is also a loss of interpretability as coefficients shrunk to zero, which is an essential requirement for clinical decision making.

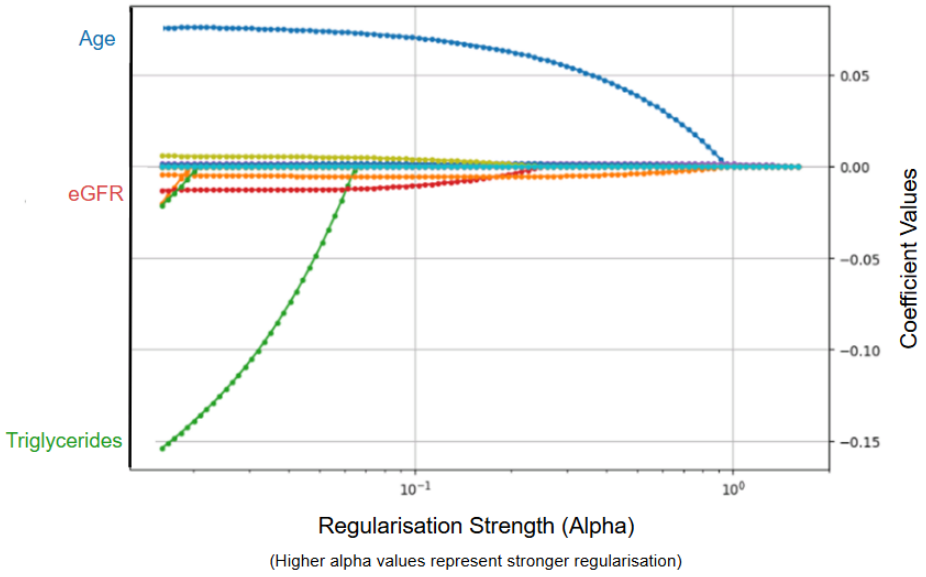


Figure 15 Elastic Net Feature Selection Baseline Model

4.5.5 Random Survival Forests in Predicting Incident Heart Failure

For this study, the Y array structure was for incident heart failure, as illustrated in **Figure 16**. The Y outcome was represented as a record array that included both the occurrence of an incident HF event and the duration until the event occurred. The random survival forest (RSF) method, was applied to assess all contributing risk factors for the development of HF.

The RSF model supports time-to-event data and censoring (Ishwaran, Udaya B. Kogalur, et al., 2008). RSF is a machine learning algorithm that combines the concepts of survival analysis and random forest mechanism. It is a non-parametric approach that can handle complex interactions and non-linear relationships between predictors and survival outcomes. This method is robust to violations of Cox proportional hazards assumption and can handle high-dimensional datasets.

The RSF model predictor is an ensemble (group) formed by combining the results of many survival trees (Ishwaran, Udaya B Kogalur, et al., 2008). The dataset is split into multiple random samples. It is an estimator that fits several survival trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. In each survival tree, the quality of a split is measured by the log-rank splitting rule.

The log-rank splitting rule is a powerful criterion used in survival trees to determine the optimal way to split data at each node. It effectively separates subsets of patients with different survival distributions, which is crucial for building accurate and informative survival *models.*

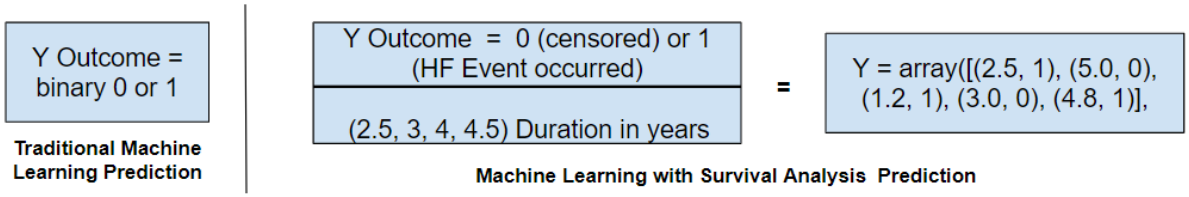


Figure 16 Y Outcome for incident HF risk prediction

4.5.6 Validation of Machine Learning Model(s)

To validate the model, the penalized Cox regression approach was applied. The penalty applied in penalized Cox regression forces the model to prioritise variables that have the strongest association with the outcome while reducing the influence of weaker, less significant predictors. In practice, this means that many variables with little or no contribution to predicting heart failure are either shrunk to near zero or entirely excluded from the model. This process effectively eliminates these less relevant features and focuses on the most significant predictors, not to analyse all contributing factors. This approach yields a focused model, suited to identifying key risk factors rather than presenting an exhaustive list of contributing variables.

4.5.7 Evaluation

Given the complexity of patient data with censored observations, specialised metrics are required to assess the performance of survival models. The prediction errors are measured by performance evaluators: c-index and time brier score The concordance index (C-index) is a metric in survival analysis that provides an assessment of the model discrimination power to predict incident heart failure. It measures the concordance between the predicted and actual event times (Alabdallah et al., 2022). A score close to 1.0 means the model is reliable and a score close to 0.5 or below the model is predicting at random (not reliable). Defined as:

$$C - index = \frac{1}{|P|} \sum_{(i,j) \in P} 1\left[\hat{h}_i > \hat{h}_j\right]$$

Where:

- P is the set of all comparable patient pairs (i.e., where one patient experiences an event before the other),
- $\hat{h}_i$ is the predicted risk or hazard for patient i.
- $1[\cdot]$ is the indicator function returning 1 if true and 0 otherwise.

The time brier score evaluates the accuracy of probabilistic predictions. It is a sum of both a calibration component and a discrimination component, with lower scores indicating improved model accuracy.

Defined as:

$$Bs(t) = \frac{1}{N} \sum_{i=1}^{N} -w_i(t) \cdot \left(\hat{s}_i(t) - \delta_i(t)\right)^2$$

Where:

- N is the number of individuals,
- $\hat{s}_i(t)$ is the predicted survival probability for individual i at time t,
- $\delta_i(t)$ is the event indicator (1 if event occurred before time t, 0 otherwise),

- $w_i(t)$ is a weighting function to handle censoring, typically using Inverse Probability of Censoring Weighting (IPCW).

4.5.8 Interpretability Methods

Stepwise Backward Selection

In stepwise backward selection, the RSF model starts with all variables included and then systematically removes variables that are found to be less important. The method works by fitting a model with all variables and then sequentially removing variables that have the least impact on the model's predictive power until only the most important variables are left. Feature permutation was the first interpretation method to determine the risk factors of incident HF in patients with type 2 diabetes. This method was carried out by measuring how the model score decreases when a feature is not available, thus the drop in the model score is indicative of how much the model depends on the feature.

Shapely Additive Explanations

To overcome these limitations, an advanced machine learning interpretability: shapely additive explanations (SHAP) was applied. SHAP is a better alternative to feature permutation, based on the magnitude of feature attributions rather than the decrease in model performance. Each variable is measured independently, avoiding collinearity (highly correlated variables). SHAP identified risk factors of incident HF. A numerical value is assigned to each feature that represents its contribution to the predicted outcome. The direction is a positive or negative value, indicating whether the factor increases or decreases the risk of incident HF. The process was repeated until the desired level of model performance was achieved.

An extended feature of SHAP called the "Tree Explainer" supports tree-based machine learning models. TreeExplainer provides an understanding of the global (entire dataset) model structure based on many local (individual) explanations. Local explanations represent an overview of each patient profile as shown in **Figure 17** compared to traditional black-box prediction (Louhichi et al., 2023). Local explanations are critical for personalised insights, especially understanding the reasoning behind a risk score for an individual patient.

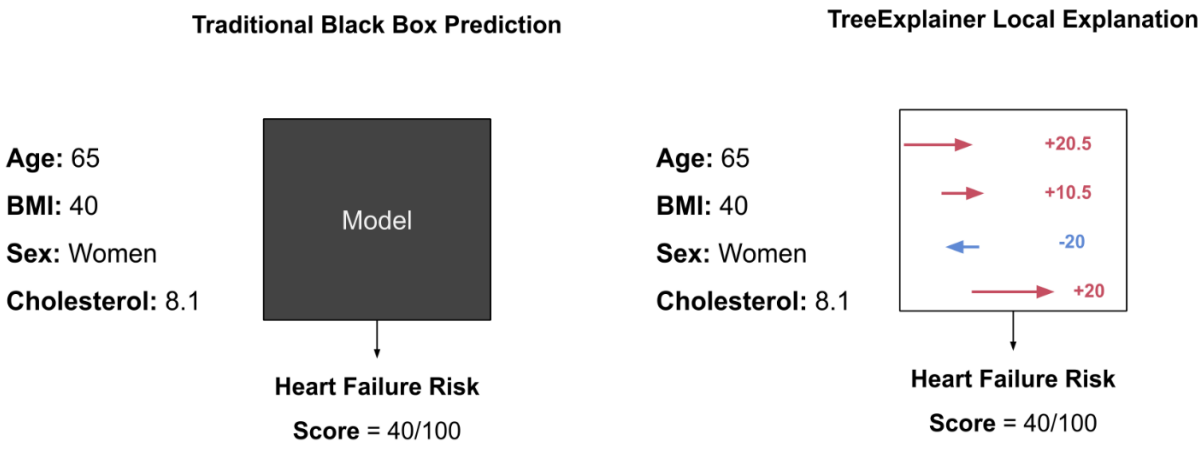


Figure 17 SHAP analysis for local explanation (for individual patient)

4.6 Results

4.6.1 Baseline Characteristics of T2DM Patients in Glasgow for Incident Heart Failure

Of 29,868 patients with T2DM aged ≥50 years, the median [quartiles] age was 63 [56 to 70] years, 45% were women, 88% were White, 30% were current smokers, 41% were in the most deprived quintile of the Scottish population and 5,367 (18%) had established atherosclerotic heart disease (AHD). Of those with BMI measured, 50% were classified as obese or severely obese. As expected, patients had high blood glucose concentrations and most had an elevated HbA1c (42-47 mmol/mol indicates pre-diabetes or well-managed diabetes; >4.8 mmol/mol is considered diagnostic for diabetes). On average serum cholesterol and triglycerides were not elevated, perhaps because many patients were on lipid lowering therapies. Many patients were on anti-hypertensive agents but the dataset does not include blood pressure. On average, renal function was impaired with most patients being in chronic kidney disease (CKD) stage III (eGFR 30-59 ml/min/1.73m^2). Liver function tests, haemoglobin, white cell counts were normal, on average. The baseline population characteristics are shown in **Table 20**.

Table 20 Incident Heart Failure in Patients with T2DM prescribed or not prescribed loop diuretics.

Demographics at Baseline (%) or Median (25/75)	Overall Glasgow Population N= 29,868	Not Prescribed Loop Diuretics at time of enrolment N = 23,546	Prescribed Loop Diuretics at time of enrolment N = 6,322	P-value
Age, (y)	63 (56 – 70)	62 (56 – 69)	67 (60 – 74)	<0.001
Sex				<0.001
Men	16,335 (55%)	13,495 (57%)	2,840 (45%)	
Women	13,533 (45%)	10,051 (43%)	3,482 (55%)	
Ethnicity				<0.001
White	26,332 (88%)	20,617 (88%)	5,715 (90%)	
Asian	1,707 (6%)	1,440 (6%)	267 (4%)	
Other	1,537 (5%)	1,251 (5%)	286 (5%)	
Unknown	292 (1%)	238 (1%)	54 (1%)	
***Body Mass Index (BMI)**	30 (27 – 34)	30 (27 – 34)	31 (27 – 37)	
BMI Classification				
Normal	4,324 (14%)	3,468 (15%)	856 (14%)	
Overweight	10,241 (34%)	8,337(35%)	1,904 (30%)	
Obese	12,659 (42%)	9,899 (42%)	2,760 (44%)	
Severely Obese	2,384 (8%)	1,660 (7%)	724 (12%)	
Underweight	260 (1%)	182 (1%)	78 (1%)	
***Current Smoker (yes)**	9,061 (30%)	7,141 (30%)	1,920 (30%)	<0.001
Socioeconomic Status (SIMD)				0.04
Quintile 1 – Most Deprived	12,009 (41%)	9,315 (40%)	2,784 (44%)	
Quintile 2	5,515 (18%)	4,345 (18%)	1,170 (19%)	
Quintile 3	4,084 (14%)	2,823 (14%)	884 (14%)	
Quintile 4	3,539 (12%)	3,200 (12%)	768 (12%)	
Quintile 5 – Least Deprived	4,631 (16%)	3,863 (16%)	764 (11%)	
Comorbidities n(%)				
Atherosclerotic Heart Disease (yes)	5,367 (18%)	3,748 (16%)	1,619 (26%)	<0.001
Angina (yes)	4,156 (14%)	2,831 (12%)	1,325 (14%)	0.12
Atrial Fibrillation (yes)	3,573 (12%)	2,058 (9%)	1,515 (12%)	<0.001
Chronic Obstructive Pulmonary Disease (yes)	2,865 (9%)	1,742 (7%)	1,123 (10%)	<0.001
Chronic Kidney Disease (yes)	1,457 (5%)	730 (3%)	727 (6%)	<0.001
Hyperkalaemia (yes)	1,452 (5%)	789 (3%)	663 (5%)	0.03
***Hypertension (Primary Care) (yes)**	18,251 (60%)	14,086 (60%)	4,165 (66%)	<0.001
Myocardial Infarction (yes)	3,241 (11%)	2,196 (9%)	1,045 (17%)	<0.001
Peripheral Artery Disease (yes)	933 (3%)	581 (2%)	352 (6%)	0.73
Stroke/TIA (yes)	2,678 (9%)	1,877 (8%)	801 (13%)	<0.001
Lab Tests * 6 months prior to or upon a Diabetes Diagnosis				
Plasma Glucose (mmoL)	8.8 (6.7 – 10.1)	9.1 (7.1 – 10.4)	9 (6.9 – 10.8)	0.04
Haemoglobin A1C (mmol/mol)	54 (46– 63)	54 (46– 63)	54 (46 – 63)	0.28
Haemoglobin (g/L)				<0.001
Men	139 (134 – 152)	140 (136 – 152)	136 (129 – 149)	
Woman	136 (124 – 140)	126 (136 – 140)	118 (131 – 137)	
Total Cholesterol (mmol)	4.1 (3.7-5.0)	4.2 (3.7 – 5.1)	4.0 (3.6 – 4.8)	0.14

Triglycerides (mmol)	1.8 (1.3 – 2.3)	1.8 (1.3 – 2.3)	1.8 (1.5 – 2.3)	<0.001
Serum Albumin (g/L)	38 (36 – 40)	38 (36 – 40)	37 (34 - 39)	<0.001
eGFR (mL/min/1.73m2)	53 (43 – 60)	53 (44 - 60)	49 (39 – 58)	<0.001
Alanine Transaminase – ALT (U/L)	23 (16 – 31)	24 (17 – 32)	20 (14 - 27)	<0.001
Aspartate Transaminase – AST (U/L)	21 (17 – 26)	21 (17 – 26)	19 (16 – 24)	<0.001
Alkaline Phosphatase (U/L)	88 (71 – 104)	87 (71 – 102)	91 (72-108)	<0.001
Neutrophils (x10^9/L)	5.0 (4.1 – 6.0)	5.1 (4.0 – 5.7)	5.0 (3.8 – 5.5)	<0.001
Lymphocytes (x10^9/L)	2.0 (2.2 – 2.3)	2.0 (1.8 - 2.2)	2.0 (1.6 – 2.4)	<0.001
Bilirubin (µmol/L)	10 (7 – 13)	10 (8 - 13)	10 (7 – 13)	<0.001
Potassium (mmol)	4.3 (4.0 – 4.6)	4.3 (4.0 – 4.6)	4.3 (4.0 – 4.7)	<0.001
Medications +/- 180 days of diabetes diagnosis, n (%)				
Metformin (yes)	10,006 (32%)	7,140 (30%)	2,866 (45%)	0.26
DPP4i (yes)	4,155 (14%)	3,100 (13%)	1,055 (17%)	0.18
Insulin (taken with Glucose-lowering Drug)	3,058 (10%)	2,109 (7%)	949 (13%)	<0.001
Sulphonylureas (yes)	6,825 (22%)	4,872 (21%)	1,980 (31%)	0.72
SGTL2i (yes)	3,681 (12%)	3,093 (13%)	588 (9%)	<0.001
Statins (yes)	14,678 (48%)	10,271 (44%)	4,407 (70%)	<0.001
Beta Blockers (yes)	6,498 (21%)	4,616 (19%)	1,882 (30%)	<0.001
ACEi or ARBS (yes)	12,737 (63%)	8,693 (59%)	4,044 (78%)	0.11
MRAs (yes)	1,354 (4%)	477 (2%)	877 (14%)	<0.001
Calcium Channel Blockers	2,344 (8%)	956 (4%)	1,388 (22%)	<0.001
Antiplatelets (yes)	6, 852 (22%)	4,872 (21%)	1,980 (31%)	0.72
Anticoagulants (yes)	2,344 (8%)	956 (4%)	1,388 (22%)	0.09
Thiazides (yes)#	7,823 (26%)	5,818 (25%)	2,005 (32%)	<0.001

***Primary Care utilises patient READ CODES**

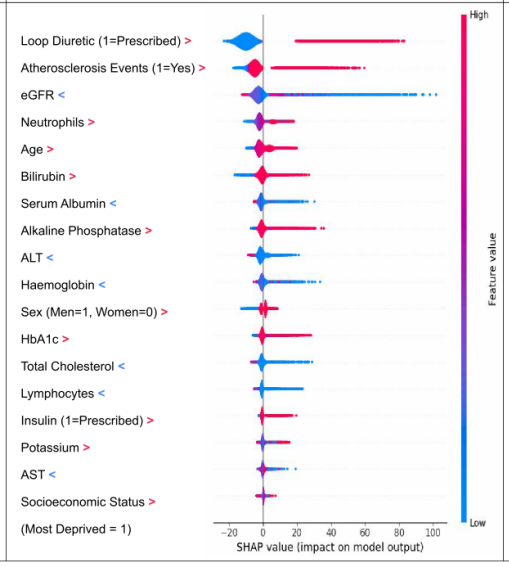
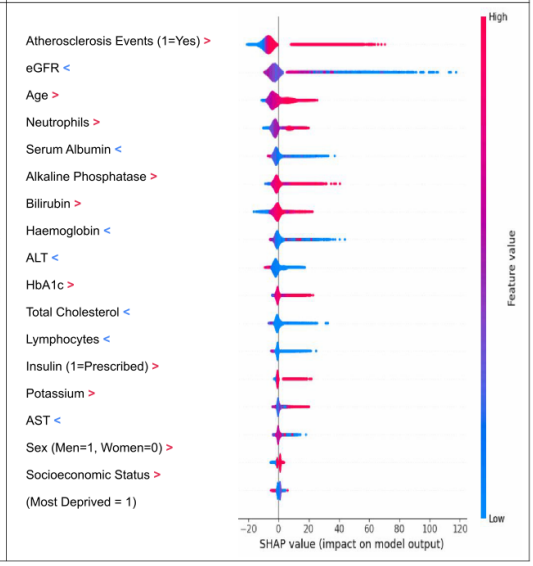
~ It is likely that most patients were switched from thiazide to loop diuretics in the preceding month rather than taking loop and thiazide diuretics at the same time as this is an extremely powerful combination that would probably not be tolerated by patients unless they had end-stage renal disease or severe heart failure.

4.6.2 Incident Heart Failure Risk Prediction Model(s)

Table 21 presents the key factors predicting incident HF in the first RSF prediction model, which included use of loop diuretics, history of atherosclerosis (myocardial infarction, angina, stroke, peripheral artery disease), estimated glomerular function ratio (eGFR), higher neutrophil counts (suggesting inflammation) and older age. Socioeconomic status of those most deprived also contributed to the risk prediction. Model interpretation was carried out utilising shapely values. The absolute SHAP value shows us how much a single feature affected the prediction displayed on the x-axis. It takes the mean average value for each feature. Here, all the values on the left represent the observations that shift the predicted value in the negative direction while the points on the right contribute to shifting the prediction in a positive direction. All the features are on the left y-axis. The arrows in red show that increasing values of a variable are associated with HF and arrows in blue show that decreasing values of a variable are associated with HF.

The second model excluded the strongest risk predictor in the initial analysis, loop diuretics, for two reasons. Firstly, loop diuretics might be considered to indicate undiagnosed HF rather than being a predictor of incident disease. Secondly, eliminating a strong predictor might reveal new predictors that had some association with loop diuretics. Model performance fell substantially to 0.79 (C-index) when loop diuretics were excluded from the model but no new strong predictor was identified.

Table 21 Random Forest Survival Baseline Model Results including and excluding Loop diuretics

	Model 1 (N=29,868) Including Loop Diuretics **C-index Score:** 0.83 ***Time Brier Score:** 0.02	**Model 2 (N=29,868) Excluding Loop Diuretics** **C-index Score:** 0.79 ***Time Brier Score:** 0.03
Key predictors of HF in rank order. **The x-axis shows how much each factor contributes to predict incident HF.** **>** Indicates increasing values of this variable associated with developing HF **<** Indicates decreasing values of this variable associated with developing HF ***C-index: Model discrimination.** ***Time Brier: Measure of calibration for time-dependent models. A lower Brier score indicates greater accuracy.**		

4.6.3 Advanced Cox Regression – Elastic Net

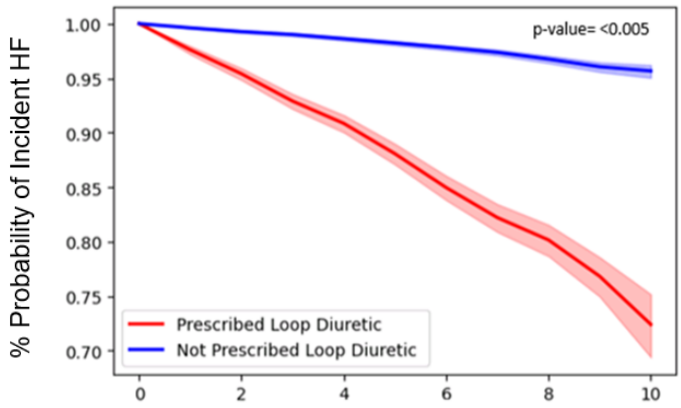
The Cox Elastic Net model automatically selects the most important variables for predicting the development of HF, while down weighting the influence of less important variables. **Table 22** presents the results of an Elastic Net regression model used to predict HF. The model's C-index score was 0.82, indicating strong discriminative ability, and a Time Brier score of 0.02, demonstrating excellent calibration and predictive accuracy over time. The top ten predictors of incident HF are listed in rank order based on their contribution to the model. Variables marked with a red arrow indicate that higher values are associated with an increased risk of heart failure, while variables with a blue arrow indicate that lower values are associated with increased risk.

Table 22 Elastic Net Model Validation

	Cox Elastic Net regression: Baseline Model (N=28,868)
C-index Score	0.82
Time Brier Score	0.02
Key predictors of HF in rank order. > Indicates increasing values of this variable associated with developing HF < Indicates decreasing values of this variable associated with developing HF	1. Loop Diuretic (1=Prescribed) > 2. Age > 3. Atherosclerosis Events (1=Yes) > 4. eGFR < 5. HbA1c > 6. Serum Albumin < 7. Haemoglobin < 8. Alkaline Phosphatase > 9. Neutrophils > 10. Bilirubin >

4.6.4 Kaplan Meier & Cox Proportional Hazards

Figure 18 shows that patients with T2DM prescribed a loop diuretic at baseline (in red) had a higher probability of receiving a diagnosis of H F. In a Kaplan-Meier analysis, the number of individuals at risk of experiencing a HF event may change over time due to censoring. When an individual is censored before experiencing a HF event, they are removed from the at-risk population at the time of censoring.



Patients At Risk (Baseline)						
Prescribed Loop Diuretic	6,316	4,873	3,336	2,081	921	0
Not Prescribed Loop Diuretic	23,537	19,344	14,523	9,525	4,284	0

Figure 18 Kaplan Meier for patients with T2DM prescribed or Not prescribed Loop diuretic

The Cox proportional hazards model investigates associations between the prescription of loop diuretics and the risk of incident HF, shown in **Table 23**. Men have a 48% higher risk of a diagnosis of HF compared to women (HR = 1.48, 95% CI 1.30 to 1.69; $p < 0.005$). For older age, the risk of heart failure increases by 4% per year (HR = 1.04; 95% CI 1.04 to 1.05; $p < 0.005$). Individuals prescribed loop diuretics had a 5-fold higher risk of incident HF than those who were not (HR: adjusted for age and sex 5.40 [95% CI 4.72 – 6.17 (<0.005)]). Note that there was not an early stepwise increase in the diagnosis of heart failure in those receiving loop diuretics at baseline that might be expected if loop diuretics had a close temporal association with a diagnosis of heart failure. This observation neither supports nor refutes the concept that loop diuretic use is merely a marker for a missed diagnosis of heart failure.

Table 23 Cox proportional hazards Model to investigate associations between the prescription of Loop diuretics

Cox Proportional Hazards Model: Investigating Loop diuretics					
Variable	**Coefficient (coef)**	**Hazard Ratio (exp(coef))**	**95% CI (HR)**	**Z-value**	**P-value**
Sex	0.39	1.48	1.30 - 1.69	6.01	< 0.005
Age	0.04	1.04	1.04 - 1.05	11.51	< 0.005
Loop Diuretic	1.69	5.40	4.72 - 6.17	24.73	< 0.005

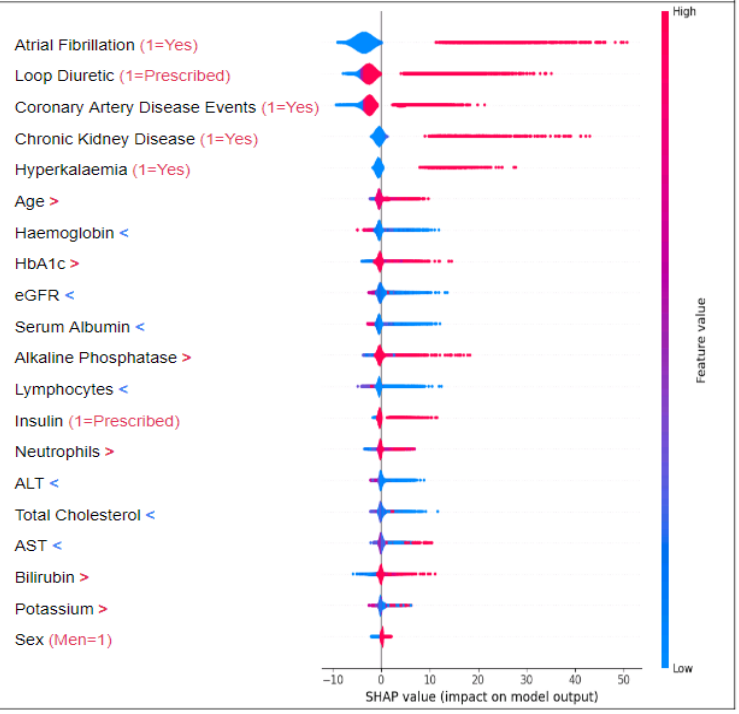
4.6.5 Investigation of Comorbidities

Finally, the initial random survival forest model was re-run (**Table 24)**. Additional comorbidities, including atrial fibrillation (AF), coronary artery disease events (ICD-10 codes: myocardial infarction, angina, ischemic heart disease) and chronic kidney disease (CKD), were included. These conditions are recognised contributors to the pathophysiology of HF, either through direct cardiac involvement or through indirect effects on the cardiovascular system. The inclusion of these variables improved the model's overall performance, as evidenced by a higher concordance index and lower Brier score, signifying better risk discrimination and calibration.

Of note CKD and low eGFR both predicted a greater risk of incident HF. From a clinical perspective, CKD indicates a low eGFR, but clinicians vary on which threshold to use. Some will use <30 and others <60 mL/min/1.73m^2 and others will also take urinary albumin excretion into account. Accordingly, a generalisable model should logically prefer use of eGFR and albuminuria rather than be subject to the vagaries of subjective clinical classification.

Most patients did not have AF, represented by the large blue shape. A minority of patients had atrial fibrillation and for some it was a strong predictor of risk (large positive SHAP value). AF was a strong predictor of incident HF. AF may be associated with the initiation of loop diuretics even in the absence of a diagnosis of HF (Zakeri et al., 2021a) which may partially account for its strong predictive association with incident HF. AF may precipitate congestion, leading to the initiation of loop diuretics but only some patients receive an immediate diagnosis of HF, whilst it is delayed in others.

Table 24 Additional Comorbidities included in the Random Survival Forest Baseline Model

	Model 3 (N=29,868) Investigating Comorbidities - Random Survival Forest **C-index Score:** 0.87 ***Time Brier Score:** 0.02
Key predictors of HF in rank order. **The x-axis shows how much each factor contributes to predict incident HF.** > Indicates increasing values of this variable associated with developing HF < Indicates decreasing values of this variable associated with developing HF ***C-index: Model discrimination.** ***Time Brier: Measure of calibration for time-dependent models. A lower Brier score indicates greater accuracy.**	

4.7 Discussion

This analysis suggests that loop diuretics are a key risk marker for developing HF. Congestion (excess water and salt retention) due to cardiac dysfunction is a key feature of the heart failure syndrome and cause of symptoms such as exertional breathlessness and ankle swelling. Loop diuretics, which promote water and sodium excretion by the kidneys, are the most important treatment for congestion, relieving symptoms and signs (Friday et al., 2024). If the reasons for initiating loop diuretics are not investigated, the diagnosis of HF may be missed. Patients treated with loop diuretics have a higher mortality, whether or not they receive a diagnosis of HF. This might be because the diagnosis of HF, a condition with a prognosis worse than many cancers, has been missed. Without the correct diagnosis, treatment is unlikely to be optimal. However, loop diuretics also increase urinary potassium loss, impair renal function and activate the renin-angiotensin system, which could drive the progression of cardiovascular disease and increase the risk of sudden death. Loop diuretics may serve as both a marker and a potential contributor to HF progression. Their use often reflects a more advanced stage of HF, even without a formal diagnosis (Mullens et al., 2019; Felker et al., 2020).

This study included loop diuretics as a risk predictor for incident heart failure, however the role of loop diuretics in the casual pathway remains debatable, as they may indicate undiagnosed heart failure rather than serve as an independent predictor. To provide further insights into loop diuretics as a risk factor or risk marker for incident HF, the random survival forest model was re-run excluding this variable. Exclusion of loop diuretics reduced the model's overall predictive performance. Although other predictors, such as other adverse CVD events, increased in importance, none fully compensated for the exclusion of loop diuretics and no new variable strongly associated with outcome was identified.

This analysis included patients with multiple chronic conditions to assess the impact of comorbidities on the risk of developing HF. By focusing on individuals with a wide range of comorbidities—such as atrial fibrillation, hypertension, hyperkalaemia, chronic kidney disease and adverse cardiovascular events—the research captured the cumulative burden these conditions impose on cardiovascular health. The risk of developing heart failure increased with each additional chronic condition. This highlights the need for integrated care for patients with multi-morbidity.

The higher cumulative incidence of HF in those prescribed loop diuretics in this study aligns with existing knowledge about the complex interplay between diuretic use, kidney function and HF risk.

Patients with kidney dysfunction (eGFR <30ml/min) had a much higher risk of developing HF, especially if prescribed loop diuretics (Mullens et al., 2019; Felker et al., 2020). The adverse effect of loop diuretics on kidney function may increase with long-term use, or when patients are dehydrated, are on other medications that can impair kidney function and in those who already have poor kidney function (Guo et al., 2023). Obesity may also be an under-recognised risk factor for chronic kidney disease (CKD) (Verma et al., 2023), perhaps because obesity is associated with poorer glycaemic control, hypertension, dyslipidaemia and low-level chronic inflammation. This analysis shows that most patients with T2DM in GG&C are over-weight, obese or severely obese (**Chapter 6.7**).

The relationship between HF and impaired kidney function is well-established. Studies like the ALLHAT trial (Khayyat-Kholghi et al., 2021) demonstrate that worsening kidney function, indicated by a decline in estimated glomerular filtration rate (eGFR), is an important contributor to the development of HF. As the heart weakens, its ability to maintain adequate blood flow to the kidneys diminishes, leading to impaired kidney function. This reduction in renal perfusion exacerbates the cycle of worsening heart failure, as the kidneys' ability to excrete sodium and water becomes compromised, contributing to fluid overload and further stressing the heart. Rapid declines in kidney function have been identified as strong predictors of incident heart failure, even in individuals who initially present with normal renal function (Bueno Junior et al., 2023). This highlights the importance of recognising kidney impairment as a risk factor for developing heart failure.

Identifying heart failure development in patients with T2DM requires time-to-event modelling. It accommodates right-censoring, time-dependent risks and interactions between clinical factors. The Elastic Net linear regression model highlighted a small number of key predictors of incident HF. This method was model-specific and does not fully account for the multitude of contributing factors involved in HF development. While powerful in its simplicity and clarity, it may overlook some of the complex interactions between variables in large datasets. Elastic Net assumes linear relationships only.

The RSF model approach implemented in this study excels in integrating and processing a large number of variables, including those that may interact in complex and non-linear ways (Miao et al., 2015), without the need for manual selection, allowing it to capture a broader and potentially more accurate picture of risk. It identified patterns and relationships between variables that may not be immediately apparent. This makes the RSF model valuable in clinical settings where the richness of data can be fully leveraged to improve predictive accuracy.

Traditional risk score models, such as the Framingham Heart Failure Risk Score or Systematic COronary Risk Evaluation 2 (SCORE2) (Shahlan Kasim et al., 2023), require manual input of a limited set of predefined risk factors. These models focus on a narrow range of variables, such as age, blood pressure, and cholesterol levels. This may not account for the full spectrum of clinical data available in a patient's medical record. Subsequently, the Cox regression models and multivariable techniques, including Elastic Net, do not perform well in this context due to issues with collinearity, which compromises the reliability of variable selection and inflates variance estimates. In contrast, our approach focuses on the interpretability of the most important risk predictors leveraging the full breadth of clinical data stored in EMRs. As the RSF model is a tree-based approach, with the use of SHAP, the model risk prediction score is able to show the factors making the greatest contribution to the development of heart failure, improving on the limited interpretability of traditional models.

4.8 Conclusion

Overall, this analysis identifies key risk factors (or risk markers) for the development of incident HF in patients with T2DM. The strongest marker was treatment with a LD. It is possible that many patients treated with LD have a missed diagnosis of HF but it is also possible that LD drive the progression of disease by activating the renin-angiotensin and other neuro-endocrine systems and by causing renal and possibly vascular dysfunction. By excluding loop diuretics from the predictive model, the overall model performance decreased, confirming its central role as a risk factor or risk marker of incident HF. This analysis also emphasised the importance of managing multiple chronic conditions, with a focus on patients with comorbidities such as chronic kidney disease, atrial fibrillation, and coronary artery disease. The RSF model excelled in capturing interactions among variables and contributed to a more accurate and thorough assessment of HF risk.

Chapter 5 "Predicting incident Heart Failure in Type-2 diabetes Mellitus: External Validation using EMRs in Hong Kong"

Abstract

Introduction: The incidence of heart failure (HF) is higher among people with type-2 diabetes mellitus (T2DM) compared to the general population. Symptoms and signs of HF often go unrecognised until severe. Novel machine-learning (ML) tools may help clinicians to identify patients with T2DM who are at high risk of developing HF or already have unrecognised HF.

Methods: We obtained electronic medical records (EMRs) from a diverse population (Hong Kong), including demographics, medical history, blood and urine test results and medications. Incident HF was defined as a primary or contributory diagnosis of HF using International Classification of Diseases, 9th and 10th Revision codes. We implemented time-dependent machine-learning models to predict incident HF. We integrated state-of-the-art artificial intelligence interpretability with clinical expertise to provide concise reasoning for a patient's increased risk of HF. We carried out propensity score matching and inverse probability weighting to investigate causality.

Results: Of 262,687 patients aged ⩾50 years with T2DM, 8,515 (3%) had incident HF between 2009-19, of whom 3,142 (37%) had prior coronary artery disease (CAD). For incident HF, the baseline model c-statistic and time brier scores were 0.88 and 0.07 respectively (with 1.0 and 0.0 being perfect scores). Important predictors were treatment with loop diuretics, insulin, lower serum albumin, haemoglobin, lymphocyte counts and eGFR and higher serum potassium, total cholesterol, neutrophil counts, and alkaline phosphatase. Those prescribed loop diuretics had a substantially higher incidence of heart failure (HR adjusted for age and sex: 4.68 [95% CI 4.47 – 4.90, $p < 0.005$]).

Conclusion: Models using readily available patient information predict the risk of incident HF for patients with T2DM in Hong Kong. The results are very similar to analyses of EMR from patients with T2DM in Scotland, which differs markedly in terms of culture, climate and ethnicity, suggesting that these results might be generalisable to diverse populations.

5.1 Introduction

Heart failure (HF) usually goes unrecognised until symptoms are severe. There is a growing awareness that HF is a common complication of T2DM (Ziaeian and Fonarow, 2016). Identifying markers that either anticipate the development of HF or alert clinicians to its existence is critical for improving early detection and management. There are few reports investigating the incidence HF in Chinese populations (Wang et al., 2021; Fu et al., 2023) with T2DM but there is a high prevalences of T2DM in Hong Kong, with low public awareness of HF (Leung et al., 2015a; Fan et al., 2022). For this analysis, routinely collected EMRs were obtained from a large population of patients with T2DM in Hong Kong to carry out external validation. Utilising another diverse T2DM cohort will strengthen this research by applying survival-based machine learning risk prediction and interpretation from the previous chapter.

5.2 Aim

This chapter performs external validation of the incident heart failure risk prediction model using a diverse population from Hong Kong, achieving robustness and generalisability.

5.3 Data Sources

Records of patients aged ≥50 years with any new diagnosis of T2DM between January 1st 2009 and December 31st 2019 were obtained from the Clinical Data Analysis and Reporting System (CDARS), which is a territory-wide electronic health database operated by the Hospital Authority (HA) of Hong Kong. Incident cases of HF were identified based on a first hospitalisation with HF recorded in any diagnostic position using the International Classification of Diseases, 9th Revision (ICD-9) codes.

5.4 Patient Information

From the CDARS record, demographic information, comorbidities, defined by ICD-9 codes, recorded at the time or prior to diagnosis, including hypertension, chronic kidney disease (CKD), coronary artery disease (CAD), myocardial infarction (MI), peripheral artery disease (PAD), atrial fibrillation (AF), chronic obstructive pulmonary disease (COPD) and stroke. Smoking was poorly recorded (~10%) due to the low prevalence in Hong Kong.

Results of routinely collected blood and urine tests were included in the analysis. The first available result in the 6 months after diagnosis was used where available but, if not available, a result from the 12 months prior to diagnosis could be used. Tests included haemoglobin A1c, haemoglobin, lipid panel (total cholesterol and triglycerides), renal panel (serum potassium, creatinine and albumin, estimated glomerular filtration rate (eGFR)), liver function (ALT, AST alkaline phosphatase and bilirubin), neutrophils, lymphocytes and potassium.

(Note that these tests may not be specific for a particular condition, for instance a low serum albumin can reflect general illness or specific liver or renal disease and alkaline phosphatase can reflect bone as well as liver disease). Treatments for diabetes or cardiovascular and lipid-lowering medications used at any time within the 6 months after diagnosis of T2DM were included.

Figure 1 shows the consort diagram for this study. Overall, 273,876 patients with a new diagnosis of T2DM were identified in public hospital or clinic records over a period of 11 years. Of these, 11,189 had a prior history of HF (look-back period of 5 years) and were excluded from analyses of incident HF, leaving 2,687 patients eligible for analysis. Only 23,973 patients had a record of body mass index (BMI) available and therefore BMI was not included in the primary analysis.

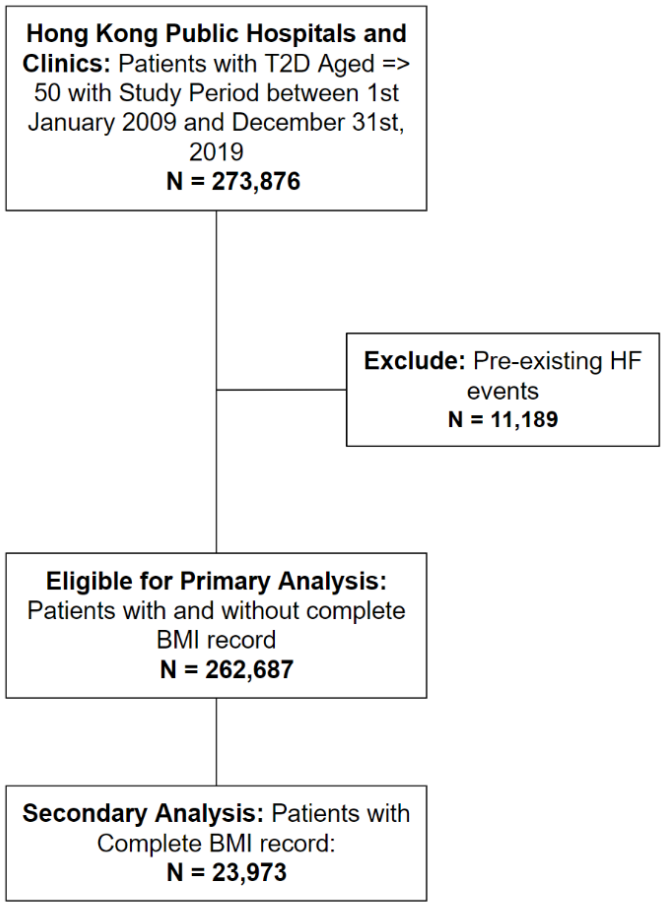


Figure 19 Consort Diagram showing selection of patients from Hong Kong with T2DM to predict incident HF

5.5 Methods

The methods established in Chapter 4 were adapted and applied to analyse the outcome of incident heart failure in the Hong Kong population, to validate the reliability and robustness of findings from the Greater Glasgow & Clyde (GG&C) population.

5.5.1 Clinical Variables

The clinical variables from section 4.6.2 chosen by clinical experts on heart failure and T2DM for the GG&C population analysis were used, whether or not they were associated with incident HF to allow comparisons across different populations (see Chapter 7).

5.5.2 Survival Analysis – Kaplan Meier

A Cox Proportional Hazards model was applied adjusting for age, sex and the most important risk predictors. For further insight patients were investigate by age group for the following: prescribed LD with HF, prescribed LD with HF, not prescribed LD with HF and not prescribed LD without HF (Neither) .

5.5.3 Development of Random Survival Forest for Incident Heart Failure Risk Overtime

The RSF model(s) with interpretation utilised in chapter 4 was implemented for comparison. Furthermore, a second RSF model was implemented with the application of propensity score matching. This was to reduce confounding by balancing characteristics between patients prescribed LD and not prescribed LD.

Propensity Score Matching (PSM)

Objective: To estimate the causal effect of loop diuretics on incident heart failure by patients prescribed and not prescribed LD with similar propensity scores to reduce confounding.

- PSM is a method used in an attempt to reduce bias by matching treated and control cases based on their propensity scores (Deb et al., 2016).
- Firstly, the propensity score for each patient is estimated using a logistic regression model, which predicts the likelihood of being prescribed loop diuretics based on baseline characteristics. These scores were then used to match patients for incident heart failure prediction:

$$logit\big(P(T = 1|X)\big) = \beta_{0^{+}1}x_1 + \beta_2 x_2 + \cdots + \beta_k X_K$$

Where:

- T = Treatment indicator (1 if prescribed loop diuretics, 0 if not prescribed).
- X = Baseline covariates that may be associated with loop diuretic use and risk of developing heart failure, such as age, sex, eGFR and prior CVD conditions.
- β = Coefficients estimated from the logistic regression model.

The propensity score P is the probability of being prescribed loop diuretics calculated as (using logistic regression):

$$P(T = 1|X) = \frac{e\beta_0 + \beta_1 X_1 + \beta_2 X_2 \ + \cdots + B_k X_k}{1 + \ e\beta_0 + \beta_1 X_1 + \beta_2 X_2 \ + \cdots + B_k X_k}$$

- $P(T = 1|X)$ is the **propensity score**, i.e., the probability of receiving loop diuretics T=1 given the covariates X
- e is the base of the natural logarithm (approximately 2.71828)

By using these scores for matching, the study ensures that patients prescribed loop diuretics and those not prescribed are balanced in terms of their baseline characteristics.

Matching Procedure

Once propensity scores are calculated, patients prescribed loop diuretics are matched with patients not prescribed based on their scores. The matching is carried out with Nearest-Neighbour (NN) using the Euclidean distance:

$$d(i,j) = |\ P(x_i) - P\big(x_j\big)|$$

- $P(x_i)$ is the propensity score for individual i in the prescribed group.
- $P\big(x_j\big)$ is the propensity score for individual j in the not prescribed group.
- $d(i,j)$ is the distance between the two individuals' scores.

The individual from the not prescribed group closest to the prescribed individual is selected as the nearest neighbour. In other terms, for each patient prescribed loop diuretics, find a patient who is not prescribed loop diuretics with the closest propensity score **Figure 20** illustrates a high-level example of the NN matching, where patient profiles are matched based on characteristics and propensity score (patient A 0.75 and patient B 0.78).

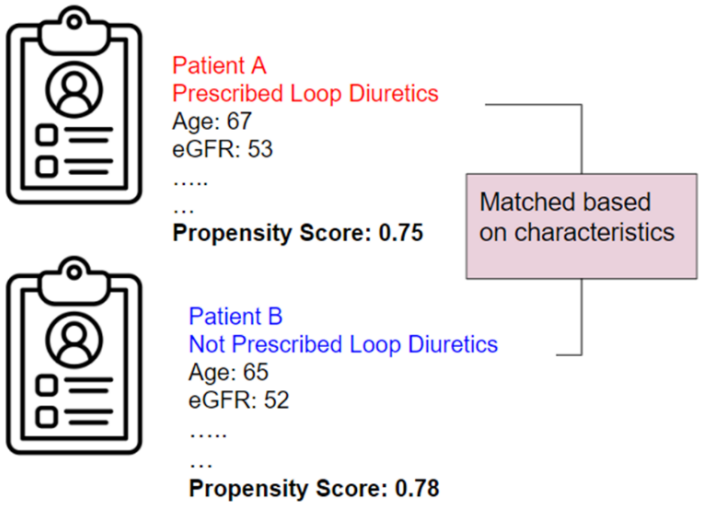


Figure 20 Nearest Neighbour Matching of Patients

Checking Balance of Covariates (patient characteristics)

The matched dataset is structured by the standardised mean difference (SMD) which measures covariate balance between groups before and after propensity score matching. SMD assesses the balance and compare groups. It ensures that matched individuals have similar distributions of baseline characteristics. Additionally, to evaluate balance between the treated and untreated groups prior to weighting, a Kernel Density Estimation (KDE) plot was generated to compare their propensity score distributions. Finally, the RSF model was implemented and shapely values were applied to the matched dataset.

Inverse Probability Weighting (IPW)

IPW was also applied in the RSF model to further improve confounding in the prediction of incident heart failure. IPW involves assigning weights to each individual based on the inverse of the probability of receiving loop diuretics. This method helps balance covariates across groups. The inclusion of IPW ensures a more robust and unbiased prediction of HF risk, accounting for any imbalances in covariates across the Hong Kong T2DM population.

5.6 Results

5.6.1 Baseline Characteristics of T2DM Patients in Hong Kong for Incident Heart Failure

Of 262,687 patients aged ⩾50 years with T2DM, 8,515 (3%) developed new-onset HF between 2009-19, of whom 3,142 (37%) had prior coronary artery disease. Patients prescribed loop diuretics were older, more likely to be women, more likely to have CKD, had lower haemoglobin and higher neutrophils counts (**Table 25:** Note that patients with a diagnosis of HF at baseline were excluded from this analysis**)**.

Table 25 Incident Heart Failure in Hong Kong Patients with T2DM prescribed or not prescribed loop diuretics

(%) or Median (25/75)	Overall Hong Kong Population N= 262,687	Not Prescribed Loop Diuretics N = 210,066	Prescribed Loop Diuretics N = 52,621	P-value
Age, (y)	65 (56 – 75)	63 (55 – 73)	72 (63 – 79)	<0.001
Sex				<0.001
Men	126,843 (48%)	13,495 (49%)	24,645 (47%)	
Women	135,844 (52%)	10,051 (51%)	27,976 (53%)	
Ethnicity				<0.001
Asian (Chinese)	26,332 (92%)	193,260 (92%)	48,412 (92%)	
Other	23,642 (8%)	16,806 (8%)	4,208 (8%)	
***Body Mass Index (BMI)**	25 (23 – 26)	24 (23 – 26)	24 (23 – 27)	
Comorbidities n(%)				
Coronary Heart Disease (yes)	20,566 (8%)	13,064 (6%)	7,502 (14%)	<0.001
Atrial Fibrillation (yes)	4,852 (2%)	2,684 (1%)	2,168 (4%)	<0.001
Chronic Obstructive Pulmonary Disease (yes)	629 (0.2%)	328 (0.15%)	301 (0.5%)	<0.001
Chronic Kidney Disease (yes)	2,635 (1%)	1,078 (0.5%)	1,557 (3%)	<0.001
Hypertension (yes)	57,000 (22%)	38,132 (18%)	18,868 (36%)	<0.001
Peripheral Artery Disease (yes)	256 (0.1%)	136 (0.06%)	129 (0.02%)	0.73
Stroke/TIA (yes)	8,017 (3%)	5,517 (3%)	2,500 (5%)	<0.001
Lab Tests within 6 months of Inclusion				
Haemoglobin A1C (mmoL)	56 (51– 63)	56 (51– 62)	56 (50 – 63)	0.28
Haemoglobin (g/L)				<0.001
Men	131 (132 – 139)	131 (134 – 139)	130 (129 – 138)	
Women	129 (122 – 136)	128 (123 – 135)	126 (118 – 134)	
Total Cholesterol (mmol)	4.8 (4.3 – 5.2)	4.7 (4.2 – 5.1)	4.7 (4.1 – 5.1)	0.14

Triglycerides (mmol)	1.5 (1.1 – 1.9)	1.4 (1.2 – 1.9)	1.1 (1.5 – 2.1)	<0.001
Serum Albumin (g/L)	40 (38 – 42)	40 (38 – 42)	37 (40 - 42)	<0.001
Serum Creatinine	85 (73 – 98)	85 (73 – 97)	87 (75 – 101)	
eGFR (mL/min/1.73m2)	60 (51 – 72)	61 (53 – 73)	59 (51 – 70)	<0.001
Alanine Transaminase – ALT (U/L)	23 (18 – 31)	24 (18 – 32)	21 (15 - 27)	<0.001
Aspartate Transaminase – AST (U/L)	25 (21 – 36)	24 (20 – 30)	25 (21 – 28)	<0.001
Alkaline Phosphatase (U/L)	74 (65 – 86)	73 (65 – 83)	76 (66 – 90)	<0.001
Neutrophils (x10^9/L)	5.3 (4.4 – 7.1)	5.2 (4.4 – 6.9)	5.5 (4.5 – 9.8)	<0.001
Lymphocytes (x10^9/L)	1.9 (1.6 – 2.4)	1.9 (1.6- 2.3)	1.9 (1.6 – 2.9)	<0.001
Potassium (mmol)	4.2 (4.0 – 4.4)	4.2 (4.0 – 4.4)	4.5 (4.2 – 4.5)	<0.001
Bilirubin (µmol/L)	10.3 (9.2 – 12.7)	9.6 (9.2 – 12.4)	11.7 (9.2 – 12.9)	<0.001
Medications within 6 months of inclusion, n (%)				
Metformin (yes)	180,516 (69%)	145,712 (69%)	34,804 (66%)	0.26
DPP4i (yes)	311 (0.1%)	247 (0.1%)	64 (0.1%)	0.18
Insulin (taken with Glucose-lowering Drug)	26,573 (10%)	16,690 (8%)	9,9883 (19%)	<0.001
Sulphonylureas (yes)	165,856 (63%)	129,936 (62%)	35,920 (68%)	0.72
Statins (yes)	56,254 (21%)	40,001 (19%)	16,253 (31%)	<0.001
Beta Blockers (yes)	86,302 (33%)	63,241 (30%)	23,061 (44%)	<0.001
ACEi or ARBS (yes)	113,426 (57%)	84,230 (40%)	29,196 (55%)	0.11
Calcium Channel Blockers	103,617 (40%)	74,998 (36%)	28,619 (54%)	<0.001
Thiazides (yes)	43,912 (17%)	26,782 (13%)	17,130 (33%)*	<0.001

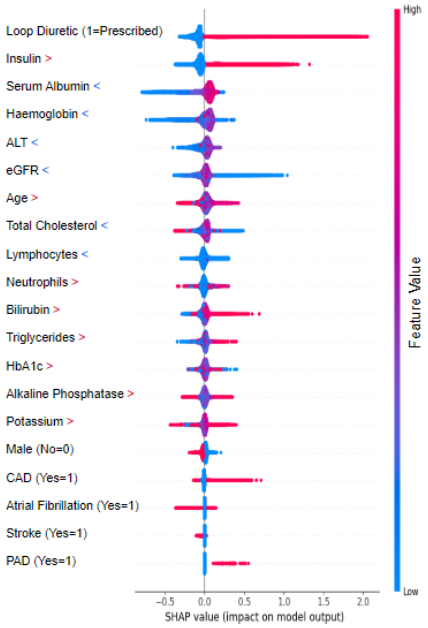
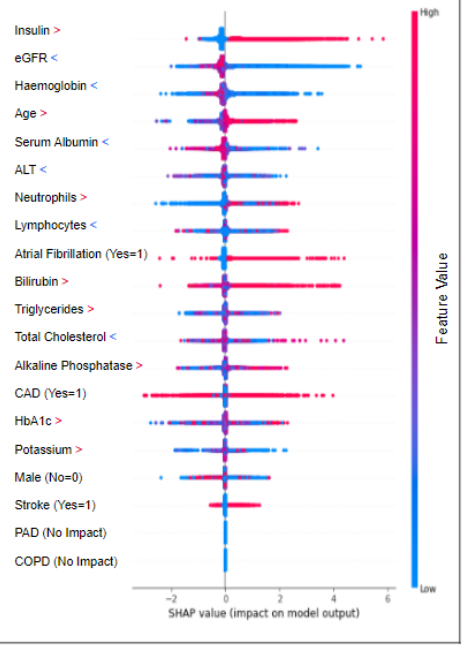
5.6.2 Incident Heart Failure Risk Prediction Model(s)

Table 26 shows variables strongly associated with incident HF in the first RSF prediction model. Older patients, treatment with loop diuretics or insulin, lower serum albumin, ALT, haemoglobin and eGFR, a history of CAD, atrial fibrillation, stroke and peripheral artery disease were all associated with an increased risk of a new diagnosis of heart failure with a C-index of 0.88 and time-brier score of 0.02.

The second model applied the RSF model with causal inference on a dataset matched by propensity scores for the prescription of loop diuretics (N=105,242). Key predictors of incident HF were use of insulin, lower eGFR, haemoglobin, serum albumin and ALT, atrial fibrillation, CAD (which included MI) and stroke. Model performance decreased to 0.85 (C-index) and time brier of 0.07. Survival analysis integrated evaluation metrics are described in chapter 4.

Modelling was repeated excluding loop diuretic as a variable, in order to balance the confounders between patients who were prescribed loop diuretics and those who were not. Figure 21 and 22 shows the application of inverse probability weighting (IPW). The X-axis represents the propensity scores. The scores range from 0 to 1. A score of 0 means there is a 0% probability that the patient receiving loop diuretics. A score of 1 means there is 100% probability of the patient receiving loop diuretics. The Y-axis represents the density of subjects at each propensity score. The values on the Y-axis are densities. Density represents the distribution of the data. The analysis of patients with complete BMI is shown in **Appendix B1**.

Table 26 Results for predicting Incident HF with and without Causal inference

Hong Kong Cohort	Random Survival Forest without Causal Inference (N=262,687)	Random Survival Forest with Causal Inference (matched_dataset according to Loop Diuretics Prescribed or Not) (N=105,242)
*C-statistic Score	0.88	0.85
*Time Brier Score	0.02	0.07
Key predictors of HF in rank order. **The x-axis shows how much each factor contributes to predict incident HF.** > Indicates higher values of this variable increase risk of HF < Indicates lower values of this variable increase risk of HF ***C-index: Model discrimination.** ***Time Brier: Measure of calibration for time-dependent models. A lower Brier score indicates greater accuracy.**		

Propensity Score Distribution Before Weighting:

The treated group (red) has a higher concentration of patients with propensity scores close to 1.0, indicating that these individuals had a high likelihood of being prescribed loop diuretics based on their baseline covariates. Conversely, the untreated group (blue) had a higher density of patients with propensity scores near 0, showing that they were less likely to receive the treatment. Limited overlap between the red and blue curves, indicating an imbalance between the treated and untreated groups before weighting. The untreated group has a higher concentration of patients at lower propensity scores, while the treated group is more spread out. This imbalance shows that before weighting, the treated and untreated groups had differences in baseline characteristics, which could confound the analysis of covariates on incident heart failure. Density curves may extend slightly below zero due to kernel density estimation artifacts. All true propensity scores in the analysis lie between 0 and 1.

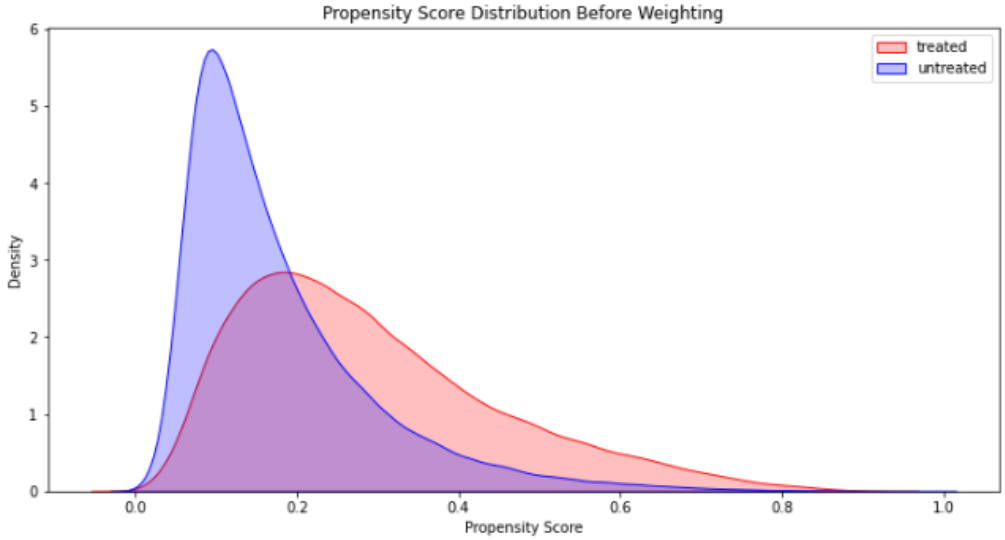


Figure 21 Propensity Score Distribution Before Weighting

After applying IPTW the distribution of propensity scores for both the treated (red) and untreated (blue) groups did not improve. This indicates that the weighting did not successfully adjust for the differences in baseline covariates between the two groups. This suggests that balance has not been achieved to estimate the effect of baseline covariates on incident heart failure without the confounding effect of loop diuretics prescription.

Propensity Score Distribution After Weighting:

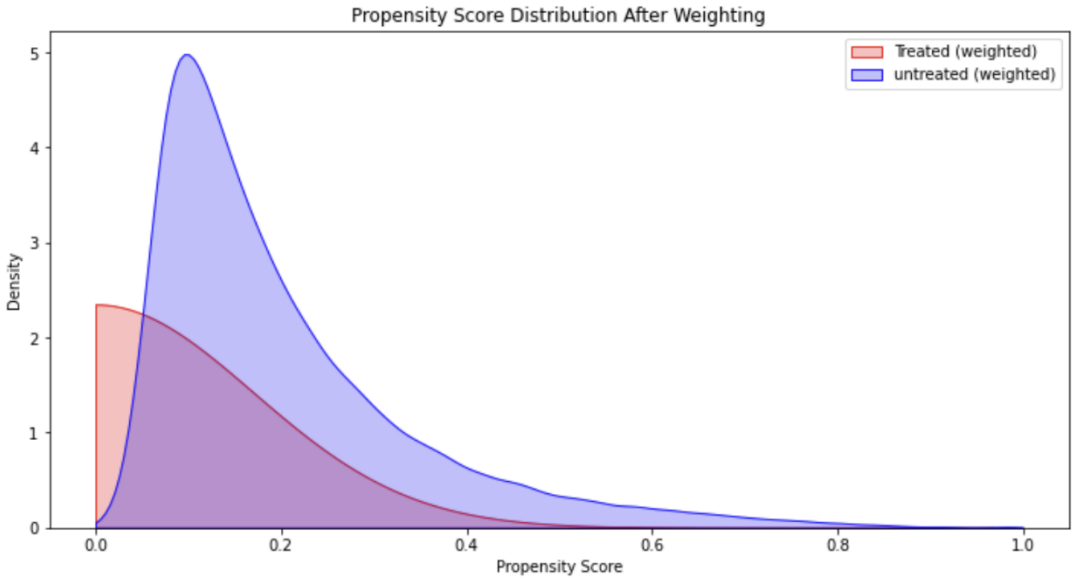
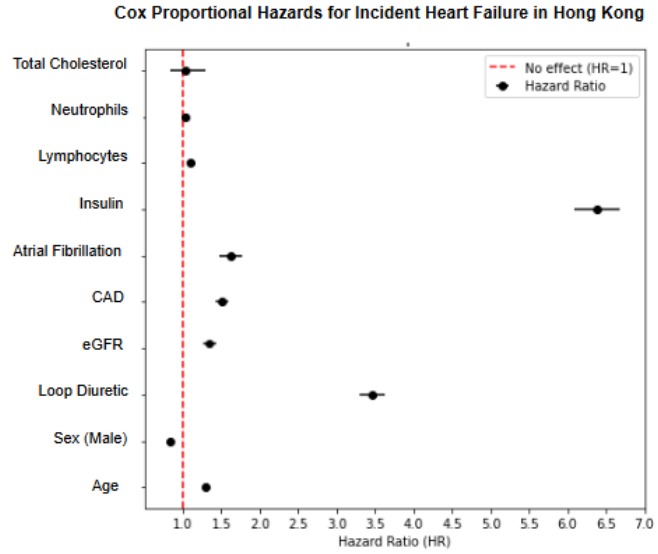


Figure 22 Propensity Score Distribution After Weighting

5.6.3 Cox Proportional Hazards & Age Groups

Figure 23 Cox Proportional Hazards model confirmed the association of key risk factors for incident heart failure in patients with type 2 diabetes. The highest risk factors included insulin use, which strongly increased the risk of heart failure (HR = 6.38, $p < 0.001$), and loop diuretic use, which was associated with more than three times the risk. Moderate risk factors included age, with older age significantly increasing the risk by quartile (HR = 1.34, $p < 0.001$), atrial fibrillation by 62%, coronary artery disease (HR = 1.51, $p < 0.001$) Lower eGFR, which was associated with a 48% increased risk. Low-risk factors included lymphocyte and neutrophil counts in quartiles, which had minimal impact on risk (HR = 1.01 and 1.00, respectively, $p < 0.001$) and total cholesterol, which showed no significant effect on heart failure risk.



Results Table of Cox Proportional Hazards				
Variables *In Quartiles	Coefficient (coef)	Hazard Ratio (exp(coef))	95% CI (HR)	p-value
*Total Cholesterol (mmol)	0.01	0.98	0.98 - 1.01	0.50
*Neutrophils (x10^9/L)	0.08	1.09	1.07 - 1.11	p < 0.001
*Lymphocytes (x10^9/L)	0.08	1.01	1.01 - 1.01	p < 0.001
Insulin (Prescribed)	1.85	6.38	6.09 - 6.67	p < 0.001
Atrial Fibrillation (Event)	0.48	1.62	1.48 - 1.78	p < 0.001
CAD (Event)	0.41	1.51	1.42 - 1.59	p < 0.001
*eGFR (mL/min/1.73m2)	0.38	1.48	1.40 - 1.52	p < 0.001
Loop Diuretic (Prescribed)	1.24	3.46	3.30 - 3.62	p < 0.001
Sex (Male)	-0.17	0.84	0.81 - 0.88	p < 0.001
*Age (Years)	0.30	1.34	1.31 - 1.38	p < 0.001

Figure 23 Cox Proportional Hazards Model

The bar chart in **Figure 24** illustrates the percentage of loop diuretic LD usage with and without HF across different age categories. In those aged <55 years, most patients did not have HF and were not prescribed LD. Similarly, in those aged 55-64 years, 84.9% had neither condition, while 13% were prescribed LD without HF. As age increases, the proportion of patients prescribed LD without HF rises, with 21.2% in the 65-75 group and 29.4% in those aged 75 and above. In contrast, the percentage of patients with HF, whether prescribed LD or not, remains low across all age categories, peaking in the 75+ group where 3.5% were prescribed LD with HF and 2.2% were not prescribed LD despite having HF. Overall, LD usage without HF increases with age, especially in those over 75.

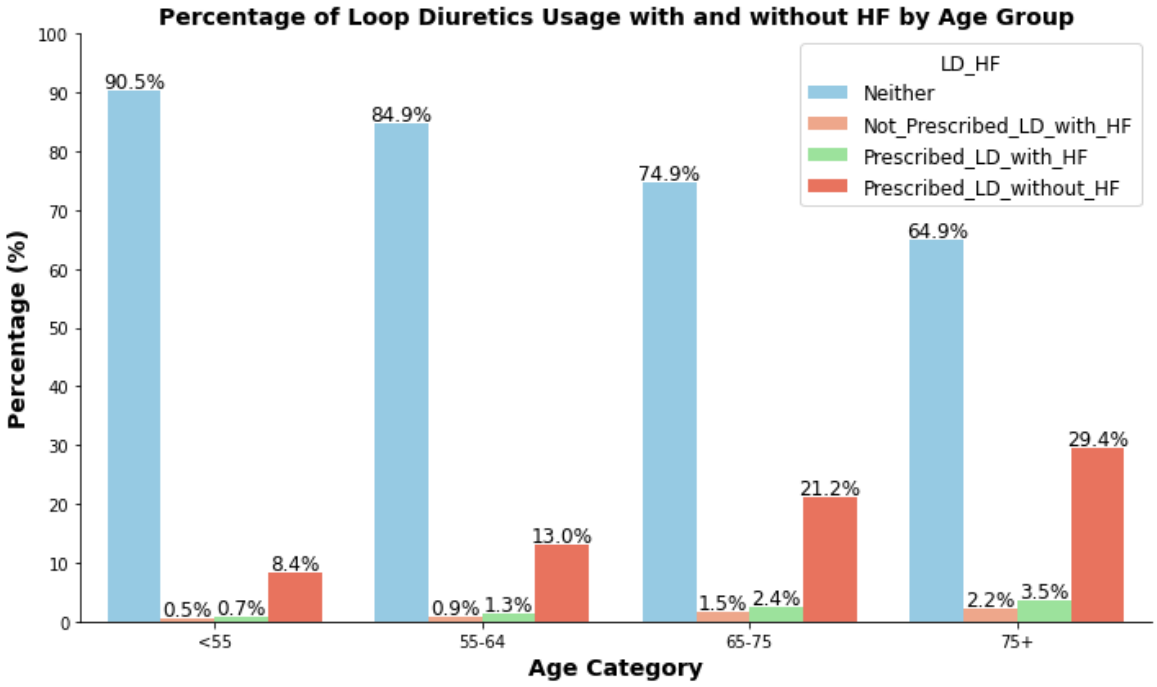


Figure 24 Loop Diuretics Usage by Age Group Bar Chart

5.6.4 Novel HF Risk Assessment Support Tool

Figure 25 shows a prototype interface for individual patient risk assessment in support of a precision medicine approach to patient care. A patient's demographic profile is shown with updated information from EMR including prescriptions and hospitalisations. The reasoning for risk of incident HF is derived from section 4.5.5 utilising shapely values, highlighting the risk factors making the greatest contribution for each individual patient. If a patient's risk score is 45.2, it means that, based on the aggregation of SHAP values from various clinical factors, the patient has a 45.2% chance of developing heart failure within the prediction window of 5 years. This score combines the contribution of all individual risk factors (prescribed loop diuretic, hyperkalaemia, etc.), illustrating the patient's risk. By using the RSF model, the survival probabilities are extracted at 5 years and converted into absolute risk (1 – survival probability) and then these values are normalised to a 0–100 scale for easier interpretation.

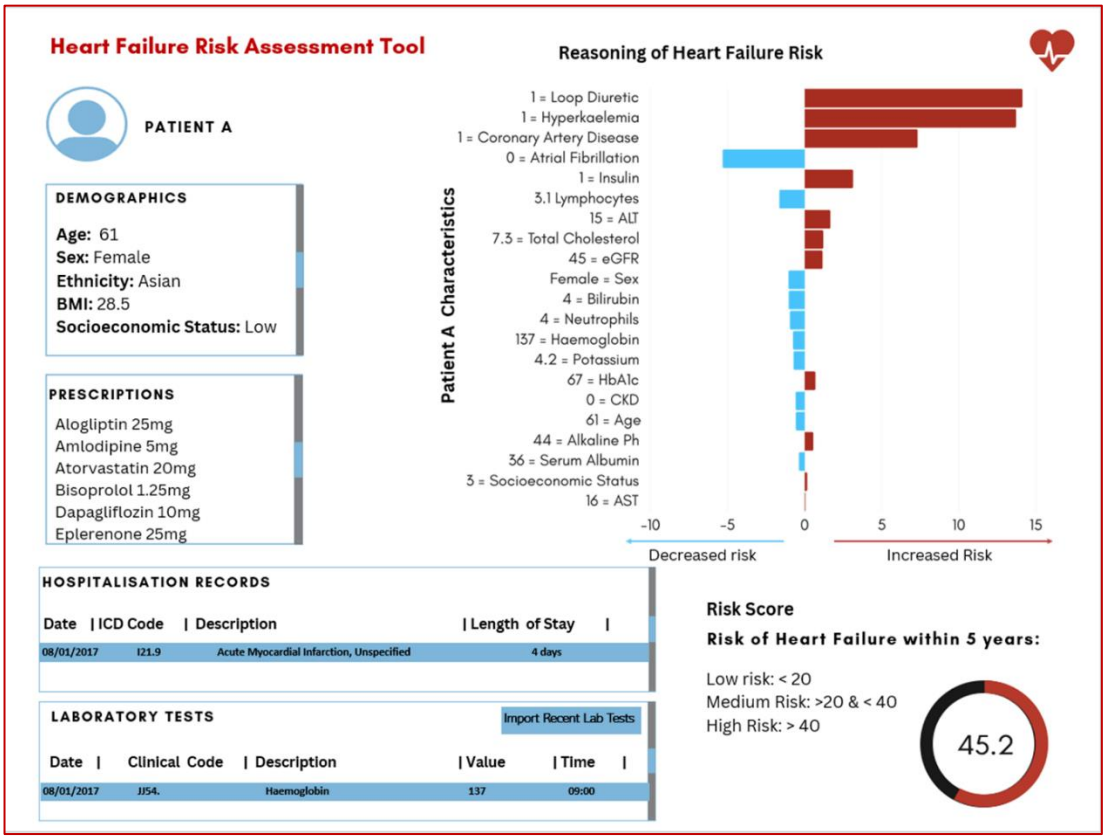


Figure 25 Interface for HF Risk Assessment Tool

5.7 Discussion

This chapter provides external validation of the support tool for predicting incident heart failure in Glasow, UK. By investigating diverse EMRs from the Hong Kong, China T2DM population, results were validated and causality was introduced into the risk prediction models.

Firstly, the analysis of incident HF in a T2DM Hong Kong population showed that many patient charactistics consistent with those observed in the Glasgow population. In both populations, women were more frequently prescribed loop diuretics, which is also confirmed in a propensity-matched cohort study in the UK (Cuthbert et al., 2024). However, there is no evidence of loop diuretics usage in Hong Kong studies. This raises important clinical concerns regarding misdiagnoses or over-prescription, especially in older patients or those with mild renal impairment. As evidence is a concern, results from this study highlighted individuals prescribed loop diuretics were more likely to have experienced CKD events. Even when the prescribed population was smaller, those receiving loop diuretics had a higher percentage of CKD events compared to those who were not prescribed loop diuretics.

Few observational studies have found associations between diuretic use and decline in eGFR in patients with pre-existing CKD (Fitzpatrick et al., 2022; Emmens et al., 2022). An example of this is illustrated in the Epidemiology of Acute Kidney Injury in 3,044,244 Chinese retrospective study (L. Zhou et al., 2022). Authors found that exposure to loop diuretics was associated with a significantly increased risk of hospital-acquired acute kidney injury, with a hazard ratio (HR) of 1.61 (95% CI, 1.55–1.67). This consistency in the prescribing patterns of loop diuretics, as well as the associated higher rates of CKD, suggests common clinical approaches and outcomes across different healthcare systems and ethnicities. The inclusion of a different ethnicity-based population in the risk prediction model was critical. Ethnicity plays an important role in heart failure development, many times this is overlooked or undervalued in studies, despite ethnic disparities in CVD risk (van Apeldoorn et al., 2024b). Other times, clinical risk prediction models are built solely for one population or utilising limited clinical trial datasets. Incident heart failure rates vary even among different Asian ethnicity sub-groups (Lam et al., 2016; Cheng et al., 2024).

Therefore, this research supports the need to investigate diverse patient populations and adapt to precision medicine approaches. Such evidence-based support tools are essential, especially in cases where certain ethnicities may otherwise be underrepresented or excluded.

Subsequently, the clinical variable selection remained consistent, allowing for direct comparisons. This promotes replicability (Wang et al., 2022): the ability to confirm results in different populations and improves the generalisability of the support tool. Majority patient characteristics were similar in both populations, where differences were in the order of importance. Interestingly, a key risk factor in this study was insulin use. Insulin is prescribed in combination with other anti-diabetic therapies in patients with T2DM. It causes sodium retention and hypoglycaemia (Cosmi et al., 2018), both of which may contribute to adverse outcomes. Hypoglycaemia induces stress on the cardiovascular system through mechanisms such as increased sympathetic nervous activity and elevated heart rate.

This relationship was previously explored in a large-scale analysis (Cosmi et al., 2018) involving two datasets: one from randomised trials comprising 24,012 HF patients and another from an administrative database of 4 million individuals, including 103,857 with HF. By applying propensity score matching, resulted showed that patients prescribed insulin had a significantly higher risk of all-cause mortality and HF hospitalisation compared to those not using insulin. Moreover, the prescription of insulin is given to T2DM patients when they cannot produce enough insulin known are insulin resistance. Another study found that elevated fasting insulin levels were significantly associated with an increased risk of incident HF (Banerjee et al., 2013), independent of other cardiovascular risk factors. Of 4,425 participants, a 10% higher risk of HF was observed per standard deviation increase in fasting insulin. The study also linked higher fasting insulin levels to structural heart abnormalities, such as increased left atrial size and left ventricular mass. However, these previous studies were limited to Cox proportional hazards (CPH) method and propensity score matching for risk estimation. In contrast, our approach overcomes the limitations of CPH, providing a more robust prediction of incident HF. This allows for enhanced evidence-based clinical decision-making.

The results from the first RSF risk prediction model highlight several clinical factors, consistent with the Glasgow Cohort, strongly associated with an increased risk of incident heart failure, including older age, comorbidities such as CAD, atrial fibrillation and stroke, as well as lower levels of serum albumin, ALT, haemoglobin and eGFR. These findings show the importance of monitoring these risk factors in diverse patients, as they contribute to the development of heart failure.

Recently, PSM is used to address biases in observational data, reduce confounding and strengthen the causal interpretation of the findings. Some diabetes patients related studies in Hong Kong apply PSM with the CDARS datasets (Lee et al., 2022; J. Zhou, Lee, Liu, et al., 2022; J. Zhou, Lee, Lakhani, et al., 2022). These studies illustrate the usefulness of PSM in some clinical use cases.

This study utilised PSM by matching patients with similar characteristics of those who were prescribed loop diuretics or not. The aim was for a clearer estimation of the causal impact of loop diuretics on incident HF risk. Results compared to the initial RSF model without PSM decreased in performance, bringing many concerns about the reliability of PSM. This approach tries to mimic the conditions of a randomised controlled clinical trial (RCT), by matching patients based on clinical characteristics of those prescribed or not prescribed loop diuretics, reducing sample size. Clinicians may debate that simulated patient matching does not fully capture the complexities of patient symptoms and signs in clinical practice settings. PSM may be somewhat valuable in this study, but not clinically complete (Reiffel, 2020).

To address the artificial matching of patients, this study also applied IPW which assigns weights to each patient based on the inverse probability of receiving loop diuretics or not. This technique avoids the exclusion of unmatched patients and enhances precision estimates (Vock et al., 2016). Evidently, even after applying IPW results still illustrate the imbalance in the distribution of propensity scores. The weighting adjusts for the treatment allocation by giving more weight to individuals who are underrepresented in the treated or untreated groups. However, this method still highlights a key issue: patients are treated for specific reasons, and the propensity scores reflect those underlying treatment decisions. This imbalance in the weighted distribution suggests that the treated group is still different from the untreated group, despite the IPW weighting.

Furthermore, applying PSM and IPW to study the effect of covariates on incident heart failure is problematic because it does not account for immortal time bias. A more appropriate approach would involve studying incident loop diuretic use in patients before they were treated. However, this would introduce its own challenges, such as immortal time bias, making the current method less valid. While the method is valuable for understanding treatment effects in certain settings, it may not be suitable for analysing covariate effects on incident heart failure in this case.

To further assess the relationship of loop diuretics may be a misdiagnosis of heart failure, the outcome shifted to focus on incident loop diuretics use with and without heart failure in **Appendix B2**. This change was crucial in confirming that the same risk factors were involved in both outcomes. Specifically, the model can effectively predict the development of heart failure and identify individuals who are at risk before they reach a stage where they require diuretic therapy. This early identification can support clinicians intervene sooner, preventing further deterioration in patient health.

Most importantly, using the Hong Kong cohort for external validation, allows a more defined assessment of HF risk for each patient. This approach supports the emerging application of precision medicine. The novel support tool developed in this chapter tells a patient's story with information from EMRs and provides an individualised risk assessment. It is a user-friendly risk assessment tool for clinicians, ensuring they can quickly access patient information, which can help with decision-making. With automation of direct data-capture from EMRs, manual input of data is not required. The ability to directly utilise EMRs ensures that the information is current.

5.8 Conclusion

Diverse Hong Kong EMRs improve generalisability and contributes to a robust incident heart failure risk prediction support tool. This research also highlights the importance of including diverse ethnic populations in predictive models, as shown by the consistency of results between the Hong Kong and Glasgow populations. The robust methodology ensures effective application across diverse populations. The model accounts for key patient characteristics, such as the higher prescription rates of loop diuretics among women and their association with chronic kidney disease events. However, PSM was not able to reduce confounding in assessing the true effects of risk factors on incident heart failure in those prescribed and not prescribed loop diuretics. Overall, this research supports precision medicine approaches and emphasises the need for transparent, reliable tools in clinical practice for heart failure prevention and management. The novel support tool developed in this analysis could help clinicians to enhance personalised patient care by harnessing the full potential of clinical data stored in EMRs.

Chapter 6 "Risk Stratification of Socioeconomic Groups in West of Scotland to predict Mortality "

Abstract

Background: Patients with type 2 diabetes mellitus (T2DM) have a reduced life-expectancy that may be made worse by socioeconomic deprivation (N Kaur et al., 2023).

Aim: to investigate drivers of prognosis in people with T2DM according to socioeconomic status using conventional statistics and a state-of-the-art machine learning (ML) model.

Methods: We obtained routinely collected electronic medical records (EMR) for patients with T2DM aged ≥50 years from the National Health Service (NHS) Scotland. Using Cox proportional hazards and random survival forest models, we assessed variations in mortality amongst socioeconomic deprivation quintiles, as determined by the Scottish Index of Multiple Deprivation (SIMD). The Shapely Additive Explanations (SHAP) interpretability method was used to identify key prognostic factors associated with survival within each subgroup.

Results: of 46,031 people with a newly recorded diagnosis of T2DM between 2009-2019, 11,727 died within 10 years. Compared to those in the most affluent quintile, patients with T2DM in the most deprived quintile had a 36% higher mortality risk (adjusted HR (95%CI): 1.36, 1.24 – 1.50, $p < 0.005$). Prescription of loop diuretics, increasing age, decreasing serum albumin, alanine transaminase and worsening renal function (c-index 0.83, brier score 0.07) were associated with mortality across all quintiles. Chronic obstructive pulmonary disease strongly correlated with mortality in the most deprived quintile, strokes in the most affluent.

Conclusion: Greater socioeconomic deprivation is associated with a worse prognosis in patients with T2DM. Readily available clinical information such as age and treatment with loop diuretics, allied to commonly available blood test results, predict mortality risk across all deprivation groups for people with T2DM.

6.1 Introduction

Developing models to predict prognosis is important for several reasons. Modifiable characteristics that are strongly associated with outcome, such as high blood pressure or smoking, might be therapeutic targets. Identification of individuals at high risk of events can inform healthcare policy, potentially focussing resources on patients at high risk of events, for whom more expensive treatments might be more cost-effective. Predictive models are also useful for auditing the quality of care. Healthcare systems that deliver better or worse outcomes than predicted can be investigated to discover the cause, copying good practice and remedying poor practice.

T2DM is a well-established risk factor for cardiovascular morbidity and mortality (Liane Ong et al., 2023). Poor dietary choices, lack of exercise, tobacco use, high blood pressure, high cholesterol and impaired kidney function, contribute to the development of many chronic conditions, including T2DM. Lower socioeconomic status (SES) is associated with a less healthy lifestyle, lower educational opportunity and attainment, lower rates of employment and income, poorer housing and less access to healthcare, all of which may increase the risk of developing T2DM and exacerbate its adverse effect on health outcomes (Jackson et al., 2012),·(Kimenai et al., 2022; Moody et al., 2016; Schultz et al., 2018),·(Stringhini et al., 2013). Accordingly, it is important to include SES in predictive models, along with other potential predictors, to improve identification of high-risk individuals and to investigate possible interactions with other risk factors (Tan et al., 2020).

6.2 Aim

This thesis chapter investigates risk factors for all-cause mortality according to SES in patients with new-onset T2DM using a novel approach to identifying key factors.

6.3 Study Data

Scotland has developed and operationalised a system of regional "Safe Havens" that provide researchers with secure access to deidentified, routinely collected, NHS electronic medical records (EMR) linked to national prescribing, hospitalisation and death records and the Scottish Care Information: Diabetes (SCI-Diabetes) registry. Registration into SCI-Diabetes occurs automatically when a patient is assigned a Read Code [10] (a coded thesaurus of clinical terms used in the NHS since 1985) for diabetes mellitus in a primary or secondary care health care information system. The registry is estimated to capture over 99% of all patients assigned a diagnostic Read Code for Diabetes (Livingstone et al., 2012). Coding may be extended to include patients with a high blood glucose measurement or raised haemoglobin A1c, a measure of longer-term blood glucose control. Linking these various sources of data enables large-scale population-level studies.

6.4 Patient Information

EMR from the Greater Glasgow and Clyde (GG&C) population from 1st of January 2009 to 31st December 2019 was used to identify people aged ≥50 years with an incident diagnosis of T2DM during this period. Patient characteristics at the time of enrolment includes demographic details such as age, sex, socioeconomic status and ethnicity. Socioeconomic status was determined by a quintile score based on the Scottish Index of Multiple Deprivation (SIMD), with the first quintile containing individuals living in the most deprived areas (data zones) and the fifth quintile the least deprived for the entire Scottish population. SIMD measures the extent to which an area is deprived across seven domains: income, employment, education, health, access to services, crime and housing. The GG&C region has a disproportionately large number of residents in the most deprived quintile for the Scottish population.

Smoking status, diagnosis of hypertension and BMI were extracted from primary care read codes. Prevalent comorbidities (chronic kidney disease (CKD), chronic obstructive pulmonary disease (COPD), atherosclerotic heart disease, heart failure, hyperkalaemia, peripheral artery disease, stroke, myocardial infarction, atrial fibrillation, and angina) were defined by the International Classification of Diseases, 10th Revision (ICD-10) codes.

The following blood and urine test results closest to the date of enrolment in the SCI-Diabetes registry enrolment were identified: glucose, haemoglobin A1c, haemoglobin, total cholesterol, triglycerides, serum albumin, serum creatinine, estimated glomerular filtration rate (eGFR), urine albumin-to-creatinine ratio, potassium, lymphocytes, neutrophils, AST, ALT, alkaline phosphatase and bilirubin. Any of the following treatments dispensed within 6 months of inclusion in the SCI-Diabetes registry were recorded; treatments for diabetes, loop diuretics (repeat prescribing only unless death occurred within 90-days of first dispensing), cardiovascular and lipid-lowering medications.

Figure 26 shows the consort diagram for this study. After excluding duplicate entries, there were 47,396 unique patients with a diagnostic label of diabetes. The first record of T2DM was used as the incidence date except for 4,947 (9%) patients who had a test showing impaired glucose tolerance or raised fasting plasma glucose and subsequently received treatment for diabetes without a formal diagnosis of T2DM; for these patients the date of entry into the database was used as the incidence date. Patients treated solely with insulin (n=1,365) were excluded as these were likely to have a Type-1 Diabetes Mellitus, as it would be very unusual to treat new-onset T2DM with insulin without any oral therapy.

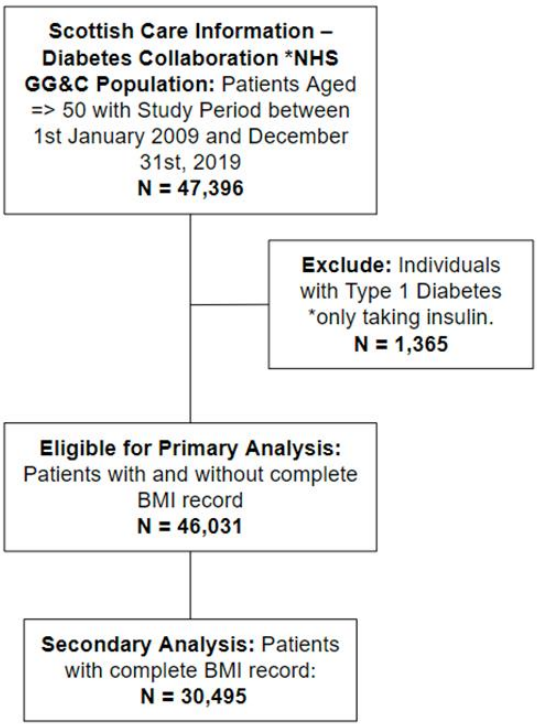


Figure 26 Consort Diagram for T2DM and Mortality in Glasgow

From the study period of 1st January 2009 and 31st December 2019, 46,031 patients were eligible for the primary analysis, including 30% with missing body mass index (BMI). A secondary analysis for 30,495 individuals with complete BMI records was done. The index date was the date of T2DM diagnosis; and the last day of follow-up was December 31st, 2019.

6.5 Methods

The methods established in Chapter 4 were adapted and applied to analyse the outcome of all-cause mortality. To investigate patients with different SES, the RSF model was implemented for the whole population and each SES group. This approach was selected due to its superior performance in capturing complex interactions between variables. Subsequently, the model(s) predictive ability was evaluated using survival-based performance metrics: C-index and the Time-dependent Brier Score. Time-dependent AUC measures the accuracy of the model's predicted probabilities of survival at specific time points. These metrics confirm whether or not the model reliably predicts survival outcomes across diverse patient populations. SHAP analysis was conducted to provide an understanding of feature importance in predicting incident heart failure. SHAP values highlight the variables most strongly associated with outcome. Kaplan Meier plots showing survival probability among BMI categories. A bar chart was used to show cause-specific mortality.

6.6 Results

6.6.1 Baseline Characteristics

We identified 46,031 people (median age of 64 years old, men: 54%) who had a newly recorded diagnosis of T2DM between 2009 and 2019. During a median of 4 years, 11,727 died.

Baseline demographics, blood and urine test results, medications, primary and secondary care diagnostic codes and death records are shown in **Table 27** (overall population with or without BMI). Overall, the median [IQR] age at diagnosis was 64 (57-72) years, with a slight preponderance of men (54%). Most were White, 20% were current smokers and 41% were in the most deprived SIMD quintile and only 15% were in the most affluent. Of those with a recorded BMI, 85% were overweight or obese. The most common cardiovascular condition was hypertension (41%), although this may have been an underestimate as about 70% of patients were taking anti-hypertensive medicines. The median [IQR] serum cholesterol was fairly low (4.1 (3.7-5.0) mmol/L), consistent with a substantial proportion receiving statins (52%). Most patients had impaired kidney function with a median [IQR] eGFR of 54 (44 - 61) mL/min/1.73m^2.

Only 70% of patients were prescribed a treatment for diabetes within 6 months of diagnosis, perhaps reflecting attempts to control blood sugar concentrations by diet and exercise alone. The prevalence of heart failure was 10% but 25% of patients were dispensed loop diuretics.

Table 28 shows the baseline characteristics of patients according to quintile of SIMD. Compared to those who were most affluent, patients with the most deprived SES were more likely to be women, were more likely to be current smokers and to have COPD, had a higher BMI and were more likely to be receiving loop diuretics. However, there was little difference in the age of onset of diabetes. Although variations in some other characteristics were highly statistically significant across quintiles, the absolute magnitude was often small, owing more to the size of the population rather than to clinical meaningful differences. Findings were similar for patients whose BMI was available **(Appendix C1)**. Compared to more affluent patients, those with the most deprived SES had a 36% higher risk of mortality (HR: adjusted for age and sex 1.36 [95% CI 1.24 – 1.50 (<0.005)]).

Table 27 Baseline characteristics of patients with T2DM with and without a BMI record

Demographics at Baseline (%) or Median (25/75)		Overall GG&C Population N=46,031	With BMI Record N=30,495	Without BMI Record N=15,536
Age, (y)		64 (57 – 72)	63 (56 – 70)	66 (58 – 75)
Sex	Men	24,664 (54%)	16,715 (55%)	7,949 (51%)
	Women	21,367 (46%)	13,780 (45%)	7,587 (49%)
Ethnicity	White	39,290 (85%)	26,897 (88%)	12,393 (80%)
	Asian	338 (5%)	1,725 (6%)	577 (4%)
	Other	593 (6%)	1,572 (5%)	1,050 (6%)
	Unknown	105 (4%)	301 (1%)	1,516 (10%)
Socioeconomic Status (SIMD)				
Quintile 1 – Most Deprived		18,517 (41%)	12,389 (41%)	6,128 (39%)
Quintile 2		8,355 (18%)	5,638 (18%)	2,727 (18%)
Quintile 3		6,360 (14%)	4,162 (13%)	2,198 (14%)
Quintile 4		5,643 (12%)	3,602 (12%)	2,041 (13%)
Quintile 5 – Least Deprived		7,156 (16%)	4,714 (15%)	2,442(16%)
***Body Mass Index (BMI)**		30 (27 – 34)	30 (27 – 34)	Missing
BMI Classification	**Normal**	4,408 (10%)	4,408 (14%)	
	Overweight	10,449 (23%)	10,449 (34%)	
	Obese	12,926 (28%)	12,926 (42%)	
	Severely Obese	2,444 (5%)	2,444 (8%)	
	Underweight	268 (1%)	268 (1%)	
	Unknown	15,536 (33%)	N/A	
***Current Smoker (yes)**		9,416 (20%)	9,259 (30%)	157 (1%)
Comorbidities n(%)				
Atherosclerotic Heart Disease (yes)		7,106 (16%)	5,595 (18%)	1,511 (10%)
Angina (yes)		5,621 (13%)	4,306 (8%)	1,315 (8%)
Atrial Fibrillation (yes)		6,083 (13%)	3,827 (13%)	2,256 (15%)
Chronic Obstructive Pulmonary Disease (yes)		4,395 (8%)	2,970 (10%)	1,425 (9%)
Chronic Kidney Disease (yes)		2,549 (6%)	1,543 (5%)	1,006 (6%)
Heart Failure (yes)		4,675 (10%)	2,852 (10%)	1,823 (12%)
Hyperkalaemia (yes)		2,336 (5%)	1,517 (5%)	819 (5%)
***Hypertension (Primary Care) (yes)**		18,999 (41%)	18,634 (61%)	365 (2%)
Myocardial Infarction (yes)		4,545 (11%)	3,358 (11%)	1,187 (8%)
Peripheral Artery Disease (yes)		1,650 (4%)	1,074 (4%)	576 (4%)
Stroke/TIA (yes)		4,010 (9%)	2,755 (10%)	1,255 (8%)
Plasma Glucose (mmol/L)		8.8(6.7 – 10.1)	9.1 (7.1 – 10.5)	7.9 (6.3 – 9.4)
Haemoglobin A1C (mmol/L)		55 (46– 61)	54 (46– 63)	56 (44– 56)
Haemoglobin (g/L)				
	Men	138 (134 – 151)	139 (134 – 152)	134 (129 – 149)
	Women	134 (123 – 139)	134 (124 – 140)	134 (120 – 138)
Total Cholesterol (mmol/L)		4.1 (3.7-5.0)	4.1 (3.7-5.0)	4.1 (3.9-5.1)
Triglycerides (mmol/L)		1.5 (1.0 – 2.2)	1.6 (1.1 – 2.3)	1.3 (1.1 – 2.4)
Serum Albumin (g/L)		37 (35-39)	38 (36 – 40)	36 (34-39)
eGFR (mL/min/1.73m^2)		54 (44 - 61)	55 (44 - 62)	53 (42-60)

Alanine Transaminase – ALT (U/L)	22 (16-30)	23 (15-31)	22 (15-28)
Aspartate Transaminase – AST (U/L)	20 (16-26)	20 (16-26)	21 (17-26)
Alkaline Phosphatase (U/L)	89 (72-105)	88 (71-105)	91 (72-108)
Neutrophils (x10^9/L)	5.0 (3.8-5.6)	5.1 (3.8-5.9)	5.0 (3.8 – 5.5)
Lymphocytes (x10^9/L)	2.0 (2-2.3)	2.0 (1.4-2.2)	2.0 (1.6-2.4)
Bilirubin (µmol/L)	10 (7 - 13)	10 (7 - 13)	11 (8 - 13)
Potassium (mmol/L)	4.3 (4.0 – 4.6)	4.3 (4.0 – 4.6)	4.2 (4.0 – 4.6)
Metformin (yes)	14,545 (32%)	10,266 (34%)	4,279 (28%)
DPP4i (yes)	5,033 (11%)	4,269 (14%)	764 (5%)
Insulin (taken with Glucose-lowering Therapies)	2,801 (6%)	1,708 (6%)	1,093 (7%)
Sulphonylureas (yes)	10,204 (22%)	7,046 (23%)	3,158 (20%)
SGTL2i (yes)	3,977 (13%)	3,742 (2%)	235 (2%)
Statins (yes)	23,802 (52%)	15,087 (49%)	8,715 (56%)
Beta Blockers (yes)	10,243 (22%)	6,640 (22%)	3,603 (23%)
ACEi or ARBS (yes)	20,549 (60%)	13,150 (43%)	7,399 (55%)
MRAs (yes)	2,528 (5%)	1,531 (5%)	997 (6%)
Calcium Channel Blockers	4,309 (9%)	2,511 (8%)	1,798 (12%)
Antiplatelets (yes)	10,204 (22%)	7,0406 (23%)	3,158 (20%)
Anticoagulants (yes)	4,309 (9%)	2,511 (8%)	1,798 (12%)
Thiazides (yes)	12,021 (26%)	7,969 (26%)	4,052 (26%)
Loop Diuretic (yes)	11,403 (25%)	6,683 (22%)	4,720 (30%)

Table 28 Clinical Characteristics of Primary Analysis stratified by Socioeconomic Deprivation Status.

Overall Population N=46,031 Demographics at Baseline (Secondary) **(%) or Median (25/75)**	**SIMD Quintile 1 (Most Deprived)**	**SIMD Quintile 2**	**SIMD Quintile 3**	**SIMD Quintile 4**	**SIMD Quintile 5 (Least Deprived)**
N=Scottish Index of Multiple Deprivation (SIMD) Group	N=18,517	N=8,355	N=6,360	N=5,643	N=7,156
Age (y)	63 (57 – 71)	64 (57 – 72)	64 (57 – 73)	64 (58 – 73)	64 (58-72)
Sex					
Women	8,974 (48%)	4,011 (48%)	2,887 (45%)	2,470 (44%)	3,025 (42%)
Men	9,543 (52%)	4,344 (52%)	3,473 (55%)	3,173 (56%)	4,131 (58%)
Ethnicity					
White	16,512 (89%)	7,223 (86%)	5,237 (82%)	4,642 (82%)	5,676 (79%)
Asian	485 (4%)	394 (5%)	463 (7%)	377 (7%)	583 (8%)
Other	831 (4%)	446 (6%)	393 (7%)	340 (6%)	433 (7%)
Unknown	510 (3%)	292 (3%)	267(4%)	284 (5%)	464 (6%)
***Body Mass Index (BMI)**	31 (27 – 35)	30 (26 – 32)	26 (26 – 31)	29 (26 – 31)	26 (25 – 30)
Missingness:	6,128 (33%)	2,727 (33%)	2,198 (34%)	2,041 (36%)	2,442 (34%)
***Current Smoker (yes)**	4,917 (27%)	1,796 (31%)	1,148 (27%)	748 (13%)	807 (17%)

Comorbidities n(%)					
Atherosclerotic Heart Disease (yes)	3,042 (16%)	1,274 (15%)	1,006 (16%)	790 (14%)	994 (14%)
Angina (yes)	2,551 (14%)	1,025 (12%)	781 (12%)	574 (10%)	690 (10%)
Atrial Fibrillation (yes)	2,530 (13%)	1,089 (13%)	866 (14%)	740 (13%)	864 (12%)
Chronic Obstructive Pulmonary Disease (yes)	2,488 (13%)	766 (9%)	508 (8%)	340 (6%)	293 (4%)
Chronic Kidney Disease (yes)	1,093 (6%)	453 (5%)	394 (6%)	307 (5%)	302 (4%)
Heart Failure (yes)	2,065 (11%)	860 (10%)	651 (10%)	520 (9%)	579 (8%)
Hyperkalaemia (yes)	1,001 (5%)	456 (5%)	301 (5%)	281 (9%)	297 (4%)
***Hypertension (yes)**	7,518 (41%)	3,598 (43%)	2,641 (42%)	2,254 (40%)	2,988 (42%)
Myocardial Infarction (yes)	1,990 (11%)	824 (10%)	632 (10%)	505 (9%)	594 (8%)
Peripheral Artery Disease (yes)	833 (4%)	282 (3%)	220 (3%)	175 (3%)	140 (2%)
Stroke/TIA (yes)	1,721 (9%)	751 (9%)	577 (9%)	462 (8%)	499 (7%)
Lab Tests within 6 months of inclusion, n (%)					
Plasma Glucose (mmoL)	8.9(6.7 – 10.4)	8.8 (6.7 – 10.1)	9 (6.7– 10.3)	7.9 (6.5 – 11.1)	7.6 (6.4 – 11)
Haemoglobin A1C (mmoL)	55 (46– 62)	55 (46– 61)	55 (46– 60)	51 (45– 63)	50 (44 – 61)
Haemoglobin (g/L)					
Men	139 (134 – 152)	139 (134 – 151)	140 (134 – 151)	140 (134 – 152)	140 (134 – 151)
Women	134 (124 – 140)	134 (123 – 139)	134 (124 – 140)	134 (124 – 140)	134 (124 – 140)
Total Cholesterol (mmol)	4.1 (3.7 – 5)	4.1 (3.7-5)	4.1 (3.8-5)	4.4 (3.7-5.3)	4.4 (3.7-5.3)
Triglycerides (mmol)	1.6 (1.6 – 2.3)	1.6 (1 – 2.3)	1.5 (1 – 2.2)	1.6 (1.2 – 2.3)	1.6 (1.1 – 2.2)
Serum Albumin (g/L)	37 (35 – 39)	37 (35-40)	37 (35-39)	38 (35-40)	38 (36-40)
eGFR (mL/min/1.73m2)	52 (44 - 60)	54 (44 - 61)	53 (44 – 61)	53 (44 – 61)	53 (44 – 61)
Alanine Transaminase – ALT (U/L)	22 (15-30)	21 (16-30)	21 (15 – 30)	22 (16-32)	23 (16-32)
Aspartate Transaminase – AST (U/L)	21 (16-26)	21 (17-26)	20 (16-26)	21 (16-27)	21 (17-27)
Alkaline Phosphate (U/L)	91 (73 -109)	89 (72-106)	86 (71-107)	84 (69-104)	82 (67-101)
Neutrophils (x10^9/L)	4.7 (3.7-6)	4.7 (3.7 – 6)	4.6 (3.6-5.9)	4.5 (3.5-5.8)	4.3 (3.4-5.5)
Lymphocytes (x10^9/L)	1.9 (1.4-2.5)	1.9 (1.5-2.5)	1.9 (1.4-2.4)	1.9 (1.4-2.4)	1.8 (1.4 -2.4)
Bilirubin (µmol/L)	10 (7 - 13)	10 (8 - 13)	10 (7 - 14)	10 (8 - 14)	11 (8 – 15)
Medications within 6 months of inclusion, n (%)					
Metformin (yes)	6,190 (33%)	2,705 (32%)	2,085 (33%)	1,684 (30%)	1,881 (26%)
Insulin (with Glucose-lowering Drug)	1,246 (7%)	503 (6%)	411 (6%)	322 (6%)	319 (4%)
Sulphonylureas (yes)	4,314 (23%)	1,857 (22%)	1,502 (24%)	1,188 (21%)	1,343 (19%)
SGTL2i (yes)	1,741 (9%)	758 (9%)	518 (8%)	432 (8%)	528 (11%)
DPP-4 inhibitor (yes)	2,125 (11%)	938 (11%)	717 (11%)	549 (10%)	704 (10%)
Statin (yes)	9,926 (54%)	4,396 (53%)	3,397 (53%)	2,794 (50%)	3,289 (46%)
Beta Blockers (yes)	4,164 (22%)	1,930 (23%)	1,444 (23%)	1,216 (23%)	1,489 (21%)
ACEi or ARBS (yes)	8,408 (61%)	3,873 (63%)	2,915 (62%)	2,432 (59%)	2,921 (57%)
MRAs (yes)	1,098 (6%)	494 (6%)	344 (10%)	281 (5%)	311 (4%)
Calcium Channel Blockers (yes)	1,773 (10%)	776 (10%)	589 (9%)	543 (10%)	628 (9%)
Antiplatelets (yes)	4,314 (23%)	1,857 (22%)	1,502 (24%)	1,188 (21%)	1,343 (19%)
Anticoagulants (yes)	1,773 (10%)	776 (9%)	589 (9%)	543 (10%)	628 (9%)

Thiazides (yes)	4,848 (26%)	2,241 (27%)	1,643 (26%)	1,462 (26%)	1,827 (26%)
Loop Diuretic (yes)	4,990 (27%)	2,093 (25%)	1,619 (25%)	1,351 (24%)	1,350 (19%)

6.6.2 Factors Associated with All-cause Mortality

For each SIMD quintile in **Figure 27**, the use of loop diuretics (LD), older age, lower serum concentrations of albumin and alanine transaminase (ALT) and estimated glomerular filtration rate (eGFR) were strong predictors of death. Some differences were also identified, for instance chronic obstructive pulmonary disease (COPD) was strongly associated with mortality for the most deprived quintiles, whilst a history of stroke was strongly associated with mortality for the least deprived quintiles (**Appendix C2**).

Patients with a history of heart failure were much more likely to be taking loop diuretics, which is expected as loop diuretics are an essential treatment for symptoms and signs of congestion due to heart failure. Despite this strong relationship, both loop diuretics and a diagnosis of heart failure were associated with mortality, although treatment with loop diuretics was the stronger and more consistent predictor. This would be consistent with many, but not all, patients treated with loop diuretics having undiagnosed heart failure. Low lymphocyte counts have been associated with congestion and could be another early manifestation of heart failure. COPD may also reflect misdiagnosed heart failure. A low haemoglobin is common in older people but especially those with heart failure and anaemia contributes to the development of heart failure. Atrial fibrillation often contributes to the development of heart failure. Low serum albumin and transaminase (ALT) and raised alkaline phosphatase might all reflect liver dysfunction due to congestion, although a raised alkaline phosphatase might also reflect disordered bone metabolism. A low eGFR indicates impaired kidney function, another precipitating factor for heart failure. A high cholesterol is associated with an increased risk of coronary artery disease and myocardial infarction, important risk factors for left ventricular systolic dysfunction and heart failure. In summary, the predictors of prognosis are all associated with factors that increase the risk of heart failure, a condition associated with a high mortality. Accordingly, the set of predictors is biologically plausible and coherent, with remarkably little variation by SES, even though prognosis was worse in those who were most deprived. As for the baseline model SES was not an important predictor of all-cause mortality.

The fact that blood tests and prescription of loop diuretics were such strong predictors of outcome may reflect, at least in part, the accuracy and completeness of such data in the EMR, whereas the diagnostic record may be less complete and less accurate. Also, almost all patients will have blood tests but only a minority of patients will have a specific diagnosis.

	Baseline Model - Primary Analysis (N=46,031)	SIMD Quintile 1 (N=18,517) (Most Deprived)	SIMD Quintile 2 (N=8,355)	SIMD Quintile 3 (N=6,360)	SIMD Quintile 4 (N=5,643)	SIMD Quintile 5 (N=7,156) (Least Deprived)
N = Deaths until end of the study	**11,727 (25%)**	**5,156 (27%)**	**2,067 (25%)**	**1,685 (26%)**	**1,443 (26%)**	**1,376 (19%)**
***C-statistic Score**	0.82	0.86	0.85	0.85	0.87	0.86
****Brier Score**	0.08	0.10	0.10	0.10	0.09	0.08
Key predictors influencing survival in rank order. > Indicates higher values of this variable associated with survival < Indicates lower values of this variable associated with survival	1. Loop Diuretic (Prescribed=1) 2. Age >3. Serum Albumin <4. eGFR5. ALT <6. Alkaline Ph >7. Heart Failure Event (Yes = 1)8. Total Cholesterol >9. Lymphocytes <10. COPD Event (Yes=1)11. Haemoglobin <12. Atrial Fibrillation (Yes=1)	1. Loop Diuretic (Prescribed=1) 2. Age >3. Serum Albumin <4. Neutrophils >5. Haemoglobin <6. Lymphocytes <7. Coronary Artery Disease (Yes =1)8. Glucose >9. ALT <10. Alkaline Ph >11. Total Cholesterol >12. COPD Event (Yes=1)	1. Age >2. Loop Diuretic (Prescribed=1)3. Serum Albumin <4. Haemoglobin <5. Neutrophils6. Hypertension7. Coronary Artery Disease (Yes =1)8. eGFR <9. Lymphocytes <10. Alkaline Ph >11. Heart Failure Event (Yes = 1)12. Total Cholesterol >	1. Age >2. Loop Diuretic (Prescribed=1)3. Serum Albumin <4. Haemoglobin <5. Lymphocytes <6. Neutrophils >7. Hypertension8. ALT <9. Glucose >10. HbA1c >11. eGFR <12. Alkaline Ph >	1. Age >2. Serum Albumin <3. Loop Diuretic (Prescribed=1)4. Haemoglobin <5. Neutrophils >6. Lymphocytes <7. Total Cholesterol >8. Alkaline Ph >9. ALT <10. Glucose >11. HbA1c >12. eGFR <	1. Age >2. Serum Albumin <3. Haemoglobin <4. Neutrophils >5. Lymphocytes <6. Alkaline Ph >7. Hypertension8. eGFR <9. Total Cholesterol >10. ALT <11. Glucose >12. Loop Diuretic (Prescribed=1)
***C-statistic Score: Model discrimination.** ****Time Brier Score: Measure of calibration for time-dependent models. A lower Brier score (0-1) indicates greater accuracy.**						

Figure 27 (Primary analysis) Factors predicting all-cause mortality in patients with T2DM, stratified by SES Quintile

6.6.3 Model Validation

The RSF model outperformed the Cox Regression Elastic Net model for both discrimination and calibration. **Figure 27** compares survival prediction performance of the Elastic Net and RSF with tenfold cross-validation. Primary analysis and Secondary excluding patients with missing BMI are shown. Model comparison is further carried out by the time-dependent areas under the curve (AUC) which measures the ability of a model to discriminate between different event times or durations (**Figure 28**). This shows time-varying prediction accuracy for time-to-event models. The RSF has a higher performance of 0.84. As expected in almost any scientific field, the performance even of a good predictive model tails off over time unless the model is updated with new information. For each year the AUC shows good performance > 0.82.

	Primary Analysis: Baseline Model (46,031)	Quintile 1 (N=18,517) (Most Deprived)	Quintile 2 (N=8,355)	Quintile 3 (N=6,360)	Quintile 4 (N=3,602)	Quintile 5 (N=7,156) (Least Deprived)
Random Survival Forest						
C-statistics Score		**0.81**	**0.80**	**0.83**	**0.83**	**0.80**
Brier Score		**0.10**	**0.10**	**0.10**	**0.09**	**0.08**
Cox Elastic Net						
C-statistic Score		**0.77**	**0.78**	**0.81**	**0.81**	**0.75**
Brier Score		**0.11**	**0.11**	**0.11**	**0.11**	**0.09**
	Secondary Analysis: Baseline Model (N=30,495)	**Quintile 1 (N=12,389) (Most Deprived)**	**Quintile 2 (N=5,628)**	**Quintile 3 (N=4,162)**	**Quintile 4 (N=3,602)**	**Quintile 5 (N=4,714) (Least Deprived)**
Random Survival Forest						
C-statistics Score	0.83	0.82	0.81	0.85	0.84	0.81
Brier Score	0.07	0.08	0.076	0.077	0.07	0.06
Cox Elastic Net						
C-statistic Score	0.79	0.79	0.79	0.82	0.83	0.77
Brier Score	0.08	0.08	0.07	0.7	0.08	0.06

Figure 28 Survival Prediction of the Elastic Net and Random Survival Forest

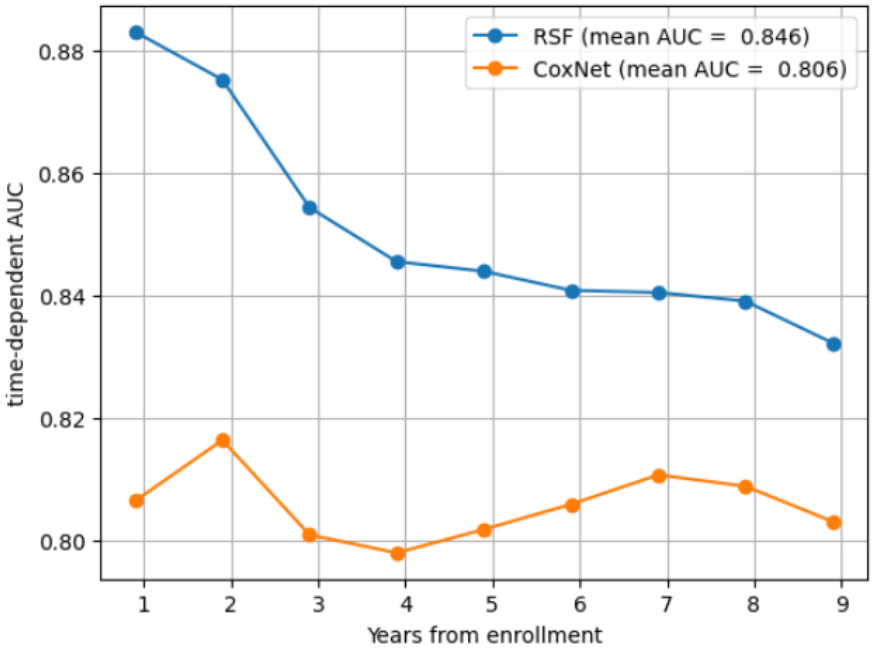


Figure 29 time-dependent AUC validation for Baseline Model (Secondary analysis)

6.7 Discussion

Routinely collected data obtained from EMR can be used to predict survival for patients with new-onset T2DM with a fair degree of accuracy. SES predicted outcome in an age and sex adjusted model, although not in a fully adjusted multi-variable model, but there were only subtle differences in predictive variables when analysed by SIMD quintile. This suggests that the association between social deprivation and an adverse prognosis can be explained by other variables included in the model.

The variable rankings varied by SIMD quintile. For instance, loop diuretic use was ranked highest in Q1 (most deprived) however dropped to 12th in Q5 (least deprived), suggesting a differential association with mortality risk across SES. This variation in rankings implies that the strength of association (i.e., the hazard) for specific variables like loop diuretics is not uniform across deprivation quintiles.

Many of the findings are aligned with those of randomised controlled trial (RCTs). However, our population includes many who would typically be excluded from RCTs, providing an opportunity to identify key differences stratified by the level of social affluence/deprivation. Previously, SES has not been investigated in this way or to this extent, which should inform future clinical studies.

Loop diuretics (LD) were a strong predictor of all-cause mortality in patients with T2DM regardless of SES. This may be because loop diuretics are a marker of undiagnosed heart failure, although it is also possible that loop diuretics accelerate the progression of cardiovascular and renal disease and increase the risk of sudden death (Rosano et al., 2017). Similar to the results of the analyses presented in this these, in the EMPA-REG OUTCOME trial (Pellicori et al., 2021) patients with T2DM prescribed LD had higher rates of cardiovascular events and mortality even if they were not reported to have HF. A recent study of 198,898 patients with cardiovascular disease, also suggested that those receiving loop diuretic therapy had higher cardiovascular and all-cause mortality, possibly due to missed heart failure diagnoses, association with other serious conditions, or inappropriate use (Friday et al., 2024). Furthermore, according to the NICE Guidelines (Moran et al., 2022) loop diuretics can also exacerbate T2DM.

Lower serum concentrations of albumin were strongly related to outcome in all SIMD quintiles. Albumin is a blood protein synthesised by the liver that has many biological functions but most importantly helps keeps the water in the circulation because, in health, it is not filtered across the capillary membrane. Low blood levels of albumin may reflect reduced synthesis due to liver dysfunction, due either to congestion (heart failure), disease (e.g. alcohol) or malnutrition (although this probably only when severe). Renal disease can cause increased loss of albumin into the urine; this can be massive (nephrotic syndrome).

Inflammatory disease and heart failure itself may also lead to albumin leaking out into the tissues; there is a lot of albumin in oedema fluid.

Prolonged bed rest may also cause serum albumin concentration to drop, although it is unclear whether this is due to reduced liver synthesis or increased circulating volume of water. Many studies have shown that low serum albumin is associated with a worse prognosis (Cleland et al., 2014; Eline Bretscher et al., 2022). Low serum concentrations of another routinely available marker of liver function, serum alanine aminotransferase (ALT), also predicted prognosis. This suggests that reduced liver function rather than liver disease, which would be expected to increase ALT, is at play. Other studies of large populations of older people have also shown that low serum concentrations of transaminases are associated with a poorer prognosis (Ndrepepa and Kastrati, 2019; Ramaty et al., 2014). However, obesity, hyperlipidaemia and T2DM have all been associated with mild-to-moderate ALT elevation (Ho et al., 2022), making this observation all the more remarkable. Raised plasma concentrations of serum alkaline phosphatase, a marker of biliary tract obstruction or several types of bone disease, were also associated with a poorer prognosis. The significance of this is unclear.

Neutrophil to lymphocyte ratio (NLR) is an indicator of systemic inflammation (high neutrophils) and congestion (low lymphocytes), with high ratios associated with a poor prognosis (Mikolasch et al., 2022). Elevated NLR levels are also associated with COPD (Paliogiannis et al., 2018), reflecting the inflammatory nature of the disease. The results of this study showed the prevalence of COPD is higher in the most deprived socioeconomic groups. COPD is strongly associated with smoking and air pollution, often impairs quality of life and is associated with an increased risk of respiratory infections, cardiovascular disease and a poorer prognosis (Balbirsingh et al., 2022).

Based on the requirements of analysis, the random survival forest was able to identify underlying relationships that linear models cannot (Miao et al., 2015). There was no assumption of proportional hazards, which is critical in traditional survival models. This method also accommodated right-censoring. Most studies limit their methodology to standardised Cox regression, although, tree-based models are known to be a lot more reliable for feature selection in other clinical settings (Spooner et al., 2020a). One of the main aspects of this study investigates the prognostic drivers of disease based on survival prediction.

Currently, medical studies utilise multivariable regression which provides direct interpretability of coefficients and allows for statistical inference but including specific assumptions which may not capture complex interactions.

To overcome these limitations, the use of SHAP values, offered a model-agnostic approach to interpretability, accounting for interactions and providing a unified measure of feature importance. SHAP is useful for explaining survival probability predictions, especially for sub-group analyses and understanding the relative contributions of predictors. Phi correlation was also key to discover key clusters between laboratory variables, which has not yet been applied in the medical field. Techniques to capture non-linear relationships were used, which are not typically practiced in studies. This approach introduces targeted interventions that are tailored to the specific factors that are most relevant for each socioeconomic group.

This study carried out risk estimation, revealing the interpretation of prognostic factors driving all-cause mortality in T2DM patients across diverse SES ranks. A survival machine learning model could provide valuable assistance in the decision-making process once a thorough patient profile is established, including socioeconomic status. It provides fairness (including patients from different socioeconomic backgrounds) in risk modelling and addresses systemic biases that can arise from socioeconomic disparities. The approach ensures that risk prediction models do not inadvertently disadvantage certain socioeconomic groups, improving the reliability and generalisability of clinical insights. This also aligns with ethical standards, ensuring that all patients, regardless of their SES, receive appropriate care recommendations.

Further work is required to include a higher percentage of ethnicities other than White. A prospective cohort study between 2006 and 2010 of 500,000 participants using UK Biobank data found that South Asians had the highest T2DM prevalence (17.9%), followed by the Black (11.7%) and White (5.5%) ethnic groups (Deepali Nagar et al., 2021). However further data collection is required to examine environmental factors in clinical studies. External validation is necessary to ensure the generalisability of the approach adopted here. Unfortunately, the Hong Kong Cohort used throughout this research for external validation did not have a reliable measure of SES. See **Appendix C3**.

BMI

More than 80% of patients with a measured BMI were either overweight or obese. It is well known that obesity is strongly associated with and a cause of insulin resistance and T2DM. Unfortunately, measurement of BMI was missing for >48% patients. Missing data reduces the quality and reliability of predictive models and may be informative. A sensitivity analysis revealed that including patients without a recorded BMI skewed the risk prediction model results. BMI was missing not at random (MNAR) (Heymans and Twisk, 2022). Patients with missing BMI values were older, were less likely to have ethnicity recorded had a substantially higher mortality **(Appendix C4).** A previous analysis of the Scottish Diabetes Registry (Read et al., 2017) found that individuals with missing BMI data had more comorbidities and lower survival rates. BMI data completeness and patient survival are somehow connected. It is likely that BMI is more likely to be recorded when the patient is obese and least likely to be recorded when it is in the normal range.

Patients who were obese had a better survival (the obesity paradox (Costanzo et al., 2015; NICE Guidelines, 2014) **(Appendix C5)**). Those in the "Underweight" group had a 50% higher risk of all-cause mortality. Patients who were "Obese", "Severely Obese" or "Overweight" had a lower all-cause mortality than those with a "Normal" weight. The poor prognosis of those with a missing BMI is consistent with a high proportion of these patients have a normal BMI. A similar 'obesity paradox' has been noted in older patients with hypertension, CAD, AF or heart failure (Cullington et al., 2014). However, amongst younger people, obesity is associated with the earlier onset of CVD and a higher mortality. This is a controversial area. Obese people may develop CVD at a younger age.

Younger age then confers a better prognosis. Alternatively, patients with T2DM who are not obese may have a more severe metabolic disease. Obese individuals might receive more medical attention and monitoring, potentially leading to earlier detection and management of health issues, although there is little evidence to support this hypothesis (Fmedsci et al., 2023). There is growing evidence that treatments for obesity reduce cardiovascular risk in patients with T2DM.

6.8 Conclusion

This analysis shows that variables collected in routine EMR can predict the mortality of patients with new-onset T2DM with reasonable accuracy. The analysis also shows the prognostic importance of socioeconomic deprivation for patients with T2DM, although the risk factors for mortality were very similar across SIMD quintiles. In agreement with many previous reports, obesity was associated with a better survival, although RCTs suggest that effective treatment of obesity reduces risk – leading to a paradox within the obesity paradox.

Chapter 7 "Treatment with Loop Diuretics is Strongly Associated with Prognosis of Patients with Type-2 Diabetes Mellitus in Two Different Geographies"

Abstract

Introduction: Type 2 diabetes mellitus (T2DM) is associated with accelerated development of atherosclerosis and a reduced life expectancy. Applying machine learning (ML) to administrative electronic medical records (EMRs) might identify novel characteristics that predict outcome, thereby improving prognostic precision and suggesting new targets for investigation and treatment (N Kaur et al., 2024).

Purpose: We adapted an ML survival analysis approach using random survival forests investigate the use of EMRs to predict all-cause mortality in two ethnically and geographically different populations; one from the Greater Glasgow & Clyde (GG&C) region in Scotland and the other from Hong Kong.

Methods: EMRs included information on demographics, prior comorbidities, laboratory measurements, medications, and mortality. Multivariable Cox regression and time-dependent random forest model were used to identify predictors of all-cause mortality. Subsequently, we applied a state-of-the-art ML interpretability method, to gain further insight into the predictors.

Results: In GG&C, 46,031 individuals received a new diagnosis of T2DM between 2009 and 2019. Their median age was 66 (interquartile range: 56 to 75) years. Within 10 years, 11,727 (25%) had died. In Hong Kong, 273,876 patients with a first-attendance with T2DM at public hospitals or clinics were included, with follow-up until December 2019. The median age of the patients was 64 (interquartile range of 57 to 72) years. Within 10 years, 91,155 (33%) had died. For both T2DM populations, the strongest association with all-cause mortality was use of loop diuretics (Figure 1). For GG&C, other important predictors were greater age, lower serum albumin, elevated alanine transaminase (ALT), increased alkaline phosphatase, and lower estimated glomerular function rate (eGFR) (c-index: 0.83; Brier score: 0.07). For Hong Kong, predictive variables were similar and included greater age, lower eGFR, lower haemoglobin and lymphocytes, lower serum albumin, and elevated alkaline phosphatase (c-index: 0.85; Brier score: 0.06). Multivariable Cox regression adjusting for age, sex and key predictors showed a

higher mortality amongst those prescribed loop diuretics compared to those who were not (GG&C: adjusted hazard ratio: 2.93, (95% CI: 2.821 to 3.04); Hong Kong: adjusted hazard ratio: 1.75 (95% CI: 1.72 to 1.77). Only a minority of patients prescribed loop diuretics had a diagnosis of heart failure, end-stage renal disease or resistant hypertension.

Conclusion: Amongst patients with recent-onset T2DM, prescription of loop diuretics was the feature most strongly associated with all-cause mortality in both GG&C and Hong Kong. Prescription of loop diuretics might be a pharmacological marker of congestion and undiagnosed heart failure or might itself have deleterious effects on prognosis.

7.1 Introduction

This chapter investigates the prescription of loop diuretics for all-cause mortality comparing two distinct populations. Identifying markers of an adverse prognosis in patients with type-2 diabetes mellitus (T2DM) could improve clinical care and identify potential new therapeutic targets. Machine learning (ML) enables the analysis of electronic medical records (EMRs) to identify novel factors associated with mortality and enhance prognostic precision. It is important to validate findings. Often this is done by withholding a random sample of people from the population of interest. This can validate the internal consistency of a model within a population but does not provide information on whether it can be extrapolated to other populations. Cross-validation of prognostic models in widely different populations, in terms of geography, healthcare systems and ethnicity, provides evidence of the generalisability of the model.

Accordingly, routinely collected data from EMRs for two large populations of patients with T2DM were obtained, one from the Greater Glasgow & Clyde (GG&C) region in the West of Scotland and the other from Hong Kong. Of particular interest was prescribing of loop diuretics (LD), widely used for managing heart failure, a serious but often undiagnosed complication of cardiac dysfunction. By testing the model's performance in diverse settings, this chapter assessed whether the predictive factors identified in Chapter 6 are consistent and relevant across different demographic groups for all-cause mortality in patients with T2DM.

7.2 Aim

The main aim of this chapter is to develop separate prediction models for mortality in patients with T2DM from GG&C and in Hong Kong using machine learning (ML) and to compare them, with a special focus on LD prescribing.

7.3 Study Data

In GG&C, a operationalised database called Safe Haven provides de-identified, routinely collected, National Health Scotland (NHS) EMRs. The Scottish Care Information (SCI)-Diabetes registry includes all patients when they are first assigned a Read Code [10] (a coded thesaurus of clinical terms used in the NHS since 1985) for diabetes mellitus in a primary or secondary care health care information system, which is estimated to capture >99% of all patients in Scotland with diabetes mellitus (Livingstone et al., 2012). In order to access the data, approval is required from the Local Privacy Advisory Committee of the West of Scotland Safe Haven, which requires the application to focus on the population of interest. Accordingly, the request was limited to people with diabetes mellitus aged ≥50 years because T2DM is less common in younger patients and they have a relatively good medium-term prognosis. In Hong Kong, EMRs are extracted from an integrated health database (Clinical Data Analysis and Reporting System) operated under the Hospital Authority in Hong Kong. T2DM was defined as a primary or secondary diagnoses in public healthcare institutes. Patients are classified in Hong Kong using the International Classification of Primary Care (Ho Wong et al., 123AD) (ICPC) code T90 (Diabetes; non-insulin-dependent) and ICD-9 250 Code. For both geographies, patients aged ≥50 years with a first record of diabetes between 1st of January 2009 to 31st December 2019 were enrolled.

7.4 Patient Information

Baseline patient characteristics included age, sex and ethnicity and smoking. However, due to only 10% smoking records in the Hong Kong population, it was excluded from this analysis due to unreliability. Common comorbidities in both GG&C and Hong Kong included chronic kidney disease (CKD), chronic obstructive pulmonary disease (COPD), coronary artery disease (CAD), heart failure, hypertension, peripheral artery disease (PAD), stroke, atrial fibrillation (AF) was defined by ICD-9 Codes). Routinely collected laboratory tests included haemoglobin, lymphocyte and neutrophil white blood cell counts, haemoglobin A1c, total cholesterol, triglycerides, serum albumin, estimated glomerular filtration rate (eGFR), potassium, AST, ALT, alkaline phosphatase and bilirubin. Results in the same month as first diagnosis were preferred but the time window could be extended to one year if a relevant test was otherwise not available. For both GG&C (NICE, 2024a) and Hong Kong (Yang et al., 2022), a Glycated Haemoglobin (HbA1c) value of $\geqslant$6.5% (48 mmol/mol) was used to define T2DM as advised by international guidelines.

Appendix D1 shows the laboratory tests measured for both populations (NICE, 2024b; SIGN, 2023; 'Hong Kong Diabetes Association', 2024). Mortality was classified using the International Classification of Diseases, 9th Revision (ICD-9) and 10th Revision (ICD-10) codes obtained from death certificates completed by doctors.

In GG&C, medicines for diabetes and cardiovascular disease, including lipid-lowering medications and loop diuretics were extracted from a Prescribing Information System (PIS) that covers all NHS medicines, classified using the British National Formulary (BNF). In Hong Kong, medicines for diabetes and cardiovascular disease were extracted directly from the Clinical Data Analysis and Reporting System. **Table 29** provides a comparison of LD dispensing and prescribing practices.

Table 29 Dispensing and Prescribing Loop Diuretics in GG&C and Hong Kong

Dispensing and Prescribing Loop Diuretics in West of Scotland and Hong Kong		
Aspect	**Glasgow (West of Scotland)**	**Hong Kong**
Population Estimate (2019)	~5.4 million (Scotland); ~1.8 million (West of Scotland)	~7.5 million
Managing Authority	National Health Service (NHS) Scotland	Hospital Authority (HA)
Primary Indications	Heart failure, CKD, hypertension	Heart failure, CKD, hypertension
Prescription Frequency	~1 million prescriptions annually (estimate: 150,000 patients prescribed loop diuretics, with ~6 prescriptions per year)	Information not available.
Number of Patients Prescribed Loop Diuretics	~150,000 prescriptions annually, predominantly in older adults	~90,000 patients prescribed annually in public hospitals (estimates for private healthcare and general outpatient care are likely higher)
Most Common Medicine	Furosemide	Furosemide
Comorbidity Management	Often prescribed with treatment for diabetes or hypertension	Prescriptions commonly co-managed with CKD or hypertension
Monitoring and Auditing	NHS provides regular audits and adherence to guidelines	Hospital Authority regulates prescription practices through public healthcare
Patient Population	Older adults, often with multiple comorbidities	Similar to Scotland, with a high prevalence in older patients
Information derived from the following guidelines: Public Health Scotland, National Health Services (NHS) Scotland, NICE guidelines and Hospital Authority (HA).		

Figure 29 shows the consort diagram for both populations in this analysis.

GG&C: The Scottish Care Information - Diabetes Collaboration population included 47,396 patients aged 50 years or older from the NHS Greater Glasgow and Clyde (GG&C) region. Patients treated with insulin only, who were presumed to have Type-1 Diabetes, were excluded (1,365 individuals), leaving 46,031 patients eligible for analysis, with or without an available measure of body mass index (BMI).

Hong Kong Population: A total of 273,876 patients aged 50 years or older with T2DM from public hospitals and clinics in Hong Kong were included. All patients were eligible for analysis, although some did not have an available measurement of BMI. No patient was treated with insulin alone.

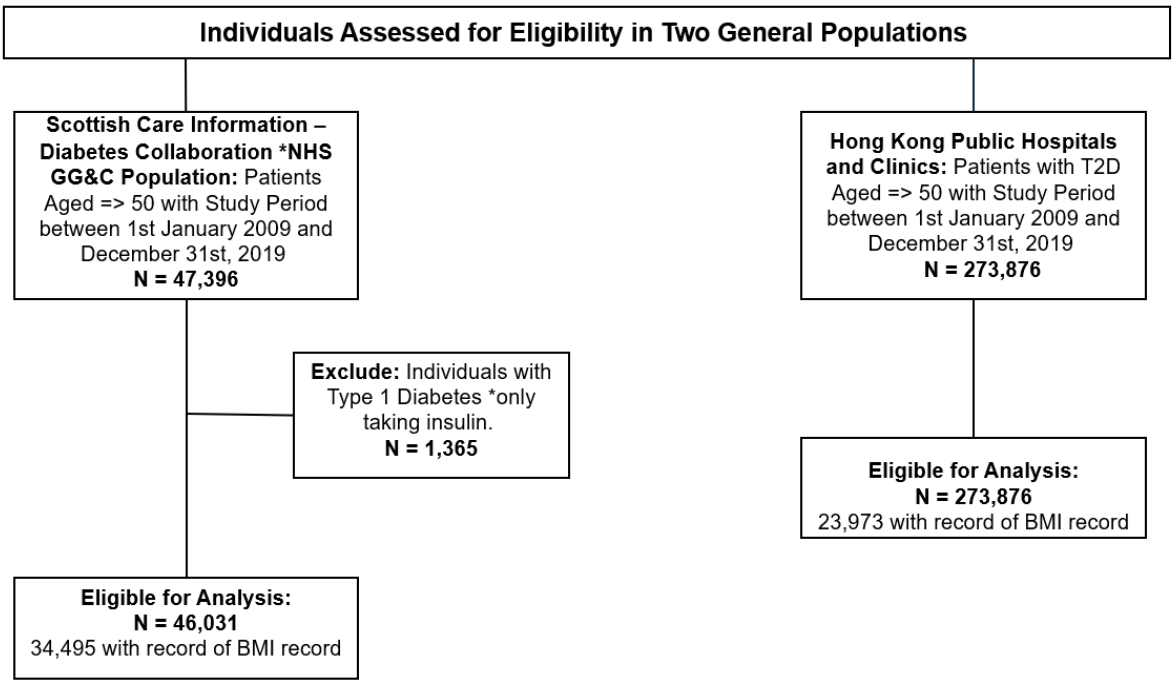


Figure 30 Consort Diagram of T2DM in GG&C and Hong Kong

7.5 Methods

The methods outlined in Chapter 4, including the use of random survival forest models, were adapted to analyse the outcome of mortality in both GG&C and Hong Kong populations. In addition to the random survival forest, Kaplan-Meier survival analysis and Cox proportional hazards models were employed to assess all-cause mortality risk across these distinct populations. SHAP (SHapley Additive exPlanations) values were used to enhance the interpretability of the random survival forest, providing insights into the key factors driving mortality predictions at the individual patient level. A table was included to show potential reasons for prescribing LD in both populations, which might help explain the association between LD and mortality.

7.5.1 Gradient Boosting with Cox Proportional Hazards (CPH)

To enhance the predictive performance of survival analysis, a hybrid approach was employed that combines Gradient Boosting with the Cox Proportional Hazards (CPH) model (Spooner et al., 2020). This integrated method improves modelling of survival data, particularly in predicting all-cause mortality, while accommodating complex relationships among covariates. Gradient Boosting is an ensemble machine learning technique that sequentially combines multiple weak learners, typically decision trees, to create a robust predictive model. By iteratively fitting the model to the residuals of the previous iteration, Gradient Boosting captures complex patterns in the patient data. The basic formulation (Friedman, 2001) of the model is represented as follows:

$$F(x) = F_0(x) + \sum_{m=1}^{m} \gamma_m h_m(x)$$

where $F_0(x)$ is the initial prediction, $h_m(x)$ represents the weak learners, γ_m and denotes the associated weights.

Integration with Cox Proportional Hazards

The integration of Gradient Boosting with the CPH model capitalizes on the strengths of both approaches. The Cox model provides a semi-parametric framework that describes the relationship between covariates and the hazard function. Gradient boosting survival analysis implementation was utilised to fit the model to the training dataset. This approach allowed for the effective modelling of the baseline hazard function while including the advantages of machine learning to capture non-linear relationships among the covariates. It optimises the partial likelihood iteratively, handles high-dimensional data well and includes regularisation to prevent overfitting.

The model was initialised with parameters such as the number of estimators, learning rate, and maximum depth of the trees. The fitting process was performed using the training data, which consisted of survival times and associated covariates. Following training, the model's performance was evaluated using the concordance index (c-index), which assesses the model's ability to rank survival times correctly.

7.6 Results

7.6.1 Baseline Characteristics Results of GG&C and Hong Kong for all-cause mortality

In GG&C, 46,031 individuals with a new diagnosis of T2DM between 2009 and 2019 were included, median age was 64 (interquartile range: 57 to 72) years. Within 10 years, 11,727 (25%) had died. In Hong Kong, 273,876 patients with a first-attendance for T2DM at public hospitals or clinics were included, with follow-up until December 2019. The median age of the patients was 65 (interquartile range of 56 to 75) years. Within 10 years, 91,155 (33%) had died.

Appendix D2 shows characteristics stratified by sex for each population with and without a missing measurement of BMI.

The Hong Kong T2DM population was predominantly of Chinese ethnicity (92%), were slightly older, had a lower median BMI and included a higher proportion of women compared to GG&C, of which 85% were of White ethnicity. Patients in the GG&C were more likely to have a record of hypertension, CKD, atrial fibrillation, coronary, peripheral or cerebrovascular disease and heart failure, which could reflect a higher burden of comorbidities in the GG&C or more complete reporting.

Compared to the GG&C, men and women from Hong Kong had a lower haemoglobin and were more likely to fulfil the World Health Organisation's definition of anaemia (see **Table 30**), which is consistent with previous reports of differences in haemoglobin between people of European ancestry and the Chinese population. However, patients from GG&C also had a lower eGFR, which may also be associated with a higher prevalence of iron deficiency anaemia.

Treatment patterns were very different between GG&C and Hong Kong. The principal glucose lowering agents in Hong Kong were metformin and sulfonylureas, whereas only slightly more than half of patients in GG&C received these agents. The principal anti-hypertensive agents used in both GG&C and Hong Kong were ACE inhibitors or ARBs but, in Hong Kong, many more were treated with calcium channel blockers, beta-blockers and thiazides. Patients in GG&C were much more likely to receive a statin, which might account for the slightly lower serum cholesterol concentration for patients from the GG&C compared to Hong Kong.

Table 30 Mortality in GG&C and Hong Kong Patients with T2DM

Categorical data shown as percentages and continuous data as median and quartiles Substantial differences between T2DM populations are shown in bold because some differences, although highly statistically different, are of doubtful clinical relevance.

Demographics at Baseline:	GG&C N = 46,031	Hong Kong N= 273,876	P-value
Age (years)	64 (57 – 72)	65 (56 – 75)	<0.001
Sex			<0.001
Men	24,664 (54%)	132,040 (48%)	
Women	21,367 (46%)	141,836 (52%)	
Ethnicity			<0.001
Chinese	N/A	251,966 (92%)	
White	39, 290 (85%)	N/A	
Other	6,741 (15%)	21,910 (8%)	
***Body Mass Index (BMI)**	26 (26 – 31)	25 (23 – 26)	<0.35
Smoker (Yes)	9,416 (20%)	N/A	N/A
Comorbidities n(%)			
Hypertension (yes)	18,999 (41%)	64,246 (23%)	<0.001
Chronic Kidney Disease (yes)	2,549 (6%)	3,381 (1%)	<0.001
Hyperkalaemia (yes)	2,336 (5%)	N/A	N/A
Atrial Fibrillation (yes)	6,083 (13%)	7,772 (3%)	<0.001
COPD (yes)	4,395 (8%)	818 (0.3%)	<0.001
Coronary Heart Disease (yes)	7,106 (16%)	26,423 (10%)	<0.001
Myocardial Infarction (yes)	4,545 (11%)	N/A	N/A
Peripheral Artery Disease (yes)	1,650 (4%)	346 (0.1%)	0.73
Stroke/TIA (yes)	4,010 (9%)	8,986 (3%)	<0.001
Heart Failure (yes)	4,675 (10%)	11,189 (4%)	<0.001
Anaemia	2, 302 (5%)	19,425 (6%)	<0.001
Lab Tests within 6 months of inclusion, n (%)			
Haemoglobin A1C (mmoL/mol)	55 (46– 61)	56 (51– 63)	0.28
Haemoglobin (g/L)			
Men	138 (134 – 151)	131 (132 – 139)	
Women	134 (123 – 139)	129 (122 – 136)	<0.001
Lymphocyte count (x10^9/L)	2.0 (2.0 – 2.3)	1.9 (1.7 – 2.4)	<0.001
Neutrophil count (x10^9/L)	5.0 (3.8-5.6)	5.3 (4.4 – 7.1)	<0.001
Total Cholesterol (mmol/L)	4.1 (3.7-5.0)	4.7 (4.3 – 5.2)	0.14
Triglycerides (mmol/L)	1.5 (1.0 – 2.2)	1.5 (1.1 – 1.9)	<0.001
Serum Albumin (g/L)	37 (35-39)	40 (38 – 42)	<0.001
eGFR (mL/min/1.73m^2)	54 (44 - 61)	64 (53 – 77)	<0.001
Potassium (mmol/L	4.3 (4.0 – 4.6)	4.2 (4.0 – 4.4)	<0.001
Alanine Transaminase – ALT (U/L)	22 (16-30)	23 (17 – 30)	<0.001

Aspartate Transaminase – AST (U/L)	20 (16-26)	25 (21 – 39)	<0.001
Alkaline Phosphatase (U/L)	89 (72-105)	75 (65 – 87)	<0.001
Bilirubin (µmol/L)	10 (7 - 13)	10.3 (9.2 – 12.8)	<0.001
Medications within 6 months of inclusion, n (%)			
Metformin (yes)	14,545 (32%)	185,881 (68%)	<0.001
Sulphonylureas (yes)	10,204 (22%)	173,525 (63%)	<0.001
DPP4i (yes)	5,033 (11%)	325 (0.1%)	<0.001
GLP1-receptor antagonists (yes)	2,139 (5%)	17	<0.001
Insulin (with other Glucose-Lowering Agent)	2,801 (6%)	29,697 (11%)	<0.001
Statins (yes)	23,802 (52%)	61,401 (22%)	<0.001
Beta Blockers (yes)	10,243 (22%)	92,309 (34%)	<0.001
ACEi or ARBS (yes)	20,549 (60%)	121,786 (44%)	<0.001
Calcium Channel Blockers	4,309 (9%)	109,225 (40%)	<0.001
Thiazides (yes)	12,021 (26%)	52,096 (19%)	<0.001
Loop Diuretics	11,403 (25%)	60,152 (22%)	<0.001

7.6.2 All-Cause Mortality Risk Prediction Model(s)

For both T2DM populations, using a random survival forest approach, the strongest predictor for all-cause mortality was being treated with loop diuretics (**Table 31**). For GG&C, other important predictors were greater age, lower serum albumin, lower alanine transaminase (ALT), increased alkaline phosphatase, and lower estimated glomerular function rate (eGFR) (c-index: 0.83; Brier score: 0.07). For Hong Kong, predictive variables were similar and included greater age, lower eGFR, lower haemoglobin and lymphocytes, lower serum albumin, and elevated alkaline phosphatase (c-index: 0.85; Brier score: 0.06).

Table 31 Results for predicting all-cause mortality in Glasgow and Hong Kong

	Glasgow: Baseline RSF Model (N=46,031)	**Hong Kong: Baseline RSF Model (N=273,876)**
***C-statistic Score**	0.82	0.85
****Time Brier Score**	0.08	0.06
Key predictors of all-cause mortality in rank order. **>** Indicates higher values of this variable increase risk of all-cause mortality **<** Indicates lower values of this variable increase risk of all-cause mortality	1. Prescribed Loop Diuretic **>** 2. Age **>** 3. Serum Albumin **<** 4. eGFR **<** 5. ALT **<** 6. Alkaline Phosphatase **>** 7. Heart Failure 7. Total Cholesterol **>** 8. Lymphocytes **<** 9. COPD 10. Haemoglobin **<**	1. Prescribed Loop Diuretic **>** 2. eGFR **<** 3. Age **>** 4. Haemoglobin **<** 5. Alkaline Phosphatase **>** 6. Lymphocytes **<** 7. ALT **<** 8. Serum Albumin **<** 9. Neutrophils **>** 10. Heart Failure
***C-statistic Score: Model discrimination.** ****Time Brier Score: Measure of calibration for time-dependent models. A lower Brier score (0-1) indicates greater accuracy.** **Abbreviations: random survival forest (RSF) estimated glomerular filtration rate (eGFR) and Alanine aminotransferase (ALT), Chronic Obstructive Pulmonary Disease (COPD)**		

7.6.3 Model Validation: Gradient Boosting with Cox Proportional Hazards

Table 32 presents the implementation and results of the gradient boosting CPH model for key predictors of mortality. The ordering of the risk factors was different. Loop diuretics were the strongest predictor of mortality in the GG&C T2DM population followed by age, eGFR, atrial fibrillation, serum albumin, ALT, ALP, heart failure event, haemoglobin and lymphocytes. The predictive performance was good (c-index: 0.81) in the GG&C T2DM population.

A maximum tree depth of 4 was used for the Gradient Boosting Cox model. This depth allows the algorithm to capture moderate interactions between predictors, which is important given the complexity of clinical risk factors. It provides a balance between model expressiveness and generalisability, avoiding the overfitting risk associated with deeper trees while outperforming very shallow trees (e.g. depth = 1), which are too simplistic.

For the Hong Kong T2DM population, the strongest predictors were serum albumin, eGFR, loop diuretics, older age, haemoglobin, ALP, lymphocytes, ALT, heart failure and total cholesterol. The C-Index of 0.84 indicates strong predictive performance. The similarity of the variables most strongly associated with mortality in two very different patient T2DM populations using two different statistical approaches suggests that the results are reliable.

Table 32 Model Validation using Gradient Boosting with CPH model

Gradient Boosting CPH	**Results of Model Validation**	
	Glasgow (N=46,031)	**Hong Kong (N=273,876)**
Number of estimators	100	100
Learning Rate	1.0	1.0
Maximum Depth of Trees	4	4
C-Index	0.81	0.84
Key Predictors of all-cause mortality	Loop diuretics, Age, eGFR, Atrial fibrillation event, Serum Albumin, ALT, ALP, Heart Failure event, Haemoglobin and Lymphocytes	Serum Albumin, eGFR, Loop Diuretics, Age, Haemoglobin, ALP, Lymphocytes, ALT, Heart Failure event and Total Cholesterol

7.6.4 Kaplan Meier & Cox Proportional Hazards

Figure 30 and Figure 31 Kaplan Meier plots showing all-cause mortality in the GG&C and Hong Kong T2DM populations prescribed LD and with heart failure (blue), prescribed LD but without heart failure (orange), not prescribed LD but with heart failure HF (green) and neither prescribed LD nor with a diagnosis of heart failure (red).

GG&C T2DM population: The survival curves show a clear difference in survival probability between these groups confirmed by a multivariable log-rank test ($p < 0.005$).

- **Blue Curve:** approximately 65% of patients with heart failure who were prescribed LD died within 5 years.
- **Orange Curve:** approximately 45% of patients prescribed LD but without heart failure died within 5 years.
- **Green Curve:** approximately 25% of patients who were not prescribed LD but were diagnosed with heart failure died within 5 years.
- **Red Curve:** approximately 15% of patients who were not prescribed LD nor were diagnosed with heart failure died within 5 years.

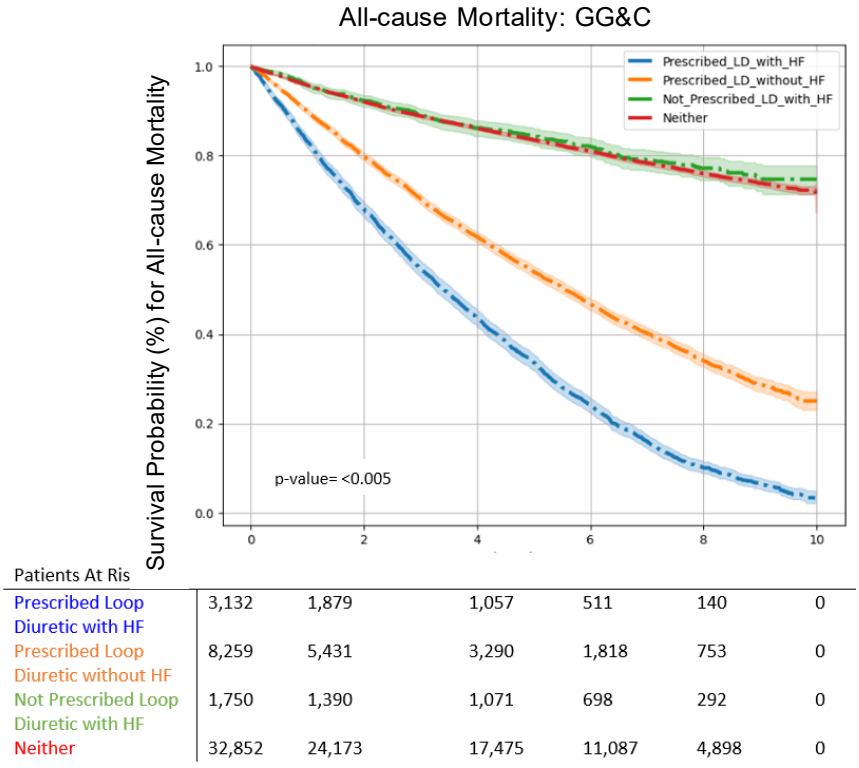
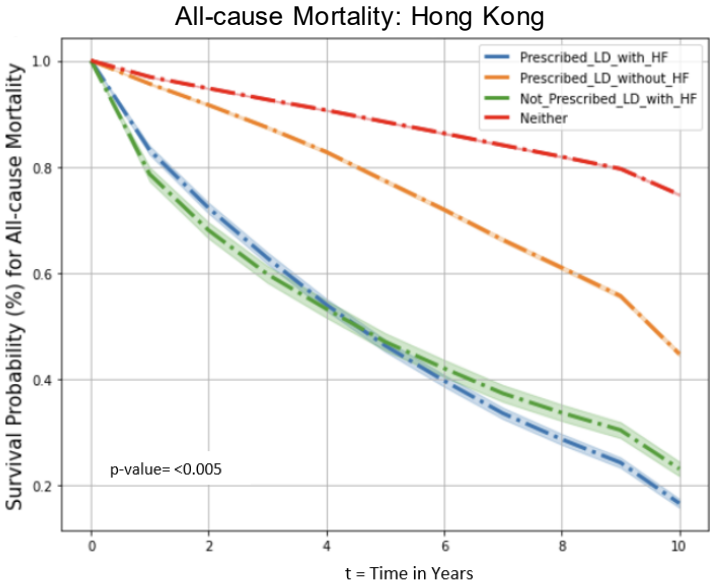


Figure 31 Kaplan Meier plot for all-cause mortality in the GG&C population

Hong Kong T2DM population: The prognosis of patients with heart failure and treated with loop diuretics was similar in Hong Kong (55%) and GG&C (65%) at 5 years as was the prognosis of patients who had neither feature (10% and 15% respectively). However, the prognosis of patients with heart failure in Hong Kong was similar whether or not they were receiving loop diuretics, which is very different from the outcome in GG&C. Patients treated with loop diuretics but without a diagnosis of heart failure had a better prognosis than those with heart failure, albeit still markedly impaired, which was somewhat different to the findings from GG&C.

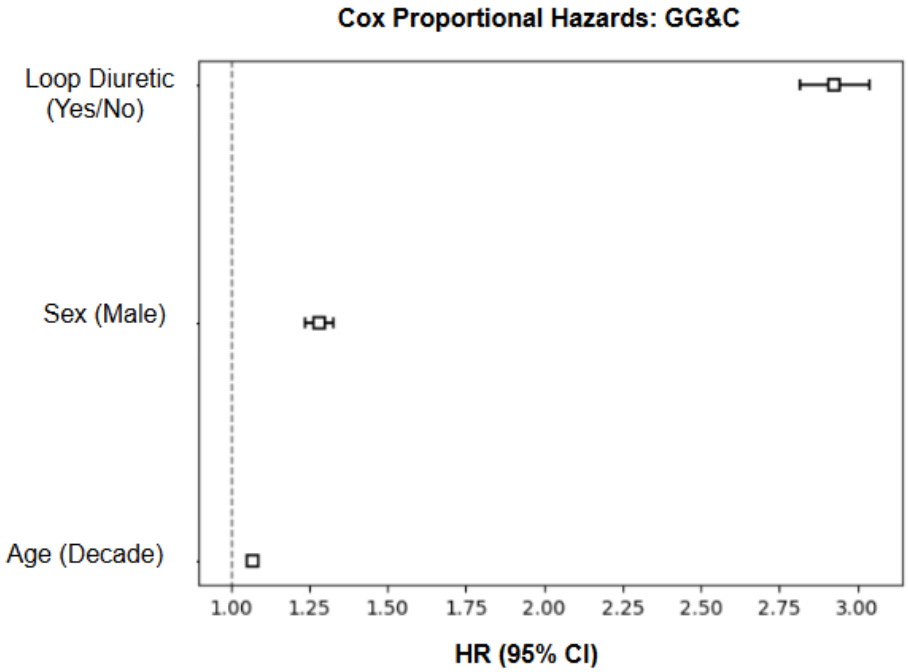
- **Blue Curve:** approximately 55% of patients with heart failure who were prescribed LD died within 5 years.
- **Orange Curve:** approximately 25% of patients without heart failure but who were prescribed loop diuretics died within 5 years.
- **Green Curve:** approximately 55% of patients with heart failure who were not prescribed LD died within 5 years.
- **Red Curve (Neither):** approximately 10% of patients without heart failure who were not prescribed LD died within 5 years.



Patients At Risk						
Prescribed Loop Diuretic with HF	7,531	5,437	4,064	2,985	2,151	0
Prescribed Loop Diuretic without HF	52,621	481,182	43,502	37,764	31,982	0
Not Prescribed Loop Diuretic with HF	3,658	2,488	1,945	1,536	1,229	0
Neither	210,066	198,982	190,253	181,030	171,768	0

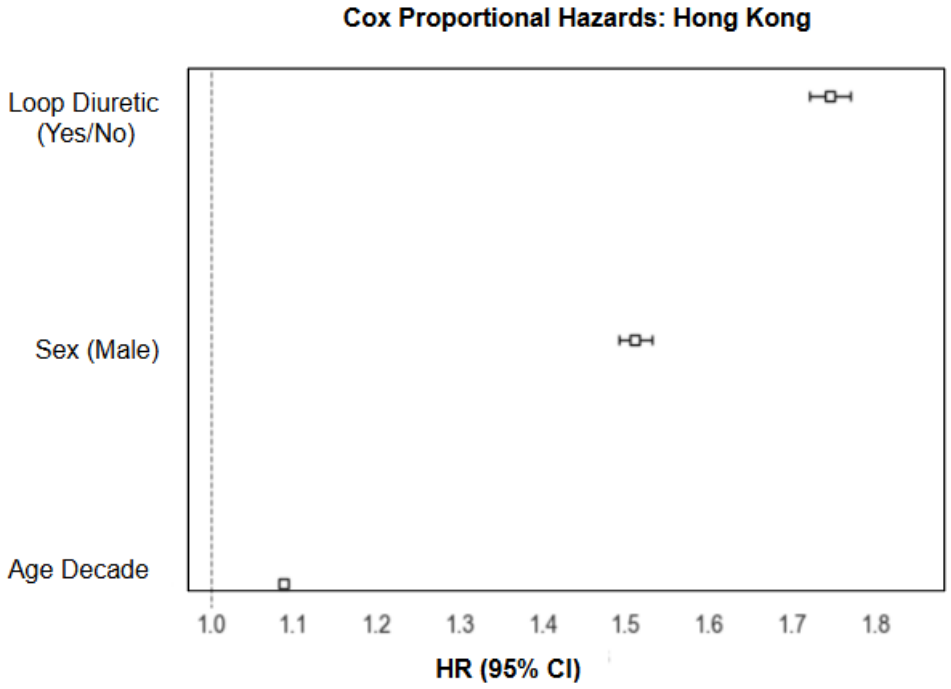
Figure 32 Kaplan Meier plot for all-cause mortality in Hong Kong

To strengthen and confirm results from the RSF model and Kaplan Meier visualisation further investigation of loop diuretics was carried out. **Figure 32** shows multivariable Cox regression adjusting for age and sex, showing a higher mortality amongst those prescribed loop diuretics in GG&C compared to those who were not. Men had a 28% higher mortality compared to women. Older age increased risk by 7% by decade. **Figure 33** shows the Hong Kong population where the use of LD was associated with a 1.75-fold increase in mortality compared to those not on LD. Men had a 51% higher risk compared to women and older age increased risk by 9% per year. For both populations the z-values and p-values indicate all variables are significantly associated with mortality ($p < 0.005$).



Cox Proportional Hazards Model Table: GG&C					
Variable	**Coefficient (coef)**	**Hazard Ratio (exp(coef))**	**95% CI (HR)**	**z-value**	**p-value**
Sex (Male)	0.25	1.28	1.24 - 1.33	13.23	p < 0.005
Age	0.07	1.07	1.06 - 1.07	68.83	p < 0.005
Loop Diuretic	1.07	2.93	2.82 - 3.04	54.98	p < 0.005

Figure 33 Cox Proportional Hazards: GG&C for All-cause Mortality



Cox Proportional Hazards Model Table: Hong Kong					
Variable	**Coefficient (coef)**	**Hazard Ratio (exp(coef))**	**95% CI (HR)**	**z-value**	**p-value**
Sex (Male)	0.41	1.51	1.43 - 1.59	6.61	$p < 0.005$
Age	0.08	1.09	1.00 - 1.09	2.13	$p < 0.005$
Loop Diuretic	0.56	1.75	1.72 - 1.77	7.98	$p < 0.005$

Figure 34 Cox Proportional Hazards: Hong Kong for All-cause Mortality

7.6.5 Potential Reasons for Prescribing Loop Diuretics in GG&C and Hong Kong

Only a minority of patients prescribed loop diuretics had a diagnosis of heart failure, end-stage renal disease or resistant hypertension. **Table 33** show the potential reasons for prescribing loop diuretics in GG&C and Hong Kong. Columns in **blue** are not mutually exclusive; a patient could have heart failure, end-stage renal disease and hypertension and would therefore count in all three columns. Resistant hypertension was defined as ≥3 anti-hypertensive therapies. End-stage renal disease was defined as eGFR <30 mL/min/1.73m2.

In GG&C, of 11,403 patients prescribed a loop diuretic 45% had one or more of the above reasons for their use, with heart failure (27%) being the major reason. In Hong Kong, of 60,152 prescribed a loop diuretic 40% had one or more of the above reasons for their use, with resistant hypertension being the major reason. However, in most cases there was no obvious reason for their prescription

Table 33 Potential reasons for Prescribing Loop diuretics in GG&C and Hong Kong

Potential Reasons for Prescribing Loop Diuretics: GG&C						
	Total	**No Conditions**	**≥1 Condition**	**Heart Failure**	**Resistant Hypertension**	**End Stage Renal Disease**
Population N=	46,031	36,295 (79%)	9,736 (21%)	4,675 (10%)	4,859 (11%)	2,549 (6%)
Not Prescribed Loop Diuretic	34,628	30,049 (87%)	4,579 (13%)	1,587 (5%)	3,347 (10%)	1,746 (5%)
Prescribed Loop Diuretic	11,403	6,246 (55%)	5,157 (45%)	3,088 (27%)	1,512 (13%)	803 (7%)
Potential Reasons for Prescribing Loop Diuretics: Hong Kong						
	Total	**No Conditions**	**≥1 Condition**	**Heart Failure**	**Resistant Hypertension**	**End Stage Renal Disease**
Population N=	273,876	220,883 (81%)	52,993 (19%)	11,189 (4%)	46,402 (17%)	3,381 (1%)
Not Prescribed Loop Diuretic	213,724	184,482 (86%)	29,142 (14%)	3,658 (2%)	26,402 (12%)	1,332 (1%)
Prescribed Loop Diuretic	60,152	36,301 (60%)	23,851 (40%)	7,531 (13%)	20,000 (33%)	2,049 (3%)

7.7 Discussion

This analysis of a large number of patients with T2DM drawn from for two very different geographical, culturally and ethnically populations shows that a rather similar set of variables predicts mortality using different statistical and ML models. These models consistently found that age, treatment with loop diuretics, serum albumin and eGFR were amongst the five strongest predictors of an adverse prognosis.

In terms of demographics, the age and sex distribution of the two population was remarkably similar, although due to the large number of patients small differences were nonetheless statistically significant. However, one key demographic difference was smoking. Hong Kong's effective tobacco control measures have contributed to the lowest smoking prevalence among high-income regions since 1990, which may have contributed to increased life expectancy (Ni et al., 2021). There were very few records on smoking (10%). Smoking is a well-established risk factor for cardiovascular disease and its exclusion from the analysis—represents a limitation. This may introduce unmeasured confounding, potentially biasing risk estimates. However, including unreliable or sparsely recorded data allows greater noise into the models, justifying its exclusion on methodological ground.

Regarding few records, the Hong Kong population in this study showed less comorbidities than the Glasgow population, which may have contributed to slightly better health outcomes. There was 90% missing BMI in the Hong Kong T2DM population, however medical literature suggests that Chinese Asians have a lower BMI cut-off for health risks (Wise, 2021) than Caucasian populations. This may be tied to differences in body composition across ethnic groups.

At slightly higher proportion of patients in Hong Kong were women. This could be attributed to various demographic and cultural factors. The Women's Commission (WoC) of Hong Kong (a central body that promotes women's development and well-being) indicates that women tend to have more frequent healthcare visits compared to men (WoC, 2019). Among high-income populations, Hong Kong recorded the lowest cardiovascular mortality and one of the lowest cancer mortalities in women (Ni et al., 2021). Greater healthcare engagement among women may explain their increased representation in this T2DM population.

Patients with T2DM were included regardless of co-morbidity or treatment, which may account for the high use of loop diuretics and provide insights into why they were prescribed and the outcome of patients who are treated with them. Women may be prescribed loop diuretics more than men due to several factors, including greater under-recognition of HF, leading to delayed diagnosis and more frequent prescription of diuretics as a symptom-management strategy before an official HF diagnosis is confirmed (Zakeri et al., 2021b).

Furthermore, in both the UK (Theresa A. McDonagh et al., 2021) and Chinese (Guo et al., 2016) populations, women are more likely to present with heart failure with preserved ejection fraction (HFpEF), which is a more difficult diagnosis than heart failure with reduced ejection Fraction (HFrEF). Additionally, women tend to live longer and develop comorbid conditions like hypertension and atrial fibrillation, which can also lead to increased prescription of diuretics (Rosengren, 2024).

In both GG&C and Hong Kong, patients with HF and taking loop diuretics had the worst prognosis and patients taking loop diuretics but without HF had a much worse prognosis than those who had neither feature. However, patients with HF who were not taking loop diuretics had a very different outcome between the two populations. This was the least common profile in both datasets but still comprised several thousand patients, making it likely that the observation is real. It is possible that loop diuretic prescriptions were missed in some patients from Hong Kong. However, it is also possible that the criteria used to diagnose HF differed between these population. Surprisingly, there is no robust definition of HF (Cleland et al., 2021) making this a distinct possibility.

The green and red survival curves overlapping in the Hong Kong cohort suggests that some patients with diagnosed heart failure, but not prescribed loop diuretics, have survival trajectories nearly identical to patients without heart failure or diuretic use. This could reflect a subgroup of patients with milder HF, potential misclassification, or treatment decisions based on low clinical risk.

Furthermore, the differing hazards of all-cause mortality associated with loop diuretics between the Glasgow and Hong Kong populations likely reflect variations in prescribing practices, disease severity at initiation, and healthcare system context.

In Scotland, loop diuretics are frequently prescribed, even in the absence of a formal heart failure diagnosis (Friday et al., 2024). In contrast, prescribing patterns in Hong Kong may reflect a more selective use, potentially indicating more advanced or acute presentations (Leung et al., 2015b). These differences highlight the importance of interpreting medication-related associations within the clinical and systemic context of each setting.

Higher HbA1c and cholesterol levels in the Hong Kong population may be linked to both dietary and genetic factors, as well as differences in diabetes management and treatment guidelines across regions. For example, East Asians are known to have different responses to certain treatments that affect glucose and lipid metabolism (Kodama et al., 2013). They are more insulin sensitive than other ethnicities but develop pancreatic beta-cell dysfunction with reduced secretion of insulin (do Vale Moreira et al., 2021), resulting in poorer blood glucose control and elevated HbA1c levels. In people of European origin, obesity and insulin resistance is the more common clinical phenotype.

This study highlights many similarities and some differences in prescribing for patients with T2DM in GG&C and Hong Kong. Patients in Hong Kong were more likely to be prescribed treatments to reduce blood glucose, whereas in Glasgow many patients were not started on pharmacological therapy in the first 6 months; presumably dietary measures were tried initially. The most popular choices for pharmacological control of diabetes in both GG&C and Hong Kong were metformin and sulfonylureas. Patients in GG&C were more likely to be prescribed statins, perhaps reflecting the higher prevalence of atherosclerotic disease.

Patients in GG&C were more likely to receive ACEi/ARB but less likely to receive CCB, agents that are prescribed predominantly for hypertension (Wong et al., 2008). A similar proportion of patients were prescribed loop diuretics in both geographies, overall (25% in GG&C; 22% in Hong Kong) and amongst patients with HF (66% and 67% respectively). In summary, despite differences in geography, ethnicity, culture and healthcare infrastructure, the treatment of these two populations was remarkably similar.

7.8 Conclusion

Applying a novel machine learning approach to EMR for patients with T2DM in GG&C and Hong Kong identified a similar set of predictors for mortality and predictive models that performed similarly well. This suggests that the models developed should be generalisable to other populations. In both GG&C and Hong Kong, treatment with loop diuretics was strongly associated with mortality, which is a novel finding for patients with T2DM. This may be because many of these patients had undiagnosed heart failure, but inappropriate use of loop diuretics might accelerate the progression of CVD and increase the risk of sudden death due to hypokalaemia.

Chapter 8 Discussion & Conclusions

8.1 Introduction to the Discussion

This research examined the incidence and prediction of heart failure and all-cause mortality in patients with T2DM using EMRs from two from two geographically, ethnically and clinically diverse populations. Differences across cohorts was explored, with a focus on cardiovascular comorbidities. The significance of prescribing loop diuretics was thoroughly investigated. Additionally, the importance of social deprivation was explored in the Glasgow cohort. Through a precision medicine approach, this research constructed robust risk prediction models. These models were built with survival analysis, including the time-to-event aspect. This allowed for risk predictions that account whether or when an event (heart failure or death) is likely to occur. Unlike simpler models like logistic regression, which provide only a binary outcome over a fixed time, survival analysis handles censored data. It also identified key risk factors for incident heart failure and mortality with interpretability through explainable artificial intelligence techniques and clinical expertise.

8.2 Summary of Key Findings

In both Glasgow and Hong Kong loop diuretics were an important predictor for both incident heart failure and mortality, possibly because loop diuretics are a marker of prevalent but undiagnosed disease or because loop diuretics accelerate disease progression. In both cohorts, older women and those with impaired renal function were more likely to be prescribed loop diuretics. The robustness and generalisability of the predictive models varied slightly between populations with slightly greater predictive accuracy in Hong Kong. Models based on interpretable machine learning methodology outperformed traditional regression models and provided insights into the importance of various risk factors/predictors for an individual patient.

8.3 Interpretation of Baseline Characteristics Differences

The statistical significance of minor differences in the characteristics between large cohorts should be interpreted with care (**Chapter 7**). They may reflect subtle differences in demographic profile, genetics, culture and diet and access to healthcare and hence duration of disease before diagnosis and may not be clinically meaningful.

Comparing the Glasgow and Hong Kong populations, age and sex distributions were remarkably similar. The Glasgow dataset had richer demographic information, with an excess of patients with T2DM in the most deprived quintile of the population. An area-based marker of socioeconomic status (SES) was used, rather than individual-level indicators. The area-based measure can reflect collective environmental and social characteristics that may directly influence health outcomes. For example, more deprived areas may have reduced access to healthy food options, green spaces, or recreational facilities, while facing greater exposure to fast food options, alcohol and tobacco advertising and other environmental stressors. However, area-based measures like SIMD can misclassify individuals, as not everyone in a deprived area is personally disadvantaged. Still, they are useful and justified for population-level studies.

Although there was missing BMI is both populations, Glasgow had better coverage (70% complete records). Furthermore, a slightly higher percentage of women in the Hong Kong cohort may reflect cultural factors that influence healthcare engagement (Women's Comission, 2021). Hong Kong's Women's commission organisation helps increase women's healthcare awareness, which may encourage healthcare utilisation and earlier diagnosis. In the UK, there may be delays in the diagnosis of T2DM (Sattar, 2013) and in the detection and management of cardiovascular disease (Bakker, 2019). This highlights the need for increased awareness to address challenges women face in the diagnosis and treatment of T2DM and cardiovascular events.

Glasgow and the West of Scotland (Health Intelligence Team, 2024b) have high age-adjusted rates of cardiovascular disease and COPD compared to the rest of the UK. The prevalence of coronary artery disease and COPD was also higher in Glasgow than in Hong Kong in the current analysis. Both cardiovascular and respiratory disease (WHO, 2024) are strongly associated with smoking (Public Health Scotland, 2024). In the Glasgow cohort, 20% of patients were current

smokers. Although data on smoking was only 10% recorded in the Hong Kong EMRs, education and health policy have reduced the prevalence of smoking in Hong Kong to <10% (Hackshaw et al., 2018; Socrates Y WU1, 2021).The general adult smoking prevalence in Scotland was approximately 11% in 2021, with higher rates (24%) observed in more deprived areas, including Greater Glasgow and Clyde (Scottish Government, 2023). The cohort data and general population estimates differ in base definitions, the figures suggest that smoking is likely to be more prevalent in the Glasgow cohort than in the general population of Hong Kong.

Only 10% of patients from Hong Kong and 70% from Glasgow had a recorded value for BMI. The substantial amount of missing BMI data is a major limitation. In the Glasgow cohort, most patients were overweight or obese. The prognosis of patients with T2DM and a normal BMI or who were underweight was much worse than the prognosis of those who were obese.

Obesity is a stress that increases blood glucose (Klein et al., 2022). The development of T2DM in the absence of such a stress might indicate more severe metabolic disease. Patients with missing BMI values had a similarly poor prognosis to those who had a normal BMI or were underweight, indicating that missing BMI values were informative. Perhaps, when BMI is normal it is considered unremarkable and is less likely to be recorded. Asians are reported to be at risk of developing T2DM at a lower BMI threshold, possibly due to higher levels of visceral adiposity (Ma and Chan, 2013), and therefore BMI classifications for obesity should be ethnicity specific.

There were minor differences in clinical laboratory tests between populations. Blood concentrations of HbA1c, total cholesterol, AST and neutrophils were higher in the Hong Kong cohort when compared to the Glasgow cohort. Higher HbA1c levels indicates poorer glycaemic control among T2DM patients in Hong Kong, which may reflect delays in diagnosis and treatment (diet, exercise and pharmacological). A study of T2DM (Wan et al., 2023) from Hong Kong and the UK found that poor HbA1c control was associated with worse cardiovascular outcomes. A larger proportion of Hong Kong patients fell into the HbA1c (≥8%) subgroup compared to the UK cohort. The higher HbA1c levels observed in Asian populations might increase the risk of CVD and higher mortality (Wan et al., 2016), further exacerbated by higher serum cholesterol, either due to higher intrinsic values (Seah et al., 2023) or lower rates of statin use. Serum AST concentrations were also higher in Hong Kong which might reflect the effects

of T2DM and metabolic syndrome on liver function ("fatty" liver disease) or more liver disease due to a higher prevalence of Hepatitis C. Increased neutrophil counts also suggest a higher state of inflammation, which has been associated with worse cardiovascular outcomes (He et al., 2023) in T2DM populations.

One other important difference was the lower eGFR in the Glasgow cohort. However, other reports suggest that some Asian groups with T2DM may be more susceptible to developing impaired renal function compared to those of European descent (Wen et al., 2022). In the UK, kidney failure is up to five times more common in people from ethnic minority backgrounds(Kidney Research UK, 2018). Renal dysfunction is known to be an important risk factor for incident heart failure and mortality whether or not patients have T2DM.

8.4 Investigation of Loop Diuretics

Congestion is an essential feature of heart failure, causing symptoms such as breathlessness and swelling of the legs. Loop diuretics increase renal water and salt excretion and are the mainstay of treatment of congestion and its symptoms and signs (McDonagh et al., 2023). This research highlights that many patients with T2DM are treated with loop diuretics without first receiving a diagnosis of heart failure. These patients are not only more likely to be diagnosed with heart failure at a later date but also have a prognosis similar to patients with heart failure, although they often die without ever receiving such a diagnosis. In the Glasgow cohort, 25% of patients were prescribed loop diuretics, compared to 23% in the Hong Kong cohort. Older women were more likely to be prescribed loop diuretics than other groups. In both cohorts, those prescribed loop diuretics had on average, poorer renal function. The strongest predictor of mortality in both populations was loop diuretic use. In both Glasgow and Hong Kong, the exclusion of loop diuretics from the predictive model reduced its performance.

A growing body of evidence confirms that loop diuretic prescriptions often reflect a missed diagnosis of HF and that those treated with loop diuretics have, at least in some populations, a similar prognosis whether or not they receive a diagnosis of heart failure (Friday et al., 2024; Cuthbert et al., 2024). However, it is unclear whether loop diuretics are merely a marker of undiagnosed heart failure, or whether they accelerate the progression of cardiovascular disease

and renal dysfunction (Amatruda et al., 2022; Wilcox et al., 2020) by activating the renin-angiotensin-aldosterone (RAAS) and sympathetic nervous systems or whether electrolyte disturbances, particularly a low serum potassium, increase the risk of sudden death. All of these might be true, the importance of each varying from one patient to the next.

Women are more likely to have heart failure with preserved ejection fraction (HFpEF) (Sotomi et al., 2021), a phenotype that is commonly underdiagnosed because the left ventricular ejection fraction is not reduced and heart function may not appear severely impaired to the non-expert eye. However further analyses are required to support this idea. Given the increasing emphasis on personalised medicine, understanding the role of loop diuretics in underdiagnosed HF phenotypes could provide new opportunities to improve outcomes in these high-risk T2DM patients.

The adverse renal effects of loop diuretic therapy are well-established. Loop diuretics can provide rapid relief from HF symptoms, but their long-term use is linked to electrolyte imbalances, activation of the RAAS and progression of CKD (Amatruda et al., 2022; Wilcox et al., 2020). In this research, patients in both Glasgow and Hong Kong who were prescribed LDs experienced higher rates of CKD events (Chapter 4 and 5). The pathophysiological link between LDs and CKD highlights the importance of regular renal function monitoring in patients with T2DM, who are at high risk of renal complications. Current guidelines (McDonagh et al., 2023) recommend using LDs appropriately and combining them with RAAS inhibitors to mitigate their adverse effects. Overall, the findings show the need for greater efforts to differentiate HF from other conditions and for population-specific research to optimise LD use in diverse populations.

8.5 Justification of Risk Prediction Model(s) Methods

The model(s) in this thesis were thoroughly validated and calibrated. The risk prediction models showed only small differences in c-index scores differs between cohorts, with slightly stronger predictive performance in Hong Kong compared to the Glasgow cohort for incident heart failure (C-index 0.88 and 0.87) and all-cause mortality (0.85 and 0.83). Both models identified similar risk factors, including loop diuretics use, older age, lower serum albumin and ALT

concentrations, lower haemoglobin, and lower eGFR, coronary artery disease, atrial fibrillation and stroke. The consistency of risk factors and model performance in two very different cohorts of T2DM, confirm the importance of these clinical variables for predicting outcomes in patients with new-onset T2DM. Moreover, many of these risk factors are biologically plausible mechanisms underlying disease progression.

The risk prediction models developed in this research for incident heart failure and mortality outperformed others (Basu et al., 2017; Yang et al., 2008; Hippisley-Cox and Coupland, 2015; Dong et al., 2024; Quan et al., 2019), the majority of which reported a c-index below 0.80.

The novelty of this research lies in advancing beyond traditional regression analyses that have been the mainstay of previous clinical risk prediction models, despite their limitations (Barraclough et al., 2011; Jiang et al., 2024). Harnessing the random survival forest method reflects an intentional emphasis on tree-based decision-making over traditional regression mechanisms. To date, few studies have confirmed that machine learning-based survival models outperform cox regression models for incident of HF (Segar et al., 2019) or mortality (Lee et al., 2021). Tree-based methods readily adapt to diverse patient populations (Ishwaran, Udaya B. Kogalur, et al., 2008), making them efficient for personalised risk prediction.

The model developed in this thesis yielded superior discrimination and calibration metrics, demonstrated by low time-Brier scores. Applying a second evaluation metric strengthens the results rather than relying solely upon the c-index. Complex models may achieve high c-index scores by overfitting the data (Hartman et al., 2023), creating a false sense of reliability in model performance. The time-Brier score addresses this limitation by assessing both discrimination and calibration. This approach to model development ensures that predictions are accurate not only in ranking risk but also in reflecting time-to-events. Whereas standard evaluation metrics do not account for time-to-events prediction. RSF ensures that predictive accuracy is maintained over time, a critical feature for CVD disease progression.

Subsequently, the use of SHAP (SHapley Additive exPlanations) (Brosula et al., 2024) provided patient-specific risk interpretations, overcoming the limitations of conventional regression models that fail to capture nonlinear interactions between variables. In this research, the inclusion of critical comorbidities such as atrial fibrillation and hyperkalaemia (from the

Glasgow cohort) enhanced performance metrics. RSF manages complex EMRs, which regression models, constrained by their linear assumptions, struggle to achieve. SHAP explains the contributions of specific variables rather than describing association. It focuses on the reasoning of each variable on the model's prediction. SHAP's ability to explain individual contributions (**Chapter 4.5.8**) makes it valuable in healthcare, where understanding why a model predicts a high risk for an individual patient can guide personalised interventions.

This research also applied causal inference methods, such as propensity score matching and inverse probability weighting, to try to address biases associated with loop diuretics prescriptions. Propensity matching reduced model performance in the Hong Kong EMRs for incident HF. However, propensity matching to investigate the importance of a therapeutic intervention that is already in place and that alters prognostically important patient characteristics might introduce systematic bias rather than reduce it. Propensity matched analyses have rarely been replicated by randomised trials in cardiovascular medicine, perhaps because the matching process fails to consider the effects of treatment on other risk markers. This might be overcome if the patient characteristics prior to the intervention of interest are used for propensity matching, but this is rarely done.

8.6 The Inclusion of Socioeconomic Status

Socioeconomic deprivation is associated with poorer education, a higher prevalence of cardiovascular risk factors (e.g., smoking, alcohol excess, obesity and hypertension), poorer access to healthcare and an increased morbidity and mortality rates (Witte et al., 2018; Health Scotland, 2015; Wright et al., 2019; Rosengren et al., 2019). In the Glasgow cohort, 41% of patients with T2DM were in the most deprived socioeconomic quintile, with a high proportion of women, smokers, people with obesity, lung disease and treatment with loop diuretics. Patients with T2DM in the most deprived quintile had a 36% higher mortality compared to those in the least deprived quintile, even after adjusting for age and sex (HR: 1.36 [95% CI 1.24–1.50, $p<0.005$]). These finding highlights the need to include SES in predictive models. This supports algorithmic fairness where most deprived individuals are not excluded. It is also important to encourage better SES measurement strategies in populations, as the Hong Kong cohort did not have a useable measure of social deprivation.

8.7 Strengths and Limitations

The key strength of this research is applying machine learning-based survival models using EMR from two diverse populations. The results appear clinically relevant, robust and generalisable. In studies using EMRs with variable data quality and completeness, external validation strengthens the reliability of predictions and enhances their applicability to broader clinical settings.

Throughout this research, integrating clinical expertise with computational methods, introduced the concept of collaborative intelligence, leveraging the strengths of both approaches, with iterative refinement of the risk model. Moreover, the integration of computational decision-making and clinical expertise helps ensure the model's clinical validity and acceptance (Sirocchi et al., 2024). For example, the correlations analysis in this research was guided by clinicians and results were iteratively checked. This evidence also introduced a clinical support tool for heart failure risk assessment in patients with T2DM (**Chapter 5**).

The tool has not yet been implemented in clinical practice, but it addresses a recognised need for early heart failure risk assessment in patients with T2DM. However, the models have some limitations that should be considered. Several key variables, including albumin-to-creatinine ratio, a key marker of kidney function, BMI, alcohol consumption and blood pressure were either not available or were incomplete but with informative missingness. Addition of these variable would likely improve the accuracy of the model and increase its value in terms of explaining the importance of risk factors, especially those that are modifiable. Information on cause of death was lacking in the Hong Kong cohort. However, it was important not to use bias imputation methods (Sterne et al., 2009) as these skews results and true patterns may be masked.

Future research should aim to address these gaps by including these missing variables. Moreover, while this thesis showed the potential of AI-driven support tools (Narinder Kaur et al., 2024), future work should focus on validating these models across more diverse populations and exploring the feasibility of deploying them in real-world healthcare settings.

8.8 Future Work

There are several opportunities for further research. One is the development of AI-powered evidence-based support tools that leverage EMRs. The rich data-environment of the NHS Safehaven and CDARS EMRs included many patient characteristics, including medical history, laboratory results, prescriptions and demographic information, provides a robust foundation for creating patient-specific risk prediction models, capable of identifying early indicators of disease and guiding targeted interventions for prevention and treatment. To enhance the clinical utility of AI models for T2DM, future research should prioritise improving model generalisability across more diverse patient populations.

Leveraging large-scale, longitudinal datasets that capture patient trends over time will be crucial for accurately predicting CVD disease progression and complications.

Time series analysis presents a promising approach in this context, as it allows for the capture of temporal patterns and fluctuations in patient data. By integrating time-dependent factors such as blood glucose levels, medication adherence and other biomarkers, time series models can offer dynamic, evolving risk assessments that adapt to the changing health status of individual patients.

Moreover, conformal prediction, where calibration is conducted on separate training and testing sets could further strengthen these models (Angelopoulos and Bates, 2021). This technique improves the reliability of predictions by quantifying uncertainty, ensuring that the model's confidence aligns with observed data. By integrating both time series analysis and conformal prediction into AI-driven clinical decision support tools, the transparency, accuracy and interpretability of predictions can be enhanced. This, in turn, would contribute to more precise, timely interventions, improving patient outcomes and enabling the effective application of precision medicine in managing T2DM and its associated complications.

Furthermore, these support tools may eventually be deployed into healthcare organisations, facilitating widespread adoption and integration into routine clinical practice, thus improving decision-making and care delivery across diverse healthcare settings.

8.9 Conclusions

The development of real-time, AI-powered risk model with a user-friendly clinical interface holds great promise for informing patients and their clinicians, about outcomes. Moreover, the proposed risk model can also provide personalised information about potentially modifiable risk factors requiring further investigation and management to improve outcomes. This or similar tools for T2DM and many other conditions should now be developed and integrated into existing EMR for further validation before deployment at scale. Development of intelligent systems that learn to improve prediction will further increase the accuracy of the model and increase the utility of decision-support tools to help patients and clinicians to choose the best treatment for their needs.

Appendices
Appendix A Chapter 3

A1: Data Ethics

The research project required ethical approval since patient data was examined. Throughout the execution of the research project, it was vital to confirm the regulations of The UK Data Protection Act 2018. Based on the regulations, patient data was kept precise, adequate and up to date throughout the entire research project (NHS, 2019). Lastly, safety of patient data is assured by accessing a security server (ISO 27001) with frequent password renewal. Before accessing the NHS SafeHaven environment, Medical Research Council – research, GDPR & confidentiality training was undertaken in July 2021. Ethical approval was given by NHS Greater Glasgow & Clyde (project reference number is GSH/20/CE/004).

A2: Diagnostic Descriptions

This table represents the categorisation of diagnostic description from SCI Diabetes registry.

Diagnostic Description	***Category***
Type 2 Diabetes Mellitus	Type 2
o Impaired Glucose Tolerance o Impaired Fasting Glucose o Impaired Glucose Metabolism and Other Not Known	Risk of Type 2
o Diabetes in Remission o Diabetes Resolved o Other o Diseases of exocrine pancreas Secondary - Medicine Induced o Induced by steroids. o Secondary - Pancreatic Pathology Medicine- or Chemical-induced Immune-mediated (LADA) o Pancreatitis	Other Types

o Latent Autoimmune Diabetes of Adulthood Malnutrition-related diabetes mellitus Gestational Diabetes (Current) Fibrocalculous pancreatopathy Haemochromatosis o Cystic fibrosis o Induced by non-steroid medicines o Type 2 diabetes-former diagnosis o Stress-induced hyperglycaemia Neoplasia	
o Diabetes not confirmed o Not Diabetic o No o Not defined	Not Defined
Type 1 Diabetes Mellitus	Type 1

A3: Extracting Prescriptions from Glasgow SafeHaven Dataset

```
import time

import pandas as pd

from datetime import datetime, timedelta

# Start timing the data loading

start_time = time.time()

# Load necessary columns from the CSV file

pharm_df = pd.read_csv(

    "/path/to/your/03_Extract_Pharmacy.csv",

    usecols=['SafeHavenID', 'DISP_DATE', 'PI_BNF_Section_Code',
'PI_BNF_Section_Description', 'Dispensed_Quantity'],

    encoding='latin-1',
```

```
    dtype={
        'PI_BNF_Section_Code': 'category',
        'PI_BNF_Section_Description': 'category'
    }
)

# Print the time taken to load the data
end_time = time.time()
print(f"Time taken to load: {end_time - start_time:.2f} seconds")
# Display the shape of the loaded DataFrame
print(f"Pharmacy DataFrame shape: {pharm_df.shape}")

# Merging with another DataFrame (example)
# Assuming `df` contains patient event data
merged_df = pd.merge(
    df.reset_index(drop=True),
    pharm_df.reset_index(drop=True),
    how='left'
)
# Check the shape of the merged DataFrame
print(f"Merged DataFrame shape: {merged_df.shape}")
# Convert date columns to datetime format
merged_df['DATE'] = pd.to_datetime(merged_df['DATE'])
```

```
merged_df['PRESC_DATE'] = pd.to_datetime(merged_df['PRESC_DATE'])

# Calculate differences in days between diagnosis date and prescription date

merged_df['presc_days'] = (merged_df['PRESC_DATE'] - merged_df['DATE']).dt.days.abs()

# Display a summary of prescription years

merged_df['Presc_Year'] = merged_df['PRESC_DATE'].dt.year

print(merged_df['Presc_Year'].value_counts())

# Output the processed DataFrame for further analysis

print(merged_df.head())
```

A4: Calculating Mortality Outcome

```
# Convert date columns to datetime format
df['Date_of_Diabetes'] = pd.to_datetime(df['Date_of_Diabetes'])
df['Date_of_Death'] = pd.to_datetime(df['Date_of_Death'])

# Calculate the duration in days between Date_of_Diabetes and Date_of_Death
df['Duration'] = (df['Date_of_Death'] - df['Date_of_Diabetes']).dt.days

# Display the DataFrame
print(df[['SafeHavenID', 'Date_of_Diabetes', 'Date_of_Death', 'Duration']])
```

A5: Heart Failure at any Diagnostic Position

In [1773]:

```
# Define a dictionary where keys are condition names and values are lists of keywords for each conditio

condition_keywords = {
    'incident_HF': ['I500', 'I5009', 'I5099', 'I5091', 'I5000', 'I110', 'I130',
                    'I139', 'I132', 'I509', 'I50', 'I42'],
    }

# Create new columns for each condition and populate with 1 or 0 based on
# events before diabetes diagnosis

for condition, keywords in condition_keywords.items():

    # Create a boolean mask for each keyword in the condition

    keyword_masks = [
        hf_df['DIAG1'].str.contains(keyword, case=False, na=False) |
        hf_df['DIAG2'].str.contains(keyword, case=False, na=False) |
        hf_df['DIAG3'].str.contains(keyword, case=False, na=False)
        for keyword in keywords

    ]

    # Combine keyword masks with logical OR to check if any keyword matches

    condition_mask = pd.concat(keyword_masks, axis=1).any(axis=1)

    # Apply additional filtering based on the date of diabetes diagnosis

    #condition_mask = condition_mask & (filtered_df['DATE'] < filtered_df ['DATE'])

    hf_df[condition] = condition_mask.astype(int)
```

Appendix B Chapter 5

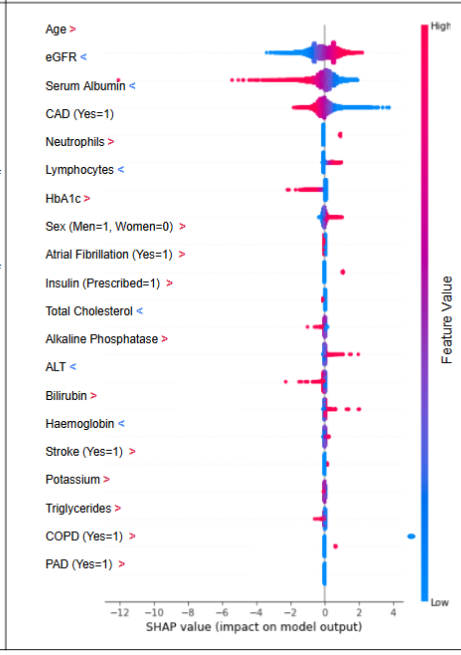
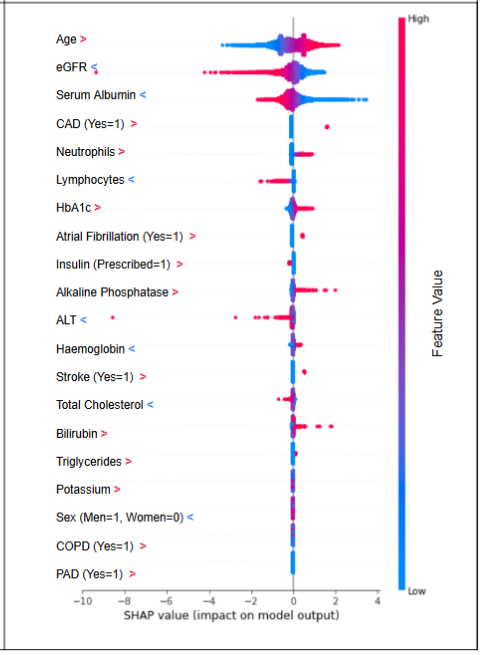
B1: A Secondary Analysis of Patients with complete BMI only

This table shows patients with T2DM whom had a complete record of BMI only. In this case the most important risk factor for developing heart failure was those with established coronary artery disease and older age. There is a clinical hypothesis that those with measured BMI are patients who were carefully monitored as there is 90% missingness in the Hong Kong population. With a c-index score of 0.90 and time brier of 0.07.

	Random Survival Forest with Complete BMI Only (N=23,973)
***C-statistic Score**	0.90
***Time Brier Score**	0.07
Key predictors of HF in rank order. **The x-axis shows how much each factor contributes to predict incident HF.** **>** Indicates higher values of this variable increase risk of HF **<** Indicates lower values of this variable increase risk of HF	1. CAD (Yes=1) 2. Age **>** 3. Loop Diuretic (Prescribed=1) 4. Lymphocytes **<** 5. Atrial Fibrillation (Yes=1) 6. eGFR **<** 7. Neutrophils **>** 8. Insulin (Yes=1) 9. Total Cholesterol **>** 10. Potassium **>** 11. Alkaline Phosphatase **>** 12. Serum Albumin **<**
***C-statistic Score: Model discrimination.** ***Time Brier Score: Measure of calibration for time-dependent models. A lower Brier score (0-1) indicates greater accuracy.** **Abbreviations:, coronary artery disease (CAD), estimated glomerular filtration rate (eGFR), haemoglobin A1c (HbA1c) and heart failure (HF).**	

B2: Incident Loop Diuretics in Hong Kong

This table presents incident loop diuretics in Hong Kong patients with T2DM. The table shows results for including HF as a variable and excluding HF. Patients who had previous prescriptions for loop diuretics prior to their diagnosis of T2DM were excluded from the analysis to ensure that the study focused on incident use of loop diuretics following T2DM diagnosis. 3,008 (1.4%) patients were newly prescribed loop diuretics. Results show consistency in patient characteristics for predicting incident HF.

Hong Kong Cohort	**Model 1 Predicting Incident Loop Diuretics without Heart Failure (N=210,066)** ***C-statistic Score:** 0.87 ***Time Brier Score:** 0.03	**Model 2 Predicting Incident Loop Diuretics with Heart Failure (N=210,066)** ***C-statistic Score:** 0.88 ***Time Brier Score:** 0.02
Key predictors of Incident Loop Diuretics in rank order. **The x-axis shows how much each factor contributes to predict incident Loop Diuretic.** **>** Indicates higher values of this variable increases the likelihood of being prescribed Loop Diuretics **<** Indicates lower values of this variable increases the likelihood of being prescribed Loop Diuretics ***C-index: Model discrimination.** ***Time Brier: Measure of calibration for time-dependent models. A lower Brier score indicates greater accuracy.**		

Appendix C Chapter 6

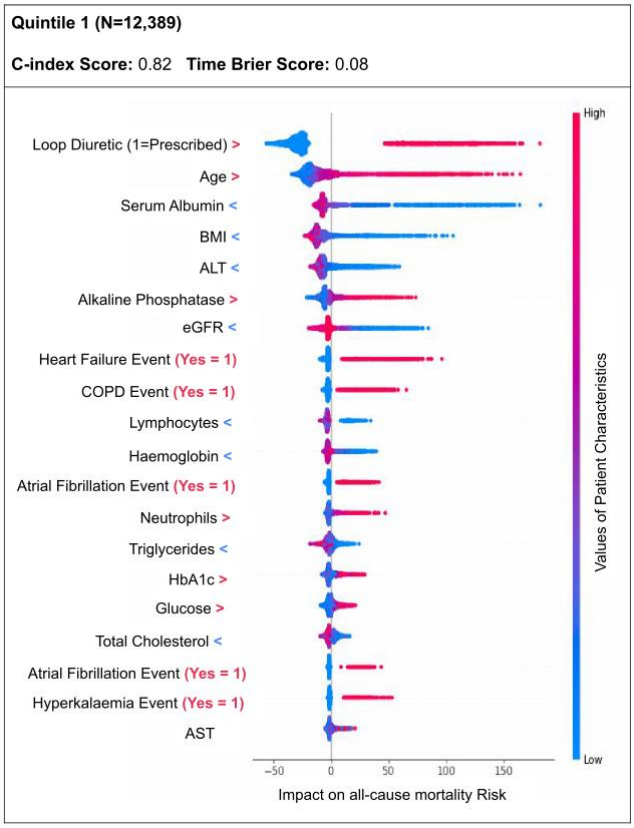
C1: Clinical characteristics of T2DM patients with a measurement of BMI stratified by socioeconomic deprivation status

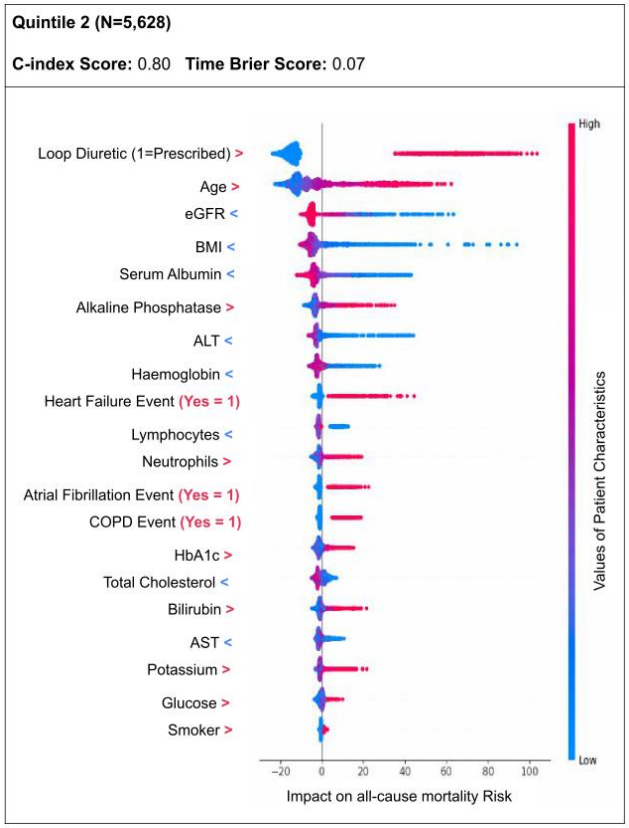
Overall Population N=30,495 Demographics at Baseline (Secondary) (%) or Median (25/75)	SIMD Quintile 1 (Most Deprived)	SIMD Quintile 2	SIMD Quintile 3	SIMD Quintile 4	SIMD Quintile 5 (Least Deprived)	P-value Overall	P-value Q1 & Q5
N=Scottish Index of Multiple Deprivation (SIMD) Group	N=12,389	N=5,628	N=4,162	N=3,602	N=4,714		
Age (y)	63 (57 – 71)	63 (57 – 70)	63 (57 – 71)	63 (57 – 70)	64 (58-72)	0.06	0.05
Sex						0.26	0.55
Women	5,892 (46%)	2,635 (45%)	1,845 (44%)	1,516 (42%)	1,892 (40%)		
Men	6,497 (54%)	2,993 (55%)	2,317 (56%)	2,086 (58%)	2,822 (60%)		
Ethnicity						0.67	0.15
White	11,353 (85%)	4,985 (88%)	3,552 (85%)	3,096 (86%)	3,911 (83%)		
Asian	338 (5%)	296 (5%)	351 (8%)	288 (8%)	452 (10%)		
Other	593 (6%)	296 (5%)	221 (6%)	185 (5%)	203 (5%)		
Unknown	105 (4%)	51 (1%)	38 (1%)	33 (1%)	74 (2%)		
*Body Mass Index (BMI)	31 (27 – 35)	30 (27 – 34)	30 (26 – 34)	29 (26 – 33)	29 (26-33)	0.10	0.06
*Current Smoker (yes)	4,836 (39%)	1,768 (31%)	1,124 (27%)	739 (21%)	792 (17%)	<0.001	<0.001
Comorbidities n(%)							
Atherosclerotic Heart Disease (yes)	2,346 (19%)	1,018 (18%)	809 (20%)	609 (17%)	813 (17%)	<0.001	<0.001
Angina (yes)	1,924 (16%)	802 (14%)	586 (14%)	425 (12%)	569 (14%)	0.13	0.66
Atrial Fibrillation (yes)	1,593 (13%)	704 (11%)	534 (13%)	442 (12%)	554 (12%)	0.37	<0.001
Chronic Obstructive Pulmonary Disease (yes)	1,704 (14%)	527 (9%)	345 (8%)	189 (5%)	196 (4%)	0.45	0.61
Chronic Kidney Disease (yes)	661 (5%)	279 (5%)	234 (6%)	180 (5%)	189 (4%)	<0.001	<0.001
Heart Failure (yes)	1,273 (10%)	551 (10%)	380 (9%)	283 (8%)	365 (8%)	<0.001	<0.001
Hyperkalaemia (yes)	645 (5%)	304 (5%)	180 (4%)	183 (5%)	205 (4%)	<0.001	<0.001
*Hypertension (yes)	7,372 (40%)	3,540 (63%)	2,583 (38%)	2,204 (39%)	2,935 (60%)	<0.001	<0.001
Myocardial Infarction (yes)	1,438 (11%)	622 (11%)	465 (11%)	376 (10%)	457 (10%)	<0.001	<0.001
Peripheral Artery Disease (yes)	556 (2%)	186 (1%)	140 (0.4%)	99 (0.3%)	93 (0.3%)	0.26	0.28
Stroke/TIA (yes)	1,160 (9%)	548 (11%)	378 (10%)	312 (9%)	357 (10%)	<0.001	<0.001
Lab Tests within 6 months of inclusion, n (%)							
Plasma Glucose (mmoL)	8.2(6.6 – 12)	8.2 (6.6 – 11.5)	8.1(6.6 – 11.5)	8.3 (6.7 – 11.7)	8.1 (6.6 – 11.1)	<0.001	<0.001
Haemoglobin A1C (mmoL)	53 (46– 67)	53 (45– 65)	53 (45– 65)	53 (45– 64)	52 (45 – 64)	<0.001	<0.001
Haemoglobin (g/L)						<0.001	<0.001
Men	144 (132 – 153)	144 (133 – 154)	144 (133 – 154)	145 (134 – 155)	145 (134 – 155)		
Woman	132 (122 – 142)	132 (121 – 141)	132 (121 – 142)	132 (122 – 141)	133 (122 – 141)		
Total Cholesterol (mmol)	4.2 (3.6-5.1)	4.2 (3.6-5)	4.2 (3.6-5)	4.3 (3.6-5.1)	4.5 (3.8-5.1)	<0.001	<0.001
Triglycerides (mmol)	1.8 (1.3 – 2.6)	1.7 (1.3 – 2.4)	1.7 (1.2 – 2.4)	1.6 (1.2 – 2.3)	1.6 (1.1 – 2.2)	<0.001	<0.001
Serum Albumin (g/L)	38 (35-40)	38 (35-40)	38 (35-40)	38 (35-40)	38 (36-40)	<0.001	<0.001

eGFR (mL/min/1.73m2)	55 (46 – 62)	54 (45 – 61)	54 (45 – 61)	54 (45 – 61)	53 (44 – 61)	<0.001	<0.001
Alanine Transaminase – ALT (U/L)	21 (15-31)	21 (15-31)	21 (15-31)	21 (15-31)	25 (17-35)	<0.001	<0.001
Aspartate Transaminase – AST (U/L)	20 (16-26)	20 (16-26)	20 (16-26)	20 (16-26)	21 (17-27)	<0.001	<0.001
Alkaline Phosphate (U/L)	88 (72-109)	86.5 (70-10	85 (70-105)	84 (69-102)	78 (64-94)	<0.001	0.10
Neutrophils (x10^9/L)	4.8 (3.7-6.1)	4.6 (3.7-5.9)	4.6 (3.6-5.7)	4.6 (3.5-5.6)	4.2 (3.3-5.4)	<0.001	0.10
Lymphocytes (x10^9/L)	2 (1.5-2.6)	2 (1.5-2.5)	2 (1.5-2.5)	1.9 (1.5-2.5)	1.9 (1.5 -2.4)	<0.001	<0.001
Bilirubin (µmol/L)	9 (7 - 13)	10 (7 - 13)	10 (7 - 14)	10 (8 - 14)	11 (8 – 14.5)	<0.001	0.10
Medications within 6 months of inclusion, n (%)							
Metformin (yes)	4,262 (34%)	1,963 (35%)	1,468 (35%)	1,154 (32%)	1,419 (30%)	0.12	0.24
Insulin (with Glucose-lowering Drug)	1,306 (11%)	552 (10%)	424 (10%)	330 (9%)	420 (9%)	<0.001	0.11
Sulphonylureas (yes)	4,316 (35%)	1,957 (35%)	1,446 (35%)	1,212 (34%)	1,524 (32%)	<0.001	0.46
SGTL2i (yes)	1,612 (13%)	727 (2%)	488 (2%)	412 (1%)	503 (11%)	<0.001	0.34
DPP-4 inhibitor (yes)	1,762 (6%)	796 (3%)	600 (2%)	482 (2%)	629 (2%)	0.40	<0.001
Statin (yes)	11,508 (93%)	5,217 (93%)	3,824 (93%)	3,267 (93%)	4,230 (90%)	0.65	0.51
Beta Blockers (yes)	3,683 (30%)	1,699 (30%)	1,262 (30%)	1,039 (29%)	1,321 (28%)	<0.001	<0.001
ACEi or ARBS (yes)	5,362 (26%)	2,533 (12%)	1,854 (9%)	1,480 (7%)	1,921 (9%)	0.94	0.34
MRAs (yes)	655 (5%)	314 (6%)	208 (5%)	163 (5%)	191 (4%)	<0.001	0.68
Calcium Channel Blockers (yes)	1,021 (8%)	464 (8%)	359 (9%)	357 (10%)	310 (7%)	0.44	0.51
Antiplatelets (yes)	4,316 (35%)	1,957 (34%)	1,446 (35%)	1,212 (34%)	1,524 (32%)	<0.001	0.46
Anticoagulants (yes)	1,997 (16%)	908 (16%)	662 (16%)	582 (15%)	787 (17%)	0.44	<0.001
Thiazides (yes)	4,156 (34%)	1,990 (35%)	1,421 (33%)	1,192 (33%)	1,606 (34%)	0.17	0.29
Loop Diuretic (yes)	2,959 (24%)	1,243 (22%)	929 (22%)	804 (22%)	748 (16%)	<0.001	<0.001
***Primary Care utilises patient READ CODES**							

C2: SIMD quintiles for the Secondary cohort showing all prognostic factors predicting all-cause mortality.

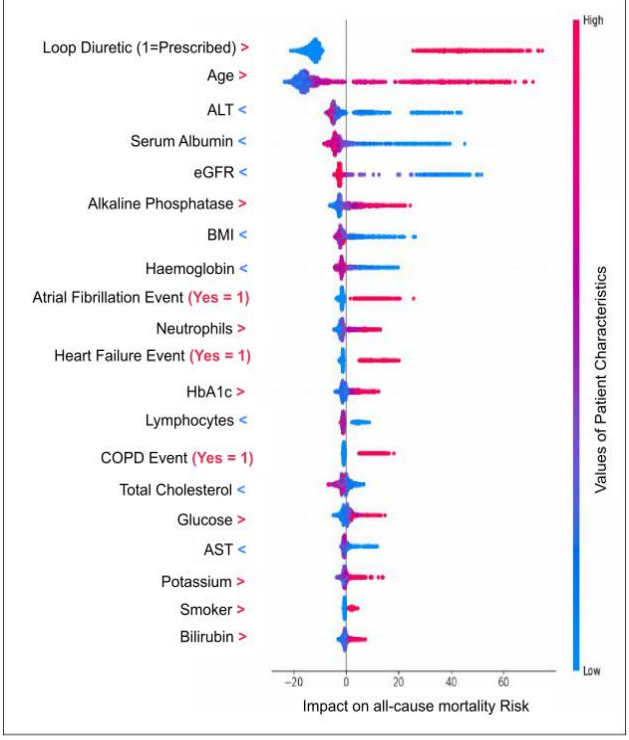
These graphs show a bar in the right axis: high is red and low is blue which represents the feature value. The absolute SHAP value shows us how much a single feature affected the prediction displayed on the x-axis. It takes the mean average value for each feature. Here, all the values on the left represent the observations that shift the predicted value in the negative direction while the points on the right contribute to shifting the prediction in a positive direction. All the features are on the left y-axis. For example, increased AGE on the x-axis has a high impact.





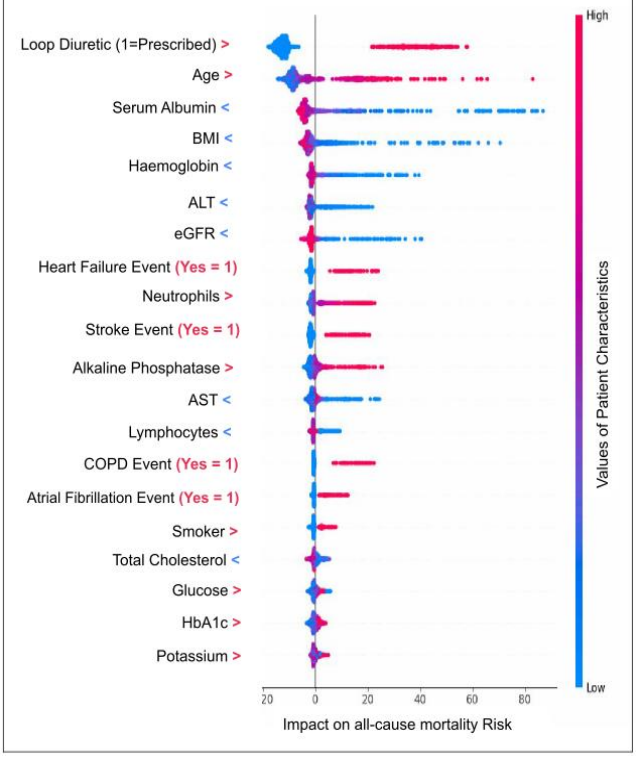
Quintile 3 (N=4,162)

C-index Score: 0.85 **Time Brier Score:** 0.07



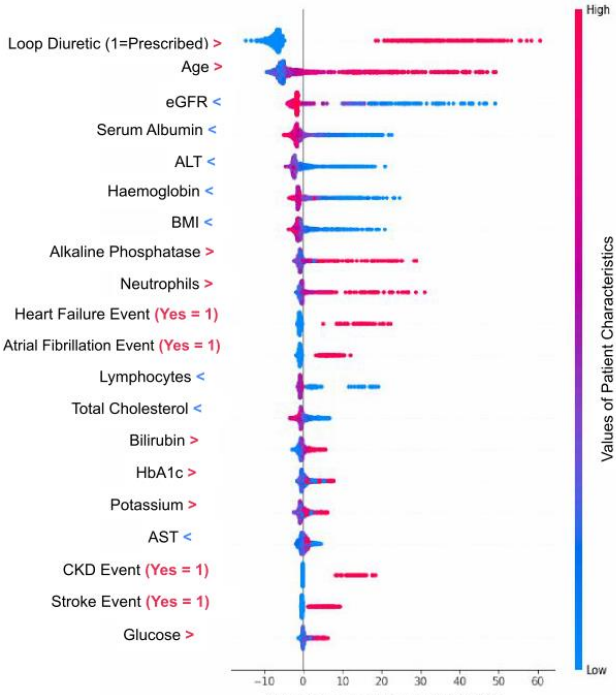
Quintile 4 (N=3,602)

C-index Score: 0.84 **Time Brier Score:** 0.07



Quintile 5 (N=4,714)

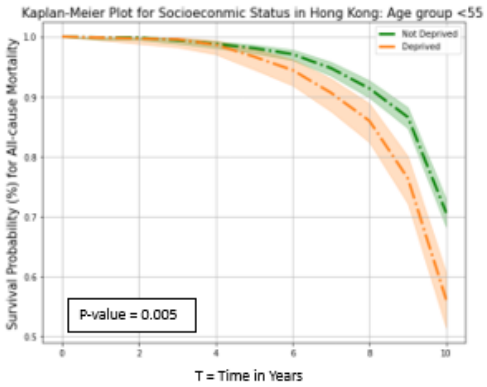
C-index Score: 0.80 **Time Brier Score:** 0.06



C3: Deprivation Status in Hong Kong

There were no events in the first four years when analysis by a measure of social deprivation was included suggesting a form of immortal time bias due to a delay in assigning SES. Only 64,380 patients (median age: 74 (67 – 79) years) were assigned a CSSA status, of whom 35,988 (55%) died within 10 years. Patients who were eventually classified as "Deprived" or “Not Deprived” had to survive until this point.

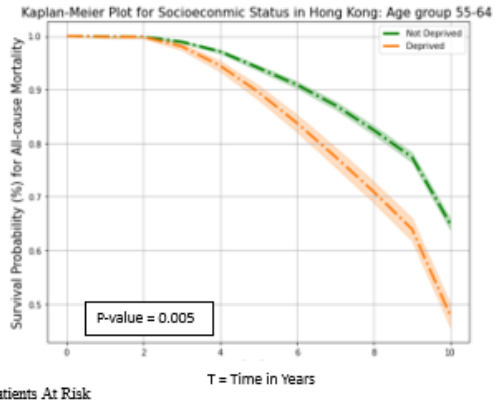
1. SES & Age Group < 55 (log-rank test confirms significance)



Patients At Risk

Not Deprived	1,507	1,503	1,485	1,459	1,372	0
Deprived	448	448	442	423	385	0

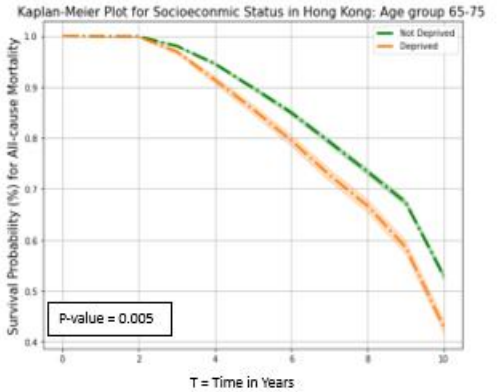
2. SES & Age Group 55-64 (log-rank test confirms significance)



Patients At Risk

Not Deprived	7,980	7,959	7,732	7,233	6,555	0
Deprived	2,024	2,019	1,909	1,691	1,431	0

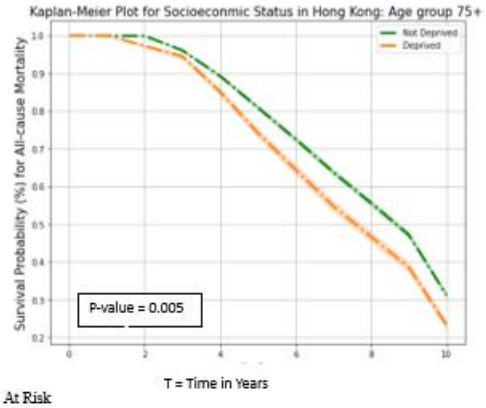
3. SES & Age Group 65 – 75 (log-rank test confirms significance)



Patients At Risk

Not Deprived	20,152	20,122	19,031	17,093	14,753	0
Deprived	5,378	5,368	4,911	4,273	3,582	0

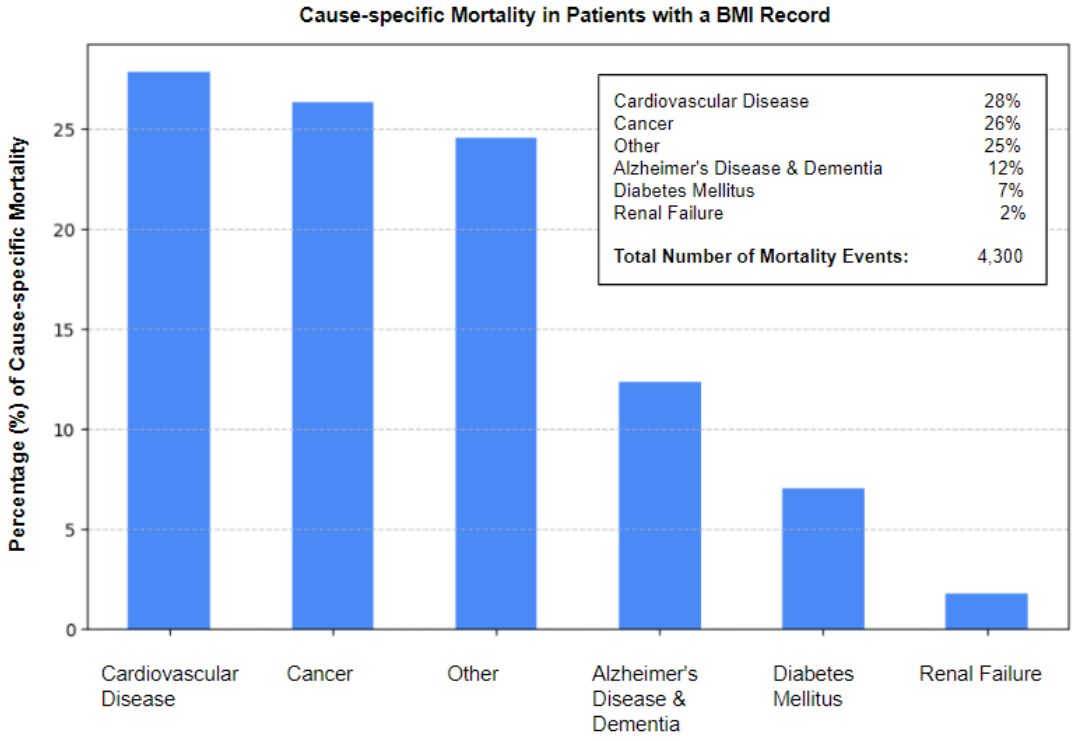
4. SES & Age Group 75+ (log-rank test confirms significance)

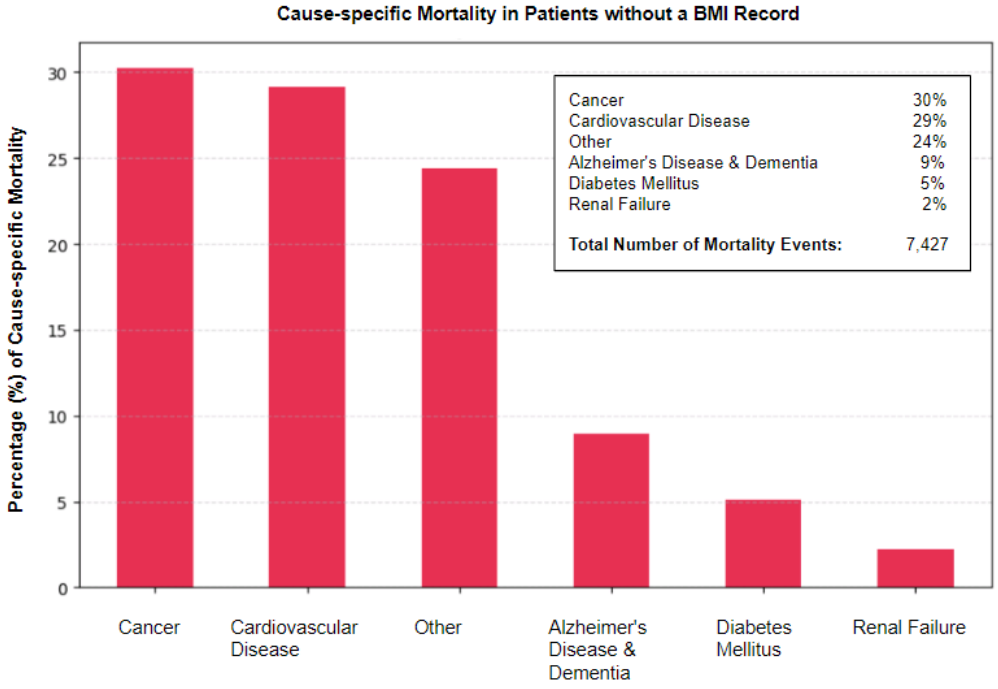


Patients At Risk

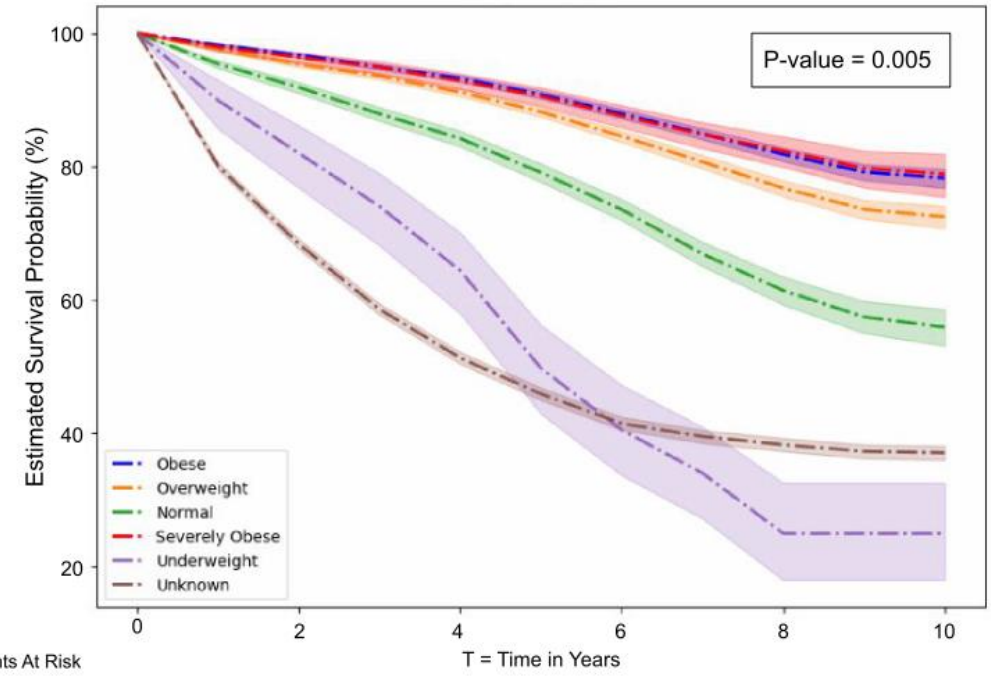
Not Deprived	20,934	20,910	18,681	15,179	11,637	0
Deprived	5,967	5,966	5,076	3,846	2,786	0

C4: Cause-specific mortality in Patients with and without BMI record





C5: Kaplan Meier Survival Estimate for All-cause mortality stratified by BMI Categories



Patients At Risk

Obese	12,926	9,579	6,798	4,114	1,504	0
Overweight	10,449	7,703	5,505	3,361	1,216	0
Normal	4,408	3,216	2,286	1,316	465	0
Severely Obese	2,444	1,801	1,256	784	288	0
Underweight	268	195	123	56	14	0
Unknown	15,536	7,736	4,541	2,405	993	0

Appendix D Chapter 7

D1: Laboratory Test Measurements in Glasgow and Hong Kong

Laboratory Test Measurements Frequency for T2D Patients		
Laboratory Test	**Glasgow (West of Scotland)**	**Hong Kong**
Haemoglobin (g/L)	Haemoglobin (g/L) Measured annually or semi-annually depending on diabetes control	Measured regularly (every 6 months), more frequently if anaemia is suspected
Total Cholesterol (mmol/L)	Measured every 6-12 months, part of cardiovascular risk assessment	Measured every 6-12 months, as part of routine lipid profile
HDL-C (mmol/L)	Measured annually or every 6 months	Assessed every 6 months as part of lipid profile
LDL-C (mmol/L)	Measured every 6-12 months	Measured every 6 months or more frequently if cholesterol-lowering medication is prescribed
VLDL (mmol/L)	Less commonly measured directly, calculated from triglycerides	Rarely measured directly, part of lipid assessment
Triglycerides (mmol/L)	Measured annually or every 6 months	Measured every 6 months
Serum Albumin (g/L)	Measured if kidney disease or malnutrition suspected	Part of routine tests, measured annually or bi-annually
Serum Creatinine (µmol/L)	Measured every 3-6 months, especially if kidney disease risk is present	Measured regularly, at least every 6 months
eGFR (mL/min/1.73 m²)	Measured every 3-6 months	Measured every 6 months
Urine Albumin-to-Creatinine Ratio (mg/g)	Annual or bi-annual test for diabetic nephropathy	Measured annually or every 6 months for kidney health monitoring
Potassium (mmol/L)	Measured along with kidney function tests	Monitored every 6 months, more frequent if on certain medications like diuretics
Lymphocytes (×10^9/L)	Part of full blood count if infection or immune issues suspected	Measured if infections or immune issues are of concern
Neutrophils (×10^9/L)	Part of routine blood work in diabetic care	Regularly assessed as part of full blood count
AST (Aspartate Aminotransferase) (U/L)	Measured annually or bi-annually	Liver function test, measured every 6 months
ALT (Alanine Aminotransferase) (U/L)	Measured annually or bi-annually	Liver function test, measured every 6 months
Alkaline Phosphatase (U/L)	Measured if liver or bone health concerns arise	Monitored along with liver function tests every 6 months
Bilirubin (µmol/L)	Part of liver function panel, measured annually	Liver function, monitored as part of bi-annual check-ups
Information derived from the following guidelines: NICE (National Institute for Health and Care Excellence), SIGN (Scottish Intercollegiate Guidelines Network), HKDF (Hong Kong Diabetes Federation) and ADA (American Diabetes Association, International Standards)		

D2: Sex Differences T2DM with and without BMI in Glasgow and Hong Kong

Baseline Characteristics of Men with T2DM with and without BMI for Glasgow and Hong Kong population

Men	Patients with T2DM		Patients with T2DM & BMI	
Population (N=)	Glasgow N=24,664	Hong Kong N=132,040	Glasgow N=16,715	Hong Kong N=10,547
Demographics (%) or Median (25/75)				
Age	63 (56 – 70)	64 (55 – 73)	62 (56 – 69)	71 (65 – 77)
*Body Mass Index (BMI)	N/A	N/A	30 (27 – 33)	24 (22 – 27)
*Current Smoker	5,258 (21%)	N/A	5,172 (31%)	N/A
Death within 5 years of T2DM Diagnosis	4,916 (20%)	19,668 (15%)	1,381 (8%)	554 (5%)
Comorbidities n (%)				
Atherosclerotic Heart Disease (yes)	4,729 (19%)	14,834 (11%)	3,751 (22%)	1,115 (11%)
Angina (yes)	3,332 (14%)	N/A	2,594 (16%)	N/A
Atrial Fibrillation (yes)	3,301 (13%)	3,429 (3%)	2,145 (13%)	238 (2%)
Chronic Obstructive Pulmonary Disease (yes)	1,990 (8%)	591 (0.44%)	1,317 (8%)	70 (1%)
Chronic Kidney Disease (yes)	1,231 (5%)	1,854 (1%)	738 (4%)	0
Heart Failure (yes)	2,684 (11%)	5,197 (4%)	1,704 (10%)	291 (3%)
Hyperkalaemia (yes)	1,060 (4%)	N/A	709 (4%)	N/A
*Hypertension (yes)	9,887 (40%)	28,799 (23%)	9,739 (58%)	2,296 (22%)
Myocardial Infarction (yes)	2,914 (12%)	N/A	2,199 (13%)	N/A
Peripheral Artery Disease (yes)	1,054 (4%)	204 (0.15%)	694 (4%)	0
Stroke/TIA (yes)	2,150 (9%)	6,230 (5%)	1,514 (9%)	484 (4%)
Lab Tests * 6 months prior to or upon a Diabetes Diagnosis				
Plasma Glucose (mmoL)	9 (6.8 – 10.4)	N/A	9 (7.1 – 10.7)	N/A
Haemoglobin A1C (mmoL)	55 (46– 62)	56 (51– 63)	55 (46– 64)	56 (47 – 63)
Haemoglobin (g/L)	138 (134 – 151)	131 (123– 139)	139 (134 – 152)	130 (123 – 137)
Total Cholesterol (mmol)	4 (3.6 – 4.8)	4.7 (4.2-5.1)	4 (3.5 - 4.8)	4.7 (4.2-5.1)
Triglycerides (mmol)	1.5 (1 – 2.3)	1.4 (1.1 – 1.9)	1.6 (1.1 – 2.4)	1.4 (1.3 – 2)
Serum Albumin (g/L)	38 (36 – 40)	40 (38-42)	38 (36- 40)	40 (38-41)
eGFR (mL/min/1.73m2)	51 (42 – 59)	33 (27 - 39)	51 (43-59)	32 (26 – 37)
Alanine Transaminase – ALT (U/L)	25 (17 - 33)	24 (18-32)	25 (18-34)	23 (18 – 30)
Aspartate Transaminase – AST (U/L)	22 (16 – 27)	25 (21-43)	22 (17-27)	61 (24 – 73)
Alkaline Phosphate	85 (69 – 101)	74 (65-86)	84 (68-100)	74 (63 – 85)
Neutrophils	5.1 (4.2 – 6.3)	5.3 (4.3-7)	5.1 (4 – 5.5)	5.4 (5 – 8)
Lymphocytes	2 (2 – 2.3)	2.3 (1.6-2.3)	2 (2.1-2.4)	2 (1.6 – 2.3)
Bilirubin (µmol/L)	11 (8 – 14)	9 (9 – 12.4)	11 (8 - 14)	11 (8 – 11)
Potassium (mmol)	4.3 (4 – 4.6)	4 (4 – 4.4)	4.3 (4 – 4.6)	4 (4.2 – 4.4)

Medications +/- 180 days of diabetes diagnosis				
Metformin (yes)	7,872 (32%)	86,534 (66%)	5,532 (33%)	8,312 (21%)
DPP4i (yes)	2,747 (11%)	182 (0.13%)	2,319 (14%)	4 (0.3%)
Insulin (taken with Glucose-lowering Drug)	1,478 (6%)	14,633 (11%)	893 (5%)	370 (4%)
Sulphonylureas (yes)	5,585 (23%)	84,414 (64%)	3,836 (23%)	7,367 (70%)
SGLT2i (yes)	2,333 (9%)	N/A	2,207 (13%)	N/A
Statins (yes)	12,746 (52%)	30,538 (23%)	8,126 (49%)	2,014 (19%)
Beta Blockers (yes)	5,299 (21%)	41,677 (32%)	3,603 (23%)	3,620 (34%)
ACEi or ARBS (yes)	11,196 (63%)	62,157 (47%)	7,270 (67%)	5,397 (51%)
Antiplatelets (yes)	5,585 (23%)	N/A	3,836 (23%)	N/A
Anticoagulants (yes)	2,329 (9%)	N/A	1,387 (8%)	N/A
Thiazides (yes)	5,466 (22%)	20,110 (15%)	3,681 (22%)	1,835 (17%)
Loop Diuretic (yes)	5,256 (21%)	28,014 (21%)	3,045 (18%)	3,350 (32%)

Baseline Characteristics of Woman with T2DM with and without BMI for Glasgow and Hong Kong population

Women	**Patients with T2DM**		**Patients with T2DM & BMI**	
Population (N=)	**Glasgow** N= 21,367	**Hong Kong** N=141,836	**Glasgow** N=13,780	**Hong Kong** N=13,426
Demographics (%) or Median (25/75)				
Age	65 (58 – 74)	68 (58 – 77)	64 (57 – 72)	73 (67 – 78)
***Body Mass Index (BMI)**	N/A	N/A	30.4 (26.4 – 35)	24.4 (22 – 27)
***Current Smoker (yes)**	4,158 (19%)	N/A	4,087 (30%)	N/A
Death within 5 years of T2D Diagnosis	4,346 (20%)	18,306 (13%)	1,200 (9%)	500 (4%)
Comorbidities n (%) at Diagnosis				
Atherosclerotic Heart Disease (yes)	2,377 (11%)	11,589 (8%)	1,844 (13%)	1,006 (7%)
Angina (yes)	2,289 (11%)	N/A	1,712 (12%)	N/A
Atrial Fibrillation (yes)	2,782 (13%)	4,343 (3%)	1,682 (12%)	280 (2%)
Chronic Obstructive Pulmonary Disease (yes)	2,405 (11%)	277 (0.16%)	1,653 (12%)	19 (0.14%)
Chronic Kidney Disease (yes)	1,318 (6%)	1,527 (1%)	805 (6%)	0
Heart Failure (yes)	1,991 (9%)	5,992 (4%)	1,148 (8%)	299 (3%)
Hyperkalaemia (yes)	1,276 (6%)	N/A	808 (6%)	N/A
***Hypertension (yes)**	9,112 (65%)	35,447 (25%)	8,895 (64%)	3,259 (24%)
Myocardial Infarction (yes)	1,631 (8%)	N/A	1,159 (8%)	N/A
Peripheral Artery Disease (yes)	596 (4%)	142 (0.1%)	380 (3%)	0
Stroke/TIA (yes)	1,860 (9%)	6,020 (4%)	1,241 (9%)	490 (4%)
Lab Tests within 6 months of inclusion, n (%)				
Plasma Glucose (mmoL)	8.5 (6.6 – 9.8)	N/A	9 (6.9 - 10.2)	N/A

Haemoglobin A1C (mmoL)	54 (46– 60)	56 (51– 63)	54 (46– 62)	56 (51– 63)
Haemoglobin (g/L)	134 (123 – 139)	122 (129 – 136)	134 (124 – 140)	130 (123 – 136)
Total Cholesterol (mmol)	4.3 (4.0-5.3)	4.8 (4.4-5.3)	4.3 (3.9-5.2)	4.8 (4.4-5.3)
Triglycerides (mmol)	1.5 (1 – 2.2)	1.5 (1.2 – 2.1)	1.6 (1.1 – 2.3)	1.6 (1.2 – 2)
Serum Albumin (g/L)	37 (35-39)	40 (38-41)	37 (35-39)	40 (38 – 42)
eGFR (mL/min/1.73m2)	57 (48 - 64)	37 (31 - 44)	58 (49-65)	35 (30 – 41)
Alanine Transaminase – ALT (U/L)	20 (14-27)	22 (17 – 30)	20 (15-27)	21 (16 – 28)
Aspartate Transaminase – AST (U/L)	20 (16-25)	25 (21 – 36)	20 (16-25)	33 (24 – 73)
Alkaline Phosphate	93 (75-110)	75 (66 – 87)	93 (75-111)	75 (66 – 86)
Neutrophils	5 (3.8-5.1)	5.3 (4.4 – 7.4)	5 (3.1 – 5.5)	5.4 (4.5 – 7.7)
Lymphocytes	2 (2-2.3)	2 (1.7 – 2.5)	2 (1.6-2.4)	2 (2 – 3)
Bilirubin (µmol/L)	9 (7 - 11)	9 (8.6 – 12.4)	9 (7 - 11)	11 (8 – 11)
Potassium (mmol)	4.3 (4 – 4.6)	4.2 (4 – 4.4)	4.3 (4 – 4.6)	4.2 (4 – 4.6)
Medications within 6 months of inclusion, n (%)				
Metformin	6,673 (31%)	42,489 (30%)	4,734 (34%)	11,126 (83%)
DPP4i	2,286 (11%)	143 (0.1%)	1,950 (14%)	2 (0.01%)
Insulin (taken with Glucose-lowering Drug)	1,323 (6%)	15, 064 (11%)	815 (6%)	421 (3.1%)
Sulphonylureas	4,619 (22%)	52,725 (37%)	3,210 (23%)	9,391 (70%)
SGTL2i	1,644 (8%)	N/A	1,535 (11%)	N/A
Statins	10,311 (52%)	30,863 (22%)	6,855 (50%)	2,440 (18%)
Beta Blockers	4,944 (23%)	50,632 (37%)	3,113 (23%)	5,255 (39%)
ACEi or ARBS	9,353 (57%)	59,629 (42%)	5,880 (61%)	5,951 (44%)
Antiplatelets	4,619 (22%)	N/A	3,210 (23%)	N/A
Anticoagulants	1,980 (9%)	N/A	1,124 (8%)	N/A
Thiazides	6,555 (31%)	27,047 (19%)	4,288 (31%)	3,104 (23%)
Loop Diuretic	6,147 (29%)	32,138 (23%)	3,638 (26%)	4,244 (32%)

Bibliography

Ahmad, A. et al. (2024) Precision prognostics for cardiovascular disease in Type 2 diabetes: a systematic review and meta-analysis. *Communications Medicine 2024 4:1*. [Online] 4 (1), 1–28. [online]. Available from: https://www.nature.com/articles/s43856-023-00429-z (Accessed 5 November 2024).

Alabdallah, A. et al. (2022) *The Concordance Index decomposition A measure for a deeper understanding of survival prediction models*.

Alvarez-Madrazo, S. et al. (2016) Data Resource Profile: The Scottish National Prescribing Information System (PIS). *International Journal of Epidemiology*. [Online] 45 (3), 714. [online]. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC5005947/ (Accessed 25 October 2024).

Amatruda, J. G. et al. (2022) Renin-Angiotensin-Aldosterone System Activation and Diuretic Response in Ambulatory Patients With Heart Failure. *Kidney Medicine*. [Online] 4 (6), 100465. [online]. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2590059522000784.

Angelopoulos, A. N. & Bates, S. (2021) *A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification*. [online]. Available from: https://arxiv.org/abs/2107.07511v6 (Accessed 30 December 2024).

Anker, S. D. et al. (2023) Patient phenotype profiling in heart failure with preserved ejection fraction to guide therapeutic decision making. A scientific statement of the Heart Failure Association, the European Heart Rhythm Association of the European Society of Cardiology, and the European Society of Hypertension. *European Journal of Heart Failure*. [Online] 25 (7), 936–955. [online]. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/ejhf.2894 (Accessed 8 July 2024).

Anon (2024) *Cerebrovascular Disease - AANS* [online]. Available from: https://www.aans.org/patients/conditions-treatments/cerebrovascular-disease/ (Accessed 4 July 2024).

Anon (2024) *Hong Kong Diabetes Association* [online]. Available from: https://www-diabetes--hk-org.translate.goog/home?_x_tr_sl=zh-TW&_x_tr_tl=en&_x_tr_hl=en&_x_tr_pto=sc&_x_tr_sch=http (Accessed 8 January 2025).

Anon (2024) *PHARMACY & POISONS BOARD OF HONG KONG - Pharmacy and Poisons Ordinance* [online]. Available from: https://www.ppbhk.org.hk/eng/ordinance/138.html (Accessed 28 October 2024).

van Apeldoorn, J. A. N. et al. (2024a) Adding ethnicity to cardiovascular risk prediction: External validation and model updating of SCORE2 using data from the HELIUS population cohort. *International journal of cardiology*. [Online] 417. [online]. Available from: https://pubmed.ncbi.nlm.nih.gov/39244095/ (Accessed 9 December 2024).

van Apeldoorn, J. A. N. et al. (2024b) Adding ethnicity to cardiovascular risk prediction: External validation and model updating of SCORE2 using data from the HELIUS population cohort. *International Journal of Cardiology*. [Online] 417132525.

Asri, H. et al. (2020) Big data and reality mining in healthcare: Promise and potential. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. [Online] 12119 LNCS122–129. [online]. Available from: https://link.springer.com/chapter/10.1007/978-3-030-51935-3_13 (Accessed 21 October 2024).

Austin, P. C. et al. (2021) Missing Data in Clinical Research: A Tutorial on Multiple Imputation. *Canadian Journal of Cardiology*. [Online] 37 (9), 1322–1331.

Baak, M. et al. (2019) *A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics*.

Bakker, J. (2019) *Heart attack gender gap is costing women's lives - BHF* [online]. Available from: https://www.bhf.org.uk/what-we-do/news-from-the-bhf/news-archive/2019/september/heart-attack-gender-gap-is-costing-womens-lives (Accessed 12 November 2024).

Balbirsingh, V. et al. (2022) *Cardiovascular disease in chronic obstructive pulmonary disease: a narrative review State of the art review*. [Online] [online]. Available from: http://thorax.bmj.com/ (Accessed 3 July 2023).

Banerjee, D. et al. (2013) Insulin resistance and risk of incident heart failure cardiovascular health study. *Circulation: Heart Failure*. [Online] 6 (3), 364–370. [online]. Available from: https://www.ahajournals.org/doi/10.1161/CIRCHEARTFAILURE.112.000022 (Accessed 3 October 2024).

Barnett, G. O. et al. (1979) COSTAR—A Computer-Based Medical Information System for Ambulatory Care. *Proceedings of the IEEE*. [Online] 67 (9), 1226–1237.

Barraclough, H. et al. (2011) Biostatistics primer: what a clinician ought to know: hazard ratios. *J Thorac Oncol*. [Online] 6 (6), 978–982.

Basu, S. et al. (2017) Development and validation of Risk Equations for Complications Of type 2 Diabetes (RECODe) using individual participant data from randomised trials. *The Lancet Diabetes & Endocrinology*. [Online] 5 (10), 788–798.

Bewick, V. et al. (2004) Statistics review 12: Survival analysis. *Critical Care*. [Online] 8 (5), 389. [online]. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC1065034/ (Accessed 7 January 2025).

Boyd, A. D. et al. (2023) Potential bias and lack of generalizability in electronic health record data: reflections on health equity from the National Institutes of Health Pragmatic Trials Collaboratory. *Journal of the American Medical Informatics Association : JAMIA*. [Online] 30 (9), 1561–1566. [online]. Available from: https://pubmed.ncbi.nlm.nih.gov/37364017/ (Accessed 10 November 2024).

Brosula, R. et al. (2024) Pathophysiological Features in Electronic Medical Records Sustain Model Performance under Temporal Dataset Shift. *AMIA Summits on Translational Science Proceedings*. 202495. [online]. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC11141811/ (Accessed 15 November 2024).

Bueno Junior, C. R. et al. (2023) Rapid kidney function decline and increased risk of heart failure in patients with type 2 diabetes: findings from the ACCORD cohort: Rapid kidney function decline and heart failure in T2D. *Cardiovascular Diabetology*. [Online] 22 (1), 1–12. [online]. Available from: https://cardiab.biomedcentral.com/articles/10.1186/s12933-023-01869-6 (Accessed 22 September 2024).

Di Cesare, M. et al. (2024) The Heart of the World. *Global Heart*. [Online] 19 (1), 11. [online]. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC10809869/ (Accessed 7 January 2025).

Chalmers, J. (2005) ADVANCE—Action in Diabetes and Vascular Disease: patient recruitment and characteristics of the study population at baseline. *Diabetic Medicine*. [Online] 22 (7), 882–888. [online]. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1464-5491.2005.01596.x (Accessed 5 November 2024).

Chandola, T. & Conibere, R. (2015) Social Exclusion, Social Deprivation and Health. *International Encyclopedia of the Social & Behavioral Sciences: Second Edition*. [Online] 285–290.

Cheng, Y. et al. (2024) Heart Failure Among Asian American Subpopulations. *JAMA Network Open*. [Online] 7 (9), e2435672–e2435672. [online]. Available from: https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2824102 (Accessed 2 October 2024).

Clark, A. L. (2022) What is heart failure? *Oxford Textbook of Heart Failure*. [Online] 3–8.

Clark, T. G. et al. (2003) Survival Analysis Part I: Basic concepts and first analyses. *British Journal of Cancer*. [Online] 89 (2), 232. [online]. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC2394262/ (Accessed 30 December 2024).

Cleland, J. G. et al. (2014) Predictors of postdischarge outcomes from information acquired shortly after admission for acute heart failure: a report from the Placebo-Controlled Randomized Study of the Selective A1 Adenosine Receptor Antagonist Rolofylline for Patients Hospitalized With Acute Decompensated Heart Failure and Volume Overload to Assess Treatment Effect on Congestion and Renal Function (PROTECT) Study. *Circulation. Heart failure*. [Online] 7 (1), 76–87. [online]. Available from: https://pubmed.ncbi.nlm.nih.gov/24281134/ (Accessed 2 January 2025).

Cleland, J. G. F. et al. (2021) The struggle towards a Universal Definition of Heart Failure—how to proceed? *European Heart Journal*. [Online] 42 (24), 2331–2343. [online]. Available from: https://dx.doi.org/10.1093/eurheartj/ehab082 (Accessed 3 January 2025).

collaboration, S. working group and E. C. risk et al. (2021) SCORE2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe. *European Heart Journal*. [Online] 42 (25), 2439–2454. [online]. Available from: https://academic.oup.com/eurheartj/article/42/25/2439/6297709 (Accessed 23 October 2021).

Committee, A. D. A. P. P. (2022) 10. Cardiovascular Disease and Risk Management: Standards of Medical Care in Diabetes—2022. *Diabetes Care*. [Online] 45 (Supplement_1), S144–S174. [online]. Available from: https://dx.doi.org/10.2337/dc22-S010 (Accessed 10 November 2024).

Copetti, M. et al. (2021) All-cause mortality prediction models in type 2 diabetes: applicability in the early stage of disease. *Acta Diabetologica*. [Online] 58 (10), 1425–1428. [online]. Available from: https://link.springer.com/article/10.1007/s00592-021-01746-2 (Accessed 21 October 2022).

Cosmi, F. et al. (2018) Treatment with insulin is associated with worse outcome in patients with chronic heart failure and diabetes. *European Journal of Heart Failure*. [Online] 20 (5), 888–895. [online]. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/ejhf.1146 (Accessed 3 October 2024).

Costanzo, P. et al. (2015) The obesity paradox in type 2 diabetes mellitus: relationship of body mass index to prognosis: a cohort study. *Annals of internal medicine*. [Online] 162 (9), 610–618. [online]. Available from: https://pubmed.ncbi.nlm.nih.gov/25938991/ (Accessed 14 April 2022).

Cullington, D. et al. (2014) Is heart rate important for patients with heart failure in atrial fibrillation? *JACC. Heart failure*. [Online] 2 (3), 213–220. [online]. Available from: https://pubmed.ncbi.nlm.nih.gov/24952686/ (Accessed 3 January 2025).

Cuthbert, J. J. et al. (2024) Outcomes in patients treated with loop diuretics without a diagnosis of heart failure: a retrospective cohort study. *Heart*. [Online] 110 (12), 854–862. [online]. Available from: https://heart.bmj.com/content/110/12/854 (Accessed 2 October 2024).

Dang, V. N. et al. (2024) Fairness and bias correction in machine learning for depression prediction across four study populations. *Scientific Reports 2024 14:1*. [Online] 14 (1), 1–12. [online]. Available from: https://www.nature.com/articles/s41598-024-58427-7 (Accessed 10 December 2024).

DD, S. et al. (1997) Dietary assessment in Whitehall II: the influence of reporting bias on apparent socioeconomic variation in nutrient intakes. *European journal of clinical nutrition*. [Online] 51 (12), 815–825. [online]. Available from: https://pubmed.ncbi.nlm.nih.gov/9426356/ (Accessed 20 August 2021).

Deb, S. et al. (2016) A Review of Propensity-Score Methods and Their Use in Cardiovascular Research. *Canadian Journal of Cardiology*. [Online] 32 (2), 259–265.

Deepali Nagar, S. et al. (2021) *Socioeconomic deprivation and genetic ancestry interact to modify type 2 diabetes ethnic disparities in the United Kingdom*. [Online] [online]. Available from: https://doi.org/10.1016/j.eclinm.2021.100960 (Accessed 15 May 2023).

Department of Health (2020) *Department of Health - Drug Office* [online]. Available from: https://www.dh.gov.hk/english/main/main_ps/main_ps.html (Accessed 7 January 2025).

Di, J. et al. (2022) Considerations to address missing data when deriving clinical trial endpoints from digital health technologies. *Contemporary Clinical Trials*. [Online] 113106661.

Diao, J. A. et al. (2022) National Projections for Clinical Implications of Race-Free Creatinine-Based GFR Estimating Equations. *Journal of the American Society of Nephrology : JASN*. [Online] 34 (2), 309. [online]. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC10103103/ (Accessed 26 October 2024).

Disease Branch, N. et al. (2021) *Non-Communicable Diseases Watch November 2021 - Screening and Prevention of Diabetes*. [online]. Available from: http://www.chp.gov.hk.

Dong, W. et al. (2024) Development and validation of 10-year risk prediction models of cardiovascular disease in Chinese type 2 diabetes mellitus patients in primary care using interpretable machine learning-based methods. *Diabetes, Obesity and Metabolism*. [Online] 26 (9), 3969–3987.

Dzau, V. J. et al. (2006) The cardiovascular disease continuum validated: Clinical evidence of improved patient outcomes: Part I: Pathophysiology and clinical trial evidence (risk factors through stable coronary artery disease). *Circulation*. [Online] 114 (25), 2850–

2870. [online]. Available from: https://www.ahajournals.org/doi/abs/10.1161/circulationaha.106.655688 (Accessed 4 June 2022).

Einarson, T. R. et al. (2018) Prevalence of cardiovascular disease in type 2 diabetes: A systematic literature review of scientific evidence from across the world in 2007-2017. Cardiovascular Diabetology 17 (1).

Eline Bretscher, C. et al. (2022) *Admission serum albumin concentrations and response to nutritional therapy in hospitalised patients at malnutrition risk: Secondary analysis of a randomised clinical trial*. [Online] [online]. Available from: https://doi.org/10.1016/j. (Accessed 9 August 2023).

Ellsworth, M. A. et al. (2016) Early Computerization of Patient Care at Mayo Clinic. *Mayo Clinic Proceedings*. [Online] 91 (7), e93–e101. [online]. Available from: http://www.mayoclinicproceedings.org/article/S0025619616301057/fulltext (Accessed 18 October 2024).

Emmens, J. E. et al. (2022) Worsening renal function in acute heart failure in the context of diuretic response. *European Journal of Heart Failure*. [Online] 24 (2), 365–374.

Fan, Y. Y. K. et al. (2022) Trends in contemporary advanced heart failure management: an in-depth review over 30 years of heart transplant service in Hong Kong. *Korean Journal of Transplantation*. [Online] 36 (4), 267. [online]. Available from: /pmc/articles/PMC9832593/ (Accessed 25 September 2024).

Felker, G. M. et al. (2020) Diuretic Therapy for Patients With Heart Failure: JACC State-of-the-Art Review. *Journal of the American College of Cardiology*. [Online] 75 (10), 1178–1195.

Fitzpatrick, J. K. et al. (2022) Loop and thiazide diuretic use and risk of chronic kidney disease progression: a multicentre observational cohort study. *BMJ open*. [Online] 12 (1), . [online]. Available from: https://pubmed.ncbi.nlm.nih.gov/35105612/ (Accessed 2 October 2024).

Fmedsci, S. et al. (2023) School of Cardiovascular and Metabolic Health Treating chronic diseases without tackling excess adiposity promotes multimorbidity. *www.thelancet.com/diabetes-endocrinology*. [Online] 1158–62. [online]. Available from: www.thelancet.com/diabetes-endocrinology (Accessed 10 November 2023).

Fonarow, G. C. (2007) The global burden of atherosclerotic vascular disease: Commentary. *Nature Clinical Practice Cardiovascular Medicine*. [Online] 4 (10), 530–531.

Foster, H. M. E. et al. (2018) The effect of socioeconomic deprivation on the association between an extended measurement of unhealthy lifestyle factors and health outcomes: a

prospective analysis of the UK Biobank cohort. *The Lancet Public Health*. [Online] 3 (12), e576–e585. [online]. Available from: http://www.thelancet.com/article/S2468266718302007/fulltext (Accessed 7 July 2021).

Foundation, B. H. (2020) *Risk factors for heart and circulatory diseases - BHF*. [online]. Available from: https://www.bhf.org.uk/informationsupport/risk-factors.

Friday, J. M. et al. (2024) Loop diuretic utilisation with or without heart failure: impact on prognosis. *European Heart Journal*. [Online]

Friedman, J. H. (2001) Greedy function approximation: A gradient boosting machine. *https://doi.org/10.1214/aos/1013203451*. [Online] 29 (5), 1189–1232. [online]. Available from: https://projecteuclid.org/journals/annals-of-statistics/volume-29/issue-5/Greedy-function-approximation-A-gradient-boosting-machine/10.1214/aos/1013203451.full (Accessed 26 March 2025).

Fu, L. et al. (2023) Prevalence and incidence of heart failure among community in China during a three-year follow-up. *Journal of Geriatric Cardiology*. [Online] 20 (4), 284–292.

Gao, L. et al. (2021) Linking cohort-based data with electronic health records: a proof-of-concept methodological study in Hong Kong. *BMJ Open*. [Online] 11 (6), e045868. [online]. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC8220454/ (Accessed 21 October 2024).

George, J. et al. (2017) *Ethnicity and the first diagnosis of a wide range of cardiovascular diseases: Associations in a linked electronic health record cohort of 1 million patients*. [Online] [online]. Available from: https://doi.org/10.1371/journal.pone.0178945.

Gonzalez, L. L. et al. (2018) Type 2 diabetes - An autoinflammatory disease driven by metabolic stress. *Biochimica et biophysica acta. Molecular basis of disease*. [Online] 1864 (11), 3805–3823. [online]. Available from: https://pubmed.ncbi.nlm.nih.gov/30251697/ (Accessed 5 November 2024).

Griffiths, K. et al. (2023) Interpreting an estimated glomerular filtration rate (eGFR) in people of black ethnicities in the UK. *BMJ*. [Online] 380. [online]. Available from: https://www.bmj.com/content/380/bmj-2022-073353 (Accessed 26 October 2024).

Groenewegen, A. et al. (2020) Epidemiology of heart failure. *European Journal of Heart Failure*. [Online] 22 (8), 1342–1356. [online]. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/ejhf.1858 (Accessed 10 July 2022).

Guan, H. et al. (2024) Advances in secondary prevention mechanisms of macrovascular complications in type 2 diabetes mellitus patients: a comprehensive review. *European Journal of Medical Research 2024 29:1*. [Online] 29 (1), 1–34. [online]. Available from:

https://eurjmedres.biomedcentral.com/articles/10.1186/s40001-024-01739-1 (Accessed 5 November 2024).

Guo, L. et al. (2023) Diuretic resistance in patients with kidney disease: Challenges and opportunities. *Biomedicine & Pharmacotherapy*. [Online] 157114058.

Guo, L. et al. (2016) Prevalence and Risk Factors of Heart Failure with Preserved Ejection Fraction: A Population-Based Study in Northeast China. *International Journal of Environmental Research and Public Health 2016, Vol. 13, Page 770*. [Online] 13 (8), 770. [online]. Available from: https://www.mdpi.com/1660-4601/13/8/770/htm (Accessed 14 October 2024).

Guthrie, B. et al. (2012) Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. *The Lancet*. [Online] 38037–43. [online]. Available from: www.thelancet.com (Accessed 6 July 2024).

Hackshaw, A. et al. (2018) Low cigarette consumption and risk of coronary heart disease and stroke: meta-analysis of 141 cohort studies in 55 study reports. *BMJ*. [Online] 360. [online]. Available from: https://www.bmj.com/content/360/bmj.j5855 (Accessed 12 November 2024).

Hartman, N. et al. (2023) Pitfalls of the concordance index for survival outcomes. *Statistics in Medicine*. [Online] 42 (13), 2179–2190. [online]. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/sim.9717 (Accessed 15 November 2024).

Hassaine, A. et al. (2020) Untangling the complexity of multimorbidity with machine learning. *Mechanisms of Ageing and Development*. [Online] 190.

He, J. et al. (2023) High absolute neutrophil count with type 2 diabetes is associated with adverse outcome in patients with coronary artery disease: A large-scale cohort study. *Frontiers in Endocrinology*. [Online] 141129633.

Health Intelligence Team, B. (2024a) *BHF UK CVD Factsheet*.

Health Intelligence Team, B. (2024b) *BHF UK CVD Factsheet*.

Health Scotland, N. (2015) *Health inequalities: What are they? How do we reduce them? Health inequalities: What are they? How do we reduce them?* [online]. Available from: www.healthscotland.com (Accessed 19 August 2021).

Hernán, M. A. & Robins, J. M. (2016) Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *American Journal of Epidemiology*. [Online] 183 (8), 758–764. [online]. Available from: https://dx.doi.org/10.1093/aje/kwv254 (Accessed 10 November 2024).

Heymans, M. W. & Twisk, J. W. R. (2022) Handling missing data in clinical research. *Journal of Clinical Epidemiology*. [Online] 151185–188.

Hill-Briggs, F. et al. (2021) *World Health Organization (WHO) Commission on Social Determinants of Health (28), Healthy People 2020 (29,30), the Diabetes Care*. [Online] 44. [online]. Available from: https://doi.org/10.2337/dci20-0053 (Accessed 3 July 2023).

Hippisley-Cox, J. & Coupland, C. (2015) *Development and validation of risk prediction equations to estimate future risk of heart failure in patients with diabetes: a prospective cohort study*. [Online] [online]. Available from: http://www.qresearch.org.

Ho, F. K. et al. (2022) Association of gamma-glutamyltransferase levels with total mortality, liver-related and cardiovascular outcomes: A prospective cohort study in the UK Biobank. *eClinicalMedicine*. [Online] 48101435. [online]. Available from: https://doi.org/10.1016/j. (Accessed 9 August 2023).

Ho Wong, M. et al. (123AD) Prevalence and factors associated with diabetes-related distress in type 2 diabetes patients: a study in Hong Kong primary care setting. *Scientific Reports* |. [Online] 1410688.

Holzinger, A. et al. (2019) Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. [Online] 9 (4), .

Hong Kong Authority (2015) *Drug Formulary Management* [online]. Available from: https://www.ha.org.hk/hadf/en-us/Drug-Formulary-Management.html (Accessed 28 October 2024).

Ishwaran, H., Kogalur, Udaya B., et al. (2008) Random survival forests. *https://doi.org/10.1214/08-AOAS169*. [Online] 2 (3), 841–860. [online]. Available from: https://projecteuclid.org/journals/annals-of-applied-statistics/volume-2/issue-3/Random-survival-forests/10.1214/08-AOAS169.full (Accessed 8 August 2022).

Ishwaran, H., Kogalur, Udaya B, et al. (2008) RANDOM SURVIVAL FORESTS 1. *The Annals of Applied Statistics*. [Online] 2 (3), 841–860.

Jackson, C. A. et al. (2012) Area-based socioeconomic status, type 2 diabetes and cardiovascular mortality in Scotland. *Diabetologia*. [Online] 55 (11), 2938–2945. [online]. Available from: https://pubmed.ncbi.nlm.nih.gov/22893029/ (Accessed 8 August 2023).

JGF, C. et al. (2021) The struggle towards a Universal Definition of Heart Failure-how to proceed? *European heart journal*. [Online] 42 (24), 2331–2332. [online]. Available from: https://pubmed.ncbi.nlm.nih.gov/33791787/ (Accessed 25 November 2024).

Jiang, N. et al. (2024) Limitations of using COX proportional hazards model in cardiovascular research. *Cardiovascular Diabetology*. [Online] 23 (1), 1–2. [online]. Available from: https://cardiab.biomedcentral.com/articles/10.1186/s12933-024-02302-2 (Accessed 15 November 2024).

John Hopkins Medicine (2021) *Cardiomyopathy | Johns Hopkins Medicine* [online]. Available from: https://www.hopkinsmedicine.org/health/conditions-and-diseases/cardiomyopathy (Accessed 18 August 2021).

Jones, N. R. et al. (2020) Screening for atrial fibrillation: a call for evidence. *European Heart Journal*. [Online] 41 (10), 1075–1085. [online]. Available from: https://academic.oup.com/eurheartj/article/41/10/1075/5663566 (Accessed 18 August 2021).

Kaur, Narinder et al. (2023) 140 Use of machine learning to predict drivers of incident heart failure in patients with type 2 diabetes mellitus. *Heart*. [Online] 109 (Suppl 3), A162–A163. [online]. Available from: https://heart.bmj.com/content/109/Suppl_3/A162 (Accessed 7 January 2025).

Kaur, Narinder et al. (2024) Harnessing Artificial Intelligence For Predicting Cardiovascular And Disease Complications: An Overview Of Innovative Support Tools. *Journal of the Hong Kong College of Cardiology*. [Online] 31 (5), . [online]. Available from: https://www.jhkcc.com.hk/journal/vol31/iss5/2.

Kaur, N et al. (2024) Predicting mortality of type-2 diabetes mellitus by applying machine learning to electronic medical records in the west of scotland and hong kong. *European Heart Journal*. [Online] 45 (Supplement_1), . [online]. Available from: https://dx.doi.org/10.1093/eurheartj/ehae666.3478 (Accessed 8 January 2025).

Kaur, N et al. (2023) Use of machine learning to predict mortality in patients with type 2 diabetes mellitus, according to socioeconomic status. Health Economics

Kengne, A. P. (2013) The ADVANCE cardiovascular risk model and current strategies for cardiovascular disease risk evaluation in people with diabetes. Cardiovascular Journal of Africa 24 (9) p.376–381.

Khayyat-Kholghi, M. et al. (2021) Worsening Kidney Function Is the Major Mechanism of Heart Failure in Hypertension: The ALLHAT Study. *JACC: Heart Failure*. [Online] 9 (2), 100–111.

Kidney Research UK (2018) *Kidney disease in people from minority ethnic groups* [online]. Available from: https://www.kidneyresearchuk.org/kidney-health-information/about-kidney-disease/am-i-at-risk/kidney-disease-in-minority-ethnic-groups/ (Accessed 13 November 2024).

Kimenai, D. M. et al. (2022) Socioeconomic Deprivation: An Important, Largely Unrecognized Risk Factor in Primary Prevention of Cardiovascular Disease. *Circulation*. [Online] 146 (3), 240–248. [online]. Available from: https://www.ahajournals.org/doi/abs/10.1161/CIRCULATIONAHA.122.060042 (Accessed 11 August 2023).

Klein, S. et al. (2022) Why does obesity cause diabetes? *Cell Metabolism*. [Online] 34 (1), 11–20.

Kodama, K. et al. (2013) Ethnic Differences in the Relationship Between Insulin Sensitivity and Insulin ResponseA systematic review and meta-analysis. *Diabetes Care*. [Online] 36 (6), 1789–1796. [online]. Available from: https://dx.doi.org/10.2337/dc12-1235 (Accessed 15 October 2024).

Kokotailo, R. A. & Hill, M. D. (2005) Coding of Stroke and Stroke Risk Factors Using International Classification of Diseases, Revisions 9 and 10. *Stroke*. [Online] 36 (8), 1776–1781. [online]. Available from: https://www.ahajournals.org/doi/10.1161/01.STR.0000174293.17959.a1 (Accessed 3 November 2024).

Kolb, H. & Martin, S. (2017) Environmental/lifestyle factors in the pathogenesis and prevention of type 2 diabetes. *BMC Medicine 2017 15:1*. [Online] 15 (1), 1–11. [online]. Available from: https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-017-0901-x (Accessed 5 November 2024).

Kraemer, H. C. (1995) Statistical issues in assessing comorbidity. *Statistics in Medicine*. [Online] 14 (8), 721–733. [online]. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/sim.4780140803 (Accessed 2 September 2021).

Kwok, W. et al. (2024) Validation of diagnostic coding for chronic obstructive pulmonary disease in an electronic health record system in Hong Kong. *Hong Kong Medical Journal*. [Online]

Kyrou, I. et al. (2020) Sociodemographic and lifestyle-related risk factors for identifying vulnerable groups for type 2 diabetes: A narrative review with emphasis on data from Europe. *BMC Endocrine Disorders*. [Online] 20 (1), 1–13. [online]. Available from: https://bmcendocrdisord.biomedcentral.com/articles/10.1186/s12902-019-0463-3 (Accessed 5 November 2024).

Lai, Y. et al. (2013) Survival analysis by penalized regression and matrix factorization. *The Scientific World Journal*. [Online] 2013.

Lam, C. S. P. et al. (2016) Regional and ethnic differences among patients with heart failure in Asia: the Asian sudden cardiac death in heart failure registry. *European Heart Journal*.

[Online] 37 (41), 3141–3153. [online]. Available from: https://dx.doi.org/10.1093/eurheartj/ehw331 (Accessed 2 October 2024).

Lawson, C. A. et al. (2021) Outcome trends in people with heart failure, type 2 diabetes mellitus and chronic kidney disease in the UK over twenty years-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/). *EClinicalMedicine*. [Online] 32100739. [online]. Available from: https://doi.org/10.1016/j.eclinm.2021.100739 (Accessed 15 July 2024).

Lee, S. et al. (2021) Development of a predictive risk model for all-cause mortality in patients with diabetes in Hong Kong. *BMJ Open Diab Res Care*. [Online] 91950. [online]. Available from: http://drc.bmj.com/.

Lee, Y. H. A. et al. (2022) Risk of New-Onset Prostate Cancer for Metformin Versus Sulfonylurea Use in Type 2 Diabetes Mellitus: A Propensity Score-Matched Study. *JNCCN Journal of the National Comprehensive Cancer Network*. [Online] 20 (6), 674–682.

Leung, A. W. et al. (2015a) Management of heart failure with preserved ejection fraction in a local public hospital in Hong Kong. *BMC Cardiovascular Disorders*. [Online] 15 (1), 1–7. [online]. Available from: https://bmccardiovascdisord.biomedcentral.com/articles/10.1186/s12872-015-0002-8 (Accessed 25 September 2024).

Leung, A. W. et al. (2015b) Management of heart failure with preserved ejection fraction in a local public hospital in Hong Kong. *BMC Cardiovascular Disorders*. [Online] 15 (1), 1–7. [online]. Available from: https://bmccardiovascdisord.biomedcentral.com/articles/10.1186/s12872-015-0002-8 (Accessed 1 May 2025).

Levey, A. S. & Stevens, L. A. (2010) Estimating GFR Using the CKD Epidemiology Collaboration (CKD-EPI) Creatinine Equation: More Accurate GFR Estimates, Lower CKD Prevalence Estimates, and Better Risk Predictions. *American Journal of Kidney Diseases*. [Online] 55 (4), 622–627. [online]. Available from: http://www.ajkd.org/article/S0272638610004762/fulltext (Accessed 26 October 2024).

Li, Y. et al. (2020) Roles and mechanisms of renalase in cardiovascular disease: A promising therapeutic target. *Biomedicine and Pharmacotherapy*. [Online] 131.

Liane Ong, K. et al. (2023) *Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021*. [Online] [online]. Available from: https://doi.org/10.1016/S0140-6736 (Accessed 3 July 2023).

Libby, P. et al. (2002) Inflammation and atherosclerosis. *Circulation*. [Online] 105 (9), 1135–1143. [online]. Available from: https://www.ahajournals.org/doi/10.1161/hc0902.104353 (Accessed 5 November 2024).

Lin, B. et al. (2024) Younger-onset compared with later-onset type 2 diabetes: an analysis of the UK Prospective Diabetes Study (UKPDS) with up to 30 years of follow-up (UKPDS 92). *The lancet. Diabetes & endocrinology*. [Online] [online]. Available from: http://www.ncbi.nlm.nih.gov/pubmed/39461360.

Livingstone, S. J. et al. (2012) *Risk of Cardiovascular Disease and Total Mortality in Adults with Type 1 Diabetes: Scottish Registry Linkage Study*. [Online] [online]. Available from: www.plosmedicine.org (Accessed 30 August 2022).

Louhichi, M. et al. (2023) Shapley Values for Explaining the Black Box Nature of Machine Learning Model Clustering. *Procedia Computer Science*. [Online] 220806–811.

Lundberg, S. M. et al. (2017) *A Unified Approach to Interpreting Model Predictions*. [online]. Available from: https://github.com/slundberg/shap (Accessed 8 August 2023).

Ma, R. C. W. & Chan, J. C. N. (2013) Type 2 diabetes in East Asians: similarities and differences with populations in Europe and the United States. *Annals of the New York Academy of Sciences*. [Online] 1281 (1), 64. [online]. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC3708105/ (Accessed 12 November 2024).

Malachias, M. V. B. et al. (2020) Nt-probnp by itself predicts death and cardiovascular events in high-risk patients with type 2 diabetes mellitus. *Journal of the American Heart Association*. [Online] 9 (19), .

Marx, N. et al. (2023) 2023 ESC Guidelines for the management of cardiovascular disease in patients with diabetes. *European Heart Journal*. [Online] 44 (39), 4043–4140.

McDonagh, Theresa A. et al. (2021) 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. *European heart journal*. [Online] 42 (36), 3599–3726. [online]. Available from: https://pubmed.ncbi.nlm.nih.gov/34447992/ (Accessed 15 October 2024).

McDonagh, Theresa A et al. (2021) 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failureDeveloped by the Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC) With the special contribution of the Heart Failure Association (HFA) of the ESC. *European Heart Journal*. [Online] 42 (36), 3599–3726. [online]. Available from: https://academic.oup.com/eurheartj/article/42/36/3599/6358045 (Accessed 10 July 2022).

McDonagh, T. A. et al. (2023) 2023 Focused Update of the 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: Developed by the task force

for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC) With the special contribution of the Heart Failure Association (HFA) of the ESC. *European Heart Journal*. [Online] 44 (37), 3627–3639. [online]. Available from: https://dx.doi.org/10.1093/eurheartj/ehad195 (Accessed 11 July 2024).

Mcdonald, C. J. & Tierney, W. M. (1988) Computer-Stored Medical Records: Their Future Role in Medical Practice. *JAMA*. [Online] 259 (23), 3433–3440. [online]. Available from: https://jamanetwork.com/journals/jama/fullarticle/372415 (Accessed 18 October 2024).

Miao, F. et al. (2015) Is random survival forest an alternative to cox proportional model on predicting cardiovascular disease? *IFMBE Proceedings*. [Online] 45740–743.

Michaud, G. F. & Stevenson, W. G. (2021) Atrial Fibrillation Caren G. Solomon (ed.). *https://doi.org/10.1056/NEJMcp2023658*. [Online] 384 (4), 353–361. [online]. Available from: https://www.nejm.org/doi/full/10.1056/NEJMcp2023658 (Accessed 18 August 2021).

Mikolasch, T. A. et al. (2022) *Multi-center evaluation of baseline neutrophil-to-lymphocyte (NLR) ratio as an independent predictor of mortality and clinical risk stratifier in idiopathic pulmonary fibrosis*. [online]. Available from: www.thelancet.com (Accessed 9 August 2023).

Mills, K. T. et al. (2020) The global epidemiology of hypertension. *Nature reviews. Nephrology*. [Online] 16 (4), 223. [online]. Available from: /pmc/articles/PMC7998524/ (Accessed 16 August 2021).

Mok, J. et al. (2013) Quality Assurance of LOINC Mapping for Laboratory Tests – A Local Experience with People, Process and Technology. *Studies in Health Technology and Informatics*. [Online] 192 (1–2), 975–975. [online]. Available from: https://ebooks.iospress.nl/doi/10.3233/978-1-61499-289-9-975 (Accessed 29 October 2024).

Moody, A. et al. (2016) Social inequalities in prevalence of diagnosed and undiagnosed diabetes and impaired glucose regulation in participants in the Health Surveys for England series. *BMJ Open*. [Online] 610155. [online]. Available from: http://bmjopen.bmj.com/ (Accessed 26 June 2023).

Moran, G. M. et al. (2022) Type 2 diabetes: summary of updated NICE guidance. *BMJ*. [Online] o775. [online]. Available from: https://www.bmj.com/lookup/doi/10.1136/bmj.o775 (Accessed 4 June 2022).

Mullens, W. et al. (2019) The use of diuretics in heart failure with congestion — a position statement from the Heart Failure Association of the European Society of Cardiology. *European Journal of Heart Failure*. [Online] 21 (2), 137–155. [online]. Available from:

https://onlinelibrary.wiley.com/doi/full/10.1002/ejhf.1369 (Accessed 21 September 2024).

Nagar, S. D. et al. (2021) Socioeconomic deprivation and genetic ancestry interact to modify type 2 diabetes ethnic disparities in the United Kingdom. *eClinicalMedicine*. [Online] 37. [online]. Available from: http://www.thelancet.com/article/S2589537021002406/fulltext (Accessed 6 December 2024).

Ndrepepa, G. & Kastrati, A. (2019) Alanine aminotransferase—a marker of cardiovascular risk at high and low activity levels. *Journal of Laboratory and Precision Medicine*. [Online] 4 (0), . [online]. Available from: https://jlpm.amegroups.org/article/view/5118/html (Accessed 8 January 2025).

Nghiem, N. et al. (2024) Predicting the risk of diabetes complications using machine learning and social administrative data in a country with ethnic inequities in health: Aotearoa New Zealand. *BMC medical informatics and decision making*. [Online] 24 (1), 274.

NHS (2019) *General Data Protection Regulation (GDPR) - information - NHS Digital*.

NHS BMI (n.d.) *Body mass index (BMI) | NHS inform* [online]. Available from: https://www.nhsinform.scot/healthy-living/food-and-nutrition/healthy-eating-and-weight-management/body-mass-index-bmi/ (Accessed 23 November 2024).

Ni, M. Y. et al. (2021) Understanding longevity in Hong Kong: a comparative study with long-living, high-income countries. *The Lancet Public Health*. [Online] 6 (12), e919–e931.

NICE (2024a) *Diagnosis in adults | Diagnosis | Diabetes - type 2 | CKS | NICE* [online]. Available from: https://cks.nice.org.uk/topics/diabetes-type-2/diagnosis/diagnosis-in-adults/ (Accessed 7 October 2024).

NICE (2024b) *NICE Guidelines* [online]. Available from: https://www.nice.org.uk/ (Accessed 7 October 2024).

NICE (2023) *Overview | Cardiovascular disease: risk assessment and reduction, including lipid modification | Guidance*.

NICE Guidelines (2014) *Obesity: identifying, assessing and managing obesity in adults, young people and children Information for the public*. [online]. Available from: www.nice.org.uk.

of Scotland, T. N. R. (2019) National Records of Scotland. National Records of Scotland

Ormazabal, V. et al. (2018) Association between insulin resistance and the development of cardiovascular disease. Cardiovascular Diabetology 17 (1).

Østergaard, H. B. et al. (2023) Estimating individual lifetime risk of incident cardiovascular events in adults with Type 2 diabetes: an update and geographical calibration of the DIAbetes Lifetime perspective model (DIAL2). *European Journal of Preventive Cardiology*. [Online] 30 (1), 61–69. [online]. Available from: https://dx.doi.org/10.1093/eurjpc/zwac232 (Accessed 10 November 2024).

Paliogiannis, P. et al. (2018) *Neutrophil to lymphocyte ratio and clinical outcomes in COPD: recent evidence and future perspectives*. [Online] [online]. Available from: https://doi.org/10.1183/16000617.0113-2017 (Accessed 9 August 2023).

Pellicori, P. et al. (2021) Use of diuretics and outcomes in patients with type 2 diabetes: findings from the EMPA-REG OUTCOME trial. *European Journal of Heart Failure*. [Online] 23 (7), 1085–1093.

Ponikowski, P. et al. (2016) 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 Journal of Heart Failure*. [Online] 18 (8), 891–975.

Prinja, S. et al. (2010) Censoring in Clinical Trials: Review of Survival Analysis Techniques. *Indian Journal of Community Medicine : Official Publication of Indian Association of Preventive & Social Medicine*. [Online] 35 (2), 217. [online]. Available from: /pmc/articles/PMC2940174/ (Accessed 5 June 2022).

Psaltopoulou, T. et al. (2017) Socioeconomic status and risk factors for cardiovascular disease: Impact of dietary mediators. *Hellenic Journal of Cardiology*. [Online] 58 (1), 32–42.

Public Health Scotland (2024) *Smoking - Public Health Scotland* [online]. Available from: https://publichealthscotland.scot/population-health/improving-scotlands-health/smoking/overview/ (Accessed 12 November 2024).

Pylypchuk, R. et al. (2021) Cardiovascular risk prediction in type 2 diabetes before and after widespread screening: a derivation and validation study. *The Lancet*. [Online] 397 (10291), 2264–2274. [online]. Available from: http://www.thelancet.com/article/S0140673621005729/fulltext (Accessed 8 November 2024).

Quan, J. et al. (2019) Risk Prediction Scores for Mortality, Cerebrovascular, and Heart Disease Among Chinese People With Type 2 Diabetes. *The Journal of Clinical Endocrinology & Metabolism*. [Online] 104 (12), 5823–5830. [online]. Available from: https://dx.doi.org/10.1210/jc.2019-00731 (Accessed 8 November 2024).

Rahimi, K. et al. (2018) *Cardiovascular disease and multimorbidity: A call for interdisciplinary research and personalized cardiovascular care*. [Online] [online].

Available from: https://doi.org/10.1371/journal.pmed.1002545 (Accessed 17 August 2021).

Ramaty, E. et al. (2014) Low ALT blood levels predict long-term all-cause mortality among adults. A historical prospective cohort study. *European Journal of Internal Medicine*. [Online] 25 (10), 919–921.

Rawlins, M. (1999) In pursuit of quality: the National Institute for Clinical Excellence. *The Lancet*. [Online] 353 (9158), 1079–1082.

Razaghizad, A. et al. (2022) Clinical Prediction Models for Heart Failure Hospitalization in Type 2 Diabetes: A Systematic Review and Meta-Analysis. *Journal of the American Heart Association*. [Online] 11 (10), .

Razieh, C. et al. (2022) Differences in the risk of cardiovascular disease across ethnic groups: UK Biobank observational study Nutrition, Metabolism & Cardiovascular Diseases. *Nutrition, Metabolism and Cardiovascular Diseases*. [Online] 322594–2602. [online]. Available from: http://creativecommons.org/licenses/by/4.0/ (Accessed 5 July 2024).

Read, S. H. et al. (2017) Measuring the association between body mass index and all-cause mortality in the presence of missing data: Analyses from the Scottish national diabetes register. *American Journal of Epidemiology*. [Online] 185 (8), 641–649.

Reiffel, J. A. (2020) Propensity Score Matching: The 'Devil is in the Details' Where More May Be Hidden than You Know. *The American Journal of Medicine*. [Online] 133 (2), 178–181.

Ribeiro, M. T. et al. (2016) 'Why Should I Trust You?': Explaining the Predictions of Any Classifier. *NAACL-HLT 2016 - 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session*. [Online] 97–101. [online]. Available from: https://arxiv.org/abs/1602.04938v3 (Accessed 6 January 2025).

Rosano, G. M. et al. (2017) *Heart Failure in Patients with Diabetes Mellitus*. [Online] 3 (1), 52–57. [online]. Available from: www.CFRjournal.com.

Rosengren, A. (2024) Loop diuretics in cardiovascular disease: friend or foe? *European Heart Journal*. [Online] 45 (37), 3850–3852. [online]. Available from: https://dx.doi.org/10.1093/eurheartj/ehae483 (Accessed 14 October 2024).

Rosengren, A. et al. (2019) Socioeconomic status and risk of cardiovascular disease in 20 low-income, middle-income, and high-income countries: the Prospective Urban Rural Epidemiologic (PURE) study. *The Lancet Global Health*. [Online] 7 (6), e748–e760. [online]. Available from: www.thelancet.com/lancetgh.

Ryan, T. J. et al. (1996) ACC/AHA Guidelines for the Management of Patients with Acute Myocardial Infarction. A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Committee on Management of Acute Myocardial Infarction). *Journal of the American College of Cardiology*. [Online] 28 (5), 1328–1419. [online]. Available from: https://pubmed.ncbi.nlm.nih.gov/8890834/ (Accessed 31 July 2024).

Sanchis-Gomar, F. et al. (2016) Epidemiology of coronary heart disease and acute coronary syndrome. *Annals of Translational Medicine*. [Online] 4 (13), 256–256. [online]. Available from: https://atm.amegroups.org/article/view/10896/html (Accessed 27 June 2024).

Sattar, N. (2013) Gender aspects in type 2 diabetes mellitus and cardiometabolic risk. *Best Practice & Research Clinical Endocrinology & Metabolism*. [Online] 27 (4), 501–507.

Sattar, N. et al. (2023) Twenty Years of Cardiovascular Complications and Risk Factors in Patients With Type 2 Diabetes: A Nationwide Swedish Cohort Study. *Circulation*. [Online] 147 (25), 1872–1886.

Sattar, N. & Gill, J. M. R. (2014) Type 2 diabetes as a disease of ectopic fat? *BMC Medicine*. [Online] 12 (1), .

Savarese, G. et al. (2022) *Heart failure with mid-range or mildly reduced ejection fraction*. [Online] [online]. Available from: www.nature.com/nrcardio (Accessed 11 July 2024).

Schultz, W. M. et al. (2018) Socioeconomic status and cardiovascular outcomes: Challenges and interventions. *Circulation*. [Online] 137 (20), 2166–2178. [online]. Available from: /pmc/articles/PMC5958918/?report=abstract.

Scotland Census (2022) *Scotland's Census 2022 - Ethnic group, national identity, language and religion - Chart data | Scotland's Census* [online]. Available from: https://www.scotlandscensus.gov.uk/documents/scotland-s-census-2022-ethnic-group-national-identity-language-and-religion-chart-data/ (Accessed 5 July 2024).

Scottish Diabetes Group (2021) *Scottish Diabetes Survey 2021*. [online]. Available from: www.diabetesinscotland.org.uk/wp-content/uploads/2023/02/Diabetes-Scottish-Diabetes-Survey-2021-final-version.pdf (Accessed 24 November 2024).

Scottish Government (2023) *Smoking Prevalence* . [online]. Available from: https://www.gov.scot/publications/scottish-health-survey-2022-volume-1-main-report/pages/11/ (Accessed 1 May 2025).

Seah, J. Y. H. et al. (2023) Differences in type 2 diabetes risk between East, South, and Southeast Asians living in Singapore: the multi-ethnic cohort. *BMJ Open Diabetes Research and Care*. [Online] 11 (4), .

Segar, M. W. et al. (2019) Machine Learning to Predict the Risk of Incident Heart Failure Hospitalization Among Patients With Diabetes: The WATCH-DM Risk Score. *2298 Diabetes Care*. [Online] 42. [online]. Available from: http://care.diabetesjournals.org/lookup/suppl/.

Segerer, S. & Seeger, H. (2018) Epidemiology and chronic kidney disease as a cardiovascular risk factor. *ESC CardioMed*. [Online] 979–981.

Shahlan Kasim, S. et al. (2023) *Validation of the general Framingham Risk Score (FRS), SCORE2, revised PCE and WHO CVD risk scores in an Asian population*. [online]. Available from: www.thelancet.com.

SIGN (2023) *Scottish Intercollegiate Guidelines Network* [online]. Available from: https://www.sign.ac.uk/ (Accessed 7 October 2024).

Sirocchi, C. et al. (2024) Medical-informed machine learning: integrating prior knowledge into medical decision systems. *BMC Medical Informatics and Decision Making*. [Online] 24 (Suppl 4), .

van Smeden, M. et al. (2021) Approaches to addressing missing values, measurement error, and confounding in epidemiologic studies. *Journal of Clinical Epidemiology*. [Online] 13189–100.

Smokefree HK (2024) *Smoking Prevalence | COSH* [online]. Available from: https://www.smokefree.hk/smoking-trend.php?lang=en (Accessed 28 October 2024).

Socrates Y WU1, K. C. W. , D. Y. C. , S. H. , H. S. T. , V. W. L. 3 , T. L. , M. W. (2021) *Tobacco Control Policy-related Survey 2020*. (COSH Report NO.29), 12. [online]. Available from: https://www.smokefree.hk/uploadedFile/COSHRN_E29.pdf (Accessed 28 October 2024).

Sotomi, Y. et al. (2021) Sex differences in heart failure with preserved ejection fraction. *Journal of the American Heart Association*. [Online] 10 (5), 1–20. [online]. Available from: https://www.ahajournals.org/doi/10.1161/JAHA.120.018574 (Accessed 13 November 2024).

Spooner, A. et al. (2020a) A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. *Scientific Reports 2020 10:1*. [Online] 10 (1), 1–10. [online]. Available from: https://www.nature.com/articles/s41598-020-77220-w (Accessed 14 July 2021).

Spooner, A. et al. (2020b) A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. *Scientific Reports 2020 10:1*. [Online] 10 (1), 1–10. [online]. Available from: https://www.nature.com/articles/s41598-020-77220-w (Accessed 3 January 2025).

Srujana, B. et al. (2022) *Machine Learning vs. Survival Analysis Models: a study on right censored heart failure data*. [Online] [online]. Available from: https://www.tandfonline.com/action/journalInformation?journalCode=lssp20.

Sterne, J. A. C. et al. (2009) Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *The BMJ*. [Online] 338 (7713), b2393. [online]. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC2714692/ (Accessed 30 December 2024).

Stevens, L. A. et al. (2010) Comparative Performance of the CKD Epidemiology Collaboration (CKD-EPI) and the Modification of Diet in Renal Disease (MDRD) Study Equations for Estimating GFR Levels Above 60 mL/min/1.73 m2. *American journal of kidney diseases : the official journal of the National Kidney Foundation*. [Online] 56 (3), 486. [online]. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC2926290/ (Accessed 26 October 2024).

Steyerberg, E. W. et al. (2018) Poor performance of clinical prediction models: the harm of commonly applied methods. *Journal of Clinical Epidemiology*. [Online] 98133–143.

Stringhini, S. et al. (2013) Association of Lifecourse Socioeconomic Status with Chronic Inflammation and Type 2 Diabetes Risk: The Whitehall II Prospective Cohort Study. *PLoS Medicine*. [Online] 10 (7), .

Suchting, R. et al. (2017) Using Elastic Net Penalized Cox Proportional Hazards Regression to Identify Predictors of Imminent Smoking Lapse. *Nicotine & Tobacco Research*. [Online] 21 (2), 173. [online]. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC7962780/ (Accessed 10 December 2024).

Suzuki, K. et al. (2024) Genetic drivers of heterogeneity in type 2 diabetes pathophysiology. *Nature*. [Online] 627 (8003), 347–357.

Tan, M. et al. (2020) Including social and behavioral determinants in predictive models: Trends, challenges, and opportunities. JMIR Medical Informatics 8 (9).

The Scottish Government (2017) *Scottish Index of Multiple Deprivation 2016: introductory booklet*. (January), 121–131. [online]. Available from: https://www.gov.scot/publications/scottish-index-multiple-deprivation-2016/ (Accessed 7 January 2025).

Timmis, A. et al. (2022) European Society of Cardiology: cardiovascular disease statistics 2021. *European Heart Journal*. [Online] 43716–799. [online]. Available from: https://doi.org/10.1093/eurheartj/ehab892 (Accessed 4 July 2024).

Tonekaboni, S. et al. (2019) *What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use*. [online]. Available from: http://arxiv.org/abs/1905.05134.

Tse, G. et al. (2024) Healthcare Big Data in Hong Kong: Development and Implementation of Artificial Intelligence-Enhanced Predictive Models for Risk Stratification. *Current Problems in Cardiology*. [Online] 49 (1), 102168.

Tsoi, M. F. et al. (2020) Epidemiology of gout in Hong Kong: A population-based study from 2006 to 2016. *Arthritis Research and Therapy*. [Online] 22 (1), 1–9. [online]. Available from: https://arthritis-research.biomedcentral.com/articles/10.1186/s13075-020-02299-5 (Accessed 28 October 2024).

Tziomalos, K. et al. (2010) Endothelial dysfunction in metabolic syndrome: prevalence, pathogenesis and management. *Nutrition, metabolism, and cardiovascular diseases : NMCD*. [Online] 20 (2), 140–146. [online]. Available from: https://pubmed.ncbi.nlm.nih.gov/19833491/ (Accessed 5 November 2024).

do Vale Moreira, N. C. et al. (2021) Race/ethnicity and challenges for optimal insulin therapy. *Diabetes Research and Clinical Practice*. [Online] 175. [online]. Available from: http://www.diabetesresearchclinicalpractice.com/article/S0168822721001820/fulltext (Accessed 15 October 2024).

Verma, S. et al. (2023) Effects of once-weekly semaglutide 2.4 mg on C-reactive protein in adults with overweight or obesity (STEP 1, 2, and 3): Exploratory analyses of three randomised, double-blind, placebo-controlled, phase 3 trials. *eClinicalMedicine*. [Online] 55. [online]. Available from: http://www.thelancet.com/article/S2589537022004667/fulltext (Accessed 10 December 2024).

Verma, S. et al. (2024) Inflammation in Obesity-Related HFpEF: The STEP-HFpEF Program. *Journal of the American College of Cardiology*. [Online] 84 (17), . [online]. Available from: https://pubmed.ncbi.nlm.nih.gov/39217564/ (Accessed 25 November 2024).

Vock, D. M. et al. (2016) Adapting machine learning techniques to censored time-to-event health record data: A general-purpose approach using inverse probability of censoring weighting. *Journal of Biomedical Informatics*. [Online] 61119–131.

Wan, E. Y. F. et al. (2016) Association of Hemoglobin A1c Levels With Cardiovascular Disease and Mortality in Chinese Patients With Diabetes. *Journal of the American College of Cardiology*. [Online] 67 (4), 456–458. [online]. Available from: https://www.jacc.org/doi/10.1016/j.jacc.2015.11.020 (Accessed 13 November 2024).

Wan, E. Y. F. et al. (2023) Diabetes with poor-control HbA1c is cardiovascular disease 'risk equivalent' for mortality: UK Biobank and Hong Kong population-based cohort study. *BMJ Open Diabetes Research and Care*. [Online] 11 (1), .

Wang, H. et al. (2021) Prevalence and Incidence of Heart Failure among Urban Patients in China: A National Population-Based Analysis. *Circulation: Heart Failure*. [Online] 14 (10), E008406.

Wang, S. V et al. (2022) *Reproducibility of real-world evidence studies using clinical practice data to inform regulatory and coverage decisions*. [Online] [online]. Available from: https://doi.org/10.1038/s41467-022-32310-3 (Accessed 2 October 2024).

Wells, S. et al. (2017) Cohort Profile: The PREDICT cardiovascular disease cohort in New Zealand primary care (PREDICT-CVD 19). *International Journal of Epidemiology*. [Online] 46 (1), .

Wen, H. et al. (2022) Comparison of trend in chronic kidney disease burden between China, Japan, the United Kingdom, and the United States. *Frontiers in Public Health*. [Online] 10999848.

WHO (2024) *Chronic obstructive pulmonary disease- WHO* [online]. Available from: https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd) (Accessed 12 November 2024).

Wilcox, C. S. et al. (2020) Pathophysiology of Diuretic Resistance and Its Implications for the Management of Chronic Heart Failure. *Hypertension*. [Online] 76 (4), 1045–1054. [online]. Available from: https://www.ahajournals.org/doi/10.1161/HYPERTENSIONAHA.120.15205 (Accessed 14 November 2024).

Wise, J. (2021) Diabetes: BMI cut-offs designed to trigger action are too high for some ethnic populations, say researchers. *BMJ*. [Online] 373. [online]. Available from: https://www.bmj.com/content/373/bmj.n1217 (Accessed 12 October 2024).

Witte, K. K. et al. (2018) Socioeconomic deprivation and mode-specific outcomes in patients with chronic heart failure. *Heart*. [Online] 104 (12), 993–998. [online]. Available from: https://heart.bmj.com/content/104/12/993 (Accessed 19 August 2021).

WoC (2019) *Hong Kong Women in Figures 2019*. [online]. Available from: https://www.women.gov.hk/en/publications/reports.html (Accessed 13 October 2024).

Women's Comission (2021) *Women's Commission - Reports and Publications* [online]. Available from: https://www.women.gov.hk/en/publications/reports.html (Accessed 13 October 2024).

Wong, M. C. S. et al. (2008) Health services research in the public healthcare system in Hong Kong: An analysis of over 1 million antihypertensive prescriptions between 2004-2007 as an example of the potential and pitfalls of using routinely collected electronic patient

data. *BMC Health Services Research*. [Online] 8. [online]. Available from: http://dx.doi.org/10.1186/1472-6963-8-138 (Accessed 15 October 2024).

Wright, M. A. et al. (2019) What is the Impact of Social Deprivation on Physical and Mental Health in Orthopaedic Patients? *Clinical Orthopaedics and Related Research*. [Online] 477 (8), 1825. [online]. Available from: /pmc/articles/PMC7000003/ (Accessed 19 August 2021).

Yang, A. et al. (2022) Glucose-lowering drug use, glycemic outcomes, and severe hypoglycemia: 18-Year trends in 0¢9 million adults with Diabetes in Hong Kong (2002−2019). *The Lancet Regional Health - Western Pacific*. [Online] 26100509. [online]. Available from: https://doi.org/10.1016/j. (Accessed 7 October 2024).

Yang, X. et al. (2008) *Cardiovascular Diabetology Development and validation of a risk score for hospitalization for heart failure in patients with Type 2 Diabetes Mellitus*. [Online] [online]. Available from: http://www.cardiab.com/content/7/1/9.

Zakeri, R. et al. (2021a) Under-recognition of heart failure in patients with atrial fibrillation and the impact of gender: a UK population-based cohort study. *BMC Medicine*. [Online] 19 (1), .

Zakeri, R. et al. (2021b) Under-recognition of heart failure in patients with atrial fibrillation and the impact of gender: a UK population-based cohort study. *BMC medicine*. [Online] 19 (1), . [online]. Available from: https://pubmed.ncbi.nlm.nih.gov/34372832/ (Accessed 14 October 2024).

Zannad, F. et al. (2020) SGLT2 inhibitors in patients with heart failure with reduced ejection fraction: a meta-analysis of the EMPEROR-Reduced and DAPA-HF trials. *The Lancet*. [Online] 396 (10254), 819–829. [online]. Available from: http://www.thelancet.com/article/S0140673620318249/fulltext (Accessed 25 November 2024).

Zheng, Y. et al. (2018) Global aetiology and epidemiology of type 2 diabetes mellitus and its complications. *Nat Rev Endocrinol*. [Online] 14 (2), 88–98.

Zhou, J., Lee, S., Lakhani, I., et al. (2022) Adverse Cardiovascular Complications following prescription of programmed cell death 1 (PD-1) and programmed cell death ligand 1 (PD-L1) inhibitors: a propensity-score matched Cohort Study with competing risk analysis. *Cardio-Oncology*. [Online] 8 (1), .

Zhou, J., Lee, S., Liu, X., et al. (2022) Hip fractures risks in edoxaban versus warfarin users: A propensity score-matched population-based cohort study with competing risk analyses. *Bone*. [Online] 156116303.

Zhou, L. et al. (2022) Loop Diuretics Are Associated with Increased Risk of Hospital-Acquired Acute Kidney Injury in Adult Patients: A Retrospective Study. *Journal of Clinical Medicine*. [Online] 11 (13), 3665. [online]. Available from: https://www.mdpi.com/2077-0383/11/13/3665/htm (Accessed 2 October 2024).

Ziaeian, B. & Fonarow, G. C. (2016) Epidemiology and aetiology of heart failure. *Nature Reviews Cardiology 2016 13:6*. [Online] 13 (6), 368–378. [online]. Available from: https://www.nature.com/articles/nrcardio.2016.25 (Accessed 25 September 2024).