DEVELOPMENT OF A CARDIOMETABOLIC DISEASE POLICY MODEL: LEVERAGING THE ROLE OF UK PRIMARY CARE DATA

Septiara Putri
BSc., MPH

Submitted in fulfillment of the requirements for the Degree of
Doctor of Philosophy (PhD)

Health Economics and Health Technology Assessment (HEHTA)
School of Health and Wellbeing
College of Medical, Veterinary, and Life Sciences
University of Glasgow

2025

Abstract

Cardiometabolic diseases, including type 2 diabetes mellitus (T2DM) and cardiovascular disease (CVD), pose a growing public health burden globally and in the UK. Effective policy responses require robust modeling tools to evaluate the long-term clinical and economic impacts, particularly for preventative interventions. This thesis presents the development of a cardiometabolic disease (CMD) policy model designed to simulate the natural history and progression of major cardiometabolic conditions.

The model adopts a multi-state survival analysis model with semi-Markov structure, by utilising real-world patient-level data from the Clinical Practice Research Datalink (CPRD) Aurum, linked with Hospital Episode Statistics (HES), mortality records, and the Index of Multiple Deprivation (IMD). It estimates transition probabilities across key health states: disease-free, T2DM, first and recurrent cardiovascular events, and death.

Both parametric and flexible survival models are explored to estimate transition risks and enable long-term extrapolation. The model also incorporates time-dependent covariates, allowing risks to evolve as patient characteristics change. Model performance is assessed through rigorous diagnostics and validation.

A key feature of this model is its hybrid approach, which combines cohort-based transitions with microsimulation components. This structure captures both population-level trends and individual-level heterogeneity, enhancing the model's flexibility and relevance for policy analysis. Model outputs include life-years, quality-adjusted life years (QALYs), and healthcare costs, with also the extended ability to assess outcomes across different ethnic groups and explore health inequalities.

This CMD policy model offers a flexible, real-world-informed decision-support tool for policymakers, health economists, and public health planners. Its hybrid structure provides a foundation for supporting the long-term clinical and economic impacts of interventions to reduce the burden of cardiometabolic diseases in the UK population.

Table of contents

List of tables

List of figures

List of equations

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my supervisors Dr. Claudia Geue, Dr. Giorgio Ciminata, and Professor Jim Lewsey for their invaluable guidance, encouragement, and patience throughout the course of this research. I can honestly say this was an incredible supervisory team. I always felt genuinely backed, understood, and uplifted by them during this extraordinary journey.

I am profoundly grateful to Indonesian Endowment Fund for Education (LPDP), Ministry of Finance, Indonesia, for granting me the full scholarship that made my PhD study possible.

My sincere thanks to Dr. Bhautesh Jani from General Practice & Primary Care (GPPC), for facilitating access to CPRD data and offering consistent guidance and clinical expertise particularly in model conceptualisation part.

To my PhD friends and colleagues at HEHTA, thank you for creating such an incredibly vibrant, welcoming, and encouraging environment.

I would also extend my appreciation to my mentors and colleagues at the Health Policy and Administration Department, as well as the Center for Health Economics and Policy Studies (CHEPS), Faculty of Public Health, University of Indonesia, for their ongoing support and encouragement.

To Grant, thank you for being my calm in the chaos, my constant emotional support, and my quiet supporter behind the scenes.

To my family, my parents (my late Dad A. Zaini and my Mom Zaina Y.), my sister (Opialeta) and her family, and my three adorable nephews. Your love and support, from afar, lifted me through the hardest days and reminded me of where I came from. You are the only ones who didn't ask *"So when are you finishing?"* (at least..not too often).

Completing a PhD is like running a marathon, starting it during a global pandemic made the journey even more surreal. Finally, thank you for showing up, learning, adapting, and enduring. You made it to the finish line.

Publications, working papers, presentations

Articles arising from this thesis:

Putri, S., Ciminata, G., Lewsey, J., Kamaruzaman, H.F.B., Duan, Y. and Geue, C., 2025. Policy models for preventative interventions in cardiometabolic diseases: a systematic review. *BMC Health Services Research*, 25(1), p.635.

Putri, S., Ciminata, G., Lewsey, J., Jani, B., McMeekin, N. and Geue, C., 2024. The conceptualisation of cardiometabolic disease policy model in the UK. *BMC Health Services Research*, 24(1), p.1060.

Articles which were also published during the PhD period:

Putri, S. and Masytoh, L.S., 2024. A decade of HTA in Indonesia: Methodological challenges in conducting economic evaluation. *Health Policy and Technology*, 13(4), p.100894.

Putri, S., Saldi, S.R.F., Khoe, L.C., Setiawan, E., Megraini, A., Santatiwongchai, B., Nugraha, R.R., Permanasari, V.Y., Nadjib, M., Sastroasmoro, S. and Armansyah, A., 2023. Cetuximab as first-line treatment for metastatic colorectal cancer (mCRC): a model-based economic evaluation in Indonesia setting. *BMC Cancer*, 23(1), p.731.

Falkowski, A., Ciminata, G., Manca, F., Bouttell, J., Jaiswal, N., Farhana Binti Kamaruzaman, H., Hollingworth, S., Al-Adwan, M., Heggie, R., **Putri, S**. and Rana, D., 2023. How least developed to lower-middle income countries use health technology assessment: a scoping review. *Pathogens and Global Health*, 117(2), pp.104-119.

Putri, S., Nugraha, R.R., Pujiyanti, E., Thabrany, H., Hasnur, H., Istanti, N.D., Evasari, D. and Afiatin, 2022. Supporting dialysis policy for end stage renal disease (ESRD) in Indonesia: an updated cost-effectiveness model. *BMC Research Notes*, 15(1), p.359.

Putri, S., Setiawan, E., Saldi, S.R.F., Khoe, L.C., Sari, E.R., Megraini, A., Nadjib, M., Sastroasmoro, S. and Armansyah, A., 2022. Adding rituximab to chemotherapy for diffuse large B-cell lymphoma patients in Indonesia: a cost utility and budget impact analysis. *BMC Health Services Research*, 22(1), p.553.

Nadjib, M., Dewi, R.K., Setiawan, E., Miko, T.Y., **Putri, S.**, Hadisoemarto, P.F., Sari, E.R., Pujiyanto, Martina, R. and Syamsi, L.N., 2022. Cost and affordability of scaling up tuberculosis diagnosis using Xpert MTB/RIF testing in West Java, Indonesia. *PLoS One,* 17(3), p.e0264912.

Presentations

The conceptualisation of cardiometabolic disease policy model in the UK. The Professional Society for Health Economics and Outcomes Research (ISPOR), Copenhagen, 12th -15th November 2023.

Cardiometabolic diseases prevention policy models: a systematic review to inform conceptual model development. The Professional Society for Health Economics and Outcomes Research (ISPOR), Copenhagen, 12th -15th November 2023.

Initial handling for CPRD Data. Data Science Clinic-School of Health and Wellbeing (SHW), University of Glasgow, Glasgow 21st April 2023.

The development of cardiometabolic disease policy model in the UK. SHW PGR presentation. 11th January 2022.

Author's declaration

I declare that, except where explicit reference is made to the contribution of others, this dissertation is the result of my own work and has not been submitted for any other degree at the University of Glasgow or any other institution.

Chapters 3 and 4 have been published in peer-reviewed journals. All drafting and analysis were undertaken by myself, and contributions by collaborating authors were limited to providing input and comments on manuscript drafts. The revised content and additional work during peer-reviewed process and inclusion on this thesis were carried out independently by me.

For the case studies presented in Chapter 7, the health interventions concept/ideas are based on published Technology Appraisal (TA) reports by the National Institute for Health and Care Excellence (NICE) UK. All subsequent analysis, including model development and updated relevant parameters, was carried out by myself using my own model.

Signature:

Septiara Putri

Abbreviations

AF	:	Atrial Fibrillation
AFT	:	Accelerated Failure Time
AIC	:	Akaike Information Criterion
BIC	:	Bayesian Information Criterion
BHF	:	British Heart Foundation
BMI	:	Body Mass Index
BNF	:	British National Formulary
CBA	:	Cost Benefit Analysis
CCA	:	Complete Case Analysis
CEA	:	Cost-Effectiveness Analysis
CEAC	:	Cost-Effectiveness Acceptability Curve
CHD	:	Coronary Heart Disease
CIF	:	Cumulative Incidence Function
CMA	:	Cost Minimisation Analysis
CUA	:	Cost Utility Analysis
CI	:	Confidence Interval
CMD	:	Cardiometabolic Disease
CMS	:	Cardiometabolic Syndrome
CPRD	:	Clinical Practice Research Datalink
CVD	:	Cardiovascular Disease
CSH	:	Cause-specific Hazard
DBP	:	Diastolic Blood Pressure
DSA	:	Deterministic Sensitivity Analysis
DCEA	:	Distributional Cost-Effectiveness Analysis
EHR	:	Electronic Health Record
HBA1C	:	Hemoglobin A1c (glycated hemoglobin)
HDL	:	High-Density Lipoprotein (good cholesterol)
HES	:	Hospital Episode Statistics
HR	:	Hazard Ratio
HRQoL	:	Health Realted Quality of Life
HTA	:	Health Technology Assessment
ICER	:	Incremental Cost-Effectiveness Ratio
ICD-10	:	International Classification of Diseases, 10th Revision
IMD	:	Index of Multiple Deprivation

LDL	:	Low-Density Lipoprotein
KM	:	Kaplan-Meier
MI	:	Myocardial Infarction
MSM	:	Multi-State Model
NHS	:	National Health Service
NICE	:	National Institute for Health and Care Excellence
NRT	:	Nicotine Replacement Therapy
ONS	:	Office for National Statistics
PH	:	Proportional Hazard
PRISMA	:	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PSA	:	Probabilistic Sensitivity Analysis
QALY	:	Qauality Adjusted Life Years
RCT	:	Randomised Controlled Trial
RP	:	Royston-Parmar
RR	:	Relative Risk
RWD	:	Real World Data
RWE	:	Real World Evidence
SR	:	Systematic Review
STM	:	State Transition Model
SSB	:	Sugar-Sweetened Beverage
TA	:	Technology Appraisal
T2DM	:	Type 2 Diabetes
WHO	:	World Health Organization
WTP	:	Willingness to Pay

Chapter 1 Introduction

1.1 Overview

Cardiometabolic diseases (CMD), encompassing conditions such as type 2 diabetes mellitus (T2DM) and cardiovascular disease (CVD), represent a major and growing global health concern. Driven by demographic shifts, urbanisation, and changes in lifestyle behaviours, the prevalence of CMD has increased markedly in recent decades. This trend poses significant challenges not only to population health but also to the sustainability of healthcare systems and the broader socioeconomic landscape.

The World Health Organization (WHO) has identified T2DM and CVD as leading contributors to global morbidity and mortality.[1,2] These conditions are closely interrelated, sharing a cluster of modifiable risk factors, including obesity, poor diet, physical inactivity, tobacco use, and dyslipidaemia, therefore often conceptualised collectively under the term cardiometabolic disease. Together, they account for a substantial proportion of preventable illness and healthcare expenditure worldwide.[3]

This chapter provides a structured overview of cardiometabolic diseases, including their clinical and epidemiological characteristics, associated disease burden, and economic implications. It also examines key risk factors, current strategies for prevention and management, and the role of policy and economic modelling in informing healthcare decision-making. Particular attention is given to the use of real-world evidence (RWE) in enhancing the applicability and impact of such models in policymaking.

1.2 Cardiometabolic disease: a definition

Cardiometabolic disease (CMD) is an umbrella term used to describe a spectrum of interrelated conditions that affect metabolic processes, vascular function, and cardiovascular health. It encompasses disorders such as, hypertension, dyslipidaemia, and obesity, which collectively increase the risk of type 2 diabetes (T2DM) disease and/or cardiovascular disease (CVD). Historically, the term cardiometabolic syndrome (CMS) was used to refer to a cluster of metabolic abnormalities, including insulin resistance, impaired glucose tolerance, dyslipidaemia, central adiposity, and elevated blood pressure.[4-6] Although CMS remains a recognised clinical construct, there has been a shift toward framing these conditions more broadly under the CMD umbrella due to their shared pathophysiological pathways and cumulative impact on cardiovascular outcomes.

The World Health Organization (WHO), the National Cholesterol Education Program (NCEP), and the American Association of Clinical Endocrinologists (AACE) have contributed to the conceptual development of this framework and recognise the clinical significance of cardiometabolic risk clustering.[7]

The pathophysiology of CMD is multifaceted, involving an intertwined collective mechanism between genetic predisposition, metabolic pathways, and environmental exposure, compounded by lifestyle factors and hormonal imbalances. Insulin resistance plays a central role in the development of these conditions, leading to impaired glucose and lipid metabolism and contributing to the accumulation of visceral fat.[6-9] In turn, this inflammatory process further damages metabolic and cardiovascular function. While each of these conditions can co-occur (Figure 1.1), this significantly raises the risk of developing type 2 diabetes (T2DM) and cardiovascular diseases (CVDs).[7,10,11] Identifying and understanding the risk factors are crucial for further prevention, early detection, and management.

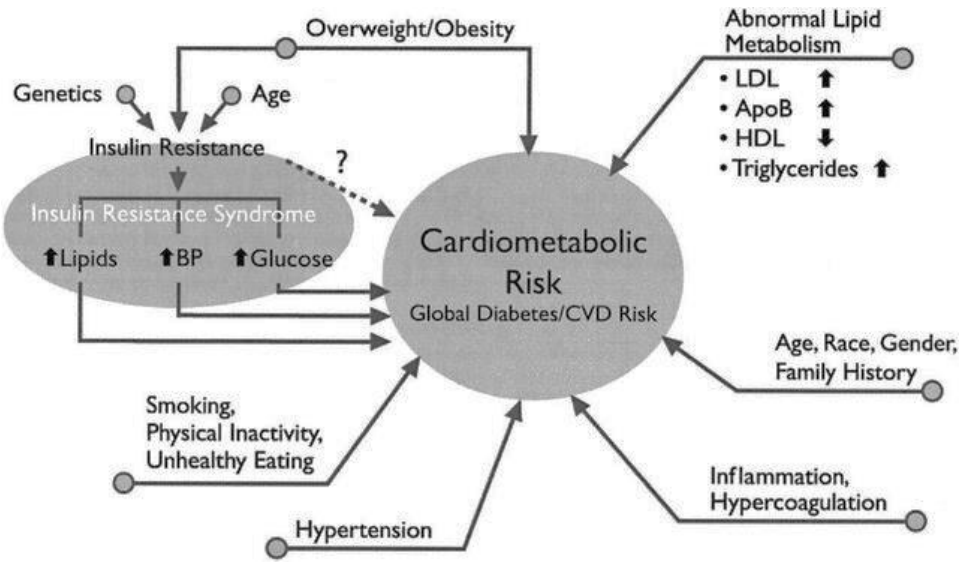


Figure 1.1 Factors contributing to the cardiometabolic risk[10]

The complex interplay of non-modifiable and modifiable risk factors influences cardiometabolic diseases.[6] Non-modifiable risk factors are crucial for assessing overall risk and tailoring preventive measures. These include age, sex, genetic predispositions, and ethnicity. For example, the risk of developing CMD increases with age, and men generally have a higher risk of CVD, while women's risk increases and may surpass men's risk after menopause.[12,13] Ethnicity is also associated with T2DM and CVD, people from South Asian, African, and Hispanic origin have a higher risk of developing the disease compared with individuals of White/European origin.[14–16]

On the other hand, low physical activity, an unhealthy diet, smoking, and alcohol consumption are considered modifiable risk factors [16,17]. These risks can be altered, reduced, or eliminated through behavioural changes and lie within an individual's control. Targeting the modifiable risk factors can decrease the likelihood of developing CMD.

1.3 Burden of cardiometabolic disease

1.3.1 Disease burden

The global prevalence of diabetes has increased rapidly, with over 536 million people living with diabetes, and it is projected to increase to 643 million in 2030.[18] In 2021, diabetes was the direct cause of approximately 1.6 million deaths, with more than 90% having T2DM.[2] Diabetes also doubles the risk of CVD mortality.[19] Meanwhile, CVDs (including coronary heart disease, stroke, and peripheral artery disease) account for an estimated 17.9 million deaths annually.[1] Based on these recent estimates, CVDs remain the leading cause of death, contributing 32% of all global deaths.[1]

Regarding regional variation, T2DM is highly prevalent in high-income countries (HICs), but the rate of increase has somewhat stabilised due to the improved healthcare system and disease management strategies.[20,21] The incidence of CVDs has generally declined as well.[22,23] However, CVDs remain a leading cause of death and disability.[23] In contrast, T2DM and CVDs are on the rise in LMICs, driven by lifestyle change, urbanisation, infrastructure development challenges, and the ageing population. This causes a double disease burden in these countries, both from communicable and non-communicable diseases.[20,24,25]

In the UK, there are approximately 4.6 million people diagnosed with diabetes, of which 90% of adults are diagnosed with T2DM (Table 1.1). The prevalence of T2DM increases significantly with age, rising to 16% among adults aged 75 and over.[26] In addition, an estimated 850,000 people are currently undiagnosed. The National Institute for Health and Care Excellence (NICE) estimates that more than 5 million people in the UK will be diagnosed with diabetes (both types) by 2025.[27]

Approximately 7.6 million people are living with CVD in the UK, including around 4 million males and 3.6 million females (Table 1.1).[28] When all diseases of the circulatory system are considered, including congenital heart disease, vascular dementia, and cardiovascular conditions originating in the perinatal period—CVD was recorded as the underlying cause of 163,888 deaths in 2019, accounting for 27.1% of all deaths in the UK.

As expected, mortality was heavily concentrated in older age groups, with 73.6% of CVD deaths occurring among individuals aged 75 years and over. However, more than 43,000 deaths occurred before the age of 75, indicating a substantial burden of premature mortality.[29]

Table 1.1 Number of people diagnosed with diabetes and cardiovascular disease by UK nation (2023)[28]

Nation	Diabetes	CVD
England	3.8 million	6.4 million
Scotland	310,000	730,000
Wales	220,000	340,000
Northern Ireland	110,000	225,000
UK	**4.6 million**	**7.6 million**

Diabetes (both type 1 and type 2) contributes to a substantial burden of mortality. With millions living with the condition, diabetes increases the risk of developing other serious health conditions that could lead to premature death.

Figure 1.2 shows the mortality rate from diabetes in the UK between 2000 and 2021, measured in deaths per 100,000 population. From 2014 to 2019, there was a slight increase in mortality, nearing 10 deaths per 100,000, before dropping slightly again. In 2021, the rate was recorded at 9 deaths per 100,000 population.[30]

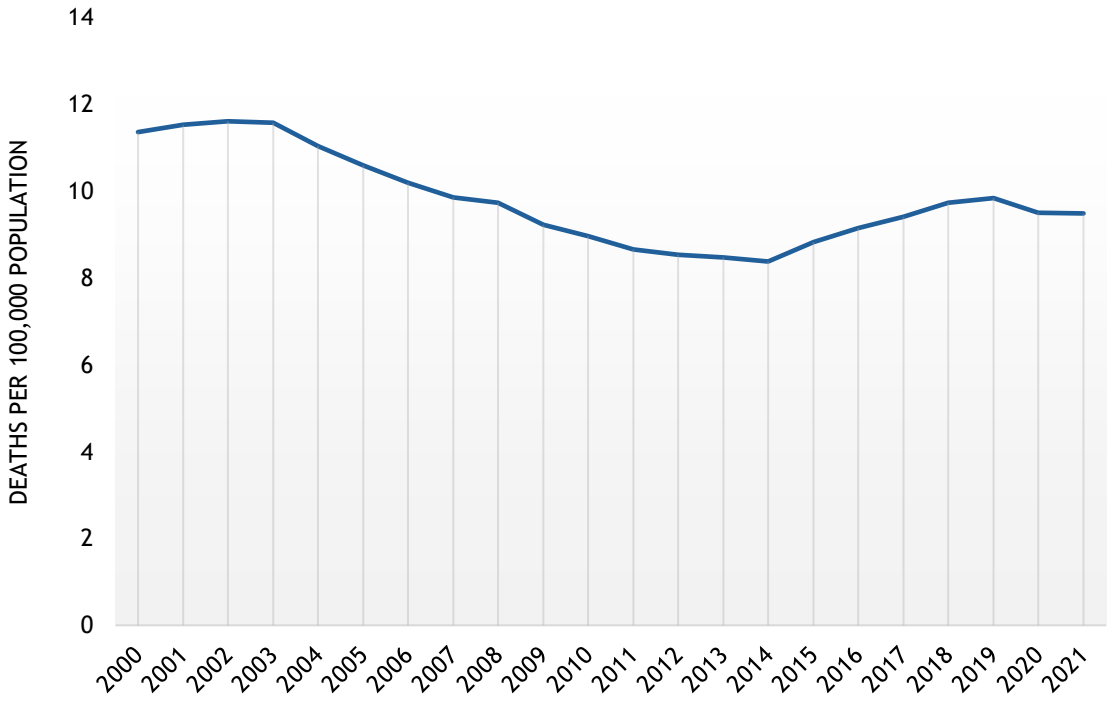


Figure 1.2 Mortality rates from diabetes in the UK 2000-2021 (per 100,000 population)[30]

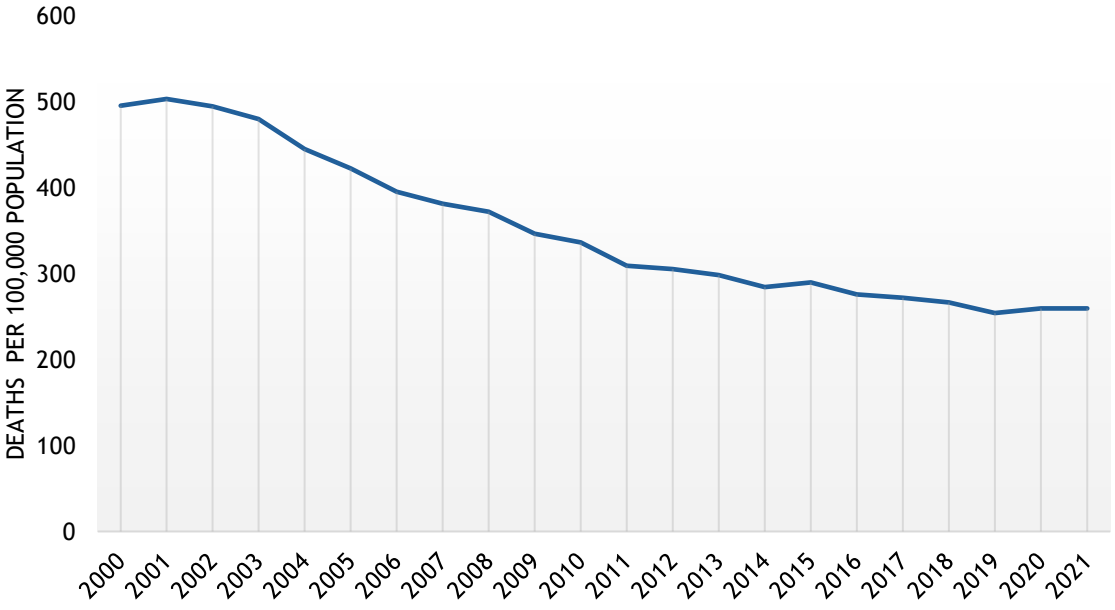


Figure 1.3 Mortality rates from cardiovascular disease (CVD) in the UK 2000-2021 (per 100,000 population)[31]

For CVD, the mortality burden is declining in the UK, particularly due to the reduction in coronary heart disease (CHD) and stroke. In 2021, there were 259 deaths per 100,000 population from CVD, the lowest rate over the period (Figure 1.3).[31] Despite this improvement, the number of people living with CVD remains high. The British Heart Foundation (BHF) notes that heart and circulatory diseases cause more than a quarter of all deaths in the UK, or nearly 170,000 deaths each year.[28]

An increase in the aging population and population growth potentially contribute to a continuous rise in CVD events. Older adults are more susceptible to CVD, which means the overall burden of these diseases may not diminish even if mortality rates improve.[32,33] Additionally, the disparities in different socio-economic regions within the UK also contributed to CVD prevalence. For example, areas with higher levels of deprivation tend to have higher rates of CVD.[34]

1.3.2 Economic burden

Cardiometabolic disease (CMD) imposes a significant economic burden, affecting direct and indirect medical costs and productivity.[35–37] People with T2DM or CVDs are likely to have more healthcare visits, and medication not only for treating the disease but also its complications, have a higher probability of being hospitalised, and require long-term/social/informal care compared to people with no disease.[35–37]

Studies in the UK have estimated the current and future economic burden of T2DM and CVD and highlighted the consistently growing cost of CMD. The current annual total cost associated with diabetes is approximately £23.7 billion, projected to increase to £39.8 billion by 2035/6 (Figure 1.4).[38] A current cost of illness study showed that direct medical costs of diabetes in 2021/22 are estimated at £10.7 billion with more than 80% of these costs being incurred by T2DM patients. Estimates for indirect costs associated with T2DM reached £3.3 billion.[39]

Currently, around 10% of the NHS budget is spent on treating diabetes and its complications. However, if current trends continue, this figure is forecasted to

rise to as much as 17% by 2035/36[40] Beyond direct healthcare costs, diabetes also imposes a broader economic impact through lost productivity and wider societal costs associated with managing the condition and its long-term complications.[38,39]

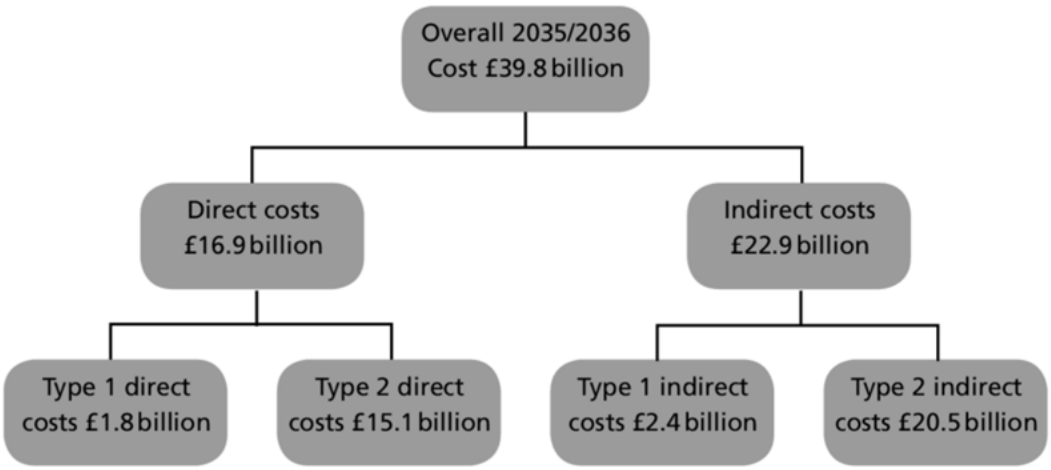


Figure 1.4 Estimated direct and indirect costs of diabetes in the UK 2035/36[38]

Similar to type 2 diabetes, cardiovascular disease (CVD) places a substantial and growing economic burden on both the NHS and wider society. Annual NHS healthcare costs amount to about £10 billion, while the overall economic impact, including long-term care, disability, informal care, and premature mortality is approximately £24 billion each year.[28] UK-wide data for 2021/22 further indicate a total societal cost of £29.0 billion, comprising £16.6 billion in direct healthcare spending and £12.4 billion in indirect costs.[41]

Efforts to address the disease and economic burden of CMD in the UK require comprehensive strategies which include improving cost-effective disease preventative strategies, ensuring healthcare capacity, improving health promotion and self-management education, and improving lifestyle behaviour. Strengthening these strategies would not only mitigate the impact on the UK healthcare system but also improve quality of life and financial capability for those with or at risk of CMD.

1.4 Prevention and management of cardiometabolic disease

1.4.1 Types of public health prevention

Public health prevention encompasses a spectrum of interventions aimed at reducing the risk of developing the disease and promoting well-being within populations. These are categorised into five levels, each with a distinct focus and approach.[42,43]

1. Primordial prevention: targeting the root causes of diseases by addressing social, economic, and environmental determinants of health. It aims to create conditions that prevent the development of risk factors. For example: the implementation of a sugar tax and the development of walking paths that promotes a healthier living environment.
2. Primary prevention: involves interventions that aim to prevent the onset of diseases, focus on reducing risk factors, and promote protective factors. This is crucial in reducing the incidence of the disease at the population level. For example: health education, vaccination programmes.
3. Secondary prevention: this focuses on early detection and intervention to halt the progression of diseases. Secondary prevention plays a vital role in reducing the impact of diseases on individual or population levels. For example: screening programmes.
4. Tertiary prevention: the intervention aims to improve the quality of life and reduce the complications of established disease. For example: diabetic foot care.
5. Quaternary prevention: focuses on avoiding unnecessary interventions and over-medicalisation. It includes ethical considerations, patient-centred care, and shared decision-making to prevent harm from excessive/unnecessary medical procedures. For example: the use of hormone replacement therapy that is susceptible to over-treatment.

The five levels of prevention provide a comprehensive framework for addressing diseases at different stages, from addressing the root cause to minimising harm from unnecessary interventions. This section, however, only focuses on preventative strategies as the thesis aims to develop a model that can accommodate the evaluation of early prevention (including prevention levels 1-3). Hence, this thesis will not cover clinical management and medication strategies for CMD.

1.4.2 Prevention guidance for cardiometabolic diseases

In the UK, cardiometabolic disease prevention and treatment guidelines are still under development.[44] However, there are prevention guidelines available for the prevention of T2DM and CVD.[45,46]

Current NICE guidelines primarily focus on identifying individuals at high risk of developing type 2 diabetes mellitus (T2DM). Risk assessment can be conducted using blood tests that measure fasting plasma glucose or HbA1c levels (to indicate prediabetes), validated risk assessment tools, and consideration of ethnicity-related risk factors. Key interventions for high-risk individuals include promoting healthy eating (a balanced intake of fat, sugar, and other nutrients), increasing physical activity (at least 150 minutes of moderate-intensity exercise per week), and maintaining a healthy weight. Prioritisation is given to those at the highest risk of T2DM for referral to intensive lifestyle-change programmes, particularly individuals with HbA1c levels between 44–47 mmol/mol (6.2–6.4%) or fasting plasma glucose levels between 6.5–6.9 mmol/L.[47,48]

For CVD prevention, current guidelines focus on modifiable risk factors, including a healthy balanced diet, physical activity, smoking cessation, and limited alcohol intake. NICE recommends using risk assessment tools to identify individuals at increased risk of CVD. The QRISK (Quantifying Risks in Individuals with Systolic Hypertension and Kindred) is a primary assessment tool for CVD risk, a widely used algorithm that calculates a patient's 10-year risk of experiencing a cardiovascular event (heart attack or stroke). Based on this assessment, personalised preventative strategies can be developed including blood pressure control (generally below 140/90 mmHg for people under 80 years old, lipid management (use of statin for lowering LDL cholesterol), and

management of other conditions such as addressing diabetes or obesity that contribute to CVD.[49,50]

Additional considerations are socioeconomic factors and the impact of social deprivation on CVD risk, and addressing these factors in building public health strategies.[34] Another important consideration is family history. Individuals with a family history of premature CVD are at significantly increased risk. In such cases, more proactive and aggressive risk management strategies may be needed. These may include earlier and more frequent cardiovascular screening, stricter control of blood pressure and lipid levels, lifestyle interventions targeting diet, physical activity, and smoking cessation, and, in some cases, earlier initiation of pharmacological treatments, even when traditional risk factors are only moderately elevated.[50]

1.5 The importance of policy models

1.5.1 Defining a Disease Policy Model

Given the significant and growing burden of T2DM and CVD on both individuals and the healthcare system, there is a pressing need for informed decision-making around prevention and management strategies. As healthcare resources are limited, policy-makers require robust tools to prioritise interventions that are not only clinically effective but also economically sustainable. This is where policy models become essential. They help synthesise clinical and economic evidence, enabling simulation of long-term outcomes and estimation of the value of interventions under varying scenarios. Understanding and applying such models is critical for guiding policies aimed at reducing the impact of these chronic conditions.

In a broad sense, a “model” is a simplified representation of reality. It is designed to provide understanding, analysis, and prediction from complex systems, behaviour, and phenomena.[51,52] A policy model in this context is “a model that can evaluate the effectiveness and cost-effectiveness of interventions and inform policy decisions.” In this thesis, the terms “policy model”, “decision model”, “health economic model” may be used

interchangeably. The distinction between them depends on the model's purpose and the context in which it is applied.

There are some examples of policy models developed. The Sheffield Type 1 Diabetes Policy Model is a mathematical simulation model that was specifically developed to evaluate health outcomes and the cost-effectiveness of treatments and interventions for T1DM patients. This model integrates various health states and complications associated with T1DM, such as hypoglycaemia, kidney disease, and CVD. This model used data from clinical trials and other studies to simulate how new health technologies (glucose monitors, insulin therapies), might influence the progression of the disease, the quality of life, and the economic costs involved.[53]

The Scottish Type 2 Diabetes Model is another example of a simulation model that is specifically developed for the T2DM population.[54] Like the Sheffield model[53], this model incorporates a range of clinical and economic inputs, including disease progression, complications, and the effectiveness and costs of treatments. This model simulates outcomes such as life expectancy, quality-adjusted life years (QALYs), and the costs associated with diabetic complications like CVDs, renal failure, and diabetic retinopathy.[54]

Furthermore, a model in the same Scotland context is The Scottish Cardiovascular Disease (CVD) Policy Model, a model that is designed to analyse and predict the impact of CVD, taking into account the unique demographic and healthcare system characteristics in Scotland's population.[55] The Scottish CVD model is a state transition model that simulates the progression of CVD within the population over time. Data was used from hospital records, national health surveys, and mortality statistics. The model can accommodate the estimation of economic costs as well.[55,56]

Initiated by the University of California San Francisco (UCSF) team, The Cardiovascular Disease (CVD) Policy Model is a model which applies a cohort-based approach that is designed to evaluate the health outcomes and economic impacts of various CVD events, including CHD and stroke, within specific populations. The model is intended to simulate the lifetime health outcomes,

healthcare costs, and cost-effectiveness of interventions aimed at preventing and managing CVDs. The model draws on various data sources, including epidemiological studies, clinical trials, health surveys, and national health statistics.[57]

Most of these policy models have been used in several important studies and evaluations to inform and strengthen the justification regarding the decision-making process. Models have not only been applied to evaluating various health technologies or interventions but have also been adopted in different settings and contexts.[58–60]

1.5.2 Common challenges and limitations of disease policy models

Despite the benefits of constructing a policy model, there are diverse methodological challenges and inherent limitations that can affect the model's accuracy and applicability. These include poor generalisability, low model quality, lack of transparency, as well as inconsistency in conclusions.[61–63]

One of the most significant challenges is the availability and quality of data. Data may be incomplete, lack detail, or be outdated, which can limit the generalisability of results to other settings.[63–68] This not only includes the clinical or epidemiological data but also cost data for economic models. Uncertainties may arise due to these limitations.

Another challenge is the simplification and assumptions that are often necessary to represent complex realities and make the model computationally feasible. These assumptions might include disease progression, treatment adherence, or patient behaviours.[64,65,69,70] While necessary, these simplifications can introduce bias and inaccuracies that may not fully capture the real-world view of an intervention. Additionally, the choice of time horizon in a model critically influences its outcomes.[71,72] A model that only considers short-term effects may overlook long-term benefits, potentially misrepresenting the value of preventive or chronic disease interventions. Determining the appropriate time horizon can be challenging too, especially when long-term data is unavailable, or current data is less credible to incorporate into a model.

Capturing the non-health outcomes and societal costs, as well as considering the dynamic complexity of health interventions and incorporating human behaviour, might still be a common limitation of policy models.[62,73] Moreover, since policy models often focus on epidemiological outcomes and cost-effectiveness, broader considerations such as equality, equity, and accessibility are often overlooked.[74,75]

1.6 The role of real-world evidence

1.6.1 Real-world data (RWD) and Real-world evidence (RWE)

Real-world data (RWD) refers to data collected from sources other than traditional randomised controlled trials (RCTs). RCTs have been widely viewed as a "gold standard" in health and medical research. However, there are several limitations in terms of the practicality of RCTs, such as intensive resources, restricted inclusion/exclusion criteria, controlled environment, and mostly conducted in a short study period resulting in limited generalisability.[76,77] Hence, the use of RWD can complement evidence generated from RCTs by offering more insight from real-world clinical practice.

RWD includes electronic health records (EHR), administrative databases, patient registries, claim data, and other sources that reflect real-life patient experiences and healthcare delivery. The evidence that is derived from real-world data is often called real-world evidence (RWE).[78]

RWE, which is generated from RWD provides valuable information, for instance: treatment patterns, medication effectiveness, comparative patient outcomes in authentic clinical settings, regulatory process, as well as broader clinical and health decision-making.[76,79] Moreover, with proper use of RWE, a comprehensive understanding of existing constraints outside a controlled environment can be achieved. [77,80]

Several types of RWD have been used extensively to support clinical and health decision making such as NHS Digital Health[81], Clinical Practice Research Datalink (CPRD)[82], Hospital Episodes Statistics (HES)[83], UK Biobank[84], Scottish Health Research Register (SHARE)[85], and more. These databases have been utilised for a

wide variety of research projects covering clinical, epidemiological, as well as cost-effectiveness studies.

In recent years the use of RWE has gained attention in health and medical research, including cardiometabolic disease. Razieh et al. (2022) explored the association of sociodemographic, lifestyle, environmental, and clinical factors with the risk of CVD across different ethnic groups.[86] A study by Buckland et al. (2023)[87] and Eriksen et al. (2018)[88] examined adherence to UK dietary guidelines and nutrient profiling with cardiometabolic risk markers, emphasising the importance of diet quality in managing cardiometabolic health.

As highlighted by Dobson & Prendergast (2022)[89], the UK's national registries, such as the UK TAVI registry for transcatheter aortic valve implantation (TAVI), have provided valuable real-world patient data to inform clinical practice. Furthermore, utilising CPRD data, Canoy et al. (2021)[90] assess the association between myocardial infarction (MI), stroke and diabetes with excess mortality. It was confirmed that other comorbidities are also strongly related to this excess mortality risk. Additionally, the study by Gulliford et al. (2020)[91] compares antibiotic prescribing records in two UK primary care EHR systems, highlighting the potential of combining CPRD GOLD and CPRD Aurum data for research purposes.

1.6.2 Utilising RWD for disease policy models: an opportunity

In health economics and policy models, RWD has become increasingly important due to its ability to provide a comprehensive understanding of real-world care settings, effectiveness, and value of healthcare interventions.[92] National bodies such as NICE UK utilise RWD to guide clinical decision-making and health technology assessment (HTA).[93]

As mentioned in the NICE strategy 2021-2026, RWD and RWE have potential to address the knowledge gap and engage further access to patient innovations. These gaps include the limited generalisability of RCTs, the underrepresentation of certain patient populations, and the lack of long-term outcome data. The NICE RWD framework has been developed to support these initiatives and offer guidance for identifying appropriate data sources to reduce uncertainties in

evidence generation and to strengthen clinical and economic recommendations. It also outlines best practice standards for conducting and reporting RWE studies, aiming to improve the quality, transparency, and policy relevance of the evidence produced.[93]

In addition, the pharmaceutical industry, regulatory agencies, and payers also recognise the value of RWD as a complementary source or approach that can work hand in hand with RCTs for establishing more robust evidence in clinical practice.[78,94] Current technological advancement and improved data governance have strengthened the potential use of real-world health data to benefit patient care and further health services, decision-making, and patient outcomes. [78]

In terms of practicality, RWD also offers the advantage of providing large sample sizes at a low cost, making it a practical and cost-effective resource for generating medical evidence.[77,80,94] For instance, the secondary use of electronic health records (EHRs), patient registries, and insurance claims or billing data enables researchers to examine patient outcomes in routine clinical settings. This not only enhances the relevance and generalisability of findings but also holds significant value for healthcare economic modelling, where understanding the real-world impact and cost-effectiveness of interventions is crucial for informing policy and resource allocation.[93,95]

While the use of RWD has great potential for improving understanding in routine settings, challenges and limitations remain to be addressed to generate robust results and improve evidence quality.[80,96]

Although various policy models have been developed to inform prevention and management of chronic diseases, there remains a need for modelling approaches that address the complexity of cardiometabolic disease (CMD) as an interconnected condition. Given the long-term nature of CMD progression and its broad population impact, modelling plays a critical role in supporting evidence-based decision-making, particularly for evaluating the health and economic implications of preventive strategies. When informed by real-world data (RWD), models can more accurately reflect routine care, diverse patient populations, and the cumulative burden of disease. This thesis seeks to contribute to ongoing

efforts by exploring how policy modelling, supported by current and locally relevant RWD, can enhance the relevance and utility of prevention-focused evaluations for CMD.

1.7 Conclusions

Cardiometabolic disease remains a public health concern in the UK, contributing to a significant burden on morbidity, mortality, and healthcare costs. Well-targeted and cost-effective preventative strategies are needed to improve cardiometabolic health in the population. Therefore, to understand or compare the costs and benefits of CMDs preventative strategies, developing a model can be beneficial to generate evidence by simulating and predicting health and economic outcomes. Incorporating RWD offers opportunities to increase representativeness and generalisability of the policy model.

By leveraging the use RWD, the model potentially provides a foundation for assessing the real-world impact of preventative strategies, optimising resource allocation, and informing evidence-based decision-making that could improve patient outcomes and healthcare efficiency.

Chapter 2 Research aims and objectives

2.1 Overview

Building on the rationale presented in Chapter 1, this chapter sets out the research objectives that guide the development of a policy model for cardiometabolic disease (CMD) prevention. Given the complex and interrelated nature of CMD conditions, and the potential of real-world data (RWD) to improve the relevance of policy models, there is a need for modelling approaches that reflect real-life care settings and long-term disease trajectories. This chapter outlines the aims of the study, which seeks to develop a real-world-data-informed model capable of evaluating preventative strategies across the CMD continuum.

RWD-informed modelling is especially valuable for preventative strategies, which involve complex disease pathways and long-time horizon. By capturing routine care patterns, such models can better estimate the long-term health and economic impact of early interventions, supporting evidence-based decision-making.

The following sections outline the specific research aims (Section 2.2) and the overall structure and layout of the thesis (Section 2.3).

2.2 Aims and objectives of the thesis

This thesis aims to develop a Cardiometabolic Disease (CMD) Policy Model, a framework that contributes to future health and economic analyses. It explores the opportunities and challenges associated with utilising UK primary care data, acknowledging both its potentials and limitations.

The specific objectives of this thesis are:

1. To critically review existing published cardiometabolic disease policy and health economic models.
2. To propose and conceptualise a Cardiometabolic Disease Policy Model
3. To construct and analyse a multi-state model using UK primary care data
4. To demonstrate and simulate the cost-effectiveness of preventive strategies using the Cardiometabolic Disease Policy Model

2.3 Thesis structure

Chapter 3 presents a systematic review of cardiometabolic disease (CMD) policy models, specifically those addressing type 2 diabetes (T2DM) and/or cardiovascular disease (CVD) prevention at population level. The review examines and critically appraises a range of modelling approaches found in the existing literature, focusing on model structures, data sources, and validation methods. It highlights the strengths and limitations of current models and provides foundational insights to inform the development of the conceptual model in the following chapter.

Chapter 4 introduces the conceptual model that forms the foundation for the development of the CMD policy model. This conceptual model represents the initial step in systematically capturing and communicating the contextual understanding of the problem, including disease progression, relevant evidence, and the rationale behind modelling choices and structure. The model is informed by findings from the systematic review (Chapter 3), current clinical guidelines, and expert input, ensuring that it reflects both the theoretical and practical dimensions of cardiometabolic disease prevention.

Chapter 5 describes the data preparation process using Clinical Practice Research Datalink (CPRD) Aurum. It summarises how the raw data were cleaned, processed, and transformed to be suitable for statistical analysis and model implementation. This includes steps taken to ensure data structure, covariates

standardisation, and overall cohort identification in accordance with the requirements of the modelling framework developed in earlier chapters.

Chapter 6 presents the detail of the cardiometabolic disease (CMD) Policy Model development, including the application of multi-state survival analysis. It simulates different model specifications (such as non-parametric, semi-parametric, and parametric approaches) with a particular focus on implementing a semi-Markov framework for parametric modelling. The chapter also discusses model diagnostics and evaluates the adequacy and performance of each modelling approach. This chapter shows how the developed model can support the long-term analysis and projections, which are essential for assessing the impact of public health preventative strategies.

Chapter 7 discusses the application of the CMD Policy Model to evaluate a set of hypothetical public health interventions, including dietary change initiatives and smoking cessation programmes. The purpose of these analyses is not to estimate the real-world effectiveness or cost-effectiveness of the interventions, but to demonstrate how the model can be used to simulate and compare outcomes across different policy scenarios. This chapter showcases the flexibility and practical utility of the CMD Policy Model in epidemiological and health economic evaluation.

Finally, **Chapter 8** concludes the thesis by summarising the key findings, methodological contributions, and policy implications of the research. It reflects on the research questions posed at the outset and evaluates how they were addressed through the development of the CMD Policy Model. The chapter also outlines areas for future research, including opportunities to enhance the model further and expand its application in supporting evidence-based decision-making for CMD prevention.

Chapter 3 Cardiometabolic diseases policy models: a systematic review

3.1 Overview

Chapter 3 covers the systematic review (SR) of the published cardiometabolic diseases policy model (T2DM and/or CVD). The SR is performed to identify current gaps in evidence and knowledge and shape the future direction of conceptualising a Cardiometabolic Disease Policy Model by consolidating and critically appraising multiple published articles.

3.2 Rationale for systematic review

As previously mentioned in Chapter 1, cardiometabolic disease (CMD), including T2DM and CVD are major contributors to morbidity and mortality, imposing substantial health and economic burden both for healthcare systems and society.[5,6] As these conditions are relatively preventable, there is a need for effective public health strategies and policies that can address the risk factors and manage the disease burden across the population.

Given the finite resources and competing needs, not all strategies or interventions can easily be offered or implemented. Resources are scarce, and consequently, the assessment and prioritisation of prevention initiatives should be carefully considered.[73,97] In addition to considerations around an intervention's effectiveness in health/clinical settings, 'value for money' needs to be assessed in order to decide how to allocate resources optimally for producing maximum benefit to society.[73] Thus, the assessment of a health intervention is required as a part of the decision-making process.

To facilitate this, modelling has become largely used to reflect and simulate the disease, intervention, and economic outcomes.[62,69] A model in the context of medical research is defined as "analytic methodology that accounts for events over time and across populations, that is based on data drawn from primary and/or secondary sources, and whose purpose is to estimate the effects of an intervention on valued health consequences and costs".[52] Through these models, policymakers can gain predictive insights into the effectiveness and cost-effectiveness of health interventions, making them indispensable in the planning and evaluation of health policies.

Modelling also transcends the complexities in RCTs design which has limitations in terms of resources, time and generalisability.[62] It can generate long-term and generalisable evidence, as well as overcome issues related to limited observable time.[52] It is, therefore, useful to assess chronic conditions like CMD, which are characterised by their presence of competing risks, complications, and long-term morbidity.

Several policy models have been developed, particularly for T2DM and CVD, including state transition models, discrete event simulations (DES), microsimulations, or other mathematical simulations.[98–105] All these models can facilitate the evaluation of various public health strategies. However, current models are often tailored to medium to high-risk patients, primarily evaluating clinical interventions, mostly on medication strategies instead of population-wide early preventative strategies, and predominantly focused on the summary of cost-effectiveness outcomes rather than providing a detailed modelling appraisal.[100,104,106–108]

Hence, there is an opportunity to summarise policy models capable of capturing the full continuum of cardiometabolic disease progression across the entire population, beyond high-risk groups or individuals with established diagnoses, and of jointly evaluating the two main CMDs (T2DM and CVD) and their interrelationships. In addition, this review will place particular emphasis on the modelling approaches themselves, as the suitability of a given modelling framework depends on the decision problem, disease complexity, and type of intervention being evaluated. Different modelling approaches offer distinct

strengths and limitations in representing long-term disease progression, intervention effects, and population-level outcomes, all of which are critical considerations for policy-relevant decision-making.

3.3 Methods

The preferred reporting items for systematic reviews and meta-analyses (PRISMA) guidelines were followed.[109] The review is registered in PROSPERO with registration number CRD42022354399.[110]

3.3.1 Eligibility criteria

A policy model in this SR is defined as any mathematical/simulation model/framework that can predict health outcomes, costs, and cost-effectiveness. The model can explain and evaluate preventative strategies which guide decision-makers. Dietary policies were chosen as an example to illustrate the application of these models. These include sugar taxes, pack labelling, and food reformulation, which are designed to create healthier environments and reduce CMD risk before metabolic disturbance occur.

This review included models that start with a general or low-risk population (i.e., those without clinically diagnosed CMD) to assess the impact of primordial preventative strategies before disease onset. Also, this review required models that are able to predict long-term or lifetime outcomes (≥10 years) since policy interventions often have delayed effects on population health. Furthermore, models focusing on specific subgroups (e.g., obese adults or hypertensive individuals) and those assessing primary prevention with medication, as our interest lies in regulatory and public health measures rather than clinical interventions are excluded.

Table 3.1 Eligibility criteria for included studies in the review

Inclusion	Exclusion
• Models starting with a general or low-risk population and without any CMDs (without clinical diagnoses of CVD/ T2DM—disease free) • The model predicts long-term/lifetime outcomes (>10 years) • Adult population (≥ 18 years) • Mathematical models that can accommodate both health and economic outcomes (cost-effectiveness evidence) • Only models which were assessing and evaluating primordial preventative strategies (restricted to regulations/policy for population dietary, limited to sugar/salt/sodium and fruit/vegetables public health policies) targeting the whole population or population-based prevention	• Clinical studies, cell and animal studies • Models starting with CMD and only including specific subgroups (obese adults, people with hypertension) • Models focussing on accuracy or cost-effectiveness of diagnostic tools, primary prevention with medication (i.e.: statin use) • Models reporting effectiveness only • Models that were published as presentations, abstracts, commentaries, letters, and review articles

3.3.2 Search strategy and study selection

A systematic search strategy was developed and run on 6th December 2022 (updated search on 31st May 2024) in MEDLINE (Ovid), EMBASE (Ovid), CINAHL, Google Scholar, and Open Grey with restricting the publication year from 1st January 2000 to May 2024, applying Medical Subject Heading (MesH). The search strategy is presented in the Appendix 1. To minimise the risk of excluding relevant articles, hand-searching the reference lists of previous systematic/literature reviews was performed using the snowball technique.[111] The search strategy has been developed with the support of a University of Glasgow subject librarian as well as thesis supervisors. Article management and duplicate removal were undertaken using Zotero®.

3.3.3 Data extraction

Data from fully eligible studies were extracted using a standardised matrix in a spreadsheet Microsoft Excel®. Items for data extraction include author/model name, year of publication, country, model type and structure, perspective, events, outcomes (clinical and economic), data sources, time horizon, validity, and sensitivity analysis. One reviewer (SP) performed data extraction, and double extraction[112] was done independently for 20% of the total number of included papers by supervisors. Disagreements were resolved by team discussion.

3.3.4 Quality Assessment

Three independent reviewers (SP, HF, YD) assessed the quality of decision models and economic evaluation studies using the Phillips et al. checklist.[69] If there were any disagreements, these were resolved by seeking advice from supervisors. Findings from this assessment are illustrated in the checklist tables and are also presented visually and in a narrative format.

3.4 Results

3.4.1 Selection process

The PRISMA flow diagram (Figure 3.1) visually depicts the article selection process. An initial search yielded 1109 records, which were reduced to 217 following the removal of duplicates and screening of titles and abstracts. A thorough full-text assessment of these 217 articles resulted in 32 studies that met the established inclusion criteria. A characteristics summary of these included articles is provided in Table 3.2.

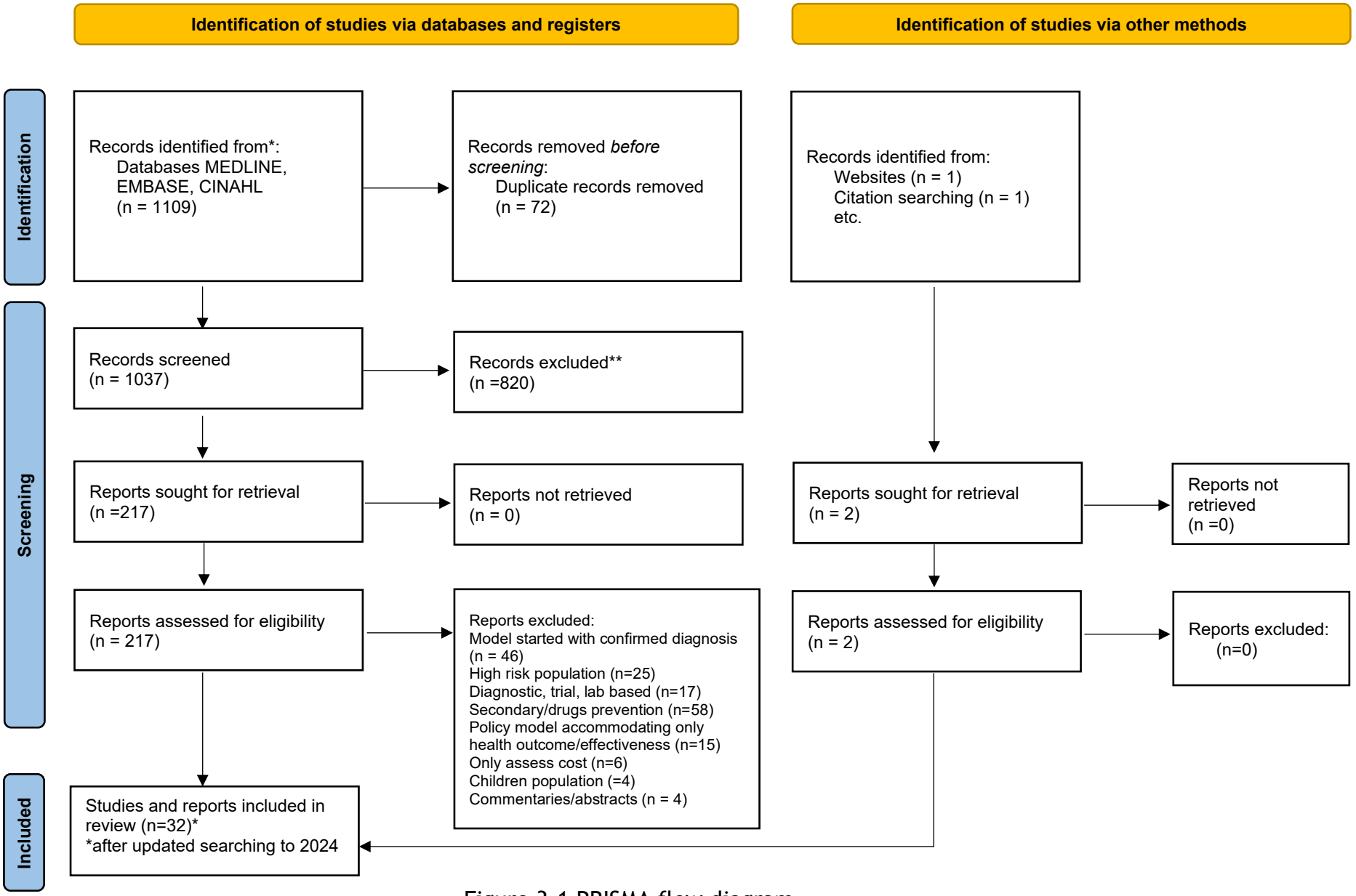


Figure 3.1 PRISMA flow diagram

Table 3.2 Description of included studies

No.	Author (year)	Country	Model's name	Policy assessment/ scenarios/evaluation	Perspective	Model types	Simulation level	Time horizon, cycle	Disease states/ measurement
1.	Moran et al. (2008)[113]	China	CHD Policy Model-China	Estimation and assessment of the CHD events based on to demographic changes	N/A	Markov	Cohort	30 years, annual	Free CHD, CHD, CHD death, non-CHD death.
2.	Moran et al. (2010 [114]	China	CHD Policy Model-China	Estimation of future risk factors on CHD and stroke	N/A	Markov	Cohort	20 years, annual	Free CHD, person with CHD, CHD death, non-CHD death.
3.	Bibbins-Domingo et al. (2010)[115]	US	CHD Policy Model	Estimation of benefits (rates, costs and cost-effectiveness) of salt reduction intervention	Healthcare	Markov	Cohort	10 years, annual	Free CHD, person with CHD, CHD death, non-CHD death.
4.	Wang et al. (2012)[116]	US	CHD Policy Model	Estimation of potential health impact and spending of a penny-per-ounce excise nationwide tax policy	Healthcare	Markov	Cohort	10 years, annual	Free CHD, person with CHD, CHD death, non-CHD death.
5.	Basu et al. (2013)[117]	US	-	Estimation of health effects and cost-effectiveness SNAP programme	Government	Microsimulation	Individual	10 years, annual	CVD mortality
6.	Konfino et al. (2013)[60]	Argentina	CVD Policy Model-Argentina	Assessment of the impact of sodium reduction policies	N/A	Markov	Cohort	10 years, annual	Free CHD, person with CHD, CHD

No.	Author (year)	Country	Model's name	Policy assessment/ scenarios/evaluation	Perspective	Model types	Simulation level	Time horizon, cycle	Disease states/ measurement
									death, non-CHD death.
7.	Basu et al. (2014)[118]	India	-	Estimation of the health effect on SSB taxation policy	Government	Microsimulation	Individual	10 years, annual	T2DM incidence
8.	Collins et al. (2014)[119]	England	CHD IMPACT Model	Cost-effectiveness analysis of four population health policies on salt intake	Health sector	Cell-based model	Cohort	N/R	CHD death
9.	Mason et al. (2014)[120]	Tunisia, Syria, Palestine, Turkey	CHD IMPACT Model	Cost-effectiveness analysis of population-based salt reduction policies in four Eastern Mediterranean countries	Public/ private sector, healthcare	Cell-based model	Cohort	N/R	CHD death
10.	Lewsey et al. (2015)[55]	Scotland	Scottish CVD Policy Model	The development of CVD policy model that predicts life expectancy and incorporating socioeconomic deprivation	-	Markov	Cohort	Potentially lifetime, annual	CVD event free, non-fatal CHD, non-fatal CBVD, fatal CVD, fatal non-CVD, fatal all cause
11.	Manyema et al. (2015)[121]	South Africa	-	Estimation of the effect of 20% SSB tax on the diabetes burden	Healthcare	Markov-multi state life table	Cohort	20 years, annual	BMI changes, diabetes

No.	Author (year)	Country	Model's name	Policy assessment/ scenarios/evaluation	Perspective	Model types	Simulation level	Time horizon, cycle	Disease states/ measurement
12.	Wilcox et al. (2015)[122]	Syria	CHD IMPACT Model	Cost-effectiveness analysis of salt reduction policies	Public/ private sector, healthcare	Cell-based model	Cohort	10 years, annual	CHD death
13.	Collins et al. (2015)[123]	England	-	Projection of 20 % of sugary drinks duty impact on disease events	Healthcare	Microsimulation	Individual	20 years, annual	Diabetes, stroke, CHD
14.	Lawson et al. (2016)[56]	Scotland	Scottish CVD Policy Model	The development of model for conducting economic evaluation	N/A	Markov	Cohort	Potentially lifetime, annual	CVD event free, non-fatal CHD, non-fatal CBVD, fatal CVD, fatal non-CVD, fatal all cause
15.	Sa´nchez-Romero et al. (2016)[58]	Mexico	CVD Policy Model-Mexico	Projection of SSB tax policies	N/A	Markov	Cohort	10 years, annual	No event, CVD event (MI, stroke, angina), death
16.	Wang et al., (2016)[124]	China	CVD Policy Model-China	Estimation of the effect of population-wide salt restriction in China	Healthcare system payer's	Markov	Cohort	Lifetime, annual	CVD free, acute CVD events, chronic CVD states, fatal CHD or stroke, non-CVD death
17.	Breeze et al. (2017)[125]	UK	SPHR Model	Cost effectiveness analysis of different interventions for type 2 diabetes prevention	NHS/PSS	Microsimulation	Individual	Lifetime, annual	Metabolic profile, no diabetes, diabetes, complications, CVD, cancer, osteo, depression, mortality

No.	Author (year)	Country	Model's name	Policy assessment/ scenarios/evaluation	Perspective	Model types	Simulation level	Time horizon, cycle	Disease states/ measurement
18.	Pandya et al. (2017)[126]	US	CVD-PREDICT	Description of the CVD model in detail; and performed model validation analyses	N/A	Microsimulation	Individual	Potentially lifetime	Disease free, CHD, stroke, death
19.	Mozaffarian et al. (2018)[127]	US	CVD-PREDICT	Estimation of the health impact and cost-effectiveness in SNAP program	Societal and government	Microsimulation	Individual	5-20 years and lifetime, annual	No CVD, acute CHD, chronic CHD, repeat MI or CVA, acute CVA, chronic CVA, CVD/non-CVD death
20.	Riveros et al. (2018)[128]	Brazil	Adaptation of Scottish CVD Policy Model	Calibration of Brazilian CVD model	N/A	Markov	Cohort	N/R	CVD event free, non-fatal CHD, non-fatal CBVD, fatal CVD, fatal non-CVD, fatal all cause
21.	Schönbach et al. (2018)[129]	Germany	DYNAMO-HIA	Estimation of health impact of tax on processed meat	N/A	Markov (extended to microsumulation)	Individual	10 years, annual	Prevalence in CHD, diabetes, cancer
22.	Huang et al. (2019)[130]	US	CHD IMPACT model	Estimation of the health impact and cost-effectiveness added sugar labelling on all packaged food and beverages	Healthcare and societal	Cell-based model	Cohort	20 years, annual	CHD incidence, stroke incidence, T2DM incidence
23.	Salgado et al. (2019)[131]	Argentina	CVD Policy Model-Argentina	The update Argentina CVD Policy Model	N/A	Markov	Cohort	Lifetime, annual	CVD free, acute and chronic CVD states,

No.	Author (year)	Country	Model's name	Policy assessment/ scenarios/evaluation	Perspective	Model types	Simulation level	Time horizon, cycle	Disease states/ measurement
									fatal CHD/ stroke, non-CVD death
24.	Wilde et al. (2019)[132]	US	CVD-PREDICT	Estimation of the health impact and cost-effectiveness of a national penny per-ounce SSBs tax	Healthcare and societal	Microsimulation	Individual	Lifetime, annual	Disease free, CHD, stroke, death
25.	Broeks et al. (2020)[133]	Netherlands	DYNAMO-HIA	Estimation of the effects of a tax on meat and a subsidy on fruit and vegetables (F&V) consumption	Societal	Markov	Cohort	30 years, annual	Healthy, disease, death
26.	Lee et al. (2020)[134]	US	CVD-PREDICT	Estimation of the health impact and cost-effectiveness of three SSBs tax designs	Healthcare, government, societal	Microsimulation	Individual	Lifetime, annual	Disease free, CHD, stroke, death
27.	Liu et al. (2020) [135]	US	CVD-PREDICT	Estimation of the health impact and cost-effectiveness of menu calorie labelling policy	Healthcare and societal	Microsimulation	Individual	Lifetime, annual	Disease free, CHD, stroke, death
28.	Salgado et al. (2020)[136]	Argentina	CVD Policy Model-Argentina	Estimation of the impact of reducing SSB consumption	N/A	Markov	Cohort	10 years, annual	CVD free, acute CVD events, chronic CVD states, fatal CHD or stroke, non-CVD death

No.	Author (year)	Country	Model's name	Policy assessment/ scenarios/evaluation	Perspective	Model types	Simulation level	Time horizon, cycle	Disease states/ measurement
29.	Dehmer et al. (2020)[137]	US	-	Evaluate prospective CVD related sodium reduction targets	Healthcare	Microsimulation	Individual	10 years, annual	Disease free, hypertension, CVD, post-CVD, death
30.	Shangguan et al. (2021)[138]	US	CVD-PREDICT	Assessment of the effect of sugar reformulation policy	Healthcare and societal	Microsimulation	Individual	Lifetime, annual	Sugar intake, acute CVD, diabetes, chronic CVD, CVD or non-CVD death
31.	Thomas et al. (2022)[139]	England	SPHR model	Estimation of health benefits, costs, and equity impact of food advertising across London transport network	NHS/PSS	Microsimulation	Cohort	Lifetime	Metabolic profile, no diabetes, diabetes, complications, CVD, cancer, osteo, depression, mortality
32.	Lou et al. (2023)[140]	US	CVD Policy Model	Impact assessment of implementing SSB taxes and FV subsidies on long-term CVD outcomes and healthcare costs	Societal	Microsimulation	Individual	10 years, annual	Healthy, CHD, stroke, both CHD and stroke, CVD-related death, and non-CVD-related death

Notes: CHD: coronary heart disease, CVA: cerebral vascular accident, CVD: cardiovascular disease, DM: diabetes mellitus, F/V: food and vegetables, MI: myocardial infarction, NHS/PSS: national health services/personal social services, SSB: sugar-sweetened beverage, T2DM: type 2 diabetes

3.4.2 Description of included studies

From 32 articles retrieved, there is a diverse range of geographical study locations, including the US (n=12)[115–117,126,127,130,132,134,135,137,138,140], UK (n=6)[55,56,119,123,125,139], Netherlands (n=1)[133], Germany (n=1)[129], Latin America (n=5)[58,60,128,131,136], South Africa (n=1)[121], India (n=1)[118], China (n=3)[113,114,124], Eastern Mediterranean (n=2)[120,122].

Policy models are primarily characterised as computational simulations that utilise mathematical frameworks, such as Markov models, cell-based models, and microsimulation techniques to project population-level outcomes, including mortality, morbidity, disease burden, and associated economic costs. These models frequently quantify the financial implications of health-related outcomes and evaluate the effects of policy interventions on both health and economic dimensions. While some studies explicitly articulate the concept of a "policy model," others implicitly adopt this framework by employing decision-analytic models for economic evaluations or health outcome projections, thereby assessing the impact of interventions at a population level. This approach underscores the integration of methodological rigour and policy relevance in addressing complex public health challenges.[117–119,123]

Several studies emphasised the importance of rigorous model development, calibration, and validation. Studies by Lewsey et al. (2015)[55], Lawson et al. (2016)[56], Pandya et al. (2017)[126], and Breeze et al. (2017)[125] described their policy models' framework that allows evaluation of further primary preventions/interventions.

All included policy models met the eligibility criteria by demonstrating the capacity to incorporate both epidemiological and economic parameters. However, the scope and depth of analysis varied significantly across studies. Some models focused exclusively on clinical or health outcomes, such as CVD mortality or T2DM incidence, while others emphasised cost and outcome estimations or conducted comprehensive economic evaluations, including cost-effectiveness analyses using metrics such as the incremental cost-effectiveness ratio (ICER). The dietary policies examined encompassed a diverse range of

interventions, including sugar taxes, salt reduction initiatives, and food labelling strategies. Importantly, this systematic review prioritises the methodological aspects of model structure and application rather than the efficacy or effectiveness of specific public health interventions.

There are a few reasons why this review mostly retrieved CVD models rather than T2DM models, despite both diseases being major concerns globally—particularly under a cardiometabolic umbrella. This review includes only models that started with disease-free or healthy populations, many T2DM models may choose to start with pre-diabetic or high-risk states when the interventions might be more actionable.[100,106,141] The strategies for managing CVD can often be implemented at various system levels (community, healthcare system, policy), while diabetes prevention and management often require more individualised approaches, also the complex presence of complications in diabetes patients makes the model more focused on treatment evaluation.[19,70] The other possible reason is CVD is linked with many other health conditions, such as diabetes and obesity, or vice versa. Effective CVD intervention that is represented in the CVD model might also indirectly influence diabetes management policies, given the overlap in risk factors like diet and exercise.[99,101]

3.4.3 Modelling types and structure

Types of mathematical model

Cohort models (Markov) have been the predominant approach in this review (47%) [55,58,60,113–116,121,124,129,131,133,136] and microsimulations have been extensively performed in recent years (40%).[117,118,123,125–127,132,134,137–140] Additionally, a smaller proportion of studies (13%) employed simpler forms of microsimulation models, such as cell-based models.[119,120,122,130] The models typically begin with a ‘disease-free’ or ‘healthy’ state, progressing through disease states and culminating in death. They commonly employ an annual cycle and adopt a long-term time horizon (>10 years or lifetime), enabling the quantification of health outcomes, benefits, and associated costs over extended periods.

Mathematical model can be broadly distinguished according to whether they simulate populations at the cohort level or individuals at the patient level, as well as by the underlying model structure used to represent disease progression.

Cohort models are typically implemented as state-transition models, in which proportions of a population move between predefined health states over discrete time cycles. These state-transition cohort models may take the form of Markov, semi-Markov, or other related structures, and generally rely on aggregated data to estimate transition probabilities.[62]

In contrast, individual patient simulation models, commonly referred to as microsimulation models, simulate disease trajectories at the level of individual patients rather than aggregated cohorts. Microsimulation models may also be implemented using state-transition frameworks, including Markov or semi-Markov formulations, but transitions occur at the individual level and can depend on patient-specific characteristics, prior events, and time-varying risk factors. This structure allows microsimulation models to represent individual heterogeneity, complex interactions, and dynamic disease pathways more flexibly than cohort-based approaches.[52,62,142]

The choice between cohort-based state-transition models and individual patient simulation (microsimulation) models should be guided by the decision problem, disease complexity, intervention characteristics, and data availability, rather than by methodological preference alone. Simpler cohort models are often sufficient when disease progression can be adequately represented by a limited number of mutually exclusive health states, transition risks are relatively stable, and outcomes of interest depend primarily on current health status rather than prior history.[143,144] In such contexts, cohort models offer advantages in transparency, ease of validation, lower data requirements, and reduced computational burden, which are important considerations for policy-facing analyses.[144,145]

Microsimulation models are more appropriate when the decision problem requires explicit representation of individual heterogeneity, history-dependent risks, competing events, or time-varying covariates. These features are particularly relevant for chronic and multifactorial conditions where disease progression and intervention effects depend on accumulated risk exposure, comorbidities, or prior events. Microsimulation also enables the evaluation of interventions that target specific subgroups or operate through multiple

interacting pathways. However, these benefits come at the cost of increased model complexity, higher computational demands, and a reliance on detailed individual-level data, which may not always be available or robust.[144,145]

Importantly, increasing model complexity does not automatically lead to more accurate or policy-relevant results. Overly complex models may obscure key assumptions, hinder validation, and introduce additional uncertainty if data inputs are weak or poorly characterised. Methodological guidance therefore emphasises the principle of parsimony, whereby the simplest model capable of addressing the decision problem should be preferred.[145–148]

In many applications, simpler and more complex models can produce similar cost-effectiveness conclusions when they are appropriately specified and parameterised. However, differences may arise when individual heterogeneity, non-linear risk accumulation, or history-dependent processes materially influence costs or outcomes. In such cases, it may alter incremental cost-effectiveness estimates and, potentially, policy conclusions. Consequently, model choice should be justified based on whether additional complexity is expected to meaningfully affect decision-relevant outcomes, rather than on technical sophistication alone.[144,147,148]

Overall, the selection of modelling approach should be driven by the alignment between the model structure and the underlying clinical and policy questions, balanced against considerations of transparency, data availability, uncertainty, and feasibility.

From this systematic review, there are feature details on the major policy model applied. Each model has its strengths, limitations and potential applicability. Table .3 presents a comparison between these policy models.

Table 3.3 Key differences between cohort and microsimulation models

Model features	Cohort model	Microsimulation model
Structure	States and transitions	Individual-level simulation
Memory	Memoryless (Markovian property)	History-dependent
Granularity	Simplified representation	Detailed representation
Modelling Flexibility	Fixed time intervals, limited interactions	Flexible time intervals complex interactions Captures individual heterogeneity, complex disease modelling, dynamic risk factors
Time	Discrete time steps (e.g., annual), constant transition probability	Continuous or discrete time steps, detailed event modelling
Data Requirements	Aggregated data	Detailed individual-level data.
Limitations	Oversimplifies disease progression, less exploring patient heterogeneity, Markov assumption, lower computational demand	Computationally intensive, complex model development.

DYNAMO-HIA (Dynamic Modelling for Health Impact Assessment) is a model that quantifies policies' impact on health determinants. It employs a Markov-based modelling approach, allowing for the simulation of a real-life population by explicitly considering risk factor states.[129,133] DYNAMO-HIA focuses on assessing the health impacts of policies on non-communicable diseases (NCDs), including CVD and diabetes. Its strengths lie in its comprehensive analysis, though its complexity and substantial data requirements can pose implementation challenges.[129,133]

Meanwhile, the CVD Policy Model[58,60,136,140], CHD Policy Model[113–116], and Scottish Policy Model[55,56] evaluate cardiovascular disease interventions at the population level using a state-transition model. It is robust for evaluating population-level interventions but can be complex to adapt to new populations or interventions.

The School of Public Health Research (SPHR) University of Sheffield model has applied a state-transition approach as well.[125,139] The SPHR Diabetes Model is a predictive tool that calculates the risk of developing type 2 diabetes (T2DM). It utilises a range of demographic, clinical, and lifestyle factors to generate personalised risk assessments, aiding in the prevention and management of diabetes. The SPHR Diabetes Model models the impact of diabetes prevention and intervention strategies at the population level using a system dynamics approach, with strengths in assessing diabetes-specific interventions but limitations due to complexity and data requirements.[125]

In addition, CVD-PREDICT (Cardiovascular Disease Policy Model for Risk, Events, Detection, Interventions, Costs, and Trends) also applied a microsimulation model to assess public health prevention programmes such as sugar-sweetened beverages (SSB) tax related diseases or other dietary policies.[126,127,132,134,135,138] IMPACT study employed a cell-based policy model, a subtype of compartmental, spreadsheet-based microsimulation, which produces aggregate estimates of population dynamics over time, in this case, focusing on life-years and mortality related to CHD. This modeling approach has since been adapted to other NCDs.[119,120,122]

Those models included common risk factors and baseline parameters such as age, sex, body mass index (BMI), systolic blood pressure (SBP), low-density lipoprotein (LDL)-cholesterol, high-density lipoprotein (HDL)-cholesterol, glycated haemoglobin (HbA1c), smoking and alcohol status, and other related factors. The structure of the model depends on the policy model itself, and most of them focus on a single disease (CVD or T2DM) or assign a CVD/T2DM state as a risk factor or comorbid state. The risk factors included are illustrated below:

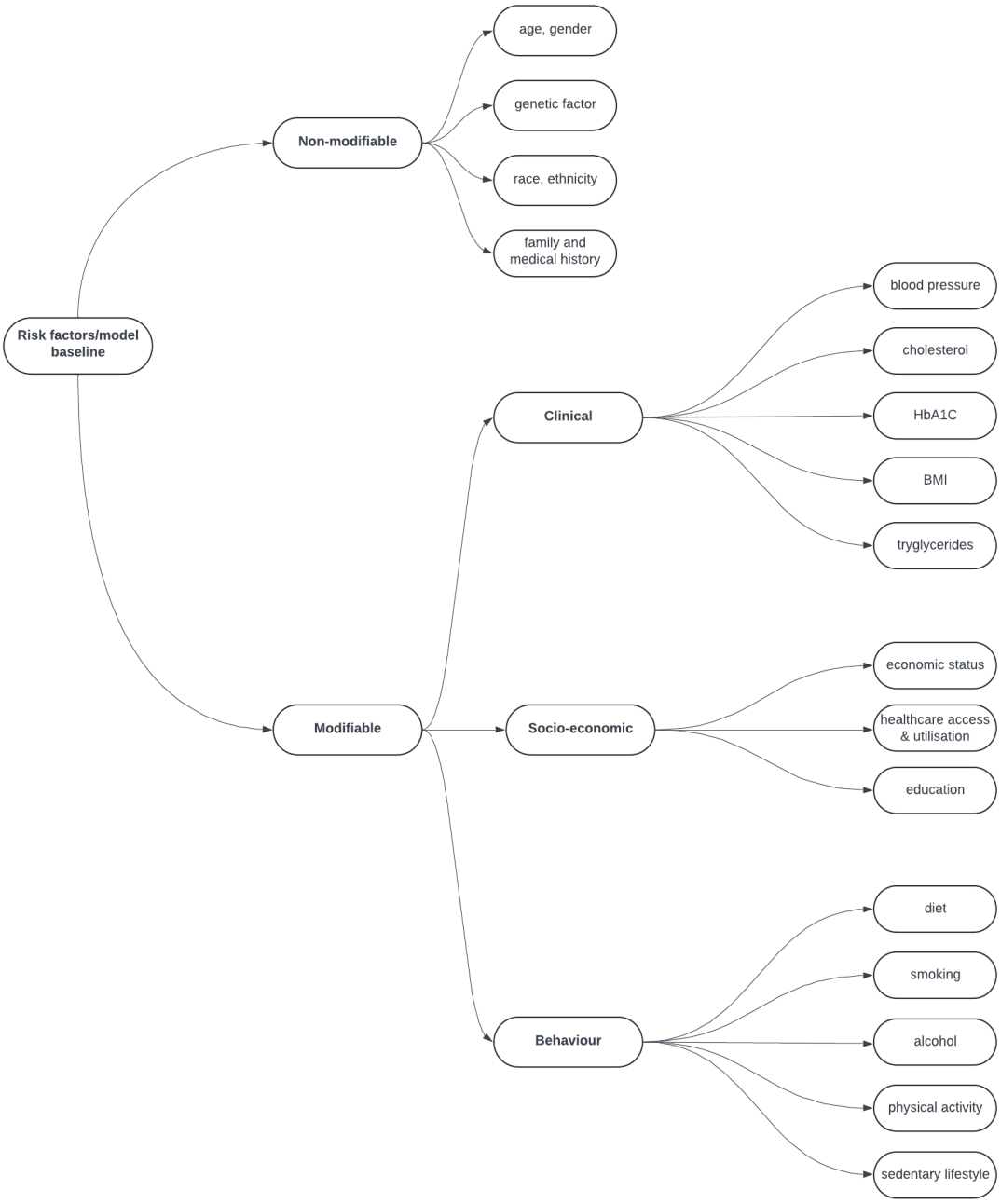


Figure 3.2 Summary of risk factors included in published policy models

Table 3.4 Comparison between cardiometabolic disease prevention policy models*

	DYNAMO-HIA	CVD Policy Model	CHD Policy model	IMPACT CHD	CVD-PREDICT	Scottish Policy model	SPHR Diabetes Model
Scope	NCDs (non-communicable diseases) including CVD, diabetes, and risk factors	CVD and related risk factors, focusing on prevention and treatment strategies	CVD and related risk factors, focusing on prevention and treatment strategies	CHD and CVD interventions, evaluating their effectiveness	CVD with a focus on prediction and risk stratification for better preventive measures	Public health with a specific focus on CVD and associated risk factors in Scotland	Diabetes and related risk factors, focusing on prevention, management, and health outcomes
Applicability	Primarily European countries, but adaptable globally	Primarily used in the US	Primarily used in the US	Applicable globally with regional adaptations	Applicable globally, with a focus on predictive analytics	Primarily used in Scotland	Primarily used in the UK
Data sources	European health surveys, epidemiological studies, and literature	National health surveys, clinical trials, epidemiological studies	National health surveys, clinical trials, epidemiological studies	National health surveys, clinical trials, epidemiological studies	National health surveys, clinical trials, epidemiological studies	Scottish health surveys, hospital records, national statistics	National health surveys, clinical trials, epidemiological studies

	DYNAMO-HIA	CVD Policy Model	CHD Policy model	IMPACT CHD	CVD-PREDICT	Scottish Policy model	SPHR Diabetes Model
Outcome of interests	Estimates incidence, prevalence, mortality, QALY health impact, under various policy scenarios	Estimates incidence, prevalence, mortality, and healthcare costs, cost - effectiveness	Estimates incidence, prevalence, mortality, QALY, health disparities healthcare costs of CHD and stroke	Estimates incidence, mortality, hospital admissions, cost-effectiveness	Estimates incidence, risk prediction, mortality, and health care costs, health outcomes, cost -effectiveness	- Estimates incidence, mortality, hospital admissions, QALE, cost-effectiveness	- Estimates incidence, prevalence, mortality, QALY, cost-effectiveness
Key strengths	Comprehensive modelling of individual and population-level effects; integration of multiple risk factors and interventions for a nuanced analysis across health outcomes	Robust framework for evaluating interventions at a population level; flexible to include various types of interventions; extensive validation with US data	Extensive validation with US data; comprehensive risk factor integration	Comprehensive evaluation of interventions; focus on real-world applicability; extensive data sources	High granularity of individual risk prediction; ability to incorporate large datasets and update predictions with real-time data.	Robust dataset specific to Scotland; focus on real-world applicability and policy impact; capable of addressing health inequalities and informing equitable policy decisions	Focus on diabetes-specific interventions and outcomes; ability to assess a wide range of potential interventions and their population-level impacts.
Key weaknesses	Complexity in adapting to non-European contexts -	Can be complex to adapt to new populations or to integrate	Requires extensive and high-quality data for accurate	May not account for all complex interactions	Requires access to high-quality, comprehensive health records;	Limited to the Scottish population, which may limit	Complexity and data requirements can limit

	DYNAMO-HIA	CVD Policy Model	CHD Policy model	IMPACT CHD	CVD-PREDICT	Scottish Policy model	SPHR Diabetes Model
	Requires extensive data input	novel interventions without substantial effort and data	projections; complexity of model may limit its accessibility for non-specialists	between risk factors and interventions; data limitations can affect accuracy	model accuracy can be affected by missing or inaccurate data	generalisability to other regions; data limitations outside of Scotland may affect model accuracy	accessibility for some users; relies on accurate input data for precise predictions

*Only models from the review that were used repeatedly for evaluating various strategies.

DYNAMO HIA=*Dynamic Modelling for Health Impact Assessment*, CVD-PREDICT= *Cardiovascular Disease Policy Model for Risk, Events, Detection, Interventions, Costs, and Trends*, SPHR= *School of Public Health Research*

Overall, these models use various types such as microsimulation, state-transition, compartmental, and system dynamics to support their specific purposes and applications. They require robust data sources like national health surveys and electronic health records for accurate predictions and assessments. While primarily used to inform policy decisions and guide public health strategies, these models vary in adaptability to different aims and health outcomes.

Model structure: examples

These are examples of model structures/frameworks from several studies included in this systematic review:

1. **DYNAMO-HIA**[129,133]

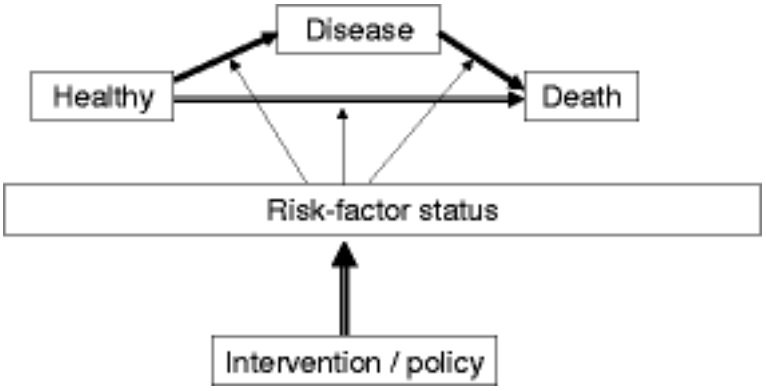


Figure 3.3 DYNAMO-HIA model structure

The model aims to estimate the long-term health effects of interventions and policy changes by altering the risk factor status of individuals, which subsequently influences their probabilities of developing diseases or experiencing mortality. This model employs health states to represent the various conditions and stages of health that individuals may experience over time, effectively capturing the progression and dynamics of NCDs with a particular emphasis on CVD and associated risk factors. By modelling the transitions between health states, the model facilitates a nuanced understanding of how preventive measures can affect public health outcomes over time.

2. CVD-PREDICT[126]

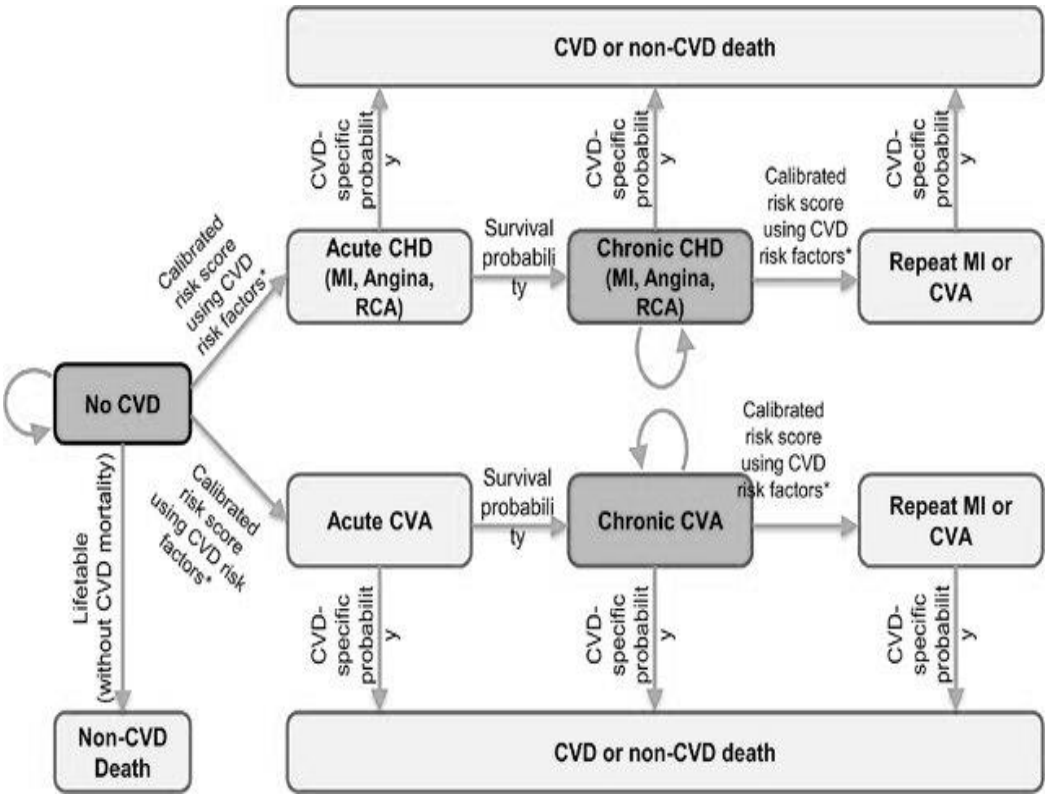


Figure 3.4 CVD-PREDICT model structure

Transitions between health states in the CVD-PREDICT model are determined by a calibrated risk score that incorporates multiple demographic, clinical, and lifestyle variables. The model uses these risk scores to simulate an individual's likelihood of progressing through different health states, capturing the dynamic nature of CVD over time, categorising individuals into health states such as "healthy," "at risk," "CVD event," "post-CVD event," and "death".

3. SPHR Diabetes Policy Model[125]

The SPHR Diabetes Model is a simulation-based framework that categorises individuals into discrete health states to represent different stages of diabetes progression, beginning with a normoglycemic state and potentially advancing to pre-diabetes, diabetes, diabetes complications, and ultimately, mortality.

For instance, individuals in a normal glycaemic state may progress to pre-diabetes if exposed to risk factors like obesity and lack of physical activity. Similarly, those with pre-diabetes may transition to diabetes with worsening glucose control and continued exposure to detrimental lifestyle factors. Poor glycaemic management can further lead to complications, including neuropathy, nephropathy, retinopathy, and cardiovascular diseases, significantly impacting quality of life and increasing mortality risk.[125]

In general, the types of modelling approaches mentioned in this chapter are suitable to represent chronic disease progression such as CMD, when the diseases are preventable, persist for a long duration but often require ongoing disease management. Brennan et al. (2006)[142] also provide a comprehensive taxonomy of model structure for supporting justification of economic evaluation studies.

Finally, the choice between modelling types largely depends on several factors, including research questions/objectives, disease characteristics, intervention being evaluated, data availability, and the desired level of model complexity. In some cases, combining or extending different modelling types can also provide a more comprehensive analysis.

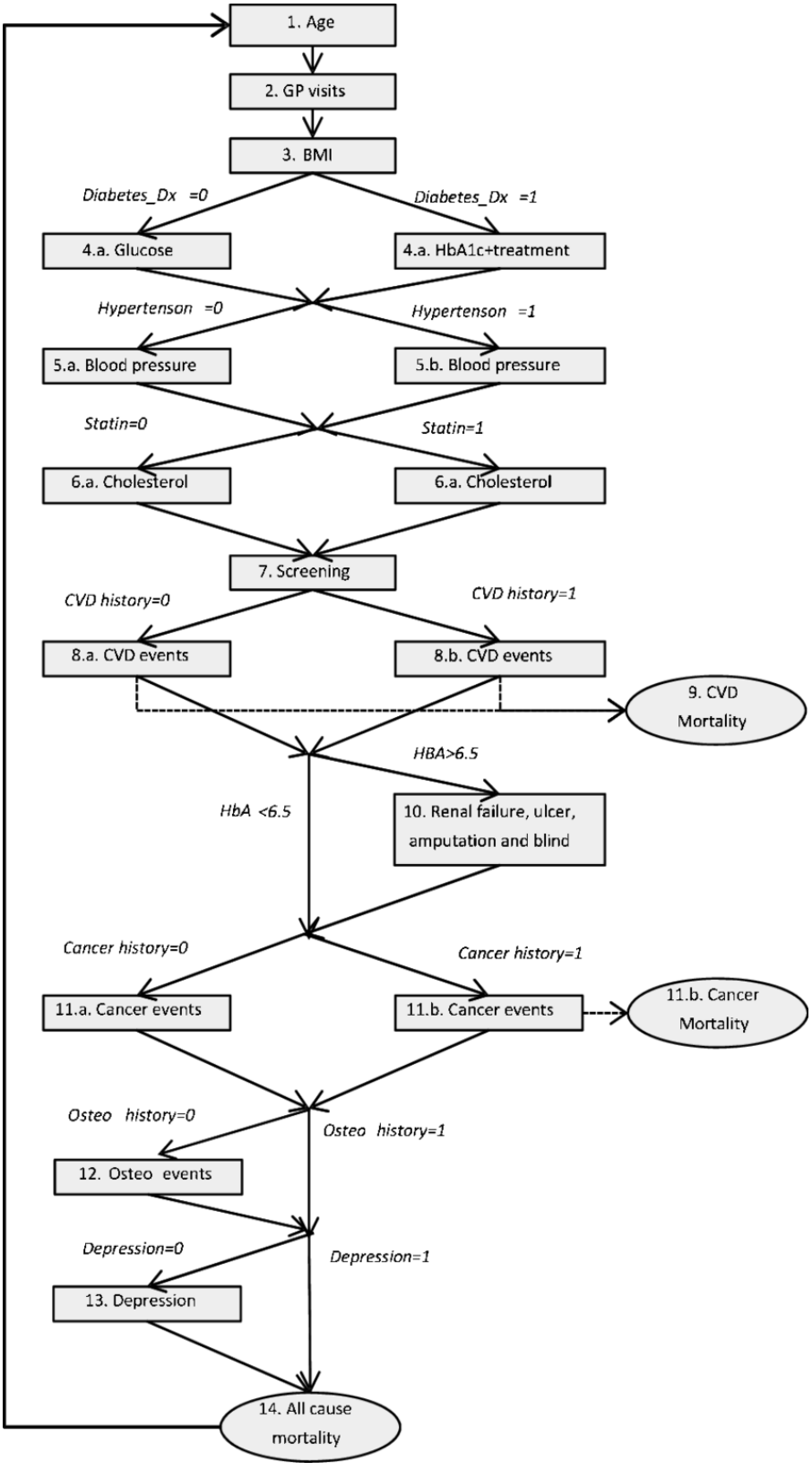


Figure 3.5 SPHR Diabetes model structure

3.4.4 Costs and Outcomes

Costs incorporated in the models were based on policy questions and perspectives defined. Direct medical costs include expenses incurred due to disease conditions, such as hospitalisation, healthcare provider services (consultations, treatments), medication use, and laboratory/diagnostic costs. Indirect costs are often associated with productivity loss due to disability or illness. Programme costs in several studies represent expenses related to the implementation of policy interventions.

About 72% of studies included direct medical costs in the analysis.[58,114–116,119–125,127,129,132–140] Indirect costs were included in some studies (38%) [119,125,127,129,132–138], depending on their analysis perspective. Most of these studies focused on the impact of dietary interventions, and 56% of studies reported their programme costs. [116–120,122–125,127,129,132–134,139,149–151] From studies retrieved in this review, monetary values were mostly reported in USD and international dollars.

Regarding the outcome measures, the majority of studies (90%) estimated disease incidence or prevalence [113,114,116,117,119,55,122,123,56,58,124-126,129-131,133-140], while 47% reported generic health outcomes such as Quality-Adjusted Life Years (QALYs) or Disability-Adjusted Life Years (DALYs). [114,115,117,121,123-125,127,130,132,134,135,139] (Table 3.5). While most studies estimated QALYs by assigning utility weights to different health states, the methodology for deriving these utility values is often poorly described. In many cases, utility values are sourced from previously published studies, but the papers do not provide detailed explanation of the methods used to derive these values-such as whether they were obtained through direct methods (e.g.: time trade-off, standard gamble) or indirect methods (e.g., EQ-5D, SF-6D).

For studies in which the decision model is used to conduct or illustrate a full economic evaluation, incremental analysis, specifically the Incremental Cost-Effectiveness Ratio (ICER) and Incremental Net Benefit (INB) were commonly applied (28%).[122,125,127,132,135,138,139] A small number of studies analysed how costs and benefits were distributed across demographic groups.[55,139]

Regarding discounting, all studies incorporated a discount rate for both costs and outcomes or only one of them. Two studies applied undiscounted simulations in their main analysis but deterministically incorporated discount rates into their scenario and sensitivity analyses.[121,152] (Table 3.5) The discount rates ranged between 0%-5% and the justifications for using those rates were mostly based on local guidelines.

All models reported their main data sources incorporated in the model. These included data from the published literature, meta-analysis evidence, local statistical data, government data, survey data, or transferability of data from other countries.

3.4.5 Model validation

Model validity is the evaluation of whether the model demonstrates proper representation of the system and whether its results could serve as a solid basis for decision-making.[153,154] Validating a model is essential in economic evaluation to ensure the reliability, accuracy, and credibility of the models used. It enhances transparency, supports evidence-based decision-making, and helps identify and address model limitations.

In this review, five types of model validation were assessed: face validity (checking if the model's structure, inputs, and outputs logically reflect known behaviours and outcomes of certain diseases), internal validity (whether the algorithms and relationships within the model correctly simulate the progression), cross-validity (confirming that the model’s findings are consistent across different samples or populations within the same study), external validity (assesses the generalisability of the model to other settings, populations, or times), and predictive validity (model or evaluation accurately predicts real-world outcomes).

Table 3.5 Cost and outcomes measured

Authors name	Costs			Outcomes#				Discounting*
	Direct healthcare costs	Indirect costs	Programme/ implementation costs	Disease cases/ event	LY/LE	QALY/ DALY	ACER/ICER/ INB/NMB	
Moran et al. (2008)[113]	√	-	-	√	-	√	-	√
Moran et al. (2010)[114]	N/R	N/R	N/R	√	-	-	-	N/R
Bibbins-Domingo et al. (2010)[115]	√	-	-	-	-	√	-	√
Wang et al. (2012)[116]	√	-	√	√	-	-	-	√
Basu et al. (2013)[117]	-	-	√	√	-	√	-	√
Konfino et al. (2013)[60]	N/R	N/R	N/R	√	-	-	-	N/R
Basu et al. (2014)[118]	-	-	√	√	-	-	-	N/R
Collins et al. (2014)[119]	√	√	√	√	√	-	-	√
Mason et al. (2014)[120]	√	-	√	√	√	-	-	√
Lewsey et al. (2015)[55]	N/R	N/R	N/R	√	√	-	-	N/R
Manyema et al. (2015)[121]	√	-	-	√	√	√	-	√
Wilcox et al. (2015)[122]	√	-	√	√	√	-	√	√
Collins et al. (2015)[123]	√	-	√	√	-	√	-	-
Lawson et al. (2016)[56]	N/R	N/R	N/R	√	√	-	-	√
Sa´nchez-Romero et al. (2016)[58]	√	-	√	√	-	-	-	N/R
Wang et al., (2016)[124]	√	-	√	√	-	√	-	√
Breeze et al. (2017)[125]	√	√	√	√	√	√	√	√
Pandya et al. (2017)[126]	N/R	N/R	N/R	√	-	-	-	N/R
Mozaffarian et al. (2018)[127]	√	√	√	-	-	√	√	√

Authors name	Costs			Outcomes#				Discounting*
	Direct healthcare costs	Indirect costs	Programme/ implementation costs	Disease cases/ event	LY/LE	QALY/ DALY	ACER/ICER/ INB/NMB	
Riveros et al. (2018)[128]	N/R	N/R	N/R	√	√	-	-	N/R
Schönbach et al. (2018)[129]	√	√	√	√	-	-	-	N/R
Huang et al. (2019)[130]	√	√	-	√	-	√	√	√
Salgado et al. (2019)[131]	N/R	N/R	N/R	√	-	-	-	N/R
Wilde et al. (2019)[132]	√	√	√	-	-	√	√	√
Broeks et al. (2020)[133]	√	√	√	√	-	-	-	√
Lee et al. (2020)[134]	√	√	√	√	-	√	-	√
Liu et al. (2020) [135]	√	√	√	√	-	√	√	√
Salgado et al. (2020)[136]	√	√	-	√	-	-	-	N/R
Dehmer et al. (2020)[137]	√	√	-	√	-	-	-	-
Shangguan et al. (2021)[138]	√	√	-	√	-	√	√	√
Thomas et al. (2022)[139]	√	-	√	√	-	√	√	√
Lou et al. (2023)[140]	√	-	√	√	-	√	√	√

LY = life years, LE = life expectancy, QALY = quality-adjusted life years, DALY = disability-adjusted life years, ACER = Average cost-effectiveness ratio, ICER = incremental cost-effectiveness ratio, INB = incremental net benefit, NMB = net monetary benefit. # Studies might have more than outcomes measured, *Discounting can be only cost or outcome or both, or part of scenario analysis

All studies conducted assessments of face and internal validity. Cross-validity was mentioned in one study[125]; however, the methodologies employed for testing were often unclear. External validation was performed in 53% of studies, indicating some efforts to evaluate the generalisability of models. [113,116,119,120,55,56,124,126,127,136,131,135,134,132] None of the included articles reported predictive validation. This omission is likely due to the fact that the studies relied on external validation procedures, which they considered sufficient for evaluating the predictive performance of the models. A summary of model validation performed is presented in Table 3.6.

3.4.6 Model uncertainty

Uncertainty is an important part of health economics and policy models. It arises from various sources and can significantly impact the results or conclusion of an analysis. Sensitivity analyses (SA) are commonly employed to explore these uncertainties, either deterministically or probabilistically.[155] Deterministic sensitivity analyses (DSA), such as one-way or scenario analyses, systematically examine the impact of uncertainty by incorporating plausible alternative values or scenarios. In contrast, probabilistic sensitivity analyses (PSA) assign probability distributions to uncertain parameters and performs multiple model simulations to produce a distribution of outcomes.

All studies included in this review reported conducting sensitivity analyses as part of their modelling process (Table 3.6). Of these, 50% (16 studies) performed both DSA and PSA, while the remainder employed only one type of sensitivity analysis. [55,56,58,58,116,117,119,123,125,127,130,132–134,136,138,139] .

Table 3.6 Validation test and uncertainty analysis

Authors	Validation test					Uncertainty analysis	
	Face validity	Internal validity	Cross-validity	External validity	Predictive validity	Deterministic SA (DSA)	Probabilistic SA (PSA)
Moran et al. (2008)[113]	√	√	-	√	-	√	-
Moran et al. (2010)[114]	√	√	-	-	-	√	-
Bibbins-Domingo et al. (2010)[115]	√	√	-	Unclear	-	√	-
Wang et al. (2012)[116]	√	√	-	√	-	√	√
Basu et al. (2013)[117]	√	√	Unclear	-	√	√	√
Konfino et al. (2013)[60]	√	√	-	-	-	√	-
Basu et al. (2014)[118]	√	√	-	-	-	√	-
Collins et al. (2014)[119]	√	√	-	√	-	√	√
Mason et al. (2014)[120]	√	√	-	√	-	√	-
Lewsey et al. (2015)[55]	√	√	-	√	-	√	√
Manyema et al. (2015)	√	√	-	Unclear	-	√	-
Wilcox et al. (2015)[122]	√	√	-	-	-	√	-
Collins et al. (2015)[123]	√	√	-	-	-	-	√
Lawson et al. (2016)[56]	√	√	-	√	-	√	√
Sa´nchez-Romero et al. (2016)[58]	√	√	Unclear	-	-	√	√
Wang et al., (2016)[124]	√	√	-	√	-	√	-
Breeze et al. (2017)[125]	√	√	√	-	-	√	√
Pandya et al. (2017)[126]	√	√	-	√	-	N/R	N/R
Mozaffarian et al. (2018)[127]	√	√	-	√	-	√	√
Riveros et al. (2018)[128]	√	√	-	√	-	-	-

Authors	Validation test					Uncertainty analysis	
	Face validity	Internal validity	Cross-validity	External validity	Predictive validity	Deterministic SA (DSA)	Probabilistic SA (PSA)
Schönbach et al. (2018)[129]	√	√	-	-	-	-	√
Huang et al. (2019)[130]	√	√	-	√	Unclear	√	√
Salgado et al. (2019)[131]	√	√	-	√	-	N/R	N/R
Wilde et al. (2019)[132]	√	√	-	√	-	√	√
Broeks et al. (2020)[133]	√	√	-	-	-	√	√
Lee et al. (2020)[134]	√	√	-	√	-	√	√
Liu et al. (2020) [135]	√	√	-	√	-	√	√
Salgado et al. (2020)[136]	√	√	-	√	-	√	√
Dehmer et al. (2020)[137]	√	√	-	√	-	√	-
Shangguan et al. (2021)[138]	√	√	-	-	-	√	√
Thomas et al. (2022)[139]	√	√	-	-	-	√	√
Lou et al. (2023)[140]	√	√	-	-	-	√	-

3.4.7 Quality appraisal

The quality of models was appraised using the Philips checklist[69] (Appendix 2), categorised into three distinct domains including structure, data, and consistency. The ‘structure’ domain assessed how well the model’s framework was constructed, including the clarity and appropriateness of the model’s design about the decision problem aims to address. The ‘data’ domain evaluates the sources, quality, and appropriateness of the data used within the model. The ‘consistency’ domain assesses the internal and external coherence of the model, ensuring that its outputs are logical and comparable with those of other models or data sources.

In the Figure 3.6, the blue colour represents "Yes" (indicating the criterion was fulfilled), orange represents "No" (indicating the criterion was not fulfilled), green indicates "Unclear" (where insufficient information or ambiguity was present), and light blue denotes "N/R" (not related or not applicable). The graph is based on cumulative percentages derived from each article's responses.

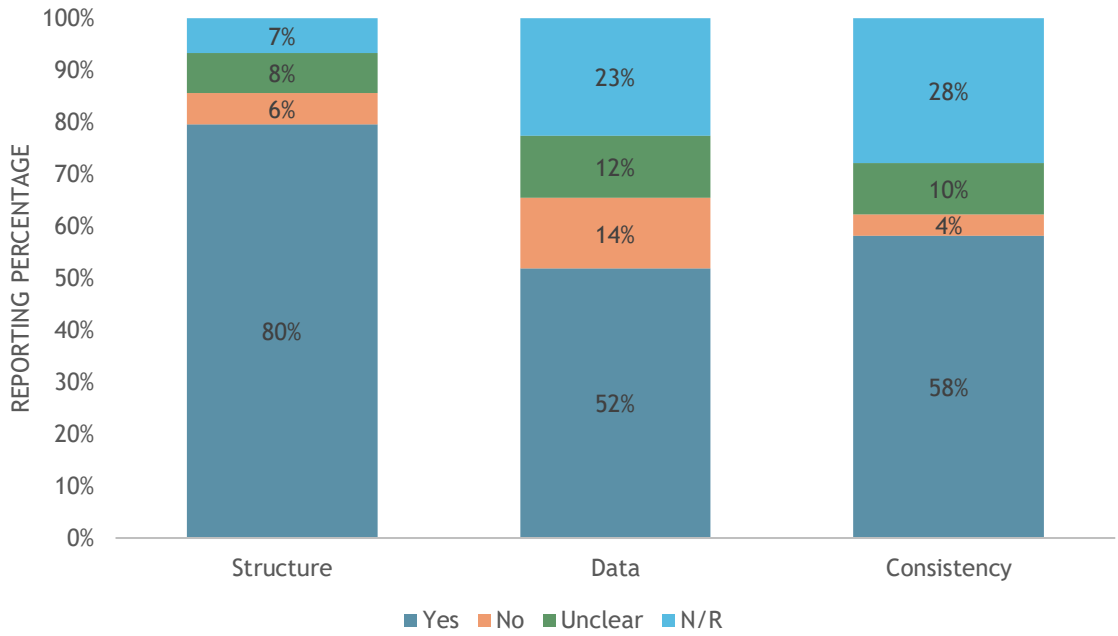


Figure 3.6 Quality appraisal of studies included in the review

Most policy models (80%) met the criteria for the 'model structure' section. This category includes the appraisal of how the decision problem was constructed, encompassing the clarity of the decision problem, the study's perspective, transparency, and consistency of model justification, input, and structural assumptions. Generally, model inputs and objectives were consistent with the stated perspectives and initial justifications. However, while the perspectives and settings were typically defined, not all models specified the decision-makers, despite the study results being intended for decision-maker use. Furthermore, most articles lacked explicit justification for the chosen time horizon and cycle length, although these were appropriately applied—likely due to the standard practice in modelling chronic diseases like CVD and T2DM. Additionally, the reasons for excluding certain options or alternative interventions were not always reported.

The cumulative quality of data and parameters used in the models was moderate (50%). This part of the appraisal focused on the data sources, the inclusion of parameters, and the methodological approaches reported in the articles. The models utilised a variety of data sources, including systematic reviews, meta-analyses, local and national epidemiological data, cost data, registries, administrative data, expert opinions, and other published sources. However, the quality assessment of the data incorporated into the models was often not clearly explained [60,116,119,121,136,137] A significant limitation was the lack of local representative data, which may have impacted final estimates and introduced high uncertainty into the results. To address this, many studies relied on data from other sources and constructed multiple assumptions. [114,116–118,124]

Although face and internal validity seems subjectively well-reported, there was less clarity regarding the transparency of validation efforts, which may have been reported elsewhere or addressed implicitly without specifying the types of validation tests performed. Despite these gaps, most models did acknowledge aspects of consistency, particularly in model structure assumption and model parameter as well as defining outcomes of interests. All models provided clear evidence of internal assessment by conducting sensitivity analysis. The cross-validity and external validation were conducted such as by calibrating against

independent data and reporting calibration results. The consistency of the articles was moderate to good (58%).

Overall, the review highlighted a moderate to good quality across different aspects of the models, with notable strengths in model structure but areas for improvement in reporting data transparency and validation.

3.5 Discussion

This systematic review offers a comprehensive critical appraisal of the methodological quality of the existing published CMD models. By evaluating the quality of these models, the findings provide valuable insights to inform and enhance the development process of a de novo policy model that can address some of the limitations identified and should be informed by a detailed conceptual model.[156]

The review enriches the evidence regarding policy models that can accommodate the analysis of preventative strategies for healthy or low-risk populations, while previous policy models are predominantly focused on general applications of summarising the evidence of particular health interventions/technologies or tailored to populations with moderate to high-risk populations. [37,70,99–101,104,106,108]

A ‘policy model’ in this review is broadly defined to encompass various modelling approaches, including epidemiology-economic models, microsimulation models, and decision models, all of which contribute to informing health policy decision [55,114,118,129,135] The distinction between policy models and decision models is often blurred, as decision models can be embedded within a broader policy modelling framework. For example, a policy model may incorporate decision-analytic components to answer specific questions—such as the cost effectiveness of an intervention—while simultaneously assessing its broader population-level and system-wide effects.

Given this overlap, this review adopts a comprehensive perspective, defining a policy model as a framework designed to evaluate clinical/health outcomes, cost, cost-effectiveness, and broader societal implications of health interventions. These models play a crucial role in guiding public health policies and programmes, aiming to reduce disease burden and improve population health by providing evidence-based projections of intervention impacts.

One of the clear advantages of modelling is the capability to estimate and simulate long-term disease progression and the impact of an intervention, which complements evidence generated in RCTs.[157,158] This SR established that models were either simulated Markov-type cohort or individual-level models (microsimulation), with different perspectives chosen, costs incurred, and sensitivity analyses performed. Cohort simulations are advantageous for their efficiency and generalisability but are limited by their inability to account for individual variability, lack of precision, potential for ecological fallacy, and challenges in modelling complex interactions. In contrast, individual-level simulations offer greater granularity and personalised insights, capturing heterogeneity and specific outcomes, but they require extensive data, are resource-intensive, may involve significant uncertainty, and can be less interpretable and generalisable. The choice between these approaches depends on the study objectives, policy questions, and data availability.

Most policy models adopt a healthcare provider perspective; however, incorporating patient perspectives and accounting for potential productivity losses could provide a more comprehensive economic evaluation[159] Given that the nature of CMD itself can significantly affect both patients' and caregivers' spending, a broader economic perspective may enhance policy relevance. However, existing studies reviewed do not provide further justification for not considering broader perspective, likely because the economic framework is typically established at the outset to align with specific policy questions.

The quality of models, as established in current appraisal does heavily rely on the quality of the data used. Many studies have highlighted concerns regarding the limited availability of representative or local data for model analysis. The lack of local clinical epidemiology data often necessitates the use of assumptions or non-local data, introducing uncertainty and raising concerns about data quality. While the use of published data from other sources can be valuable, issues regarding data transferability standards and the processes for adopting such data remain an issue. Justifications for data transferability were not consistently addressed in the reviewed studies, leading to reliance on multiple assumptions about parameters, which may introduce further limitations. Additionally, many models relied on survey and observational data (e.g., survey, self-reported non-local data), which is prone to under-reporting, selection bias, and recall bias, potentially affecting the accuracy of estimations.

The use of real-world data (RWD) and updated local data is potentially beneficial for enhancing model accuracy and representativeness. RWD reflects actual patient experiences and outcomes in routine clinical practice, providing a more accurate representation of the broader population. This improves the generalisability of findings and offers a deeper understanding of the real-world impact of healthcare interventions.[92,93,95] However, despite its significant potential, the use of RWD requires careful consideration of potential confounding variables, missing data, lead-time bias, and the inherent complexities of the data itself. Addressing these challenges effectively is crucial to maximising the benefits of RWD in modelling analysis.[80,96]

Uncertainty is inherent in every modelling exercise, underscoring the need for improved reporting and characterisation of uncertainty. Additionally, it is crucial to report clear validation tests conducted to enhance the transparency of model development.[61,62] The model validation process was mostly not extensively discussed in published articles or overlapped terms in validation itself in publication-related health economic studies, thus limiting the reporting quality.

Addressing equality and equity concerns in health economic analysis can enhance overall results. Policies that are designed solely on cost-effectiveness without considering equity can lead to interventions that are efficient in aggregate but may exacerbate existing inequalities. By integrating equity, policymakers can design more holistic interventions that balance efficiency with fairness, leading to more socially acceptable and sustainable health policies.[161,162]

The overall quality of the models in this review is relatively good. Most of the important model features are well-reported. However, in line with several current systematic literature reviews[151,163,164], not all policy models are fully comparable, due to the different model assumptions, modelling approaches, perspectives, and outcomes generated from the model.

This review is subject to some limitations. First, this review only focused our search and review on articles that defined policy or decision models in a very specific dietary policy intervention. There are probably many primordial public health strategies besides dietary intervention, such as physical activities or smoking cessation policies. Second, the various applications of the policy model objectives and parameters inputted might influence conclusions in terms of generalisability from this review. Variations in data availability and quality across studies may have influenced the reported outcomes, potentially affecting the overall reliability of the evidence based. Hence, the general interpretation of this review should be accompanied with caution since it applies a wide range of aims, model details, and reporting. It is important to note that the suitability of a policy model depends on the specific research question and data availability.

Based on this systematic review, the following recommendations are made to enhance the development of CMD policy models (Table 3.7)

Table 3.7 Evidence-based recommendations for CMD Policy Modelling

Area	Key recommendations
Model selection	State-transition models (e.g., Markov models) are commonly used for CMD progression, but analysts should align the model choice with the policy question, available data, and computational feasibility.
Integration of CMDs	Given the shared risk factors of T2DM and CVD, incorporating them into the same model can improve accuracy and capture event-related risks.
Risk factors	Models can integrate modifiable risk factors (e.g., BMI, cholesterol, lifestyle changes) to ensure more realistic projections.
Data quality	High-quality patient-level and representative epidemiological data should be prioritised. Incorporating clinical biomarkers and capturing heterogeneous effects can improve generalisability.
Economic perspective	If data permit, a societal perspective should be used when interventions generate substantial non-healthcare costs or benefits, particularly for public health and preventive interventions; otherwise, a healthcare payer perspective is generally sufficient, with societal impacts explored in sensitivity analyses where relevant.
Uncertainty analysis	Specifying uncertainty and conducting appropriate sensitivity analyses is essential for ensuring robust conclusions.
Validation	Reporting validation tests (internal, external, face validity) is recommended to improve model reliability and reproducibility.
Transparency & reporting	Clearly document model rationale, assumptions, and methodologies. Conceptual models should be well-documented to enhance credibility.
Equity & distributional analysis	Ensuring that models assess distributional impacts can support policies that reduce health inequalities.
Reproducibility & open science	Adhering to best research practices and making policy models open source can improve transparency, accessibility, and reproducibility

3.6 Conclusions

In conclusion, the policy models reviewed herein show promising insights for informing policy decisions, particularly in the context of public health preventative strategies. Based on this systematic review, several recommendations are established to enhance the development of a CMD policy model. The findings of this review directly inform the development of the conceptual model (Chapter 4), which serves as the foundation for the modelling framework developed in the subsequent chapters.

Chapter 4 The conceptualisation of a cardiometabolic disease policy model

4.1 Introduction

Developing a conceptual model is a critical aspect of mathematical/statistical works. This chapter describes how the CMD policy model is conceptualised, translating the complex ideas into a structured-concise format. Following good practice in modelling, the model development process is documented. It serves as a way to communicate problem understanding and model choice before doing further analysis. Establishing the conceptual model primarily aims to enhance model development's transparency.

The previous systematic review chapter informs the development of a conceptual CMD model in this Chapter. In Chapter 3, the existing published literature on CMD models was assessed to gain insight regarding model quality. Several aspects of modelling were critically appraised in the systematic review. Information from the review is finally summarised to inform the current conceptual model development, identifying the modelling aspects that need to be improved or can be addressed in our current model.

Chapter 4 consists of several sub-chapters, including the need for a conceptual model, the methods and guidelines used, and the process of model conceptualisation as described in section 4.2 to 4.3. The methods section explains the two-stage process model conceptualisation: conceptualising the problem and conceptualising the model structure. These included the combination of findings from the systematic review, clinical guidelines review, and expert consultation. Finally, details of the conceptual model that will be applied further for analysis are presented (sub-section 4.4.3).

4.2 The importance of a conceptual model

Over two decades, modelling techniques have been increasingly applied to assist decision-makers in various settings.[165,166] Results from modelling are utilised by stakeholders particularly when making or evaluating decisions about interventions or strategies that can improve health, both at individual and population level.

Although the use of policy models has emerged, there remains a need to improve their credibility.[167] The development process itself particularly influences the estimation of health and economic outcomes generated from these models. Appropriate development of policy models goes beyond mathematical operationalisation alone.[168] It requires an understanding of the complexity of real systems as well as the ability to translate those into credible conceptual structures. This understanding and conceptualising can be structured using a conceptual model.[167,168]

In the realm of health economic modelling, the process of conceptualisation is essential but often underreported, it might be due to the limited literature, agreed definition, and missing familiarity with the importance of this idea.[169] Nevertheless, providing a clear conceptualisation process for economic models will be helpful to ensure that decision problems are translated effectively into mathematical/quantitative models.

Developing a conceptual model has many advantages, such as improving our understanding of the decision problem, addressing the decision/policy needs, exploring current clinical/public health practice in a particular setting, and increasing the knowledge of currently available strategies.[167] A conceptual model not only allows us to visually represent the relationships between the model attributes, but also gives room for clarification of the decision problem being analysed and fosters better communications between researchers, policymakers, and stakeholders.[169–171] It is also beneficial for engaging the stakeholders' roles to obtain an agreement and approval of the problem defined as well as model structure development, considerations on relevant assumptions, and parameters incorporated in the model.[167,169–171] Hence, the

development of a conceptual model is recommended as an initial stage of the model development process.

4.3 Methods

A well-developed conceptual model serves as the foundation of robust and informative modelling. It ensures the overall clarity, validity, and credibility of the model. Tappenden (2012)[168] published guidance for conceptual model development. It consists of two distinct stages: problem-oriented and design-oriented. Problem-oriented represents the problem that exists within the system, diving into a deeper understanding of the disease and treatment pathways. Conversely, the design-oriented model articulates the envisioned framework of the model, and the proposed plan for the model structure, taking into account the available evidence.[168]

Similarly, The International Society of Pharmacoeconomics and Outcomes Research-Society of Medical Decision Making (ISPOR-SMDM) Modelling Good Research Practices Task Force-2 published guidelines for conceptualising a model.[167] Two main components of modelling processes are provided in this report (Figure 4.1). First, the conceptualisation of the problem, which covers the translation of the healthcare process knowledge into a representation of the problem (step 1 in Figure 4.1). Second, as a sequential process, model conceptualisation is used to determine which modelling types and their attributes best represent the defined problem (steps 2,3 and 4 in Figure 4.1) as well as data and parameters used (step 5 in Figure 4.1), followed by transparency and validation of the model (step 6 in Figure 4.1). The nature of the conceptual model described in these two reports is not too distinct, both set boundaries between two sequential stages in model development.[167,168]

This ISPOR-SMDM report guides researchers by outlining good research practices for developing conceptual models in health economic evaluations. It does not prescribe a specific model but offers the framework for developing and choosing a suitable model that fits specific research questions, and the decision problem addressed. This thesis follows the ISPOR-SMDM guidelines for developing a CMD Policy Model.[167] The general stages of the conceptual model process are illustrated in the following figure (Figure 4.2).

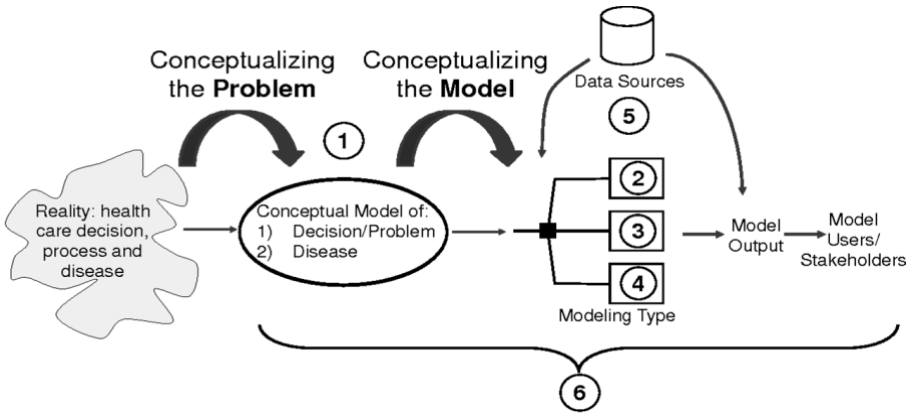


Figure 4.1 Development and construction of a model [167]

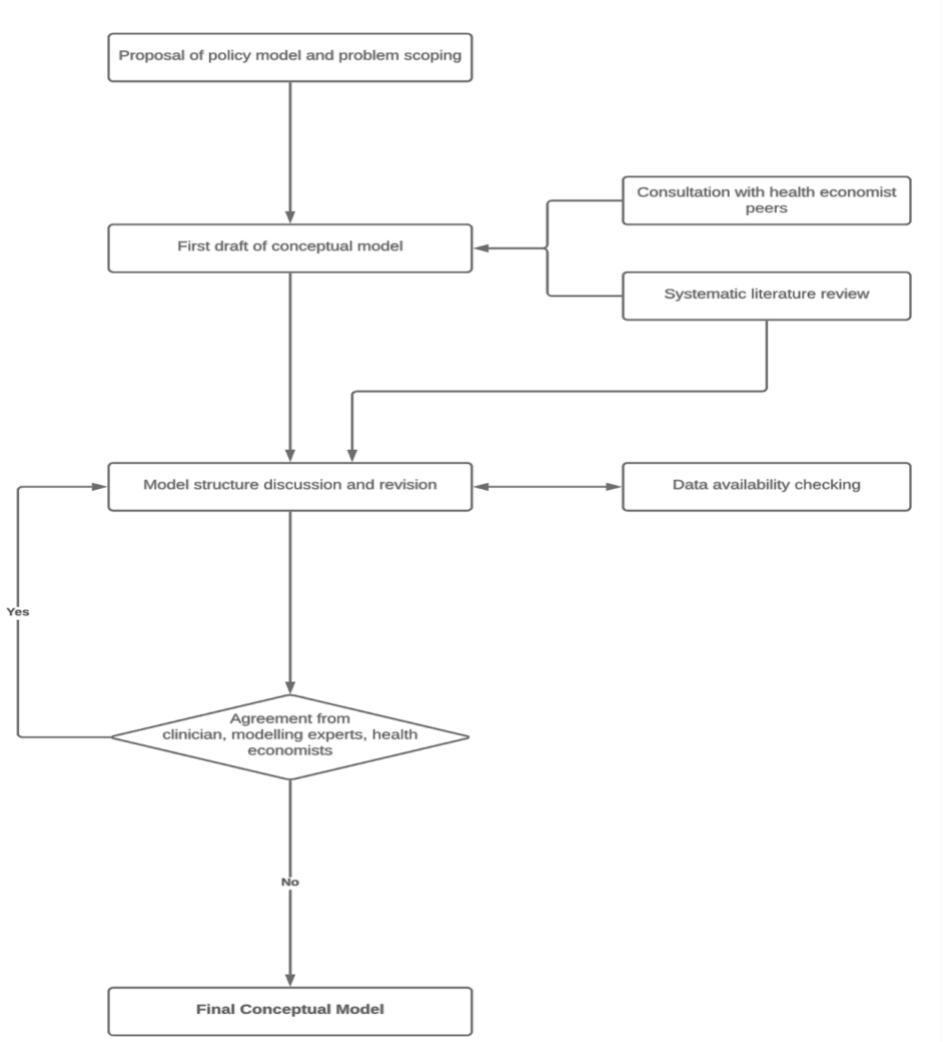


Figure 4.2 Schematic flow diagram of conceptual model development for CMD Policy Model

4.3.1 Part 1: Conceptualising the problem

A conceptualisation of the problem requires an understanding of CMD progression and prevention based on clinical and public health guidelines available in the UK. This is guided by the findings from the systematic review that has been conducted in Chapter 3.

The initial conceptual model was developed based on clinical guidelines and systematic review findings to reflect assumptions about disease progression and structure. This outlines key components of conceptual modelling framework by providing a foundation for expert input and iterative refinement (Appendix 3).

4.3.2 Part 2: Conceptualising the model

Expert opinion

The objective of the model development is not to reproduce, but to represent a simplified reality. The model development process should reflect the reality that represents the decision problem.[168] To accommodate this, the role of clinical experts is important to meet contextual relevance. This stage highlighted the significance of collaboration between clinical experts and experienced modellers to achieve consensus on the model structure.

The expert group, consisting of a clinician, two health economists, and a medical statistician, provided diverse and valuable perspectives that contributed to the refinement of the model. The clinician's input was particularly crucial in ensuring that the model remained aligned with clinical practice and relevant to real-world applications. The conceptual model draft (Appendix 3) was presented to the group, and informal feedback was gathered during the presentation. After informal consultations with these experts (2-3 meetings), the conceptual model and model structure were revised to better reflect clinical realities and to enhance its overall validity.

4.4 Results

4.4.1 Understanding the disease progression

Cardiometabolic disease (CMD) is inherently complex due to its multifactorial nature of interconnected risk factors and comorbidities. The interrelation of genetic, environmental, and lifestyle factors collectively contribute to the development and progression of cardiometabolic conditions that lead to wide pathophysiological mechanisms and clinical manifestations.[6,7]

The progression of CMD typically begins with insulin resistance, which may lead to metabolic syndrome or 'pre-diabetes'. As cardiometabolic syndrome (CMS) progresses, the body's ability to respond to insulin diminishes, compelling the pancreas to compensate by producing higher levels of insulin. However, over time, this compensatory mechanism becomes insufficient, leading to impaired glucose tolerance (IGT) and, ultimately, the onset of T2DM. CMS also doubles the risk of CVD contributing to the rising incidence of heart attacks, strokes, and coronary artery disease. The interplay of insulin resistance, dyslipidaemia, hypertension, and chronic inflammation in CMS accelerates atherosclerosis by promoting endothelial dysfunction, oxidative stress, and plaque formation. This process narrows the arteries, increasing the likelihood of myocardial infarction (MI) and stroke.[172–174]

Guidelines for identifying and diagnosing the risk of cardiometabolic syndrome are based on sources generated from the World Health Organization (WHO)[175], the European Group for the Study of Insulin Resistance (EGIR)[176], the International Diabetes Federation (IDF)[177], National Cholesterol Education Project Adult Treatment Panel III (NCEP ATP III)[178], the National Heart, Lung, and Blood Institute/American Heart Association (NHLBI/AHA)[179]. Several recent recommendations and guidelines for CMD staging have also been introduced (Table 4.1).[180,181]

Table 4.1 The Cardiometabolic Disease Staging System (CMDS)

Stage	Descriptor	Criteria
Stage 0	Metabolically Healthy	No risk factors
Stage 1	One or two risk factors	Have one or two of the following risk factors: a. high waist circumference b. elevated blood pressure or on anti-hypertensive medication c. reduced serum HDL cholesterol or on medication d. elevated fasting serum triglycerides or on medication.
Stage 2	Metabolic syndrome or prediabetes	Have only one of the following three conditions in isolation a. Metabolic Syndrome based on three or more of four risk factors. b. Impaired Fasting Glucose (IFG) c. Impaired Glucose Tolerance (IGT)
Stage 3	Metabolic syndrome + prediabetes	Have any two of the following three conditions: a. Metabolic Syndrome b. IFG c. IGT
Stage 4	T2DM and/or CVD	Have Type 2 Diabetes Mellitus (T2DM) and/or cardiovascular disease (CVD): a. T2DM (fasting glucose ≥126 mg/dL or 2-hour glucose ≥200 mg/dL or on anti-diabetic therapy) b. active CVD (angina pectoris, or status post a CVD event such as acute coronary artery syndrome, stent placement, coronary artery bypass, thrombotic stroke, non-traumatic amputation due to peripheral vascular disease)

Cardiometabolic staging involves the classification and progression of CMD, including CVD, diabetes, and associated risk factors, to facilitate targeted interventions and management strategies. The identification of distinct disease patterns and subtypes within the spectrum of CMD highlights the heterogeneity and complexity of these conditions. The complexity of CMD is also evident in the need for comprehensive guidelines and interventions to address the nature of these conditions.

As mentioned in Chapter 1, the CMD prevention and treatment guidelines are still in the development stage in the UK. [43] A screening strategy has also recently been proposed.[4] However, the diagnosis, prevention, and treatment guidelines for CVD and T2DM have been published by The National Institute of Care and Excellence (NICE)[182,183] as well as the Scottish Intercollegiate Guidelines Network (SIGN).[184] Physical activity, dietary recommendations, behavioural changes, and other primordial preventive policies are covered in the guidelines and recommendations as described. These guidelines are continuously evolving based on the latest evidence and are aimed at addressing the prevention and management of CMD in the population.

4.4.2 Summary of published evidence

The details of a systematic review of published literature have been described in Chapter 3. To prevent repetition, only result summaries are outlined here to integrate the justification of problem conceptualisation.

Model types and structure

State transition models (STMs), particularly those using a cohort-based Markov framework, are commonly employed in the analysis of chronic, long-term conditions such as CMD. In these models, health states are typically defined based on clinical guidelines and the natural progression of disease. While existing models often focus on later stages of CMD, future approaches could benefit from a more comprehensive structure that explicitly incorporates intermediate complications of type 2 diabetes mellitus (T2DM) and cardiovascular disease (CVD) events, such as myocardial infarction (MI) and stroke.[185]

Markov models are well-suited to represent transitions between health states over time. They allow for the simulation of time-dependent risks, recurrent events, and disease trajectories that involve repeated or episodic outcomes (e.g., MI, stroke).

Risk Factors

Risk factors and covariates can generally be classified into two categories: modifiable (e.g., lifestyle, biomarkers) and non-modifiable (e.g., age, family history). The systematic review in Chapter 3 identified published models that incorporate both types. Moreover, dynamically modelling changes in key biomarkers over time can enhance the realism and accuracy of estimated outcomes.[156]

Other models' features (costs, outcome, validation, sensitivity analyses)

From the review, most policy models adopt a healthcare perspective. Incorporating productivity losses from patient and caregiver viewpoints could enhance decision-making relevance. While sensitivity analyses (SA) like deterministic, probabilistic, or both are routinely reported to address uncertainty, validation procedures (e.g., face/internal validation) require greater transparency in reporting methodologies and impacts. Few studies assess equity implications, though integrating these could foster more holistic, socially sustainable policies that balance efficiency and fairness.[185]

Model reliability fundamentally depends on input data quality; a key challenge is the scarcity of representative local data for development. Heavy reliance on assumptions or external data introduces uncertainty and compromises validity. Though secondary data offer utility, inconsistent transferability standards and inadequate justification for their use limit effectiveness. Integrating fit-for-purpose real-world data (RWD) may improve model accuracy and generalisability.[185]

Expert opinion

Inputs from a clinician and experts are mostly focused on the proposed model structure and key features of the model. Initially, the general risk factors were based on findings from the systematic review. However, following a review of clinical guidelines, metabolic conditions such as obesity, hypertension and hyperlipidemia/dyslipidaemia are also considered as covariates due to their strong correlation with metabolic syndrome.[6,7,172]

Experts proposed atrial fibrillation (AF) to be included as a health state in the conceptual model, as cardiometabolic risk factors can increase the risk of AF, while AF itself can contribute to the progression of cardiometabolic conditions.[186] Another recommendation was to re-evaluate T2DM progression, in particular whether we should include states with/without diabetes complications before the final state. The onset of diabetes accelerates the development of atherosclerosis and other CVD risk factors, showing that people with diabetes also have a risk of having CVD.[172] This progression was recommended to be added to the final conceptual model since the conceptual model draft did not consider this relationship.

For CVD state (stroke and MI), defining post-CVD event must be ensured [56,126], since there may be differences in terms of utility assessment that could influence the cost-effectiveness results if we plan to conduct further cost-utility analysis (CUA). For instance, fatal MI/Stroke would have different utility values than non-fatal events.[156]

Finally, in further discussions with the health economists, it was agreed that the model should remain representative but not overly complex, and that atrial fibrillation (AF) should be considered as a potential covariate in the model. In addition, the model structure was presented at an internal meeting attended by a broader research audience. Three main points were raised. First, the model was commended for its simplicity and representativeness; however, it was noted that these should be balanced against the feasibility and time required for the analysis and modelling exercise. Second, it was recommended to explore options for incorporating productivity loss parameters into the analysis or sensitivity analyses, to capture indirect costs from patient and caregiver perspectives in addition to direct costs. Third, the model should also address equality and equity considerations, consistent with the findings from the systematic review.

4.4.3 Final conceptual model

The final conceptual model is illustrated using an influence diagram form (Figure 4.3). This proposed model will facilitate the improvement of clinical and economic representation of CMD, where metabolic dysfunction conditions could lead to various events including both CVD and T2DM.

To operationalise this conceptual model, a detailed model-structuring stage is specified (Figure 4.4). The target population comprises all adults (≥ 18 years) without a confirmed CMD at model entry. Baseline characteristics with various risk factors determine individual risk profiles. Disease states include: disease-free; type 2 diabetes mellitus (T2DM); CVD (myocardial infarction [MI], stroke); post-CVD (post-MI, post-stroke); and death. All individuals begin in the disease-free state and may transition to T2DM, MI, stroke, or directly to death. Those who develop T2DM face elevated risks of MI, stroke, or death. Patients experiencing MI or stroke either die or survive and move into the corresponding post-CVD state, where they remain at risk for subsequent events or death.

Diabetes remission was not explicitly modelled. Although recent diabetes remission programmes have shown promising results, evidence on the long-term durability of remission at the population level remains limited, with substantial relapse observed within 2–5 years.[187,188] Modelling remission would require strong assumptions regarding remission duration, relapse rates, and long-term cardiovascular risk reduction, introducing considerable structural uncertainty. Diabetes was therefore modelled as a chronic condition influencing cardiovascular risk trajectories, consistent with established economic evaluation guidance and cardiometabolic policy models that discussed in systematic reviews.[185]

The model is progressive: once a higher-severity state is reached, reversion to a less severe state is not permitted. Death functions as the sole absorbing state, accessible via multiple pathways. Key outputs include all-cause mortality, disease-specific mortality, life expectancy, quality-adjusted life expectancy (QALE)/quality-adjusted life years (QALYs), and lifetime health-care costs. This structure supports evaluation of disease progression, long-term health outcomes, and the cost-effectiveness of interventions, thereby informing survival analyses and health-policy decision-making.

Semi-parametric (e.g., Cox) and fully parametric survival models were performed, feeding into a semi-Markov state-transition framework.[189,190] While traditional cohort Markov models are commonly identified in the systematic review, the semi-Markov extension more closely aligns with the nature of CMD

and lets the model handle varying time spent in each state and time-dependent transitions, better reflecting CMD progression.

The semi-Markov framework offers several technical advantages for the CMD model. [191–193] First, it permits non-exponential sojourn distributions, capturing clinically plausible phenomena such as escalating progression risk after prolonged residence in a pre-disease or intermediate state. Second, transition probabilities may depend on both fixed and time-varying covariates thereby yielding more personalised patient trajectories. Third, integration with longitudinal data methods enables joint modelling of state transitions alongside continuous outcomes, supporting comprehensive long-term health and economic assessments. [191–193] By contrast, standard (memoryless) Markov models often oversimplify these dynamics, risking underestimation or mischaracterisation of transition timings and probabilities. The semi-Markov approach thus furnishes a more realistic and policy-relevant foundation for modelling CMD progression and evaluating interventions.

To parameterise this, Clinical Practice Research Datalink (CPRD) data was used. CPRD comprises records for 60 million patients of whom 18 million are currently registered across England, Scotland, Wales, and Northern Ireland.[194] The remaining records represent patients who were previously registered but are no longer active in the database.

The dataset captures routine clinical information on demographics, behavioural factors, signs and symptoms, diagnoses, prescriptions, immunisations, referrals, and lifestyle measures. Beyond its large, nationally representative sample, CPRD’s longitudinal design enables analysis of disease onset and progression over extended periods. Moreover, linkage to hospital and mortality registries facilitates reconstruction of complete patient pathways.[195]

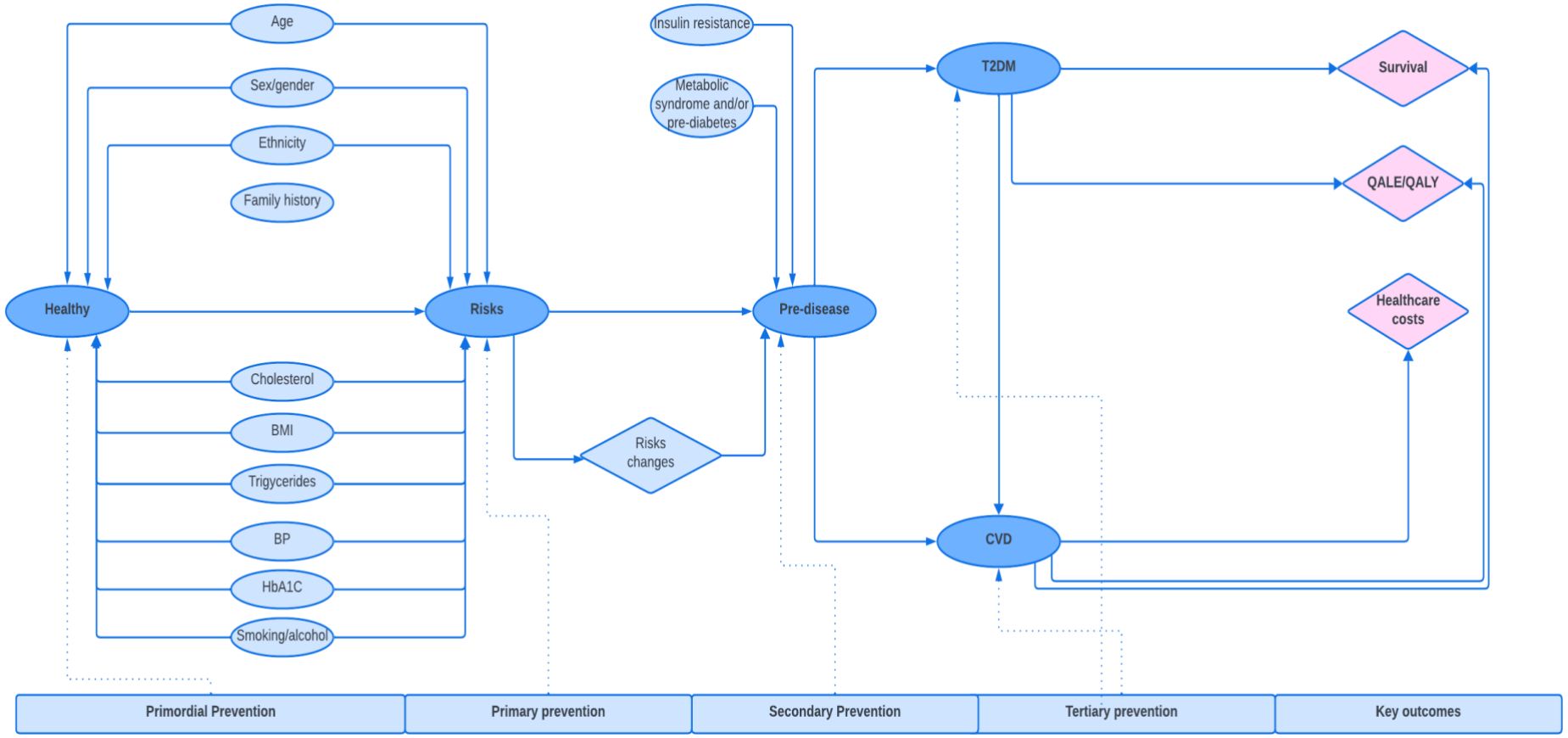


Figure 4.3 Final conceptual model

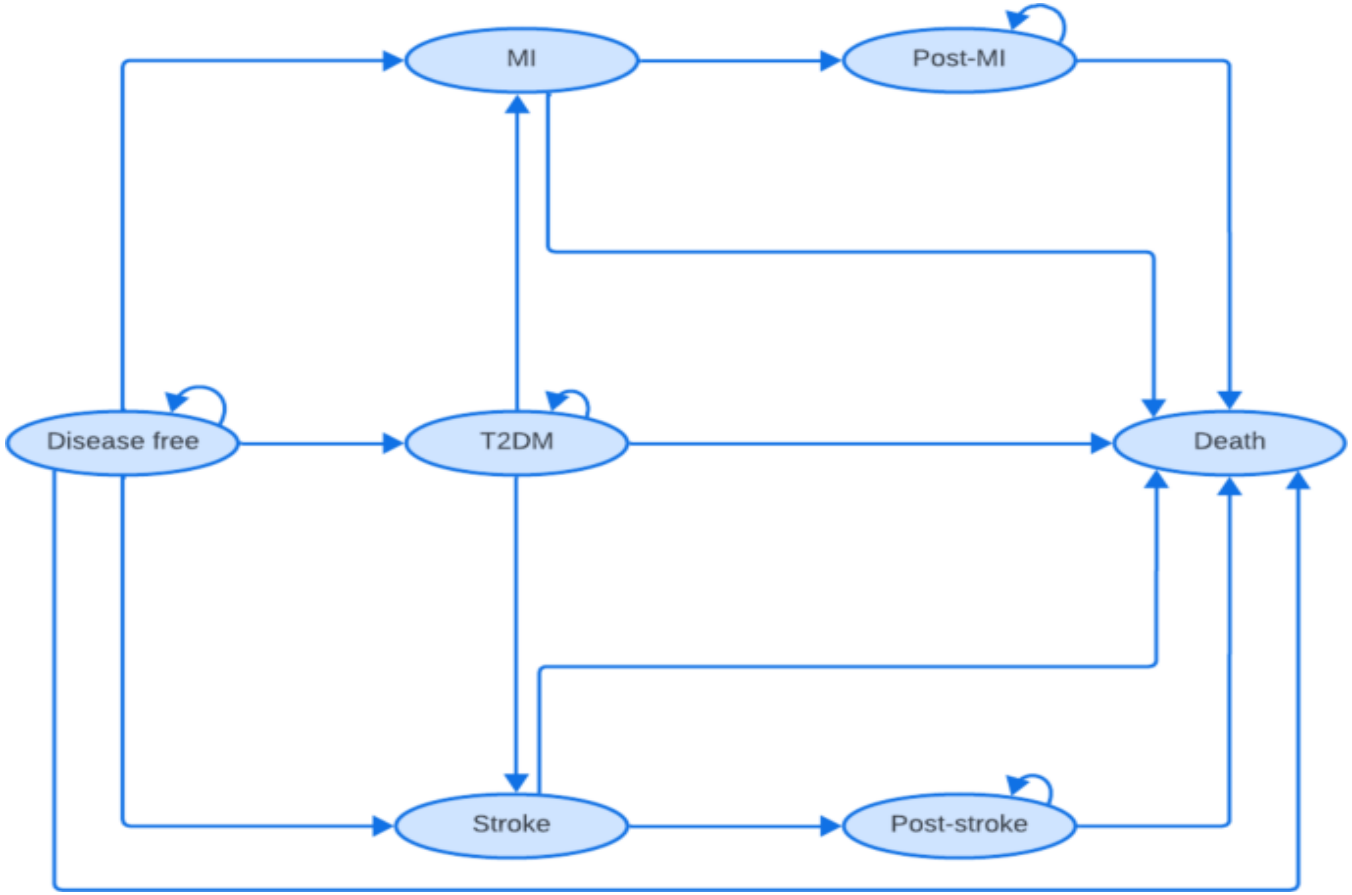


Figure 4.4 State transition model structure

4.5 Discussion

It is well-understood that models must be clearly defined and conceptualised prior to analysis.[168,167] This chapter introduces a conceptual model serving as a foundational framework for developing policy models that is both appropriate and fit for purpose, by explicitly defining core components, relationships, and underlying assumptions. This framework ensures subsequent policy models possess robust theoretical grounding and practical utility.

Given the adequate reporting quality established in the systematic review, the findings were deemed reliable and provided valuable insights for conceptual model development. Integration of clinical guidelines, systematic review evidence, and expert consultation significantly strengthened the modelling process. Particular emphasis was placed on model structuring[196], yielding a technically precise and detailed conceptual framework.

Areas which required further attention include rigorous selection and incorporation of parameters, especially leveraging high-quality routine data to enhance conclusion generalisability.[185] These considerations have been addressed in the final conceptual model. The proposed structure aligns with established CMD stages and existing economic evaluation frameworks, demonstrating methodological consistency with current practices. Based on this alignment, no major modifications appear necessary in this conceptual model.

Following development, the CMD policy model could be applied to evaluate early preventative strategies, including dietary interventions, screening programmes, and preventive medications. Subsequent analyses will incorporate structural sensitivity testing and model performance evaluation in accordance with established modelling best practices.[69]

To date, published conceptual models in this domain remain relatively limited.[170,196–199] The conceptual model presented herein advances current research on CMD, particularly within health economic modelling. It extends existing frameworks through a comprehensive and systematic

conceptualisation process, prioritising modelling transparency and methodological rigor.

Some limitations of this conceptual model are acknowledged. First, an experienced clinician and experts were asked to ensure the disease state relevance at a practice level. This is done by gathering input informally during the presentation of the modelling plan. Conducting a Delphi panel with a structured questionnaire would potentially improve the process and minimise subjectivity. Second, the structure is trying to cover both T2DM and CVD states that represent major CMD events. Unlike the second event such as the post-CVD event, the model did not consider T2DM complications as a subsequent state from T2DM state. The model assumed that patients would progress to death over time, no matter what diabetic event occurs after the initial T2DM event. Third, transitions between MI and stroke were not permitted in the model. MI and stroke were treated as competing first cardiovascular events arising from a shared atherosclerotic process. Explicitly modelling MI–stroke transitions would require robust, time-dependent estimates of conditional second-event risks, which are limited and highly uncertain. Instead, subsequent vascular risk was captured within post-MI and post-stroke states through increased mortality, costs, and utility decrements. This structure is consistent with established cardiovascular policy models but may underestimate the burden associated with multiple sequential cardiovascular events.[200,201] Fourth, a healthcare perspective is planned to be used for the model in terms of facilitating further economic analysis. Considering a societal perspective in the model may optimise societal decisions. If sufficient data are available, this economic perspective may be incorporated in sensitivity analysis.

Furthermore, the use of utility values to generate QALYs is considered, such as EQ-5D. However, the EQ-5D-5L valuation study remains ongoing for the UK general population.[202] A solution here is to use published EQ-5D-3L for each state (if a hypothetical public health intervention is conducted), or use QALE as one of the outcomes.

4.6 Conclusions

It is widely accepted that clearly defining and conceptualising the model is a crucial first step before analysis. The conceptual model developed in this Chapter serves as a first step in representing the systematic process for communicating the contextual understanding of the current problem and knowledge, disease progression, and modelling choice as well as its structure. Basically, the conceptual model describes how the decision problems are specified and how the model structure is established. It will be beneficial to provide insight to the broader audience for the modelling development process before further analysis.

Chapter 5 Data preparation

5.1 Introduction

Data preparation is a pivotal stage in the data analysis pipeline, especially when dealing with real-world data (RWD), as it transforms raw data into a structured and analysable format suitable for advanced statistical and computational analysis. Chapter 5 offers a detailed exploration of the data preparation process for (Clinical Practice Research Datalink) CPRD Aurum, outlining the pipeline and the sequence of operations such as cleaning, linking, integrating, and manipulating the data—while addressing key challenges inherent to this phase.

The dataset utilised in this study has been granted ethical approval by the UK Health Research Authority (HRA) Research Ethics Committee (REC) as part of CPRD's standard governance protocols. Specific ethical clearance for the cardiometabolic disease study was secured from the Independent Scientific Advisory Committee (ISAC) under project number 20_129.[a] Furthermore, a Data Management Plan (DMP) was submitted to the University of Glasgow in 2023, ensuring compliance with institutional standards for secure data handling and storage (Appendix 4).

All data management and analysis processes are conducted in collaboration with the Medical, Veterinary and Life Sciences (MVLS) Advanced Research System (MARS) University of Glasgow, a high-performance computing (HPC) platform that facilitates complex, computationally intensive research.[b] To ensure transparency and reproducibility, the entire data preparation process, including all relevant code, is available publicly via a GitHub repository.

[a] Cardiometabolic disease prediction using general practice consultation pattern: Use of machine learning (ML) https://www.cprd.com/approved-studies/cardiometabolic-disease-prediction-using-general-practice-consultation-pattern-use

[b] MVLS Advanced Research System. High Performance Computing. https://mars.ice.gla.ac.uk

5.2 Secondary use of routine data

Routine data, or routine health data is information collected as a part of regular healthcare delivery. The primary use of this is to directly inform the care of individuals/patients whose data was collected. This data is typically gathered by healthcare providers during visits, hospital admissions, and immunisation programmes (e.g.: electronic health records, disease registries, administrative data, claim data, epidemiologic surveillance, etc.). However, this data has a wealth of potential beyond informing individuals or clinicians. The secondary use of routine health data focuses on leveraging aggregated information to benefit a broader population.

Over two decades, there has been substantial growth in the use of routine data for public health research. The secondary use of this data plays a pivotal role in providing a more comprehensive understanding of health and disease in individuals/populations, improving clinical decision-making, medical intervention, and personalised care, as well as enhancing the wider impact of the healthcare system.[203–205] Furthermore, the use of routine data (i.e.: electronic health records) enables insights into population representativeness and the possibility of long-term follow-up analysis.[206–209]

Routine data utilisation enables a reflection of the health and disease conditions in the general population, captures the reality of clinical practice, and allows for a more granular understanding of real-world clinical practice.

In the UK, one of the routine EHR-based databases is the Clinical Practice Research Datalink (CPRD), which stands as the pre-eminent repository of primary care data, offering anonymised, comprehensively coded EHR data that are collected every month by a nationwide GP network. All patient data are securely stored in EHR software, and datasets generated from this can be utilised to support retrospective and prospective public health research.[210] For more than 30 years, CPRD data has supported the development of clinical guidance and best practices, including medicine use, drug safety investigations, disease risk factors, healthcare delivery, and the effectiveness of health policy.[211,212]

CPRD launched a new database called CPRD Aurum in 2017, and it became available to use in 2018. CPRD Aurum contains data that is predominantly sourced in England and Northern Ireland, capturing symptoms, diagnoses, tests, prescriptions, and referrals for over 20 million patients. CPRD Aurum has a different data platform system for electronic records compared to the well-known CPRD GOLD. CPRD Aurum uses the EMIS® platform data system, while CPRD GOLD uses Vision® software that has reduced in use by GPs in recent years. Despite these differences, both databases remain to provide research potential to support public health research.[210] This thesis will utilise the CPRD Aurum to build a CMD policy model, a mathematical framework, particularly a state transition model that can accommodate health and economic analysis. The details of this model conceptualisation have been presented in Chapter 4.

In general, EHR-based data are mainly inputted by end-user healthcare providers as a part of routine patient care. The vast amount of information contained in EHR-based databases result in massive database size, complex data relations and structure, and comprehensive information on individual clinical history. Given the complexity and voluminous nature of the database, researchers are often confronted with formidable challenges during the data preparation stage.[213,214]

Data preparation is the most critical phase in the analytical process, as a significant portion of coding effort is dedicated to ensuring data quality. This involves data transfer, ingestion, cleaning, handling missing values, manipulation, and structuring. These steps entail complex, iterative technical tasks, ultimately producing refined datasets that are ready and suitable for subsequent statistical analysis.

5.3 Data source profile

5.3.1 The Clinical Practice Research Datalink (CPRD)

Clinical Practice Research Datalink (CPRD) is a UK government research data service that provides real-world routine collected data that support observational clinical and public health studies. It is jointly supported by the Medicines and Healthcare Products Regulatory Agency (MHRA) and the National Institute of Health Research (NIHR), as part of the UK Department of Health and Social Care. Established in 1987, the Value-Added Medical Products (VAMP) dataset expanded to become the General Practice Research Database (GPRD) in 1993 and continuously became CPRD in 2012.[211,212]

CPRD collects anonymised patient-level data from participating GP practices across the UK and extensively recorded data for millions of patients. With reliable research standard data, CPRD represents almost 25% of the UK population, and currently encompasses more than 60 million patients, including 18 million currently registered patients.[215] Data recorded in CPRD include primary diagnosis, clinical events, prescriptions, tests, demographics, referrals, admissions, and preventative care. Primary care data are collected electronically daily by the GPs and uploaded to CPRD secure servers on a monthly basis before being released further for public health research. Moreover, CPRD datasets were linked to Hospital Episodes Statistics (HES) for admitted patients and outpatients, the Office for National Statistics (ONS) as well as Index of Multiple Deprivation (IMD).[210]

Before the data is fully released for research purposes, validation and quality checks are performed. First, the collection level validation ensures the data element is correctly captured and checks for duplication. Second, the transformation level ensures all events are linked to patients. Third, the research quality level includes a check of recording and internal consistency of key variables.[216] It is possible to extend these checks to include specific checks and validation if required, or to fit the research purpose.

5.3.2 CPRD GOLD versus CPRD Aurum

Although CPRD GOLD and CPRD Aurum are UK EHR-based large databases that collect de-identified primary care patient-level data, there are several differences between these useful databases. The main difference is the patient management software system to record routine clinical data. CPRD GOLD uses the web-based Vision® software, the data source generated from this system has been used for more than 30 years for research.[210]

Meanwhile, GP practices in the UK are gradually switching to a new system called EMIS® software. The clinical coding system for the medical dictionary is different. The medical dictionary for CPRD GOLD contains information on all medical history using read version v.2 codes referenced in the data file as 'medcode', while CPRD Aurum uses a combination of SNOMED, Read, and local EMIS® Codes. For drugs and prescriptions, it is referenced as 'prodcode' (gemscript product code) and dictionary of medicines devices (DM+D), respectively.[210]

There are similarities and differences between these two databases summarised by Jick et al. (2023)[217] This distinction allows the researcher to strengthen the consideration and justification when planning research using CPRD data.

Table 5.1 Data use consideration when using CPRD and linked data[217]

CPRD GOLD and CPRD Aurum data coverage	▪ CPRD Aurum contains more patients than Gold, especially within currently contributing practices (data form 1989 to present)[1] ▪ CPRD Aurum primarily includes English practices (data form 1989 to present)[1] ▪ CPRD GOLD covers practices from all UK nations; however, currently contributing practices are predominantly located in Scotland and Wales (data form 1989 to present)[2]
CPRD GOLD and CPRD Aurum data quality	▪ Similar data quality in GP record for CPRD GOLD and Aurum, particularly post- 2004, though variability in completeness and quality exists over time. ▪ Validation effort should be an ongoing component of research using database.
HES and ONS Death registration	▪ HES and ONS linkage are available for practices in England ▪ Most CPRD Aurum practices have linkage to HES and ONS ▪ Very few currently contributing CPRD GOLD practice have linkage to HES and ONS data (due to linkage availability for English practices only ▪ HES APC started in 1997 and ONS death registration started in 1998, the start of CPRD GOLD and Aurum data ▪ HES APC and ONS are updated approximately yearly[3]. ▪ HES OP has limited capture diagnosis information

Notes: [1,2] at the time of the paper publication (2023). Abbreviations GP: general practitioner, HES: hospital episodes statistics, HES APC: HES admitted patient care, HES OP: HES outpatient data, ONS: Office of National Statistics

In addition, CPRD GOLD have practices up to standard (UTS) date, while this is missing from Aurum. Both include derived death date and acceptable patient flag.[210] Recent publications compared CPRD GOLD and Aurum for breast cancer[218] and rheumatoid arthritis (RA)[219] in terms of the consistencies between the two databases. The information on clinical details was consistent, also the correctness and completeness of the diagnosis were similar.

Based on the information above, the selection of CPRD Aurum over CPRD GOLD for this study is justified by several key factors related to data coverage, quality, and linkage availability, as outlined in the comparative profile of the two databases. Although both CPRD GOLD and CPRD Aurum are valuable UK-based electronic health record (EHR) databases, CPRD Aurum offers distinct

advantages that align with the research objectives and requirements of this study.

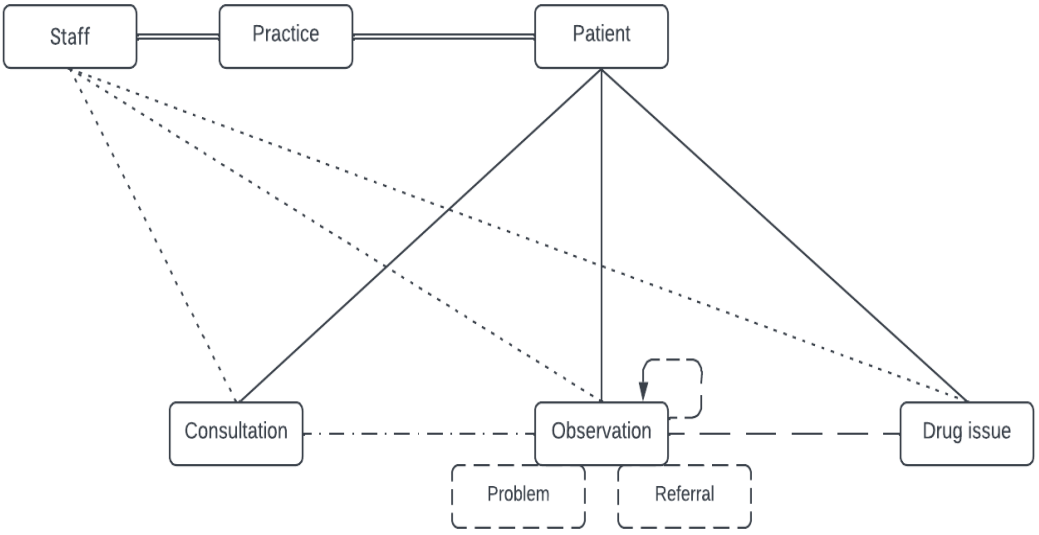
First, CPRD Aurum contains a larger patient population than CPRD GOLD, particularly within currently contributing practices (data from 1989 to present). This extensive coverage enhances the statistical power and generalisability of the study findings.[210] Second, CPRD Aurum utilises the EMIS® software, which is increasingly adopted by GP practices across the UK, reflecting a more contemporary and widely used system for recording routine clinical data. This modern coding system enhances the accuracy and relevance of the data for current clinical research.[210] Third, the superior linkage to secondary care and mortality data is critical for this study, as it enables a more holistic analysis of patient outcomes. It also provides data for longitudinal follow-up, including hospitalisations and mortality.[217]

In summary, CPRD Aurum's broader patient coverage, alignment with modern clinical coding systems, superior linkage to secondary care and mortality data, and relevance to current and future research make it the preferred choice for this study. These advantages outweigh the benefits of CPRD GOLD, particularly given the study's focus on English primary care data and the need for comprehensive linked datasets to analyse patient outcomes effectively.

5.3.3 Data structure of CPRD Aurum

Primary care data in CPRD Aurum has a complex relational data structure and format. There are eight structured separate files, and data are formatted in long format, where a patient could have multiple rows of data. The patient file contains the basic information of patient's demographics, registration details, and date of death. The practice file contains the practice region and date of data collection by practice. Meanwhile, the staff file records the job category for staff registered in CPRD Aurum.[195]

Furthermore, the observation files record the medical history of patients, including symptoms, clinical measurement, laboratory tests, and diagnosis as well as patient ethnicities. The consultation files contain the type of consultation performed by the GP (i.e: telephone, visits type), these files can be linked with observation and or the problem, referral, as drug files data relating to prescriptions (Figure 5.1)



Linked via: ____ patient id ____ practice id ..... staff id — — observation id - . - . consultation id

Note: Observation files contain symptoms, diagnoses, immunizations, tests, and lifestyle factors. Problem and referral files contain add-on information for certain types of observation. Some drug issues are linked to problem-type observation.

Figure 5.1 CPRD Aurum dataset structure

By May 2022, the CPRD Aurum database recorded over 41 million acceptable patients for research, with 38 million eligible for linkage. The total number of NHS GP practices is 1,491. In terms of coverage, this data covers approximately 13 million patients in the UK population (Table 5.2).[215]

Table 5.2 Updated version of Aurum database, May 2022[215]

Metrics	Coverage
Total number of acceptable patients[1] (including transferred out and deceased patients):	41,200,722
Current acceptable patients (i.e: registered at currently contributing practices, excluding transferred out and deceased patients):	13,300,067
Percentage UK population coverage[2] (current patients only):	13,300,067 of 67,081,000 (19.83%)
Total patients eligible for linkage	38,377,503
Available follow-up time in years since 1st January 1995[3] (all patients including transferred out and deceased): Mean (standard deviation) Median (25th and 75th percentile)	 7.93 (7.97) 4.76 (1.80-11.77)
Available follow-up time in years since 1st January 1995 (all patients including transferred out and deceased): Mean (standard deviation) Median (25th and 75th percentile)	 11.66 (9.49) 8.74 (3.25-19.85)
Total number of practices (current and historic) included in the database:	1,491
Currently contributing practices:	1,345
Percentage coverage of UK GPs (currently contributing practices only):	1,345 of 8,178 (16.45%)
Regional distribution of currently contributing practices[4]	
England	1,332 (99.03%)
Northern Ireland	13 (0.97%)
Scotland	0 (0.00%)
Wales	0 (0.00%)

[1] Permanent registration only; [2] Based on lates UK population estimates from ONS; [3] Follow-up time stated does not incorporated UTS data and the database includes records pre-dating the 1st January 1995.; [4]Expressed as a percentage of all practices currently contributing to CPRD Aurum

Apart from this, the missing data and the different GP IT systems and coding should be explored more when using CPRD Aurum data. Recently published studies reported the correctness and completeness of several diseases (pulmonary embolism, MI, T2DM, comorbidities)[204,220,221] recorded in CPRD Aurum is suitable to be used for research. We therefore arguably focus on the data cleaning stage since the data checking, quality assurance, as well as data quality and accuracy has been performed by CPRD and these published studies.

5.3.4 Variables in datasets

Variables in CPRD Aurum datasets encompass a wide range of clinical, demographic, and administrative data. The details of overall data specification are presented in CPRD Aurum data specifications. For this thesis the key sub-dataset and variables are presented in Table 5.3. These datasets are linked to each other to draw the sequence and comprehensive journey of patients recorded in CPRD. This serves as the specific information that will be utilised in the CMD model.

Table 5.3 Details on dataset used for model development and statistical analysis

Dataset	Description	Key variables
Patient	Patient and practice information, HES id and patient's ethnicity	patient id, age, gender, practice id, hes id, ethnicity, deprivation index
Observation	Covers any clinical data/measurement reported in certain observation time	observation date, medcode id, measurement values (numutid),
Diagnosis	Divided into three datasets: primary diagnosis, episodes, hospitalisation. Consists of episodes and hospitalisation information. Unique identifiers for each clinical episode in database, also the time of clinical episodes/events.	ICD, admission and discharge date, start and end date (episodes),
Linkage	List of patients who are eligible to be linked to HES and ONS death data	linked date, HES, spno, death
Death	Details of patient death information including date of death and cause of death	date of death, cause of death

HES: hospital episode statistics; ICD: international classification of diseases; medcode: A unique identifier used to represent specific medical concepts, such as diagnoses/symptoms/clinical observations selected by GP; spno: spell number uniquely identifying a hospitalisation

5.4 Data pipeline

The patient-level data was obtained directly from CPRD and securely stored within the University of Glasgow's cloud services (NextCloud). Prior to sampling, a data pipeline was constructed to facilitate the ingestion, transformation, and manipulation of CPRD Aurum data, ensuring its readiness for analysis. This stage is critical for maintaining efficient data flow and preparing the dataset for downstream statistical work.

Initially, the CPRD Aurum data was imported into a Structured Query Language (SQL) relational database management system. SQL was used primarily for data discovery, allowing the team to explore the general structure of the database and experiment with generating sample patient index values. However, due to performance limitations and inefficiencies in handling large volumes of data (N= 14,464,503), SQL was used solely for initial inspection and exploratory purposes.

With the full support of the University of Glasgow's MVLS Advanced Research System (MARS), all subsequent queries and analyses were transitioned to R, a more efficient and flexible environment for processing large-scale datasets. This shift significantly improved operational efficiency, enabling streamlined data manipulation and analysis while maintaining the integrity and security of the data throughout the pipeline (Figure 5.2).

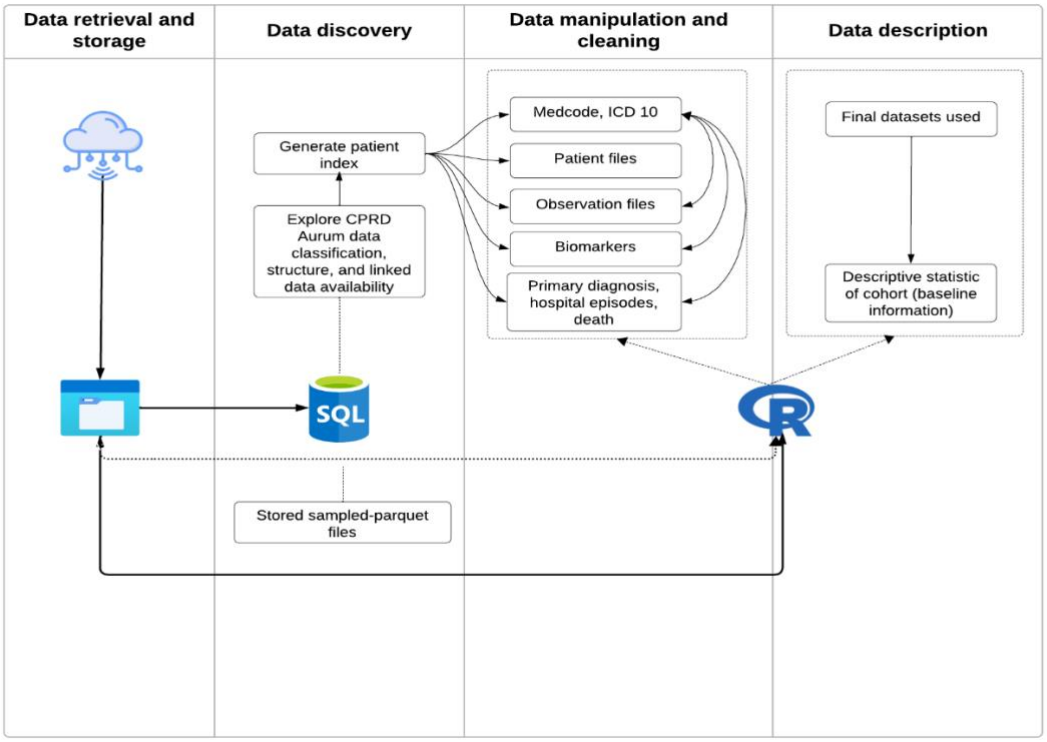


Figure 5.2 Data preparation workflow

The initial data processing was primarily conducted using the Aurum pipeline package in R, developed by Jay Hughes (2022).[222] The `aurum_pipeline ( )` function was executed to process raw electronic health records (EHRs) and store the output in Parquet format. Parquet is a highly efficient columnar storage file format, specifically optimised for handling large-scale datasets. Unlike traditional row-based storage formats (e.g., CSV or JSON), Parquet enables efficient data compression and encoding, significantly reducing storage requirements. By organising data into columns rather than rows, Parquet facilitates vectorised processing, allowing analytical queries to retrieve only the required columns instead of scanning entire rows.[223] This columnar structure enhances performance in big data analytics, making Parquet particularly well-suited for large healthcare datasets like Aurum, where millions of patients generate vast longitudinal clinical records.

Once the data was processed and stored in Parquet, it was loaded into an R data frame for further analysis. To efficiently query and manipulate these large datasets, the ‘Arrow’ package in R was applied. The Apache Arrow project provides a cross-language development platform for high-performance in-

memory computing. It offers a unified interface that allows seamless interaction with multiple data storage formats, including Parquet and csv. The Arrow package enables zero-copy reads, minimising memory overhead when handling large datasets, which is crucial when working with highly granular longitudinal EHR data (e.g., Aurum observation records where a single patient can have thousands of time-stamped clinical encounters over several years). The Arrow R package integrates seamlessly with `dplyr`[224], allowing analysts to write expressive and optimised data transformation pipelines without requiring explicit low-level memory management. This functionality is particularly beneficial for filtering, aggregating, and joining high-dimensional datasets efficiently. The full code implementation for this process is available in Github for reproducibility and transparency.

Additionally, SQL was utilised for a limited subset of data processing tasks, primarily for initial data exploration, generating patient indices, and performing inner joins with linked datasets such as HES via the patient identifier. At an earlier stage of the pipeline, data transformation was partially executed on a high-performance computing (HPC) system before transitioning to an R-based processing workflow. However, depending on system requirements and computational efficiency, analysts may choose to perform all data management within R, Python, or a hybrid approach leveraging SQL servers.

While SQL-based data warehousing solutions (e.g., cloud-based systems such as Google BigQuery, Microsoft SQL Server) provide robust scalability, they introduce additional costs and require ongoing maintenance. Furthermore, when working with sensitive patient-level data, analysts must ensure compliance with institutional governance frameworks and data protection regulations (e.g., GDPR in the UK). The decision to integrate SQL with R or Python should be based on computational efficiency, security, and data governance considerations.

5.5 Cohort identification

5.5.1 Inclusion criteria

As described in Chapter 4, the state transition model has seven states, including disease free, type 2 diabetes (T2DM), myocardial infarction (MI), post-MI, stroke and post-stroke, and death. The basic premise of this structure is to describe and analyse how individuals transition through multiple health states over time.

Given these characteristics, the thesis adopts the term of 'multi-state model' to represent the analytical framework. Unlike traditional analysis, which typically considers a single transition (e.g. from alive to death), a multi-state model accommodates complex disease trajectories, capturing the progression through intermediate health conditions.[167,225] This approach necessitates a structured data management process ensuring that transitions between states are accurately defined and modelled. The statistical method will be comprehensively presented in Chapter 6.

From the total CMD patient population within the CPRD Aurum dataset (N=14,464,503), approximately 10% (n=1,344,338) were eligible for linkage with Hospital Episode Statistics (HES) and mortality data. In this context, eligibility refers to registration with GPs that had provided consent for data linkage to external sources. Patients registered with practices that had not opted into the linkage scheme contribute complete primary care records within CPRD Aurum but cannot be linked to hospital admissions or death records for governance and data-sharing reasons.

Accordingly, linkage eligibility was determined at the practice level rather than the individual patient level, and does not reflect clinical characteristics, disease severity, or differential data completeness. The proportion of linkable patients therefore reflects the extent of practice participation in the CPRD linkage programme during the study period.

For this study, the inclusion criteria encompassed adult patients (≥18 years old) who were registered with a GP between 2000 and 2020, ensuring consistency in clinical coding practices over time. Based on these criteria, the observation

period spans from 1990 to 2020 followed by filtering the plausible data, which is justified by the number of recorded observation years per patient (Figure 5.3).

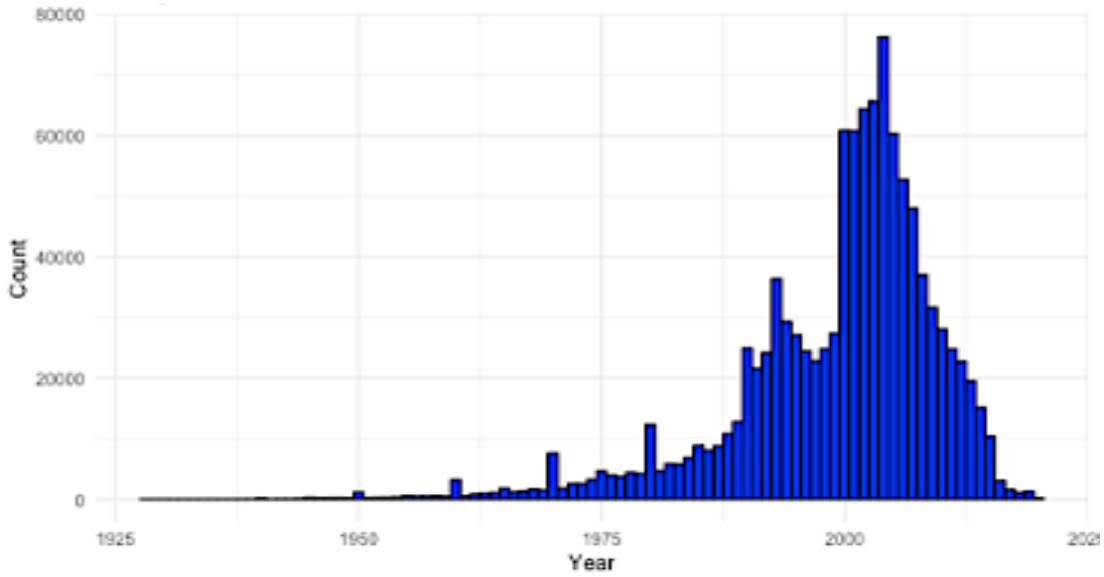


Figure 5.3 Number of patients by observation period

The cardiometabolic “disease-free” state was defined as the absence of any recorded CMD prior to study entry. This classification encompassed individuals with no prior diagnoses of T2DM, MI, coronary heart disease (CHD), stroke, transient ischaemic attack (TIA), cerebrovascular disease (CBVD), or acute coronary events, as determined by ICD-10 codes. Fatal stroke and MI were defined as cases where death occurred within 30 days of the event; otherwise, these were categorised as post-stroke or post-MI events.

The first cardiometabolic event was identified as the earliest recorded diagnosis of T2DM, stroke, or MI, irrespective of whether it was documented as a primary or secondary diagnosis, recorded within a hospital episode, associated with a single hospital admission, or listed as a comorbidity. To ensure a comprehensive assessment of patient trajectories, primary care records were linked with hospital episode data, enabling the identification of transitions between primary and secondary care settings and the determination of first consultations or hospital visits. Events that were not associated with cardiometabolic disease (CMD) were classified as non-CMD events, referring to diagnoses not included within the ICD-10 classifications for CMD-related conditions.

5.5.2 Covariates and clinical events of interest

The covariates (and clinical biomarkers) and clinical events of interest included in this CMD model development are presented in Table 5.4. The medical code (medcode) list and its description were retrieved from various published literature. Furthermore, the medcodeid was re-checked, compared between studies, and summarised as new medcode files that represent current events of interest. For the hospital data (HES), ICD-10 is used to define the clinical event of interests.

Table 5.4 Code source for variables included in the model

Variables	Description	Code	Sources
Covariates	• age, gender, deprivation index, ethnicity • smoking status, alcohol status, • family history (T2DM and CVD), presence of diagnosis (atrial fibrillation, hyperlipidaemia, hypertension • BMI (kg/m^2), blood glucose (mmol/l), blood pressure (mmHg), HbA1C (mmol/l), HDL (mmol/l), LDL (mmol/l), triglycerides (mmol/l), total cholesterol (mmol/l)	medcodeid	Exeter Diabetes Research Team codelists[226], LSHTM Data Compass[227-229], DaRe2THINK[230]
Clinical event (primary care)	T2DM, MI, Stroke	medcodeid	Exeter Diabetes Research Team codelists[226], LSHTM Data Compass[227-229], DaRe2THINK[230]
Clinical event (secondary care)	T2DM	E11, O24.1	ICD-10 dictionary[231]
	MI	I21, I22, I23, 124.1, 125.2	ICD-10 dictionary[231]
	Stroke	I60, I61, I62, I63, I64	ICD-10 dictionary[231]

After filtering by observation period and data filtering, 1,191,377 patients remained, the data underwent filtering to address potential issues related to invalid coding, duplication, and inconsistencies within clinical code lists. At this stage, all invalid and inconsistent data were removed before constructing the final standardized dataset.

Invalid coding in this context refers to patients with discrepant or implausible data entries, particularly inconsistencies between primary care records and hospital data. Examples include erroneous diagnosis entry dates (e.g., a recorded year of 1895), observation dates that match a patient's birth date, making it impossible to determine the precise timing of a diagnosis or procedure, or cases where patients were recorded as deceased but continued to have observations documented after their date of death.

A five-year lookback period was employed to ensure accurate classification of baseline health status by examining clinical event histories prior to cohort entry. This period serves to identify and exclude patients with CMD events. By doing so, the analysis minimises left-censoring bias and ensures that only individuals without CMD enter the model.

Technically (see Github), to get eligible baseline individuals, the `getQualifyingPatients()` function identified eligible individuals aged 18 or older by 1990 and without prior records of CMD events, using general practice and hospital-coded data. The `generateStateTransitionTable()` function then constructed a chronologically ordered dataset of CMD events after 1st January 1990, harmonising formats, integrating death records, and calculating time variables to support longitudinal analyses.

To ensure clinical validity, the filtering functions were also established by removing overlapping MI and stroke events within a single hospital episode and generate only the first diabetes diagnosis per patient. For CVD events, it begins by linking each event in the main event table to the corresponding hospital discharge date, based on a shared ‘spno’ number. It then iterates through each patient’s records, applying logic to retain only the first occurrence of MI or

stroke, and exclude subsequent events of the same type if they fall within the previously recorded hospital discharge window.

The `relabelFirstSecondCVD()` function enhances the model by distinguishing initial from subsequent MI or stroke events, labelling the latter as post-MI or post-stroke to capture disease progression.
The `removeCompetingCVDEvents()` function enforces a hierarchy between MI and stroke. This approach assumes that recurrent events within the same hospital episode likely represent continued care for the same incident, rather than new, independent events. By systematically updating the most recent discharge date and comparing it to the timing of subsequent events, the function ensures that only clinically meaningful, non-overlapping MI and stroke events are preserved. The dataset is then re-ordered chronologically by patient and event date, with recalculated inter-event durations.

5.6 Data cleaning and pre-processing

5.6.1 Complete case and missing data

For the development of the CMD model, a complete case analysis (CCA) approach was employed. The dataset was filtered to include only those observations with complete data for the primary covariates of interest, namely age, sex, cholesterol levels, glucose levels, and blood pressure. CCA, also referred to as listwise deletion, is a widely used statistical method in scenarios where the focus is on analysing observations with no missing values, excluding incomplete cases from the dataset.

As part of this process, the biomarkers data quality was checked and cleaned by applying EHRBiomarkr packages in R developed by Exeter Diabetes Research Team (2023).[226] This package has two main functions, `clean_biomarker_values`, which removes implausible biomarker values, or extreme value ranges, and the `clean_biomarker_units` function retains only biomarker values with appropriate unit codes (numunitid) or those with a missing unit code in CPRD Aurum. Prior to applying this, the relevant biomarker units were checked to make sure all biomarkers are already converted to standard units.

After applying the filtering criteria, a total of 184,845 patients were included as the baseline for model development and cohort construction. All covariates in the dataset were complete; however, one biomarker, HbA1c, exhibited a high proportion of missing data (70%) observed across both male and female populations. HbA1c is a critical indicator of glycaemic control in patients with diabetes and is often associated with a range of health outcomes, making its inclusion potentially valuable for the analysis. Furthermore, when examining HbA1c specifically in the context of T2DM status, the missingness remained notably high, reinforcing the decision to address this issue carefully.

Missing data can indeed be addressed through imputation techniques, and there is no universally defined threshold for the percentage of missingness that necessitates imputation. However, guidance from the literature suggests that missing data exceeding 10-15% generally warrants consideration of imputation methods.[232] In cases where missingness approaches 20-30%, more advanced techniques, such as multiple imputation, are particularly valuable due to their ability to account for the uncertainty associated with missing data. Multiple imputation creates several plausible datasets by replacing missing values with estimates based on observed data, thereby preserving the variability and relationships within the dataset.[232,233]

Given the exceptionally high missingness rate of 70% for HbA1c, the decision was made not to perform multiple imputation due to several technical and methodological concerns. First, imputing more than 70% of the data for a variable means that many of the values would be estimated rather than observed, leading to over-reliance on the imputation model and potentially distorting the true relationships within the data.[234,235] Second, with such a high proportion of missing data, there may not be sufficient observed information to generate accurate imputations, even when using advanced techniques, increasing the risk of implausible or unreliable values.[232,233,236] Third, imputed values may artificially reduce the natural variability in the data, leading to an underestimation of uncertainty and overconfidence in the results. Finally, multiple imputation assumes that the data are Missing at Random (MAR), but with >70% missingness, this assumption is often violated, as the missingness

pattern is more likely to be Missing Not at Random (MNAR).[234,237] In such cases, the missingness may depend on unobserved factors or the true values of HbA1c itself, rendering standard imputation methods inappropriate and potentially biased.

In addition, the model was designed to begin with a disease-free population, rather than being exclusively focused on diabetes patients. Given this scope, the other biomarkers included in the analysis were already deemed sufficient and important for predicting the metabolic profiles, progression, and outcomes of the model.

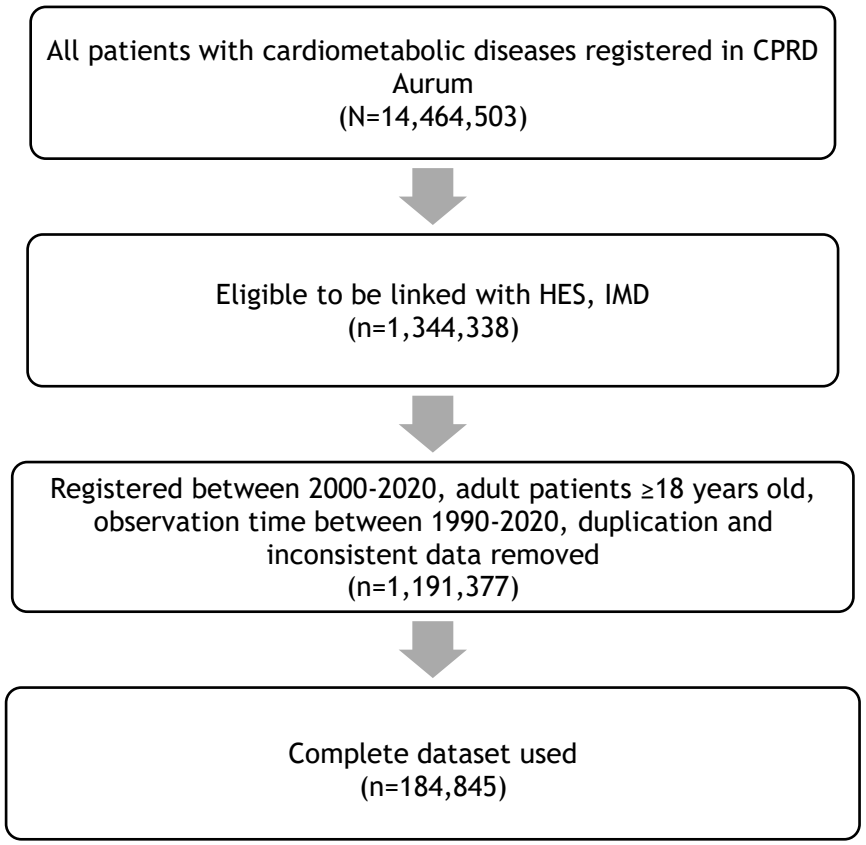


Figure 5.4 Data filtering process

5.6.2 Structuring data for multi-state model analysis

The next step in data preparation is to structure the dataset to be analysis-ready within a multi-state modelling framework. This requires ensuring the dataset captures event times and disease states, which are essential for modelling transitions between health states over time. The method of how the data will be analysed is presented in detail in Chapter 6.

The function `createTransitionMatrix()` constructs a transition matrix that defines the permissible state transitions within the multi-state model. Utilising the *transMat* function from the `mstate` package in R, it encodes the possible clinical pathways a patient may follow through various CMD events. The 'long format' table is compatible with the `mstate` package means that in the long format there are multiple rows for the same 'patid'.

Conversion from long to wide (patid occupies one row, with separate columns for each event and censoring indicator) will depend on the need of further analysis. For instance, wide format will be more straightforward for summary statistic or further single event survival model. Hence, using either long or wide format would be dependent on the analysis undertaken.

Table 5.5 Long format transition table

patid	from	to	Tstart (days)	Tstop (days)	Status	Covariates (1,2…)
001	Disease-free	Diabetes	0	1200	1	xxxxx
001	T2DM	Death	1200	1800	1	xxxxx

*Tstart is the time of entry in the state, and Tstop the event or censoring time, depending on the value of status

Based on this established table (Table 5.5), the matrix can be computed, summarising disease trajectories and number of individuals in each state, this mapping is essential for estimating future transition-specific hazards in survival analysis.

The transition matrix (Table 5.6) defines the allowed transitions between health states in the multi-state model, with each numbered transition corresponding to a specific pathway (e.g., from disease-free to T2DM is transition 1). These transitions reflect the natural progression of CMD and mortality.

Table 5.6 Transition matrix

```
print(transitionMatrix)
events(msdataContinous)
```

from	to Disease-free	T2DM	MI	Post-MI	Stroke	Post-Stroke	Death
Disease-free	NA	1	2	NA	3	NA	4
T2DM	NA	NA	5	NA	6	NA	7
MI	NA	NA	NA	8	NA	NA	9
Post-MI	NA	NA	NA	NA	NA	NA	10
Stroke	NA	NA	NA	NA	NA	11	12
Post-Stroke	NA	NA	NA	NA	NA	NA	13
Death	NA	NA	NA	NA	NA	NA	NA

Labelled as: 1= "Disease-free to T2DM", 2 = "Disease-free to MI", 3 = "Disease-free to Stroke", 4 = "Disease-free to Death",5 = "T2DM to MI", 6 = "T2DM to Stroke",7 = "T2DM to Death",8 = "MI to Post-MI", 9 = "MI to Death",10 = "Post-MI to Death", 11 = "Stroke to Post-Stroke", 12 = "Stroke to Death",13 = "Post-stroke to Death".

The accompanying event frequency table summarises (Table 5.7) the observed number of individuals who experienced each transition, along with the total number entering each state. For instance, among those initially disease-free, 59,226 developed T2DM, 11,806 experienced an MI, 12,505 had a stroke, and 21,355 died without developing any of those conditions. This structure supports a multi-state framework by clearly identifying the sequence of disease progression and allows for the estimation of transition-specific survival models using real-world patient data.

Table 5.7 Summary of events in each states

$Frequencies

from	to Disease-free	T2DM	MI	Post-MI	Stroke	Post-Stroke	Death	Total n event	Total entering
Disease-free	0	59226	11806	0	12505	0	21355	79953	184845
T2DM	0	0	2692	0	2560	0	9676	44298	59226
MI	0	0	0	2217	0	0	4245	8036	14498
Post-MI	0	0	0	0	0	0	912	1305	2217
Stroke	0	0	0	0	0	2988	5970	6107	15065
Post-Stroke	0	0	0	0	0	0	1483	1505	2988
Death	0	0	0	0	0	0	0	43641	43641

5.6.3 Assigning time-dependent covariates

The hazard (risk) of an event often depends not only on baseline values but on the most recent or cumulative values of covariates at the time the event is being assessed.[238] This dynamic relationship underscores the necessity of incorporating time-dependent covariates especially in survival and multi-state models.

There are several reasons for assigning time dependencies before conducting analysis. First, the model is intended to reflect the real-world clinical progression, many risk factors (e.g., biomarkers) evolve over time, and their current values—rather than just baseline measurements—directly influence event risks. This aligns with the previously discussed systematic review recommendation in Chapter 3, where dynamic covariates are integral to accurately modelling disease progression.Second, as an attempt to prevent survivorship bias and misclassification, for instance: if patients must survive 6 months to receive a medication, treating them as "treated" from baseline may underestimate their early risk. It would artificially inflate their apparent survival advantage by ignoring early untreated follow-up time. Third, adjustment for intermediate events is important, particularly in models with transitions between states. In such models, covariates can be updated after each transition, ensuring that the hazard function reflects the patient’s most recent clinical

status. Also, when static models may underestimate variability in risk, time-dependent models adapt to new information.

All covariates in this model are treated as time-dependent, age was computed from full date of birth and updated each interval, meaning the most recent recorded value before the start of each interval (`Tstart`) was used. This ensures that covariate values reflect the patient's most up-to-date state just before entering the event risk window, allowing the model to more accurately capture real-world disease progression.

This approach ensures the model captures the patient's state immediately preceding each event-risk window. The resulting data structure is compatible with standard survival analysis tools (e.g., the survival package in R) and means that it can be seamlessly integrated into Cox models while properly accounting for time-varying effects.

5.7 Cohort characteristics

In a cohort of 184,845 individuals (Table 5.8), comprising 89,645 males and 95,200 females, the median age was 42 years (IQR: 32–52) for both sexes. This exploratory analysis serves as a foundation for identifying transitions where sex differences may influence disease progression.

A family history of diabetes was common in both sexes, reported by 46.4% of males and 42.6% of females, while approximately one-fifth reported a family history of cardiovascular disease (CVD). Around one-third of participants had no reported family history of diabetes or CVD. The cohort was predominantly of White ethnicity (approximately 73% in both sexes), followed by Asian and Black ethnic groups. The ethnic distribution was broadly comparable between males and females, with a slightly higher proportion of females classified as "Other" ethnicity.

In terms of health-related behaviours, most participants reported low-risk alcohol consumption, although abstention was more common among males

(14.2%) than females (7.2%). Females had a higher prevalence of current and former smoking, whereas males were more likely to be never smokers. Socioeconomic status, measured using the Index of Multiple Deprivation (IMD), showed a relatively even distribution across quintiles for both sexes, with a modest concentration in the most deprived quintile (IMD 5).
Regarding baseline clinical conditions, hypertension was highly prevalent, affecting 45.0% of males and 49.3% of females. Hyperlipidaemia was present in around 17% of both sexes, while atrial fibrillation was more common in males than females (6.3% vs 4.6%).

Anthropometric and biochemical risk factors differed by sex. Overweight and obesity (BMI ≥25 kg/m^2) were highly prevalent in both groups, particularly among females. Females also had a higher proportion classified as obese (BMI ≥30 kg/m^2). Males were more likely to have low HDL cholesterol, while females more frequently exhibited higher HDL levels.

Blood pressure measurements indicated a high prevalence of elevated systolic blood pressure (≥140 mmHg) in both sexes, particularly among females. Raised diastolic blood pressure (≥90 mmHg) was more common in females than males. Finally, glycaemic measures showed that the majority of participants had blood glucose levels in the prediabetes range (5.5–7.0 mmol/L), with a higher proportion of females meeting criteria for diabetes (≥7.0 mmol/L). Triglyceride levels were generally lower among females, whereas males were more likely to have elevated triglycerides.

The categorisation of baseline characteristics in Table 5.8 was based on established clinical thresholds and common epidemiological conventions to support interpretability and comparability. Age, blood pressure, body mass index, lipid fractions, and glycaemic measures were grouped using clinically recognised cut-points rather than data-driven thresholds.[239,240] A single set of HDL cholesterol categories was applied across sexes to ensure consistency and facilitate direct comparison, with sex-specific differences reflected through stratified reporting. All categorical distributions are presented as column percentages within sex, in line with conventional practice.

Table 5.8 Baseline characteristics of study population by sex

	Male (n= 89,645)		Female (n= 95,200)	
	Median (Q1-Q3)	n (%)	Median (Q1-Q3)	n (%)
Age (years), all	42 (32-52)			
18-25		8,146 (9.08)		9,761 (10.25)
25-34		16,235 (18.11)		20,701 (21.75)
35-44		20,053 (22.37)		24,520 (25.76)
45-54		21,347 (23.81)		22,586 (23.72)
55-64		16,721 (18.65)		13,775 (14.47)
>65		7,143 (7.98)		3,857 (4.05)
Family history				
Diabetes		41,562 (46.36)		40.583 (42.63)
CVD		18,030 (20.11)		19,470 (20.45)
No family history		30,053 (33.53)		35,147 (36.92)
Ethnicity				
White		65,229 (72.76)		69,405 (72.91)
Asian		6,550 (7.31)		6,836 (7.18)
Black		4,456 (4.97)		3,536 (3.71)
Mixed		582 (0.65)		527 (0.55)
Other		12,828 (14.31)		14,896 (15.65)
Alcohol consumption				
Level 0		12,703 (14.17)		6,815 (7.16)
Level 1		73,809 (82.34)		81,172 (85.26)
Level 2		1,140 (1.27)		3,339 (3.50)
Not reported		1,018 (1.14)		837 (0.88)
Missing		975 (1.08)		3,037 (3.20)
Smoking status				
Active smoker		20,423 (22.78)		26,277 (27.60)
Ex smoker		19,114 (21.32)		30,477 (32.03)
Non-smoker		50,108 (55.90)		38,446 (40.38)
Deprivation index				
IMD 1		15,007 (16.74)		17,715 (18.61)
IMD 2		16,749 (18.68)		18,553 (19.48)
IMD 3		17,133 (19.11)		18,344 (19.27)
IMD 4		19,118 (21.33)		19,253 (20.23)
IMD 5		21,638 (24.14)		21,235 (22.31)
Presence diagnosis				
Atrial fibrillation		5,612 (6.26)		4,374 (4.59)
Hyperlipidaemia		15,811 (17.64)		15,944 (16.75)
Hypertension		40,343 (45)		46,885 (49.25)
BMI (kg/m2)				
<18.5	17.60 (16.80-18.10)	1,233 (1.45)	17.67 (17.00—18.13)	527 (0.55)
18.5-24.9	22.87 (20.13-23.37)	20,608 (52.00)	23.30 (22.00-24.20)	18,205 (19.13)
25-29.9	27.37 (26.15-28.60)	29,443 (32.84)	27.40 (26.20-28.60)	40,224 (42.25)
≥ 30	20.75 (31.88-38.22)	38, 361 (42.81)	19.61 (31.31-36.20)	36,244 (38.07)
HDL (mmol/l)				
<1.03	0.95 (0.89-1.00)	7,603 (8.48)	0.92 (0.85-0.99)	24,466 (25.70)
1.03-1.54	1.30 (1.20-1.41)	45,091 (50.30)	1.23 (1.13-1.36)	55,778 (58.60)
≥1.55	1.80 (1.66-2.05)	36,951 (41.22)	1.74 (1.62-1.95)	14,956 (15.70)
LDL (mmol/l)				
<2.6	2.20 (1.90-2.43)	24,339 (27.15)	2.17 (1.83-2.40)	31,175 (32.75)
2.6-3.3	2.99 (2.80-3.15)	27,091 (30.22)	2.97 (2.80-3.14)	29,208 (30.70)
3.4-3.9	3.60 (3.45-3.74)	20,111 (22.43)	3.60 (345-3.74)	19,285 (20.25)
4.0-4.9	4.27 (4.09-4.52)	15,528 (17.33)	4.27 (4.09-4.52)	13,668 (14.35)
≥4.9	5.38 (5.17-5.70)	2,576 (2.87)	5.33 (5.15-5.67)	1,864 (1.95)

	Male (n= 89,645)		Female (n= 95,200)	
	Median (Q1-Q3)	n (%)	Median (Q1-Q3)	n (%)
SBP (mmHg)				
<120	115.13 (110.85-118.25)	6,721 (7.50)	115.83 (111.50-118.56)	6,415 (6.74)
120-140	132.75 (128.0-136.62)	39,042 (43.55)	132.88 (128.08-136.67)	41,120 (43.20)
≥ 140	148.75 (144.12-155.00)	43,882 (48.95)	148.67 (144.12-155.00)	47,665 (50.06)
DBP (mmHg)				
<80	75.35 (71.78-78.00)	38,526 (42.97)	75.50 (71.75-78.00)	35,418 (37.20)
80>90	84.44 (82.22-86.93)	38,913 (43.41)	84.46 (82.50-87.29)	40,623 (42.67)
≥ 90	93.22 (91.44-96.08)	12,206 (13.62)	94.00 (91.82-97.30)	19,159 (20.13)
Triglycerides(mmol/l)				
<1.7	1.15 (0.90-1.40)	56,243 (32.38)	1.18 (0.92-1.42)	56,243 (59.10)
1.7-2.2	1.92 (1.80-2.06)	16,626 (8.11)	1.93 (1.80-2.06)	16,626 (17.45)
2.3-5.6	2.70 (2.40-2.18)	14,633 (7.92)	2.80 (2.45-3.38)	21,658 (22.75)
≥ 5.6	6.40 (5.95-7.55)	167 (0.09)	6.60 (6.00-7.92)	673 (0.70)
Blood glucose (mmol/l)				
<5.4	4.70 (4.50-4.90)	35,003 (39.04)	4.75 (4.53-4.90)	28,466 (29.90)
5.5-7	5.58 (5.27-6.06)	45,479 (50.73)	5.60 (5.30-6.10)	54,043 (56.77)
≥7	8.10 (7.40-10.10)	9,163 (10.23)	8.33 (7.45-10.90)	12,700 (12.33)
Total cholesterol (mmol/l)				
<5.0	4.51 (4.15-4.80)	32,004 (35.70)	4.40 (3.97-4.72)	46,459 (48.80)
5-5.5	5.27 (5.14-5.40)	18,211 (20.31)	5.25 (5.13-5.40)	18,519 (19,45)
5.6-6	5.75 (5.63-5.89)	16,310 (23.48)	5.75 (5.63-5.89)	14,073 (14.80)
≥ 6	6.52 (6.24-6.96)	23,120 (25.90)	6.48 (6.20-6.90)	16,149 (17.25)

IMD: index of multiple deprivation, BMI: body mass index, HDL/LDL: high/low density lipoprotein, SBP/DBP: systolic/diastolic blood pressure.
Note: Alcohol consumption level (0: abstinence, non-drinker, 1: moderate drinking ≤ 14 units/week, 2: ≥ 14 units/week), smoking status (active: ≥1 cigarette/day, includes e-cigarettes, ex: Stopped ≥6 months ago, non: Never smoked or <100 lifetime cigarettes).

5.8 Discussion and conclusions

Data preparation is a critical part of real-world data research, often accounting for 50–80% of the analytical workload.[241] This section outlines the key steps taken to prepare CPRD Aurum data for the cardiometabolic disease (CMD) policy model, capturing both technical and conceptual aspects. Documenting this process supports transparency, reproducibility, and methodological rigour. In the context of growing reliance on real-world evidence and open science practices, such documentation also contributes to broader efforts in promoting accountability and openness in health research.[242,243]

In addition to the computational demands, this process highlights several challenges in preparing CPRD Aurum data. The relational database structure requires careful merging, especially for time-to-event analyses, so that the data accurately captures the trajectory and hierarchy of events. Additionally, Aurum uses different clinical coding systems compared to CPRD GOLD; therefore, using validated code lists is strongly recommended. As mentioned earlier, there is also a high likelihood of implausible data entries, so double-checking data quality can be beneficial before proceeding with further data manipulation.

This data preparation documentation addresses key challenges in large-scale data analysis through several practical advantages. First, adopting the Parquet file format enhances processing speed and scalability, particularly for high-volume datasets. Parquet files in R demonstrate superior computational efficiency and reduced storage demands compared to traditional formats such as SAS, especially when managing large health records.[223,244] Second, linking and extracting data from both primary and secondary care sources enables a more complete representation of disease trajectories and cohort development, also by building new R functions replication is possible when preparing the analysis-ready dataset, particularly for multi-state modelling. Third, applying established updated R functions available for cleaning covariates and biomarkers ensures a reproducible and streamlined workflow, eliminating the need for manual adjustments.[222,226]

It is important to recognise that data preparation methods can vary significantly depending on research objectives, prior expertise, and the availability of data infrastructure. Furthermore, as computational and data management techniques continue to evolve through advancements in automation, function development, and other innovations, thus there is considerable potential to refine and improve the efficiency of data preparation processes. While the methods described here were tailored to the specific needs of this study, alternative approaches may offer greater efficiency, particularly when handling large-scale and complex datasets.

Chapter 6 Developing a Cardiometabolic Disease Policy Model

6.1 Introduction

Chapter 6 presents the development of a Cardiometabolic Disease (CMD) Policy Model to support further long-term epidemiological analysis and health economic analysis. Building on the conceptual framework outlined in Chapter 4 and the data sources (Clinical Practice Research Datalink (CPRD) Aurum and linked datasets) described in Chapter 5, this chapter applies time-to-event multi-state survival analyses to model transitions between cardiometabolic health states.

Sections 6.2 and 6.3 outline the statistical and modelling principles underpinning the CMD Policy Model, including the rationale for using survival analysis, the structure and logic of the multi-state framework, and the time-dependency assumptions that govern disease progression. The CMD Policy Model captures progression across seven health states from disease-free to type 2 diabetes mellitus (T2DM), myocardial infarction (MI), stroke, and death, via 13 clinically defined transitions.

To reflect the complexity of disease pathways and enable extrapolation, a range of survival modelling approaches is explored, including non-parametric (e.g., Kaplan-Meier), semi-parametric (e.g., Cox regression), standard parametric (e.g., exponential, Weibull), and flexible parametric (e.g., Royston-Parmar spline) models. These methods are embedded within a multi-state framework and extended using a semi-Markov structure. The model results and the final CMD Policy Model are presented in sections 6.4 and 6.5, respectively.

6.2 Overview of modelling strategy

6.2.1 Rationale for survival analysis

Definition

Survival data has been widely applied across various disciplines, particularly in medicine and public health. Also known as time-to-event (TTE) data, it provides information not only about whether an event of interest (e.g., death) has occurred, but also on how the 'hazard' of when it occurs changes over time.

Survival analysis is designed to handle TTE data, particularly in situations where traditional regression methods (e.g., logistic or linear regression) are inadequate due to censoring or the time-dependent nature of outcomes." [245,246] While conventional regression methods can assess the relationship between risk factors and the occurrence of an event, they are not well-suited for scenarios where some subjects do not experience the event within the study period.[247] Survival analysis addresses this limitation by properly accounting for censored observations and enabling estimation of the timing and risk of events over time.[245,246]

Typically, survival analyses have several objectives: 1) estimate, interpret, or compare survival and/hazard functions over time 2) identify and assess predictor (explanatory variables) of survival time (in proportional hazard assumptions) 3) handle censored data appropriately. These will be briefly discussed in following sections.

Censoring

A fundamental concept in survival analysis is censoring, which occurs when the exact time of the event of interest is not observed for some individuals. This typically arises when participants do not experience the event within the study's observation period. In such cases, it is known that the event has not yet occurred, but it remains uncertain if or when it will happen in the future.[248] Survival data analysis becomes methodologically challenging due to the variation in follow-up times across participants and the presence of censored

observations, which must be properly accounted for to avoid biased estimates of event timing and risk.

There are several reasons why censoring occurs in survival analysis. First, the study may conclude before all participants experience the event of interest, leaving their full survival times unknown. Second, some participants may drop out or be lost to follow-up, resulting in incomplete outcome data. Third, participants may withdraw consent, relocate, or experience a different event that prevents further data collection.[249-251] Additionally, individuals who have not experienced the event of interest by the end of the follow-up period are treated as right-censored, as their event time remains unknown but may occur in the future.[252]

Censoring can be categorised into three main types: right censoring, left censoring, and interval censoring.[248] Right censoring occurs when a subject's follow-up ends before they experience the event of interest (e.g., death, disease onset), this is by far the most common of censoring. The situation when a study terminates before all participants have died, those who are still alive at the end of the study are considered right censored.[253] For instance, if a participant(s) in a 10-year cohort study does not develop diabetes by the end of the follow-up period. The study ends in 2030, and some individuals remain diabetes-free until then. Since we do not know if they develop diabetes after 2030, their data is right censored in 2030.

Interval censoring arises when the exact time of the event is unknown, but it is known to have occurred within a specific time range.[248] For example, consider the case of atrial fibrillation (AF): older adults might undergo routine electrocardiograms (ECGs) annually to monitor heart health, and then a patient who had a normal ECG in January 2020, missed their 2021 follow-up, and was then found to have AF in January 2022. Since the patient was known to be free of AF in 2020 but had a confirmed diagnosis by 2022, the actual onset occurred at some unknown point within that two-year window.

The less common censoring, left censoring, happens when the event of interest has occurred before the study begins, but the exact timing is unknown..[248] This

means that the event time is only known to be less than a certain value, rather than observed precisely.[253] For example, if a participant is enrolled in a study while already hospitalised with Covid-19, but the exact date of infection is unknown, the data is considered left-censored. In this case, the Covid-19 infection clearly occurred before study entry, yet the precise timing remains unobserved. Dey et al (2020)[253] illustrated clearly these different types of censoring (the illustration is reproduced in Figure 6.1)

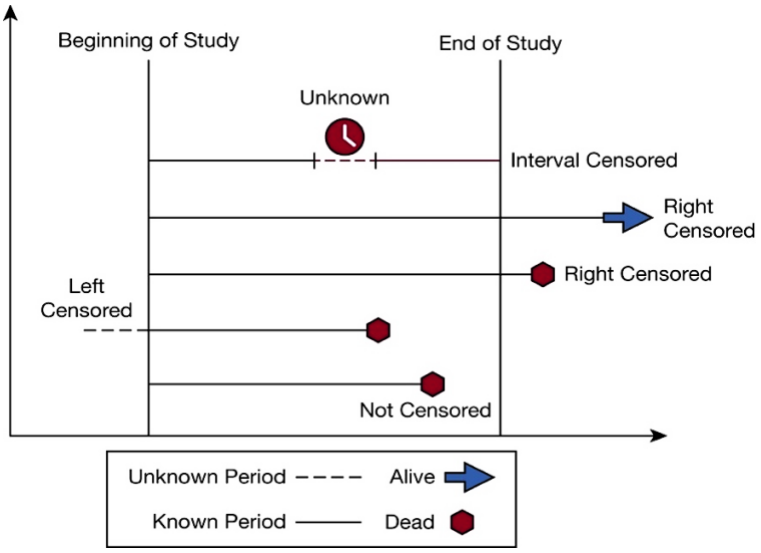


Figure 6.1 Illustration of three types of censoring

From those examples if the participant was censored simply because the study ended, it means the censoring is unrelated to the underlying risk of the events which is defined as 'non-informative' censoring.[245,249] However, there is also a possible condition where participants with worsening health are more likely to drop out, and that health decline is associated with higher event risk, then their censoring can be potentially 'informative', which can bias survival estimates if not properly handled.[250]

Finally, censoring is an unavoidable and natural feature of survival data in longitudinal research. Rather than excluding these observations, survival analysis methods are designed to incorporate them appropriately to maintain the validity and power of the study.

Hazard function and survival function

Two fundamental and related concepts in survival analysis are: survival function denoted by $S(t)$ as the probability of individual surviving to at least a certain time, and Hazard function $h(t)$ describes the instantaneous risk of the event occurring at time t, given that the individual has survived up to that time.[189,246]

Mathematically, *S(t)* is expressed as:

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

(Equation 6.1)

T is a non-negative random variable representing the time until the occurrence of the event $(T \geq 0)$. At the beginning of observation $(t = 0)$, everyone is alive $S(0) = 1$. As time progresses, $S(t)$, when t=∞ decreases because some people will experience and event, so $S(t)$ is always non-increasing and will fall to 0 when eventually nobody survives the event.

In theory, the survival $S(t)$ is a smooth curve, but when we estimate with data/models it often looks like step function (Figure 6.2). It means that the event of interest occurs in specific discrete time points, not continuously. It means $S(t)$ 'drops' when an event occurs creating a step pattern.[189]

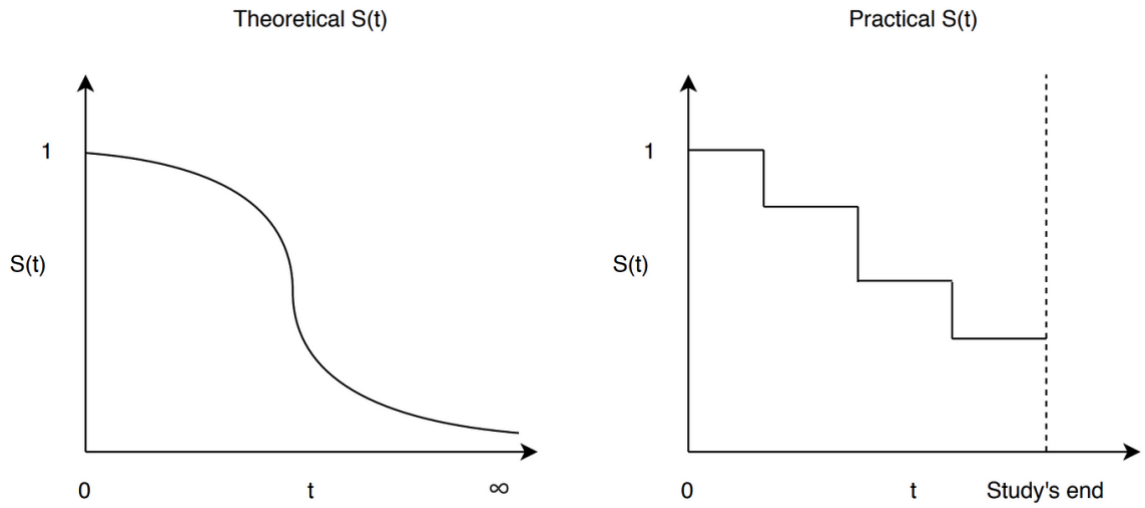


Figure 6.2 Theoretical versus practical survival curve

In contrast to the survivor function $S(t)$, the hazard function $h(t)$ focuses on the risk that the event will occur. This function describes the instantaneous rate at which an event (such as failure, death, or relapse) is expected to occur at time t, given that the individual has survived up to that time.[189,246]

To clarify the often challenging concept of the hazard function, Kleinbaum (2012)[189] uses the analogy of a car's speedometer. When a driver sees a reading of 60 mph (miles per hour), it does not mean that the driver will travel 60 miles in the next hour. Rather, it shows the instantaneous speed at that specific moment. Driver might speed up, slow down, or even stop altogether in the next hour, but at the point the driver glances at the gauge, 60 mph is the current speed.

In the same way, the hazard function provides an instantaneous rate at which an event might occur at time t, assuming the individual has survived up to that time. It does not predict whether the event will occur at that moment, just as the speedometer does not predict how far the car will travel. Instead, it gives the rate at which the event is likely to happen if current conditions continue. Mathematically, $h(t)$ is represented as follows.

$$h(t) = \lim_{dt \to 0} \frac{\Pr\,(t < T \leq t + dt | T > t)}{dt} = \frac{f(t)}{S(t)}$$

(Equation 6.2)

This equation illustrates that as the time interval approaches zero $\lim_{dt \to 0}$, the instantaneous rate is obtained, given that the individual has survived up to that point. It is important to note that the hazard function is not a 'probability', as it involves dividing a probability by a time interval, resulting in a rate rather than a probability value between 0 and 1.[245] Because it is conditional on survival up to time t, the hazard function is often referred to as the conditional failure rate.

This can also be rewritten in terms of a probability density function $f(t)$ and the survival function $S(t)$, and can use derivative of the survivor function as[246]:

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d \ln (S(t))}{d(t)}$$

(Equation 6.3)

The following formula indicates the relationship between $S(t)$, $h(t)$, and $H(t)$, with $H(t)$ being the cumulative hazard, representing the accumulated risk up to time t.

$$H(t) = -\int_0^t h(u)\, du$$

$$h(t) = -\frac{d\ ln\ (S(t))}{d(t)}$$

$$H(t) = -ln\ (S(t))$$

$$S(t) = e^{-H(x)}$$

(Equation 6.4)

In summary, the equation above shows that the survival function decreases exponentially as the cumulative hazard increases. When cumulative hazard is low, the survival probability remains high; as cumulative hazard increase over time, survival drops more steeply.

While the formula above may not be essential for routine data analysis since the statistical software can easily compute transformations between functions—it is important to understand the conceptual relationship between them, as they are mathematically linked.

6.2.2 Multi-state framework and its suitability for CMD Policy Model

Based on the explanation in previous sub-sections, standard survival models are effective for analysing time to a single event (e.g., death). However, they are often inadequate for chronic and progressive conditions like cardiometabolic disease (CMD). CMD typically involves multiple intermediate events, such as the

development of type 2 diabetes mellitus (T2DM), myocardial infarction (MI), and stroke, before reaching a death event. Standard survival models do not account for the order, timing, or recurrence of such events, nor can they model transitions between intermediate disease states. In summary, standard survival modelling is too restrictive to analyse overall complex CMD processes (Chapter 4).

In contrast, multi-state models (MSMs) provide a more flexible and clinically realistic framework by allowing individuals to transition between multiple defined health states over time. This enables the estimation of transition-specific hazards and accommodates time-dependent covariates, making them more appropriate for capturing the complexity of CMD progression.[254–256] Traditional survival analysis, can be viewed as a simple form of a MSM, for example, modelling the transition from being 'alive to dead' only (Figure 6.3a).

Before delving deeper, it is important to clarify that MSMs serve as an umbrella framework in survival analysis, capable of representing a wide range of disease or life-course processes through transitions between well-defined states. These states may include intermediate stages (such as disease onset or recovery) and absorbing states (such as death). This clarification helps prevent confusion, particularly around the concept of competing risks (will be discussed later). While competing risks can appear conceptually similar to multi-state models, they are best understood as a special case within the broader multi-state framework. Although their definitions and applications may sometimes overlap, they differ in both structure and analytical scope.[257–260]

Therneau et al. (2024)[261] presented a series of diagrams illustrating MSMs (Figure 6.3). Each diagram offers a different perspective depending on the structure of the available dataset and the specific research questions being addressed.

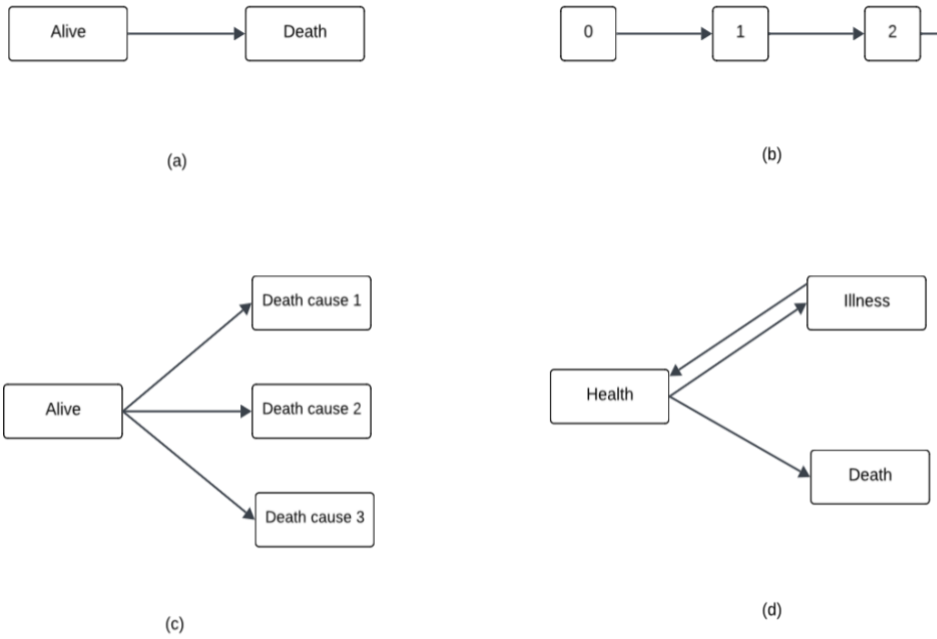


Figure 6.3 Four multi-state models

Figure 6.3 (a) depicts the simplest survival model, where an individual begins in the "alive" state and transitions to the "dead" state; this reflects the traditional survival analysis framework. Figure 6.3 (b) illustrating ordered, repeated, or progressive events, such as stages of disease or increasing severity, with state 0 representing study entry and subsequent states indicating sequential transitions. Figure 6.3 (c) represents a typical competing risks scenario, in which an individual starts in a single initial state (e.g., "alive") and may transition to one of several mutually exclusive terminal states (e.g., different causes of death), with no further transitions possible. Lastly, Figure 6.3 (d) presents the well-known illness-death model, characterised by transitions between "health" and "illness", allowing for recovery or relapse, before reaching the absorbing state of "death." This model is particularly relevant in the study of chronic diseases and long-term prognoses.

Understanding the framework of multi-state models (MSMs) allows for diverse perspectives and definitions in both estimation and modelling. MSMs provide a flexible approach to estimating various measures of interest, such as the probability of being in a specific state (or set of states) over time, the likelihood of transitioning between states, the expected duration spent in a particular

state, the probability of ever entering a given state, and transition-specific hazard rates or hazard ratios.

A notable key feature within this MSM framework is the ability to account for competing risks. Competing risks commonly arise in clinical studies using time-to-event (TTE) data when multiple potential outcomes can occur, and the occurrence of one event precludes the occurrence of others. For example, in the context of cardiovascular disease (CVD), a patient may die due to stroke, myocardial infarction (MI), or another cause.

A paper by Putter et al. (2007)[260] outlines several scenarios involving competing risks. The classical competing risks framework, as illustrated in Figure 6.7 (c), describes situations where multiple mutually exclusive events can occur, and the occurrence of one precludes the others. Another important scenario is the illness-death model, in this setting, an individual may develop an illness (e.g., diabetes) and subsequently die. Death competes with illness when analysing ‘time to illness’, since death may occur before the illness develops. However, illness does not compete with death, as death can still occur afterward. This asymmetry is referred to as "semi-competing."[260]

Putter et al. (2007)[260] also describe scenarios where a non-fatal event is the primary outcome of interest and death acts as a competing risk. The other possible scenario, for example death may be considered a competing event when the event of interest is hospitalisation. These examples highlight that competing risks either fully prevent the occurrence of the event of interest or simply preclude it from occurring first, which is crucial in selecting an appropriate modelling approach.[258,260]

Several foundational studies across various disease areas have demonstrated that ignoring competing risks can significantly distort both the estimation of survival probabilities and the interpretation of covariate effects.[262–266] Those emphasised that relying on Kaplan–Meier estimates and standard Cox models (within a simple alive–dead framework) can lead to inflated survival probabilities. It tends to overestimate cumulative incidence in long-term studies because they treat

competing events as censored, thus assuming individuals remain at risk indefinitely for the event of interest.

The consequences of neglecting competing risks goes beyond statistical inaccuracy; they carry direct clinical implications. When treatment efficacy is evaluated without considering competing risks, researchers or clinicians may be misled about patients' true survival prognosis.[257,259,260,267,268] Having awareness about the competing risk scenario then can be helpful to decide which statistical model or scenario aligns with study objectives. The statistical approach to handle this condition will be presented in the following section.

6.2.3 State-transition model structure

The state-transition structure of the CMD Policy Model comprising seven disease states and thirteen clinically plausible transitions was first introduced in Chapter 4 (Figure 4.4). That conceptual model provides the foundation for the statistical multi-state modelling presented in this chapter.

Each of the thirteen transitions is treated as a distinct time-to-event (TTE) process and was analysed using appropriate survival modelling approaches. These methods enable the estimation of transition-specific hazards, the inclusion of relevant covariates, time-dependent effects, and support for extrapolation beyond the observed follow-up period.

For clarity, the full state-transition structure is recalled from Chapter 4, along with corresponding transition labels used in subsequent sections, and is summarised in Figure 6.4. The labelling here serves to map transition movements and to inform subsequent statistical analyses.

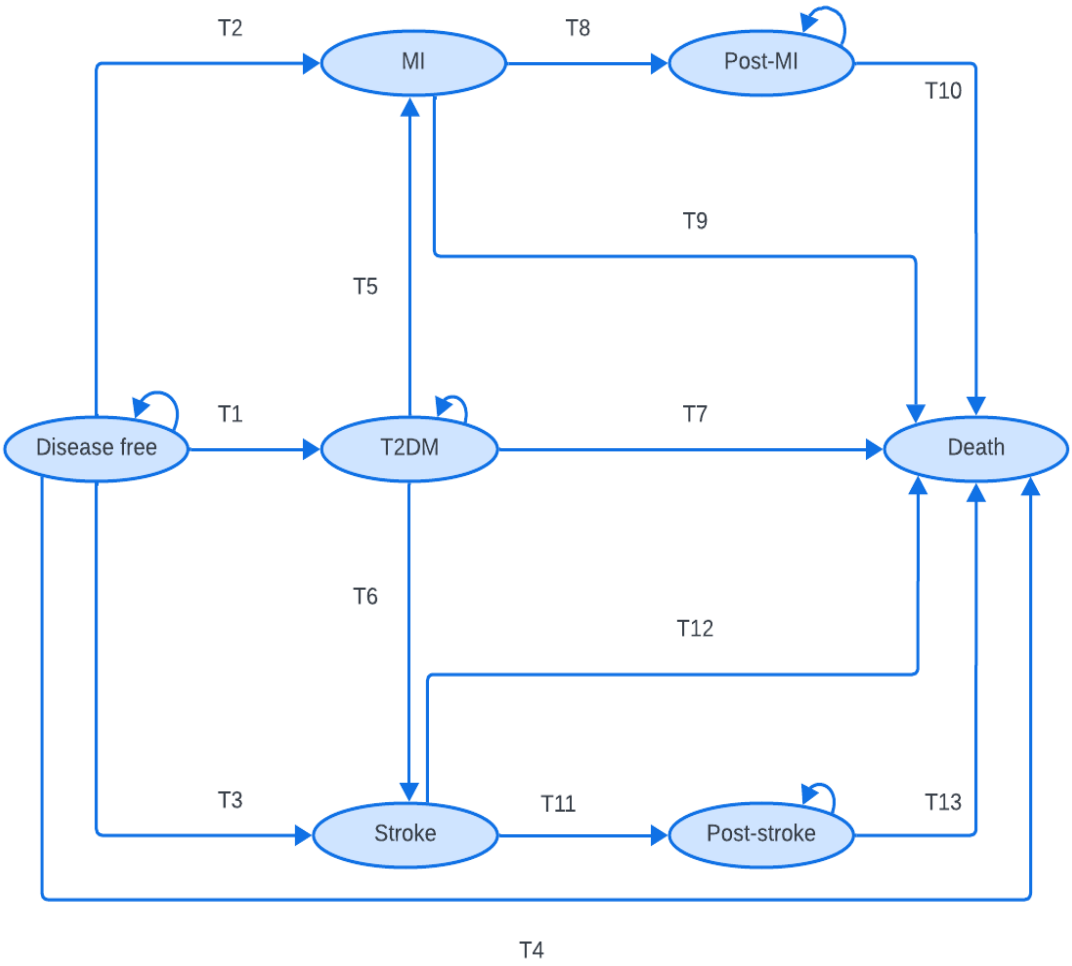


Figure 6.4 State transition model (labelled)

Note: T1 = "Disease-free to T2DM", T2 = "Disease-free to MI", T3 = "Disease-free to Stroke", T4 = "Disease-free to Death",T5 = "T2DM to MI", T6 = "T2DM to Stroke", T7 = "T2DM to Death", T8 = "MI to Post-MI", T9= "MI to Death", T10 = "Post-MI to Death", T11 = "Stroke to Post-Stroke", T12 = "Stroke to Death", T13 = "Post-stroke to Death"

The survival framework outlined here sets the foundation for modelling time-to-event outcomes across multiple disease states. This approach is extended through a multi-state structure to reflect the complexity and progression of cardiometabolic disease. The implementation depends on how transition risks are assumed to evolve over time, particularly in relation to Markov versus semi-Markov assumptions, which have been justified in the conceptual model (Chapter 4). The rationale and statistical approaches for survival analysis and Markov assumptions are detailed in Section 6.3.

6.3 Statistical method for multi-state survival analysis model

6.3.1 Non-parametric model

Kaplan-Meier estimator

The Kaplan-Meier (KM) is a non-parametric technique used to estimate the survival function when the time to an event varies across individuals.[245,269] Introduced in a 1958 paper by Edward L. Kaplan and Paul Meier, this method provided a practical solution for handling censored data.[270] The KM estimator, also known as the 'product-limit' method has since become the most common technique used for survival analysis.[271]

The central idea of KM is that time intervals are not predetermined but are instead based on the actual occurrence of events. This allows for a stepwise construction of survival curve, with the probability of survival recalculated each event time and the curve remaining constant between events.[270] Then the estimator adjusts for censoring by appropriately modifying the number of individuals at risk at each time point. For example, the primary endpoint over 1 year period of a cohort study is stroke occurrence, if individuals experience more than one non-fatal stroke (after 5 months, and after 10 months), the KM only include the first occurrence of stroke, which is at 5 months. However, does not accommodate multiple or recurrent events unless extended models are applied.

Under the assumption that events happen independently, the overall survival probability at any time point can be estimated by multiplying the conditional probabilities of surviving each interval. Specifically, the survival probability at time t_i, denoted $\hat{S}(t_i)$ is derived from the survival probability at the previous event time $\hat{S}(t_{i-1})$, the number of individuals at risk just before time t_i, (denoted n_i), and the number of events occurring at that time (d_i). The KM estimator uses the formula[271]:

$$\hat{S}(t_i) = \hat{S}(t_{i-1})\left(1 - \frac{d_i}{n_i}\right)$$

(Equation 6.5)

with $S(0) = 1$ as the starting condition. Between observed event times, the survival probability remains unchanged, resulting in a characteristic stepwise curve. This method ensures that each participant contributes survival information up to the point of event or censoring. In the absence of censoring, the survival estimate simplifies to the proportion of individuals who remain event-free at each time point.

One of the strengths of the KM method is that it does not require assumptions about the underlying survival time distribution.[245,269,270] This makes KM flexible for analysing data where distributional assumptions may not hold. Also, the KM handles censored data efficiently by allowing the inclusion of individuals who are lost to follow up or have not yet experienced the event by the end of study.[269] The generation of KM survival curves provide a visualisation of survival probability over time and is useful for estimating metrics such as median survival.[272]

Figure 6.5 illustrates the example of a hypothetical KM survival curve (with confidence interval) between treatment and control group over specific time. It illustrates that the treatment group has higher survival probability compared to a control group. For example, 60% of individuals are still event-free at the 10 months in treatment group, while only 45% event-free individuals in control group.

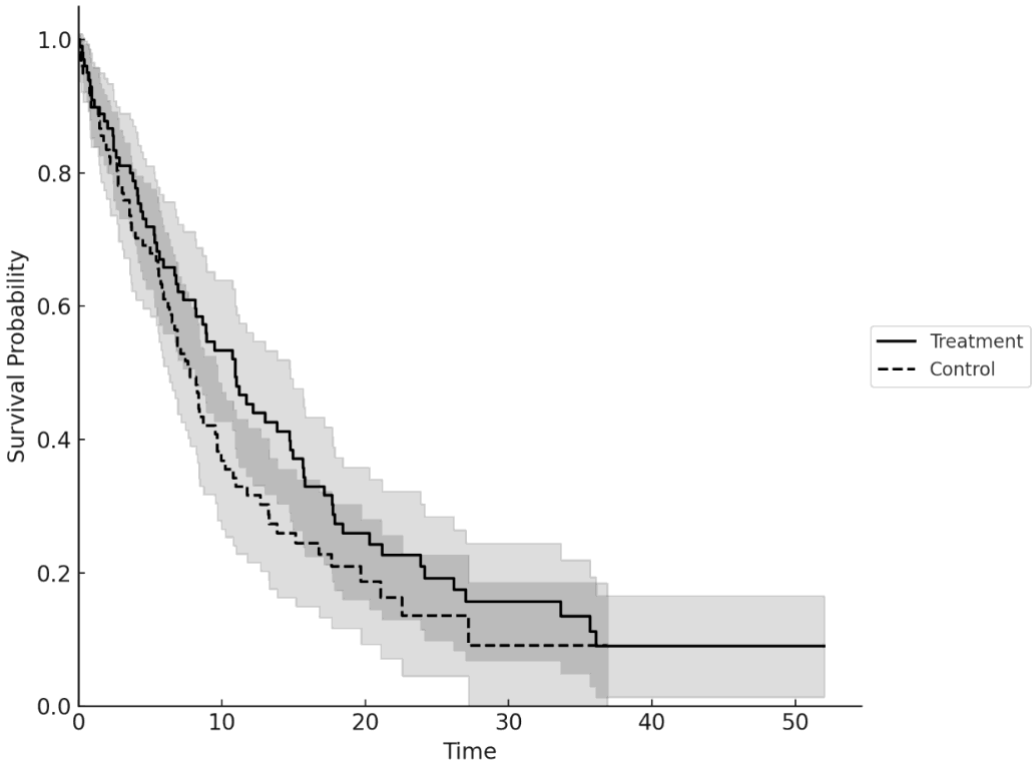


Figure 6.5 Kaplan-Meier survival probability curve

The KM approach has several limitations. Akey limitation is its inability to adjust for covariates, which may limit studies aiming to measure causal effects or account for prognostic differences between groups.[270]

Furthermore, KM assumes non-informative censoring, which may not always be realistic. If censoring is related to the probability of the event, survival estimates may be biased.[269,273] KM also assumes that there are no competing risks, meaning that, meaning that the event of interest is the only possible outcome.[189] If another event, such as death from an unrelated cause, prevents the main event from happening, the KM estimates may not reflect the true survival experience.[274] Lastly, is that KM curves become less reliable with small sample sizes or heavy censoring, as the number of individuals at risk decreases over time.[272]

Log-rank test

Often paired with KM, the Log-rank test is a non-parametric test that provides a formal statistical comparison between groups, by comparing the observed and expected number of events in each group at each time point under the null hypothesis. If the observed differences are large enough, it is indicating a statistically significant difference in survival.[189,245,275]

At each time an event occurs the method calculates the observed number of events in each group and compares it to the expected number under the null hypothesis of no group difference. Conceptually, these expected values are summed over all event times to obtain the total expected number of events for each group, denoted as E_i. The actual number of events observed in each group is called O_i. The log-rank test then compares the observed (O_i) and expected (E_i) values using a test statistic that follows a chi-square (X^2) distribution. This allows calculation of a p-value to determine whether the differences in survival across the groups are statistically significant.[245]

$$X^2 = \sum_{i=1}^{g} \frac{(O_i - E_i)^2}{E_i}$$

(Equation 6.6)

If only two groups are compared, the Log-rank test is assessing the null hypothesis, whether the ratio of the hazard rates in the two group (hazard ratio) is equal to 1. The ratio $^{O_i}/_{E_i}$ represents the estimated relative (or excess) hazard in group i. A hazard ratio (HR) of 1 indicates no difference in survival between the groups, while values above or below 1 suggest higher or lower risk, respectively.

$$HR = \frac{O_1/E_1}{O_2/E_2}$$

(Equation 6.7)

The Log-rank has several important limitations, despite its non-parametric strengths. One major drawback is that it does not adjust for covariates, making it less suitable when multiple risk factors need to be considered simultaneously[189,276,277] Additionally, because this test is based on rankings rather than raw data, it is also less sensitive to extreme values that might otherwise influence the analysis.[278]

While the Log-rank test is the most commonly used non-parametric method for comparing survival curves, several alternative non-parametric tests exist. These include the Wilcoxon (Breslow) test[278], which places greater emphasis on early survival differences; the Fleming-Harrington test[279], which allows for flexible, weight-based comparisons across the survival curve; and the Tarone-Ware test, which offers a balanced weighting approach between early and late events.[276] However, this thesis limits its scope to the use of the KM estimator in conjunction with the log-rank test, as it represents the most widely accepted and applied method in non-parametric survival analysis.

Nelson-Aalen estimator

Another non-parametric method is the Nelson-Aalen (NA) estimator, a method used to estimate the cumulative hazard function $H(t)$ in survival analysis.[280] It is especially helpful when the focus is on modelling hazard rates over time rather than survival probabilities, as explained in the previous section.

$$\widehat{H}(t) = \sum_{t_i \leq t} \frac{d_1}{n_i}$$

(Equation 6.8)

Where t_i represents each distinct time an event occurs, d_i is the number of observed events (e.g., deaths) at time t_i, and n_i is the number of individuals at risk just before time t_i.

This additive approach builds up the cumulative hazard over time by summing small risk contributions at each event time. It is especially helpful in visualising and comparing hazard patterns across different groups or periods. While it does

not provide a formal statistical test for group comparisons like the Log-rank test, it plays an important descriptive role and can also serve as a basis for estimating survival through the relationship $\hat{S}(t) = e^{-\hat{H}(x)}$

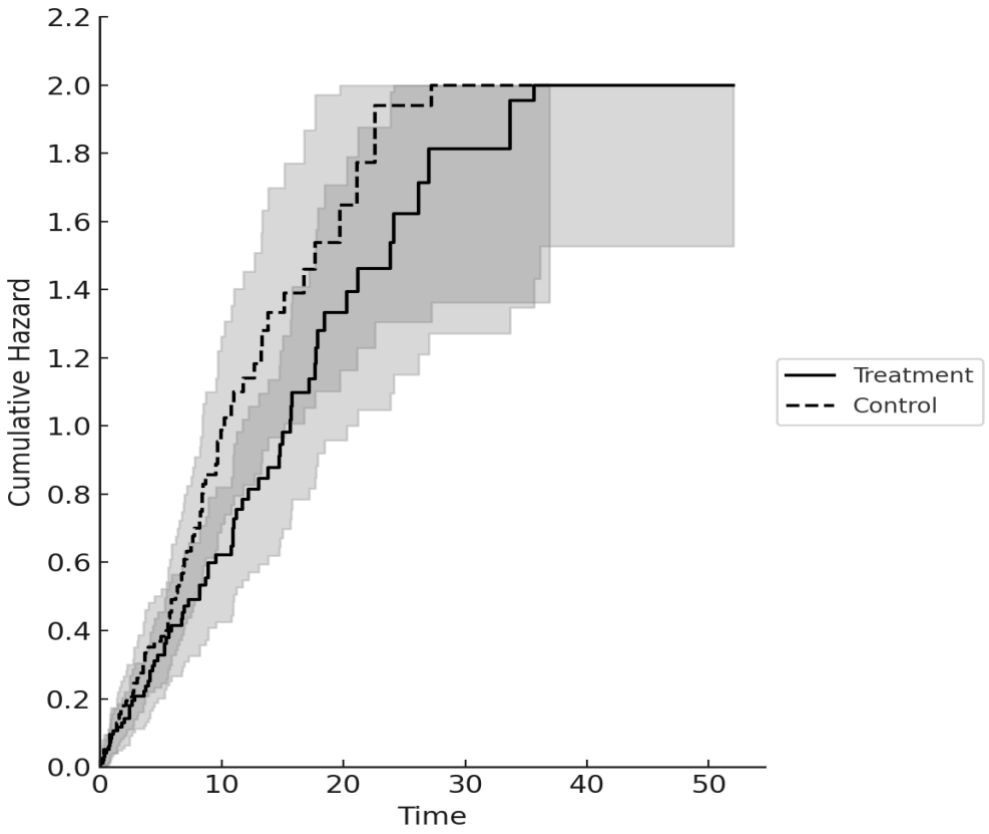


Figure 6.6 Nelson-Aalen cumulative hazard curve

Figure 6.6 (with confidence interval) indicates that the treatment group had a lower cumulative hazard over time compared to the control group, indicating the treatment may be protective or effective in reducing the risk of the event (i.e., death). The control group experienced more events which generated high cumulative hazard. Compared to KM, the NA describes the accumulation of risk rather than individual survival probabilities. When the cumulative hazard exceeds 1 (for example, reaching 2), it indicates that, on average, each individual has accumulated “two units” of risk. This does not directly represent the proportion of people who experienced the event. Thus, this non-linear outcome is more challenging to interpret and less intuitive. To interpret them

meaningfully, it is often necessary to transform them back to survival probability, or pair them with KM plot.

Sharing the same non-parametric, univariable, and descriptive nature with KM, the NA estimator also has similar key limitations. The method cannot account for multiple risk factors or perform covariate-adjusted analysis.[189,274,278] To address this, more advanced methods such as semi-parametric models like the Cox proportional hazards model are required.[281]

6.3.2 Semi-parametric model

Cox Model: preliminaries

In health and medical research, there are many situations where multiple variables (known as covariates) can influence an individual's prognosis.[189,275] For example, consider a study comparing two groups (treatment versus control), individuals may have a condition such as with and without family history of diabetes, as well as varying ages. Any observed differences in survival outcomes beyond treatment effect itself could be due to older age, family history, or combination of both. Therefore, when examining the effect of a particular factor on survival, it's important to adjust for other variables that might also impact the outcome. Such adjustments can improve the precision of estimates, particularly when assessing the effect of a treatment or exposure.

In 1972, David Cox presented his paper entitled "Regression Models and Life-Tables" which presented a regression method for analysing survival data.[190] The purpose of the method is to investigate several variables on survival simultaneously, known as proportional hazard (PH) regression analysis. A common way of referring to this well recognised and most applied survival analysis method is the 'Cox model'. In this thesis, the general term 'Cox model' will be used to represent term such as Cox regression and Cox PH model, since these refer to the same underlying method.

The Cox model is categorised as semi-parametric, simply because it has two parts: parametric and non-parametric.[248,252,282] A parametric part is the model assumes a specific form of how covariates affect the hazard (through linear combination), just like in regular regression.

The Cox model does not require a specified functional form for the baseline hazard, and it does not estimate it directly. Instead, the model focuses on estimating the relative effects of covariates on the hazard. No assumptions are made about the shape of the baseline hazard, it can increase, decrease, or vary in any form over time.

Assumptions and interpretations of Cox model

The Cox model is represented by the conditional hazard function $h(t|X)$. In this case, this $h(t|X)$ can be interpreted as the risk of having an event at time t[190]:

$$h(t|X) = h_0(t).exp\ (\beta_1 X_1 + \beta_2 X_2 + \cdots . + \beta_p X_p)$$

$$h(t|X) = h_0(t).exp\ (\beta^T X)$$

(Equation 6.9)

Where t is the survival time, h_0 is the baseline hazard function, representing the hazard when all covariates equal to zero. This is as the reminder that the baseline hazard can change over time. $X_1, X_2, .. X_p$ are covariates and the $\beta_1, \beta_2, .. \beta_p$ are the coefficients that measure the impact of the covariates on hazard. Thus, the $exp\ (\beta^T X)$ is known as hazard ratio (HR) for covariate X_i, indicating the relative change in hazard for a one-unit increase in that variable. For interpretation[189]:

- HR = 1 indicates no effect on the hazard.
- HR > 1 suggests an increased hazard (event occurs sooner).
- HR < 1 indicates a decreased hazard (delayed event).

In other words, a HR greater than 1 means that as the value of a covariate increases, the event (e.g., death, disease) is more likely to happen sooner. Conversely, a HR less than 1 indicates a protective effect, delaying the event.

To understand assumptions on proportionality, consider the HR between two individuals with X_1 and X_2 :

$$HR = \frac{h(t|X_1)}{h(t|X_2)} = \frac{h_0(t).exp\ (\beta^T X_1)}{h_0(t).exp\ (\beta^T X_2)} = \exp\ (\beta^T (X_1 - X_2))$$

(Equation 6.10)

The fact that time t disappears from the HR expression is exactly what the proportional hazard refers to.[252] It means that the hazard ratio does not change over time. The effect of the covariates (that represent through HR) is constant throughout the follow up-period.

A key strength of the Cox model is that it does not require the specification of the functional form of the baseline hazard. That is, the shape of $h_0(t)$ is left unspecified and estimated non-parametrically from the data.[190,252,283] This characteristic gives the Cox model its 'semi-parametric' nature which makes the model very flexible. A simple illustration of this semi-parametric model is for example when studying how smoking and blood pressure (BP) affect myocardial infarction (MI). The Cox model will estimate how smoking and BP change the relative risk, but it would not assume how the baseline risk of MI changes over time.

Despite its flexibility, the Cox proportional hazards model relies on several key assumptions that must be satisfied to ensure valid and interpretable results. The most fundamental is the proportional hazards assumption[189,252], as stated above. If this assumption is violated, the estimated HRs may be biased, and the model may not accurately reflect the relationship between covariates and the event risk.[190,278]

Another important assumption is the independence of survival times across individuals, meaning that the occurrence of an event in one subject does not influence the risk in another. This assumption can be problematic in clustered or correlated data, such as patients within the same hospital or community, and

may require advanced modelling approaches like frailty models.[284] Additionally, the model assumes a linear relationship between covariates and the log-hazard function; non-linearity can lead to misestimation of effects and may be addressed using transformations or spline functions.[285]

Finally, the model assumes non-informative censoring, if censoring is informative, specialised methods or sensitivity analyses may be required.[286,287] Careful evaluation of these assumptions through residual diagnostics and model checking is essential to ensure the robustness and reliability of the Cox model's findings.

Other semi-parametric method for handling competing risk

In the presence of competing risks, standard Cox regression (if alive to death state) can be extended using two main semi-parametric approaches: the cause-specific hazard (CSH) model and the sub-distribution hazard model, also known as the Fine–Gray (FG) model.[267,288] It estimates the effect of covariates on the instantaneous risk of experiencing a specific type of event, while treating other competing events as censored.[289,290]

In the context of MSM, each transition between defined health states is modelled separately, and these transition-specific hazards align with CSH, assuming the transitions from a given state are mutually exclusive.[290] That said, it is important to be clear about the framing. Competing risk models typically begin from a single starting state and consider multiple mutually exclusive end events, such as different causes of death. In contrast, multi-state models go further by allowing for multiple intermediate and absorbing states, which means they can capture sequential events and more complex pathways.[261]

The sub distribution hazard model, commonly known as the Fine–Gray (FG) model, was introduced in 1999 to directly model the cumulative incidence function (CIF) for a specific event, while appropriately accounting for the presence of competing risks.[288] Unlike cause-specific hazard models, the FG model does not censor individuals who experience competing events. Instead, it includes them in the risk set using adjusted weighting, which allows for a more accurate estimation of event probabilities over time.[259,267]

However, for long-term analysis, both CSH and FG model have limitations.[267,268,289,291] Because it is a semi-parametric model and does not impose a parametric form on the baseline hazard, it cannot easily extrapolate beyond the observed follow-up period. Like the standard Cox model, it relies heavily on the observed data, and the cumulative incidence estimates are constrained to the time frame of the available follow-up.[292]

Since this thesis primarily focuses on developing a model capable of projecting long-term outcomes and supporting extrapolation, the discussion of competing risks methods is limited to a conceptual overview. Ultimately, the CMD Policy Model is designed to prioritise flexibility in extrapolation through a multi-state framework. While Cox models within the multi-state structure are still used to assess covariate effects, the emphasis of the model lies in capturing disease progression over time rather than modelling mutually exclusive terminal events.

Model diagnostics

Schoenfeld residuals are used to evaluate whether the proportional hazard assumption holds by examining the relationship between residuals and time. Specifically, these residuals assess whether the effect of a covariate changes over time by testing for a correlation between the residuals and event time.[293] A statistically significant correlation suggests a potential violation of the proportional hazards' assumption for that covariate.[294]

Schoenfeld residuals are calculated for each covariate and for each individual who experiences the event of interest (i.e., uncensored observations). Each residual represents the difference between the observed value of the covariate for an individual who had the event and the expected value of that covariate across all individuals who were at risk at the time of the event. If the proportional hazards assumption holds, these residuals should be randomly scattered with no systematic pattern over time. Graphical inspection and formal testing can help identify whether any covariate violates this key assumption.[295,296]

In addition, Martingale residuals are primarily used to assess whether the functional form of continuous covariates is appropriately specified in the

model.[297] The Cox model assumes that the effect of each covariate is linearly related to the log-hazard.[190] To check this assumption, Martingale residuals are plotted against continuous covariates. If the relationship is truly linear, the plot should show no systematic pattern, the points would appear randomly scattered. However, if the plot shows a curved or non-random trend, this may indicate non-linearity in the covariate's effect.[294]

Martingale residuals are defined for all individuals, regardless of whether they experienced the event or were censored, and they typically take values between $-\infty$ and 1.[297] The residual tends to be closer to 1 for individuals who experienced the event and much smaller (or negative) for censored observations.[297]

Deviance residuals are used to identify outliers or influential observations that may disproportionately affect the model's estimates, especially for identifying data points that the model fits poorly. Large positive or negative deviance residuals suggest the observed survival time deviates substantially from what the model predicts. Such observations could be outliers or influential cases where data points with high leverage that may shift the estimated coefficients significantly if removed. Plotting deviance residuals against fitted values or covariates can reveal which individuals may be problematic and deserve further investigation.

Another test to assess model adequacy is Cox-Snell residual.[298] For each individual, the residual is defined as the estimated cumulative hazard at their observed event or censoring time. Under the assumption that the model is correctly specified, these residuals should follow a unit exponential distribution.[299] In practice, the model is assessed by plotting cumulative hazard function of the Cox-Snell residuals, typically using the NA estimator—against the residual values themselves. A well-fitting model will produce a plot that closely aligns with the 45-degree line, reflecting agreement between the observed data and the model's hazard predictions.[300]

Although the Cox model is semi-parametric, it allows for the computation of cumulative hazards through methods such as Breslow's estimator.[301] Unlike diagnostic tools such as Schoenfeld or Martingale residuals, which target specific assumptions (e.g., proportional hazards or covariate functional form), Cox-Snell residuals provide a more general assessment of model fit.[298] Their use adds an important layer of validation to ensure that the model offers a reasonable representation of the underlying survival process.

6.3.3 Standard parametric model

In parametric survival models, all parts of the model are fully specified, both the hazard function and the covariates' effect. Parametric survival analysis assumes that survival times follow specific statistical distributions (e.g., Log-normal, Exponential, Weibull). These assumptions define the shape of hazard and survival curves using mathematical equations. The general form of a parametric survival model (proportional hazards) is[189]:

$$h(t|X) = h_0(t;\theta).\exp\left(\beta^T X\right)$$

(Equation 6.11)

which is equivalent as:

$$S(t|X) = S_0(t;\theta)^{\exp(\beta^T X)}$$

(Equation 6.12)

Similar definition of survival formula, the $h_0(t;\theta)$ is the parametrically specified baseline hazard function with parameter(s) θ, and $S_0(t;\theta)$ is the corresponding baseline survival function. As it fully determines the hazard shape and survival functions, the choice of distribution for survival times is important. Different distributions make distinct assumptions about how the hazard rate 'behaves' over time. Choosing an appropriate distribution depends on the nature of the event process being modelled and the shape suggested by the data.

In addition to the PH formulation, parametric survival models can also be expressed using the accelerated failure time (AFT) formulation. The AFT model expresses the effect of covariates as a direct acceleration (or deceleration) of the survival time:

$$log(T) = \mu + \beta^{\tau}X + \varepsilon$$

(Equation 6.13)

where T is the survival time, μ is the intercept term, $\beta^{\tau}X$ represents the linear predictor, and ε is a random error term following a specific distribution. In this formulation, covariates act to stretch or shrink the time scale, rather than modifying the hazard multiplicatively as in PH models.

The two modelling approaches, proportional hazard (PH) and accelerated failure time (AFT) models represent two different ways of understanding how covariates affect survival time.[302] In PH model, the idea is that covariates affect the risk of an event happening at any moment in time.[295] For example, there are two patients: one who smokes and one who does not. If smoking doubles the risk of death, this doubling stays the same over time, at every day, every month, and every year, the smoker's risk is always twice as high. Covariates in PH models multiply the hazard (the risk) but do not change how the risk evolves over time.

In contrast, an AFT model is formulated differently (see equation 6.12). Instead of focusing on the risk at each moment, it focuses on the entire survival time. Covariate effects in AFT models stretch or compress the survival timeline.[302,303] For example, a treatment might double the survival time compared to no treatment, meaning patients live twice as long, but the shape of the survival curve stays the same.

Table 6.1 summarises the common parametric distributions used in survival analysis, presenting their mathematical forms, parameterisations, and typical applications in research. These models can be easily estimated using statistical

software, which also enables graphical visualisation (curves) of their survival and hazard functions.

Regarding how to specify these distributions, in simple terms, the main differences between these distributions lie in how the underlying hazard over time are modelled. Some distributions assume a constant risk, while others allow the risk to increase, decrease, or vary in more complex ways.[304,305] In addition to the primary regression parameters, many distributions include ancillary parameters that govern the shape, variability, or higher-order moments of the distribution.[304] These ancillary parameters critically influence the hazard function, allowing it to assume a variety of forms, such as constant, monotonic, or hump-shaped, depending on the distribution chosen[189] (Figure 6.7).

6.3.4 Flexible parametric model

Another alternative in parametric survival modelling is the flexible parametric survival model (FPM), which was popularised by Royston and Parmar and is often referred to as the Royston-Parmar (RP) model.[306,307] Methodologically, this approach models the baseline hazard using smooth functions, typically restricted cubic splines (also known as natural splines) applied to the log cumulative hazard.[308]

This spline-based framework enables the model to flexibly capture complex hazard patterns, allowing for smooth hazard function estimation without the need to assume a specific parametric distribution. As a result, FPMs combine the interpretability and structure of traditional parametric models with the adaptability to fit non-linear and non-monotonic hazard shapes.[306]

Unlike standard parametric models that impose a fixed functional form (e.g., always increasing or decreasing hazard), flexible parametric models allow the data to guide the shape of the hazard (Equation 6.14), making them especially valuable when the true hazard function is unknown or varies over time.[309] These models let the data “speak” about the hazard shape.

$$\log H(t) = \eta(t) = \gamma_0 + \gamma_1 s_1(\log t) + \cdots + \gamma_k s_k(\log t) + x^T \beta$$

(Equation 6.14)

Where $H(t)$ is cumulative hazard, s_1 is spline functions of log-time, *x*= covariates and β is coefficients for covariates.

This flexibility, however, relies on careful specification of the spline structure especially the number and placement of knots, which define where the spline can bend to capture changes in the hazard function. The choice of knots is critical: too few can lead to underfitting and missed hazard features, while too many can result in overfitting and unstable estimates.[306,310] In practice, knot placement is commonly based on quantiles of the log event times, and the degree of flexibility is controlled by specifying either the degrees of freedom (df) or the number of internal knots (k), depending on the software implementation. A common approach is to begin with a model equivalent to a Weibull distribution (e.g., df = 1 or k = 1) and then gradually increase complexity, evaluating improvements in fit.[309]

A key advantage of RP model lies in its flexibility, particularly in the direct estimation of baseline and cumulative hazard functions, as well as their ability to provide analytic expressions that facilitate further inference—such as the computation of cumulative incidence functions or restricted mean survival times[306,309,311,312]. This flexibility enhances both model fit and interpretability, allowing researchers to visually and statistically assess the shape of the baseline hazard, manage the risk of overfitting when incorporating time-varying effects, and compare alternative model specifications.

Table 6.1 Parametric distributions

Distribution	Metric	Survival function $S(t)$	Parameterisation	Ancillary parameters	Common use
Exponential	PH & AFT	$S(t) = exp\ (-\lambda t)$	$\lambda_i = exp\ (X_i\beta)$	None	Models constant risk over time. Used in mechanical failures, rare in clinical studies.
Weibull	PH & AFT	$S(t) = exp\ (-\lambda t^{\gamma})$	$\lambda_i = exp\ (X_i\beta)$	γ (shape)	Models monotonic increasing or decreasing hazards. Common in chronic disease survival.
Gompertz	PH	$S(t) = exp\ -\left(\frac{\lambda}{\gamma}(e^{\gamma} - 1)\right)$	$\lambda_i = exp\ (X_i\beta)$	γ (growth rate)	Models exponentially increasing hazard. Typical for human aging and mortality data.
Log-normal	AFT	$S(t) = 1 -\ \phi\left(\frac{log(t) - \mu}{\sigma}\right)$	$\mu_i = X_i\beta$	σ (scale)	Hazard rises then falls. Useful for time to disease recurrence or epidemics.
Log-logistic	AFT	$S(t) = (1 + (\lambda t)^{\gamma})^{-1}$	$\lambda_i = exp\ (-X_i\beta)$	γ (shape)	Models' long-term survival patterns. Good for chronic conditions with heavy-tailed survival.
Gamma	AFT	$S(t) = 1 - F(t) = 1 - \frac{\gamma(k, \lambda t)}{\Gamma(k)}$	$\mu_i = X_i\beta$	σ (scale)	Models waiting times for multi-stage biological processes (e.g., disease progression)
Generalised gamma	AFT	*If* $\kappa > 0$, $S(t) = 1 - I(\gamma, u)$ *if* $\kappa < 0$, $S(t) = 1 - I(\gamma, u)$, *where* $u = (e^{\beta t})^{\lambda}$	$\mu_i = X_i\beta$	σ, κ	Extremely flexible; nests Weibull, log-normal, gamma. Used for extrapolation in economic models.

PH: proportional hazard, AFT: accelerated failure time

$S(t)$: survival function, X_i = covariate, β= regression coefficient, λ_i= hazard function modified by covariates under proportional hazard, μ_i: location parameter for models with AFT interpretation, γ: shape parameters, can be defined as increasing/decreasing hazard or positive hazard=increasing, ϕ = normal cumulative distribution function (CDF), $\gamma(k, \lambda t)$ incomplete gamma function, $\Gamma(k)$= gamma function, u: transform time variable, σ and κ: additional shape and flexibility parameter

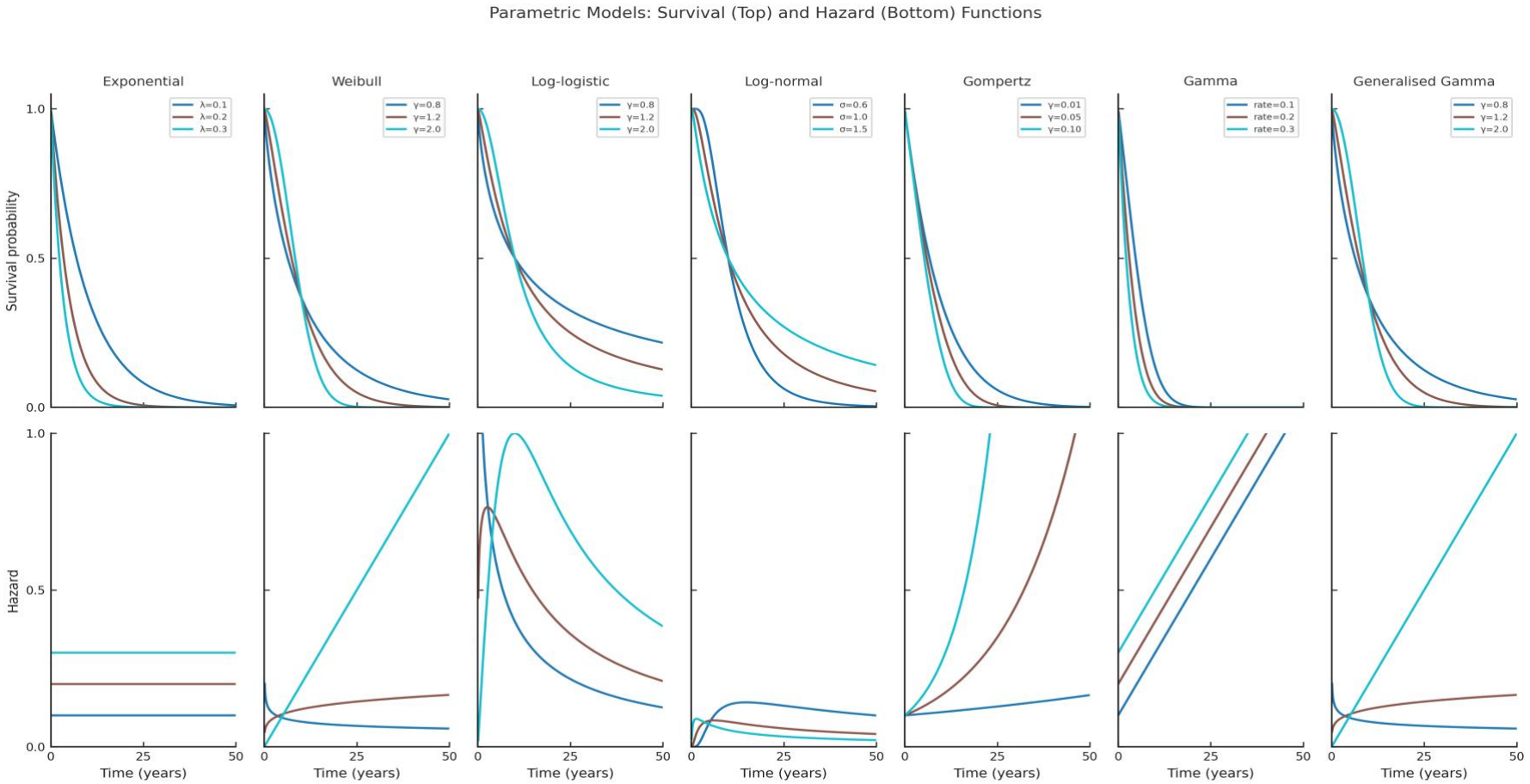


Figure 6.7 Survival and hazard function curves of parametric models

6.3.5 Model selection and evaluation for parametric modelling

Like semi-parametric models discussed earlier, it is essential to evaluate whether a parametric survival model adequately represents the underlying data after fitting. This evaluation helps to enduse that the model prediction is accurate and reliable.

One common approach for evaluating model fit is through goodness-of-fit statistics, particularly the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC).[304,305] Both metrics balance model fit and complexity by penalising the number of estimated parameters to reduce the risk of overfitting. Their general formulas are:

$$AIC = -2\log(L) + 2k$$
$$BIC = -2\log(L) + k\log(n)$$

(Equation 6.15)

Where L is the likelihood of the model, k is the number or parameters in the model, and n is the sample size.

Both criteria incorporate a penalty for model complexity, but BIC imposes a stronger penalty as it scales with the logarithm of the sample size $(\log(\mathrm{n}))$, whereas AIC uses a constant multiplier. As a result, BIC tends to favour simpler models more heavily, particularly in large datasets. Lower values of AIC or BIC indicate a better balance between model fit and parsimony, with the preferred model being the one with the lowest score among the candidates.

In addition to numerical criteria, visual inspection is also a valuable tool for assessing model fit. This involves comparing the estimated survival function from the parametric model to the non-parametric Kaplan-Meier (KM) survival curve.[313] A close alignment between the two curves indicates that the parametric model captures the observed survival pattern well. Conversely, systematic deviations suggest misspecification of the hazard function or distributional form.

Figure 6.8 is a visual example when comparing the KM survival curve with seven common parametric models. Among these, the Weibull distribution demonstrates the best overall fit, closely aligning with the non-parametric estimate across entire study period. This is further supported by its lowest AIC and BIC values among the standard distributions considered, indicating better model parsimony and goodness-of-fit. While the Gamma distribution also provides reasonable approximation, its slightly higher AIC/BIC values suggest that Weibull is statistically more appropriate for the data presented. Consequently, the Weibull model was selected for subsequent analyses and simulations.

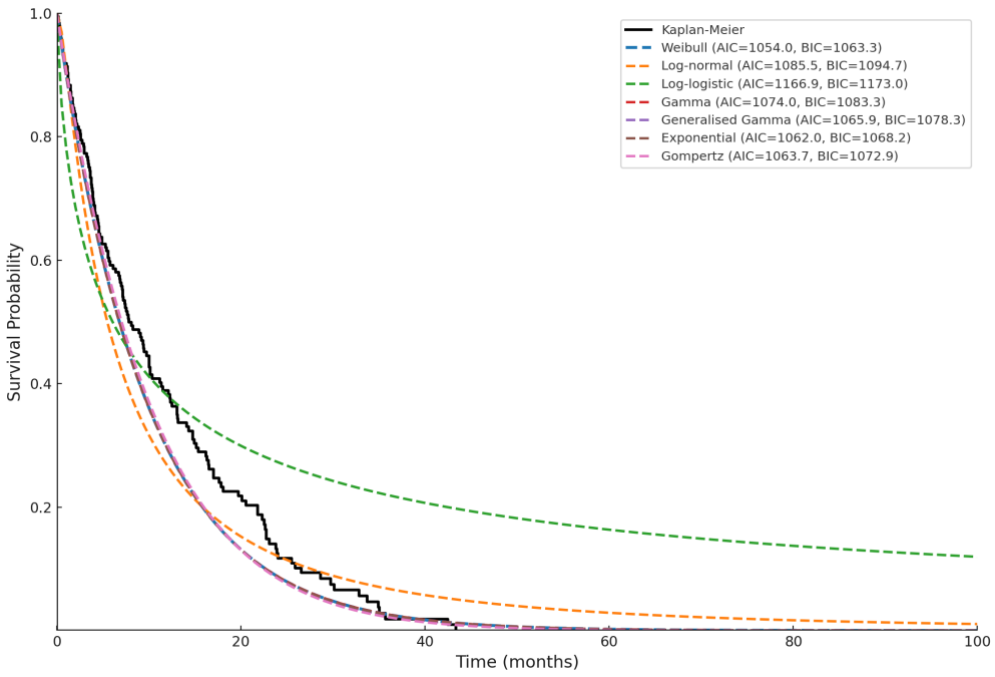


Figure 6.8 Comparison of parametric survival model and Kaplan-Meier estimation

In the flexible parametric context, the fit of the RP model can also be evaluated by visually comparing it with the KM curve as illustrated in Figure 6.8. A practical demonstration involves fitting RP models with increasing spline complexity (e.g., using one, two, or three internal knots) to show how added flexibility affects model fit..[312] Beyond visual comparison, RP models also allow formal evaluation using information criteria such as AIC/BIC, enabling assessment of relative fit across different specifications.[314]

These RP models alongside traditional parametric alternatives, can be both plotted against KM estimates and compared using numerical criteria to aid model selection.

Residual-based methods, such as Cox-Snell and deviance residuals, can help identify systematic departures from model assumptions and are applicable in both parametric and semi-parametric survival models.[296,298] However, the key difference lies in the underlying assumptions: parametric models rely on a fully specified distribution for survival times, making residual diagnostics particularly important for assessing the appropriateness of the chosen distribution. In contrast, semi-parametric models like the Cox model do not assume a specific baseline hazard function, so residuals are typically used to assess proportional hazards assumptions and covariate effects rather than the shape of the survival distribution itself.

6.3.6 Time-dependent covariates

Covariate effects are often assumed constant over time, particularly in traditional survival analysis, however in many real-world settings, individual characteristics such as clinical biomarkers, treatment status, or health behaviours change over the course of follow-up.[315,316]

Time-dependent covariates allow the model to more realistically represent evolving risk profiles. For example, a patient's cholesterol level may be elevated at baseline but improve with statin therapy, or a diabetes diagnosis may lead to lifestyle modifications that influence subsequent risk of stroke or myocardial infarction (MI). Failing to account for such changes risks misclassifying exposure status, which can result in biased estimates, attenuated hazard ratios, or misleading associations. By updating covariate values over time and aligning them with the appropriate intervals, survival models can provide more accurate estimates of how these variables impact the hazard function.[317,318]

Time-dependent covariates can be broadly classified as either external—such as calendar-based events or policy changes that are unrelated to the individual's event history—or internal, which include updated lab values or clinical conditions that may be influenced by the underlying disease process.[319,320]

One practical implementation involves incorporating the time since entry into a state (referred to as sojourn time) as a covariate in a Cox-type survival model to estimate transition hazards alongside evolving individual characteristics. This allows incorporation of updated information such as age, treatment status, or biomarker changes. The model relies on the baseline hazard, representing the event risk when covariates are zero, and may use the log cumulative hazard, a transformed version of the cumulative risk, to support model estimation.

6.3.7 Summary of different statistical models

The different feature of each model is summarised in Table 6.2. Based on the explanation provided earlier, survival analysis methods differ in term of assumptions, model structure, flexibility and practical application.

Non-parametric models provide empirical survival estimates without assuming any specific hazard structure; however, they do not support covariate adjustment or extrapolation beyond the observed data.[269] Semi-parametric models, incorporate covariate effects without specifying the baseline hazard function, though they depend on the proportional hazards' assumption.[190] Parametric models assume a predefined distributional for both the baseline hazard and covariate effects, offering ease of interpretation and the ability to extrapolate, albeit with limited flexibility in capturing complex hazard patterns.[313] Flexible parametric models, allow for more adaptable hazard shapes, the inclusion of time-dependent effects, and improved extrapolation capabilities. This makes them especially useful for modelling complex survival data and projecting long-term outcomes, even though in practice this method is still underutilised.[321]

In summary, the choice between non, semi, or flexible parametric in survival models depends critically on the study objectives, complexity of the survival data as well as modelling requirements. Key considerations include whether the analysis is focused solely on estimating covariate effects or also aims to perform long-term extrapolation.

Table 6.2 Comparison of statistical models

Feature	Non-parametric	Semi-parametric	Parametric	Flexible parametric
Model characteristics	No assumptions; empirical estimation	Parametric covariate effects, unspecified baseline hazard	Fully parametric: baseline and covariate effects specified	Parametric covariate effects, spline-based hazard
Baseline hazard	Not modelled explicitly	Not estimated directly	Assumed known form (e.g., exponential, Weibull etc)	Estimated via restricted cubic splines
Hazard shape assumption	None	None (but assumes proportional hazards)	Fixed functional form	Flexible, data-driven
Proportional hazard assumption	Not applicable	Required (unless extended)	Often assumed	Can be relaxed
Covariate inclusion	Not supported	Supported	Supported	Supported
Time-dependent effect	Not supported	Can be included (extended Cox)	Difficult to implement	easily included via spline interactions
Interpretability	Easy to interpret survival probabilities	Covariate effects interpretable; baseline hazard is abstract	High interpretability if model fits well	Interpretation more complex due to splines
Model fit assessment	Visual (e.g., survival curves)	Residuals (e.g., Schoenfeld)	AIC/BIC, likelihood, visual fit	AIC/BIC, visual fit, spline tuning
Extrapolation capability	Not possible	Not recommended	Possible	Possible

Flexibility	High for empirical survival, low overall	Moderate	Low to moderate	High

6.3.8 Markov and semi-Markov assumptions

From the analytical standpoint, multi-state models (MSMs) are built upon either Markov or semi-Markov processes, where the assumptions about memory of past transition differs. Markov models assume that future transitions depend solely on the present state, while semi-Markov allow the sojourn time in a given state to influence future transitions.[322]

This conceptual difference is often reflected in the choice of time scale, typically described as either 'clock-forward' or 'clock-reset'.[260] In a clock-forward model (aligned with the Markov assumption), the hazard of transition depends on the time since the beginning of the process (e.g., since study entry or baseline). Time continuously accumulates as an individual moves through different states. In contrast, a 'clock-reset' model (used in semi-Markov settings) resets the time counter each time a new state is entered, meaning that the hazard of transition depends on the time since entering the current state.[323]

For example, in modelling disease progression, the risk of death in a 'clock-forward' model might depend on the total time since diagnosis, while in a 'clock-reset model', the risk of death after a stroke would depend specifically on the time since the stroke occurred. These assumptions have important implications for how risk evolves over time and must be carefully matched to the nature of the disease and available data.

Markov models

A Markov model assumes that the probability of moving from one state to another depends on the current state, not on the time already spend in that state. This feature known as 'memoryless property', In simple terms, the chance of transitioning out of state depends on the current state, but not on how long the individual has been there.[324]

In a discrete-time Markov model, time moves forward in fixed steps (e.g.: monthly, annual). At each step, a person can either stay in the same state or move to a different one.[325] The model then uses a transition probability matrix

to show the chance of moving between states. Each number in the matrix p_{rs}, tells the probability of going from state r to state s in one time step.

$$p_{rs} = P(X_{t+1} = s \mid X_t = r)$$

(Equation 6.16)

This means, what is the chance that someone will be in state s at the next time step, given that they are currently in state r. Each row of matrix adds up to 1, because it includes all possible outcomes for someone in a given state.

In a continuous-time Markov model, changes between states can happen at any moment in time, not just at fixed intervals.[325] Instead of transition probabilities, this model uses transition rates, written in a matrix called Q. Each value q_{rs} tell about how quickly someone in state r is expected to move to state s. The higher the number, the faster the expected transition. The rows are set up so that everything balances out (i.e., each row sums to zero).

If the transition rates do not change over time, the model is called a time-homogeneous continuous-time Markov model.[326] In this case, the transition intensity matrix Q remains constant over time. However, in the time-inhomogeneous version, the transition rates depend on time, meaning the chance of moving between states can vary as time progresses[327] , for example, the risk of death may increase with age or disease duration.

In the time-homogeneous case, the transition probability matrix P_t which gives the probability of being in each state at time t, can be calculated using the matrix exponential:

$$P(t) = \exp(Qt)$$

(Equation 6.17)

In contrast, for time-inhomogeneous models where Qt varies with time, Pt must be calculated by solving the Kolmogorov forward differential equation:

$$\frac{dP(t)}{dt} = (Pt) \,.\, (Qt)$$

(Equation 6.18)

To help illustrate this, a simple illness-death model structure is adopted as an example (Figure 6.8). This setup can be used in either discrete or continuous time models to describe how people move through different stages of health. For example, someone might stay healthy, become diseased, or die. These transitions then can be simulated over time depending on the structure and assumptions of the model.

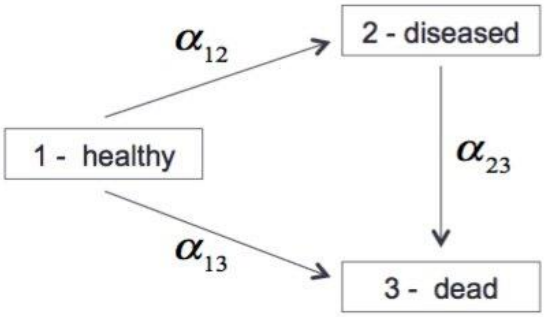


Figure 6.9 Three states Markov model

The general transition matrix can be illustrated as below. The format is common for discrete-time Markov models, where each row sums to 1, representing the probabilities of moving from a current state to all possible states in the next time period.

$$P = \begin{pmatrix} 1 - \alpha_{12} - \alpha_{13} & \alpha_{12} & \alpha_{13} \\ 0 & 1 - \alpha_{23} & \alpha_{23} \\ 0 & 0 & 1 \end{pmatrix}$$

(Equation 6.19)

and then re-parameterised the matrix with actual transition probabilities:

$$P = \begin{pmatrix} 0.85 & 0.10 & 0.05 \\ 0 & 0.80 & 0.20 \\ 0 & 0 & 1.00 \end{pmatrix}$$

From this matrix, it can be interpreted that there are 85% chance of remaining in healthy state, 10% chance of progressing to diseased, and 5% chance of directly transitioning to dead.

If this were in continuous time, the matrix would instead represent transition intensities (rates), typically called a Q-matrix, and the matrix exponentiation $P(t) = exp\ (Qt)$ would be used to derive probabilities over time.

Semi-Markov models

Standard Markov models rely on the *memoryless* assumption, a key consequence of this assumption is that sojourn times are exponentially distributed and "non-ageing," with a constant hazard rate that does not change as time in state accrues.[260,328] This is a stringent limitation in many clinical applications, since the risk of an event often varies with the length of time a patient has spent in a given health state.

In other words, the Markov model ignores any *duration* effect – a patient who has just entered a state is treated the same as one who has stayed there for months or years. Empirical evidence and theory have long shown that such an assumption can be unrealistic: the hazard of disease progression or death can increase or decrease as a function of time already spent in the current state. Consequently, a pure Markov approach may misrepresent the natural course of disease when past "time in state" is a strong determinant of what happens next.[329]

In a semi-Markov model, the transition process is still governed by a transition intensity matrix or by transition probabilities, but the waiting time distributions between transitions are no longer restricted to the exponential distribution (as in continuous-time Markov models).[257,329] This flexibility allows for more realistic modelling of TTE data, particularly when empirical evidence suggests that hazards change with time in state.

$$Q_{ij}(t) = p_{ij}G_{ij}(t)$$

(Equation 6.20)

A semi-Markov, where p_{ij} probability of transition from i to j. G_{ij} probability of transitioning **within** time t, given entry to state i. These *G* functions are survival distributions.

$$Q(t) = \begin{pmatrix} G_{11}(t) & G_{12}(t) & G_{13}(t) \\ 0 & G_{22}(t) & G_{23}(t) \\ 0 & 0 & 1.00 \end{pmatrix}$$

(Equation 6.21)

It should be noted that the $G_{ii}(t)$ are typically not defined, it is just implied that the individual staying at the same state or not leaving the state.

By allowing hazard rates to vary with the time spent in a given state, semi-Markov models provide a more flexible and clinically realistic framework for multi-state survival analysis. They are particularly well-suited to capturing the natural history of diseases in which the risk of progression or adverse outcomes evolves over time.

A commonly cited example is the illness–death model used in chronic disease contexts. In such cases, patients who remain longer in an intermediate or 'ill' state often face an increased risk of death, a feature that cannot be adequately captured under the memoryless assumption of a standard Markov model. Semi-Markov models overcome this limitation by incorporating sojourn time into the hazard function. Thus, if empirical data suggest that longer time in a progression state corresponds to higher mortality risk, the semi-Markov model can reflect this by allowing transition hazards to increase with sojourn time.

This enhanced clinical realism makes semi-Markov models particularly valuable for health policy modelling and medical decision-making. In health economics, Markov decision models are widely used to simulate patient trajectories and

evaluate the cost-effectiveness of interventions. However, it is increasingly recognised that the assumption of time-homogeneous (memoryless) transitions, inherent in standard Markov models, may lead to oversimplification and misrepresentation of disease processes. Semi-Markov models address this by enabling transition probabilities or hazards to depend explicitly on the time elapsed since entering a state, thus allowing simulations to be conducted in continuous time rather than through fixed-cycle approximations.[248,318,319]

This is especially important for long-term modelling where the timing of events such as disease progression, relapse, or death has a significant impact on clinical outcomes, resource use, and costs.[323,324] By explicitly modelling how long patients remain in each state before transitioning, semi-Markov models can more realistically project outcomes for interventions whose effectiveness or cost-effectiveness depends on when events occur. For example, an intervention that aims to delay the progression of a disease will have different implications if progression is postponed by a few months versus several years. A semi-Markov approach can capture these differences by accounting for the distribution of sojourn times in the pre-progression state.

Similarly, time-sensitive policies (such as earlier screening or rapid treatment escalation after a diagnosis) can be evaluated in a framework that reflects the natural timing of disease events.

6.4 Model Results

Based on the state transition model structure presented in Figure 6.4 (and described in Chapters 4 and 5), this section applies a range of survival modelling strategies using data from the Clinical Practice Research Datalink (CPRD) Aurum. The results are organised by modelling approach: non-parametric, semi-parametric, standard parametric, and flexible parametric to estimate event-specific risks across the model's 13 transitions. For each modelling strategy, transition-specific hazard estimates are presented alongside diagnostic assessments to identify the most appropriate model specifications.

It is important to note that these results do not represent a direct comparison of which modelling strategy is "superior," as each method offers distinct advantages. The Kaplan-Meier (KM) method offers a descriptive approach for estimating crude survival probabilities without adjusting for covariates. The Cox model allows for the inclusion of covariates to examine how individual characteristics affect transition risks while making minimal assumptions about the baseline hazard. Standard parametric models provide interpretable hazard functions and are suitable when the underlying risk can be reasonably approximated by predefined distributions. In contrast, flexible parametric models offer greater adaptability and can better accommodate non-proportional or non-linear risk patterns observed in empirical data.

In the context of current objective for policy modelling, the selection of survival modelling approaches was guided by a trade-off between model complexity, data availability, and the requirement for accurate long-term extrapolation. KM analysis was employed as an exploratory tool to visualise observed survival patterns and assist in identifying suitable parametric survival distributions. However, due to its inability to incorporate covariates or extrapolate beyond the observed follow-up period, it was not used directly in the policy model. Instead, modelling approaches were selected to account for multiple covariates, reflect the complexity of CMD progression, and support reliable extrapolation over a lifetime horizon.

As background, the study population comprises 184,845 individuals (89,645 males and 95,200 females), with baseline characteristics detailed in Chapter 5. Figures 6.10 and 6.11 summarise the distribution of modifiable and non-modifiable risk factors, stratified by sex. Stratification by sex was undertaken given the well documented differences in cardiometabolic disease (CMD) risk profiles, presentation, and progression between males and females. Non-modifiable factors (e.g., age, sex), along with behavioural characteristics (e.g., smoking status, alcohol consumption), are included due to their established associations with CMD risk. While their inclusion does not ensure model accuracy, these covariates help reflect individual-level heterogeneity and strengthen the clinical relevance of the estimated transition hazards.

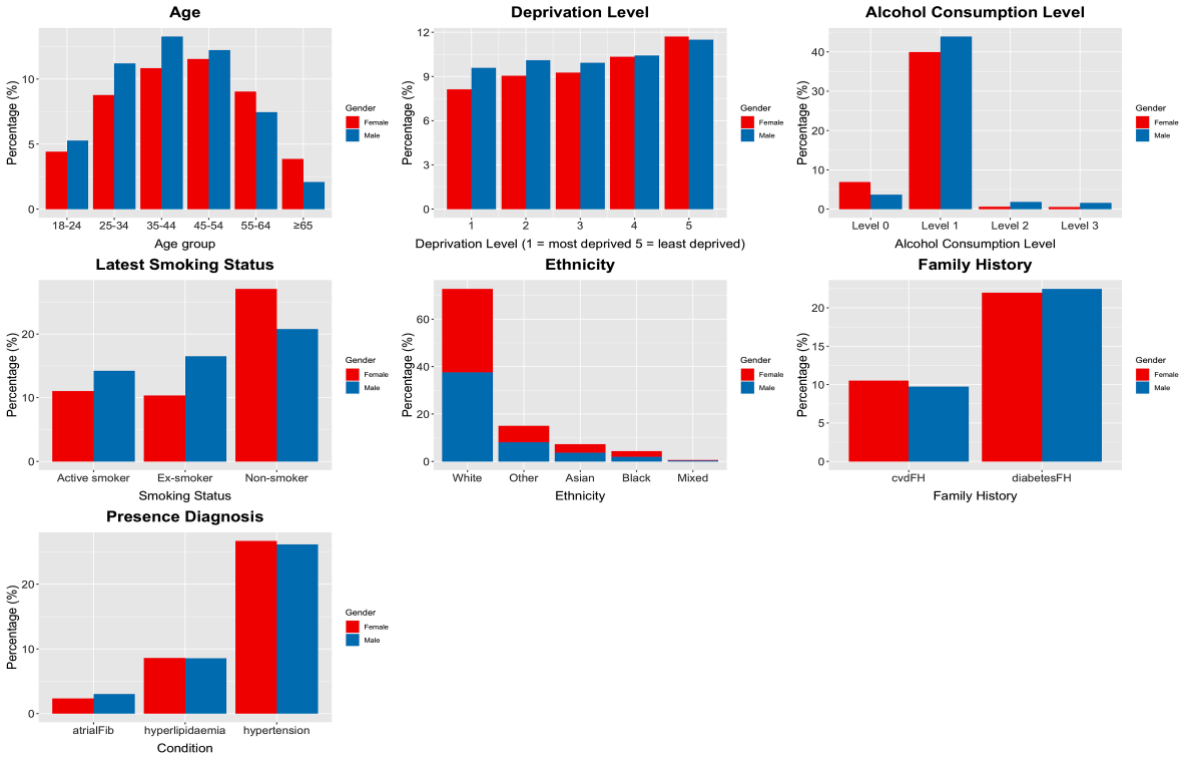


Figure 6.10 Non-modifiable and behavioural covariates by sex

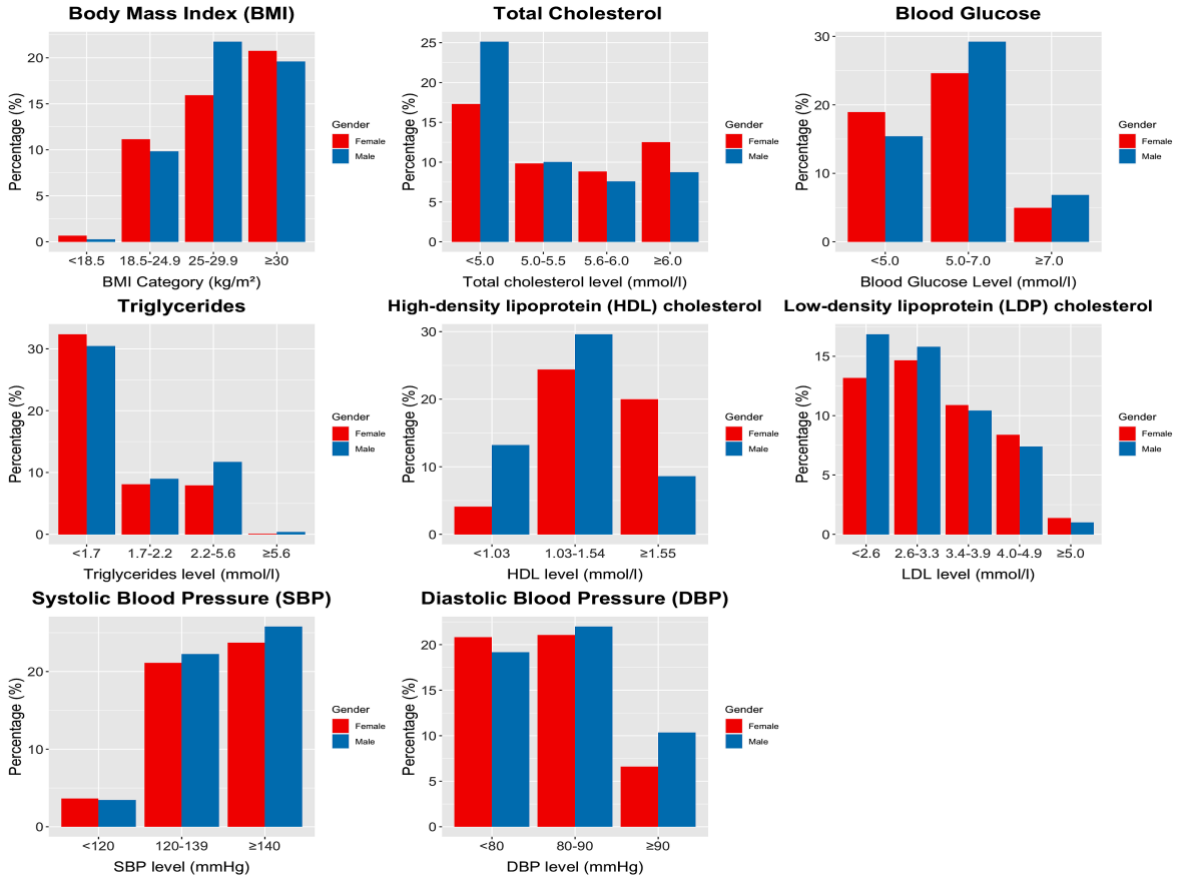


Figure 6.11 Modifiable biomarker covariates by sex

6.4.1 Kaplan-Meier results

As an initial step in exploring time-to-event patterns across the 13 transitions in the multi-state model, Kaplan-Meier (KM) survival analysis was conducted. This non-parametric method enables visualisation and comparison of survival distributions without assuming an underlying hazard structure, making it well-suited for preliminary investigation of event timing across subgroups.[189] In this case, sex-based differences were assessed, with log-rank tests used to evaluate statistical significance between male and female survival curves for each transition.

KM curves (Figure 6.12) indicated sex-based differences in progression across several transitions. Females exhibited faster progression from disease-free to T2DM (Transition 1) and from T2DM to death (Transition 7), while males showed more rapid transitions to MI from both disease-free (Transition 2) and diabetic states (Transition 5). In post-MI and post-stroke transitions, males consistently demonstrated longer survival, particularly in transitions to death (Transitions 9, 10, 12, and 13). Conversely, no statistically significant differences were observed in transitions such as disease-free to stroke (Transition 3), disease-free to death (Transition 4), and stroke to post-stroke (Transition 11), suggesting that sex-related disparities may be more pronounced in cardiovascular progression and mortality than in stroke onset or general survival.

Furthermore, Figure 6.13 presents cumulative hazard curves across all transitions. Cumulative hazard functions provide a complementary perspective to survival probabilities by illustrating the accumulated risk of an event occurring over time. Compared to KM curves, cumulative hazard plots emphasise the intensity and pace at which events accumulate, which is particularly informative when evaluating long-term risk trajectories.

The cumulative hazard plots reinforce the patterns observed in the survival analysis. Males exhibit a slower cumulative risk of progressing to T2DM (Transition 1) but a faster accumulation of MI risk from both disease-free (Transition 2) and diabetic states (Transition 5). Females demonstrate steeper

cumulative hazards for death following T2DM (Transition 7), MI (Transition 9), and post-MI (Transition 10), highlighting increased mortality risk in these transitions. These differences are statistically significant, consistent with the log-rank p-values. For other transitions such as disease-free to death (Transition 4), T2DM to stroke (Transition 6), and stroke to post-stroke (Transition 11), the cumulative hazard curves are largely parallel, indicating no significant sex-based divergence in event accumulation.

However, as non-parametric methods, KM and cumulative hazard estimates do not adjust for potential confounding variables such as age, socioeconomic status, comorbidities, or biomarker profiles (factors that can significantly affect transition risks).[270,271] To accommodate this, further analysis is conducted using Cox proportional hazards models.[190] The Cox model enables the estimation of hazard ratios and can incorporate both fixed and time-dependent covariates, providing a more comprehensive and interpretable framework for understanding the determinants of cardiometabolic disease progression in diverse subpopulations.

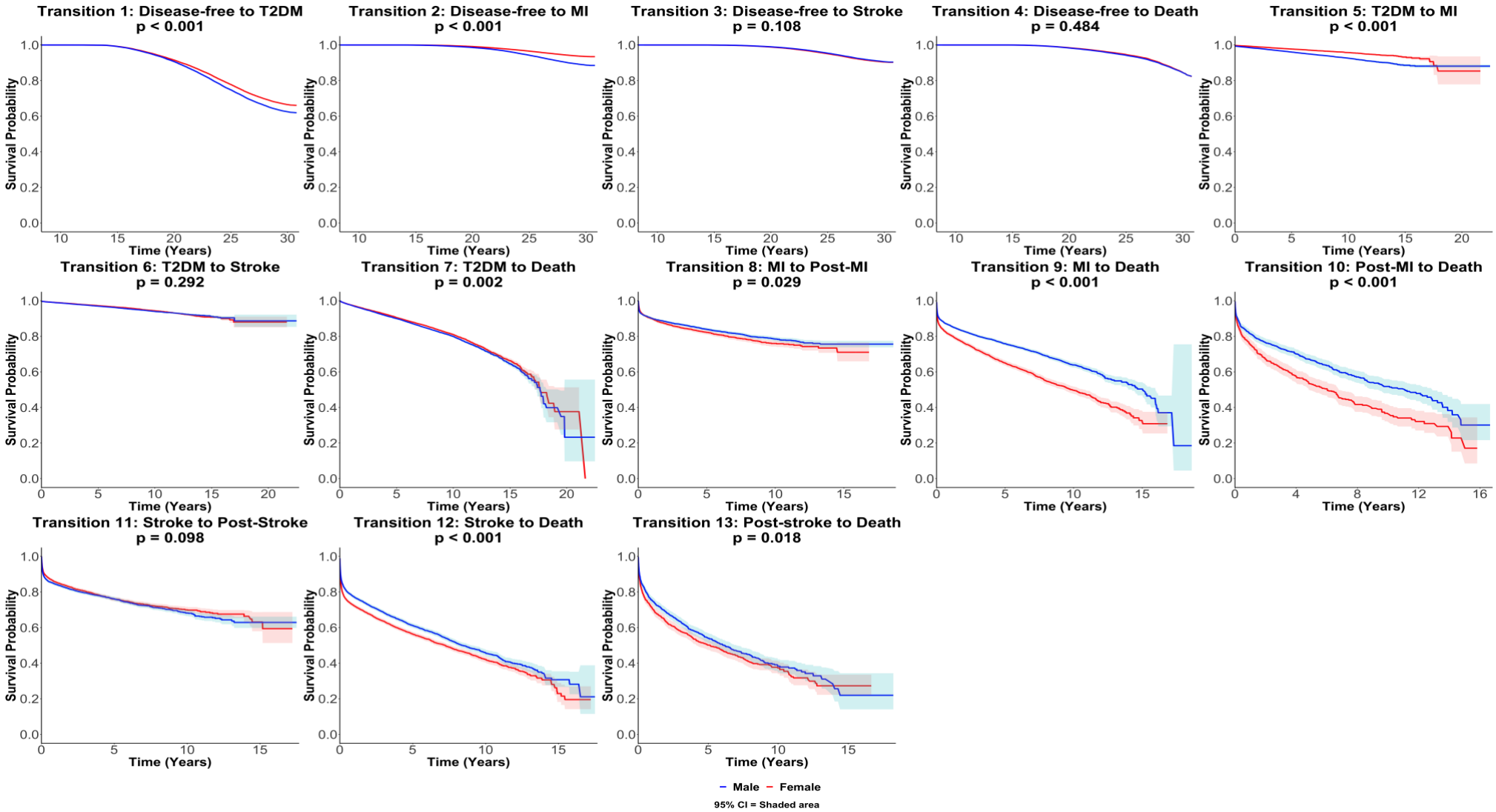

Transition 1: Disease-free to T2DM
p < 0.001
Transition 2: Disease-free to MI
p < 0.001
Transition 3: Disease-free to Stroke
p = 0.108
Transition 4: Disease-free to Death
p = 0.484
Transition 5: T2DM to MI
p < 0.001
Transition 6: T2DM to Stroke
p = 0.292
Transition 7: T2DM to Death
p = 0.002
Transition 8: MI to Post-MI
p = 0.029
Transition 9: MI to Death
p < 0.001
Transition 10: Post-MI to Death
p < 0.001
Transition 11: Stroke to Post-Stroke
p = 0.098
Transition 12: Stroke to Death
p < 0.001
Transition 13: Post-stroke to Death
p = 0.018
Survival Probability
Time (Years)
Male
Female
95% CI = Shaded area

Figure 6.12 Kaplan-Meier survival probability across all transitions

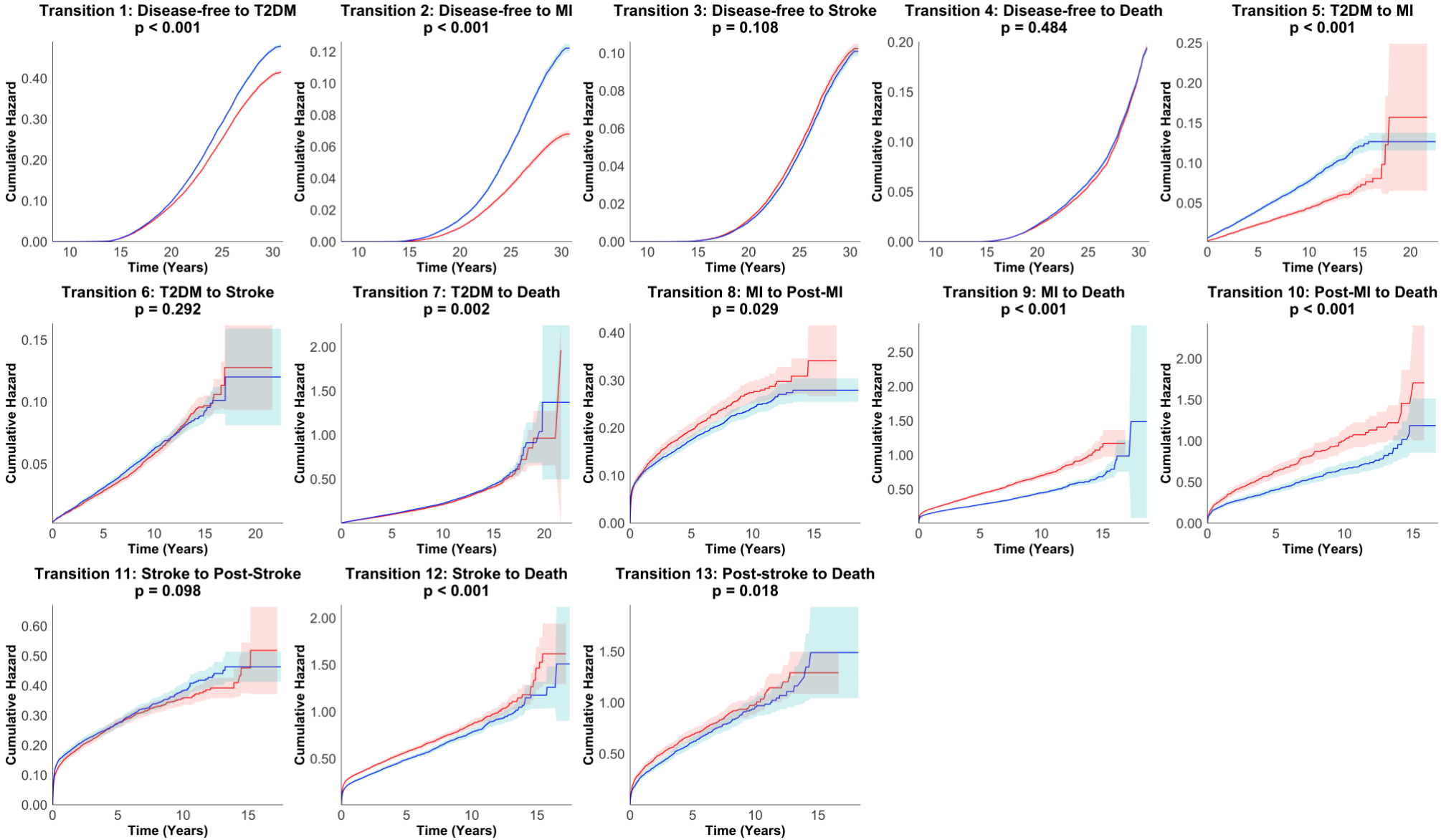

Transition 1: Disease-free to T2DM
p < 0.001
Transition 2: Disease-free to MI
p < 0.001
Transition 3: Disease-free to Stroke
p = 0.108
Transition 4: Disease-free to Death
p = 0.484
Transition 5: T2DM to MI
p < 0.001
Transition 6: T2DM to Stroke
p = 0.292
Transition 7: T2DM to Death
p = 0.002
Transition 8: MI to Post-MI
p = 0.029
Transition 9: MI to Death
p < 0.001
Transition 10: Post-MI to Death
p < 0.001
Transition 11: Stroke to Post-Stroke
p = 0.098
Transition 12: Stroke to Death
p < 0.001
Transition 13: Post-stroke to Death
p = 0.018
Cumulative Hazard
Time (Years)

Figure 6.13 Cumulative hazard across all transitions

6.4.2 Cox regression results

The pre-specified covariates were included in the Cox proportional hazards analysis, with appropriate reference groups assigned for categorical variables. Figures 6.14 to 6.20 present the results of Cox proportional hazards models. Each figure displays hazard ratios (HRs) with corresponding 95% confidence intervals (Cis). Statistically significant associations ($p < 0.05$) are highlighted in blue, whereas non-significant effects are displayed in red. In addition, the details of survival and hazard curves of Cox model results are presented in Appendix 5.

From the analysis, age shows a consistent and significant association with increased mortality risk across transitions, particularly from stroke and T2DM to death, which is well-documenter in most longitudinal studies.

High-density lipoprotein (HDL) is uniformly protective, while total cholesterol and low-density lipoprotein (LDL) are strongly associated with increased risk of T2DM, MI and stroke[332–335]. Smoking, both current and former, significantly elevates risk of CVD and death across states which is frequently observed in published epidemiological studies.[336,337]

Similarly, the findings regarding body mass index (BMI) and its complex relationship with cardiovascular outcomes emphasise that BMI is a significant but not straightforward indicator of cardiovascular health risks. BMI increases the risk of developing T2DM but is inversely associated with MI and stroke incidence, reflecting the well-documented 'obesity paradox' in cardiovascular outcomes. The obesity paradox refers to the surprising and counterintuitive finding that, in some populations or clinical conditions, people classified as overweight or mildly obese (based on BMI) seem to have better health outcomes than those with a "normal" BMI. This terminology is introduced by Lavie et al (2009)[338], as his paper observed that overweight and obese patients with heart failure had better survival rates than leaner patients. This phenomenon has since been widely discussed, as similar findings have been reported in several studies on CVD.[339–341]

The obesity paradox can happen due to BMI measurement itself that does not distinguish between fat and muscle, or fat distribution (visceral vs subcutaneous). The other case may be due to metabolic activity, when extra body fat might provide energy during illness or stress, or sicker people may lose weight (reverse causality), or the "normal BMI" group might include people who are underweight due to disease.[339–341]

In terms of presence of disease, atrial fibrillation consistently shows the strongest association with higher risk across nearly all events, especially for stroke-related transitions and cardiovascular death, aligning with extensive clinical evidence. Hypertension also emerges as a key risk factor, particularly after the onset of disease (e.g., post-T2DM or post-MI), reinforcing its cumulative impact on downstream complications.

Ethnicity plays a key role, with individuals of Black and Asian backgrounds generally exhibiting lower hazards for MI, stroke, and death, especially Black ethnicity (e.g., HR = 0.37 for MI). These patterns are consistent with findings from UK-based datasets like CPRD[342], ONS[343], and UK Biobank[344]. Socioeconomic deprivation (higher IMD quintiles) is also consistently associated with elevated risk of adverse outcomes, particularly in transitions from disease-free to T2DM, MI, stroke, or death. The impact of socioeconomic deprivation, as indicated by indices of multiple deprivation, further emphasises the association of lower social standing with poorer health outcomes.

Compared to non-drinkers, moderate intake of alcohol is not showing increase hazard while heavy use increases mortality risk, matching global burden of disease findings.[345] Moderate drinking appears protective while heavy consumption yields adverse outcomes, aligning with findings from various systematic reviews.[346,347]

From these results, it can be concluded that the use of time-dependent Cox model offers a robust framework for incorporating evolving covariate values and capturing dynamic risk profiles across disease transitions. This semi-parametric approach effectively highlights the relative strength of covariate effects on transition hazards. However, one key limitation is that the Cox model does not

specify the baseline hazard function, which restricts its interpretability in terms of absolute risk estimation and long-term extrapolation. Additionally, the model's validity may be compromised if the PH assumption is violated.

To address these PH limitations and assess the adequacy of the model, diagnostic procedures were undertaken. Specifically, Schoenfeld residuals[296], and Martingale residuals were examined for all transitions.[297]

Model diagnostics

The Schoenfeld test evaluates the proportional hazards assumption by testing whether the effect of a covariate on the hazard function remains constant over time. A significant p-value ($p < 0.001$) indicates a violation of this assumption, suggesting that the covariate's effect varies over time. The details on this is presented in Appendix 6.

The global test was significant across all transitions, indicating that for each transition, at least one covariate violated the proportional hazards assumption. Covariates exhibited consistent violations at the individual level, despite the use of time-updated values. This underscores an important limitation: while time-updated covariates capture changes in exposure status (e.g., biomarker levels), they do not account for changes in the effect of those covariates over time. In other words, the hazard ratio itself may vary with time, which is not addressed by simply updating covariate values. To meet the PH assumption in such cases, it may be necessary to include explicit time, covariate interaction terms or apply stratified Cox models that allow baseline hazards to differ across strata.[189,301]

Another reason may be the large statistical power inherent in this dataset. With such a large sample size, even minor deviations from proportionality unlikely to be clinically meaningful can produce statistically significant p-values in PH tests. Importantly, the hazard ratio represents an average relative effect over the follow-up period. Even when the proportional hazards assumption is not strictly valid across the entire duration, the estimated hazard ratios may still provide a meaningful summary of the average benefit over time. This interpretation aligns with conclusions from the WOSCOPS study, which acknowledged the potential violation of the PH assumption but still considered the HRs to reflect a valid average effect over the follow-up period.[348]

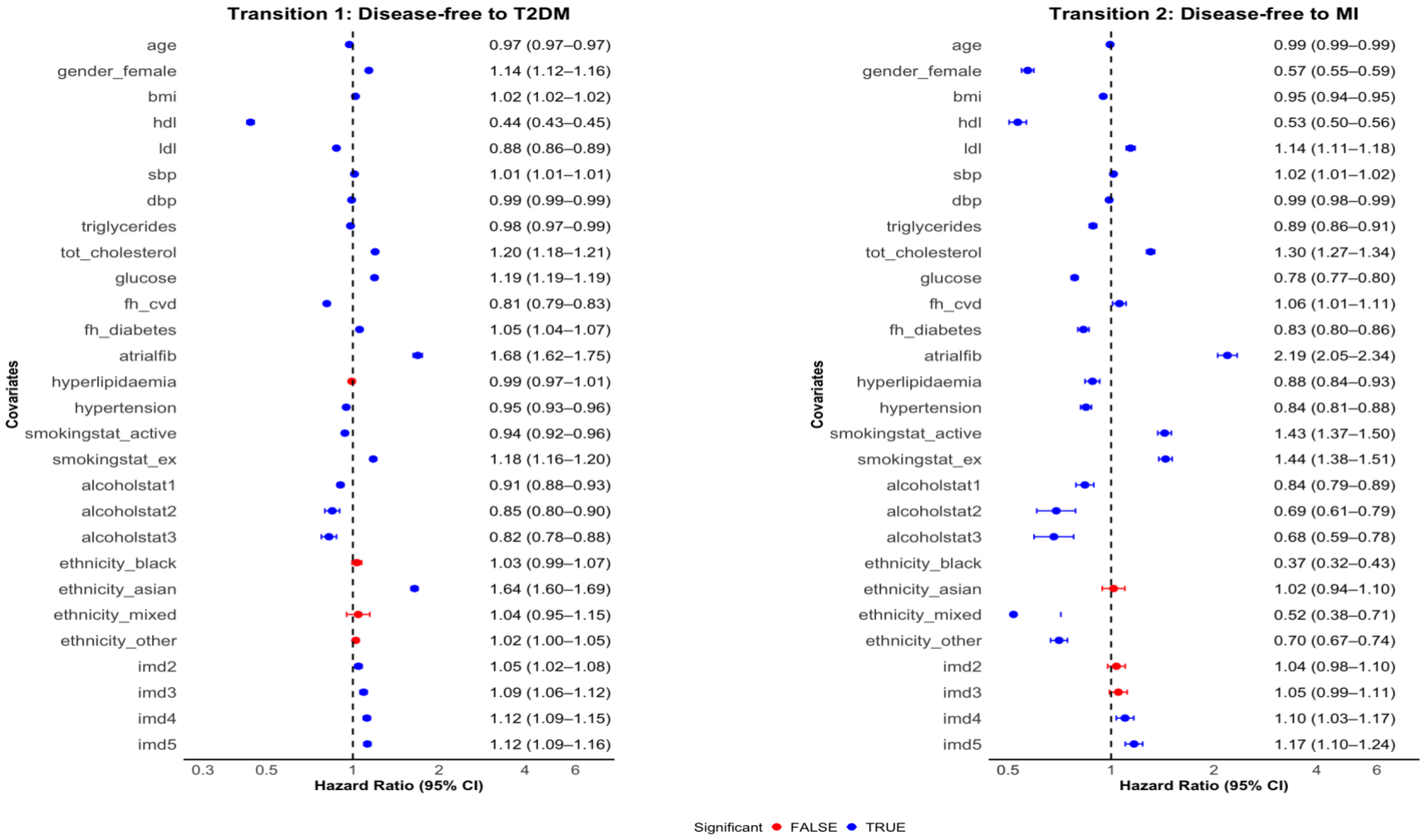
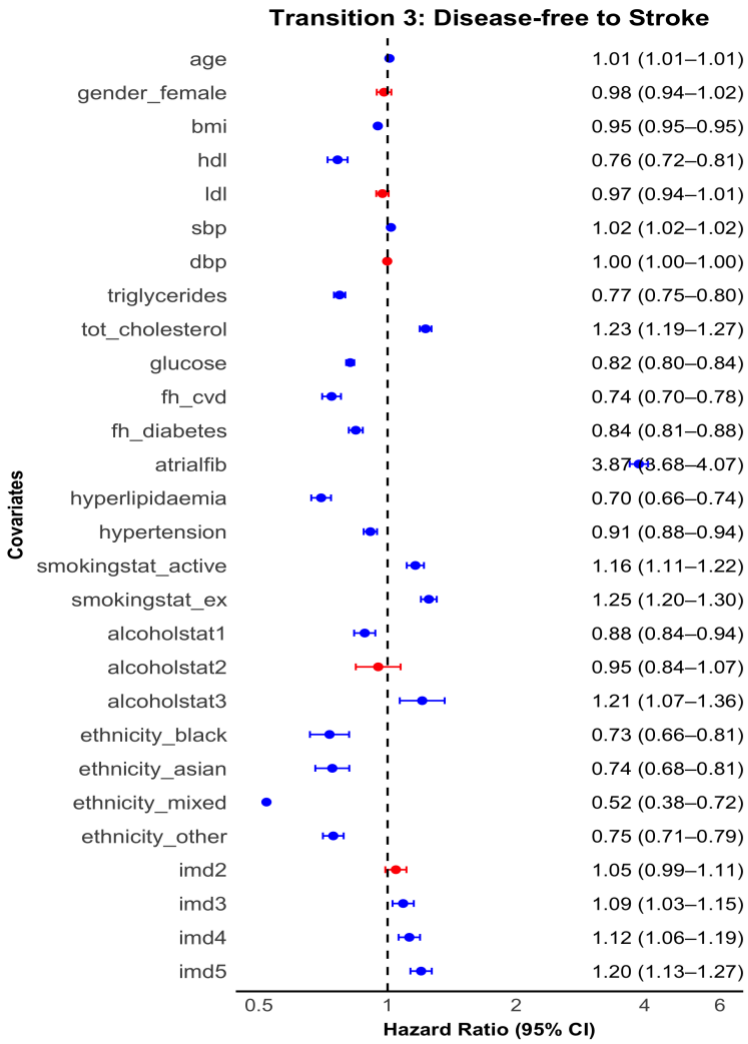
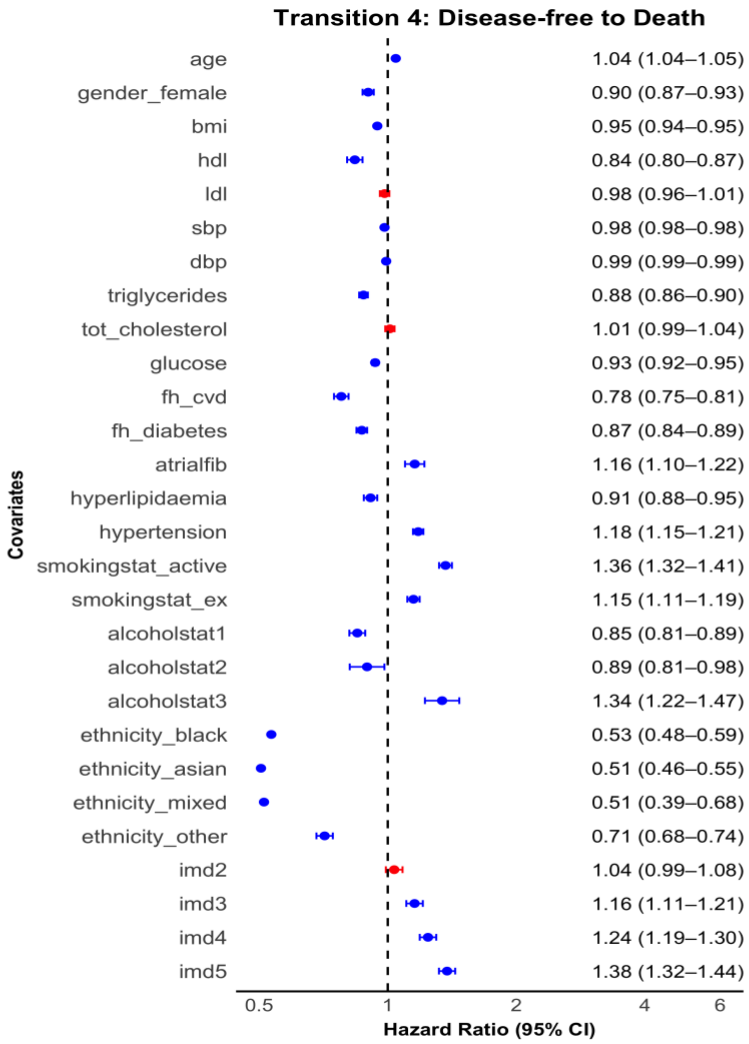


Figure 6.14 Forest plot on hazard ratios (HRs) from Cox regression (transition 1-2)


Transition 3: Disease-free to Stroke
age 1.01 (1.01–1.01)
gender_female 0.98 (0.94–1.02)
bmi 0.95 (0.95–0.95)
hdl 0.76 (0.72–0.81)
ldl 0.97 (0.94–1.01)
sbp 1.02 (1.02–1.02)
dbp 1.00 (1.00–1.00)
triglycerides 0.77 (0.75–0.80)
tot_cholesterol 1.23 (1.19–1.27)
glucose 0.82 (0.80–0.84)
fh_cvd 0.74 (0.70–0.78)
fh_diabetes 0.84 (0.81–0.88)
atrialfib 3.87 (3.68–4.07)
hyperlipidaemia 0.70 (0.66–0.74)
hypertension 0.91 (0.88–0.94)
smokingstat_active 1.16 (1.11–1.22)
smokingstat_ex 1.25 (1.20–1.30)
alcoholstat1 0.88 (0.84–0.94)
alcoholstat2 0.95 (0.84–1.07)
alcoholstat3 1.21 (1.07–1.36)
ethnicity_black 0.73 (0.66–0.81)
ethnicity_asian 0.74 (0.68–0.81)
ethnicity_mixed 0.52 (0.38–0.72)
ethnicity_other 0.75 (0.71–0.79)
imd2 1.05 (0.99–1.11)
imd3 1.09 (1.03–1.15)
imd4 1.12 (1.06–1.19)
imd5 1.20 (1.13–1.27)
Covariates
0.5
1
2
4
6
Hazard Ratio (95% CI)


Transition 4: Disease-free to Death
age 1.04 (1.04–1.05)
gender_female 0.90 (0.87–0.93)
bmi 0.95 (0.94–0.95)
hdl 0.84 (0.80–0.87)
ldl 0.98 (0.96–1.01)
sbp 0.98 (0.98–0.98)
dbp 0.99 (0.99–0.99)
triglycerides 0.88 (0.86–0.90)
tot_cholesterol 1.01 (0.99–1.04)
glucose 0.93 (0.92–0.95)
fh_cvd 0.78 (0.75–0.81)
fh_diabetes 0.87 (0.84–0.89)
atrialfib 1.16 (1.10–1.22)
hyperlipidaemia 0.91 (0.88–0.95)
hypertension 1.18 (1.15–1.21)
smokingstat_active 1.36 (1.32–1.41)
smokingstat_ex 1.15 (1.11–1.19)
alcoholstat1 0.85 (0.81–0.89)
alcoholstat2 0.89 (0.81–0.98)
alcoholstat3 1.34 (1.22–1.47)
ethnicity_black 0.53 (0.48–0.59)
ethnicity_asian 0.51 (0.46–0.55)
ethnicity_mixed 0.51 (0.39–0.68)
ethnicity_other 0.71 (0.68–0.74)
imd2 1.04 (0.99–1.08)
imd3 1.16 (1.11–1.21)
imd4 1.24 (1.19–1.30)
imd5 1.38 (1.32–1.44)
Covariates
0.5
1
2
4
6
Hazard Ratio (95% CI)


Significant TRUE FALSE

Figure 6.15 Forest plot on hazard ratios (HRs) from Cox regression (transition 3-4)

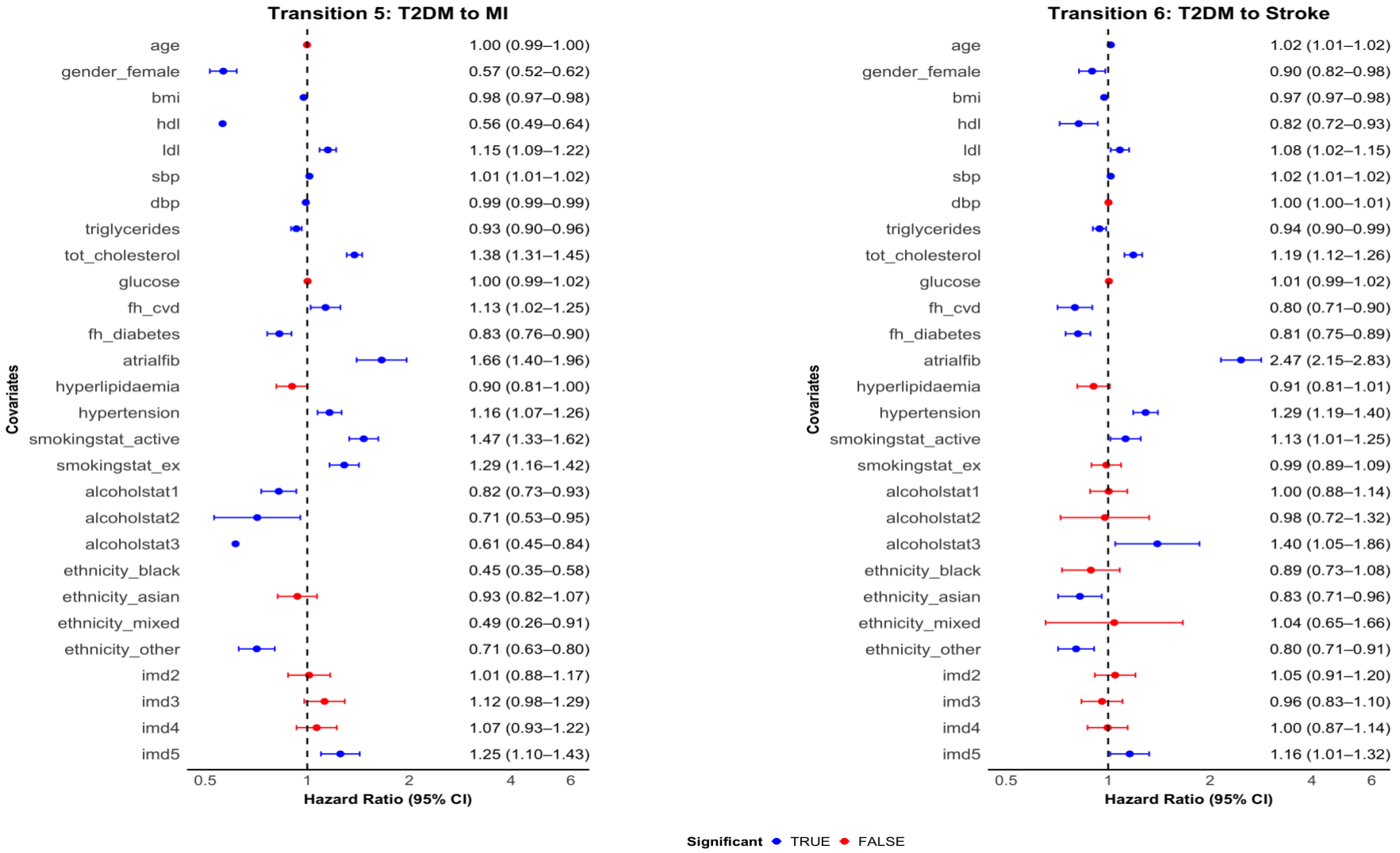
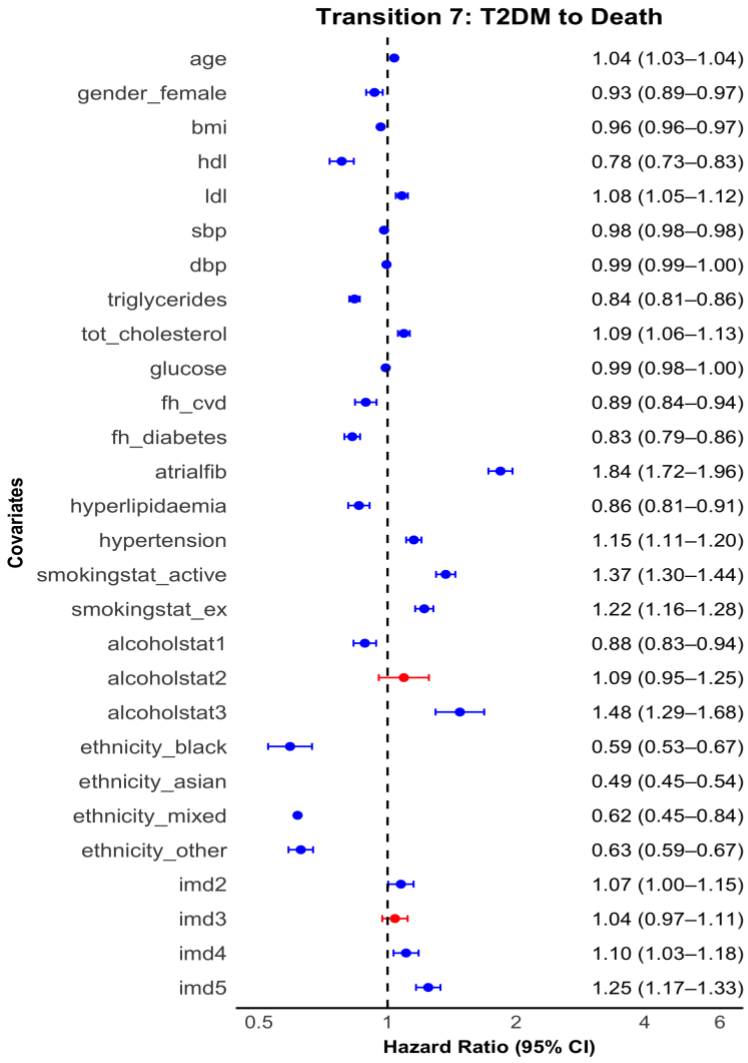
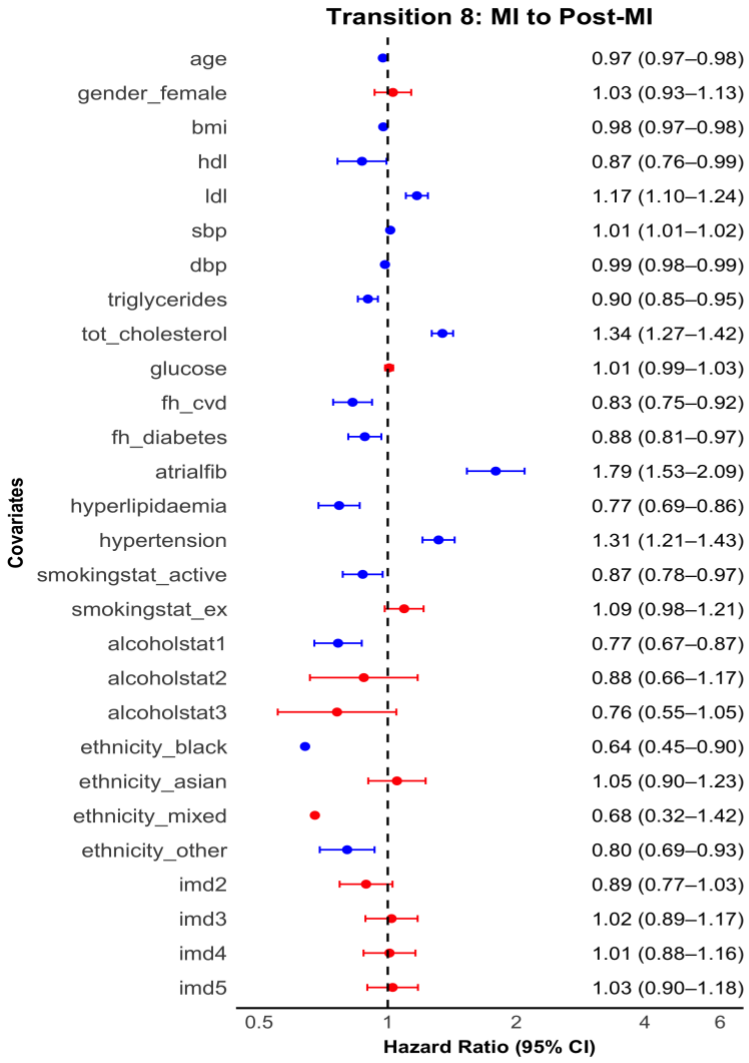


Figure 6.16 Forest plot on hazard ratios (HRs) from Cox regression (transition 5-6)


Transition 7: T2DM to Death
Covariates
age 1.04 (1.03–1.04)
gender_female 0.93 (0.89–0.97)
bmi 0.96 (0.96–0.97)
hdl 0.78 (0.73–0.83)
ldl 1.08 (1.05–1.12)
sbp 0.98 (0.98–0.98)
dbp 0.99 (0.99–1.00)
triglycerides 0.84 (0.81–0.86)
tot_cholesterol 1.09 (1.06–1.13)
glucose 0.99 (0.98–1.00)
fh_cvd 0.89 (0.84–0.94)
fh_diabetes 0.83 (0.79–0.86)
atrialfib 1.84 (1.72–1.96)
hyperlipidaemia 0.86 (0.81–0.91)
hypertension 1.15 (1.11–1.20)
smokingstat_active 1.37 (1.30–1.44)
smokingstat_ex 1.22 (1.16–1.28)
alcoholstat1 0.88 (0.83–0.94)
alcoholstat2 1.09 (0.95–1.25)
alcoholstat3 1.48 (1.29–1.68)
ethnicity_black 0.59 (0.53–0.67)
ethnicity_asian 0.49 (0.45–0.54)
ethnicity_mixed 0.62 (0.45–0.84)
ethnicity_other 0.63 (0.59–0.67)
imd2 1.07 (1.00–1.15)
imd3 1.04 (0.97–1.11)
imd4 1.10 (1.03–1.18)
imd5 1.25 (1.17–1.33)
0.5
1
2
4
6
Hazard Ratio (95% CI)


Transition 8: MI to Post-MI
Covariates
age 0.97 (0.97–0.98)
gender_female 1.03 (0.93–1.13)
bmi 0.98 (0.97–0.98)
hdl 0.87 (0.76–0.99)
ldl 1.17 (1.10–1.24)
sbp 1.01 (1.01–1.02)
dbp 0.99 (0.98–0.99)
triglycerides 0.90 (0.85–0.95)
tot_cholesterol 1.34 (1.27–1.42)
glucose 1.01 (0.99–1.03)
fh_cvd 0.83 (0.75–0.92)
fh_diabetes 0.88 (0.81–0.97)
atrialfib 1.79 (1.53–2.09)
hyperlipidaemia 0.77 (0.69–0.86)
hypertension 1.31 (1.21–1.43)
smokingstat_active 0.87 (0.78–0.97)
smokingstat_ex 1.09 (0.98–1.21)
alcoholstat1 0.77 (0.67–0.87)
alcoholstat2 0.88 (0.66–1.17)
alcoholstat3 0.76 (0.55–1.05)
ethnicity_black 0.64 (0.45–0.90)
ethnicity_asian 1.05 (0.90–1.23)
ethnicity_mixed 0.68 (0.32–1.42)
ethnicity_other 0.80 (0.69–0.93)
imd2 0.89 (0.77–1.03)
imd3 1.02 (0.89–1.17)
imd4 1.01 (0.88–1.16)
imd5 1.03 (0.90–1.18)
0.5
1
2
4
6
Hazard Ratio (95% CI)


Significant TRUE FALSE

Figure 6.17 Forest plot on hazard ratios (HRs) from Cox regression (transition 7-8)

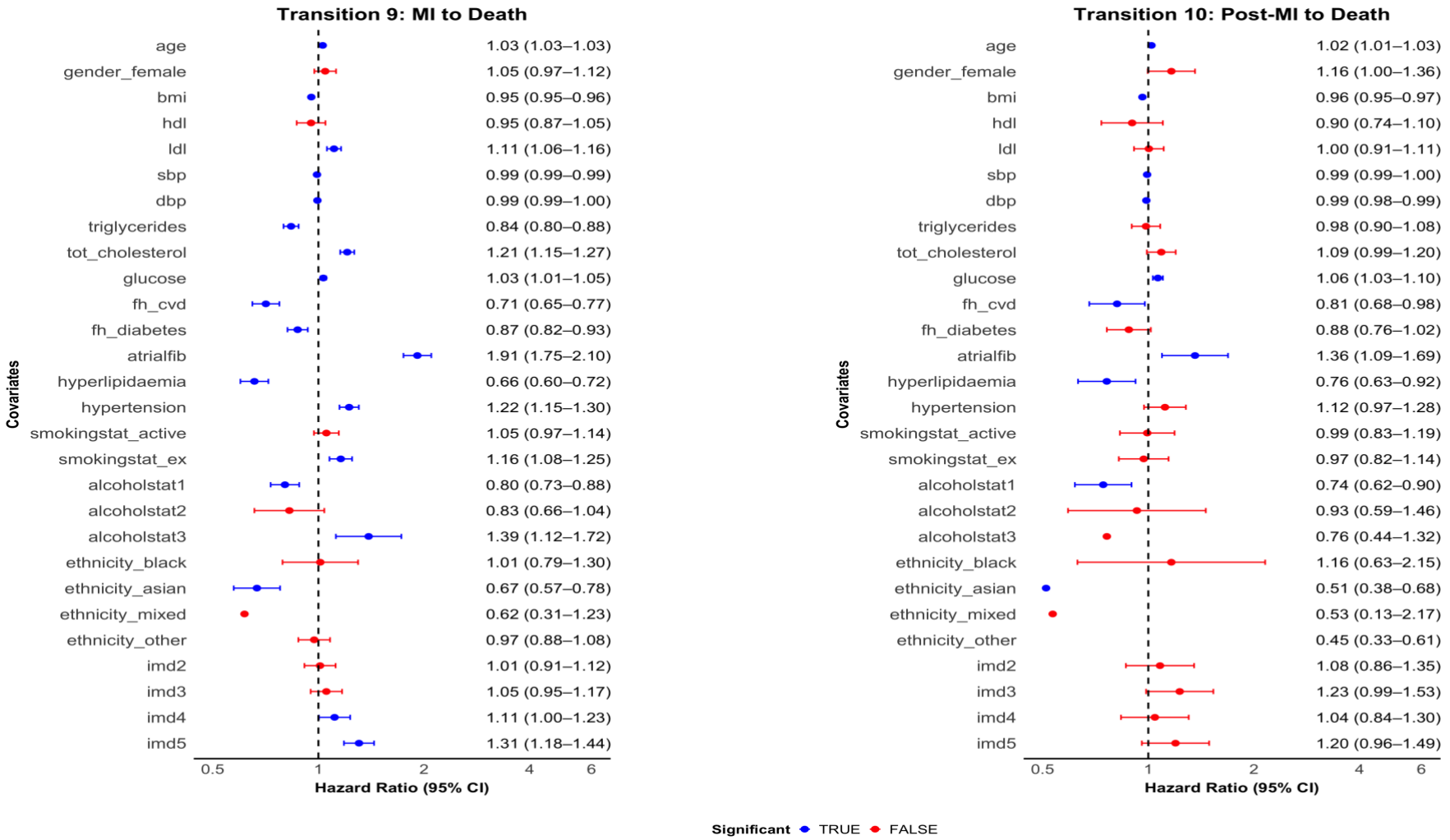
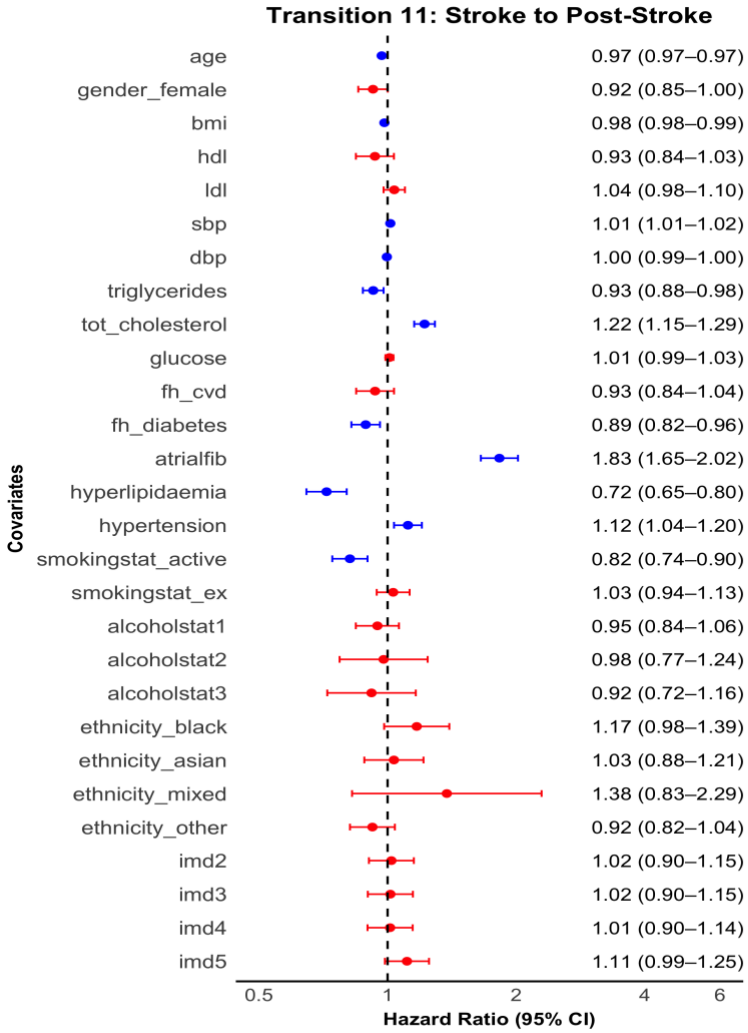
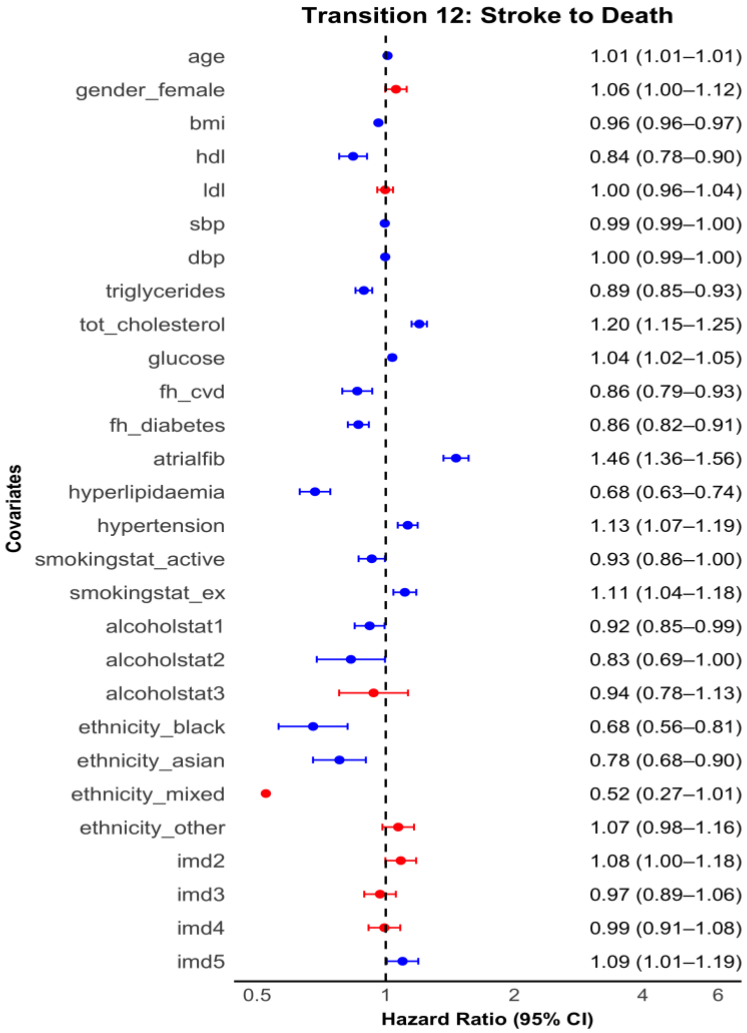


Figure 6.18 Forest plot on hazard ratios (HRs) from Cox regression (transition 9-10)


Transition 11: Stroke to Post-Stroke
Covariates
age 0.97 (0.97–0.97)
gender_female 0.92 (0.85–1.00)
bmi 0.98 (0.98–0.99)
hdl 0.93 (0.84–1.03)
ldl 1.04 (0.98–1.10)
sbp 1.01 (1.01–1.02)
dbp 1.00 (0.99–1.00)
triglycerides 0.93 (0.88–0.98)
tot_cholesterol 1.22 (1.15–1.29)
glucose 1.01 (0.99–1.03)
fh_cvd 0.93 (0.84–1.04)
fh_diabetes 0.89 (0.82–0.96)
atrialfib 1.83 (1.65–2.02)
hyperlipidaemia 0.72 (0.65–0.80)
hypertension 1.12 (1.04–1.20)
smokingstat_active 0.82 (0.74–0.90)
smokingstat_ex 1.03 (0.94–1.13)
alcoholstat1 0.95 (0.84–1.06)
alcoholstat2 0.98 (0.77–1.24)
alcoholstat3 0.92 (0.72–1.16)
ethnicity_black 1.17 (0.98–1.39)
ethnicity_asian 1.03 (0.88–1.21)
ethnicity_mixed 1.38 (0.83–2.29)
ethnicity_other 0.92 (0.82–1.04)
imd2 1.02 (0.90–1.15)
imd3 1.02 (0.90–1.15)
imd4 1.01 (0.90–1.14)
imd5 1.11 (0.99–1.25)
0.5
1
2
4
6
Hazard Ratio (95% CI)


Transition 12: Stroke to Death
Covariates
age 1.01 (1.01–1.01)
gender_female 1.06 (1.00–1.12)
bmi 0.96 (0.96–0.97)
hdl 0.84 (0.78–0.90)
ldl 1.00 (0.96–1.04)
sbp 0.99 (0.99–1.00)
dbp 1.00 (0.99–1.00)
triglycerides 0.89 (0.85–0.93)
tot_cholesterol 1.20 (1.15–1.25)
glucose 1.04 (1.02–1.05)
fh_cvd 0.86 (0.79–0.93)
fh_diabetes 0.86 (0.82–0.91)
atrialfib 1.46 (1.36–1.56)
hyperlipidaemia 0.68 (0.63–0.74)
hypertension 1.13 (1.07–1.19)
smokingstat_active 0.93 (0.86–1.00)
smokingstat_ex 1.11 (1.04–1.18)
alcoholstat1 0.92 (0.85–0.99)
alcoholstat2 0.83 (0.69–1.00)
alcoholstat3 0.94 (0.78–1.13)
ethnicity_black 0.68 (0.56–0.81)
ethnicity_asian 0.78 (0.68–0.90)
ethnicity_mixed 0.52 (0.27–1.01)
ethnicity_other 1.07 (0.98–1.16)
imd2 1.08 (1.00–1.18)
imd3 0.97 (0.89–1.06)
imd4 0.99 (0.91–1.08)
imd5 1.09 (1.01–1.19)
0.5
1
2
4
6
Hazard Ratio (95% CI)


Significant TRUE FALSE

Figure 6.19 Forest plot on hazard ratios (HRs) from Cox regression (transition 11-12)

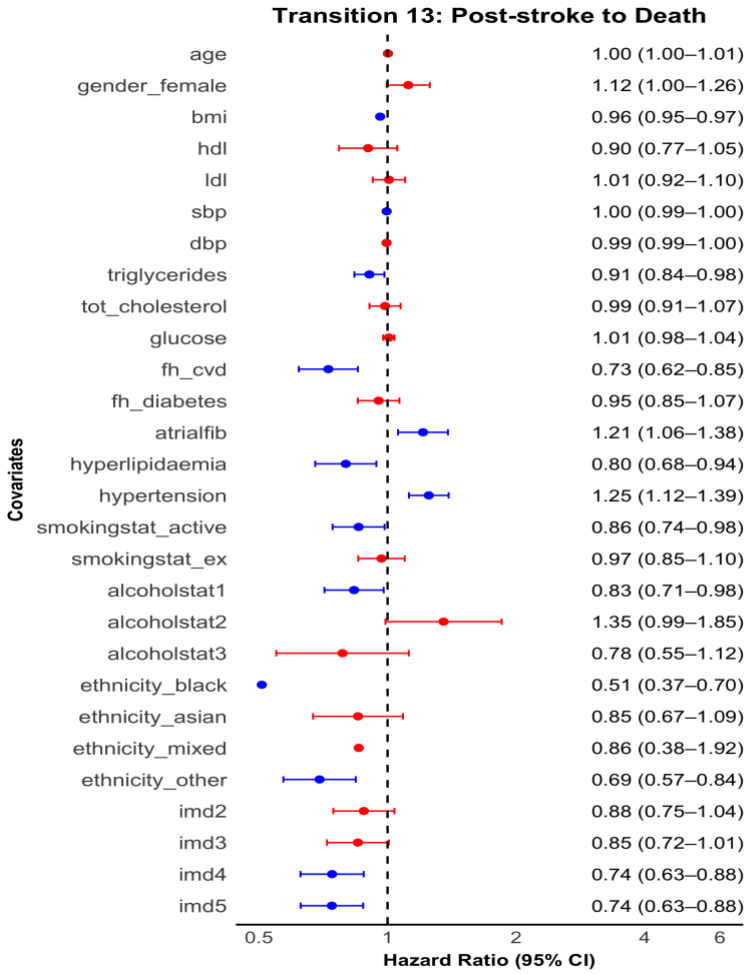


Figure 6.20 Forest plot on hazard ratios (HRs) from Cox regression (transition 13)

Covariates such as age, blood pressure, BMI, and glucose showed the strongest and most consistent evidence of non-proportionality, particularly in transitions to death (e.g., Transition 4, 7, 9, 12, and 13), where risk is likely to evolve with time since prior events. This implies that the hazard ratios associated with these covariates change meaningfully over the disease trajectory. Consequently, modelling approaches that incorporate time-by-covariate interactions or parametric (including flexible) methods may be necessary to accurately represent these dynamics. For instance, age showed highly significant deviations in nearly every transition as did blood pressure across mortality and post-event pathways. While time-updated covariates were included to reflect changing exposure values, they do not address time-varying effects of those covariates. That is, a time-updated value for glucose may capture current status, but if the hazard ratio associated with glucose changes over time, this would still violate the PH assumption. Therefore, additional diagnostics led to testing interactions between covariates and time or exploring stratification by covariates such as age group or sex.

Where violations of the proportional hazards assumption are persistent and clinically plausible, further modelling strategies should be employed. Parametric survival models, which accommodate time-varying hazards and covariate-time interactions potentially useful in transitions with complex or non-monotonic risk patterns, as observed in several post-event and mortality transitions.

It is important to note a limitation in the standard visual interpretation of Schoenfeld residual plots. When time-varying effects are subtle, for example: a log HR that increases by only 0.01 per year—the resulting curve may appear nearly flat. This is due to the plot's scale being adjusted to accommodate all transitions and long follow-up periods (e.g., over 30 years), which can obscure small but meaningful trends. Furthermore, the residual dots may overlap or obscure the fitted line, making it difficult to visually detect time-dependent effects. However, such gradual changes can still accumulate to meaningful levels over time (e.g., a 0.1 increase over ten years) and may be statistically significant, particularly in large datasets. For this reason, smoothed

visualisations of time-varying coefficients are provided alongside standard Schoenfeld plots to better illustrate these subtle trends. (Appendix 6).

In addition to Schoenfeld residuals test, residuals were also checked to evaluate the functional form of the continuous covariates using Martingale residuals test. Martingale residuals can help to detect an appropriate non-linear relationship between continuous predictors and the hazard. This diagnostic is particularly important in the present analysis due to the inclusion of several continuous variables and the evidence of time-varying effects.

Overall Martingale residuals were plotted against the linear predictor for each transition, also the plot that illustrated the residual in each continuous covariate (Appendix 7). In a well-specified model, residuals should be symmetrically scattered around zero with no clear trend, and the smoothed loess line should remain flat along the horizontal axis. Deviations from this expectation may indicate that the model fails to capture the true hazard structure, often due to non-linearity or omitted variable interactions.

In general, Transitions 1 to 4 (from the Disease-free state) show clear non-linear trends, particularly in Transitions 1 (to T2DM), 2 (to MI), and 3 (to Stroke), where the smoothed lines show strong curvature. This suggests that the combined effects of covariates are not adequately modelled by the linear predictor, likely due to non-linear biomarker-risk relationship. Transition 4 (to Death) also shows notable deviation, though with slightly less curvature, reinforcing the complexity of mortality modelling from a baseline healthy state.

Transitions 5 to 7 (from T2DM) and Transitions 8 to 13 (post-MI and post-Stroke pathways) display moderate to mild non-linearity. For instance, Transition 9 (MI to Death) and Transition 10 (Post-MI to Death) reveal upward-sloping trends in the smoothed residuals, indicating possible underestimation of risk in higher-risk patients. Some later transitions (e.g., Transitions 11–13) show relatively stable residual patterns, suggesting that the linear predictor performs more adequately in later disease stages, potentially due to more homogeneous risk profiles. All covariates have highly significant p-values ($p < 0.001$), confirming their

relevance, but also suggesting the current model may not be using the most appropriate functional form for them.

Based on this test, overall model structure appears sound, the observed non-linearities indicate that several transitions, particularly form early state (disease-free) may require more flexible modelling such as Royston-Parmar (RP) modelling to better capture the relationships between biomarkers and event risk.

Both Schoenfeld and Martingale tests results are useful tools to evaluate the Cox model. For Schoenfeld, these findings do not indicate a failure of the model but rather affirm its capacity to detect and reflect true temporal variation in covariate effects. Similarly, Martingale residual test does not invalidate the model, rather they provide critical insight that the relationship between biomarkers and risks may be complex and non-linear, that indicate more flexible modelling may needed to strengthen predictive accuracy and its clinical relevance.

6.4.3 Parametric modelling results

Fitting parametric models

Parametric models are integral to survival analysis, particularly in healthcare settings where extrapolation beyond observed follow-up is often required. Unlike semi-parametric models such as the Cox model, parametric models explicitly define the baseline hazard and the functional relationship between covariates and survival outcomes.[189,313]

Various distributional forms are employed to describe survival time in this current parametric approach. As previously mentioned, the model selection is typically guided by information criteria like the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). The lowest value of these represents the potentially best-fitting model (Table 6.5).

Information criteria test such as AIC/BIC are useful for comparing the relative fit of candidate models for a specific state transition, however, they may not fully capture overall model adequacy, especially when multiple models fit the data reasonably well.[349] Hence, the current model selection was supplemented with additional diagnostic, including visual inspection by comparing fitted parametric curves with Kaplan-Meier (KM) estimates (Figure 6.21), as well as residual-based diagnostics.

Based on the parametric model evaluation, AIC/BIC selections were consistent across all transitions. The Log-normal distribution emerged as the optimal fit for the initial transitions from the disease-free state (Transitions 1–3). This aligns with the natural progression of cardiometabolic diseases, where risk accelerates gradually over time, particularly as metabolic dysregulation worsens with aging or cumulative exposures, or potential competing risks. (Figure 6.21)

In Transition 4 (Disease free to Death), The Weibull distribution is the best fitting one, showing its flexibility by allowing for increasing, decreasing or constant hazard. For all-cause mortality from a disease state, this aligns with the known pattern of increasing mortality risk with age.

For T2DM-related events and death (Transitions 5–7), the Gompertz distribution characterised by an exponentially increasing hazard seems appropriate, as it reflects the rising likelihood of complications with longer disease duration. In individuals with diabetes, prolonged exposure to hyperglycaemia contributes to progressive vascular and neurological damage, elevating the risk of adverse outcomes over time.

Table 6.3 AIC/BIC score (parametric models)

Transition and distribution tests	Exponential		Gamma		Gen_Gamma		Gompertz		Log-logistic		Log-normal		Weibull	
	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
T1: Disease-free → T2DM	599,580	599,873	510,897	511,201	488,768	489,082	508,818	509,122	487,671	487,975	487,616	487,920	498,031	498,335
T2: Disease-free→ MI	156,986	157,279	143,595	143,899	143,114	143,428	139,070	139,374	136,964	137,268	136,282	136,586	137,223	137,527
T3: Disease-free→ Stroke	163,829	164,123	142,433	142,737	141,277	141,590	143,939	144,243	141,838	142,141	141,230	141,533	142,055	142,358
T4: Disease-free→ Death	250,743	251,036	214,451	214,755	207,388	207,702	204,652	204,955	204,428	204,732	206,064	206,368	203,633	203,937
T5: T2DM→ MI	28,965	29,226	28,969	29,238	29,616	29,894	28,952	29,221	28,967	29,236	29,169	29,438	28,967	29,237
T6: T2DM→ Stroke	28,246	28,506	28,248	28,517	28,966	29,245	28,223	28,492	28,250	28,519	28,458	28,726	29,246	28,515
T7: T2DM→ Death	85,209	85,470	86,356	86,625	89,840	85,219	84,833	85,102	85,792	86,061	87,545	87,814	85,182	85,451
T8: MI→post-MI	18,252	18,472	15,033	15,260	14,714	14,949	16,885	17,112	14,915	15,142	14,775	15,002	14,971	15,198
T9: MI→Death	27,729	27,948	23,511	23,738	23,513	23,747	26,895	27,122	23,748	23,975	23,892	24,119	23,555	23,782
T10: Post-MI→Death	5,576	5,741	4,908	5,079	4,910	5,087	5,396	5,566	4,958	5,129	4,993	5,163	4,914	5,058
T11: Stroke→post-stroke	21,911	22,131	17,199	17,427	16,648	16,883	19,661	19,888	16,987	17,215	16,771	16,999	17,090	17,318
T12: Stroke→Death	34,967	35,187	27,707	27,935	27,450	27,685	32,834	33,062	27706	27,934	27,615	27,843	27,528	27,755
T13: Post-Stroke→Death	8,483	8,657	7,359	7,539	3,802	3,988	8,029	8,209	7,401	7,580	7,408	7,588	7,357	7,536

AIC: Akaike Information Criterion; BIC: Bayesian Information Criterion
*Grey highlighted: lowest score

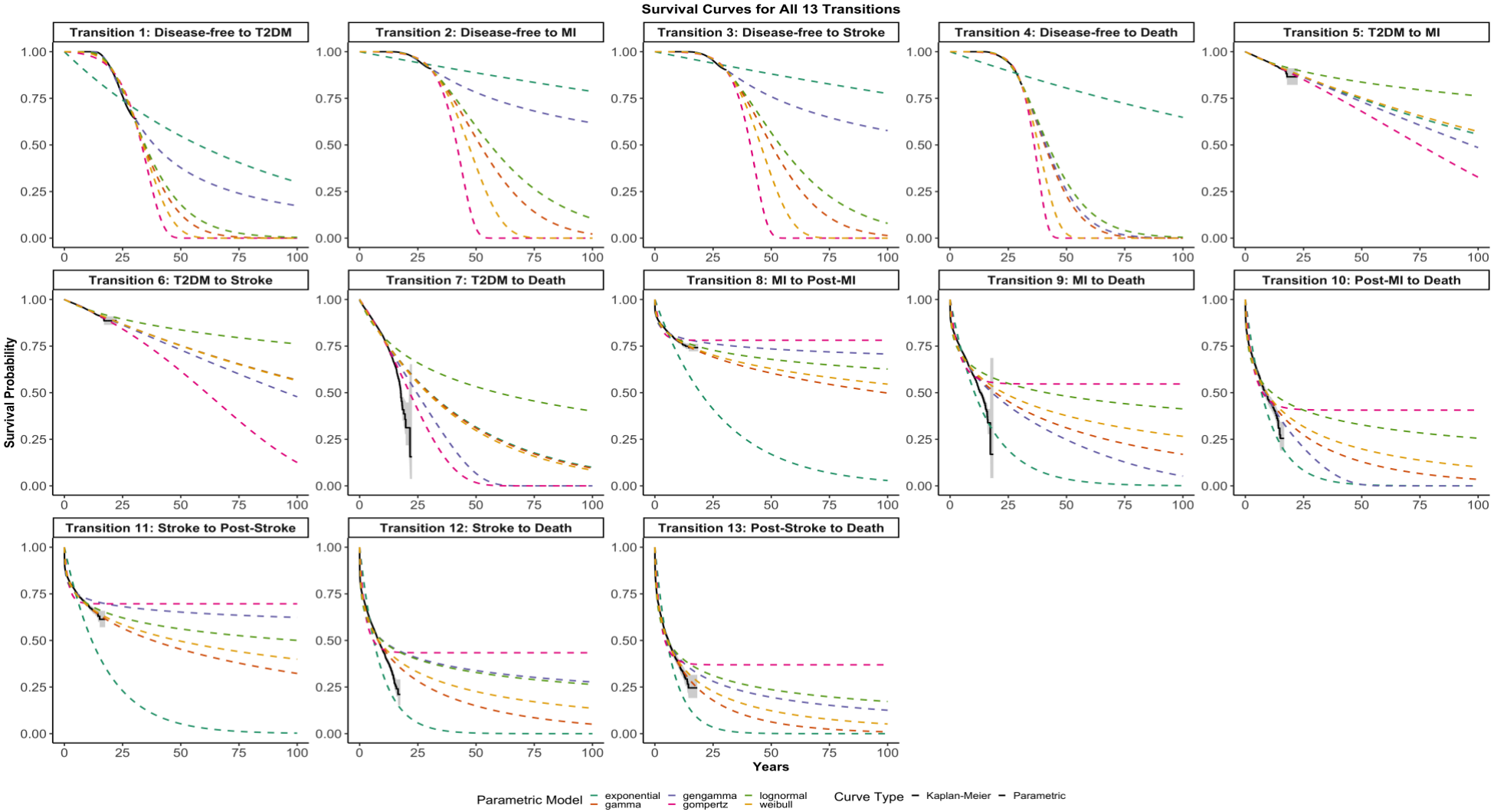


Figure 6.21 Standard Parametric vs Kaplan-Meier survival curves

For post-event transitions, such as those following MI or stroke (Transitions 8 and 11–13), the Generalised Gamma distribution consistently offered the best fit. This is supported by studies that report non-constant hazards, for example the risk of death after MI is highest in the first 30 days and then declines, or as defined as post-acute CVD events.[350,351]

It can be seen that transition 13 likely overfit (Table 6.3), with a very large difference score compared to other distributions. It might be due to small event counts in this transition render likelihood-based criteria unstable or overly responsive to model complexity, potentially affecting convergence. Nevertheless, visual inspection of this transition fit confirms that Generalised Gamma closely follows the observed KM curve throughout the follow-up period.

In terms of extrapolation, figure 6.21 demonstrates that several models fit the observed data similarly within the follow-up range, but diverge in their extrapolated tails. For instance, exponential and Gompertz distributions often showed either persistent survival (flattened curves) or overly rapid decline, particularly visible in transitions 11–13 (stroke-related mortality). These patterns were judged implausible based on known disease trajectories. Conversely, the log-normal and Weibull distributions produced extrapolations that more closely aligned with clinical expectations, such as gradual risk increase in early transitions and steeper declines in mortality-related transitions. The generalised gamma distribution also performed well in transitions 11-13, offering flexible shapes that could mimic the more clinically credible behaviour of log-normal and Weibull models, while also fitting the observed data reasonably. These visual assessments supported the final model selection process, prioritising both goodness-of-fit and credible long-term projection.

Model diagnostics

Cox-Snell residual plots were generated to assess the goodness-of-fit of the final parametric models across all 13 transitions. A well-fitting model is expected to produce a cumulative hazard plot of residuals that follows a 45-degree line (i.e., the unit exponential distribution). Overall, most transitions showed reasonable alignment with this expectation, suggesting acceptable model calibration for the

majority of transitions, particularly in early- to mid-follow-up periods. Transitions such as Transition 3, 4, 9, 10 and 12 demonstrated near-linear residual behaviour, indicating a good fit between predicted and observed hazards.

Transitions 2, 3, 7, 8, 11 exhibit minor deviations, particularly at the tails. These may indicate limited data at later time points or slight misspecification. T1, T5, T6, T13 show more noticeable divergence from the ideal line. This suggests the chosen parametric models might not fully capture the shape of the hazard over time, perhaps due to time-varying risks or heterogeneity in event patterns.

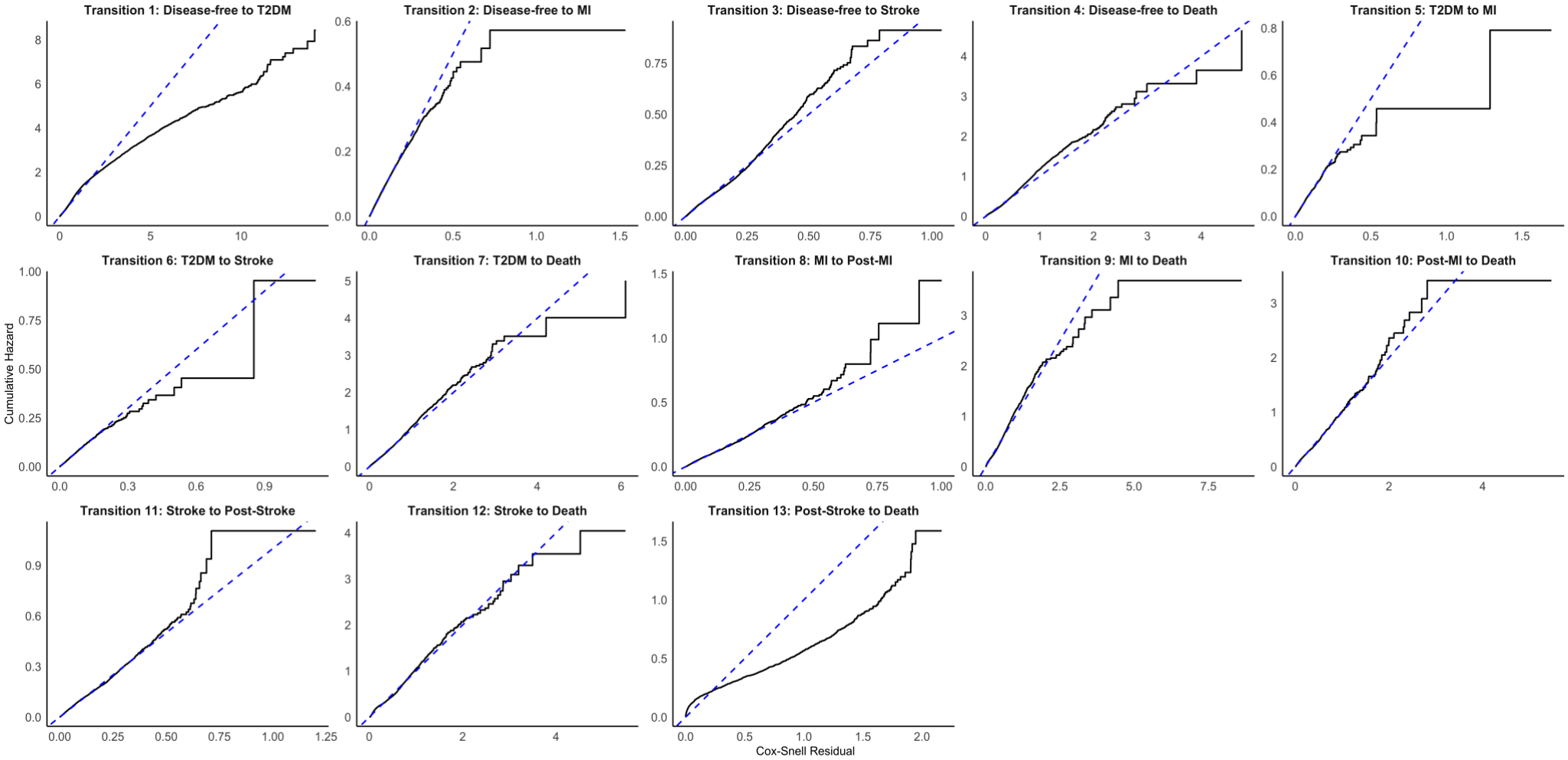
6.4.4 Flexible parametric results

Fitting flexible parametric models

While standard parametric distributions offer a foundation for extrapolation, Figure 6.21 demonstrates cases where some models inadequatly capture the empirical survival patterns, especially in the early follow-up periods (see the divergence on transition 7, 9, 10, 12), where rigid assumptions about hazard shapes limit model fit.

The flexible parametric survival model, particularly the Royston-Parmar (RP) model is introduced to address limitations of standard parametric approaches. They model the log cumulative hazard as a smooth function of time using restricted cubic splines [306], enabling the hazard to follow more realistic, non-monotonic patterns often observed in chronic and multi-phase disease processes like cardiometabolic disease.

This is achieved without introducing excessive model complexity or sacrificing interpretability. RP models allow the hazard function to bend smoothly at predefined knot points. The selection of these knot numbers was guided by the principle of parsimony: while additional knots increase the model's ability to capture complex hazard shapes, they also risk overfitting and reduced interpretability. Fewer knots (e.g., 1) yield smoother and more stable hazard curves, which are easier to interpret and generalise, particularly when sample sizes are moderate or follow-up is limited. Conversely, more knots (e.g., 3) provide greater flexibility to accommodate non-linear hazard trajectories where supported by the data.



Blue dashed line = expected if model fits perfectly

Figure 6.22 Cox-Snell Residual for best parametric models

In this thesis, limiting the knot range to 1–3 allowed for exploration of model performance under increasing complexity without excessive computational burden or risk of instability. Model selection was based on a combination of statistical criteria (AIC and BIC) and visual inspection of survival and hazard curves to ensure that improvements in fit did not come at the cost of overfitting or loss of clinical interpretability (Table 6.6).

Overall, increasing the number of knots generally improves model fit, as indicated by lower AIC/BIC values. RP models provided better fit than standard parametric models in transitions with complex or non-monotonic hazard patterns (e.g., Transition 1, 2, 3, 7), particularly when using higher knot values (k=2 or 3), aligning more closely with Kaplan-Meier estimates. For mortality-related transitions (Transition 9-13), standard models often diverged over time, while RP models more followed observed survival trends. In transitions with limited follow-up (e.g., Transition 6,8,11), it performed similarly to parametric, though RP models offered greater flexibility and reduced risk of misfit.

Table 6.4 AIC/BIC score (flexible parametric models)

Transition and knots simulation	**k=1**		**k=2**		**k=3**	
	AIC	BIC	AIC	BIC	AIC	BIC
T1: Disease-free → T2DM	487,029	487,343	486,511	486,835	484,927	485,261
T2: Disease-free→ MI	134,892	135,206	138,568	138,892	134,928	135,262
T3: Disease-free→ Stroke	139,647	139,960	139,960	140,284	139,706	140,040
T4: Disease-free→ Death	226,039	226,069	225,531	225,572	225,156	225,206
T5: T2DM→ MI	28,911	29,189	28,913	29,200	30,126	30,171
T6: T2DM→ Stroke	29,249	29,276	29,246	29,282	29,247	29,292
T7: T2DM→ Death	91,802	91,829	91,570	91,606	91,538	91,583
T8: MI→post-MI	14,866	15,101	14,423	14,665	14,365	14,615
T9: MI→Death	23,550	23,785	22,708	22,950	22,628	22,878
T10: Post-MI→Death	4,916	5,092	4,848	5,030	4,844	5,032
T11: Stroke→post-stroke	16,773	17,009	17,005	17,036	16,998	17,036
T12: Stroke→Death	27,454	27,689	26,246	26,489	26,201	26,451
T13: Post-Stroke→Death	7,349	7,534	7,258	7,450	7,250	7,448

In addition, some transitions, especially those occurring later in the disease progression, such as from post-MI or post-stroke to death show only marginal improvements or even slight deterioration in model fit with additional knots, suggesting simpler hazard structures or limited event data.

Model diagnostics

Figure 6.24 illustrates Cox-Snell residual diagnostics for RP models across all 13 transitions show generally improved model fit compared to standard parametric counterparts. In most transitions, the cumulative hazard (black line) closely follows the expected 45-degree reference line (blue dashed), indicating a good approximation to the observed data. Slight deviations are visible in transitions with sparse events or extended follow-up (e.g., Transition 11-13). However, the flexible structure of RP models makes them appealing for extrapolation scenarios where standard parametric forms may be too restrictive. Therefore, their use is justified not solely by fit, but by their capacity to reflect plausible hazard trajectories beyond observed data, especially when supported by clinical or external validation.

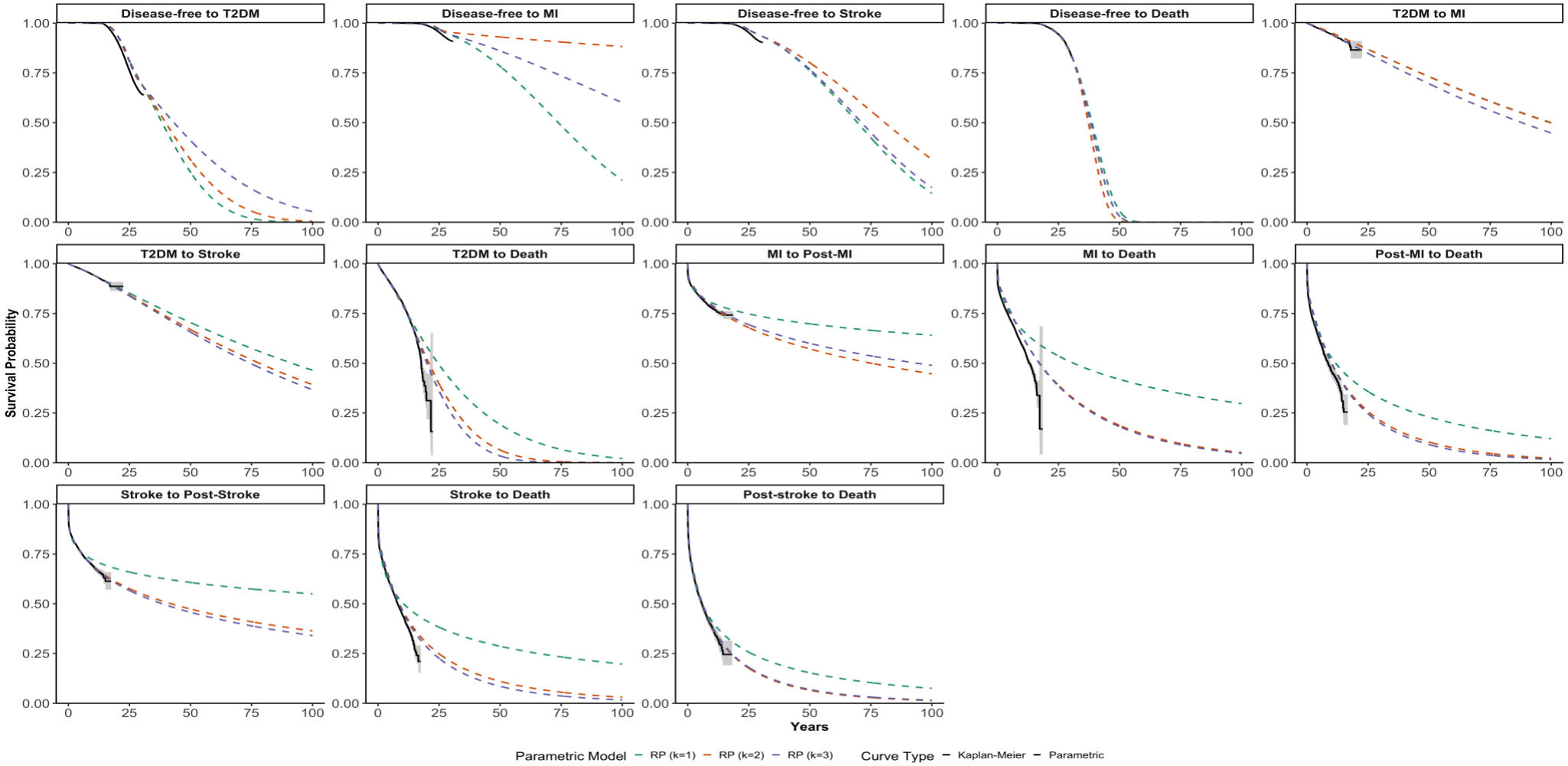

Disease-free to T2DM
Disease-free to MI
Disease-free to Stroke
Disease-free to Death
T2DM to MI
T2DM to Stroke
T2DM to Death
MI to Post-MI
MI to Death
Post-MI to Death
Stroke to Post-Stroke
Stroke to Death
Post-stroke to Death
Survival Probability
Years
Parametric Model
RP (k=1)
RP (k=2)
RP (k=3)
Curve Type
Kaplan-Meier
Parametric

Figure 6.23 Flexible Parametric vs Kaplan-Meier survival plot

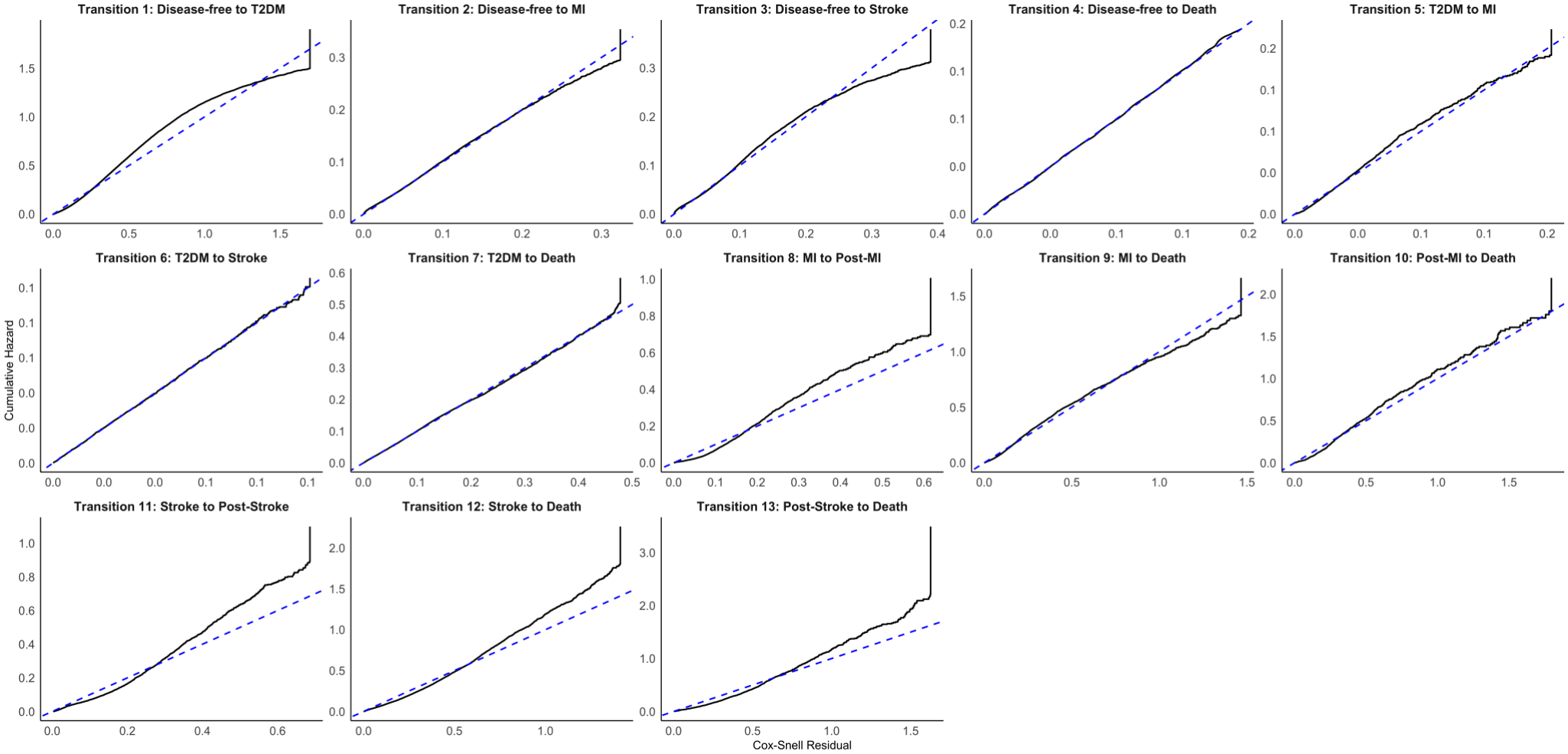


Figure 6.24 Cox-Snell Residual for Royston-Parmar

6.4.5 Incorporating semi-Markov framework for extrapolation

The semi-Markov framework provides a structured approach to modelling multi-state processes where the transition hazard depends on how long an individual has remained in a particular state (sojourn time).[260] This differs from traditional Markov models, which assume transition hazards depend only on the current state and baseline time. In this context, parametric survival models (both standard and flexible) define the underlying hazard function, shaping how transition risks evolve over time.

For example, in cardiometabolic modelling, the risk of death after a MI or stroke may change depending on how long a patient remains in the post-event state. This dynamic cannot be captured using baseline time alone but can be modelled effectively with a semi-Markov approach that incorporates parametric hazard functions. In simpler terms, while parametric models are useful for extrapolating risks beyond observed data due to their well-defined hazard functions, they are not sufficient to represent the complexities of multi-state disease processes on their own.

Therefore, integrating parametric models within a semi-Markov framework is important for several reasons. First, it ensures the correct time scale is used. In semi-Markov models, time resets at entry into each new state, so hazards depend on the time spent in that state rather than on total follow-up time.[260] Second, parametric models need to be applied to the correct time scale to work effectively. The semi-Markov structure ensures that these models align with sojourn time, which reflects real-world disease progression more accurately than baseline time.[324]

Table 6.5 summarises the final survival model selected for each transition in the CMD Policy Model, which is embedded within a semi-Markov framework. Model selection was based on a combination of statistical goodness-of-fit criteria (AIC and BIC), visual inspection of observed and fitted survival curves, and assessment of clinical plausibility of the implied hazard functions.(see section 6.3 and 6.5)

Table 6.5 Summary of final survival model selected for each transition in the CMD Policy Model

Transition	Standard parametric (distribution)	Flexible Parametric (k=number of knots)
T1: Disease-free → T2DM	Log-normal	k=3
T2: Disease-free→ MI	Log-normal	k=1
T3: Disease-free→ Stroke	Log-normal	k=1
T4: Disease-free→ Death	Log-normal	k=3
T5: T2DM→ MI	Weibull	k=1
T6: T2DM→ Stroke	Gompertz	k=2
T7: T2DM→ Death	Gompertz	k=3
T8: MI→post-MI	Generalised gamma	k=3
T9: MI→Death	Gamma	k=3
T10: Post-MI→Death	Gamma	k=3
T11: Stroke→post-stroke	Generalised gamma	k=1
T12: Stroke→Death	Generalised gamma	k=3
T13: Post-Stroke→Death	Generalised gamma	k=3

In this thesis, the `semiMarkov()` function developed by Williams et al., (2017) [324] was applied, that constructs a sojourn-time-based multi-state model. For each of the 13 transitions, the best-fitting parametric distribution and the optimal number of spline knots for RP models (see sub section 6.4.3 to 6.4.4) were identified and embedded into the function to generate transition-specific estimates.

The model produces estimated cumulative hazards $H(t)$ and state occupancy probabilities over time, both of which can be adjusted based on the chosen distributions and model parameters. These outputs can then be transformed into transition probabilities $P(t)$, enabling the estimation of the likelihood that an individual will move from one health state to another at any given time point.[352]

$$P(t) = 1 - \exp\left(-H(t)\right)$$

(Equation 6.22)

To support interpretation and link back to the model structure described in Chapter 4 and Figure 6.4, the full transition probability matrix of the CMD Policy Model is re-stated below for completeness.

Table 6.6 Transition probability matrix of CMD Policy Model

	Disease-free	T2DM	MI	Post-MI	Stroke	Post-stroke	Death
Disease-free	1- TP_1-TP_2-TP_3-TP_4	TP_1	TP_2	-	TP_3	-	TP_4
T2DM	-	1- TP_5- TP_6- TP_7	TP_5	-	TP_6	-	TP_7
MI	-	-	1-TP_8-TP_9	TP_8	-	-	TP_9
Post-MI	-	-	-	1- TP_{10}	-	-	TP_{10}
Stroke	-	-	-	-	1- TP_{11} -TP_{12}	TP_{11}	TP_{12}
Post-stroke	-	-	-	-	-	1- TP_{13}	TP_{13}
Death	-	-	-	-	-	-	1

Note: TP= Transition probability. 1 = "Disease-free to T2DM", 2 = "Disease-free to MI", 3 = "Disease-free to Stroke", 4 = "Disease-free to Death",5 = "T2DM to MI", 6 = "T2DM to Stroke", 7 = "T2DM to Death", 8 = "MI to Post-MI", 9= "MI to Death", 10 = "Post-MI to Death", 11 = "Stroke to Post-Stroke", 12 = "Stroke to Death", T13 = "Post-stroke to Death"

This transformation enables the estimation of the likelihood that an individual will transition from one health state to another at any given time point. Importantly, these outputs can be used to inform decision-analytic models, such as cost-effectiveness analyses, that require long-term risk estimation across different health states.

While most transition-specific models demonstrated good agreement with observed Kaplan-Meier estimates, a few transitions exhibited overly optimistic survival projections. The inclusion of background mortality serves to correct this bias and produce more realistic survival outcomes across the full disease trajectory.

Background mortality was incorporated to account for the risk of death from causes unrelated to the explicitly modelled disease processes. This was derived from age- and sex-specific national life tables provided by the UK Office for National Statistics (ONS)[353] and applied in addition to the disease-specific mortality transitions.

To ensure realistic mortality estimates, the cumulative background mortality risk (derived from life tables) was applied to proportionally scale down all other state probabilities at each time point. Specifically, after computing transition probabilities from the semi-Markov model, the probability of background (non-disease-related) death was subtracted from the total probability mass. The remaining probabilities for all other states were then rescaled proportionally to ensure that the full set of state probabilities still summed to one (see implementation details on GitHub). This adjustment is particularly important to avoid underestimation of overall mortality, especially in earlier disease stages or among older individuals who may face substantial non-disease-related death risks.

Using this semi-Markov framework, state occupancy probabilities are generated both for standard parametric and flexible parametric models. The simulation starts age 40 years (median age of sample population is 42 years), and all 'healthy' biomarkers and behaviour covariate levels (e.g, non-smoking, alcohol level 0).

It should be noted that this simulation started at age 40, meaning that lifetime projections are over 60 years (standard lifetime horizon). However, the figures that follow show an extrapolation period up to 100 years to help show potential implausible probabilities of being over 100 years old. In practice, analysts may only use a maximum 60 years to model natural disease progression.

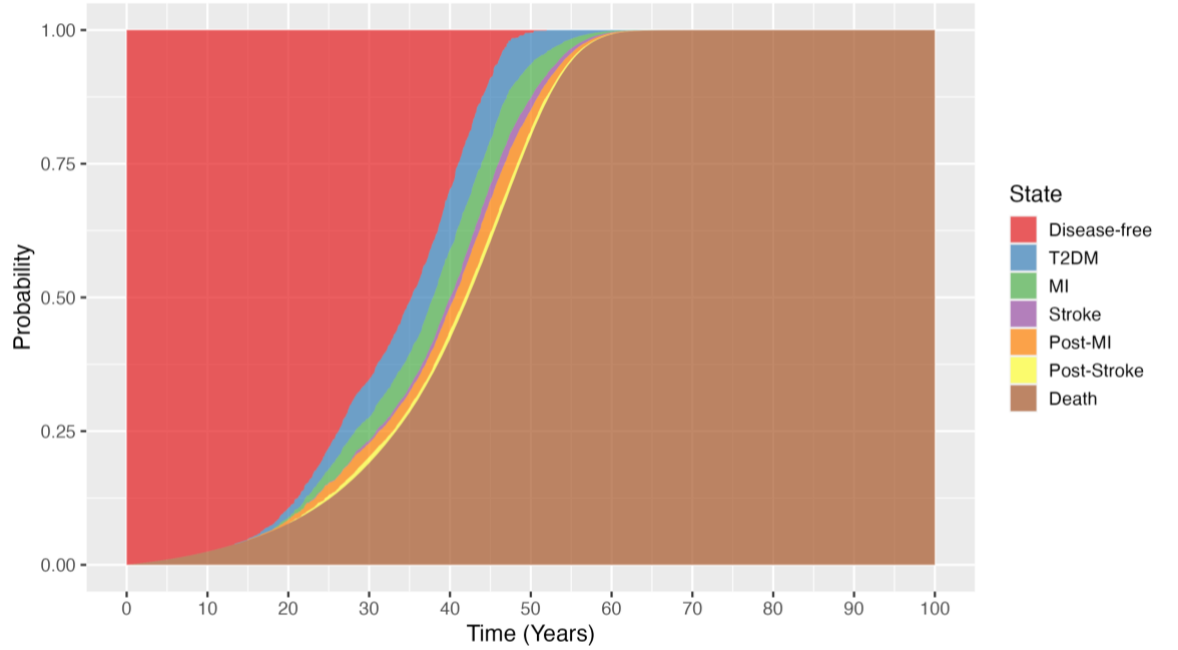


Figure 6.25 State occupancy probabilities (Semi Markov-embedded standard parametric)

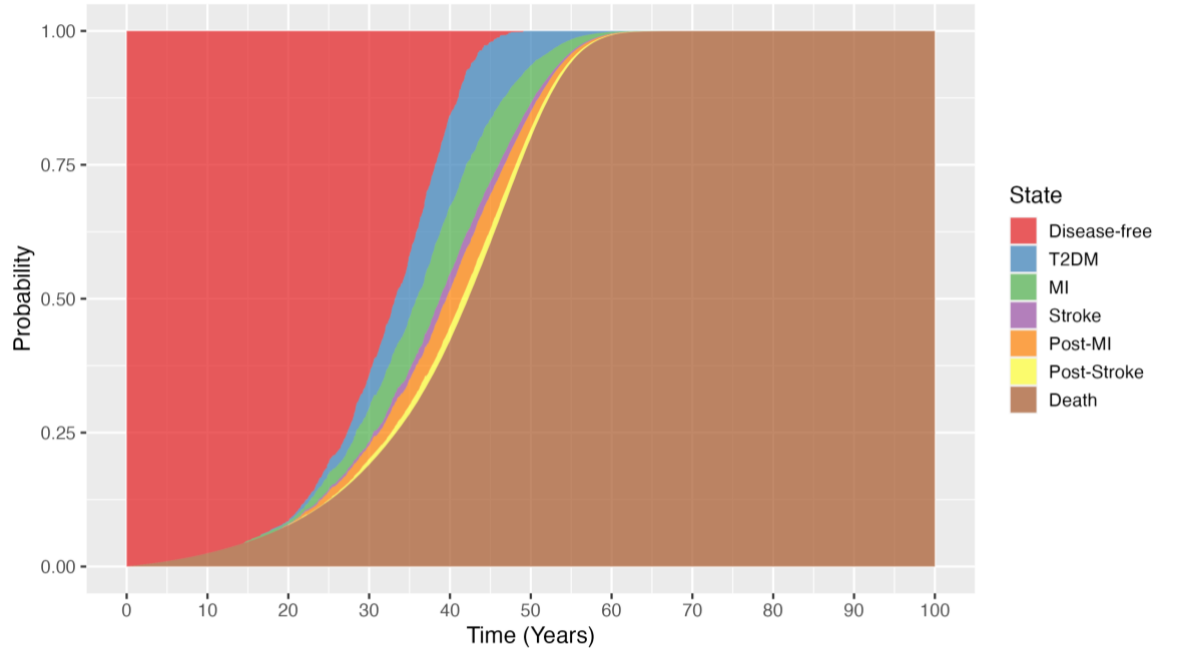


Figure 6.26 State occupancy probabilities (Semi Markov-embedded flexible parametric)

The two figures above (Figure 6.25 and 6.26) represent the estimated state occupancy probabilities over time derived from semi-Markov models using different types of parametric survival functions. The first plot is based on standard parametric distributions, while the second employs flexible parametric models that allow greater adaptability in hazard shapes over time.

In the standard model, transitions between states appear more abrupt and follow a relatively uniform pattern, with the disease-free state move to the next states earlier. This approach benefits from simplicity and computational efficiency, making it suitable for settings where hazard patterns are expected to follow known distributions.

In contrast, the flexible parametric model displays smoother and more gradual transitions, particularly in the timing of disease onset and mortality. People remain in disease-free state slightly longer, and progression through conditions such as T2DM, MI, and stroke occurs more slowly before reaching death state. This approach offers improved flexibility in capturing changes in risk over time, allowing the model to better represent real-world patterns of disease progression.

6.5 Final CMD Policy Model : discussion

In summary, the Cardiometabolic Disease (CMD) Policy model presented in this thesis is defined as a statistical model based on multi-state survival analysis that is designed to project future health trajectories and support both epidemiological and economic evaluation analysis.

Using complex, linked UK patient-level data sets[210], the model captures the progression of key cardiometabolic events such as type 2 diabetes mellitus (T2DM), myocardial infarction (MI), and stroke. These events are modelled as distinct health states within a multi-state structure, allowing the flexibility for transition probability estimation, state occupancy over time, as well as long-term impact of disease progression on both clinical and economic outcomes.

For long-term extrapolation, in a semi-Markov framework, each transition is modelled using standard or flexible parametric survival models, enabling the calculation of cumulative hazard over time, which is eventually converted to transition probabilities. This allows simulation of individual-level pathways through disease progression.

Although the CMD Policy Model follows a semi-Markov multi-state structure, it can be best described as a "hybrid model" that blends features of both microsimulation and cohort modelling. Like a traditional cohort model, it estimates state occupancy probabilities over time for a representative population, allowing aggregation of outcomes such as life expectancy or disease prevalence at the population level. These probabilities are calculated based on transition-specific hazard models, using parametric or flexible parametric survival functions applied to a defined cohort. However, the model also embodies 'microsimulation-like' characteristics. It supports individual-level risk stratification by allowing covariates (e.g., age, sex, clinical biomarkers) to influence transition hazards. This means that transitions are not governed by average population risks alone but can vary across subgroups or individuals, depending on their risk profiles. Additionally, because hazards are converted into transition probabilities at each time step, the model can simulate pathways that resemble those seen in discrete-time microsimulation, where individuals probabilistically move between states over time.

Hence, while the output is aggregated (as in cohort models), the structure and flexibility of the model, particularly in accommodating individual-level heterogeneity and time-updated covariates, make it more aligned with hybrid policy modelling approaches. This design balances computational efficiency with granular clinical plausibility, enabling its application in both population health forecasting, cost-effectiveness analysis, and inequality impact where subgroup-specific insights are essential.

Strengths and limitations of CMD Policy Model

One of key strengths of this CMD Policy Model is its use of large-scale UK patient-level data (CPRD Aurum), linking primary and secondary care with death records and deprivation indices. This enhances population representativeness and policy

relevance. The model also includes a wide range of time-varying covariates, including clinical biomarkers, behavioural risk factors (e.g., smoking, alcohol), and socioeconomic indicators, supporting analysis of health inequalities and subgroup-specific outcomes.

While CPRD Aurum has underpinned numerous epidemiological studies, its use in constructing multi-state frameworks to analyse transitions between CMD states remains underexplored.[354] This is the first CMD policy model integrating CPRD Aurum to explicitly capture these transitions, enabling UK population-level preventive evaluations and policy adoption.

Methodologically, applying a semi-Markov structure offers a clinically realistic representation of chronic disease progression. This is particularly relevant for conditions such as type 2 diabetes mellitus (T2DM), myocardial infarction (MI), and stroke, where risks evolve with disease duration. Additionally, the model adopts a hybrid approach, combining elements of microsimulation and cohort modelling.

Each transition in the model is estimated using either standard parametric distributions or flexible parametric survival models, which provides the flexibility to adapt hazard functions based on empirical data and clinical relevance. This improves both the precision and credibility of long-term risk estimates, particularly when modelling interventions or projecting disease burden over time.

Despite its strengths, the CMD Policy Model has some limitations. First, it remains dependent on the quality and completeness of the input data. Any bias or missingness in the underlying patient-level data may affect the validity and generalisability of model outputs, particularly for underrepresented subgroups. Additionally, the model does not explicitly account for intermediate complications of T2DM, such as microvascular and macrovascular events that may influence transition hazards, unless manually incorporated into the model structure.

The conceptual model underlying the CMD framework was informed by clinical literature and expert input, providing face validity. Internal diagnostic procedures were conducted, such as AIC/BIC comparisons, residual checks, and visual inspection of survival curves against Kaplan-Meier estimates which demonstrated acceptable internal model fit. However, external validation using independent datasets (e.g., CPRD GOLD) has not yet been performed.[69,153] Moreover, the current implementation of the model requires manual manipulation of code and parameters within R, which may present a barrier to uptake. A web-based application or graphical user interface could improve accessibility for broader use by researchers, clinicians, and policymakers.

When compared to other established policy models, the CMD Policy Model offers several unique advantages. For example, the UKPDS Outcomes Model is widely used in health economic evaluations of diabetes.[355] It shares a multi-state framework with the CMD model but focuses more narrowly on diabetes-related complications using trial-derived risk equations. In contrast, the CMD model is based on routine population-level data and accommodates a wider range of disease trajectories, making it better suited for early prevention and population-level policy analysis.

Similarly, the Cardiovascular Disease Policy Model (CVDPM)[356] developed in the United States uses a Markov-based structure with detailed treatment modules and heterogeneous populations. However, it assumes memoryless transitions and often requires calibration to external data. The CMD model improves upon this by employing a sojourn-time-based semi-Markov structure, enabling more dynamic and time-sensitive modelling of disease progression without reliance on empirical calibration.

Another model, The IMPACT CHD Policy Model has been valuable in quantifying the contributions of risk factor changes and treatment uptake to coronary heart disease mortality trends at the population level.[357] However, it does not include morbidity outcomes (e.g., diabetes, MI, stroke), which limits its ability to capture the full burden of disease or evaluate quality-of-life outcomes which are addressed in current CMD Policy Model. The model is well-suited for retrospective analysis of observed mortality trends but less equipped

for prospective simulation of new or complex policies, especially those affecting disease incidence, or long-term resource use.

Unlike more complex platforms such as CVD-PREDICT[126] and SPHR model[125], which employ full microsimulation modelling, the CMD Policy Model offers a computationally efficient hybrid design. It supports individual-level heterogeneity through covariates while retaining tractability for large-scale scenario analysis. This makes it more accessible for routine use in academic and policy contexts without requiring high-performance simulation environments.

In summary, the CMD Policy Model provides a transparent, flexible, and methodologically robust framework for simulating cardiometabolic disease progression and estimating population-level health and economic outcomes. While its current scope and validation status present areas for further development, the model stands out for its real-world data foundations, semi-Markov logic, and hybrid architecture, positioning it as a valuable tool for informing prevention strategies and healthcare resource planning.

6.6 Conclusions

The development of the CMD Policy Model represents a step toward more integrated and policy-relevant simulation tools for cardiometabolic disease prevention. Rather than relying on simplified assumptions or narrowly defined clinical cohorts, the model provides a structured yet flexible framework that balances methodological rigour with real-world applicability. By combining advanced statistical modelling with UK patient-level data, it enables meaningful forecasting and supports decisions that reflect the complexity of cardiometabolic care.

As policy demands grow more complex, the model provides a strong foundation for future developments, including external validation, expanded disease coverage, and user-friendly interfaces. Designed for practical use, it enables the simulation of targeted interventions, such as dietary changes, metabolic control, or smoking cessation, by adjusting covariates and estimating their impact on transition risks and outcomes, thereby supporting effective public health planning and resource allocation.

Chapter 7 Case studies

7.1 Introduction

Following the development and validation of the CMD Policy Model in Chapter 6, Chapter 7 demonstrates its practical application through a case study focused on targeted preventative strategies. Leveraging detailed individual-level healthcare data and advanced multi-state survival models, the model enables simulation of various clinical and policy scenarios, offering a platform for forecasting cardiometabolic disease (CMD) trajectories within the UK population.

The intervention scenarios examined in this chapter are inspired by priorities identified in guidance from the National Institute for Health and Care Excellence (NICE), particularly in areas related to cardiovascular and diabetes prevention. The current scenarios do not replicate existing NICE evaluations, but adopt similar intervention types and are implemented using the CMD Policy Model, with updated or adapted parameters as appropriate. The primary objective of this chapter is to demonstrate ‘how the model works’ and show its flexibility, rather than to produce results ready for direct policy use.

The chapter includes two case studies, each demonstrating a distinct application of the model. The first focuses on dietary interventions aimed at preventing type 2 diabetes (T2DM) among ethnic minority populations in England. The second examines the cost-effectiveness of smoking cessation among adult smokers. Together, these examples illustrate two core ways in which the model can be utilised: first, by simulating the impact of modifying covariate profiles before and after an intervention; and second, by projecting long-term clinical outcomes and conducting health economic evaluations.

7.2 Case study 1: behavioural weight gain prevention

The intervention in this study is informed by the NICE public health guideline (PH38)[358], which focuses on the cost-effectiveness of preventing pre-diabetes among adults in high-risk groups. Originally published in 2012, this guideline specifically assessed the impact of dietary interventions among Black and minority ethnic populations in England with low socio-economic status (SES). While this thesis does not adopt all aspects of the original NICE modelling approach, it builds on its core concepts. The NICE model evaluated a range of weight management and dietary programme scenarios (retrieved from international published studies) that influence key metabolic risk factors such as body mass index (BMI), HDL, LDL, and total cholesterol.

7.2.1 Overview of public health intervention

In the UK, individuals of Black and other minority group (e.g.,South Asian) ethnicities and those living in areas of high deprivation are recognised as high-risk groups for cardiometabolic diseases (CMD), including type 2 diabetes (T2DM), hypertension, and cardiovascular disease (CVD). National health data consistently show that Black communities, particularly of African and Caribbean descent, experience higher rates of obesity, diabetes, and stroke compared to the general population.[359–361] These disparities are further exacerbated by socio-economic inequalities, with people in the most deprived areas facing greater exposure to risk factors such as poor diet, limited access to green spaces for physical activity, and higher levels of stress and material hardship.[361–363]

Structural barriers within the healthcare system, including reduced access to preventive services and culturally appropriate care, also contribute to delayed diagnosis and poorer disease outcomes. As a result, public health preventative strategies in the UK (such as NICE guidelines above) have increasingly targeted these high-risk groups to reduce health inequalities, improve early detection, and deliver more impactful and equitable health interventions.

The CMD Policy Model allows for such targeted analysis and, in this instance, is applied to individuals of Black ethnicity living in the most deprived areas. This approach highlights the model's capacity to incorporate subgroup-specific

characteristics and evaluate the potential impact of interventions on key metabolic risk factors. In the original NICE guidance, the scenarios for metabolic risk changes were used to estimate the cost per person per intervention. However, the focus of this case study is not to perform a full economic evaluation but rather to examine the effect of changes in covariates resulting from lifestyle interventions.

7.2.2 Modelling method: risk factor modification

This case study is informed by published clinical trial data focused on weight management interventions among the Black adult population in the UK. The intervention was a 12-month behavioural programme designed to support sustainable weight loss and metabolic improvement. It included behaviour change goals, weekly self-monitoring activities, monthly counselling sessions, training materials, and access to gym facilities. The baseline population consists of adults aged 18 years and older who were classified as obese and presented with moderate to high levels of clinical biomarkers associated with cardiometabolic risk. These biomarkers include elevated BMI, fasting glucose, blood pressure, and lipid levels.[364]

Following the intervention, participants experienced measurable improvements in metabolic parameters. These changes are assumed to result from the combined effect of structured behavioural support and physical activity, as evidenced in the trial.[365] The updated covariate values used in the intervention scenario reflect the post-intervention biomarker profile and are summarised in the table below.

Table 7.1 Metabolic risk changes based on trials results

Metabolic risks	Baseline (Initial)	After Intervention	Target/Ideal
BMI (kg/m2)	32	24	18.5–24.9
Glucose (mmol/l)	7	4.8	4.5–5.0 (fasting)
SBP (mmHg)	125	115	<120
DBP (mmHg)	85	75	<80
Cholesterol (mmol/l)	7	4.8	<5.0
HDL (mmol/l)	1.2	1.6	>1.6
LDL (mmol/l)	3.5	2.5	<3.0
Triglycerides (mmol/l)	2.0	1.4	<1.7

Appendix 8 provides example R code that shows how the model works to accommodate the mean risk changes before simulation in R. The simulated individual used in this case study is defined through a set of baseline characteristics (`initialCovariateValues`) representing a high-risk profile within a Black adult population living in the most deprived areas of the UK. The individual is classified as obese with a BMI of 32, and exhibits elevated cardiometabolic biomarkers and has borderline blood pressure and sub-optimal lipid and glucose profile. Lifestyle factors include being a non-smoker and consuming alcohol within safe limits.

The intervention scenario (`interventionCovariateValues`) assumes metabolic improvements following a structured 12-month behavioural weight management programme. This intervention is expected to result in reductions in metabolic risk levels.

For the first four transitions (from disease free to T2DM and CVD) the covariates are drawn from the initialCovariateValues object. For all subsequent transitions (from diabetes or after a cardiovascular event), the covariates are based on `interventionCovariateValues`, which reflect improvements due to a lifestyle intervention. These changes simulate the metabolic benefits of a structured behavioural weight management programme and allow the model to evaluate how improved risk profiles affect progression to subsequent events and death.This approach allows the CMD Policy Model to assess how an intervention influences disease progression across different stages of cardiometabolic disease by updating the relevant covariate values at appropriate points in the disease pathway.

Then the model can be fitted using either parametric or flexible parametric models for each transition. The best models (see Chapter 6) are embedded (see Chapter 6), ordered by transition name. Once the transition-specific models are fitted, the semi-Markov simulation can be executed. This involves computing state occupancy probabilities over time, which represent the likelihood that a simulated individual will occupy each health state (e.g., disease-free, diabetes, MI, stroke, death) at each time point during the simulation horizon.

7.2.3 Simulation results

The two plots (Figure 7.1 and 7.2) illustrate the projected state occupancy probabilities over a life-time horizon for the individual profile in the CMD Policy Model. The first plot represents the baseline (pre-intervention) scenario, while the second plot shows outcomes under the intervention scenario, in which improved metabolic risk factors (e.g., reduced BMI, glucose, and blood pressure) are assumed after implementation of a 12-month behavioural weight management programme.

The comparison between the pre-and post-intervention plots reveals several important shifts in disease progression. In the pre-intervention scenario, the "Disease-free" state begins to decline earlier (around age 20 to 25) indicating earlier onset of type 2 diabetes (T2DM). In contrast, the post-intervention plot shows a delayed decline in this state, with T2DM onset postponed by approximately 5 to 10 years, reflecting the preventive effect of improved metabolic risk factors.

The progression to myocardial infarction (MI) and stroke is also slower and less pronounced after the intervention, consistent with better lipid and blood pressure control. Additionally, the burden of chronic post-event states is reduced: in the pre-intervention plot, post-MI and post-stroke segments appear earlier and are more prominent, while in the post-intervention scenario, these segments are narrower, suggesting fewer severe or repeated events. Finally, the transition to death is delayed in the post-intervention group, with a higher likelihood of survival beyond age 80, indicating potential gains in life expectancy as a result of the intervention.

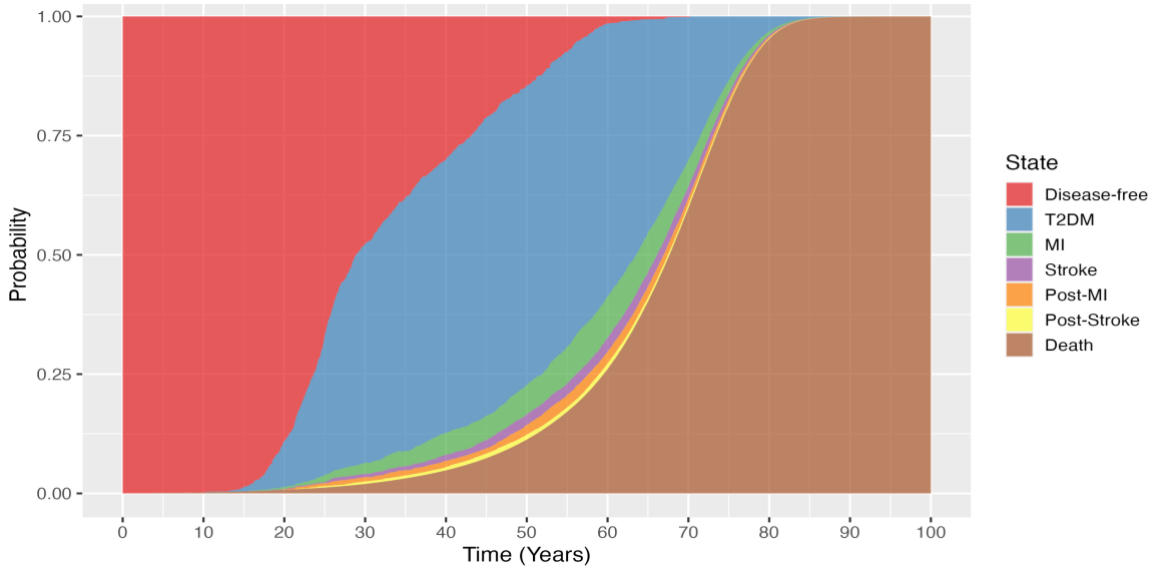


Figure 7.1 State occupancy probabilities at baseline

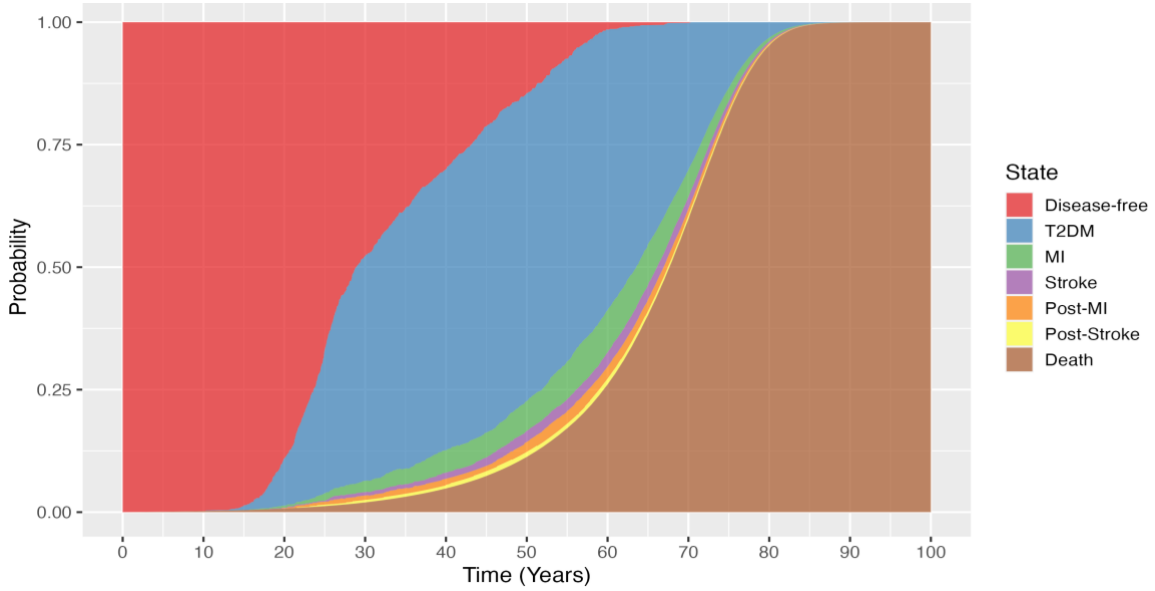


Figure 7.2 State occupancy probabilities after intervention

Although the behavioural programme improves metabolic risk factors, it does not completely eliminate the risk of developing cardiometabolic conditions. The individual in the model still begins with a high-risk profile, and while those risks are reduced, they are not completely brought down to the level of a low-risk individual.

As a result, the person is still likely to develop disease eventually, just at a later point in life. For instance, someone who would have developed diabetes at age 30 might now develop it at age 40 or 45 instead. Similarly, cardiovascular events may occur later but are not completely avoided. In the model, this leads to similar overall state trajectories, especially in the long term, since all individuals eventually transition to the "death" state.

It also can be argued with optimistic scenarios that some individuals may adopt lasting lifestyle changes following an intervention, such as sustained quit smoking or improved diet, which could reduce their lifetime risk.

This is why the shape of the state occupancy plot looks similar in both scenarios, even though the intervention is beneficial. The curves shift to the right (indicating delayed disease), but the overall structure remains intact. Hence, the intervention's impact lies in delaying disease and extending healthy life, not in preventing disease altogether.

The state occupancy probabilities data and transition probabilities generated by this model are presented in Appendix 9. From this, all modification by adding more parameters such as costs or utility can be performed, both in individual perspective or population level analysis, or simply comparing life expectancy.

7.2.4 Discussion and conclusions

This first case study demonstrates the application of the CMD Policy Model to simulate the health impact of a behavioural weight management intervention targeted at a high-risk subgroup: Black adults living in the most deprived areas of the UK. By modifying covariate inputs based on published clinical trial data, the model enables exploration of how realistic improvements in metabolic risk factors that can influence the progression of cardiometabolic diseases (CMD) over time.

Simulation results indicated that improved risk profiles can meaningfully alter transition probabilities, potentially delaying the onset of diabetes and reducing the risk of cardiovascular events or premature death. These findings support the policy relevance of lifestyle interventions and demonstrate how individual-level

risk factor changes can translate into long-term health benefits. It supports subgroup-specific simulations, making it well-suited for analysing ethnic and socio-economic disparities in health outcomes. A major advantage is its ability to dynamically update covariates over time, enabling assessment of interventions that modify risk factors. Its flexible framework accommodates both standard parametric and Royston-Parmar spline models, improving fit across diverse transitions. It has potential for inequality inspection and equity analysis, although these features were not activated in this first case study. In terms of usability, the model provides a practical and adaptable platform, allowing researchers to easily modify covariate values and simulate alternative intervention scenarios with minimal effort.

However, several limitations must be acknowledged. In this case study, the intervention effects were applied uniformly to all post-diagnosis transitions, without accounting for individual variation in adherence or behavioural relapse, factors that are common in real-world settings. Additionally, the model currently simulates a single average individual and does not capture heterogeneity across a population. While it is capable of accommodating subgroup differences based on deprivation index or ethnicity, this was not fully explored in the current scenario. Finally, although the intervention is informed by clinical trial data, the model has not yet been externally validated using long-term real-world datasets. These limitations point to opportunities for further refinement and future development.

In summary, this case study demonstrates how the model can be adapted with simple modifications to support epidemiological investigations. With the addition of cost and utility values assigned to each health state, the model could also support health economic evaluations. The next section will explore preventative strategies in more detail, using transition probabilities as the foundation for illustrating movement between health states, to offer more perspective regarding model usability. These transitions can be paired with effectiveness, costs, and utility estimates to quantify intervention impact at each stage of disease progression.

7.3 Case study 2: cost-effectiveness of smoking cessation intervention

The first case study demonstrated how the CMD Policy Model can be used to simulate changes in disease progression based on behavioural risk factor modification, this approach can also be extended to incorporate economic dimensions. By linking clinical events with associated costs and health-related quality of life, the model enables the assessment of both cost-effectiveness and budget impact. This integration allows for a more comprehensive policy analysis, supporting decision-making not only on the basis of health outcomes but also on the value for money of interventions. The following sections explore how this modelling framework can be adapted for economic evaluation, using a combination of clinical evidence, cost data, and utility weights to inform resource allocation in the context of cardiometabolic disease prevention.

7.3.1 Health economic evaluation: preliminary concept

Definition

A widely accepted definition of economic evaluation is provided by Drummond et al. (2015)[73], who describes it as "the comparative analysis of alternative courses of action in terms of both their costs and consequences." This definition captures the essence of economic evaluation: it is not merely about calculating costs or measuring outcomes in isolation, but about assessing the relative value of competing interventions in a structured and transparent manner.

Economic evaluation plays a critical role in informing public health policy by assessing the value for money of health interventions. It is a core component of evidence-based healthcare decision-making, aiming to ensure that limited resources are allocated efficiently to achieve the greatest possible health benefits. In public health and chronic disease prevention, economic evaluation is especially important because interventions often require substantial upfront investment but deliver benefits over an extended time horizon. Without such evaluations, there is a risk of under- or over-investing in interventions that may be ineffective, inefficient, or inequitable.[366,367]

Types of economic evaluation

There are several types of economic evaluation, each serving different decision-making needs. Cost-minimisation analysis (CMA) is used when two interventions are proven to have equivalent outcomes; in such cases, the focus is solely on identifying the option with the lowest cost. Cost-effectiveness analysis (CEA) compares interventions based on the cost per unit of health outcome (e.g., cost per case of diabetes prevented), while cost-utility analysis (CUA) extends this by incorporating a generic outcome such as quality-adjusted life years (QALYs), capturing both the quantity and quality of life gained. Cost-benefit analysis (CBA), in contrast, expresses both costs and outcomes in monetary terms, enabling comparisons across sectors or policy areas.[73]

Table 7.2 Types of economic evaluation

Method	Costs	Effects
Cost-minimisation analysis	Monetary unit (£)	Considered equal
Cost-effectiveness analysis (CEA)	Monetary unit (£)	Natural unit (LY, disease events prevented)
Cost-utility analysis (CUA)	Monetary unit (£)	QALYs, DALYs
Cost-benefit analysis (CBA)	Monetary unit (£)	Monetary unit

Importance of modelling in economic evaluation

The most common applied modelling frameworks in economic evaluation are decision-trees and Markov models. Decision tree models are well-suited for short-term economic evaluations involving interventions with a limited number of outcomes and a clear sequence of events. They are relatively simple to construct and interpret but become less practical when events recur or evolve over time. The Markov framework (in state transition models), on the other hand, is designed for chronic conditions and long-term interventions, allowing individuals to transition between health states across multiple time cycles. This structure makes Markov models particularly advantageous for capturing disease progression, recurrent events, and cumulative outcomes like lifetime costs and QALYs. As a result, Markov modelling offers greater flexibility and realism in evaluating complex, long-term public health strategies.[143,368]

In the UK, health economic modelling plays a central role in evidence-based decision-making, particularly through institutions such as the National Institute for Health and Care Excellence (NICE). NICE routinely employs economic models to assess the clinical and cost-effectiveness of healthcare technologies, public health interventions, and disease preventative strategies.[369–371] These models are essential for determining whether new interventions represent good value for money within the constraints of the NHS budget.

Modelling allows NICE to extrapolate trial data over long-term horizons, assess uncertainty, and compare interventions across different diseases using a common metric such as cost per QALY gained. This is especially important in chronic conditions like cardiometabolic disease, where preventive strategies may offer small but cumulative health benefits over time. By simulating alternative scenarios, economic models help policymakers understand trade-offs between immediate costs and long-term health gains, examine the distribution of benefits across population subgroups (e.g., by ethnicity or deprivation), and support equitable resource allocation. In this way, health economic modelling not only informs funding decisions but also enhances the fairness, transparency, and accountability of health policy.

In this thesis, the CMD Policy Model provides the structural foundation to integrate various modelling elements and support future cost-effectiveness evaluations of CMD preventative strategies. By incorporating clinical, epidemiological, and economic data, the model enables robust, long-term projections of value for money, thereby informing policy and resource allocation decisions.

The first case study demonstrates how changes in metabolic risk factors influence state occupancy probabilities and transition probabilities, offering an epidemiological perspective on disease progression. Building on that, this second case study illustrates how transition probabilities can be adjusted using information from published evidence, such as intervention effectiveness, (e.g: risk ratios), inputs commonly employed in long-term economic evaluations. While the two case studies differ in their perspectives, one grounded in epidemiology and the other in literature-based effectiveness evidence, both approaches are valid and compatible within an economic evaluation framework.

7.3.2 Methods

Intervention, comparator, and target population

The intervention examined in this second case study is based on the Single Technology Appraisal (STA) published by Hind et al. (2007)[358] which evaluated the clinical and cost-effectiveness of varenicline (Champix®) for smoking cessation. The appraisal explicitly recommended varenicline for adult smokers who wish to quit, noting it should ideally be prescribed alongside behavioural support, though it may still be offered when such support is declined.

The STA provided the foundation for NICE guidance, supporting the routine commissioning of varenicline in the NHS in England and Wales. As of June 2025, updated resources from the National Centre for Smoking Cessation and Training (NCSCT) continue to support varenicline as a safe, effective, and cost-effective first-line treatment for smoking cessation, reaffirming its NICE-endorsed role in clinical practice.[372,373]

Building on this context, the current thesis conducts a simple economic evaluation (cost-utility analysis) comparing varenicline to bupropion for adult smoking cessation. Bupropion was selected as the comparator in this evaluation because it has long been an established pharmacological aid for smoking cessation and was one of the first non-nicotine medications approved before the introduction of varenicline.[374] However, bupropion carries a higher risk of adverse effects, most notably insomnia, dry mouth, and a dose-dependent risk of seizures, which restricts its use in patients with seizure disorders, eating disorders, or those on interacting medications.[375]

Multiple clinical trials and meta-analyses, including those cited in the NICE technology appraisal (TA123), have demonstrated that compared to buproprion varenicline significantly improves quit rates relative to bupropion. The design of the analysis follows the PICOS framework, as outlined below:

Table 7.3 PICOS for economic evaluation

PICOS	Description
Population	Adult smokers (≥40 years)
Intervention	Varenicline Patients begin with a titration week, then proceed to 1 mg orally twice daily: - Week 1: Typically 0.5 mg once daily on days 1–3, then 0.5 mg twice daily on days 4–7. - Week 2–12: 1 mg twice daily (total 2 mg/day). Standard treatment 12 weeks, with a possibility of extending another 12 weeks for patients who successfully quit in the first course.
Comparator	Buproprion - 150 mg orally twice daily; a total daily dose of 300 mg. - Duration: Used within a 12-week treatment course, consistent with smoking-cessation protocols
Outcomes and study design	- Model based economic evaluation - Cost per QALY gained (economic outcome)

Model structure and assumptions

The model follows the current structure of the CMD Policy Model, comprising seven health states and thirteen transition probabilities as previously outlined. Individuals enter the model at age 40 years in a disease-free state who are motivated to quit smoking, with a moderate metabolic risk profile and a current smoking status. The model simulates annual transitions over a lifetime horizon. Although the original appraisal included participants aged ≥18 years, the model assumes a baseline age of 40 to better represent the demographic more likely to engage seriously with cessation support and to maintain long-term abstinence.

In contrast to earlier models such as BENESCO (BENefits of Smoking Cessation), developed for a Pfizer-commissioned technology appraisal, where smoking status (current, former, relapse) and related comorbidities (e.g., cardiovascular disease, chronic obstructive pulmonary disease) are explicitly modelled, the current model incorporates smoking status via relative risk adjustments.

The effectiveness of smoking cessation interventions is captured through the proportion of individuals who achieve smoking abstinence, and the associated risk reductions in disease onset.

The seven health states included in the model are: Disease-free, Type 2 Diabetes Mellitus (T2DM), Myocardial Infarction (MI), Post-MI, Stroke, Post-stroke, and Death. Transition probabilities were adjusted based on treatment-specific quit rates: 21.9% for varenicline and 16.1% for bupropion, based on estimates from NICE STA 123[358] Relative risk reductions associated with smoking cessation: 20% for T2DM, 30% for MI, 25% for stroke, and 15% for all-cause mortality were applied to the relevant transition probabilities.[358,376]

Cost, perspective, and dicounting

The analysis adopts the perspective of the UK National Health Service (NHS) and Personal Social Services (PSS), consistent with the NICE reference case for economic evaluations. This includes direct medical costs such as pharmacological treatment and hospitalisation, as well as long-term care costs related to complications such as post-stroke disability. Indirect costs (e.g., productivity losses) are not currently considered, but technically it is possible for them to be incorporated if a societal perspective is adopted.

It is important to emphasise that the data used in the model are primarily derived from previously published NICE appraisals. The purpose of this example is to illustrate the structure and functionality of the CMD policy model, rather than to re-evaluate the cost-effectiveness of varenicline in the current UK context.

Intervention costs were derived from the NICE STA for varenicline and were inflated to 2025 values using the UK Consumer Price Index (CPI). The cost of a 12-week course of varenicline was estimated at £204.40, while the corresponding cost for bupropion was £104.21.[358] These represent one-off treatment costs incurred at the outset of the model. Annual healthcare costs associated with each disease state were sourced from validated national datasets, including the NHS Reference Costs and the Personal Social Services Research Unit (PSSRU) Unit Costs of Health and Social Care.

Table 7.4 Input parameter

Parameters	Values	Distribution	Description	Sources
Transition probabilities	0-1	Beta	Lifetime transition probabilities	CMD Policy Model
Intervention costs				
Varenicline	£293.80	Gamma	As dosage mentioned in PICOS	BNF [NICE STA 123][358]
Buproprion	£144.63	Gamma	As dosage mentioned in PICOS	BNF [NICE STA 123][358]
Quit rate				
Varenicline	21.8%	Gamma	Quit rate at 12 months	NICE STA123; adjusted from trial data[358]
Buproprion	16.2%	Gamma	Quit rate at 12 months	NICE STA123; adjusted from trial data[358]
Healthcare costs (Annual)				
c_T2DM	£2,900	Gamma	Cost of managing T2DM	Hex et al. 2012[377]
c_MI	£4,250	Gamma	Cost of first MI	NHS Ref Costs BNF [NICE STA 123/NIHR reports][358,378-380]
c_post-MI	£810	Gamma	Cost of post-MI maintenance	NHS Ref Costs BNF [NICE STA 123/NIHR reports][358,378-380]
c_stroke	£4,840	Gamma	Cost of acute stroke	NHS Ref Costs BNF [NICE STA 123/NIHR reports][358,378-380]
c_post-stroke	£14,800	Gamma	Annual cost of post-stroke disability and care	NHS Ref Costs BNF [NICE STA 123/NIHR reports][358,378-380]
Relative risks (RR)				
rr_ T2DM	0.80	Log-normal	RR of T2DM in ex-smokers vs smokers	Pan et al., 2015[381]
rr_ MI	0.70	Log-normal	RR of MI in ex-smokers vs smokers	Critchley & Capewell, 2003[382]
rr_ stroke	0.75	Log-normal	RR of stroke in ex-smokers vs smokers	Peters et al., 2013[383], Shah and Cole (2010)[383]
Utility				
u_disease free	0.85	Beta	Utility of disease free state	Ara & Brazier (2010)[384], NIHR reports[379,385]
u_T2DM	0.78	Beta	Utility of T2DM state	Ara & Brazier (2010)[384], NIHR reports[379,385]
u_MI	0.76	Beta	Utility of MI state	Ara & Brazier (2010)[384], NIHR reports[379,385]
u_post-MI	0.65	Beta	Utility of post MI state	Ara & Brazier (2010)[384], NIHR reports[379,385]
u_stroke	0.73	Beta	Utility of stroke state	Ara & Brazier (2010)[384], NIHR reports[379,385]
u_post-stroke	0.48	Beta	Utility of post stroke	Ara & Brazier (2010)[384], NIHR reports[379,385]
Discount rate				
dr_costs	3.5%	-	Annual discount rate for costs	NICE Reference Case[369]
dr_outcomes	3.5%	-	Annual discount rate for outcomes	NICE Reference Case[369]

NICE STA: National Institute for Health and Care Excellence; *NIHR:* National Institute for Health and Care Research, BNF:British National Formulary

Health-related quality of life (HRQoL) was modelled using utility values from Ara and Brazier (2010)[384], a widely cited UK-based source for economic evaluations.It is then adjusted with NIHR reports for smoking cessation. The model also incorporated the long-term health benefits of smoking cessation by applying relative risks (RRs) to transitions into smoking-related diseases. These RRs reflect the reduced incidence of disease among ex-smokers compared to continuing smokers and were applied proportionally, based on the intervention-specific quit rates reported in NICE STA123.

No risk reductions were applied to transitions into Post-MI or Post-stroke states, as these represent downstream consequences of the primary event and are assumed to be unaffected directly by smoking cessation. This is consistent with the structure of prior economic evaluations, including NICE STA123 and international models such as BENESCO and EQUIPTMOD (European Tobacco ROI Model).

All future costs and health outcomes were then discounted at an annual rate of 3.5%, consistent with the NICE reference case for economic evaluations in the UK.[369] For probabilistic sensitivity analysis, model parameters were simulated using standard probability distributions commonly applied in health economic modelling.

Base-case analysis

In economic evaluation, Incremental Cost-Effectiveness Ratio (ICER) and Net Monetary Benefit (NMB) are two fundamental tools for assessing cost-effectiveness. The ICER is a key metric in health economic evaluation that compares the difference in costs and outcomes between two interventions. It is calculated as the difference in cost divided by the difference in effectiveness. It answers the question, "How much extra does it cost to gain one additional QALY with the new intervention?" The ICER helps decision-makers assess whether the additional benefits of a new intervention are worth the additional costs, relative to a comparator (often standard care).

$$ICER = \frac{Cost_{intervention}}{Effect_{intervention}} - \frac{Cost_{comparator}}{Effect_{comparator}}$$

(Equation 7.1)

Lower ICERs suggest more cost-effective options, and in many health systems, including the UK, ICER thresholds (e.g., £20,000–£30,000 per QALY) guide decisions on whether an intervention offers good value for money.[369] To aid interpretation, the Cost-Effectiveness (CE) Plane is used.

On the other hand, the NMB approach reformulates cost-effectiveness by translating health outcomes into monetary terms using a predefined WTP threshold.

$$NMB = (Effectiveness \times \lambda) - Cost$$

(Equation 7.2)

Where effectiveness is typically measured in QALYs, λ (lambda) is WTP threshold per QALY . Cost is the total cost of the intervention. By comparing two intervention, Incremental Net Monetary Benefit (INMB) can be calculated using formula above incorporationg difference in effectiveness and costs (ΔEffectiveness and ΔCosts).

If INMB > 0 it means the new intervention is cost-effective at the chosen WTP threshold, and if INMB < 0 means that the comparator is more cost-effective.

ICER expresses cost-effectiveness as a ratio, specifically, the additional cost per additional quality-adjusted life year (QALY) gained, and is widely used by HTA bodies such as NICE. Its main advantages lie in its intuitive interpretation and strong alignment with established WTP thresholds, making it accessible to policymakers and stakeholders. However, ICER has mathematical limitations, particularly when the incremental difference in QALYs is very small or negative, which can lead to unstable or misleading results. In contrast, the NMB framework reformulates the cost-effectiveness question into a monetary value, subtracting the cost of an intervention from the monetary value of the health benefits. This approach is statistically more robust, especially for handling uncertainty and conducting probabilistic sensitivity analyses. Moreover, NMB is always defined, even in cases where ICER fails. Nevertheless, its main drawbacks include the need to specify a willingness-to-pay threshold upfront (which may vary across contexts) and the fact that monetary valuation of health outcomes may be less intuitive for some audiences.[386,387] Despite the growing use of NMB in methodological research, ICER remains the most widely used measure in policy-making, and this case study applying ICER for final result.

In terms of practicality, the NMB calculation can still be implemented within the policy model by adding supplementary code to represent the mathematical expression.

Sensitivity analysis

Sensitivity analysis is an essential component of health economic evaluation, used to assess the robustness of model results to uncertainty in input parameters. Deterministic sensitivity analysis (DSA) involves varying one (one-way) or several (multi-way) parameters at a time within a plausible range to observe the impact on outcomes such as the ICER.[155] This helps identify key drivers of cost-effectiveness and supports transparent reporting of assumptions.

In contrast, probabilistic sensitivity analysis (PSA) simultaneously varies multiple uncertain parameters by assigning probability distributions to model inputs (e.g., costs, utilities, transition probabilities) and using Monte Carlo simulation to

generate a range of possible outcomes.[155] PSA provides a more comprehensive picture of uncertainty and allows the generation of outputs like the cost-effectiveness acceptability curve (CEAC), which shows the probability that an intervention is cost-effective at different willingness-to-pay thresholds. Both DSA and PSA are performed in this current model to highlighting how uncertainty may influence conclusions.

7.3.3 Results

Base case results

The updated model-based economic evaluation compared the lifetime costs and health outcomes of two pharmacotherapies for smoking cessation, varenicline and bupropion (based on a hypothetical UK cohort of 1,000). These results show that varenicline is more effective, offering 0.06 additional QALYs and 0.06 additional life-years per person. The resulting ICER of £1,656 per QALY gained is well within the UK's cost-effectiveness threshold of £20,000–£30,000/QALY, suggesting that varenicline is highly cost-effective from an NHS perspective.

Table 7.5 Base case result

	Lifetime costs (discounted)	LYG (discounted)	QALY (discounted)	ICER
Varenicline	33,205.16	19.60	15.63	£1,656
Buproprion	33,106.94	19.54	15.57	

It can be said that the results of this analysis are nearly identical to those reported by NICE STA 123.[358] They concluded that, on a per-person basis, varenicline was associated with lower average costs (£10,717 vs. £10,820 for bupropion) and higher average QALYs (13.27 vs. 13.25). To project these results to the population level, the per-person estimates can be multiplied by the number of eligible individuals. For example, in the NICE STA123 report, the estimated number of adult smokers in England eligible for cessation services was approximately 3,173,000 people, resulting in a total cost for varenicline of £34.02 billion and a total of 42.14 million QALYs.[358]

To extrapolate these findings to the population level, the per-person results can be multiplied by the number of eligible individuals. For instance, if we assume there are 4.09 million adult smokers aged 40 in England, this results in a total cost of approximately £135.79 billion (£33,205.16 × 4.09 million) and a total of 63.92 million QALYs (15.63 × 4.09 million) for Varenicline.

Sensitivity Analysis

Deterministic sensitivity analysis

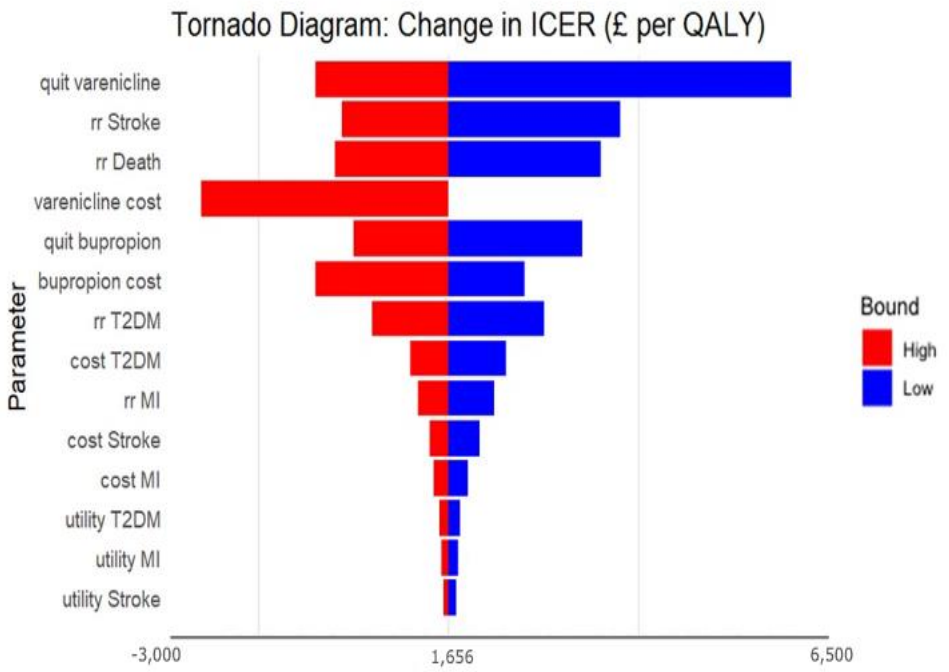


Figure 7.3 Tornado diagram

The tornado diagram presents the results of a one-way sensitivity analysis, highlighting how changes in individual model parameters affect the incremental cost-effectiveness ratio (ICER) of varenicline compared with bupropion. The most influential parameter was the quit rate for varenicline, as it directly determines the number of individuals who successfully stop smoking and therefore avoid future smoking-related diseases. A higher quit rate leads to greater health gains (increased QALYs and life-years), which reduces the ICER and makes the intervention appear more cost-effective. Conversely, a lower quit rate weakens these health benefits, driving the ICER upwards. Similarly, the relative risk reductions for death and stroke had a major impact, as these

are high-burden outcomes where small changes in risk can result in significant differences in life expectancy and healthcare costs over time. The quit rate for bupropion also influenced the ICER considerably, since greater effectiveness of the comparator diminishes the relative advantage of varenicline, making it appear less favourable economically.

Moderate effects were seen from parameters related to diabetes and MI, as well as health state utility values for stroke, MI, and diabetes. These influence the QALY outcomes but to a lesser extent because the baseline utilities and disease incidence are relatively stable or less impactful than stroke or death. In contrast, treatment costs, including the drug acquisition costs for varenicline and bupropion and the costs of managing cardiovascular and metabolic conditions, had minimal impact on the ICER. This is because small changes in unit costs are outweighed by the much larger health effects driving cost-effectiveness.

Probabilistic sensitivity analysis

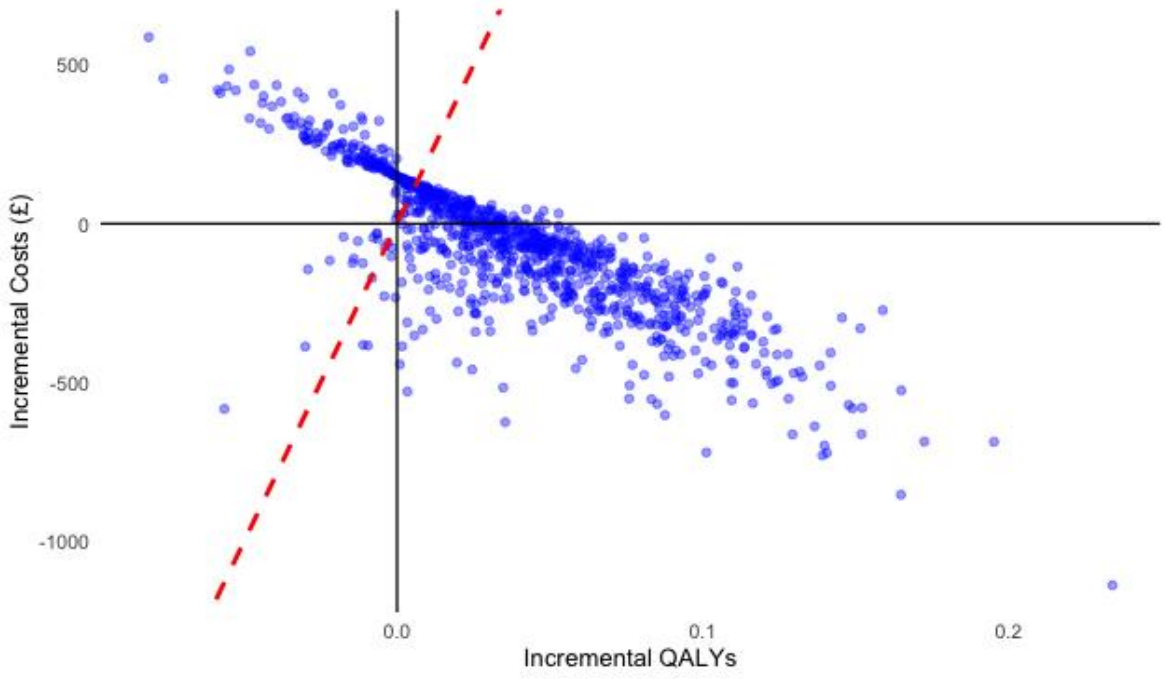


Figure 7.4 Cost-effectiveness plane

The cost-effectiveness plane illustrates the results of the PSA, where each of the 1,000 simulations represents a random draw from probability distributions for key model parameters. The x-axis shows the incremental QALYs gained by varenicline compared to bupropion, while the y-axis reflects the corresponding incremental costs. The majority of simulations fall within the southeast quadrant, indicating that varenicline is both more effective and less costly than bupropion—in other words, it is the dominant strategy in most scenarios. A small number of simulations lie in the northeast quadrant, where Varenicline is more effective but also more costly, yet still often falls below the WTP threshold represented by the red dashed line. Very few simulations lie in the northwest or southwest quadrants, which would imply that Varenicline is either dominated or offers fewer QALYs.

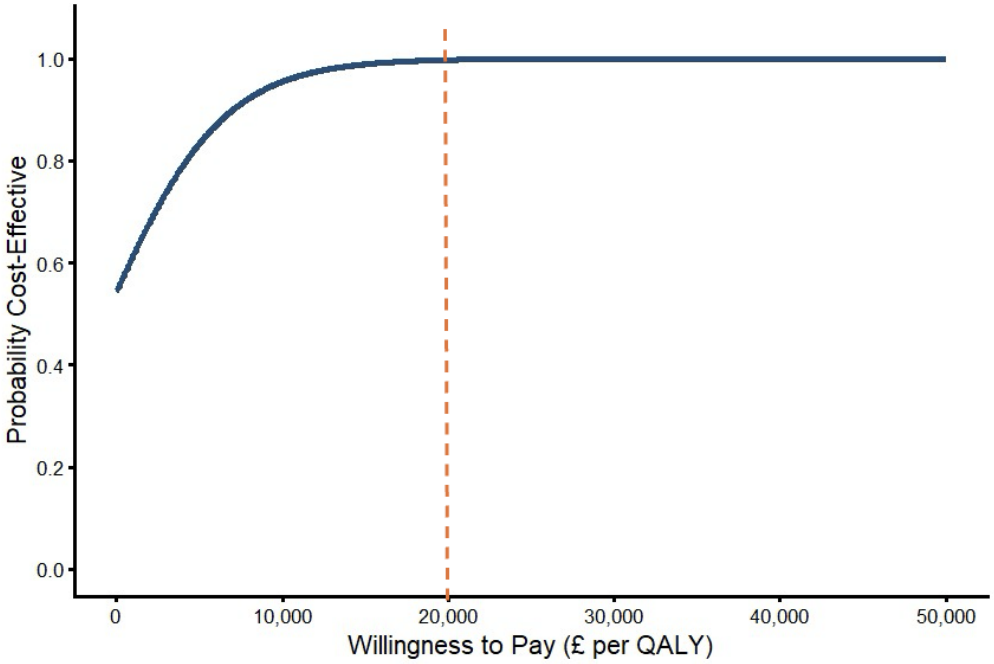


Figure 7.5 Cost-effectiveness acceptability curve (CEAC)

The cost-effectiveness acceptability curve (CEAC) illustrates the probability that Varenicline is cost-effective at varying levels of WTP thresholds/QALY gained. The curve rises steeply and shows that at a WTP of £20,000 to £30,000 per QALY (marked by the vertical red dashed line), the probability that Varenicline is cost-effective is already above 90%, exceeding the accepted decision threshold used by NICE.[369] This confirms the earlier findings from the cost-

effectiveness plane and base-case ICER, demonstrating that even when parameter uncertainty is fully accounted for through PSA, varenicline remains highly likely to be the most cost-effective option. The steep shape of the curve also suggests that the model's results are relatively insensitive to variations in the decision-maker's WTP, strengthening the policy argument for continued or expanded investment in smoking cessation interventions using varenicline within the NHS.

7.3.4 Discussion and conclusions

This analysis demonstrates not only that varenicline is a cost-effective intervention for smoking cessation, but also that the economic evaluation model performed consistently and reliably across a wide range of conditions. In the base-case scenario, varenicline generated greater health benefits than bupropion (15.63 QALYs vs. 15.57) at a slightly higher cost (£33,205 vs. £33,107), resulting in an ICER of £1,656 per QALY gained. This is well below NICE's standard willingness-to-pay threshold and confirms the intervention's strong value for money. When extrapolated to a national level, assuming 4.09 million adult smokers aged 40 years old in England, the model projected over £135.8 billion in total costs and 63.9 million QALYs. These outputs align with published findings, including those from NICE's STA123.[358] Importantly, this case study demonstrates how the model behaves in practice: it produces plausible, policy-relevant outputs while efficiently linking short-term intervention effects to long-term disease consequences and economic value.

Nonetheless, this study has several limitations that warrant consideration. Structurally, the use of a cohort-based Markov model imposes simplifying assumptions such as permanent cessation status after quitting. This excludes the possibility of relapse, which is a clinically important factor in long-term smoking cessation. Additionally, the model is heavily reliant on published literature for utility values, disease risks, and cost data. While these sources provide a solid foundation, they may not fully capture recent changes in practice or reflect the heterogeneity of real-world patients. As a result, the external generalisability of the model outputs may be limited, particularly in settings where patient characteristics, healthcare costs, or intervention delivery differ significantly from the assumptions used in this evaluation.

The NHS perspective used here also omits indirect costs and non-medical costs such as lost productivity and informal care. However, it is important to stress that the current model is fully capable of incorporating these additional cost components. Their exclusion in this analysis was a deliberate modelling choice, primarily due to the reliance on published studies that often lack robust or consistent estimates for such parameters. Including them without high-quality supporting data could introduce substantial uncertainty into the model. Future versions using real-world or patient-level datasets could confidently extend the model's perspective to capture a fuller range of economic impacts.

Recent literature, including the study by Zhou et al. (2024)[388], has highlighted the importance of estimating healthcare costs using individual-level data and regression models that account for cost variability and skewness. Applying these methods in future model updates would allow for more accurate cost estimations, particularly for conditions which significantly impact long-term healthcare expenditures. Integrating relapse dynamics and capturing individual differences in treatment adherence and disease risk would further improve the model's validity and policy relevance.

In conclusion, this analysis confirms that varenicline is a clinically effective and economically efficient option for smoking cessation. As the first case study applying the CMD policy model to a behavioural intervention, it provides a strong foundation for future modelling work. By incorporating individual-level data, relapse probabilities, and broader cost perspectives, future versions of the model can better inform tobacco control policies and support more promising decisions within the NHS and beyond.

To enhance transparency and reporting quality, the current model is accompanied by the Consolidated Health Economic Evaluation Reporting Standards (CHEERS) checklist, which is provided in Appendix 10.

Chapter 8 Main insights, policy implications, further research

8.1 Introduction

The general aim of this thesis was to develop a Cardiometabolic Disease (CMD) Policy Model by exploring the potential use of Clinical Practice Research Datalink (CPRD) Aurum and its relevant linked datasets within the UK setting. This model was designed to accommodate both epidemiological and health economic contexts in which extrapolation and long-term analyses are critical to inform decision-making in public health policy.

Throughout the preceding chapters, the model development followed a sequential and structured process and included: a systematic review to map existing models and gap (Chapter 3), the conceptual model construction (Chapter 4), data preparation and cohort definition (Chapter 5), development and analysis of multi-state survival models (Chapter 6), and case studies on preventative interventions to illustrate the application and potential of the CMD Policy model (Chapter 7).

As the final chapter, Chapter 8 is organised into three sections. First, main insights (section 8.2, covering sub-section 8.2.1-8.2.5), which provides a synthesis of the findings from Chapter 3 to 7, summarising the most important insights from the conceptual, data analysis, and modelling stages. Second, policy implications (section 8.3), providing the overall interpretation of the findings in the context of health policy, particularly for prevention and management of cardiometabolic diseases. Lastly, recommendations fo further research (section 8.5), which outlines key areas where additional investigation, data improvement, or methodological refinement is needed to enhance the accuracy, relevance, and practical utility of the CMD Policy Model.

8.2 Main insights

8.2.1 Critical literature review and conceptual model

The systematic review in Chapter 3 offers a comprehensive evaluation of existing policy models aimed at preventing cardiometabolic disease (CMDs) through dietary interventions. By analysing 32 studies published between 2000-2024, the review provides critical insight into the methodologies, applications, and quality of these models.[185]

Unlike previous reviews that extensively concentrate on clinical interventions or high-risk population, such as individuals already diagnosed with CMDs, this study highlights models assessing population-wide dietary policies. These include initiatives likes sugar taxes, salt reduction programme, and food labelling strategies for general or low-risk population. Such interventions aim to achieve primordial prevention by addressing risk factors before the onset of disease, aligning with the overarching goal of the CMD Policy model developed in this thesis.

Most of the reviewed models demonstrated moderate to good reporting quality. However, a key limitation identified was the frequent reliance on non-local data, often without adequate justification regarding its transferability. Additionally, the review emphasised the need for greater transparency in the validation of input data and assumptions.[185]

The results of this systematic review not only provide a critical appraisal of existing models and highlight the gaps in modelling practices but also serve as a foundational resource for developing conceptual models[156] such as presented in Chapter 4. By integrating insights from reviews along with clinical guidelines and experts' input, supporting both face and structural validity, the conceptual model then establishes a basis for overall model development.

Based on the findings of this review, several key recommendations emerged. These include leveraging real-world data (RWD) to enhance population representativeness, improving the transparency and quality of data inputs,

adopting a broader economic perspective, and incorporating equity considerations into policy modelling.[185]

The CMD Policy Model addresses many of these identified gaps through a series of methodological and structural advancements. First, it is explicitly grounded in UK-specific RWD, enhancing the contextual relevance and policy applicability of its outputs. Second, it incorporates time-dependent covariates to more accurately reflect dynamic risk profiles and capture the long-term impacts of behavioural interventions. Third, the model integrates both health and economic outcomes within a unified state-transition framework, enabling robust cost-effectiveness analysis over a lifetime horizon. These design features position the CMD Policy Model as a significant advancement over many of the models identified in the review. It aligns closely with emerging best practices by improving data quality, enhancing model transparency, adopting a prevention-focused perspective, and ability to applying economic evaluation analysis also with expansion of inequality outcomes generation.

8.2.2 Caveats in data preparation

Chapter 5 outlines the critical data preparation phase required to construct the CMD Policy Model using real-world data from CPRD Aurum. This stage details the steps involved in transforming raw patient-level electronic health records (EHR) into an analysis-ready dataset. While many existing studies focus on general data cleaning procedures or the construction of a baseline cohort, this chapter provides specific guidance on how the dataset was structured to fit the state-transition modelling framework used in this thesis. It provides the technical insights that are often underreported in the modelling literature, not because data preparation is unimportant, but because most attention tends to be placed on how the data are analysed and the findings that emerge from those analyses.

Three main components are addressed in data preparation section:

1) Initial data manipulation. Before inclusion criteria was applied, the raw dataset included over 14 million patients, which introduced significant computational challenges. To address this, memory-efficient strategies were employed in R environment, enabling the processing and

transformation of the data within the limitations of available computing resources.

2) Cohort construction and transition matrix development. The development of the transition matrix posed several challenges, which were addressed using a combination of custom-built R functions and existing published methods.
3) Time-dependent covariate handling. All relevant covariates were treated as time-dependent variables, allowing the model to more accurately reflect changes in patient health status over time.

8.2.3 CMD Policy Model: standard vs flexible parametric model

The CMD Policy Model developed in this thesis relies on real-world data (RWD) rather that extrapolated results from clinical trial populations, which is more representative and generalisable for modelling future disease progression. It allows also the flexibility to incorporate public health strategies and simulate them for long-term time horizons.

To model disease progression across health states, a wide range of survival analysis methods were applied. These ranged from simple non-parametric methods (e.g., Kaplan-Meier), to semi-parametric approach (e.g., Cox proportional hazards model), to standard parametric models (e.g., exponential, Weibull), and finally to flexible parametric models (e.g., Royston-Parmar splines). This methodological progression allowed for in-depth exploration of model performance across transitions.

For extrapolation beyond the observed follow-up period, the findings in this thesis demonstrated that both standard and flexible parametric models are capable of generating long-term projections. While this thesis presents results using both approaches, it does not aim to compare them head-to-head or declare one as universally superior. Instead, the modelling framework was designed to remain flexible, allowing the selection of the most appropriate method based on the specific characteristics of each transition.

The choice between standard and flexible parametric models is not based solely on predictive accuracy. Each approach has distinct strengths and limitations, and their suitability depends on factors such as the shape of the underlying hazard, the amount and quality of available data, the complexity of disease progression, and the intended use of the model (e.g., for scenario simulation or policy evaluation). In this way, the CMD Policy Model accommodates both modelling approaches, using them pragmatically and adaptively to ensure robust and policy-relevant results.

Instead, each modelling approach has its own advantages and limitations. Standard parametric models offer simplicity, computational efficiency, and ease of interpretation. However, they may lack flexibility when hazard functions are complex, time-varying, or non-monotonic.[304,313,389,390] Each standard distribution imposes a specific functional form on the hazard, for example, the exponential model assumes a constant hazard over time, while the Weibull model assumes a hazard that is either monotonically increasing or decreasing. These fixed assumptions may not accurately reflect real-world disease dynamics, potentially leading to biased estimates, particularly when extrapolating beyond observed data.

In contrast, flexible parametric models, such as Royston-Parmar spline-based approaches, are more adaptable and capable of capturing complex hazard shapes. They provide a better fit to observed data by modelling the log cumulative hazard (or log hazard) using restricted cubic splines. However, this flexibility comes at a cost: flexible models require more careful calibration and carry a higher risk of overfitting, especially when the number of spline knots is excessive or the sample size is limited. Moreover, while flexible models often perform well within the observed follow-up period, they may behave unpredictably during extrapolation, as the spline-based hazard function is not inherently constrained outside the data range.[391–393] Without appropriate validation or sensitivity analyses, this can lead to unstable or implausible long-term projections.

Furthermore, by embedding each transition-specific survival model (standard or flexible) into a semi-Markov structure, the CMD Policy Model can simulate

patient trajectories over time with greater clinical plausibility and temporal precision.

Both standard and flexible models were applied and evaluated using statistical criteria (e.g., AIC, BIC) and diagnostic plots checks (e.g., Cox-Snell residuals). The final model selection was conducted on a transition-specific basis, ensuring that each transition was represented by the most appropriate method. This approach allowed for an optimal balance between goodness-of-fit, interpretability, and the appropriateness of long-term projections.

8.2.4 Perspective on applying CMD policy models: a hybrid approach

The CMD Policy Model developed in this thesis adopts a hybrid approach, enabling both individual-level simulation and population-level aggregation. This dual perspective enhances the model's applicability for a range of public health and policy evaluation scenarios.

By using patient-level data, the model allows for customised simulations based on individual risk profiles, accounting for differences in age, sex, or clinical biomarker levels. This individual-level flexibility supports stratified analysis, which is critical for evaluating equity impacts, targeting high-risk subgroups, or tailoring interventions. At the same time, the outputs can be aggregated to reflect population-level outcomes, such as total cardiometabolc events prevented or life-years gained under various preventative scenarios.

A key feature of this model is the ability to estimate and visualise transition probabilities between health states. These probabilities are not static; they are dynamically influenced by both individual characteristics and the duration of time spent in a given state, a feature made possible through the integration of a semi-Markov framework. The semi-Markov approach provides a more realistic and clinically plausible structure by allowing time-dependent risks to evolve based on the time since entering a health state.

As reflected in the case studies, the flexibility of the CMD Policy Model enables a wide range of applications. For instance, it can simulate how changes in

metabolic risk factors influence transition probabilities before and after an intervention. The model can also be applied using fixed baseline characteristics to generate transition probabilities for cohort-based analyses, allowing the integration of various parameters which are often sourced from published literature (such as relative risks, costs, and utility values). This adaptability makes the model suitable for both clinical and economic evaluations, supporting evidence-based decision-making across diverse policy scenarios.

Beyond that, even though it is not covered yet in case studies, the individual level parameters also can be analysed such as individual costs or quality of life that can be incorporated in the models, so the state based estimation can be obtained from those, as more reflecting mean estimation based on patient level calculations. In addition, socio-economic and ethnicity information from this model can further enrich long term analyses and address equity concerns via distributional cost-effectiveness analysis.

In terms of reproducibility and scalability, this thesis optimised modular coding practices using R. Reproducibility was maintained through structured workflows, transparent documentation, and automated model-fitting loops across 13 transitions. This allows for consistent model updating as new data become available and facilitates adaptation for different settings or subpopulations. Memory-efficient techniques were used to handle large datasets, and key modelling functions were designed for generalisability and reuse.

8.3 Policy implications

The findings of this thesis have several important policy implications, particularly for the prevention and management of cardiometabolic diseases (CMD) in the UK. The development of a multi-state CMD Policy Model based on real-world data (RWD) offers a powerful tool to support evidence-based public health decision-making. By capturing disease progression through key stages (from a disease-free state to the onset of type 2 diabetes, cardiovascular events, and death), the model enables policymakers to assess the long-term impact of interventions and resource allocation planning.

First, the model underscores the importance of early prevention. The transition from a disease-free state to CMD outcomes (e.g. type 2 diabetes, myocardial infarction, stroke) can be predicted using routinely collected clinical and biomarker data. This supports the implementation of population-level strategies, such as risk-based screening, lifestyle modification programs, and preventive care pathways. The model also facilitates identification of subgroups, reinforcing the need for stratified care and targeted interventions for individuals with modifiable risk factors (e.g., elevated BMI, hypertension, or dyslipidaemia).

Second, the integration of linked datasets (e.g., HES and ONS) enables a more comprehensive understanding of patient trajectories across both primary and secondary care settings. This aligns with NHS goals for integrated care systems and allows policymakers to capture the broader health system impact of CMD, including hospitalisations, complications, and mortality. The model therefore moves beyond traditional reliance on clinical trial data, providing a framework that incorporates real-world clinical pathways and enhances relevance for service planning.

Third, the inclusion of time-dependent covariates allows for a dynamic representation of disease risk, reflecting how patients' risk profiles evolve over time. This feature aligns with real-world clinical practice and supports policies promoting continuity of care, regular monitoring, and proactive disease management. Interventions that adapt to patient risk progression rather than relying on static baseline assessments can be better evaluated using this model structure.

Fourth, the semi-Markov structure captures disease progression where risk evolves with duration in intermediate states. It supports precise evaluation of interventions targeting long-term disease management.

Fifth, the model provides a foundation for resource allocation and economic evaluation. By enabling long-term extrapolation of disease trajectories, the model supports estimation of cost-effectiveness and potential health gains associated with preventive strategies. This is especially relevant given increasing

demands on the health system and the urgent need to balance short-term costs with long-term benefits of upstream prevention.

In practical terms, policymakers might use outputs from the CMD Policy Model to inform NHS commissioning decisions or public health investment. For example, the model could be applied to evaluate the cost-effectiveness of a national diabetes screening programme, estimating the number of cases prevented, long-term healthcare costs avoided, and quality-adjusted life years (QALYs) gained across different population segments. Similarly, the model could support resource planning by projecting the impact of dietary interventions (e.g., a sugar reduction strategy) on future rates of CVD events and hospital admissions, helping NHS and local Integrated care boards prioritise funding towards interventions that deliver the greatest health return on investment.

Finally, although health inequalities were not the central focus in this thesis, the model structure allows for stratification by socioeconomic status, ethnicity, and other relevant demographic variables. This can help policymakers understand and address disparities in CMD burden and the differential impact of interventions. For instance, the model could reveal whether the benefits of lifestyle interventions are equitably distributed, or disproportionately favour lower-risk groups. In line with this, the National Institute for Health and Care Excellence (NICE) has proposed updates to its technology evaluation manual, including explicit guidance on incorporating health inequality analysis into decision-making. These updates (currently under consultation) highlight the growing importance of embedding equity considerations into public health and health technology assessments (HTA), making the model developed in this thesis highly relevant for future policy evaluation frameworks. [394]

8.4 Limitations

Although CPRD Aurum covers approximately 20% of the UK population, there are notable gaps in geographic coverage, particularly in Scotland, Northern Ireland, and Wales. This limitation may affect the geographic representativeness of the study cohort and should be considered when generalising results to the wider UK population.

From a modelling perspective, although face and internal validation have been performed, external validation was not conducted and remains an important next step. Face validity was established through expert consultation and alignment with clinical guidelines (see Chapter 4), while internal validation involved checks on model logic, parameter consistency, and calibration against observed event rates (see Chapter 6). Future work should focus on validating the model against independent datasets and exploring its performance in different population subgroups to strengthen its generalisability and policy relevance.

The baseline cohort in this thesis was constructed using a complete-case approach, with no imputation applied. While this decision simplified the modelling process and ensured internal consistency, it may pose challenges for analysts seeking to adapt the model to different datasets, particularly those with missing values. Additionally, the model includes a rich set of covariates, which adds clinical realism but may impact computational stability in large-scale analyses or when new variables are introduced.

An area of note relates to the transition from type 2 diabetes mellitus (T2DM) to death, which showed poor model fit compared to observed data. This may reflect the influence of competing risks and the absence of intermediate events, such as diabetes-related complications, in the transition structure. This limitation highlights the importance of ongoing model refinement, particularly in transitions where disease trajectories are more complex.

The CMD Policy Model was designed to capture the natural progression of cardiometabolic disease, rather than treatment-modified pathways. While this focus is appropriate for model conceptualisation and baseline risk estimation, additional parameterisation would be required to evaluate specific interventions or treatment strategies. Consequently, adapting the model for intervention-specific cost-effectiveness analyses may require further development to reflect alternative patient pathways and treatment effects.

Although the structure and codebase are transparent and can be followed by experienced analysts or modellers, the current version does not include automated functions or user-friendly interfaces for parameter adjustment or translating state occupancy probabilities into transition probabilities. Instead, guidance is provided through example code, which requires a certain level of technical skill. Future development could address this limitation by incorporating interactive features, such as graphical interfaces or guided input forms (e.g., through R Shiny), to improve accessibility for non-technical stakeholders and support wider use in public health policy decision-making.

8.5 Areas for future research

From this work, several promising areas for potential future research can be explored.

Future studies should assess the external validity of the current policy model by comparing its performance with other relevant datasets. Such comparisons would help to determine the model's generalisability across different populations and healthcare systems.

In terms of model application, the current framework mainly demonstrates 'how the model works' using hypothetical public health interventions, rather than evaluating real health policy questions. As such, further validation is required before the model can be confidently applied to inform national decision-making. External validation remains a critical next step to assess the model's predictive performance and generalisability. Nonetheless, the framework has considerable potential for application in real-world health intervention and policy evaluation. Incorporating individual-level parameters such as patient-level cost and resource use data would enhance the model's utility for health economic evaluations.

Applying a societal perspective in economic evaluation can be beneficial, and this model is capable of accommodating such an approach. However, the current

case study focuses on applying the model from the NHS and Personal Social Services (PSS) perspective, as this was the primary objective at the time. If data available, future studies could adapt the model to adopt this perspective, allowing for a broader assessment of costs and outcomes.

Additionally, the model could be expanded to support distributional analyses, a growing area of interest in health policy formulation, especially when socio-economic data are already incorporated in the model. This extended analysis can support long-term resource planning and equity focused decision-making.

Another important step forward would be making the model more accessible and interactive. Developing friendly interface would allow users to explore different scenarios, adjust parameters, and visualise outcome in real time. Alongside this, a feasibility study employing qualitative methods could investigate how analysts and policymakers use such tools in practice, thus ensuring functionality aligns with users' needs and embedding stakeholder involvement early in the design process.

8.6 Conclusions

This thesis presents a body of work aimed at developing a flexible, data-driven Cardiometabolic Disease (CMD) Policy Model to support the evaluation of early preventative strategies and inform long-term public health planning. By leveraging complex, individual-level real-world data and applying robust survival modelling techniques, the model offers a transparent and adaptable platform for simulating disease progression and assessing the impact of population-wide policy interventions.

It is anticipated that the contributions of this research will support further methodological advancements and practical applications in the field of health policy modelling. By outlining both the potential and limitations of the approach, this thesis provides a foundation for strengthening the evidence base used in cardiometabolic health decision-making.

References

1. World Health Organization (WHO). Cardiovascular diseases (CVDs). https://www.who.int/news-room/fact sheets/detail/cardiovascular-diseases-(cvds). Accessed 5th February 2025.
2. World Health Organization (WHO). Diabetes. https://www.who.int/news-room/fact-sheets/detail/diabetes. Accessed 5th February 2025.
3. Roth, G. A. *et al.* Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet* 392, 1736–1788 (2018).
4. Sattar, N., Gill, J. M. R. & Alazawi, W. Improving prevention strategies for cardiometabolic disease. *Nat. Med.* 26, 320–325 (2020).
5. Ralston, J. & Nugent, R. Toward a broader response to cardiometabolic disease. *Nat. Med.* 25, 1644–1646 (2019).
6. Kirk, E. P. & Klein, S. Pathogenesis and Pathophysiology of the Cardiometabolic Syndrome. *J. Clin. Hypertens.* 11, 761–765 (2009).
7. Castro, J. P., El-Atat, F. A., McFarlane, S. I., Aneja, A. & Sowers, J. R. Cardiometabolic syndrome: Pathophysiology and treatment. *Curr. Hypertens. Rep.* 5, 393–401 (2003).
8. Kassi, E., Pervanidou, P., Kaltsas, G. & Chrousos, G. Metabolic syndrome: definitions and controversies. *BMC Med.* 9, 48 (2011).
9. Huang, P. L. A comprehensive definition for metabolic syndrome. *Dis. Model. Mech.* 2, 231–237 (2009).
10. Brunzell, J. D. *et al.* Lipoprotein Management in Patients With Cardiometabolic Risk. *Diabetes Care* 31, 811–822 (2008).
11. American Heart Association *et al.* Diagnosis and management of the metabolic syndrome. An American Heart Association/National Heart, Lung, and Blood Institute Scientific Statement. Executive summary. *Cardiol. Rev.* 13, 322–327 (2005).
12. Nair, A. R., Pillai, A. J. & Nair, N. Cardiovascular Changes in Menopause. *Curr. Cardiol. Rev.* 17, e230421187681 (2021).
13. Marlatt, K. L. *et al.* Body composition and cardiometabolic health across the menopause transition. *Obesity* 30, 14–27 (2022).

14. Walker, R. J., Strom Williams, J. & Egede, L. E. Influence of Race, Ethnicity and Social Determinants of Health on Diabetes Outcomes. *Am. J. Med. Sci.* 351, 366–373 (2016).
15. Yu, M. G. & King, G. L. Inflammation and Incident Diabetes: The Role of Race and Ethnicity. *J. Clin. Endocrinol. Metab.* 107, e3082–e3083 (2022).
16. Modifiable Risk Factors and Cardiovascular Outcomes. *N. Engl. J. Med.* 389, 2400–2402 (2023).
17. Salzberg, L. Risk Factors and Lifestyle Interventions. *Prim. Care Clin. Off. Pract.* 49, 201–212 (2022).
18. Sun, H. *et al.* IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Res. Clin. Pract.* 183, 109119 (2022).
19. Raghavan, S. *et al.* Diabetes Mellitus–Related All-Cause and Cardiovascular Mortality in a National Cohort of Adults. *J. Am. Heart Assoc.* 8, e011295 (2019).
20. Allen, L. *et al.* Socioeconomic status and non-communicable disease behavioural risk factors in low-income and lower-middle-income countries: a systematic review. *Lancet Glob. Health* 5, e277–e289 (2017).
21. Chan, J. C. N. *et al.* The Lancet Commission on diabetes: using data to transform diabetes care and patient lives. *Lancet Lond. Engl.* 396, 2019–2082 (2021).
22. Amini, M., Zayeri, F. & Salehi, M. Trend analysis of cardiovascular disease mortality, incidence, and mortality-to-incidence ratio: results from global burden of disease study 2017. *BMC Public Health* 21, 401 (2021).
23. Vaduganathan, M., Mensah, G. A., Turco, J. V., Fuster, V. & Roth, G. A. The Global Burden of Cardiovascular Diseases and Risk. *J. Am. Coll. Cardiol.* 80, 2361–2371 (2022).
24. Ezzati, M., Pearson-Stuttard, J., Bennett, J. E. & Mathers, C. D. Acting on non-communicable diseases in low- and middle-income tropical countries. *Nature* 559, 507–516 (2018).
25. Magliano, D. J. *et al.* Trends in the incidence of diagnosed diabetes: a multicountry analysis of aggregate data from 22 million diagnoses in high-income and middle-income settings. *Lancet Diabetes Endocrinol.* 9, 203–211 (2021).

26. Diabetes UK -Know diabetes. Fight diabetes. | Diabetes UK. https://www.diabetes.org.uk/.Accessed 5th February 2025.
27. National Institute of Care and Excellence (NICE). Prevalence | Background information | Diabetes - type 2 https://cks.nice.org.uk/topics/diabetes-type-2/background-information/prevalence/.Accessed 5th February 2025.
28. British Heart Foundation (BHF). UK Factsheet. bhf-cvd-statistics-uk-factsheet.pdf.Accessed 5th February 2025.
29. Cheema, K. M., Dicks, E., Pearson, J. & Samani, N. J. Long-term trends in the epidemiology of cardiovascular diseases in the UK: insights from the British Heart Foundation statistical compendium. *Cardiovasc. Res.* 118, 2267–2280 (2022).
30. IHME. GBD Compare. Institute for Health Metrics and Evaluation (IHME) http://vizhub.healthdata.org/gbd-compare.Accessed 10th August 2024.
31. Cardiovascular disease mortality rate in the UK 2000-2021. Statista https://www.statista.com/statistics/940678/cardiovascular-disease-mortality-rate-in-the-united-kingdom-uk/.Accessed 10th August 2024.
32. Conrad, Nathalie, et al. "Trends in cardiovascular disease incidence among 22 million people in the UK over 20 years: population based study." *Bmj* 385 (2024).
33. Health matters: preventing cardiovascular disease. *GOV.UK* https://www.gov.uk/government/publications/health-matters-preventing-cardiovascular-disease/health-matters-preventing-cardiovascular-disease.
34. British Heart Foundation (BHF). How inequalities contribute to heart and circulatory diseases. British Heart Foundation https://www.bhf.org.uk/what-we-do/our-research/heart-statistics/health-inequalities-research/inequalities-in-heart-and-circulatory-diseases-in-england.Accessed 10th August 2024.
35. Luengo-Fernandez, R. *et al.* Economic burden of cardiovascular diseases in the European Union: a population-based cost study. *Eur. Heart J.* 44, 4752–4767 (2023).
36. Bommer, C. *et al.* Global Economic Burden of Diabetes in Adults: Projections From 2015 to 2030. *Diabetes Care* 41, 963–970 (2018).
37. Einarson, T. R., Acs, A., Ludwig, C. & Panton, U. H. Economic Burden of Cardiovascular Disease in Type 2 Diabetes: A Systematic Review. *Value Health* 21, 881–890 (2018).

38. Hex, N., Bartlett, C., Wright, D., Taylor, M. & Varley, D. Estimating the current and future costs of Type 1 and Type 2 diabetes in the UK, including direct health costs and indirect societal and productivity costs. *Diabet. Med.* 29, 855–862 (2012).
39. Hex, N. *et al.* Estimation of the direct health and indirect societal costs of diabetes in the UK using a cost of illness model. *Diabet. Med.* e15326 (2024) doi:10.1111/dme.15326.
40. NHS Prevention Programme cuts chances of Type 2 diabetes for thousands. https://www.england.nhs.uk/2022/03/nhs-prevention-programme-cuts-chances-of-type-2-diabetes-for-thousands/.Accessed 10th August 2024.
41. Shih, K., Herz, N., Sheikh, A., O'Neill, C., Carter, P., & Anderson, M. Economic burden of cardiovascular disease in the United Kingdom. European Heart Journal-Quality of Care and Clinical Outcomes, qcaf011(2025).
42. Outwater, A. H., Leshabari, S. C. & Nolte, E. Disease Prevention: An Overview. in *International Encyclopedia of Public Health (Second Edition)* (ed. Quah, S. R.) 338–349 (Academic Press, Oxford, 2017). doi:10.1016/B978-0-12-803678-5.00117-X.
43. AbdulRaheem, Y. Unveiling the Significance and Challenges of Integrating Prevention Levels in Healthcare Practice. *J. Prim. Care Community Health* 14, 21501319231186500 (2023).
44. National Institute of Care and Excellence (NICE). Cardiometabolic disease prevention and treatment guidelines.2025. https://www.nice.org.uk/hub/indevelopment/gid-hub10002. Accessed 3rd July 2024.
45. Scottish Intercollegiate Guidelines Network (SIGN). Risk estimation and the prevention of cardiovascular disease.2017. https://www.sign.ac.uk/assets/qrg149.pdf. Accessed 3rd July 2024
46. National Institute of Care and Excellence (NICE). Type 2 diabetes in adults: management.2022. https://www.nice.org.uk/guidance/ng28/resources/type-2-diabetes-in-adults-management-pdf-1837338615493. Accessed 3rd July 2024
47. National Institute of Care and Excellence (NICE). Type 2 diabetes in adults: management.2022.

https://www.nice.org.uk/guidance/ng28/resources/type-2-diabetes-in-adults-management-pdf-1837338615493. Accessed 3rd July 2024.

48. National Institute of Care and Excellence (NICE). Type 2 diabetes prevention: population and community-level interventions. 2011. https://www.nice.org.uk/guidance/ph35/resources/type-2-diabetes-prevention-population-and-communitylevel-interventions-pdf-1996299153349. Accessed 3rd July 2024.
49. National Institute of Care and Excellence (NICE). Cardiovascular risk assessment and lipid modification. 2015 https://www.nice.org.uk/guidance/qs100/resources/cardiovascular-risk-assessment-and-lipid-modification-pdf-75545230925509. Accessed 3rd July 2024.
50. National Institute of Care and Excellence (NICE). Cardiovascular disease prevention. 2010. https://www.nice.org.uk/guidance/ph25. Accessed 3rd July 2024.
51. Jit, M. & Cook, A. R. Informing Public Health Policies with Models for Disease Burden, Impact Evaluation, and Economic Evaluation. *Annu. Rev. Public Health* 45, 133–150 (2024).
52. Weinstein, M. C. *et al.* Principles of Good Practice for Decision Analytic Modeling in Health-Care Evaluation: Report of the ISPOR Task Force on Good Research Practices—Modeling Studies. *Value Health* 6, 9–17 (2003).
53. Thokala, P. *et al.* Assessing the cost-effectiveness of type 1 diabetes interventions: the Sheffield type 1 diabetes policy model. *Diabet. Med. J. Br. Diabet. Assoc.* 31, 477–486 (2014).
54. Chief Scientist Office (CSO). Developing a Scottish Type 2 Diabetes Policy Model and Web-based Decision Aid "Dashboard". 2021. https://www.cso.scot.nhs.uk/wp-content/uploads/HIPS1726-2.pdf. Accessed 3rd July 2024
55. Lewsey, J. D. *et al.* A cardiovascular disease policy model that predicts life expectancy taking into account socioeconomic deprivation. *Heart* 101, 201–208 (2015).
56. Lawson, K. D. *et al.* A cardiovascular disease policy model: part 2—preparing for economic evaluation and to assess health inequalities. *Open Heart* 3, e000140 (2016).

57. Cardiovascular Disease Policy Model. *Epidemiology & Biostatistics* https://epibiostat.ucsf.edu/cardiovascular-disease-policy-model.
58. Sánchez-Romero, L. M. *et al.* Projected Impact of Mexico's Sugar-Sweetened Beverage Tax Policy on Diabetes and Cardiovascular Disease: A Modeling Study. *PLoS Med.* 13, e1002158 (2016).
59. Riveros, B. S. *et al.* Brazilian Analytical Decision Model for Cardiovascular Disease: An Adaptation of the Scottish Cardiovascular Disease Policy Model. *Value Health Reg. Issues* 17, 210–216 (2018).
60. Konfino, J., Mekonnen, T. A., Coxson, P. G., Ferrante, D. & Bibbins-Domingo, K. Projected Impact of a Sodium Consumption Reduction Initiative in Argentina: An Analysis from the CVD Policy Model – Argentina. *PLOS ONE* 8, e73824 (2013).
61. Eddy, D. M. *et al.* Model Transparency and Validation: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force-7. *Value Health* 15, 843–850 (2012).
62. Caro, J. J., Briggs, A. H., Siebert, U. & Kuntz, K. M. Modeling Good Research Practices—Overview: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force-1. *Value Health* 15, 796–803 (2012).
63. Drummond, M. & Sculpher, M. Common Methodological Flaws in Economic Evaluations. *Med. Care* 43, II (2005).
64. Bodrogi, J. & Kaló, Z. Principles of pharmacoeconomics and their impact on strategic imperatives of pharmaceutical research and development. *Br. J. Pharmacol.* 159, 1367–1373 (2010).
65. Gomersall, J. *et al.* The Systematic Review of Economic Evaluation Evidence. in (2014).
66. Guerre, P. *et al.* Quality Assessment of Reporting of Economic Evaluation in Cardiac Sugery: Has it Improved? *Int. J. Technol. Assess. Health Care* 35, 45–49 (2019).
67. Edmunds, K., Ling, R., Shakeshaft, A., Doran, C. & Searles, A. Systematic review of economic evaluations of interventions for high risk young people. *BMC Health Serv. Res.* 18, 1–10 (2018).
68. Kadu, M., Ehrenberg, N., Stein, V. & Tsiachristas, A. Methodological Quality of Economic Evaluations in Integrated Care: Evidence from a Systematic Review. *Int. J. Integr. Care* 19, (2019).

69. Philips, Z., Bojke, L., Sculpher, M., Claxton, K. & Golder, S. Good Practice Guidelines for Decision-Analytic Modelling in Health Technology Assessment. *PharmacoEconomics* 24, 355–371 (2006).
70. Becker, C., Langer, A. & Leidl, R. The quality of three decision-analytic diabetes models: a systematic health economic assessment. *Expert Rev. Pharmacoecon. Outcomes Res.* 11, 751–762 (2011).
71. Drummond, M. & Sculpher, M. Common methodological flaws in economic evaluations. *Med. Care* 43, 5–14 (2005).
72. Kim, D. D. *et al.* The influence of time horizon on results of cost-effectiveness analyses. *Expert Rev. Pharmacoecon. Outcomes Res.* 17, 615–623 (2017).
73. Drummond, M. *Methods for the Economic Evaluation of Health Care Programmes*. (Oxford University Press, Oxford, United Kingdom ; New York, NY, USA, 2015).
74. Cookson, R. *et al.* Using Cost-Effectiveness Analysis to Address Health Equity Concerns. *Value Health* 20, 206–212 (2017).
75. Patikorn, C., Cho, J.-Y., Lambach, P., Hutubessy, R. & Chaiyakunapruk, N. Equity-Informative Economic Evaluations of Vaccines: A Systematic Literature Review. *Vaccines* 11, 622 (2023).
76. Gabrielle, P.-H. *et al.* From randomised controlled trials to real-world data: Clinical evidence to guide management of diabetic macular oedema. *Prog. Retin. Eye Res.* 97, 101219 (2023).
77. Dang, A. Real-World Evidence: A Primer. *Pharm. Med.* 37, 25–36 (2023).
78. Liu, F. & Panagiotakos, D. Real-world data: a brief review of the methods, applications, challenges and opportunities. *BMC Med. Res. Methodol.* 22, 287 (2022).
79. White, R. Building trust in real world evidence (RWE): moving transparency in RWE towards the randomized controlled trial standard. *Curr. Med. Res. Opin.* 39, 1737–1741 (2023).
80. Franklin, J. M. & Schneeweiss, S. When and How Can Real World Data Analyses Substitute for Randomized Controlled Trials? *Clin. Pharmacol. Ther.* 102, 924–933 (2017).
81. NHS Digital. *GOV.UK* https://www.gov.uk/government/organisations/nhs-digital (2022).

82. Clinical Practice Research Datalink. http://www.cprd.com/.Accessed 18th January 2025.
83. Hospital Episode Statistics (HES). NHS England Digital https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics. Accessed 18th January 2025.
84. UK Biobank - UK Biobank. https://www.ukbiobank.ac.uk (2024).
85. Biobank | RegisterforSHARE. Register for SHARE https://www.registerforshare.org.Accessed 18th January 2025.https://www.registerforshare.org.
86. Razieh, C. *et al.* Differences in the risk of cardiovascular disease across ethnic groups: UK Biobank observational study. *Nutr. Metab. Cardiovasc. Dis.* 32, 2594–2602 (2022).
87. Buckland, G., Taylor, C. M., Emmett, P. M. & Northstone, K. Prospective association between adherence to UK dietary guidelines in school-age children and cardiometabolic risk markers in adolescence/early adulthood in the Avon Longitudinal Study of Parents and Children (ALSPAC) cohort. *Br. J. Nutr.* 130, 1766–1778.
88. Eriksen, R. *et al.* Nutrient profiling and adherence to components of the UK national dietary guidelines association with metabolic risk factors for CVD and diabetes: Airwave Health Monitoring Study. *Br. J. Nutr.* 119, 695–705 (2018).
89. Dobson, L. E. & Prendergast, B. D. Heart valve disease: a journey of discovery. *Heart* 108, 774–779 (2022).
90. Canoy, D. *et al.* Association between cardiometabolic disease multimorbidity and all-cause mortality in 2 million women and men registered in UK general practices. *BMC Med.* 19, 1–10 (2021).
91. Gulliford, M. C., Sun, X., Anjuman, T., Yelland, E. & Murray-Thomas, T. Comparison of antibiotic prescribing records in two UK primary care electronic health record systems: cohort study using CPRD GOLD and CPRD Aurum databases. *BMJ Open* 10, e038767 (2020).
92. Makady, A. *et al.* What Is Real-World Data? A Review of Definitions Based on Literature and Stakeholder Interviews. *Value Health J. Int. Soc. Pharmacoeconomics Outcomes Res.* 20, 858–865 (2017).
93. Introduction to real-world evidence in NICE decision making | NICE real-world evidence framework | Guidance | NICE.

https://www.nice.org.uk/corporate/ecd9/chapter/introduction-to-real-world-evidence-in-nice-decision-making (2022).

94. Katkade, V. B., Sanders, K. N. & Zou, K. H. Real world data: an opportunity to supplement existing evidence for the use of long-established medicines in health care decision making. *J. Multidiscip. Healthc.* 11, 295–304 (2018).
95. Kang, J. Real-World Data in Health Technology Assessment: Do We Know It Well Enough? in *Precision Oncology and Cancer Biomarkers: Issues at Stake and Matters of Concern* (eds Bremer, A. & Strand, R.) 187–203 (Springer International Publishing, Cham, 2022). doi:10.1007/978-3-030-92612-0_12.
96. Zisis, K., Pavi, E., Geitona, M. & Athanasakis, K. Real-world data: a comprehensive literature review on the barriers, challenges, and opportunities associated with their inclusion in the health technology assessment process. *J. Pharm. Pharm. Sci.* 27, 12302 (2024).
97. O'Reilly, D., Gaebel, K., Xie, F., Tarride, J.-E. & Goeree, R. Health economic evaluations help inform payers of the best use of scarce health care resources. *Int. J. Circumpolar Health* 70, 419–427 (2011).
98. Stevanovic, J., Postma, M. J. & Pechlivanoglou, P. A systematic review on the application of cardiovascular risk prediction models in pharmacoeconomics, with a focus on primary prevention. *Eur. J. Prev. Cardiol.* 19, 42–53 (2012).
99. Suhrcke, M., Boluarte, T. A. & Niessen, L. A systematic review of economic evaluations of interventions to tackle cardiovascular disease in low- and middle-income countries. *BMC Public Health* 12, 2 (2012).
100. Li, J., Bao, Y., Chen, X. & Tian, L. Decision models in type 2 diabetes mellitus: A systematic review. *Acta Diabetol.* 58, 1451–1469 (2021).
101. Unal, B., Capewell, S. & Critchley, J. A. Coronary heart disease policy models: a systematic review. *BMC Public Health* 6, 213 (2006).
102. Hiligsmann, M., Wyers, C. E., Mayer, S., Evers, S. M. & Ruwaard, D. A systematic review of economic evaluations of screening programmes for cardiometabolic diseases. *Eur. J. Public Health* 27, 621–631 (2017).
103. Cummins, E. *et al.* Clinical effectiveness and cost-effectiveness of continuous subcutaneous insulin infusion for diabetes: systematic review and economic evaluation. *Health Technol. Assess. Winch. Engl.* 14, iii–iv, xi–xvi, 1–181 (2010).

104. Abushanab, D. *et al*. A Systematic Review of Cost-Effectiveness of Non-Statin Lipid-Lowering Drugs for Primary and Secondary Prevention of Cardiovascular Disease in Patients with Type 2 Diabetes Mellitus. *Curr. Probl. Cardiol.* 48, 101211 (2023).
105. Roberts, S. *et al*. Preventing type 2 diabetes: systematic review of studies of cost-effectiveness of lifestyle programmes and metformin, with and without screening, for pre-diabetes. *BMJ Open* 7, e017184 (2017).
106. Leal, J., Morrow, L. M., Khurshid, W., Pagano, E. & Feenstra, T. Decision models of prediabetes populations: A systematic review. *Diabetes Obes. Metab.* 21, 1558–1569 (2019).
107. Onyimadu, O. *et al*. A systematic review of economic evaluations of interventions targeting childhood overweight and obesity. *Obes. Rev.* 24, e13597 (2023).
108. Twumwaa, T. E., Justice, N., Robert, van D. M. & Itamar, M. Application of decision analytical models to diabetes in low- and middle-income countries: a systematic review. *BMC Health Serv. Res.* 22, 1–10 (2022).
109. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews | EQUATOR Network. https://www.equator-network.org/reporting-guidelines/prisma/.
110. Putri, S. et al. Cardiometabolic Diseases Prevention Policy Models: A Systematic Review. PROSPERO 2024 Available from https://www.crd.york.ac.uk/PROSPERO/view/CRD42022354399
111. A snowballing technique to ensure comprehensiveness of search for systematic reviews: A case study | Cochrane Colloquium Abstracts. https://abstracts.cochrane.org/2011-madrid/snowballing-technique-ensure-comprehensiveness-search-systematic-reviews-case-study.
112. Nussbaumer-Streit, B. *et al*. Rapid reviews methods series: Guidance on team considerations, study selection, data extraction and risk of bias assessment. *BMJ Evid.-Based Med.* 28, 418–423 (2023).
113. Moran, A. *et al*. The future impact of population growth and aging on coronary heart disease in China: projections from the Coronary Heart Disease Policy Model-China. *BMC Public Health* 8, 1–14 (2008).
114. Moran, A. *et al*. NIH Public Access. 3, 243–252 (2010).
115. Bibbins-Domingo, K. *et al*. Projected Effect of Dietary Salt Reductions on Future Cardiovascular Disease. *N. Engl. J. Med.* 362, 590–599 (2010).

116. Wang, Y. C., Coxson, P., Shen, Y. M., Goldman, L. & Bibbins-Domingo, K. A penny-per-ounce tax on sugar-sweetened beverages would cut health and cost burdens of diabetes. *Health Aff. (Millwood)* 31, 199–207 (2012).
117. Basu, S., Seligman, H. & Bhattacharya, J. Nutritional policy changes in the supplemental nutrition assistance program: A microsimulation and cost-effectiveness analysis. *Med. Decis. Making* 33, 937–948 (2013).
118. Basu, S. *et al*. Averting Obesity and Type 2 Diabetes in India through Sugar-Sweetened Beverage Taxation: An Economic-Epidemiologic Modeling Study. *PLoS Med.* 11, (2014).
119. Collins, M. *et al*. An economic evaluation of salt reduction policies to reduce coronary heart disease in england: A policy modeling study. *Value Health* 17, 517–524 (2014).
120. Mason, H. *et al*. A cost effectiveness analysis of salt reduction policies to reduce coronary heart disease in four Eastern Mediterranean countries. *PLoS ONE* 9, (2014).
121. Manyema, M. *et al*. Decreasing the burden of type 2 diabetes in South Africa: The impact of taxing sugar-sweetened beverages. *PLoS ONE* 10, 1–17 (2015).
122. Wilcox, M. L. *et al*. Cost-effectiveness analysis of salt reduction policies to reduce coronary heart disease in Syria, 2010–2020. *Int. J. Public Health* 60, 23–30 (2014).
123. Collins, B. *et al*. Modelling the health impact of an English sugary drinks duty at national and local levels. *PLoS ONE* 10, 1–13 (2015).
124. Wang, M. *et al*. Projected impact of salt restriction on prevention of cardiovascular disease in China: A modeling study. *PLoS ONE* 11, 1–16 (2016).
125. Breeze, P. R. *et al*. Cost-effectiveness of population-based, community, workplace and individual policies for diabetes prevention in the UK. *Diabet. Med.* 34, 1136–1144 (2017).
126. Pandya, A. *et al*. Validation of a Cardiovascular Disease Policy Microsimulation Model Using Both Survival and Receiver Operating Characteristic Curves. *Med. Decis. Making* 37, 802–814 (2017).
127. Mozaffarian, D. *et al*. Cost-effectiveness of financial incentives and disincentives for improving food purchases and health through the US

Supplemental Nutrition Assistance Program (SNAP): A microsimulation study. *PLoS Med.* 15, 1–25 (2018).

128. Riveros, B. S. *et al.* Brazilian Analytical Decision Model for Cardiovascular Disease: An Adaptation of the Scottish Cardiovascular Disease Policy Model. *Value Health Reg. Issues* 17, 210–216 (2018).
129. Schönbach, J.-K., Thiele, S. & Lhachimi, S. K. What are the potential preventive population-health effects of a tax on processed meat? A quantitative health impact assessment for Germany. *Prev. Med.* 118, 325–331 (2019).
130. Huang, Y. *et al.* Cost-Effectiveness of the U.S. FDA Added Sugar Labeling Policy for Improving Diet and Health. *Circulation* 139, 2613–2624 (2019).
131. Salgado, M. V. et al. Update of the Cardiovascular Disease Policy Model to Predict Cardiovascular Events in Argentina.
132. Wilde, P. *et al.* Cost-effectiveness of a US national sugar-sweetened beverage tax with a multistakeholder approach: Who pays and who benefits. *Am. J. Public Health* 109, 276–284 (2019).
133. Broeks, M. J. *et al.* A social cost-benefit analysis of meat taxation and a fruit and vegetables subsidy for a healthy and sustainable food consumption in the Netherlands. *BMC Public Health* 20, 1–12 (2020).
134. Lee, Y. *et al.* Health impact and cost-effectiveness of volume, tiered, and absolute sugar content sugar-sweetened beverage tax policies in the United States: A microsimulation study. *Circulation* 142, 523–534 (2020).
135. Liu, J. *et al.* Health and Economic Impacts of the National Menu Calorie Labeling Law in the United States: A Microsimulation Study. *Circ. Cardiovasc. Qual. Outcomes* 13, e006313 (2020).
136. Salgado, M. V. *et al.* Projected impact of a reduction in sugar-sweetened beverage consumption on diabetes and cardiovascular disease in Argentina: A modeling study. *PLoS Med.* 17, (2020).
137. Dehmer, S. P. *et al.* Health and Budgetary Impact of Achieving 10-Year U.S. Sodium Reduction Targets. https://stacks.cdc.gov/view/cdc/99828.
138. Shangguan, S. *et al.* Health Impact and Cost-Effectiveness of Achieving the National Salt and Sugar Reduction Initiative Voluntary Sugar Reduction Targets in the United States: A Micro-Simulation Study. *Circulation* 1362–1376 (2021) doi:10.1161/circulationaha.121.053678.

139. Thomas, C. *et al*. The health, cost and equity impacts of restrictions on the advertisement of high fat, salt and sugar products across the transport for London network: a health economic modelling study. *Int. J. Behav. Nutr. Phys. Act.* 19, 1–12 (2022).
140. Lou, Z. *et al*. The Health and Economic Impact of Using a Sugar Sweetened Beverage Tax to Fund Fruit and Vegetable Subsidies in New York City: A Modeling Study. *J. Urban Health Bull. N. Y. Acad. Med.* 100, 51–62 (2023).
141. Ismail, L., Materwala, H. & Al Kaabi, J. Association of risk factors with type 2 diabetes: A systematic review. *Comput. Struct. Biotechnol. J.* 19, 1759–1785 (2021).
142. Brennan, A., Chick, S. E. & Davies, R. A taxonomy of model structures for economic evaluation of health technologies. *Health Econ.* 15, 1295–1310 (2006).
143. Briggs, A. *et al*. *Decision Modelling for Health Economic Evaluation*. (Oxford University Press, Oxford, New York, 2006).
144. Caro, J. J. & Möller, J. Decision-Analytic Models: Current Methodological Challenges. *PharmacoEconomics* 32, 943–950 (2014).
145. Siebert, U. *et al*. State-transition modeling: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force--3. *Value Health J. Int. Soc. Pharmacoeconomics Outcomes Res.* 15, 812–820 (2012).
146. Sculpher, M., Fenwick, E. & Claxton, K. Assessing quality in decision analytic cost-effectiveness models. A suggested framework and example of application. *PharmacoEconomics* 17, 461–477 (2000).
147. Eddy, D. M. *et al*. Model Transparency and Validation: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force–7. *Med. Decis. Making* 32, 733–743 (2012).
148. Karnon, J. *et al*. Modeling using discrete event simulation: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force--4. *Value Health J. Int. Soc. Pharmacoeconomics Outcomes Res.* 15, 821–827 (2012).
149. Sánchez-Romero, L. M. *et al*. Projected Impact of Mexico's Sugar-Sweetened Beverage Tax Policy on Diabetes and Cardiovascular Disease: A Modeling Study. *PLoS Med.* 13, 1–17 (2016).
150. Lou, Z. *et al*. The Health and Economic Impact of Using a Sugar Sweetened Beverage Tax to Fund Fruit and Vegetable Subsidies in New York City: A Modeling Study. *J. Urban Health Bull. N. Y. Acad. Med.* 100, 51–62 (2023).

151. Liu, S., Veugelers, P. J., Liu, C. & Ohinmaa, A. The Cost Effectiveness of Taxation of Sugary Foods and Beverages: A Systematic Review of Economic Evaluations. *Appl. Health Econ. Health Policy.* 20, 185–198 (2022).
152. Veerman, J. L., Sacks, G., Antonopoulos, N. & Martin, J. The impact of a tax on sugar-sweetened beverages on health and health care costs: A modelling study. *PLoS ONE* 11, 1–10 (2016).
153. Vemer, P., Corro Ramos, I., van Voorn, G. A. K., Al, M. J. & Feenstra, T. L. AdViSHE: A Validation-Assessment Tool of Health-Economic Models for Decision Makers and Model Users. *Pharmacoeconomics* 34, 349–361 (2016).
154. Kim, L. G. & Thompson, S. G. Uncertainty and validation of health economic decision models. *Health Econ.* 19, 43–55 (2010).
155. Briggs, A. H. *et al.* Model Parameter Estimation and Uncertainty: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force-6. *Value Health* 15, 835–842 (2012).
156. Putri, S. *et al.* The conceptualisation of cardiometabolic disease policy model in the UK. *BMC Health Serv. Res.* 24, 1060 (2024).
157. Weinstein, M. C. *et al.* Principles of Good Practice for Decision Analytic Modeling in Health-Care Evaluation: Report of the ISPOR Task Force on Good Research Practices—Modeling Studies. *Value Health* 6, 9–17 (2003).
158. Briggs, A., Sculpher, M. & Claxton, K. *Decision Modelling for Health Economic Evaluation*. (OUP Oxford, 2006).
159. Drost, R. M. W. A., Putten, I. M. van der, Ruwaard, D., Evers, S. M. A. A. & Paulus, A. T. G. Conceptualizations of the societal perspective within economic evaluations: a systemtic review. Int. J. Technol. Assess. Health Care 33, 251–260 (2017).
160. Culyer, A. J. & Wagstaff, A. Equity and equality in health and health care. *J. Health Econ.* 12, 431–457 (1993).
161. Cookson, R., Griffin, S., Norheim, O. F., Culyer, A. J. & Chalkidou, K. Distributional Cost-Effectiveness Analysis Comes of Age. *Value Health* 24, 118–120 (2021).
162. Cookson, R. *et al.* Using Cost-Effectiveness Analysis to Address Health Equity Concerns. *Value Health* 20, 206–212 (2017).
163. Dötsch-Klerk, M. *et al.* Modelling health and economic impact of nutrition interventions: a systematic review. *Eur. J. Clin. Nutr.* 77, 413–426 (2023).

164. Mertens, E., Genbrugge, E., Ocira, J. & Peñalvo, J. L. Microsimulation Modeling in Food Policy: A Scoping Review of Methodological Aspects. *Adv. Nutr. Bethesda Md* 13, 621–632 (2022).
165. Wu, O. Microsimulation Model for Health Economic Evaluation of Public Health Policies: An Imperfect but Useful Tool. https://www.ahajournals.org/doi/epub/10.1161/CIRCULATIONAHA.121.056502 doi:10.1161/CIRCULATIONAHA.121.056502.
166. Kretzschmar, M. Disease modeling for public health: added value, challenges, and institutional constraints. *J. Public Health Policy* 41, 39–51 (2020).
167. Roberts, M. *et al.* Conceptualizing a Model: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force-2. *Value Health J. Int. Soc. Pharmacoeconomics Outcomes Res.* 15, 804–811 (2012).
168. Tappenden, P. Conceptual modelling for health economic model development.
169. Montague, R. Conceptual modelling and NICE requirements: a significant change or hill of beans? *Source Health Economics* https://source-he.com/conceptual-modelling-and-nice-requirements-a-significant-change-or-hill-of-beans/ (2022).
170. Huygens, S. A. *et al.* Conceptual model for early health technology assessment of current and novel heart valve interventions. *Open Heart* 3, e000500 (2016).
171. McMeekin, N. J. M. Conceptual models in health economic evaluation: a new role. (University of Glasgow, 2021). doi:10.5525/gla.thesis.82570.
172. Khan, A. R. *et al.* The Promising Frontier of Cardiometabolic Syndrome: A New Paradigm in Cardiology. *Cureus* 15, e45542.
173. Rutter, M. K., Meigs, J. B., Sullivan, L. M., D'Agostino, R. B. & Wilson, P. W. F. C-reactive protein, the metabolic syndrome, and prediction of cardiovascular events in the Framingham Offspring Study. *Circulation* 110, 380–385 (2004).
174. Ceska, R. Clinical implications of the metabolic syndrome. *Diab. Vasc. Dis. Res.* 4, S2–S4 (2007).
175. World Health Organization. (1999). Definition, diagnosis and classification of diabetes mellitus and its complications : report of a WHO consultation.

Part 1, Diagnosis and classification of diabetes mellitus. World Health Organization. https://iris.who.int/handle/10665/66040.

176. International Diabetes Federation (IDF). the IDF consensus worldwide definition of the metabolic syndrome. 2006. https://idf.org/media/uploads/2023/05/attachments-30.pdf
177. Alberti, K. G. M. M., Zimmet, P. & Shaw, J. Metabolic syndrome—a new world-wide definition. A Consensus Statement from the International Diabetes Federation. *Diabet. Med.* 23, 469–480 (2006).
178. Executive Summary of the Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults T HE THIRD REPORT OF THE EX-pert Panel on Detection, Evalu-ation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III, or ATP III) constitutes the National.
179. Grundy, S. M. *et al*. Diagnosis and management of the metabolic syndrome. *Current Opinion in Cardiology* vol. 21 1–6 (2006).
180. Mechanick, J. I., Farkouh, M. E., Newman, J. D. & Garvey, W. T. Cardiometabolic-Based Chronic Disease, Addressing Knowledge and Clinical Practice Gaps: JACC State-of-the-Art Review. *Journal of the American College of Cardiology* vol. 75 539–555 (2020).
181. Guo, F., Moellering, D. R. & Garvey, W. T. The Progression of Cardiometabolic Disease: Validation of a New Cardiometabolic Disease Staging System Applicable to Obesity. *Obes. Silver Spring Md* 22, 110–118 (2014).
182. National Institute of Care and Excellence (NICE). 2023. https://www.nice.org.uk/guidance/ng238. Accessed 25th August 2024.
183. National Institute of Care and Excellence (NICE). Overview Type 2 diabetes prevention: population and community-level interventions .https://www.nice.org.uk/guidance/Ph35 (2011).Accessed 25th August 2024.
184. Scottish Intercollegiate Guidelines Network (SIGN). Risk estimation and the prevention of cardiovascular disease.2017. https://www.sign.ac.uk/assets/qrg149.pdf. Accessed 3rd July 2024.

185. Putri, S. *et al*. Policy models for preventative interventions in cardiometabolic diseases: a systematic review. *BMC Health Serv. Res.* 25, 635 (2025).
186. Menezes, A. R. *et al*. Cardiometabolic risk factors and atrial fibrillation. *Rev. Cardiovasc. Med.* 14, e73-81 (2013).
187. Gregg, E. W. *et al*. Association of an intensive lifestyle intervention with remission of type 2 diabetes. *JAMA* 308, 2489–2496 (2012).
188. Lean, M. E. J. *et al*. Durability of a primary care-led weight-management intervention for remission of type 2 diabetes: 2-year results of the DiRECT open-label, cluster-randomised trial. *Lancet Diabetes Endocrinol.* 7, 344–355 (2019).
189. Kleinbaum, D. G. & Klein, M. Introduction to Survival Analysis. in *Survival Analysis: A Self-Learning Text* (eds Kleinbaum, D. G. & Klein, M.) 1–54 (Springer, New York, NY, 2012). doi:10.1007/978-1-4419-6646-9_1.
190. Cox, D. R. Regression Models and Life-Tables. *J. R. Stat. Soc. Ser. B Methodol.* 34, 187–220 (1972).
191. Abner, E. L., Charnigo, R. J. & Kryscio, R. J. Markov chains and semi-Markov models in time-to-event analysis. *J. Biom. Biostat.* Suppl 1, 19522 (2013).
192. Asanjarani, A., Liquet, B. & Nazarathy, Y. Estimation of semi-Markov multi-state models: a comparison of the sojourn times and transition intensities approaches. *Int. J. Biostat.* 18, 243–262 (2022).
193. Ferrer, L. *et al*. Joint modelling of longitudinal and multi-state processes: application to clinical progressions in prostate cancer. *Stat. Med.* 35, 3933–3948 (2016).
194. Clinical Practice Research Datalink. https://www.cprd.com/.
195. Wolf, A. *et al*. Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum. *Int. J. Epidemiol.* 48, 1740–1740g (2019).
196. Haji Ali Afzali, H., Bojke, L. & Karnon, J. Model Structuring for Economic Evaluations of New Health Technologies. *PharmacoEconomics* 36, 1309–1319 (2018).
197. Afzali, H. H. A. *et al*. Structuring a conceptual model for cost-effectiveness analysis of frailty interventions. *PLOS ONE* 14, e0222049 (2019).

198. Alemao, E. *et al*. Conceptual model for the health technology assessment of current and novel interventions in rheumatoid arthritis. *PLOS ONE* 13, e0205013 (2018).
199. Gonzalez-McQuire, S. *et al*. Development of an Initial Conceptual Model of Multiple Myeloma to Support Clinical and Health Economics Decision Making. *MDM Policy Pract*. 4, 2381468318814253 (2019).
200. Feng, W., Hendry, R. M. & Adams, R. J. Risk of recurrent stroke, myocardial infarction, or death in hospitalized stroke patients. *Neurology* 74, 588–593 (2010).
201. Bots, M. L. & Grobbee, D. E. Intima media thickness as a surrogate marker for generalised atherosclerosis. *Cardiovasc. Drugs Ther*. 16, 341–351 (2002).
202. Position statement on use of the EQ-5D-5L value set for England (updated October 2019) | Technology appraisal guidance | NICE guidance | Our programmes | What we do | About. *NICE* https://www.nice.org.uk/about/what-we-do/our-programmes/nice-guidance/technology-appraisal-guidance/eq-5d-5l.
203. Lee, S. *et al*. Unlocking the Potential of Electronic Health Records for Health Research. *Int. J. Popul. Data Sci*. 5, 1123.
204. Casey, J. A., Schwartz, B. S., Stewart, W. F. & Adler, N. E. Using Electronic Health Records for Population Health Research: A Review of Methods and Applications. *Annu. Rev. Public Health* 37, 61–81 (2016).
205. Morrato, E. H., Elias, M. & Gericke, C. A. Using population-based routine data for evidence-based health policy decisions: lessons from three examples of setting and evaluating national health policy in Australia, the UK and the USA. *J. Public Health Oxf. Engl*. 29, 463–471 (2007).
206. Lawson, C. A. *et al*. Outcome trends in people with heart failure, type 2 diabetes mellitus and chronic kidney disease in the UK over twenty years. *eClinicalMedicine* 32, (2021).
207. Haase, C. L. *et al*. Body mass index and risk of obesity-related conditions in a cohort of 2.9 million people: Evidence from a UK primary care database. *Obes. Sci. Pract*. 7, 137–147 (2021).
208. Clarke, G. M., Conti, S., Wolters, A. T. & Steventon, A. Evaluating the impact of healthcare interventions using routine data. *BMJ* l2239 (2019) doi:10.1136/bmj.l2239.

209. Hung, Y. W., Hoxha, K., Irwin, B. R., Law, M. R. & Grépin, K. A. Using routine health information data for research in low- and middle-income countries: a systematic review. *BMC Health Serv. Res.* 20, 790 (2020).
210. Primary care data for public health research. https://www.cprd.com/primary-care-data-public-health-research (2024).
211. Herrett, E. *et al.* Data Resource Profile: Clinical Practice Research Datalink (CPRD). *Int. J. Epidemiol.* 44, 827–836 (2015).
212. Oyinlola, J. O., Campbell, J. & Kousoulis, A. A. Is real world evidence influencing practice? A systematic review of CPRD research in NICE guidances. *BMC Health Serv. Res.* 16, 299 (2016).
213. Miao, Z. *et al.* A data preparation framework for cleaning electronic health records and assessing cleaning outcomes for secondary analysis. *Inf. Syst.* 111, 102130 (2023).
214. Grath-Lone, L. M. *et al.* What makes administrative data research-ready? : A systematic review and thematic analysis of published literature. *Int. J. Popul. Data Sci.* 7, (2022).
215. CPRD Aurum March 2024 dataset. https://www.cprd.com/doi/cprd-aurum-march-2024-dataset.
216. CPRD. Data quality. https://www.cprd.com/data-quality (2024).
217. Jick, S. *et al.* Use of the CPRD Aurum Database: Insights Gained from New Data Quality Assessments. *Clin. Epidemiol.* 15, 1219–1222 (2023).
218. Hagberg, K. W. *et al.* Correctness and Completeness of Breast Cancer Diagnoses Recorded in UK CPRD Aurum and CPRD GOLD Databases: Comparison to Hospital Episode Statistics and Cancer Registry (Companion Paper 2). *Clin. Epidemiol.* 15, 1193–1206 (2023).
219. Vasilakis-Scaramozza, Catherine, et al. "Quality of rheumatoid arthritis recording in United Kingdom clinical practice research datalink aurum." *Pharmacoepidemiology and Drug Safety* 32.1 (2023): 73-77.
220. Jick, S. S. *et al.* Quality and completeness of diagnoses recorded in the new CPRD Aurum Database: evaluation of pulmonary embolism. *Pharmacoepidemiol. Drug Saf.* 29, 1134–1140 (2020).
221. Persson, R. *et al.* CPRD Aurum database: Assessment of data quality and completeness of three important comorbidities. *Pharmacoepidemiol. Drug Saf.* 29, 1456–1464 (2020).

222. Hughes, J. Introducing a new R package to process primary care data. *The Health Foundation Data Analytics* https://medium.com/healthfdn-data-analytics/introducing-a-new-r-package-to-process-primary-care-data-88e863130ada (2021).
223. Levy, E. What is the Parquet File Format? Use Cases & Benefits. *Upsolver* https://www.upsolver.com/blog/apache-parquet-why-use (2023).
224. https://cran.r-project.org/web/packages/dplyr/vignettes/dplyr.html
225. Putter, H. Tutorial in biostatistics: Competing risks and multi-state.
226. Exeter-Diabetes/CPRD-Codelists: CPRD Aurum codelists from the Exeter Diabetes team. https://github.com/Exeter-Diabetes/CPRD-Codelists.
227. Davidson, J. Clinical codelist - CPRD Aurum - major adverse cardiovascular event. London School of Hygiene & Tropical Medicine https://doi.org/10.17037/DATA.00002196 (2021).
228. Davidson, J. Clinical codelist - CPRD Aurum - atrial fibrillation. London School of Hygiene & Tropical Medicine https://doi.org/10.17037/DATA.00002797 (2022).
229. Davidson, J., Warren-Gash, C. & Mcdonald, H. Therapy codelist - CPRD Aurum - Diabetes. London School of Hygiene & Tropical Medicine https://doi.org/10.17037/DATA.00002224 (2021).
230. Wang, X. *et al*. Systematic approach to outcome assessment from coded electronic healthcare records in the DaRe2THINK NHS-embedded randomized trial. *Eur. Heart J. - Digit. Health* 3, 426–436 (2022).
231. ICD-10 CODE. https://www.datadictionary.nhs.uk/data_elements/icd-10_code.html.
232. Enders, C. K. *Applied Missing Data Analysis*. xv, 377 (Guilford Press, New York, NY, US, 2010).
233. Carroll, O. U., Morris, T. P. & Keogh, R. H. How are missing data in covariates handled in observational time-to-event studies in oncology? A systematic review. *BMC Med. Res. Methodol.* 20, 134 (2020).
234. Sterne, J. A. C. *et al*. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 338, b2393 (2009).

235. Donders, A. R. T., van der Heijden, G. J. M. G., Stijnen, T. & Moons, K. G. M. Review: a gentle introduction to imputation of missing values. *J. Clin. Epidemiol.* 59, 1087–1091 (2006).
236. White, I. R., Royston, P. & Wood, A. M. Multiple imputation using chained equations: Issues and guidance for practice. *Stat. Med.* 30, 377–399 (2011).
237. Bennett, D. A. How can I deal with missing data in my study? *Aust. N. Z. J. Public Health* 25, 464–469 (2001).
238. Therneau, T., Crowson, C. & Atkinson, E. Using Time Dependent Covariates and Time Dependent Coefficients in the Cox Model.
239. Standards of Medical Care in Diabetes—2023. - Penelusuran Google. https://diabetes.org/newsroom/american-diabetes-association-2023-standards-care-diabetes-guide-for-prevention-diagnosis-treatment-people-living-with-diabetes.
240. von Elm, E. *et al.* Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *BMJ* 335, 806–808 (2007).
241. Barapatre, D. & Vijayalakshmi, A. Data preparation on large datasets for data science. *Asian J. Pharm. Clin. Res.* 10, 485–488 (2017).
242. Zečević, K. *et al.* Exploring factors that influence the practice of Open Science by early career health researchers: a mixed methods study. Preprint at https://doi.org/10.12688/hrbopenres.13119.2 (2021).
243. Payne, P., Lele, O., Johnson, B. & Holve, E. Enabling Open Science for Health Research: Collaborative Informatics Environment for Learning on Health Outcomes (CIELO). *J. Med. Internet Res.* 19, e6937 (2017).
244. Abdelaziz, A. I., Hanson, K. A., Gaber, C. E. & Lee, T. A. Optimizing large real-world data analysis with parquet files in R: A step-by-step tutorial. *Pharmacoepidemiol. Drug Saf.* 33, e5728 (2024).
245. Clark, T. G., Bradburn, M. J., Love, S. B. & Altman, D. G. Survival Analysis Part I: Basic concepts and first analyses. *Br. J. Cancer* 89, 232–238 (2003).
246. Emmert-Streib, F. & Dehmer, M. Introduction to Survival Analysis in Practice. *Mach. Learn. Knowl. Extr.* 1, 1013–1038 (2019).
247. Harrell, F. E. Introduction to Survival Analysis. in *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis* (ed. Harrell, Jr., Frank E.) 399–422

(Springer International Publishing, Cham, 2015). doi:10.1007/978-3-319-19425-7_17.

248. Stevenson, M. An Introduction to Survival Analysis. 2007. http://www.biecek.pl/statystykaMedyczna/Stevenson_survival_analysis_195.721.pdf
249. Sainani, K. L. Introduction to Survival Analysis. *PM&R* 8, 580–585 (2016).
250. Shreffler, J. & Huecker, M. R. Survival Analysis. in *StatPearls* (StatPearls Publishing, Treasure Island (FL), 2025).
251. Wang, X. Introduction to survival analysis. *Medium* https://xiongfeng.medium.com/introduction-to-survival-analysis-9ed63d1fa6d9 (2020).
252. Harrell, F. E. Cox Proportional Hazards Regression Model. in *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis* (ed. Harrell, F. E.) 465–507 (Springer, New York, NY, 2001). doi:10.1007/978-1-4757-3462-1_19.
253. Dey, T., Mukherjee, A. & Chakraborty, S. A Practical Overview and Reporting Strategies for Statistical Analysis of Survival Studies. *CHEST* 158, S39–S48 (2020).
254. Meira-Machado, L., de Uña-Álvarez, J., Cadarso-Suárez, C. & Andersen, P. K. Multi-state models for the analysis of time-to-event data. *Stat. Methods Med. Res.* 18, 195–222 (2009).
255. Andersen, P. K. & Ravn, H. *Models for Multi-State Survival Data: Rates, Risks, and Pseudo-Values*. (Chapman and Hall/CRC, New York, 2023). doi:10.1201/9780429029684.
256. Hougaard, P. Multi-state models: a review. *Lifetime Data Anal.* 5, 239–264 (1999).
257. Andersen, P. K., Abildstrom, S. Z. & Rosthøj, S. Competing risks as a multi-state model. *Stat. Methods Med. Res.* 11, 203–215 (2002).
258. Llopis-Cardona, F., Armero, C. & Sanfélix-Gimeno, G. Reflection on modern methods: competing risks versus multi-state models. Preprint at https://doi.org/10.48550/arXiv.2104.03671 (2021).
259. Beyersmann, J., Schumacher, M. & Allignol, A. Multistate modelling of competing risks. in *Competing Risks and Multistate Models with R* (eds Beyersmann, J., Allignol, A. & Schumacher, M.) 41–53 (Springer, New York, NY, 2012). doi:10.1007/978-1-4614-2035-4_3.

260. Putter, H., Fiocco, M. & Geskus, R. B. Tutorial in biostatistics: competing risks and multi-state models. *Stat. Med.* 26, 2389–2430 (2007).
261. Therneau, T., Crowson, C. & Atkinson, E. Multi-state models and competing risks.
262. Varadhan, R. *et al.* Evaluating Health Outcomes in the Presence of Competing Risks: A Review of Statistical Methods and Clinical Applications. *Med. Care* 48, S96 (2010).
263. Berry, S. D., Ngo, L., Samelson, E. J. & Kiel, D. P. Competing Risk of Death: An Important Consideration in Studies of Older Adults. *J. Am. Geriatr. Soc.* 58, 783–787 (2010).
264. Wu, W. *et al.* Competitive Risk Analysis of Prognosis in Patients With Cecum Cancer: A Population-Based Study. *Cancer Control J. Moffitt Cancer Cent.* 28, 1073274821989316 (2021).
265. Cai, H. *et al.* Association of age and cause-special mortality in patients with stage I/ II colon cancer: A population-based competing risk analysis. *PLoS ONE* 15, e0240715 (2020).
266. Zhang, S., Wang, Y., Zhang, P., Ai, L. & Liu, T. Cardiovascular Outcomes in the Patients With Colorectal Cancer: A Multi-Registry-Based Cohort Study of 197,699 Cases in the Real World. *Front. Cardiovasc. Med.* 9, (2022).
267. Putter, H., Schumacher, M. & van Houwelingen, H. C. On the relation between the cause-specific hazard and the subdistribution rate for competing risks data: The Fine–Gray model revisited. *Biom. J.* 62, 790–807 (2020).
268. Basak, R., Mistry, H. & Chen, R. C. Understanding Competing Risks. *Int. J. Radiat. Oncol. Biol. Phys.* 110, 636–640 (2021).
269. Dudley,1, W. N., Wickham,2, R. & COOMBS,3, N. An Introduction to Survival Statistics: Kaplan-Meier Analysis. *J. Adv. Pract. Oncol.* 7, 91–100 (2016).
270. Kaplan, E. L. & and Meier, P. Nonparametric Estimation from Incomplete Observations. *J. Am. Stat. Assoc.* 53, 457–481 (1958).
271. Stalpers, L. J. A. & and Kaplan, E. L. Edward L. Kaplan and the Kaplan-Meier Survival Curve. *BSHM Bull. J. Br. Soc. Hist. Math.* 33, 109–135 (2018).
272. Rich, J. T. *et al.* A Practical Guide to Understanding Kaplan-Meier Curves.*Otolaryngol.--Head Neck Surg. Off. J. Am. Acad. Otolaryngol.-Head Neck Surg.* 143, 331–336 (2010).

273. Hobbs, B. P. On nonparametric hazard estimation. *J. Biom. Biostat.* 6, 232 (2015).
274. Applied Survival Analysis | Wiley Series in Probability and Statistics. https://onlinelibrary.wiley.com/doi/book/10.1002/9780470258019.
275. Rai, S., Mishra, P. & Ghoshal, U. C. Survival analysis: A primer for the clinician scientists. *Indian J. Gastroenterol.* 40, 541–549 (2021).
276. Karadeniz, P. G. & Ercan, I. Examining Tests for Comparing Survival Curves with Right Censored Data. *Stat. Transit. New Ser.* 18, 311–328 (2017).
277. Emmert-Streib, F. & Dehmer, M. Introduction to Survival Analysis in Practice. *Mach. Learn. Knowl. Extr.* 1, 1013–1038 (2019).
278. Hazra, A. & Gogtay, N. Biostatistics Series Module 9: Survival Analysis. *Indian J. Dermatol.* 62, 251–257 (2017).
279. Kaneko, Y. & Morita, S. Determining the late effect parameter in the Fleming-Harrington test using asymptotic relative efficiency in cancer immunotherapy clinical trials. *J. Biopharm. Stat.* 34, 626–645 (2024).
280. McGill University. https://www.medicine.mcgill.ca/epidemiology/hanley/c609/material/NelsonAalenEstimator.pdf
281. Cox, David R. "Regression models and life-tables." Journal of the Royal Statistical Society: Series B (Methodological) 34, no. 2 (1972): 187-202.
282. Kalbfleisch, J. D. & Schaubel, D. E. Fifty Years of the Cox Model. *Annu. Rev. Stat. Its Appl.* 10, 1–23 (2023).
283. Pourhoseingholi, M. A. *et al.* Comparing Cox Regression and Parametric Models for Survival of Patients with Gastric Carcinoma.
284. Balan, T. A. & Putter, H. A tutorial on frailty models. *Stat. Methods Med. Res.* 29, 3424–3454 (2020).
285. Roshani, D. & Ghaderi, E. Comparing Smoothing Techniques for Fitting the Nonlinear Effect of Covariate in Cox Models. *Acta Inform. Medica* 24, 38–41 (2016).
286. Jin, M. Imputation methods for informative censoring in survival analysis with time dependent covariates. *Contemp. Clin. Trials* 136, 107401 (2024).
287. Ranganathan, P. & Pramesh, C. S. Censoring in survival analysis: Potential for bias. *Perspect. Clin. Res.* 3, 40 (2012).
288. Fine, J. P. & and Gray, R. J. A Proportional Hazards Model for the Subdistribution of a Competing Risk. *J. Am. Stat. Assoc.* 94, 496–509 (1999).

289. Jeon, Y. & Lee, W. K. Competing Risk Model in Survival Analysis. *Cardiovasc. Prev. Pharmacother.* 2, 77–84 (2020).
290. Lau, B., Cole, S. R. & Gange, S. J. Competing Risk Regression Models for Epidemiologic Data. *Am. J. Epidemiol.* 170, 244–256 (2009).
291. Putter, Hein. "Tutorial in biostatistics: Competing risks and multi-state models Analyses using the mstate package." Companion file for the mstate package 145 (2011).
292. Pourhoseingholi et al. - Comparing Cox Regression and Parametric Models for.pdf.
293. In, J. & Lee, D. K. Survival analysis: part II – applied clinical data analysis. *Korean J. Anesthesiol.* 72, 441–457 (2019).
294. Dessai, S. & Patil, V. Testing and interpreting assumptions of COX regression analysis. *Cancer Res. Stat. Treat.* 2, 108 (2019).
295. Grambsch, P. M. & Therneau, T. M. Proportional Hazards Tests and Diagnostics Based on Weighted Residuals. *Biometrika* 81, 515–526 (1994).
296. Kueon, J. A Comparison of Schoenfeld and Generalized Residuals in Survival Analysis. https://dr.lib.iastate.edu/handle/20.500.12876/105455 (2022) doi:10.31274/cc-20240624-26.
297. Therneau, T. M., Grambsch, P. M. & Fleming, T. R. Martingale-Based Residuals for Survival Models. *Biometrika* 77, 147–160 (1990).
298. Cox, D. R. & Snell, E. J. A General Definition of Residuals. *J. R. Stat. Soc. Ser. B Methodol.* 30, 248–275 (1968).
299. Collett., Model Checking in the cox Regression Model. https://jhanley.biostat.mcgill.ca/c681/cox/Collett_ch_4_a.pdf
300. Kausar, T., Akbar ,Atif & and Qasim, M. Influence diagnostics for the Cox proportional hazards regression model: method, simulation and applications. *J. Stat. Comput. Simul.* 93, 1580–1600 (2023).
301. Therneau, T. M. survival: Survival Analysis. 3.8-3 https://doi.org/10.32614/CRAN.package.survival (2001).
302. Wei, L. J. The accelerated failure time model: A useful alternative to the cox regression model in survival analysis. *Stat. Med.* 11, 1871–1879 (1992).
303. Pang, M., Platt, R. W., Schuster, T. & Abrahamowicz, M. Flexible extension of the accelerated failure time model to account for nonlinear and time-dependent effects of covariates on the hazard. *Stat. Methods Med. Res.* 30, 2526–2542 (2021).

304. Crowther, M. J. & Lambert, P. C. A general framework for parametric survival analysis. *Stat. Med.* 33, 5280–5297 (2014).
305. Ishak, K. J., Kreif, N., Benedict, A. & Muszbek, N. Overview of Parametric Survival Analysis for Health-Economic Applications. *PharmacoEconomics* 31, 663–675 (2013).
306. Royston, P. & Parmar, M. K. B. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Stat. Med.* 21, 2175–2197 (2002).
307. Ng, R., Kornas, K., Sutradhar, R., Wodchis, W. P. & Rosella, L. C. The current application of the Royston-Parmar model for prognostic modeling in health research: a scoping review. *Diagn. Progn. Res.* 2, 4 (2018).
308. Rutherford, M. J., Crowther ,Michael J. & and Lambert, P. C. The use of restricted cubic splines to approximate complex hazard functions in the analysis of time-to-event data: a simulation study. *J. Stat. Comput. Simul.* 85, 777–793 (2015).
309. Royston, P. & Lambert, P. C. *Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model*. (StataCorp LP, College Station, Texas, 2011).
310. Flexible Parametric Alternatives to the Cox Model, and more - Patrick Royston, 2001. https://journals.sagepub.com/doi/abs/10.1177/1536867X0100100101.
311. Lambert, P. C. & Royston, P. Further Development of Flexible Parametric Models for Survival Analysis. *Stata J.* 9, 265–290 (2009).
312. Du, X. *et al*. Comparison of the flexible parametric survival model and Cox model in estimating Markov transition probabilities using real-world data. *PLOS ONE* 13, e0200807 (2018).
313. Nemes, S. Parametric analysis and model selection for economic evaluation of survival data. *Model Assist. Stat. Appl.* 19, 123–131 (2024).
314. Hettle, R., Sackeyfio, A. & Suri, G. Achieving Optimal Fit-To-Survival Data Using Cubic Spline Models: A Case-Study Comparing The Fit of Standard and Cubic Spline Models to Survival Data from the Icon6 Study. *Value Health* 19, A347 (2016).
315. Fisher, L. D. & Lin, D. Y. Time-dependent covariates in the Cox proportional-hazards regression model. *Annu. Rev. Public Health* 20, 145–157 (1999).

316. Zhang, Z., Reinikainen, J., Adeleke, K. A., Pieterse, M. E. & Groothuis-Oudshoorn, C. G. M. Time-varying covariates and coefficients in Cox regression models. *Ann. Transl. Med.* 6, 121 (2018).
317. Meier-Hirmer, C. & Schumacher, M. Multi-state model for studying an intermediate event using time-dependent covariates: application to breast cancer. *BMC Med. Res. Methodol.* 13, 80 (2013).
318. Wong, J., Taljaard, M., Forster, A. J., Escobar, G. J. & van Walraven, C. Addition of time-dependent covariates to a survival model significantly improved predictions for daily risk of hospital death. *J. Eval. Clin. Pract.* 19, 351–357 (2013).
319. Fosen, J., Ferkingstad, E., Borgan, Ø. & Aalen, O. O. Dynamic path analysis—a new approach to analyzing time-dependent covariates. *Lifetime Data Anal.* 12, 143–167 (2006).
320. Barnett, A. & Graves, N. Competing risks models and time-dependent covariates. *Crit. Care* 12, (2008).
321. Ng, R., Kornas, K., Sutradhar, R., Wodchis, W. P. & Rosella, L. C. The current application of the Royston-Parmar model for prognostic modeling in health research: a scoping review. *Diagn. Progn. Res.* 2, 4 (2018).
322. Asanjarani, A., Liquet, B. & Nazarathy, Y. Estimation of semi-Markov multi-state models: a comparison of the sojourn times and transition intensities approaches. *Int. J. Biostat.* 18, 243–262 (2022).
323. Chen, E. Y.-T., Dickman, P. W. & Clements, M. S. A Multistate Model Incorporating Relative Survival Extrapolation and Mixed Time Scales for Health Technology Assessment. *PharmacoEconomics* 43, 297–310 (2025).
324. Williams, C., Lewsey, J. D., Briggs, A. H. & Mackay, D. F. Cost-effectiveness Analysis in R Using a Multi-state Modeling Survival Analysis Framework: A Tutorial. *Med. Decis. Making* 37, 340–352 (2017).
325. van Rosmalen, J., Toy, M. & O'Mahony, J. F. A Mathematical Approach for Evaluating Markov Models in Continuous Time without Discrete-Event Simulation. *Med. Decis. Making* 33, 767–779 (2013).
326. Wan, L., Lou, W., Abner, E. & Kryscio, R. J. A comparison of time-homogeneous Markov chain and Markov process multi-state models. *Commun. Stat. Case Stud. Data Anal. Appl.* 2, 92–100 (2016).

327. Abner, E. L., Charnigo, R. J., & Kryscio, R. J. Markov chains and semi-Markov models in time-to-event analysis. Journal of biometrics & biostatistics, (e001), 19522 (2013).
328. Titman, A. C. & Putter, H. General tests of the Markov property in multi-state models. *Biostatistics* 23, 380–396 (2022).
329. Markov and semi-Markov multi-state models. https://hesim-dev.github.io/hesim/articles/mstate.html.
330. Asanjarani, A., Liquet, B. & Nazarathy, Y. Estimation of semi-Markov multi-state models: a comparison of the sojourn times and transition intensities approaches. *Int. J. Biostat.* 18, 243–262 (2021).
331. Król, A. & Saint-Pierre, P. SemiMarkov: An R Package for Parametric Estimation in Multi-State Semi-Markov Models. *J. Stat. Softw.* 66, 1–16 (2015).
332. Gao, X. & Yuan, S. High density lipoproteins-based therapies for cardiovascular disease. *J. Cardiovasc. Dis. Res.* 1, 99–103 (2010).
333. Mooradian, A. D. Dyslipidemia in type 2 diabetes mellitus. *Nat. Rev. Endocrinol.* 5, 150–159 (2009).
334. Femlak, M., Gluba-Brzózka, A., Ciałkowska-Rysz, A. & Rysz, J. The role and function of HDL in patients with diabetes mellitus and the related cardiovascular risk. *Lipids Health Dis.* 16, 207 (2017).
335. Associations of HDL-C/LDL-C with myocardial infarction, all-cause mortality, haemorrhagic stroke and ischaemic stroke: a longitudinal study based on 384 093 participants from the UK Biobank | Stroke and Vascular Neurology. https://svn.bmj.com/content/8/2/119.
336. Kos, K. Cardiometabolic Morbidity and Mortality with Smoking Cessation, Review of Recommendations for People with Diabetes and Obesity. *Curr. Diab. Rep.* 20, 82 (2020).
337. Kar, D. *et al.* Relationship of cardiometabolic parameters in non-smokers, current smokers, and quitters in diabetes: a systematic review and meta-analysis. *Cardiovasc. Diabetol.* 15, 158 (2016).
338. Lavie, C. J., Milani, R. V., Artham, S. M., Patel, D. A. & Ventura, H. O. The obesity paradox, weight loss, and coronary disease. *Am. J. Med.* 122, 1106–1114 (2009).
339. Lavie, C. J. *et al.* Update on Obesity and Obesity Paradox in Heart Failure. *Prog. Cardiovasc. Dis.* 58, 393–400 (2016).

340. Dixon, J. B. & Lambert, G. W. The obesity paradox – A reality that requires explanation and clinical interpretation. *Atherosclerosis* 226, 47–48 (2013).
341. Horwich, T. B., Fonarow, G. C. & Clark, A. L. Obesity and the Obesity Paradox in Heart Failure. *Prog. Cardiovasc. Dis.* 61, 151–156 (2018).
342. Head, A. *et al.* Inequalities in incident and prevalent multimorbidity in England, 2004–19: a population-based, descriptive study. *Lancet Healthy Longev.* 2, e489–e497 (2021).
343. Statistics on Health Inequalities in the United Kingdom. *Office for Statistics Regulation* https://osr.statisticsauthority.gov.uk/publication/statistics-on-health-inequalities-in-the-united-kingdom/.
344. Nagar, S. D., Jordan, I. K. & Mariño-Ramírez, L. The landscape of health disparities in the UK Biobank. *Database* 2023, baad026 (2023).
345. Sohi, I. *et al.* The Global Impact of Alcohol Consumption on Premature Mortality and Health in 2016. *Nutrients* 13, 3145 (2021).
346. Zhao, J. *et al.* Association Between Daily Alcohol Intake and Risk of All-Cause Mortality: A Systematic Review and Meta-analyses. *JAMA Netw. Open* 6, e236185 (2023).
347. Tian, Y. *et al.* Alcohol consumption and all-cause and cause-specific mortality among US adults: prospective cohort study. *BMC Med.* 21, 208 (2023).
348. Ford, I. *et al.* Long-Term Follow-up of the West of Scotland Coronary Prevention Study. *N. Engl. J. Med.* 357, 1477–1486 (2007).
349. Han, H. A Method to Explore the Best Mixed-Effects Model in a Data-Driven Manner with Multiprocessing: Applications in Public Health Research. *Eur. J. Investig. Health Psychol. Educ.* 14, 1338–1350 (2024).
350. Shah, A. S. V. *et al.* Clinical burden, risk factor impact and outcomes following myocardial infarction and stroke: A 25-year individual patient level linkage study. *Lancet Reg. Health - Eur.* 7, 100141 (2021).
351. Pana, T. A. *et al.* Myocardial infarction after acute ischaemic stroke: Incidence, mortality and risk factors. *Acta Neurol. Scand.* 140, 219–228 (2019).
352. Briggs, A., Sculpher, M. & Claxton, K. *Decision Modelling for Health Economic Evaluation*. (OUP Oxford, 2006).

353. Deaths - Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths.
354. CPRD. https://www.cprd.com/search/content?keys=cardiometabolic&op=Search. 2025.
355. UKPDS Outcomes Model : Home page. https://www2.dtu.ox.ac.uk/outcomesmodel/.
356. Cardiovascular Disease Policy Model | Epidemiology & Biostatistics. https://epibiostat.ucsf.edu/cardiovascular-disease-policy-model.
357. IMPACT Coronary Heart Disease Policy Model and Prevention Policies. University of Liverpool. https://impact.ref.ac.uk/casestudies/CaseStudy.aspx?Id=3865#:~:text=Capewell's%20policy%20model%2C%20IMPACT%2C%20has,and%20validated%2C%20particularly%20since%202008.
358. Hind, D., Tappenden, P., Peters, J. & Kenjegalieva, K. Varenicline for Smoking Cessation: A Single Technology Appraisal.
359. Ntuk, U. E. *et al.* Association between grip strength and diabetes prevalence in black, South-Asian, and white European ethnic groups: a cross-sectional analysis of 418 656 participants in the UK Biobank study. *Diabet. Med.* 34, 1120–1128 (2017).
360. Topriceanu, C.-C., Chaturvedi, N., Mathur, R. & Garfield, V. Validity of European-centric cardiometabolic polygenic scores in multi-ancestry populations. *Eur. J. Hum. Genet.* 32, 697–707 (2024).
361. Patel, S. A. *et al.* Is the "South Asian Phenotype" Unique to South Asians? Comparing Cardiometabolic Risk Factors in the CARRS and NHANES Studies. *Glob. Heart* 11, (2016).
362. Delord, M., Ashworth, M. & Douiri, A. Does Ethnicity Alters the Risk of Stroke in Patients with Modifiable Cardiometabolic Risk Factors? Preprint at https://doi.org/10.1101/2024.04.08.24305017 (2024).
363. Razieh, C., Zaccardi, F., Davies, M. J., Khunti, K. & Yates, T. Body mass index and the risk of COVID-19 across ethnic groups: Analysis of UK Biobank. *Diabetes Obes. Metab.* 22, 1953–1954 (2020).
364. Hind, D., Tappenden, P., Peters, J. & Kenjegalieva, K. Varenicline in the management of smoking cessation: a single technology appraisal. *Health Technol. Assess.* 13, 9–13 (2009).

365. Bennett, G. G. *et al*. Behavioral treatment for weight gain prevention among black women in primary care practice: a randomized clinical trial. *JAMA Intern. Med.* 173, 1770–1777 (2013).
366. Rudmik, L. & Drummond, M. Health economic evaluation: Important principles and methodology. *The Laryngoscope* 123, 1341–1347 (2013).
367. Rabarison, K. M., Bish, C. L., Massoudi, M. S. & Giles, W. H. Economic Evaluation Enhances Public Health Decision Making. *Front. Public Health* 3, (2015).
368. Sun, X. & Faunce, T. Decision-Analytical Modelling in Health-Care Economic Evaluations. *Eur. J. Health Econ.* 9, 313–323 (2008).
369. NICE health technology evaluations: the manual.(2025). https://www.nice.org.uk/process/pmg36/resources/nice-health-technology-evaluations-the-manual-pdf-72286779244741
370. Owen, L., Pennington, B., Fischer, A. & Jeong, K. The cost-effectiveness of public health interventions examined by NICE from 2011 to 2016. *J. Public Health* 40, 557–566 (2018).
371. Methods for the development of NICE public health guidance (third edition).
372. Varenicline. https://www.ncsct.co.uk/publications/category/varenicline.
373. Langley, T. E. *et al*. Prescribing of smoking cessation medication in England since the introduction of varenicline. *Addiction* 106, 1319–1324 (2011).
374. The clinical effectiveness and cost effectiveness of bupropion (Zyban) and Nicotine Replacement Therapy for smoking cessation | Guidance | NICE. https://www.nice.org.uk/guidance/ta39 (2002).
375. Wilkes, S. The use of bupropion SR in cigarette smoking cessation. *Int. J. Chron. Obstruct. Pulmon. Dis.* 3, 45–53 (2008).
376. Tonstad, S. & Lawrence, D. Varenicline in smokers with diabetes: A pooled analysis of 15 randomized, placebo-controlled studies of varenicline. *J. Diabetes Investig.* 8, 93–100 (2017).
377. Hex, N., Bartlett, C., Wright, D., Taylor, M. & Varley, D. Estimating the current and future costs of Type 1 and Type 2 diabetes in the UK, including direct health costs and indirect societal and productivity costs. *Diabet. Med. J. Br. Diabet. Assoc.* 29, 855–862 (2012).
378. National Institute for Health and Care Excellence (NICE). Evidence Review for Cessation and Harm Reduction Treatments. 2021.

https://www.nice.org.uk/guidance/ng209/evidence/k-cessation-and-harm-reduction-treatments-pdf-10890777861

379. Athanasakis, K. *et al*. Cost-Effectiveness of Varenicline Versus Bupropion, Nicotine-Replacement Therapy, and Unaided Cessation in Greece. *Clin. Ther*. 34, 1803–1814 (2012).
380. Gillett, M. *et al*. Report on Use of Evidence from Effectiveness Reviews and Cost-effectiveness Modelling.
381. Pan, A., Wang, Y., Talaei, M., Hu, F. B. & Wu, T. Relation of active, passive, and quitting smoking with incident type 2 diabetes: a systematic review and meta-analysis. *Lancet Diabetes Endocrinol*. 3, 958–967 (2015).
382. Critchley, J. A. & Capewell, S. Mortality risk reduction associated with smoking cessation in patients with coronary heart disease: a systematic review. *JAMA* 290, 86–97 (2003).
383. Shah, R. S. & Cole, J. W. Smoking and stroke: the more you smoke the more you stroke. *Expert Rev. Cardiovasc. Ther*. 8, 917–932 (2010).
384. Ara, R. & Brazier, J. E. Populating an economic model with health state utility values: moving toward better practice. *Value Health J. Int. Soc. Pharmacoeconomics Outcomes Res*. 13, 509–518 (2010).
385. Davies, N. M. *et al*. Varenicline versus nicotine replacement therapy for long-term smoking cessation: an observational study using the Clinical Practice Research Datalink. *Health Technol. Assess*. 24, 1–46 (2020).
386. Abbott, J. H. *et al*. Economic evaluation: a reader's guide to studies of cost-effectiveness. *Arch. Physiother*. 12, 28 (2022).
387. Spieker, A. J., Illenberger, N., Roy, J. A. & Mitra, N. Net benefit separation and the determination curve: a probabilistic framework for cost-effectiveness estimation. *Stat. Methods Med. Res*. 30, 1306–1319 (2021).
388. Zhou, J., Williams, C., Keng, M. J., Wu, R. & Mihaylova, B. Estimating Costs Associated with Disease Model States Using Generalized Linear Models: A Tutorial. *PharmacoEconomics* 42, 261–273 (2024).
389. Therneau - 2001 - survival Survival Analysis.pdf.
390. Mokhtar, K. I., Wan Nor Arifin, Tengku Muhammad Hanis Tengku. *Chapter 12 Parametric Survival Analysis | Data Analysis in Medicine and Health Using R*.

391. Bower, H. *et al*. Capturing simple and complex time-dependent effects using flexible parametric survival models: A simulation study. *Commun. Stat. - Simul. Comput.* 50, 3777–3793 (2021).
392. Burke, K., Jones, M. C. & Noufaily, A. A Flexible Parametric Modelling Framework for Survival Analysis. *J. R. Stat. Soc. Ser. C Appl. Stat.* 69, 429–457 (2020).
393. Quartilho, A., Gore, D. M., Bunce, C. & Tuft, S. J. Royston–Parmar flexible parametric survival model to predict the probability of keratoconus progression to corneal transplantation. *Eye* 34, 657–662 (2020).
394. Health inequalities - an update to NICE's methods for health technology evaluation. *NICE website: The National Institute for Health and Care Excellence* https://www.nice.org.uk/news/blogs/health-inequalities-an-update-to-nice-s-methods-for-health-technology-evaluation (2025).

Appendices

Appendix 1. Search strategy

EMBASE/ MEDLINE (OVID)

1 exp Diabetes Mellitus/
2 exp Diabetes Mellitus, Type 2/
3 (type* adj1 ("2" or "II" or two*) adj2 (diabete* or diabeti*)).mp.
4 (T2D or T2DM).mp.
5 exp Dyslipidemias/
6 exp Insulin Resistance/
7 exp Glucose Intolerance/
8 1 or 2 or 3 or 4 or 5 or 6 or 7
9 exp Cardiovascular Diseases/
10 (CV or CVD).mp.
11 exp Stroke/
12 exp Hypertension/
13 exp Myocardial Infarction/
14 (cardiovascular disease* or heart disease* or ischaemic heart disease* or ischemic heart disease* or angina or coronary disease* or cardiac or vascular disease* or cerebrovascular or cerebral vascular).mp.
15 9 or 10 or 11 or 12 or 13 or 14
16 exp Metabolic Diseases/
17 (metabolic adj1 (disease* or syndrom* or dysfunction* or disorder*)).mp.
18 (cardiometabolic or cardio-metabolic).mp.
19 16 or 17 or 18
20 8 or 15 or 19
21 (model* adj2 (decision* or analys* or simulat* or predict* or statistic* or mathematic* or state transition or Markov or discrete event simulation*)).mp.
22 20 and 21
23 exp Health Policy/ or health polic*.mp.
24 public health policy.mp.
25 (policy model* or health policy model* or diabetes polic* or cardiovascular polic* or cardiometabolic polic*).
26 (policy adj2 (disease* or epidemiolog*)).mp.
27 23 or 24 or 25 or 26
28 22 and 27
29 limit 28 to (english language and yr="2000 - 2022")

*Similar terms to CINAHL, Google Scholar, OpenGre

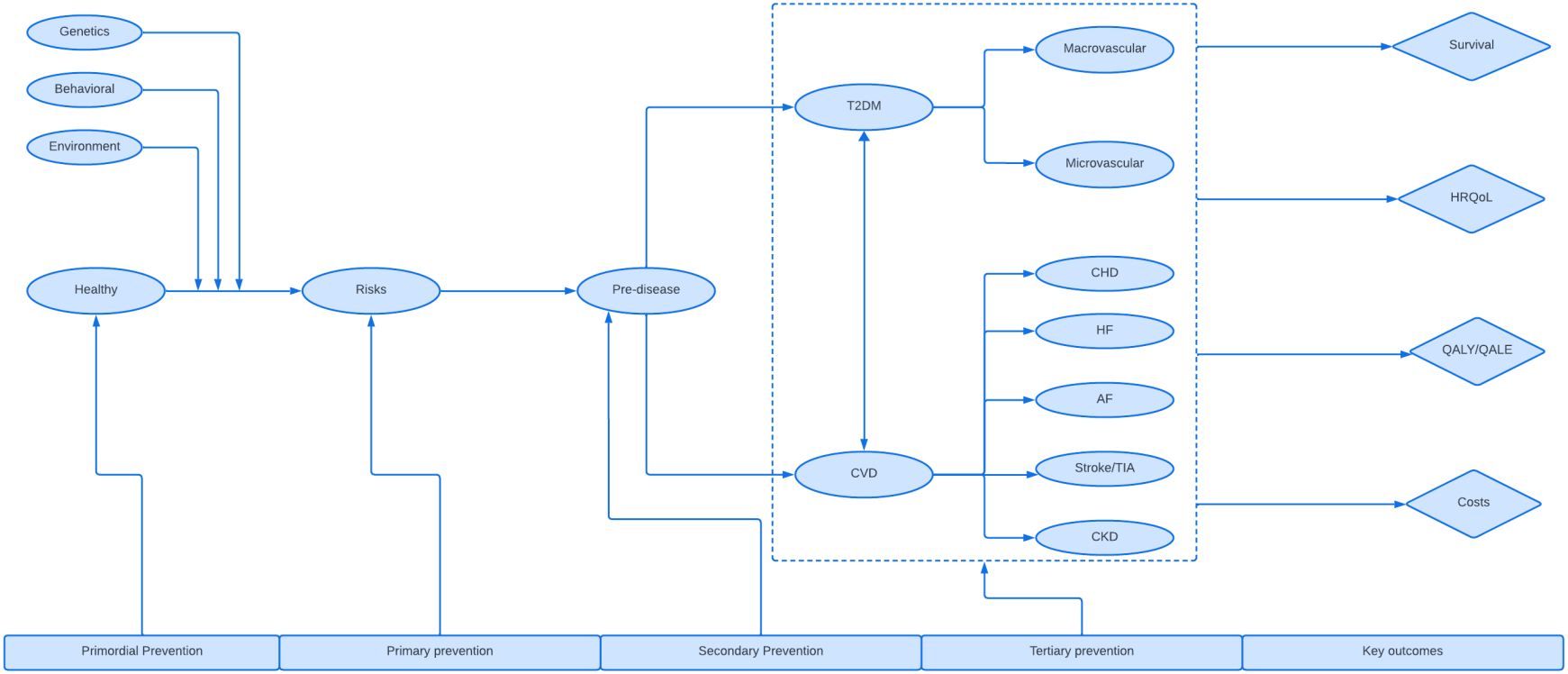
Appendix 2. Philips et al., checklist

	STRUCTURE	Yes	No	Unclear	Not related
S1	Is there a clear statement of the decision problem?				
S2	Is the objective of the evaluation and model specified and consistent with the stated decision problem?				
S3	Is the primary decision-maker specified?				
S4	Is the perspective of the model stated clearly?				
S5	Are the model inputs consistent with the stated perspective?				
S6	Has the scope of the model been stated and justified?				
S7	Are the outcomes of the model consistent with the perspective, scope and overall objective of the model?				
S8	Is the structure of the model consistent with a coherent theory of the health condition under evaluation?				
S9	Are the sources of data used to develop the structure of the model specified?				
S10	Are the causal relationships described by the model structure justified appropriately?				
S11	Are the structural assumptions transparent and justified?				
S12	Are the structural assumptions reasonable given the overall objective, perspective and scope of the model?				
S13	Is there a clear definition of the options under evaluation?				
S14	Have all feasible and practical options been evaluated?				
S15	Is there justification for the exclusion of feasible options?				
S16	Is the chosen model type appropriate given the decision problem and specified causal relationships within the model?				
S17	Is the time horizon of the model sufficient to reflect all important differences between options?				
S18	Are the time horizon of the model, the duration of treatment and the duration of treatment effect described and justified?				
S19	Do the disease states (state transition model) or the pathways (decision tree model) reflect the underlying biological process of the disease in question and the impact of interventions?				
S20	Is the cycle length defined and justified in terms of the natural history of disease?				

	DATA	Yes	No	Unclear	Not related
D1	Are the data identification methods transparent and appropriate given the objectives of the model?				
D2	Where choices have been made between data sources, are these justified appropriately?				
D3	Has particular attention been paid to identfying data fo the important parameters in the model?				
D4	Has the quality of the data been assessed appropriately?				
D5	Where expert opinion has been used are the methods described and justified?				
D6	Is the data modelling methodology based on justifiable statistical and epidemiological techniques?				
D7	Is the choice of baseline data described and justified?				
D8	Are transition probabilities calculated appropriately?				
D9	Has a half-cycle correction been applied to both cost and outcome?				
D10	If not, has this omission been justified?				
D11	If relative treatment effects have been derived from trial data, have they been synthesised using appropriate techniques?				
D12	Have the methods and assumptions used to extrapolate short-term results to final outcomes been documented and justified?				
D13	Have alternative assumptions been explored through sensitivity analysis?				
D14	Have assumptions regarding the continuing effect of treatment once treatment is complete been documented and justified? Have alternative assumptions been explored through sensitivity analysis?				
D15	Are the costs incorporated into the model justified?				
D16	Has the source for all costs been described?				
D17	Have discount rates been described and justified given the target decision-maker?				
D18	Are the utilities incorporated into the model appropriate?				
D19	Is the source for the utility weights referenced?				
D20	Are the methods of derivation for the utility weights justified?				
D21	Have all data incorporated into the model been described and referenced in sufficient detail?				
D22	Has the use of mutually inconsistent data been justified (i.e. are assumptions and choices appropriate)?				
D23	Is the process of data incorporation transparent?				
D24	If data have been incorporated as distributions, has the choice of distributions of each parameter been described and justified?				

D25	If data have been incorporated as distributions, is it clear that second order uncertainty is reflected?				
D26	Have the four principal types of uncertainty been addressed?				
D27	If not, has the omission of particular forms of uncertainty been justified?				
D28	Have methodological uncertainties been addressed by running alternative versions of the model with different methodological assumptions?				
D29	Is there evidence that structural uncertainties have been addressed via sensitivity analysis?				
D30	Has heterogenity been dealt with by running the model separately for different subgroups?				
D31	Are the methods of assessment of parameter uncertainties have been addressed via sensitivity analysis?				
D32	If data are incorporated as point estimates, are the ranges used for sensitivity analysis stated clearly and justified?				
	CONSISTENCY	**Yes**	**No**	**Unclear**	**Not related**
C1	Is there evidence that the mathematical logic of the model explained has been tested thoroughly before use?				
C2	Are any counterintuitive results from the model explained and justified?				
C3	If the model has been calibrated against independent data, have any differences been explained and justified?				
C4	Have the results of the model been compared with those of previous models and any differences in results explained?				

Appendix 3. First conceptual model



CVD: cardiovascular disease, T2DM: type 2 diabetes, CHD: chronic heart disease, HF: heart failure, AF: atrial fibrillation, TIA: transient ischemic attack, CKD: chronic kidney disease, HRQoL: health related quality of life, QALE: quality adjusted life expectancy, QALY: quality adjusted life years

Appendix 4. Data management plan (DMP)

University of Glasgow

Data Management Plan for PGR students

1. Overview	
Student name	Septiara Putri
Supervisor name	Dr. Claudia Geue, Prof. Jim Lewsey, Dr. Giorgio Ciminata
Project title	The development of cardiometabolic disease model in the UK
Funder & award number	Not applicable
Project Summary	This project's general objective is to develop and validate a cardiometabolic disease (CMD) decision analytic model using Clinical Practice Research Datalink (CPRD) data. Specifically, the state transition model of CMD would be constructed to facilitate the estimation of CMD incidence, life expectancy extrapolation, quality of life, and direct medical costs. The quantitative model is intended to aid decision making process in terms of healthcare resource allocation as well as individual healthcare context in the UK setting.

2. Data
What types of data will be collected or created?
Clinical Practice Research Datalink (CPRD) that collects routine patients' data across UK will be utilized in this study. The data encompasses 60 million patients, including 16 million currently registered patients. CPRD covers approximately 6% of UK population, the data include routine clinical practice, such as information regarding symptoms, diagnoses, prescriptions, and referrals. This large database is highly generalisable. CPRD AURUM and CPRD GOLD data. CPRD AURUM would be utilized for model development, and CPRD GOLD for model validation. General data included: - Patients' profile - Diagnosis and procedures - Observation (metabolic data) - Drug issue Linked data: - HES (Hospital Episodes Statistics) - ONS (Office for National Statistics) - IMD (Index of Multiple Deprivation) Sample will be drawn to construct the state transition model from the total data retrieved (n=2,656,165) patients.
What formats will you use?
- Data are in csv. and txt. Currently stored in Nextcloud University of Glasgow - Further data storage and transfer will be used High Performance Computing (HPC) MARS (MVLS Advanced Research) system - Initial data query and sampling will be using PostGRE SQL or SQLite - Code for data analysis will be written in R and recorded to GitHub - Visual final model result and estimation will be using RShiny
How much data will you collect?

The current data size approximately 200 GB (zipped file). It probably become higher depend on transferring process and data checking/revision following CPRD data dictionary. Hence, to anticipate this, the larger storage might be needed.

3. Documentation
How will the data be documented and described?
All CPRD data for this project will be documented under the University of Glasgow system. Code for analysis will be recorded in GitHub.
Are there any standards for this in your field of research?
CPRD has its own standard for data documentation, metadata, and linkage. (See: https://link.springer.com/article/10.1007/s10654-018-0442-4) After sampling, our documentation and descriptive standards will be depending on the project's circumstances.

4. Ethics and Intellectual Property
Who owns the data in your project?
List of the team who can use CPRD data is available at: https://cprd.com/cardiometabolic-disease-prediction-using-general-practice-consultation-pattern-use-machine-learning
Detail any ethical, legal or commercial considerations relating to your research data
Observational research undertaken using CPRD data must be for public health purposes and approved by an Independent Scientific Advisory Committee (ISAC). This study has been approved by ISAC.
How will these concerns be dealt with?
There are no legal, ethical and commercial issues as long as we follow the CPRD data use governance. - Datasets will be stored under MARS system until the end of analysis and will not be shared even though the project output is finished and published. - Patients were fully anonymised, non-identifiable, confidential - Sharing data only allowed for researchers who are stated in the projects protocol - My PhD scholarship funder has no right to see or access the data

5. Storage and organisation
How will the data be named, organised and structured?
- Since the original data were fragmented due to size reasons, we plan to combine all the data first (i.e: combine each file, building script etc.) - Two different folders for CPRD AURUM and CPRD GOLD - Folder and variable naming will be following CPRD data dictionary - Data will be structured by relational identification published by CPRD, and merged by patient id. - Each data update will be recorded with date/time to keep track the version, it would also perform in SQL and R. Folders and files will be identified by date.
How will the data be stored for the duration of the project?
- Datasets are stored in NextCloud University of Glasgow and plan to migrate them to HPC MARS system

- May maintain access in J: drive space (with supervisors) for sampled data, if needed
How will the data be backed up during the project?
- Data back-up will be under university system, only for sampled CPRD data - If possible, automatic backup services will be discussed with IT Services, so we will not rely on manual process.
Does access to the data need to be controlled for the duration of the project?
- Password protecting files in system and laptop, screen locking - Encrypted files if needed to email/cloud transfer - Sharing drive only with supervisors
Who has the right to access the data during the project?
List of the team who can use CPRD data is available at: https://cprd.com/cardiometabolic-disease-prediction-using-general-practice-consultation-pattern-use-machine-learning

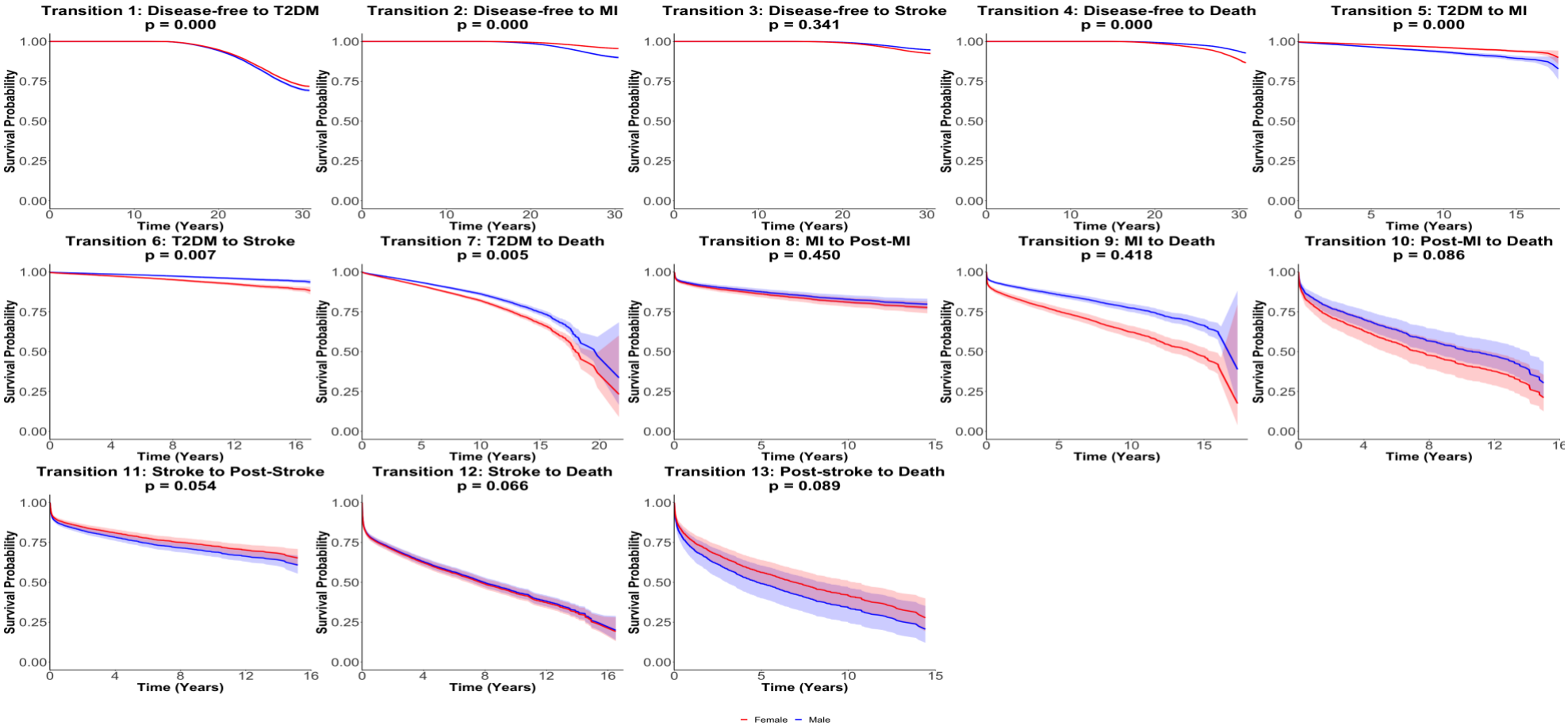
6. Deposit and long-term preservation
Which data should be retained long-term?
Since this data will be used for various aims and different researchers (outside this PhD project), it might depend on the agreement and project duration. It is also depend on the PI and supervisors' approval/agreement.
How long will data be retained for?
The University of Glasgow requires that data of long-term value should be retained for a minimum of 10 years from the end of the project. But again, it depends on the reasons above. We may consider whether is economically viable to keep, or any data governance procedures that we need to review and discuss further.
Where will the data be archived at the end of the project?
- Under repository (recommended by CPRD team) - Enlighten: research data University of Glasgow (if it is allowed)
What formats will the data be archived in?
Same format as in section 2.

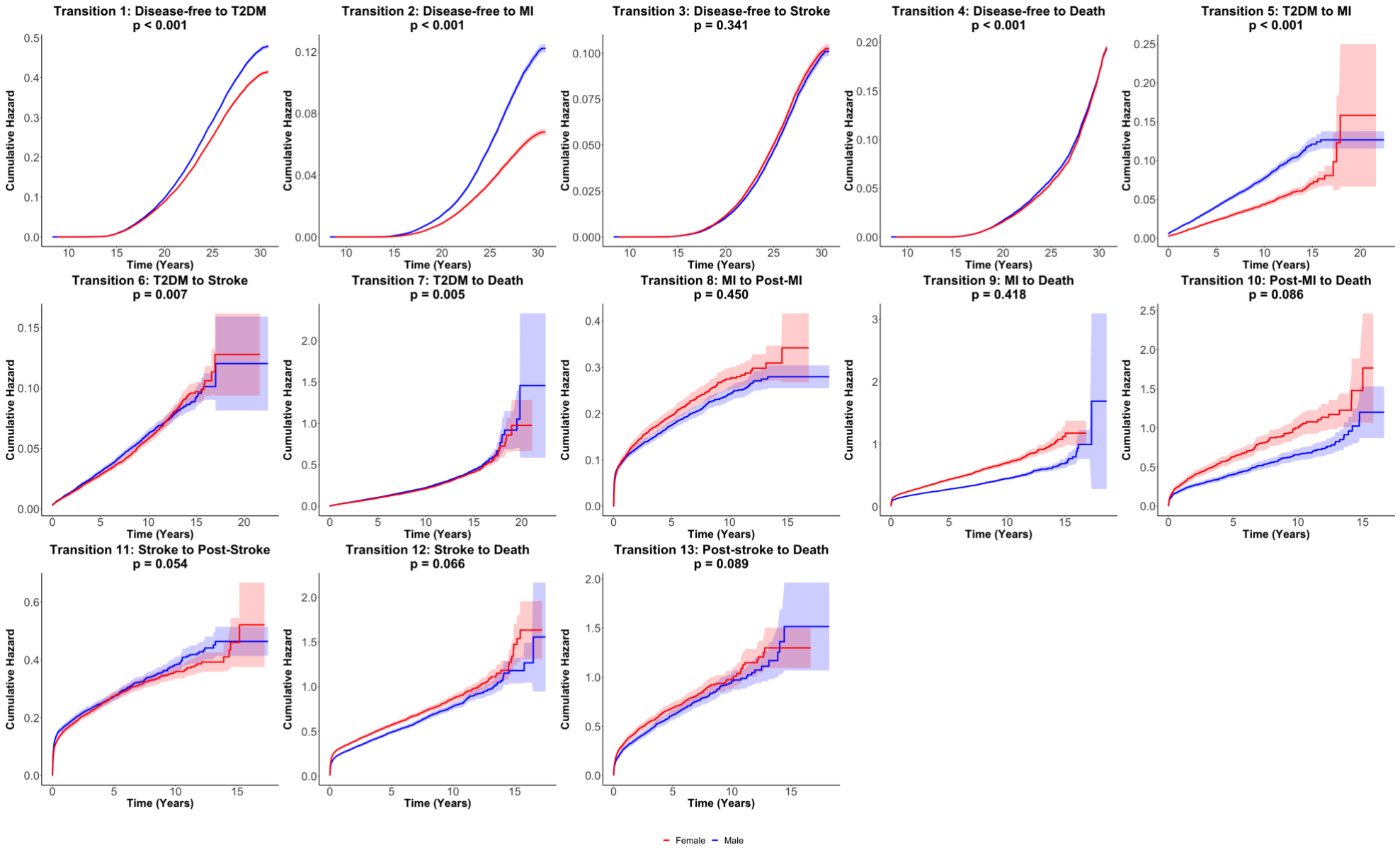
7. Data sharing
Is any of the data suitable for sharing?
No. - No archival data will be shared (i.e: for other researchers, journal supplementary material)
How will the data be shared?
N/A
Who should be able to access and use the shared data?
N/A

8. Implementation

Who is responsible for implementing this plan?
I will be responsible with data management activity, with supervision from academic supervisors as well as projects' PI.
How will this plan be kept up-to-date?
This plan will be reviewed regularly aligned with supervisory meeting or wider researchers whom using CPRD.
What actions are necessary to implement this plan?
- Contact local IT support to ensure storage provision is adequate - Ensure all data management activities are acceptable in terms of data governance and protection - Ensure data management and analysis are well reported and systematically recorded
What training or further information are needed to implement this plan?
- GDPR training - Data query training - Improve coding skill for data management - Attend trainings/workshops about managing and analysing routine data

Appendix 5. Cox model results (survival and hazard curves)

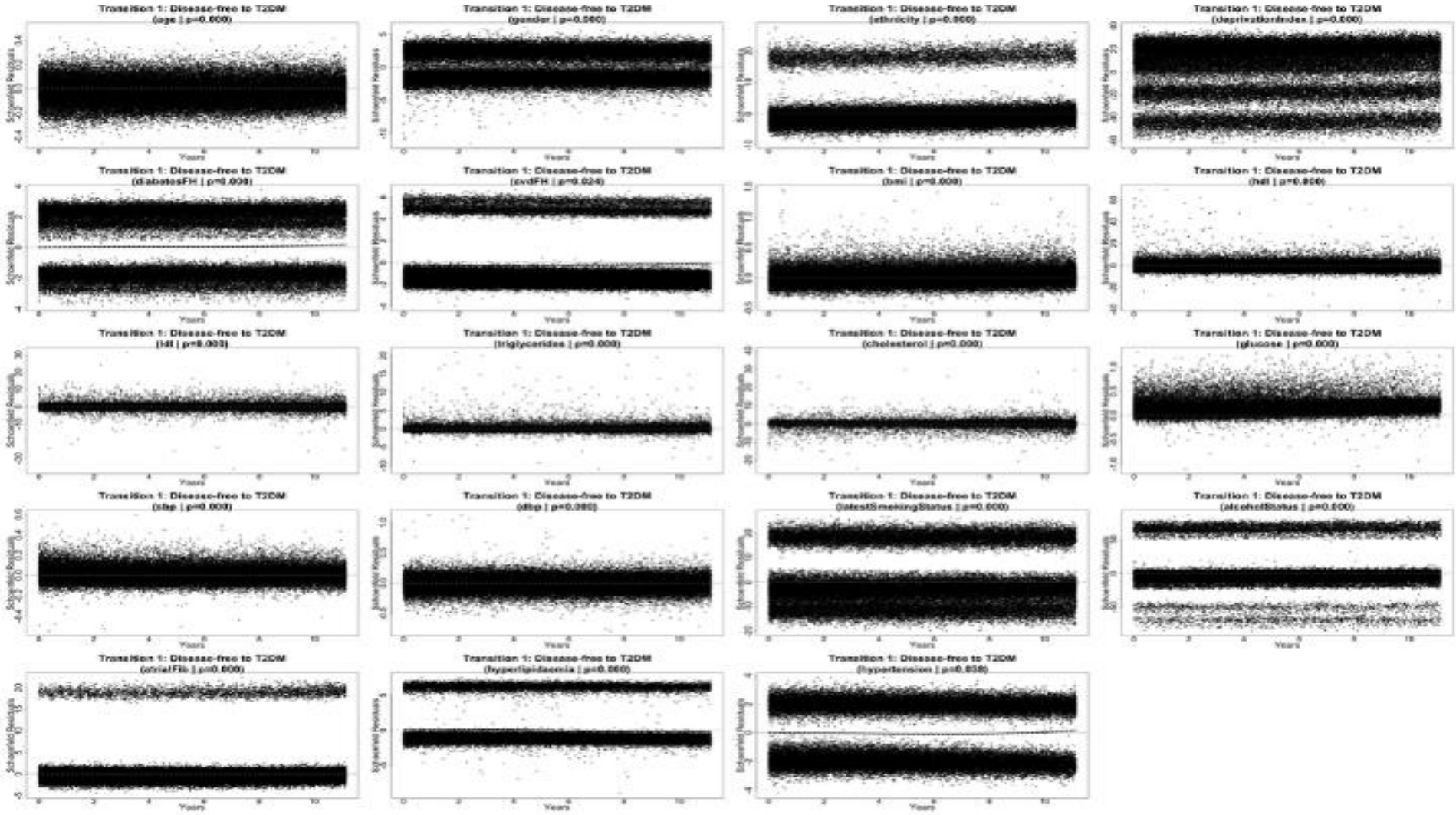


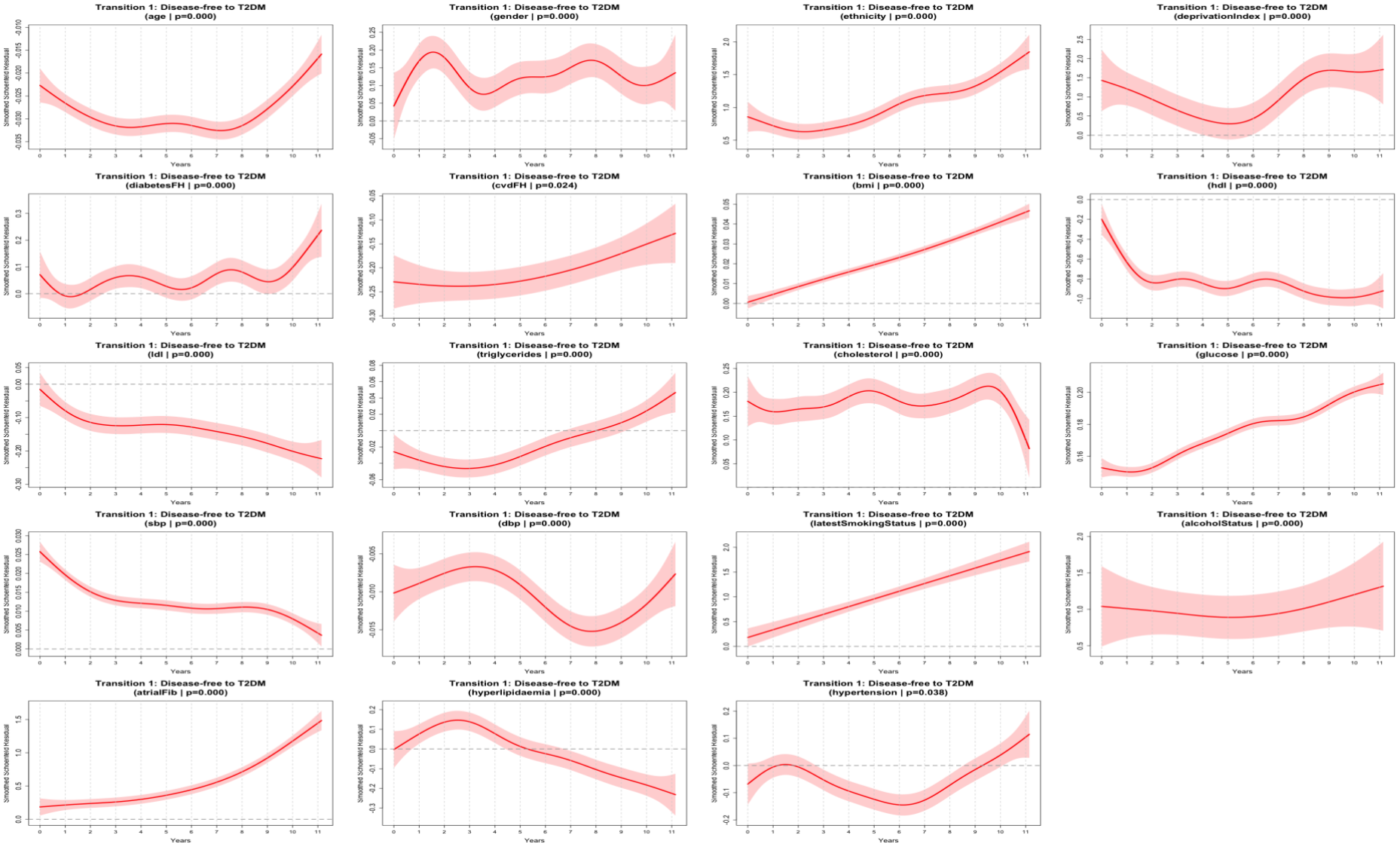

Transition 1: Disease-free to T2DM
p < 0.001
Transition 2: Disease-free to MI
p < 0.001
Transition 3: Disease-free to Stroke
p = 0.341
Transition 4: Disease-free to Death
p < 0.001
Transition 5: T2DM to MI
p < 0.001
Transition 6: T2DM to Stroke
p = 0.007
Transition 7: T2DM to Death
p = 0.005
Transition 8: MI to Post-MI
p = 0.450
Transition 9: MI to Death
p = 0.418
Transition 10: Post-MI to Death
p = 0.086
Transition 11: Stroke to Post-Stroke
p = 0.054
Transition 12: Stroke to Death
p = 0.066
Transition 13: Post-stroke to Death
p = 0.089
Cumulative Hazard
Time (Years)
Female
Male

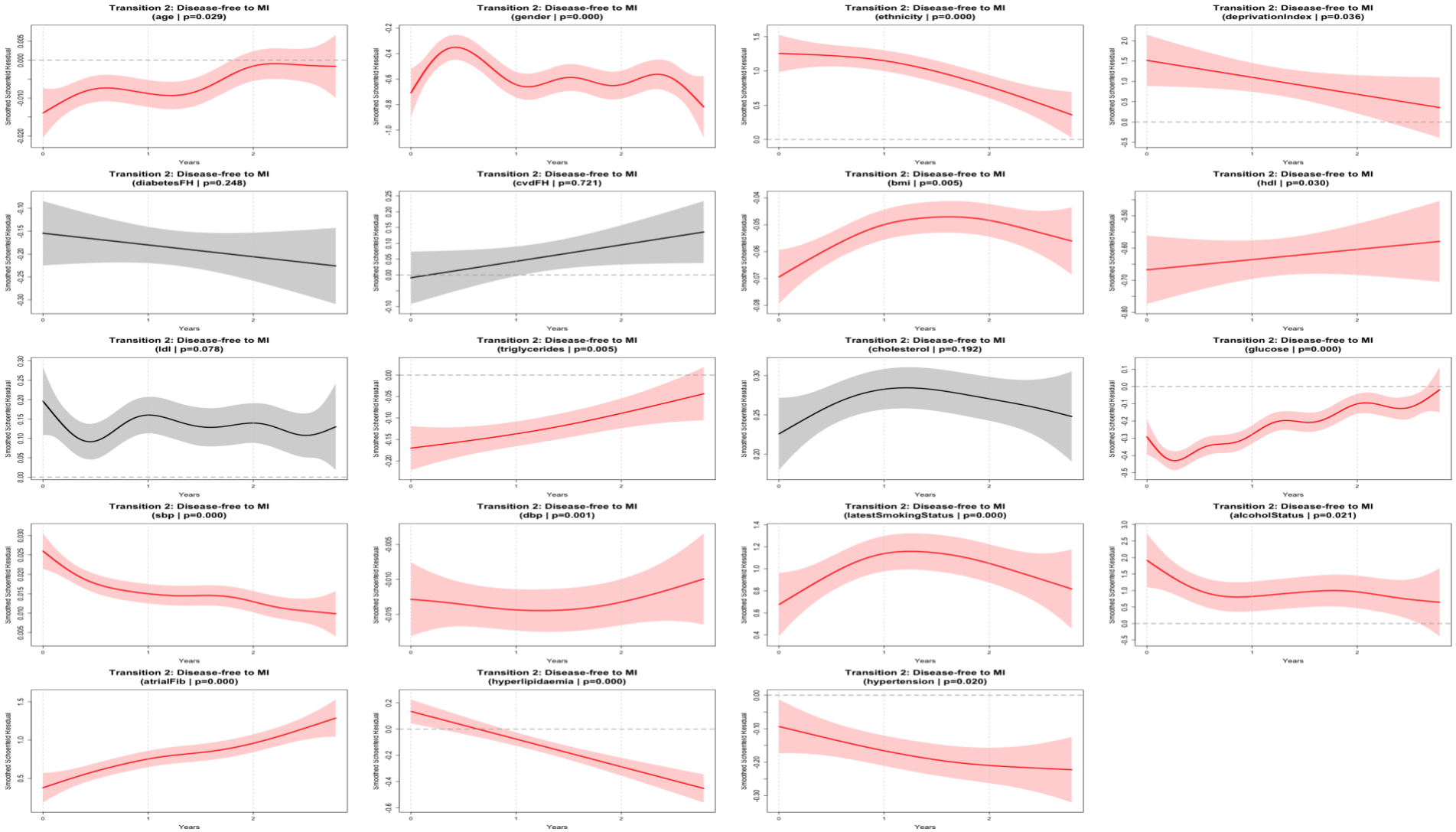
Appendix 6. Schoenfeld residuals

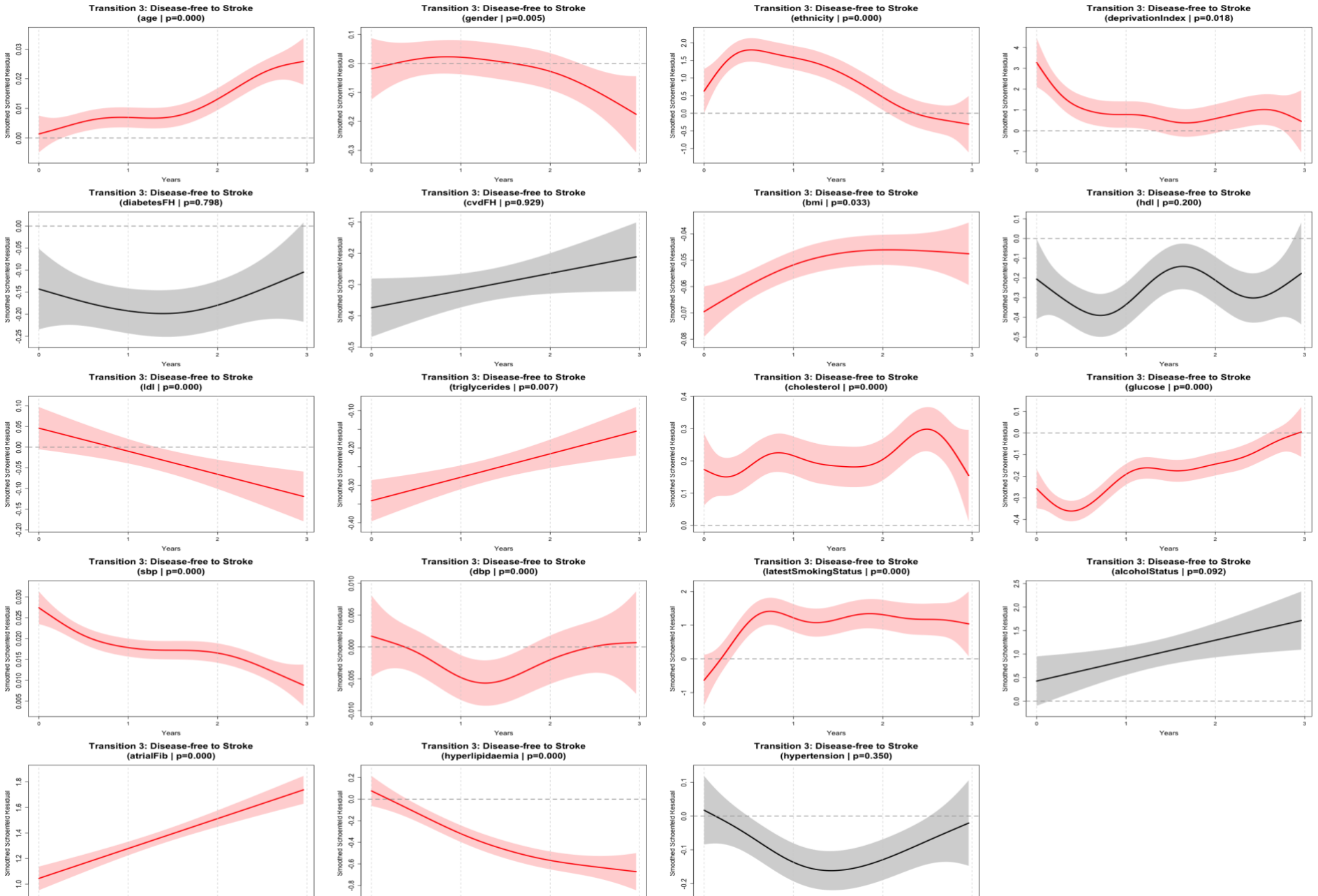
Semi-parametric model diagnostic Schoenfeld residuals

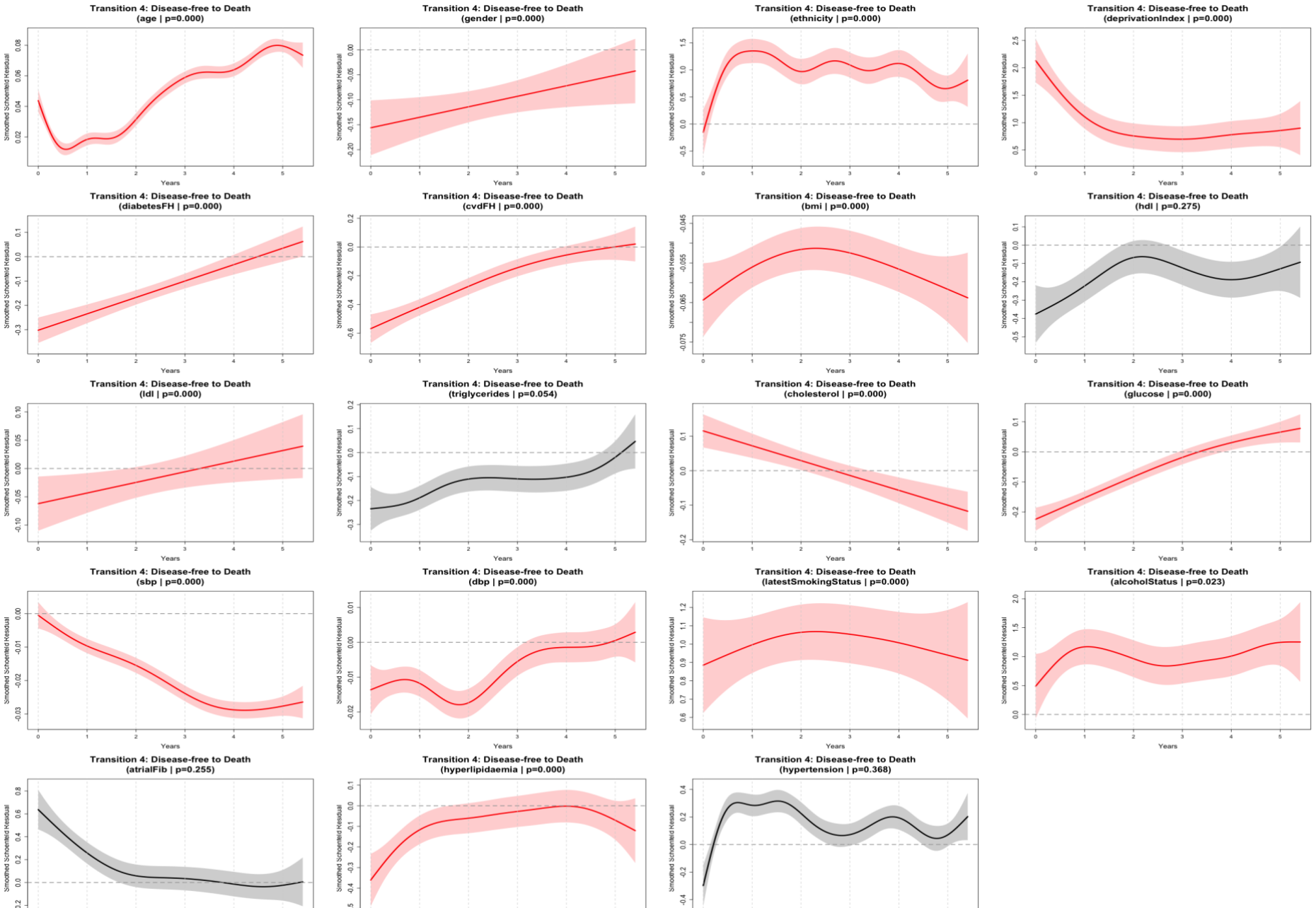
By using standard one the result of residual plot as below, which is overlapped the line due to black scattered dots. Thus, to make them clearer, the smoothed Schoenfeld residual were plotted to enhance plot clarity.

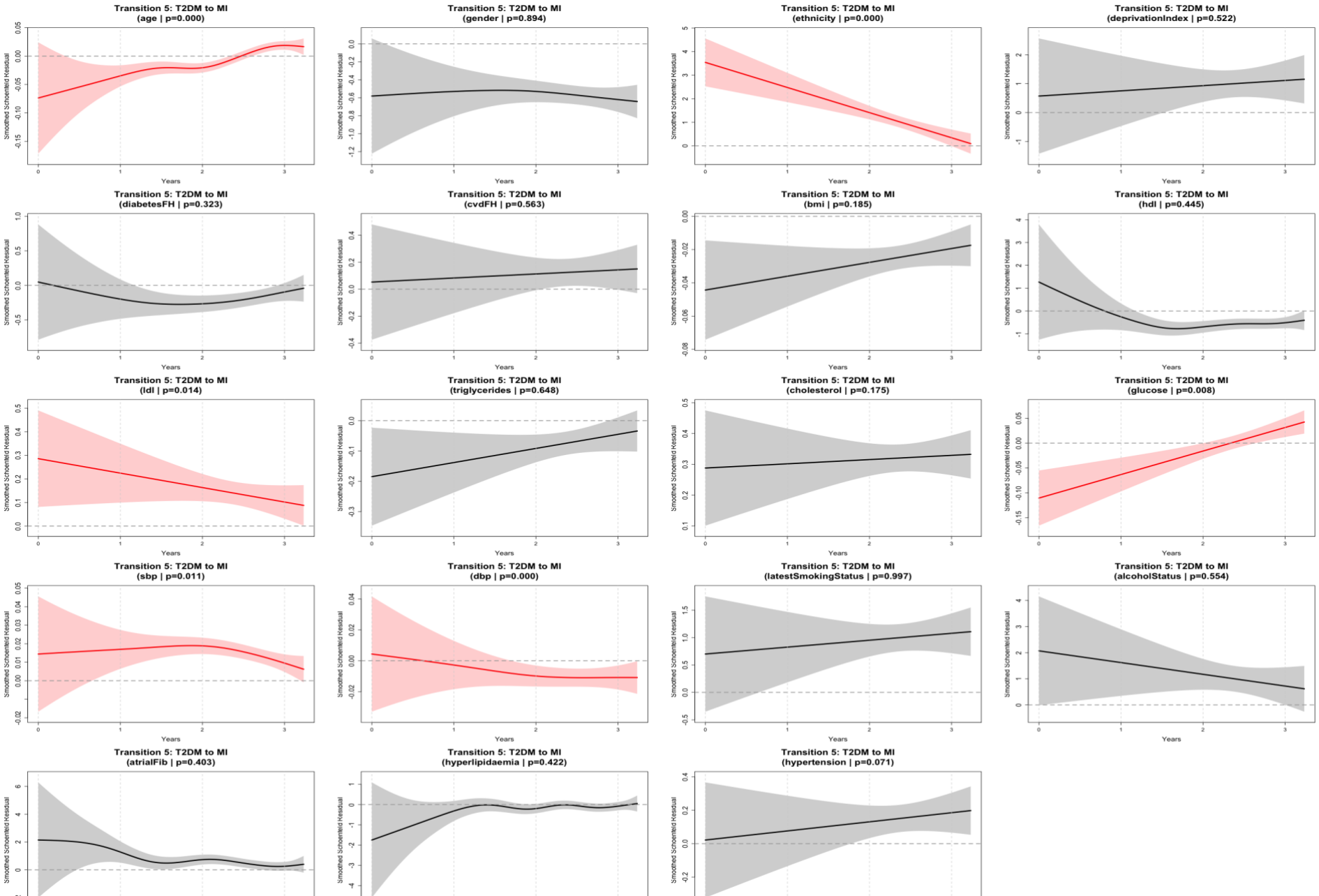

Transition 1: Disease-free to T2DM
(age | p=0.000)
Transition 1: Disease-free to T2DM
(gender | p=0.000)
Transition 1: Disease-free to T2DM
(ethnicity | p=0.000)
Transition 1: Disease-free to T2DM
(deprivationIndex | p=0.000)
Transition 1: Disease-free to T2DM
(diabetesFH | p=0.000)
Transition 1: Disease-free to T2DM
(cvdFH | p=0.024)
Transition 1: Disease-free to T2DM
(bmi | p=0.000)
Transition 1: Disease-free to T2DM
(hdl | p=0.000)
Transition 1: Disease-free to T2DM
(ldl | p=0.003)
Transition 1: Disease-free to T2DM
(triglycerides | p=0.000)
Transition 1: Disease-free to T2DM
(cholesterol | p=0.000)
Transition 1: Disease-free to T2DM
(glucose | p=0.000)
Transition 1: Disease-free to T2DM
(sbp | p=0.000)
Transition 1: Disease-free to T2DM
(dbp | p=0.000)
Transition 1: Disease-free to T2DM
(latestSmokingStatus | p=0.000)
Transition 1: Disease-free to T2DM
(alcoholStatus | p=0.000)
Transition 1: Disease-free to T2DM
(atrialFib | p=0.000)
Transition 1: Disease-free to T2DM
(hyperlipidaemia | p=0.000)
Transition 1: Disease-free to T2DM
(hypertension | p=0.038)
Years

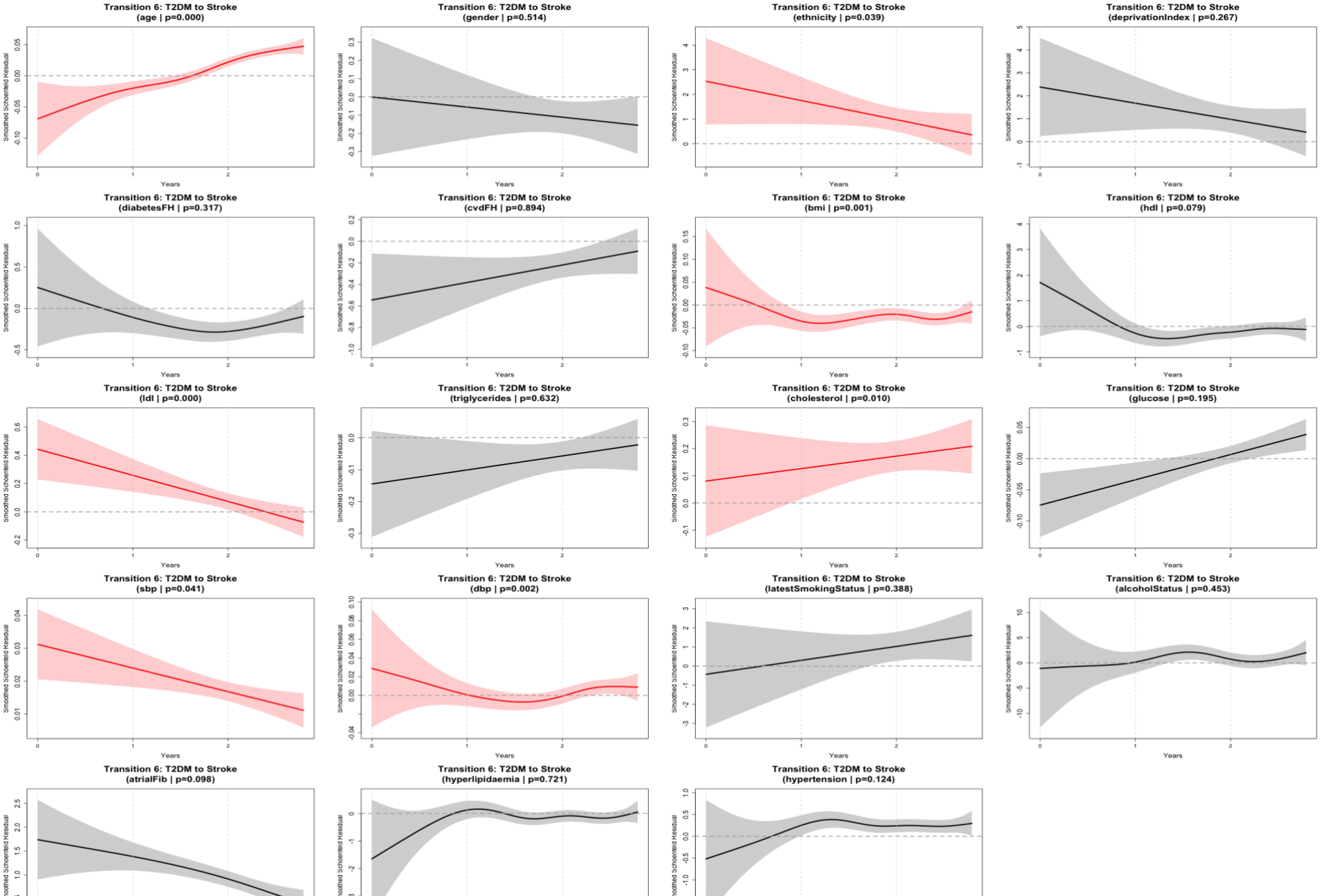

Transition 1: Disease-free to T2DM
(age | p=0.000)
Transition 1: Disease-free to T2DM
(gender | p=0.000)
Transition 1: Disease-free to T2DM
(ethnicity | p=0.000)
Transition 1: Disease-free to T2DM
(deprivationIndex | p=0.000)
Transition 1: Disease-free to T2DM
(diabetesFH | p=0.000)
Transition 1: Disease-free to T2DM
(cvdFH | p=0.024)
Transition 1: Disease-free to T2DM
(bmi | p=0.000)
Transition 1: Disease-free to T2DM
(hdl | p=0.000)
Transition 1: Disease-free to T2DM
(ldl | p=0.000)
Transition 1: Disease-free to T2DM
(triglycerides | p=0.000)
Transition 1: Disease-free to T2DM
(cholesterol | p=0.000)
Transition 1: Disease-free to T2DM
(glucose | p=0.000)
Transition 1: Disease-free to T2DM
(sbp | p=0.000)
Transition 1: Disease-free to T2DM
(dbp | p=0.000)
Transition 1: Disease-free to T2DM
(latestSmokingStatus | p=0.000)
Transition 1: Disease-free to T2DM
(alcoholStatus | p=0.000)
Transition 1: Disease-free to T2DM
(atrialFib | p=0.000)
Transition 1: Disease-free to T2DM
(hyperlipidaemia | p=0.000)
Transition 1: Disease-free to T2DM
(hypertension | p=0.038)
Smoothed Schoenfeld Residual
Years

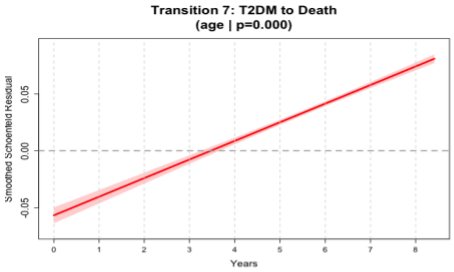

Transition 2: Disease-free to MI
(age | p=0.029)
Transition 2: Disease-free to MI
(gender | p=0.000)
Transition 2: Disease-free to MI
(ethnicity | p=0.000)
Transition 2: Disease-free to MI
(deprivationIndex | p=0.036)
Transition 2: Disease-free to MI
(diabetesFH | p=0.248)
Transition 2: Disease-free to MI
(cvdFH | p=0.721)
Transition 2: Disease-free to MI
(bmi | p=0.005)
Transition 2: Disease-free to MI
(hdl | p=0.030)
Transition 2: Disease-free to MI
(ldl | p=0.078)
Transition 2: Disease-free to MI
(triglycerides | p=0.005)
Transition 2: Disease-free to MI
(cholesterol | p=0.192)
Transition 2: Disease-free to MI
(glucose | p=0.000)
Transition 2: Disease-free to MI
(sbp | p=0.000)
Transition 2: Disease-free to MI
(dbp | p=0.001)
Transition 2: Disease-free to MI
(latestSmokingStatus | p=0.000)
Transition 2: Disease-free to MI
(alcoholStatus | p=0.021)
Transition 2: Disease-free to MI
(atrialFib | p=0.000)
Transition 2: Disease-free to MI
(hyperlipidaemia | p=0.000)
Transition 2: Disease-free to MI
(hypertension | p=0.020)
Smoothed Schoenfeld Residual
Years


Transition 3: Disease-free to Stroke
(age | p=0.000)
Transition 3: Disease-free to Stroke
(gender | p=0.005)
Transition 3: Disease-free to Stroke
(ethnicity | p=0.000)
Transition 3: Disease-free to Stroke
(deprivationIndex | p=0.018)
Transition 3: Disease-free to Stroke
(diabetesFH | p=0.798)
Transition 3: Disease-free to Stroke
(cvdFH | p=0.929)
Transition 3: Disease-free to Stroke
(bmi | p=0.033)
Transition 3: Disease-free to Stroke
(hdl | p=0.200)
Transition 3: Disease-free to Stroke
(ldl | p=0.000)
Transition 3: Disease-free to Stroke
(triglycerides | p=0.007)
Transition 3: Disease-free to Stroke
(cholesterol | p=0.000)
Transition 3: Disease-free to Stroke
(glucose | p=0.000)
Transition 3: Disease-free to Stroke
(sbp | p=0.000)
Transition 3: Disease-free to Stroke
(dbp | p=0.000)
Transition 3: Disease-free to Stroke
(latestSmokingStatus | p=0.000)
Transition 3: Disease-free to Stroke
(alcoholStatus | p=0.092)
Transition 3: Disease-free to Stroke
(atrialFib | p=0.000)
Transition 3: Disease-free to Stroke
(hyperlipidaemia | p=0.000)
Transition 3: Disease-free to Stroke
(hypertension | p=0.350)
Smoothed Schoenfeld Residual
Years

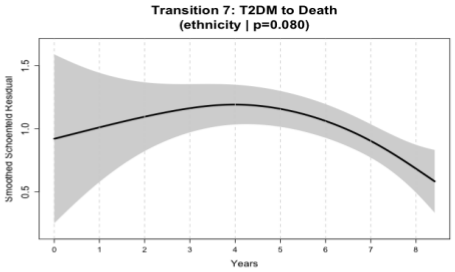

Transition 4: Disease-free to Death
(age | p=0.000)
Transition 4: Disease-free to Death
(gender | p=0.000)
Transition 4: Disease-free to Death
(ethnicity | p=0.000)
Transition 4: Disease-free to Death
(deprivationIndex | p=0.000)
Transition 4: Disease-free to Death
(diabetesFH | p=0.000)
Transition 4: Disease-free to Death
(cvdFH | p=0.000)
Transition 4: Disease-free to Death
(bmi | p=0.000)
Transition 4: Disease-free to Death
(hdl | p=0.275)
Transition 4: Disease-free to Death
(ldl | p=0.000)
Transition 4: Disease-free to Death
(triglycerides | p=0.054)
Transition 4: Disease-free to Death
(cholesterol | p=0.000)
Transition 4: Disease-free to Death
(glucose | p=0.000)
Transition 4: Disease-free to Death
(sbp | p=0.000)
Transition 4: Disease-free to Death
(dbp | p=0.000)
Transition 4: Disease-free to Death
(latestSmokingStatus | p=0.000)
Transition 4: Disease-free to Death
(alcoholStatus | p=0.023)
Transition 4: Disease-free to Death
(atrialFib | p=0.255)
Transition 4: Disease-free to Death
(hyperlipidaemia | p=0.000)
Transition 4: Disease-free to Death
(hypertension | p=0.368)
Smoothed Schoenfeld Residual
Years

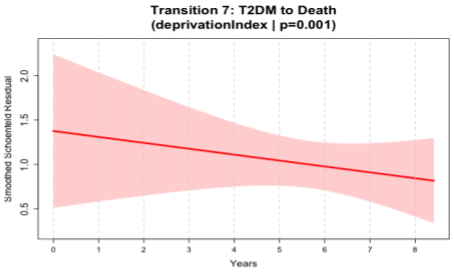

Transition 5: T2DM to MI
(age | p=0.000)
Transition 5: T2DM to MI
(gender | p=0.894)
Transition 5: T2DM to MI
(ethnicity | p=0.000)
Transition 5: T2DM to MI
(deprivationIndex | p=0.522)
Transition 5: T2DM to MI
(diabetesFH | p=0.323)
Transition 5: T2DM to MI
(cvdFH | p=0.563)
Transition 5: T2DM to MI
(bmi | p=0.185)
Transition 5: T2DM to MI
(hdl | p=0.445)
Transition 5: T2DM to MI
(ldl | p=0.014)
Transition 5: T2DM to MI
(triglycerides | p=0.648)
Transition 5: T2DM to MI
(cholesterol | p=0.175)
Transition 5: T2DM to MI
(glucose | p=0.008)
Transition 5: T2DM to MI
(sbp | p=0.011)
Transition 5: T2DM to MI
(dbp | p=0.000)
Transition 5: T2DM to MI
(latestSmokingStatus | p=0.997)
Transition 5: T2DM to MI
(alcoholStatus | p=0.554)
Transition 5: T2DM to MI
(atrialFib | p=0.403)
Transition 5: T2DM to MI
(hyperlipidaemia | p=0.422)
Transition 5: T2DM to MI
(hypertension | p=0.071)
Smoothed Schoenfeld Residual
Years

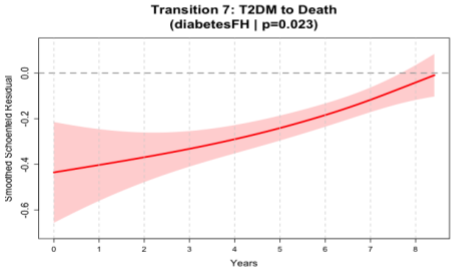

Transition 6: T2DM to Stroke
(age | p=0.000)
Transition 6: T2DM to Stroke
(gender | p=0.514)
Transition 6: T2DM to Stroke
(ethnicity | p=0.039)
Transition 6: T2DM to Stroke
(deprivationIndex | p=0.267)
Transition 6: T2DM to Stroke
(diabetesFH | p=0.317)
Transition 6: T2DM to Stroke
(cvdFH | p=0.894)
Transition 6: T2DM to Stroke
(bmi | p=0.001)
Transition 6: T2DM to Stroke
(hdl | p=0.079)
Transition 6: T2DM to Stroke
(ldl | p=0.000)
Transition 6: T2DM to Stroke
(triglycerides | p=0.632)
Transition 6: T2DM to Stroke
(cholesterol | p=0.010)
Transition 6: T2DM to Stroke
(glucose | p=0.195)
Transition 6: T2DM to Stroke
(sbp | p=0.041)
Transition 6: T2DM to Stroke
(dbp | p=0.002)
Transition 6: T2DM to Stroke
(latestSmokingStatus | p=0.388)
Transition 6: T2DM to Stroke
(alcoholStatus | p=0.453)
Transition 6: T2DM to Stroke
(atrialFib | p=0.098)
Transition 6: T2DM to Stroke
(hyperlipidaemia | p=0.721)
Transition 6: T2DM to Stroke
(hypertension | p=0.124)
Smoothed Schoenfeld Residual
Years

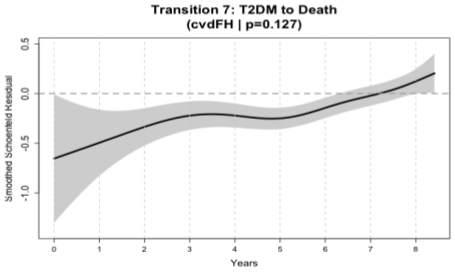

Transition 7: T2DM to Death
(age | p=0.000)
Smoothed Schoenfeld Residual
Years

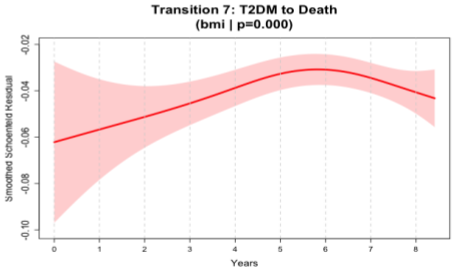

Transition 7: T2DM to Death
(gender | p=0.944)
Smoothed Schoenfeld Residual
Years

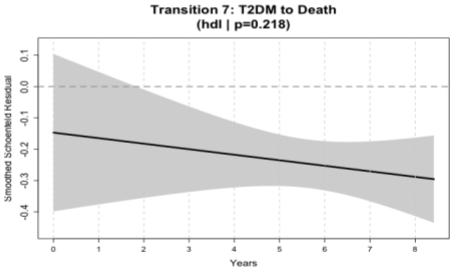

Transition 7: T2DM to Death
(ethnicity | p=0.080)
Smoothed Schoenfeld Residual
Years

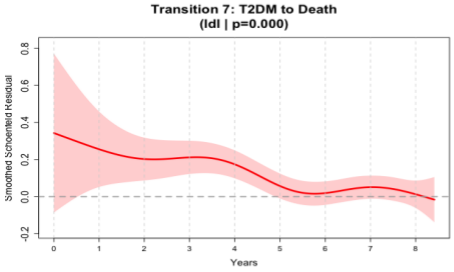

Transition 7: T2DM to Death
(deprivationIndex | p=0.001)
Smoothed Schoenfeld Residual
Years

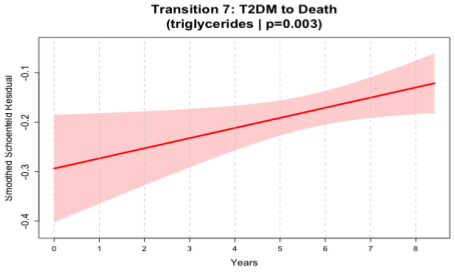

Transition 7: T2DM to Death
(diabetesFH | p=0.023)
Smoothed Schoenfeld Residual
Years

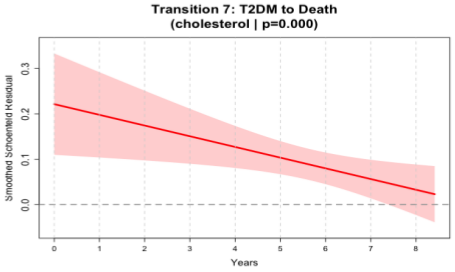

Transition 7: T2DM to Death
(cvdFH | p=0.127)
Smoothed Schoenfeld Residual
Years

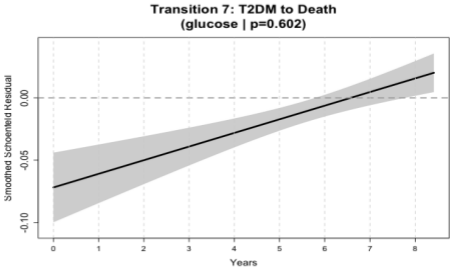

Transition 7: T2DM to Death
(bmi | p=0.000)
Smoothed Schoenfeld Residual
Years

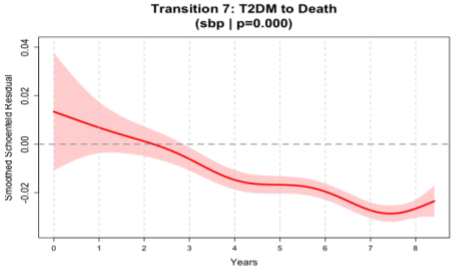

Transition 7: T2DM to Death
(hdl | p=0.218)
Smoothed Schoenfeld Residual
Years

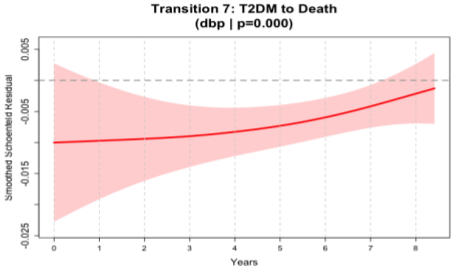

Transition 7: T2DM to Death
(ldl | p=0.000)
Smoothed Schoenfeld Residual
Years

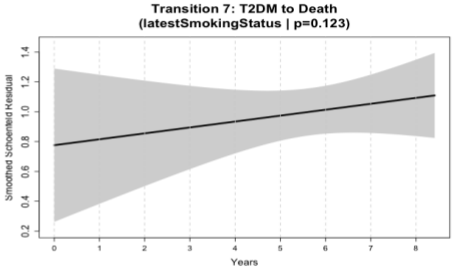

Transition 7: T2DM to Death
(triglycerides | p=0.003)
Smoothed Schoenfeld Residual
Years

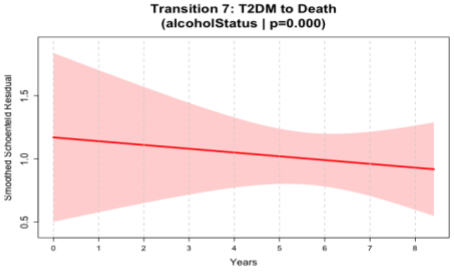

Transition 7: T2DM to Death
(cholesterol | p=0.000)
Smoothed Schoenfeld Residual
Years

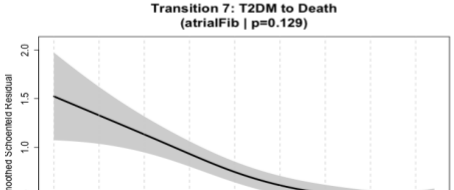

Transition 7: T2DM to Death
(glucose | p=0.602)
Smoothed Schoenfeld Residual
Years

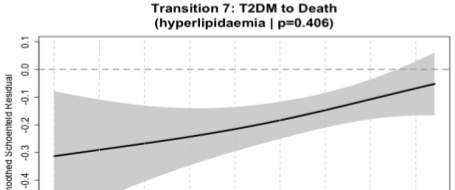

Transition 7: T2DM to Death
(sbp | p=0.000)
Smoothed Schoenfeld Residual
Years

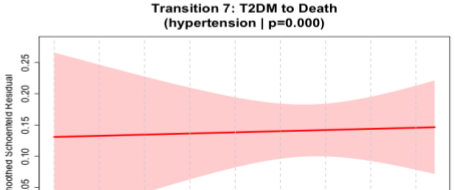

Transition 7: T2DM to Death
(dbp | p=0.000)
Smoothed Schoenfeld Residual
Years

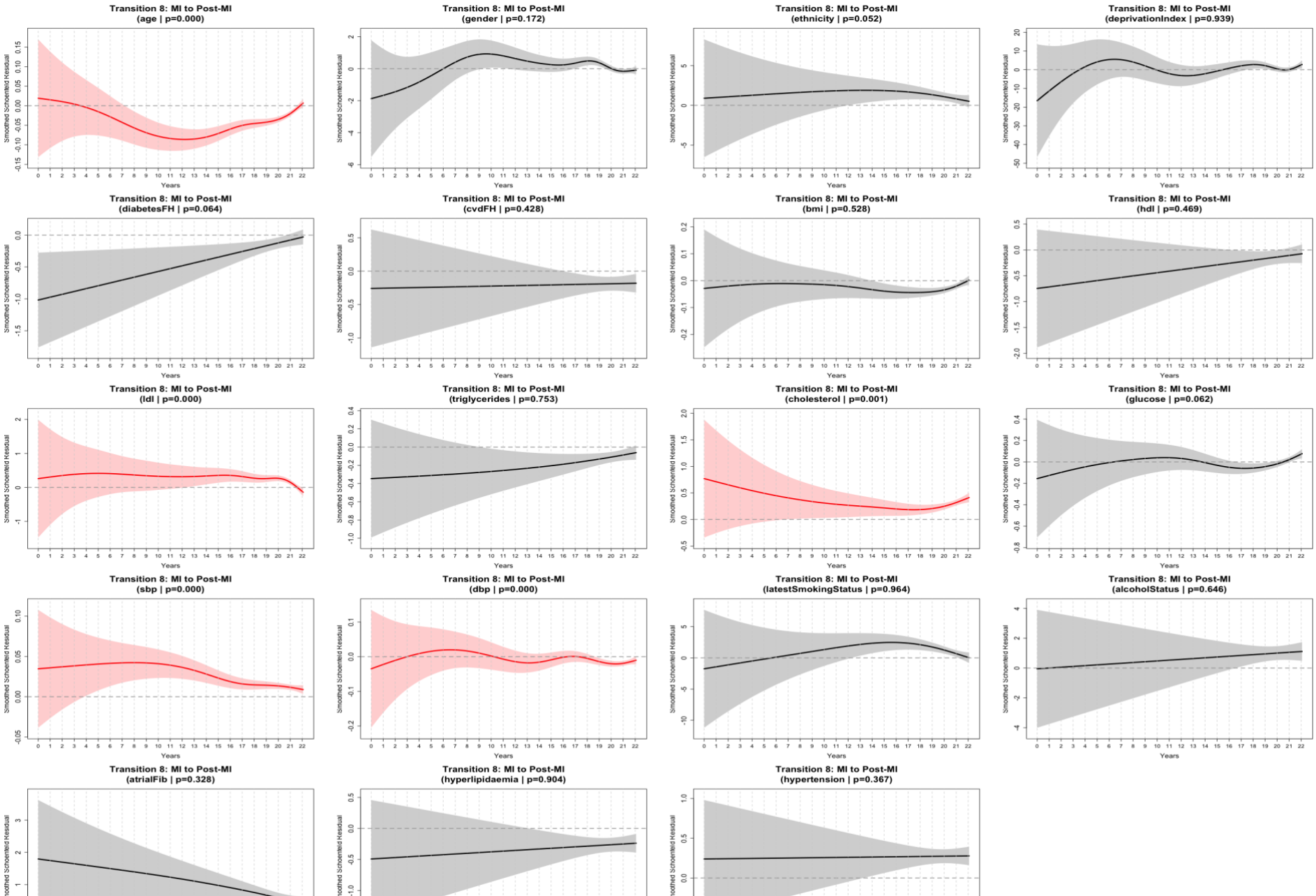

Transition 7: T2DM to Death
(latestSmokingStatus | p=0.123)
Smoothed Schoenfeld Residual
Years

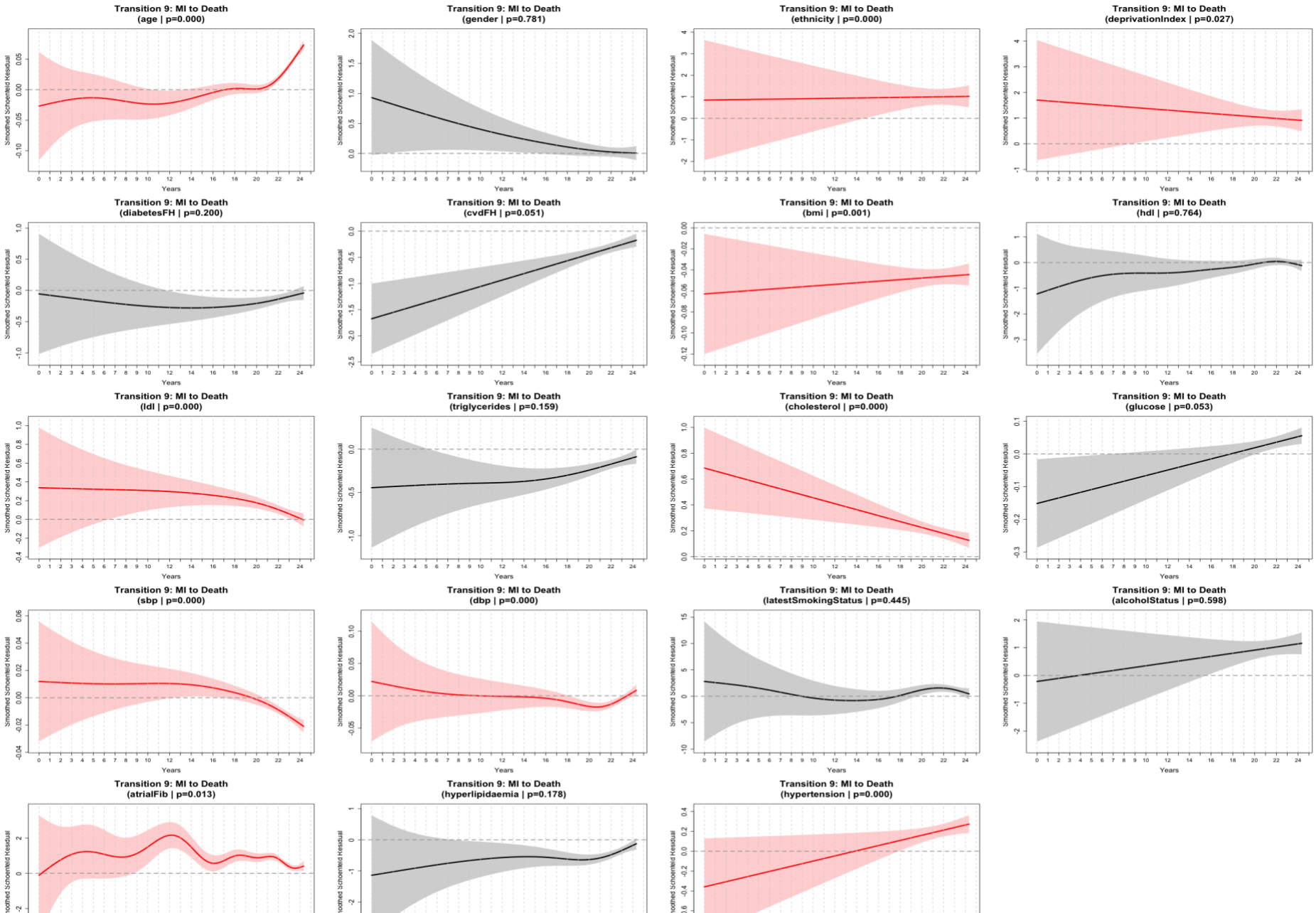

Transition 7: T2DM to Death
(alcoholStatus | p=0.000)
Smoothed Schoenfeld Residual
Years

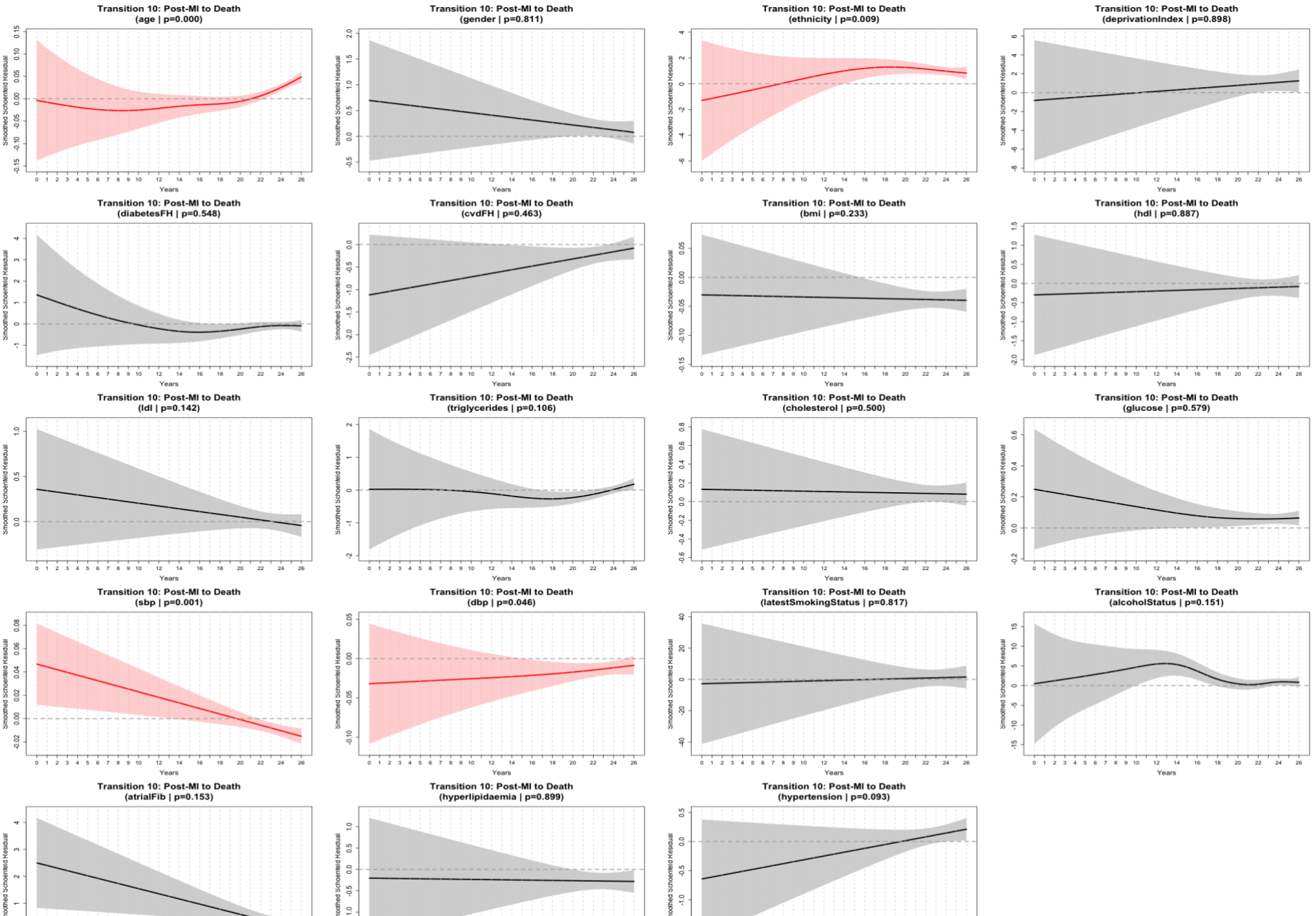

Transition 7: T2DM to Death
(atrialFib | p=0.129)
Years

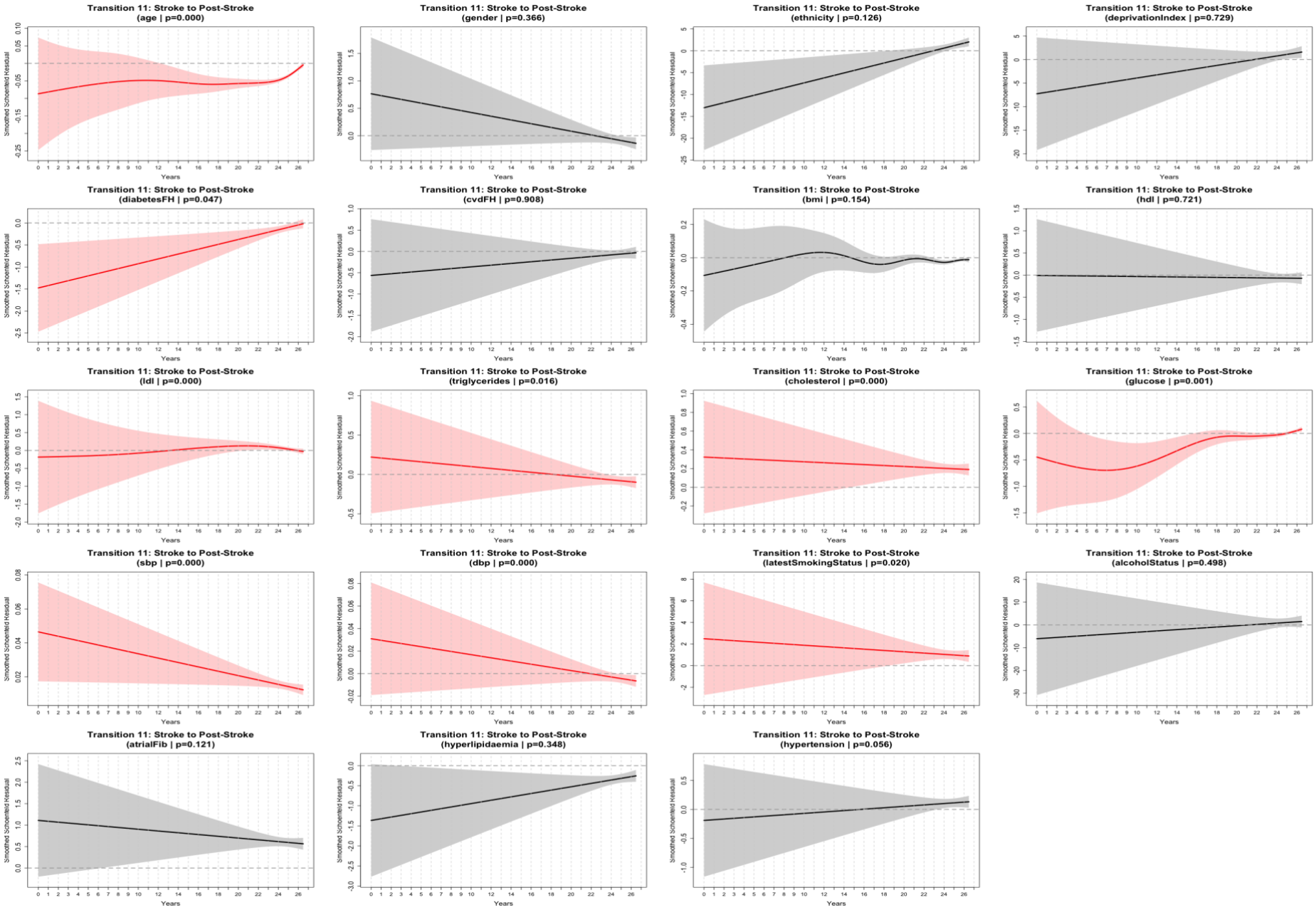

Transition 7: T2DM to Death
(hyperlipidaemia | p=0.406)
Years

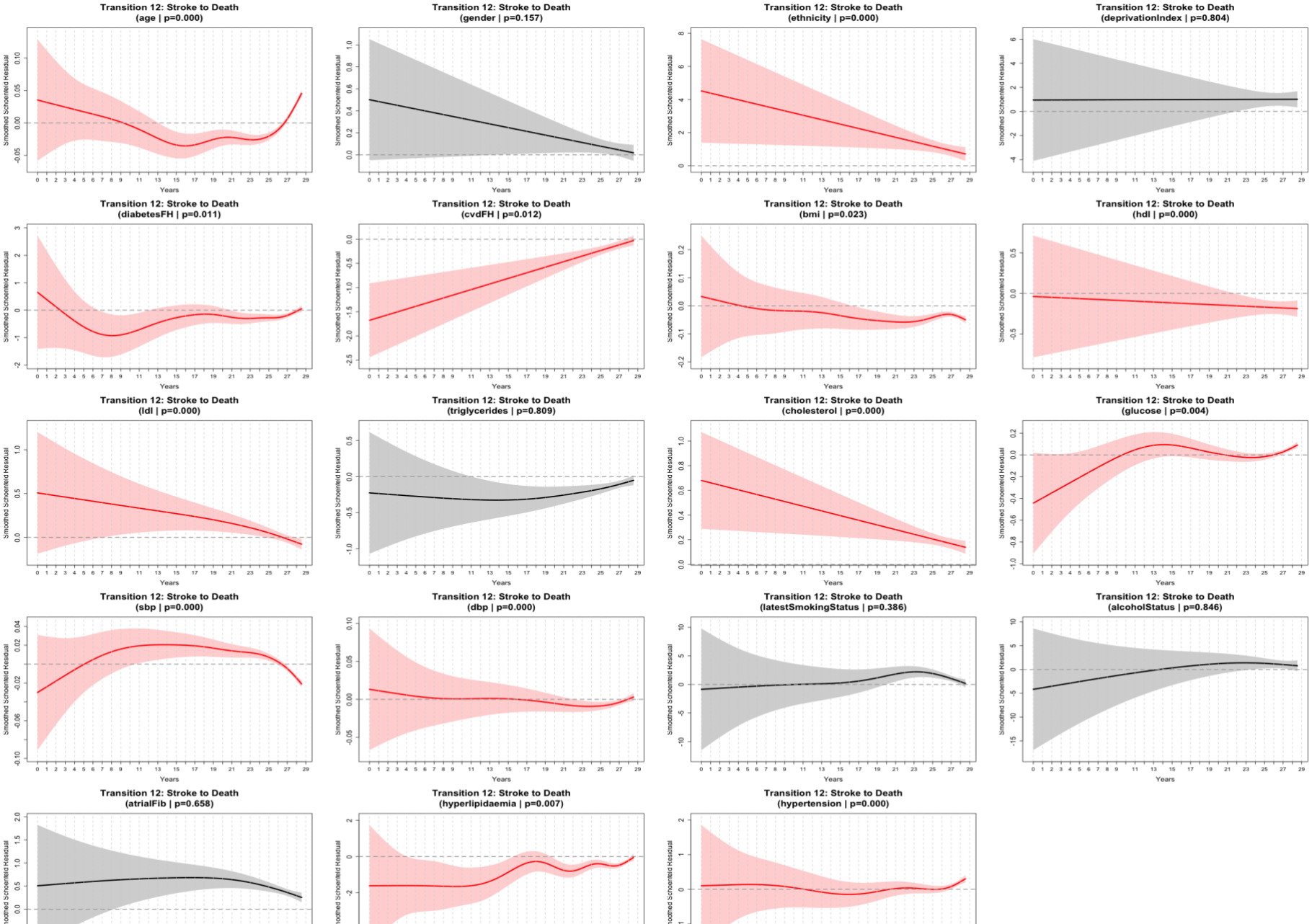

Transition 7: T2DM to Death
(hypertension | p=0.000)
Years

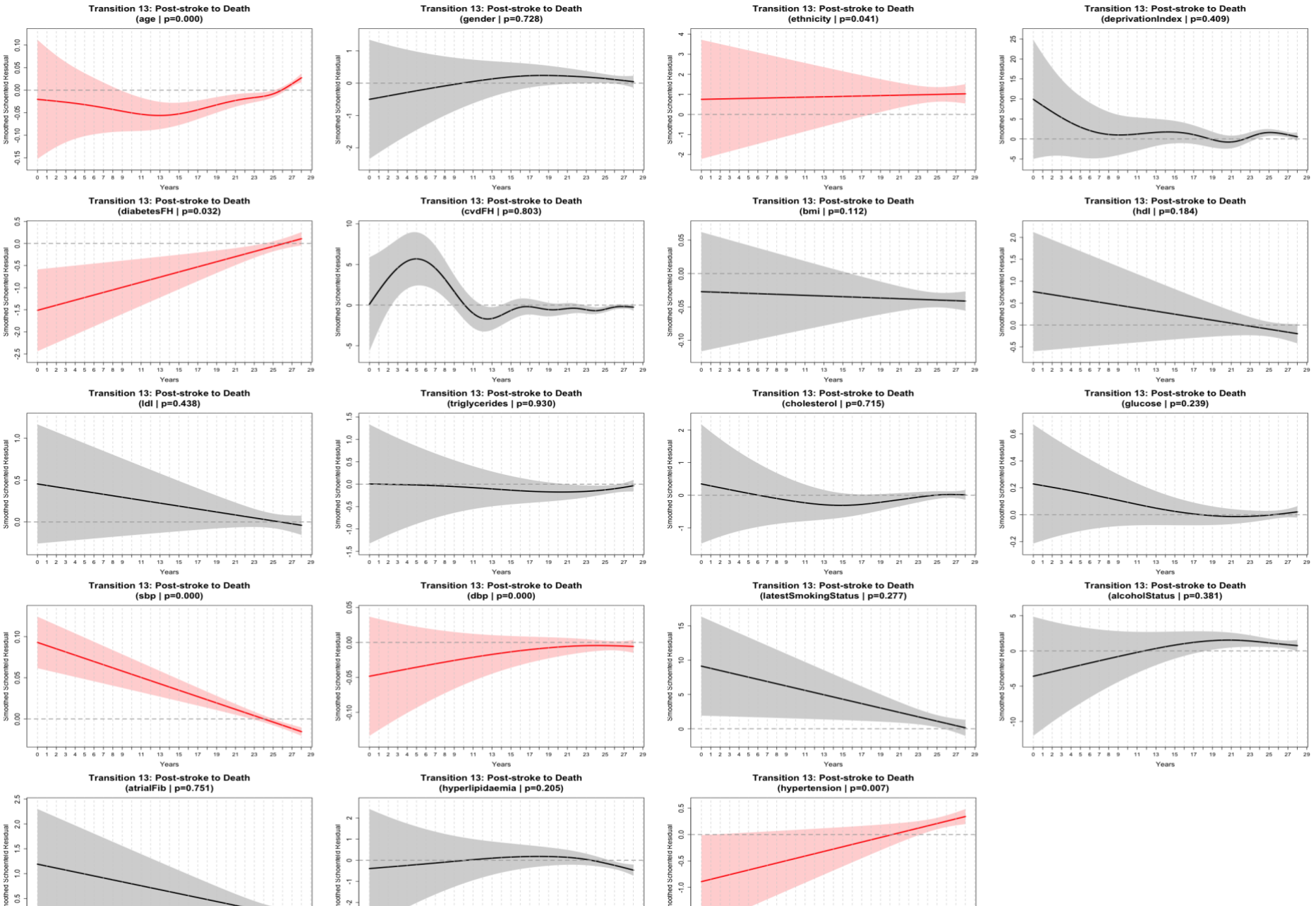

Transition 8: MI to Post-MI
(age | p=0.000)
Transition 8: MI to Post-MI
(gender | p=0.172)
Transition 8: MI to Post-MI
(ethnicity | p=0.052)
Transition 8: MI to Post-MI
(deprivationIndex | p=0.939)
Transition 8: MI to Post-MI
(diabetesFH | p=0.064)
Transition 8: MI to Post-MI
(cvdFH | p=0.428)
Transition 8: MI to Post-MI
(bmi | p=0.528)
Transition 8: MI to Post-MI
(hdl | p=0.469)
Transition 8: MI to Post-MI
(ldl | p=0.000)
Transition 8: MI to Post-MI
(triglycerides | p=0.753)
Transition 8: MI to Post-MI
(cholesterol | p=0.001)
Transition 8: MI to Post-MI
(glucose | p=0.062)
Transition 8: MI to Post-MI
(sbp | p=0.000)
Transition 8: MI to Post-MI
(dbp | p=0.000)
Transition 8: MI to Post-MI
(latestSmokingStatus | p=0.964)
Transition 8: MI to Post-MI
(alcoholStatus | p=0.646)
Transition 8: MI to Post-MI
(atrialFib | p=0.328)
Transition 8: MI to Post-MI
(hyperlipidaemia | p=0.904)
Transition 8: MI to Post-MI
(hypertension | p=0.367)
Smoothed Schoenfeld Residual
Years


Transition 9: MI to Death
(age | p=0.000)
Transition 9: MI to Death
(gender | p=0.781)
Transition 9: MI to Death
(ethnicity | p=0.000)
Transition 9: MI to Death
(deprivationIndex | p=0.027)
Transition 9: MI to Death
(diabetesFH | p=0.200)
Transition 9: MI to Death
(cvdFH | p=0.051)
Transition 9: MI to Death
(bmi | p=0.001)
Transition 9: MI to Death
(hdl | p=0.764)
Transition 9: MI to Death
(ldl | p=0.000)
Transition 9: MI to Death
(triglycerides | p=0.159)
Transition 9: MI to Death
(cholesterol | p=0.000)
Transition 9: MI to Death
(glucose | p=0.053)
Transition 9: MI to Death
(sbp | p=0.000)
Transition 9: MI to Death
(dbp | p=0.000)
Transition 9: MI to Death
(latestSmokingStatus | p=0.445)
Transition 9: MI to Death
(alcoholStatus | p=0.598)
Transition 9: MI to Death
(atrialFib | p=0.013)
Transition 9: MI to Death
(hyperlipidaemia | p=0.178)
Transition 9: MI to Death
(hypertension | p=0.000)
Smoothed Schoenfeld Residual
Years


Transition 10: Post-MI to Death
(age | p=0.000)
Transition 10: Post-MI to Death
(gender | p=0.811)
Transition 10: Post-MI to Death
(ethnicity | p=0.009)
Transition 10: Post-MI to Death
(deprivationIndex | p=0.898)
Transition 10: Post-MI to Death
(diabetesFH | p=0.548)
Transition 10: Post-MI to Death
(cvdFH | p=0.463)
Transition 10: Post-MI to Death
(bmi | p=0.233)
Transition 10: Post-MI to Death
(hdl | p=0.887)
Transition 10: Post-MI to Death
(ldl | p=0.142)
Transition 10: Post-MI to Death
(triglycerides | p=0.106)
Transition 10: Post-MI to Death
(cholesterol | p=0.500)
Transition 10: Post-MI to Death
(glucose | p=0.579)
Transition 10: Post-MI to Death
(sbp | p=0.001)
Transition 10: Post-MI to Death
(dbp | p=0.046)
Transition 10: Post-MI to Death
(latestSmokingStatus | p=0.817)
Transition 10: Post-MI to Death
(alcoholStatus | p=0.151)
Transition 10: Post-MI to Death
(atrialFib | p=0.153)
Transition 10: Post-MI to Death
(hyperlipidaemia | p=0.899)
Transition 10: Post-MI to Death
(hypertension | p=0.093)
Smoothed Schoenfeld Residual
Years

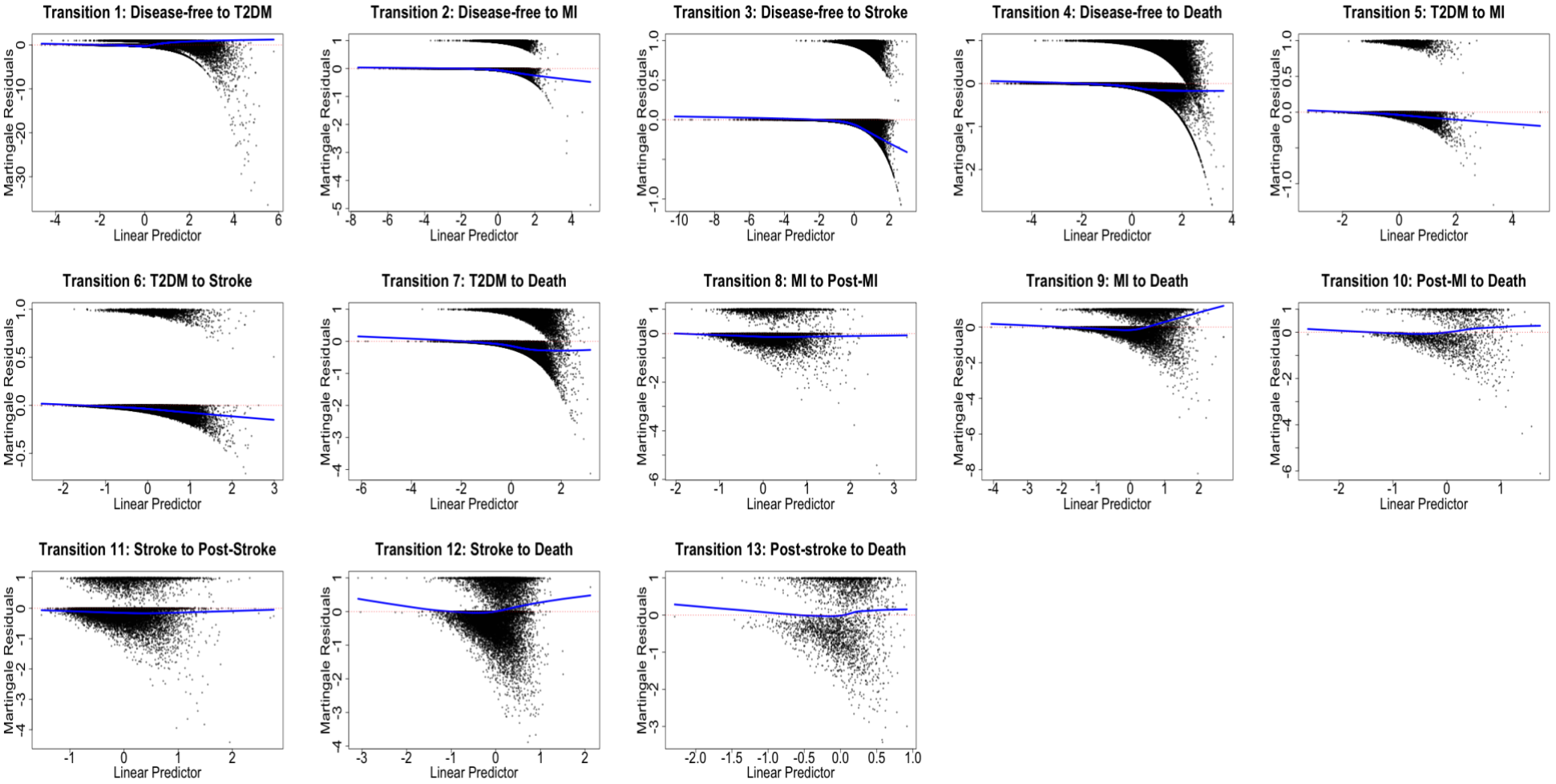

Transition 11: Stroke to Post-Stroke
(age | p=0.000)
Transition 11: Stroke to Post-Stroke
(gender | p=0.366)
Transition 11: Stroke to Post-Stroke
(ethnicity | p=0.126)
Transition 11: Stroke to Post-Stroke
(deprivationIndex | p=0.729)
Transition 11: Stroke to Post-Stroke
(diabetesFH | p=0.047)
Transition 11: Stroke to Post-Stroke
(cvdFH | p=0.908)
Transition 11: Stroke to Post-Stroke
(bmi | p=0.154)
Transition 11: Stroke to Post-Stroke
(hdl | p=0.721)
Transition 11: Stroke to Post-Stroke
(ldl | p=0.000)
Transition 11: Stroke to Post-Stroke
(triglycerides | p=0.016)
Transition 11: Stroke to Post-Stroke
(cholesterol | p=0.000)
Transition 11: Stroke to Post-Stroke
(glucose | p=0.001)
Transition 11: Stroke to Post-Stroke
(sbp | p=0.000)
Transition 11: Stroke to Post-Stroke
(dbp | p=0.000)
Transition 11: Stroke to Post-Stroke
(latestSmokingStatus | p=0.020)
Transition 11: Stroke to Post-Stroke
(alcoholStatus | p=0.498)
Transition 11: Stroke to Post-Stroke
(atrialFib | p=0.121)
Transition 11: Stroke to Post-Stroke
(hyperlipidaemia | p=0.348)
Transition 11: Stroke to Post-Stroke
(hypertension | p=0.056)
Smoothed Schoenfeld Residual
Years


Transition 12: Stroke to Death
(age | p=0.000)
Transition 12: Stroke to Death
(gender | p=0.157)
Transition 12: Stroke to Death
(ethnicity | p=0.000)
Transition 12: Stroke to Death
(deprivationIndex | p=0.804)
Transition 12: Stroke to Death
(diabetesFH | p=0.011)
Transition 12: Stroke to Death
(cvdFH | p=0.012)
Transition 12: Stroke to Death
(bmi | p=0.023)
Transition 12: Stroke to Death
(hdl | p=0.000)
Transition 12: Stroke to Death
(ldl | p=0.000)
Transition 12: Stroke to Death
(triglycerides | p=0.809)
Transition 12: Stroke to Death
(cholesterol | p=0.000)
Transition 12: Stroke to Death
(glucose | p=0.004)
Transition 12: Stroke to Death
(sbp | p=0.000)
Transition 12: Stroke to Death
(dbp | p=0.000)
Transition 12: Stroke to Death
(latestSmokingStatus | p=0.386)
Transition 12: Stroke to Death
(alcoholStatus | p=0.846)
Transition 12: Stroke to Death
(atrialFib | p=0.658)
Transition 12: Stroke to Death
(hyperlipidaemia | p=0.007)
Transition 12: Stroke to Death
(hypertension | p=0.000)
Smoothed Schoenfeld Residual
Years

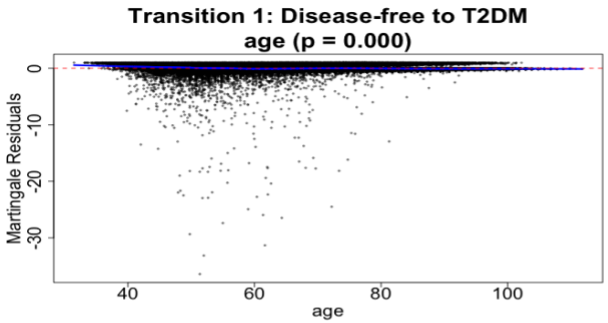

Transition 13: Post-stroke to Death
(age | p=0.000)
Transition 13: Post-stroke to Death
(gender | p=0.728)
Transition 13: Post-stroke to Death
(ethnicity | p=0.041)
Transition 13: Post-stroke to Death
(deprivationIndex | p=0.409)
Transition 13: Post-stroke to Death
(diabetesFH | p=0.032)
Transition 13: Post-stroke to Death
(cvdFH | p=0.803)
Transition 13: Post-stroke to Death
(bmi | p=0.112)
Transition 13: Post-stroke to Death
(hdl | p=0.184)
Transition 13: Post-stroke to Death
(ldl | p=0.438)
Transition 13: Post-stroke to Death
(triglycerides | p=0.930)
Transition 13: Post-stroke to Death
(cholesterol | p=0.715)
Transition 13: Post-stroke to Death
(glucose | p=0.239)
Transition 13: Post-stroke to Death
(sbp | p=0.000)
Transition 13: Post-stroke to Death
(dbp | p=0.000)
Transition 13: Post-stroke to Death
(latestSmokingStatus | p=0.277)
Transition 13: Post-stroke to Death
(alcoholStatus | p=0.381)
Transition 13: Post-stroke to Death
(atrialFib | p=0.751)
Transition 13: Post-stroke to Death
(hyperlipidaemia | p=0.205)
Transition 13: Post-stroke to Death
(hypertension | p=0.007)
Smoothed Schoenfeld Residual
Years

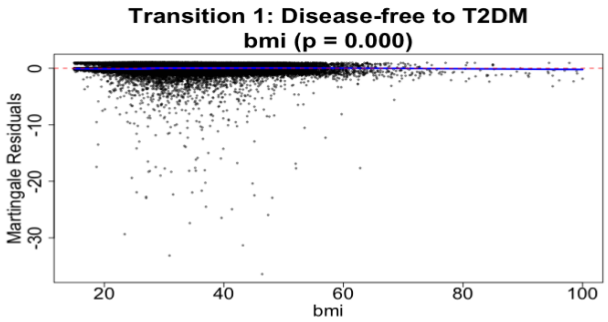
Appendix 7. Martingale residuals

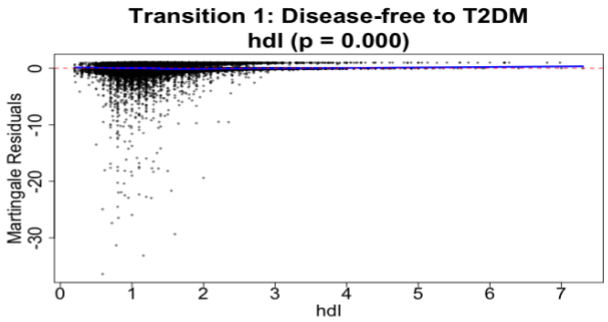
Appendix Figure 1. Overall plot

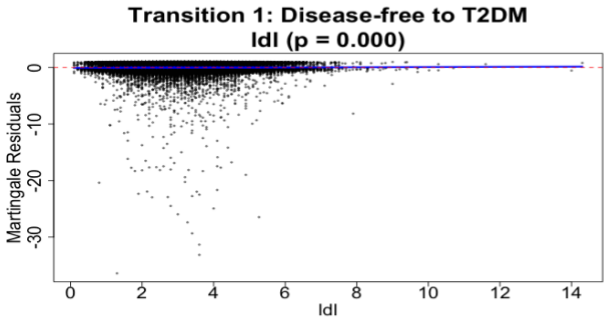

Transition 1: Disease-free to T2DM
Transition 2: Disease-free to MI
Transition 3: Disease-free to Stroke
Transition 4: Disease-free to Death
Transition 5: T2DM to MI
Transition 6: T2DM to Stroke
Transition 7: T2DM to Death
Transition 8: MI to Post-MI
Transition 9: MI to Death
Transition 10: Post-MI to Death
Transition 11: Stroke to Post-Stroke
Transition 12: Stroke to Death
Transition 13: Post-stroke to Death
Martingale Residuals
Linear Predictor

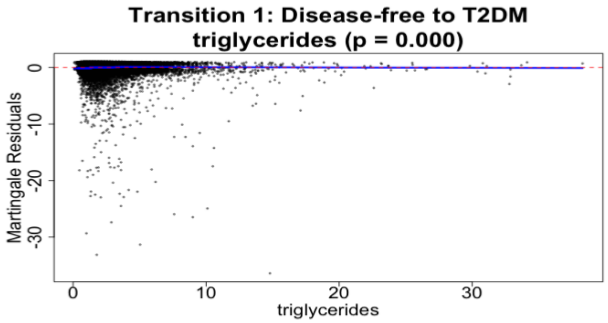
Appendix Figure 2. Individual plot (by covariates)

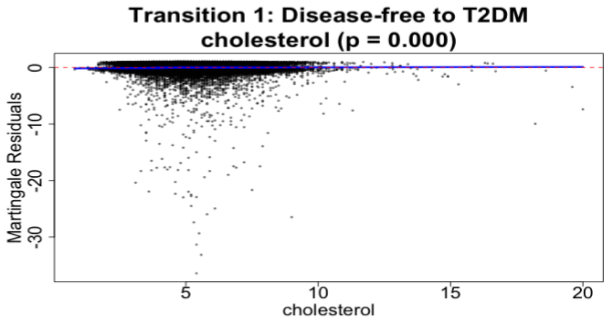

Transition 1: Disease-free to T2DM
age (p = 0.000)
Martingale Residuals
0
-10
-20
-30
40
60
80
100
age

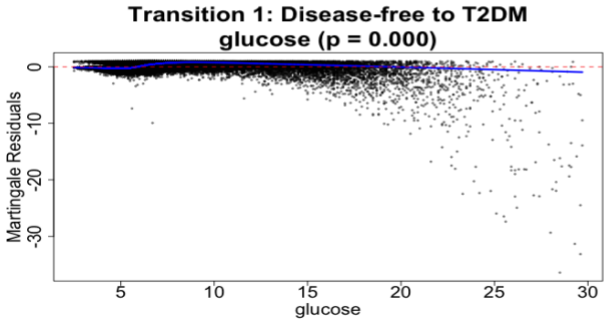

Transition 1: Disease-free to T2DM
bmi (p = 0.000)
Martingale Residuals
0
-10
-20
-30
20
40
60
80
100
bmi

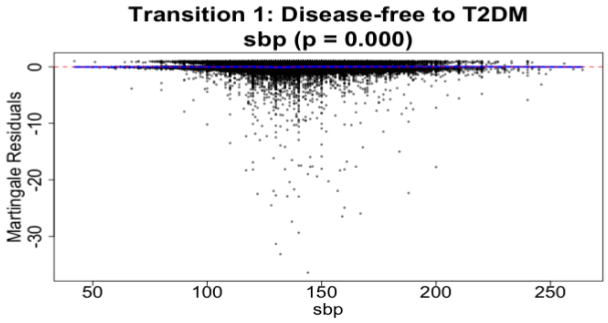

Transition 1: Disease-free to T2DM
hdl (p = 0.000)
Martingale Residuals
0
-10
-20
-30
0
1
2
3
4
5
6
7
hdl

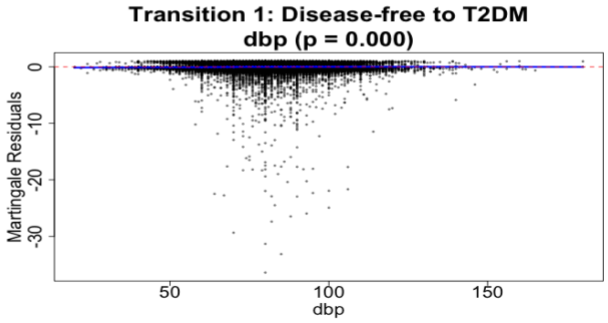

Transition 1: Disease-free to T2DM
ldl (p = 0.000)
Martingale Residuals
0
-10
-20
-30
0
2
4
6
8
10
12
14
ldl

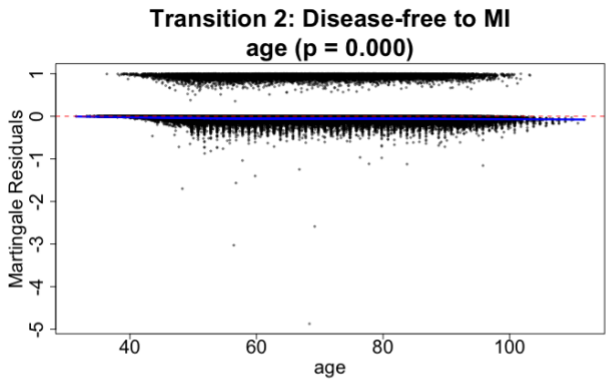

Transition 1: Disease-free to T2DM
triglycerides (p = 0.000)
Martingale Residuals
0
-10
-20
-30
0
10
20
30
triglycerides

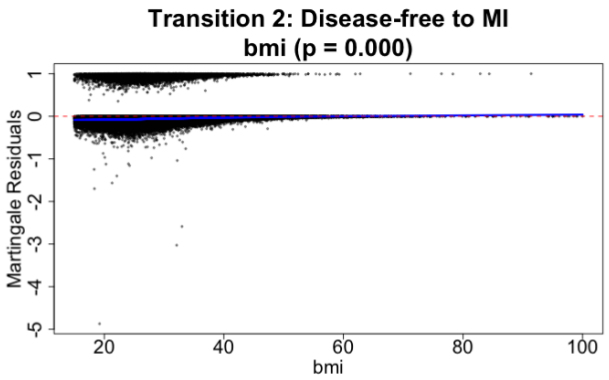

Transition 1: Disease-free to T2DM
cholesterol (p = 0.000)
Martingale Residuals
0
-10
-20
-30
5
10
15
20
cholesterol

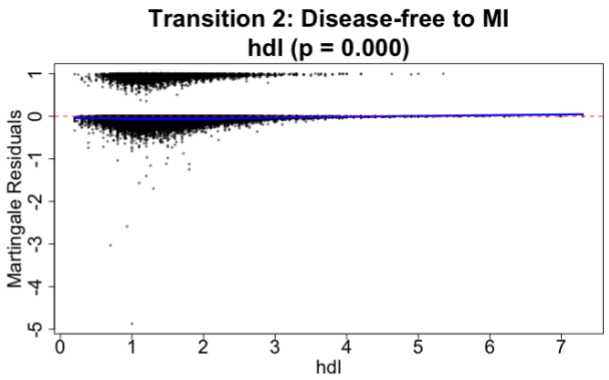

Transition 1: Disease-free to T2DM
glucose (p = 0.000)
Martingale Residuals
0
-10
-20
-30
5
10
15
20
25
30
glucose

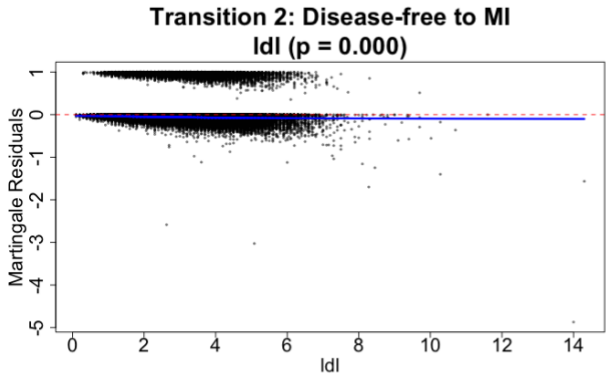

Transition 1: Disease-free to T2DM
sbp (p = 0.000)
Martingale Residuals
0
-10
-20
-30
50
100
150
200
250
sbp

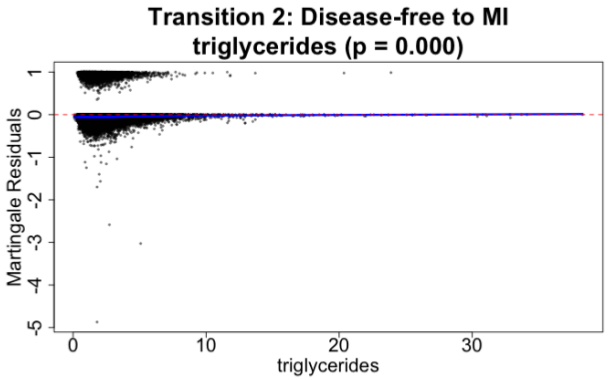

Transition 1: Disease-free to T2DM
dbp (p = 0.000)
Martingale Residuals
0
-10
-20
-30
50
100
150
dbp

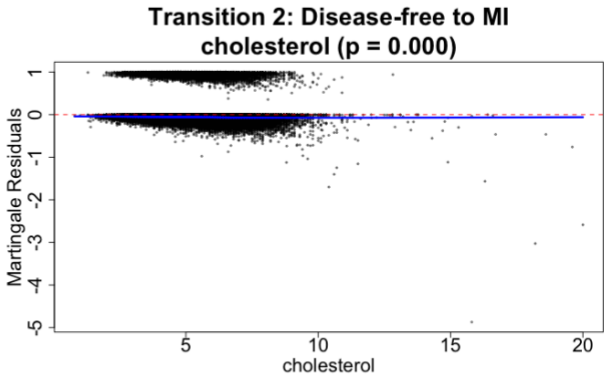

Transition 2: Disease-free to MI
age (p = 0.000)
Martingale Residuals
age

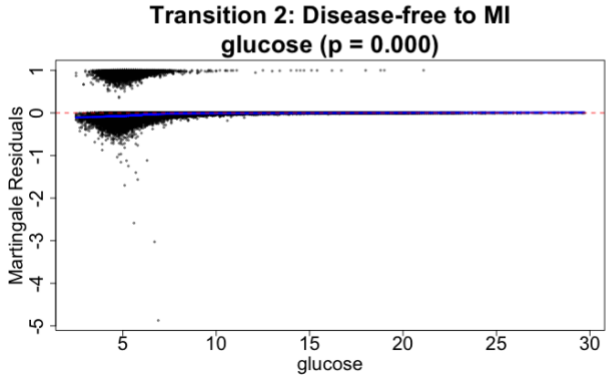

Transition 2: Disease-free to MI
bmi (p = 0.000)
Martingale Residuals
bmi

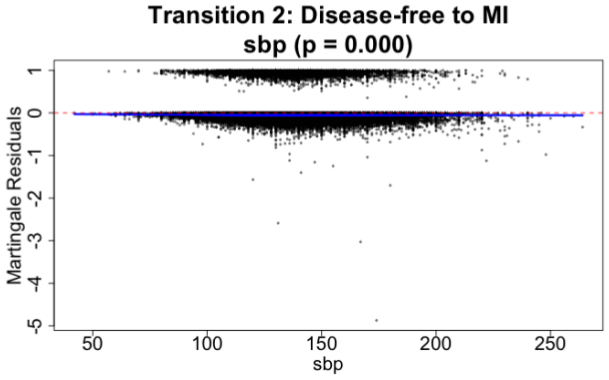

Transition 2: Disease-free to MI
hdl (p = 0.000)
Martingale Residuals
hdl

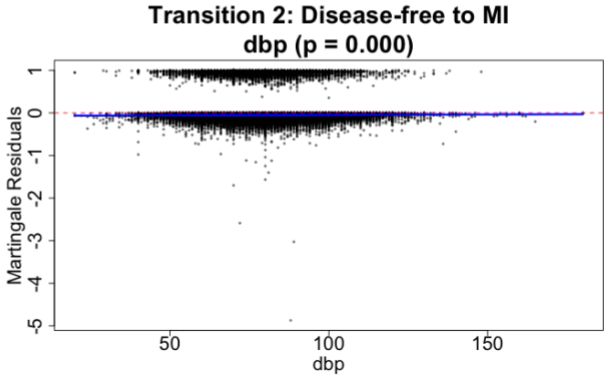

Transition 2: Disease-free to MI
ldl (p = 0.000)
Martingale Residuals
ldl

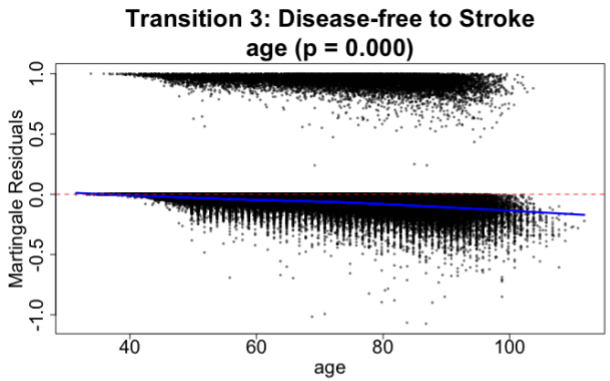

Transition 2: Disease-free to MI
triglycerides (p = 0.000)
Martingale Residuals
triglycerides

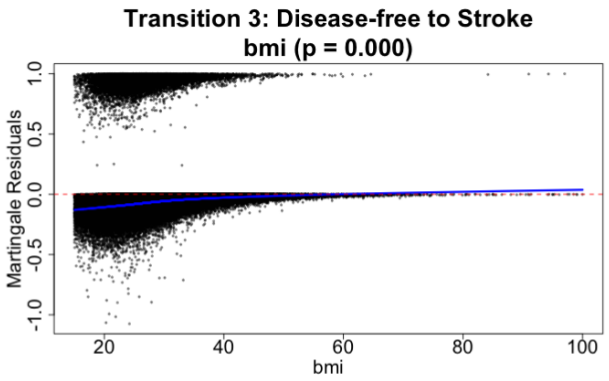

Transition 2: Disease-free to MI
cholesterol (p = 0.000)
Martingale Residuals
cholesterol

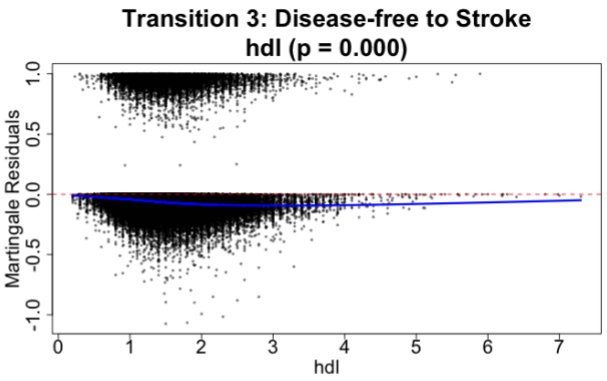

Transition 2: Disease-free to MI
glucose (p = 0.000)
Martingale Residuals
glucose

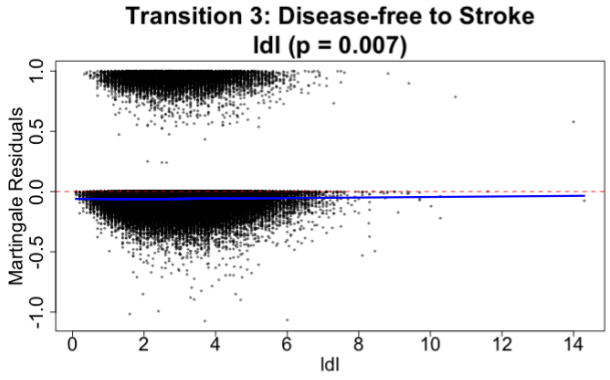

Transition 2: Disease-free to MI
sbp (p = 0.000)
Martingale Residuals
sbp

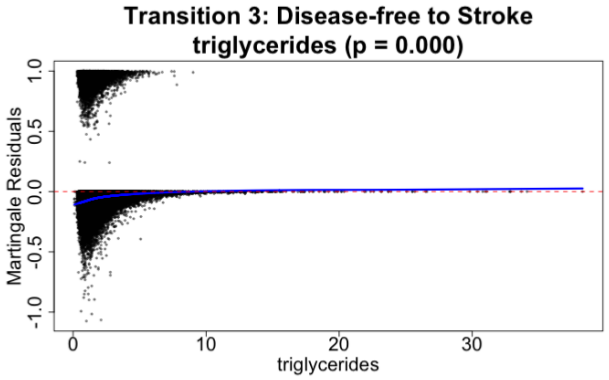

Transition 2: Disease-free to MI
dbp (p = 0.000)
Martingale Residuals
dbp

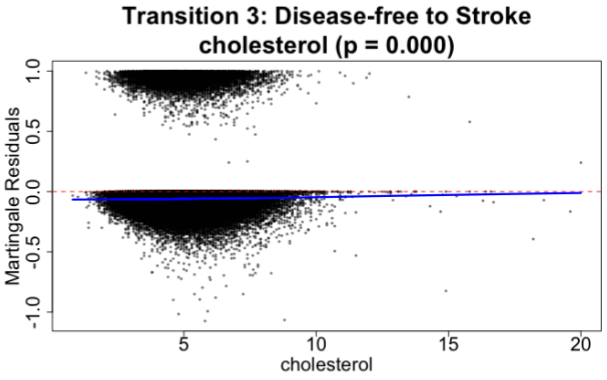

Transition 3: Disease-free to Stroke
age (p = 0.000)
Martingale Residuals
age

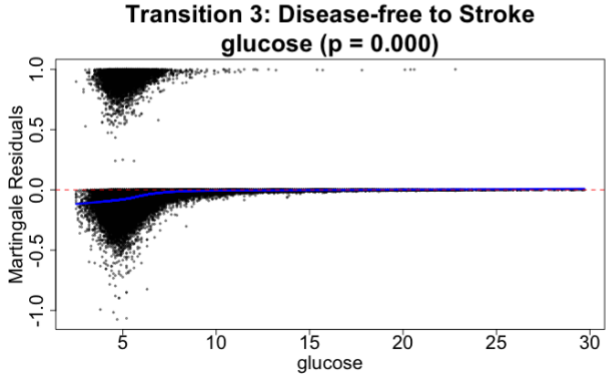

Transition 3: Disease-free to Stroke
bmi (p = 0.000)
Martingale Residuals
bmi

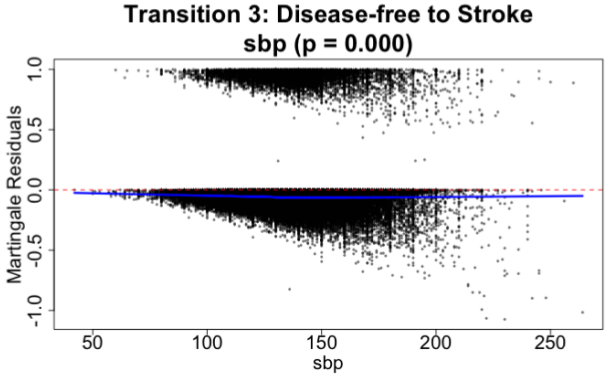

Transition 3: Disease-free to Stroke
hdl (p = 0.000)
Martingale Residuals
hdl

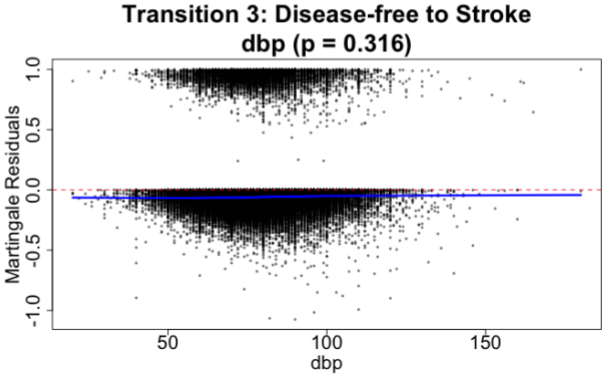

Transition 3: Disease-free to Stroke
ldl (p = 0.007)
Martingale Residuals
ldl

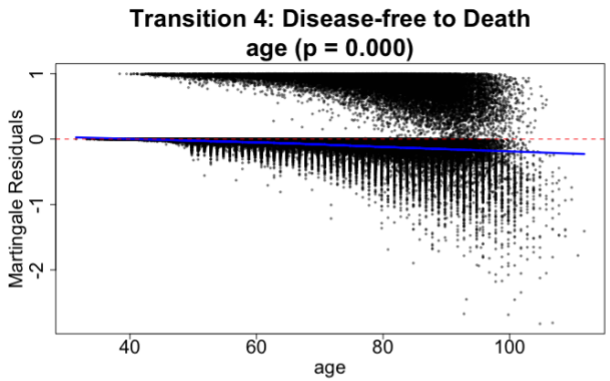

Transition 3: Disease-free to Stroke
triglycerides (p = 0.000)
Martingale Residuals
triglycerides

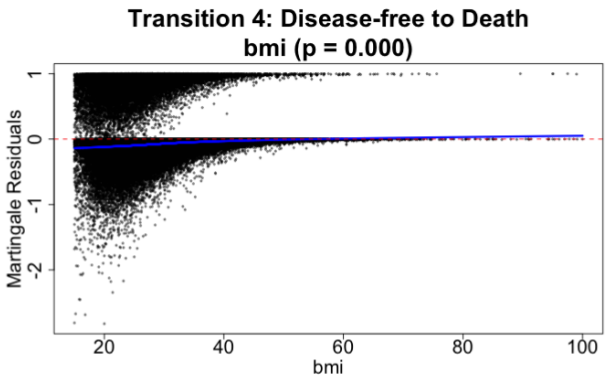

Transition 3: Disease-free to Stroke
cholesterol (p = 0.000)
Martingale Residuals
cholesterol

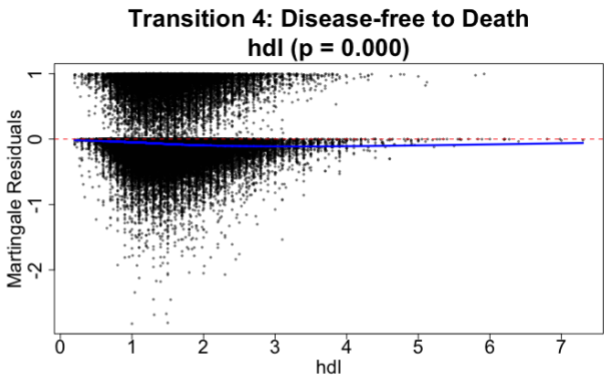

Transition 3: Disease-free to Stroke
glucose (p = 0.000)
Martingale Residuals
glucose

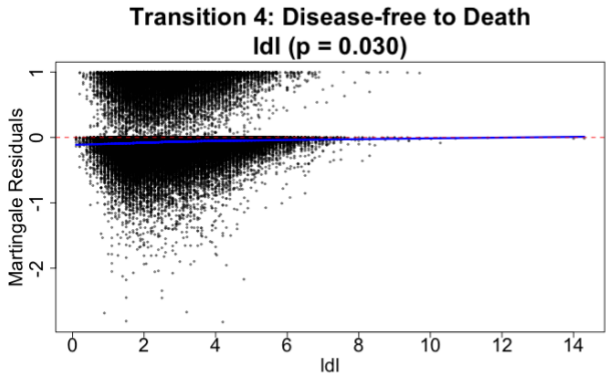

Transition 3: Disease-free to Stroke
sbp (p = 0.000)
Martingale Residuals
sbp

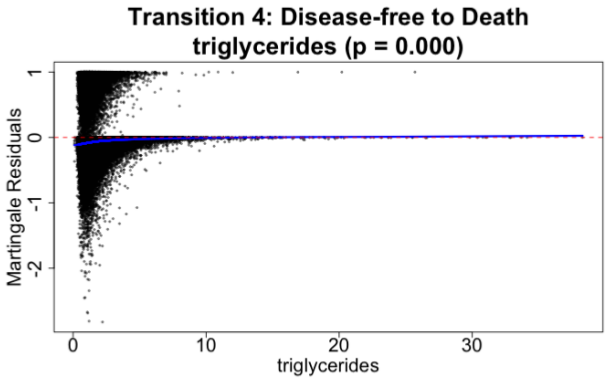

Transition 3: Disease-free to Stroke
dbp (p = 0.316)
Martingale Residuals
dbp

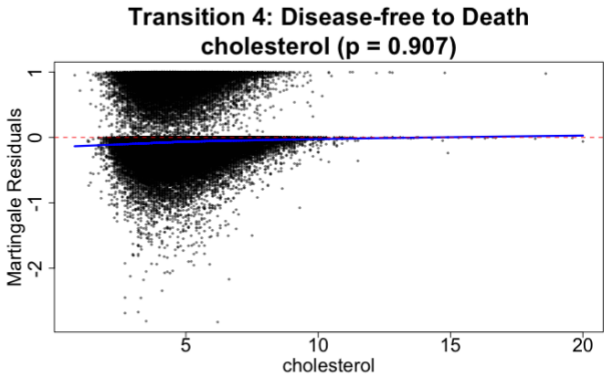

Transition 4: Disease-free to Death
age (p = 0.000)
Martingale Residuals
age

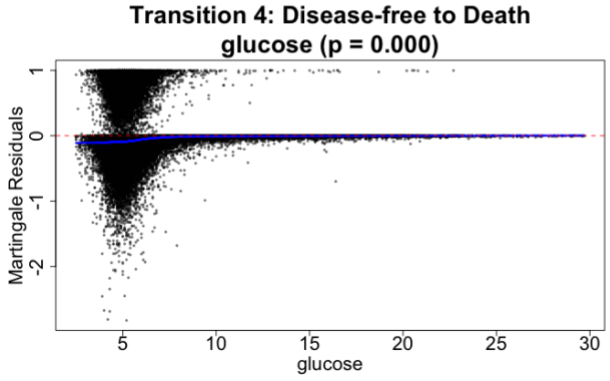

Transition 4: Disease-free to Death
bmi (p = 0.000)
Martingale Residuals
bmi

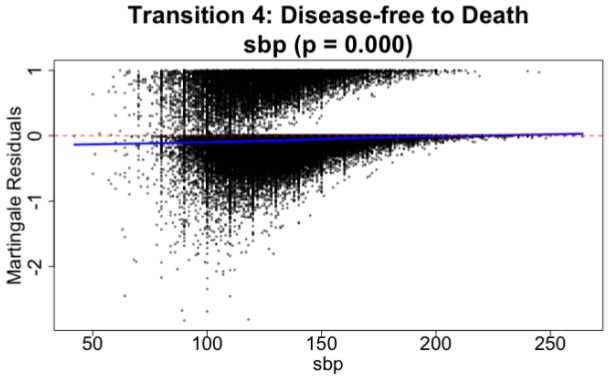

Transition 4: Disease-free to Death
hdl (p = 0.000)
Martingale Residuals
hdl

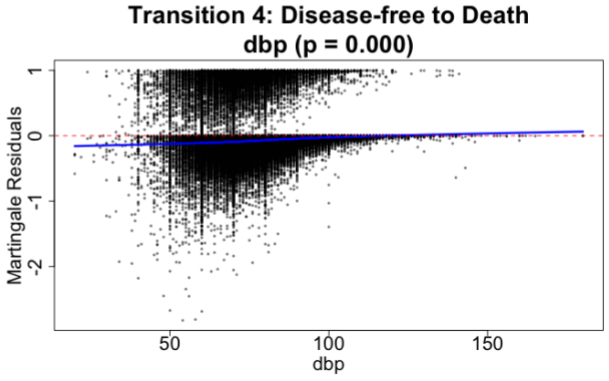

Transition 4: Disease-free to Death
ldl (p = 0.030)
Martingale Residuals
ldl

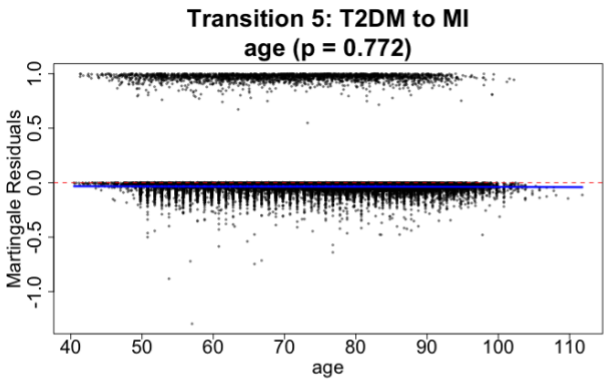

Transition 4: Disease-free to Death
triglycerides (p = 0.000)
Martingale Residuals
triglycerides

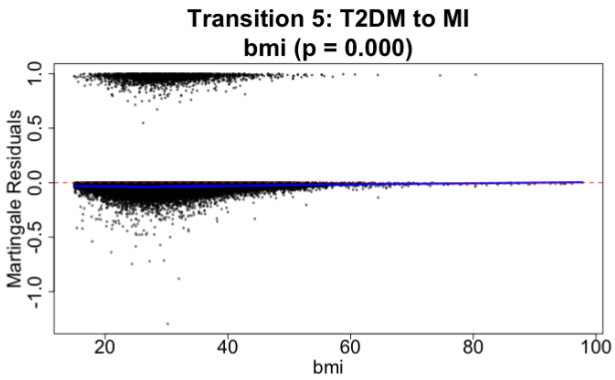

Transition 4: Disease-free to Death
cholesterol (p = 0.907)
Martingale Residuals
cholesterol

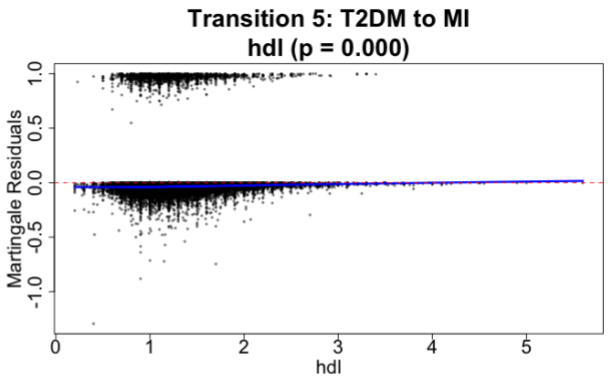

Transition 4: Disease-free to Death
glucose (p = 0.000)
Martingale Residuals
glucose

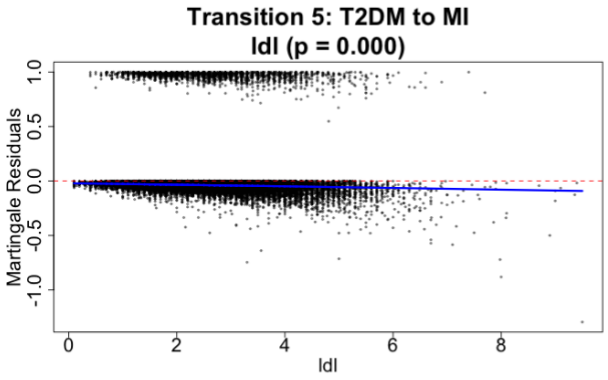

Transition 4: Disease-free to Death
sbp (p = 0.000)
Martingale Residuals
sbp

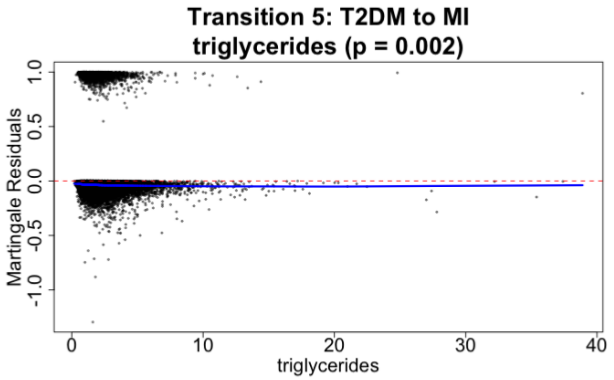

Transition 4: Disease-free to Death
dbp (p = 0.000)
Martingale Residuals
dbp

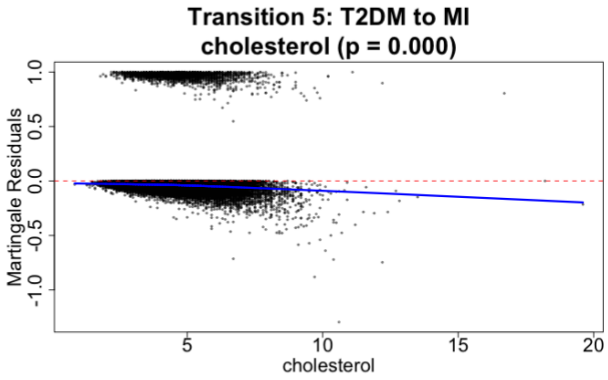

Transition 5: T2DM to MI
age (p = 0.772)
Martingale Residuals
age

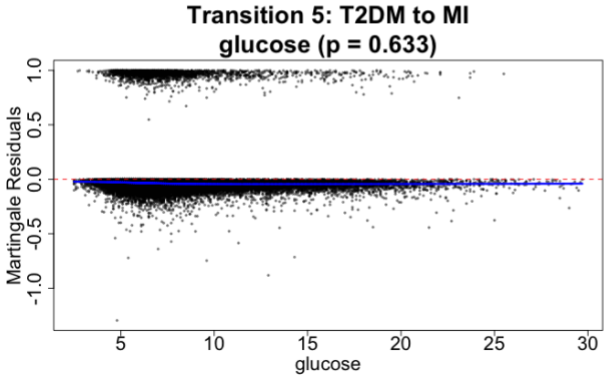

Transition 5: T2DM to MI
bmi (p = 0.000)
Martingale Residuals
bmi

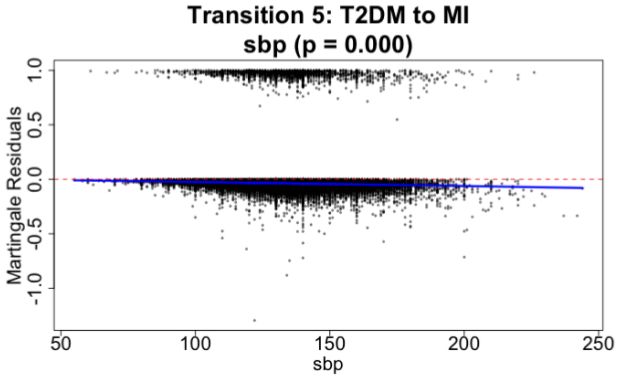

Transition 5: T2DM to MI
hdl (p = 0.000)
Martingale Residuals
hdl

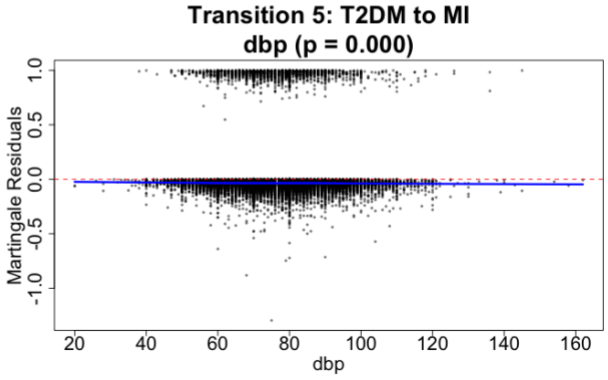

Transition 5: T2DM to MI
ldl (p = 0.000)
Martingale Residuals
ldl

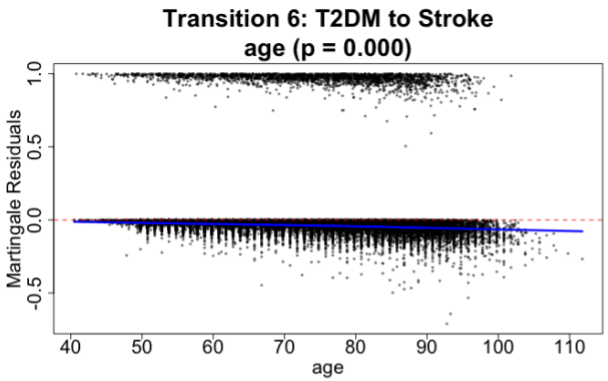

Transition 5: T2DM to MI
triglycerides (p = 0.002)
Martingale Residuals
triglycerides

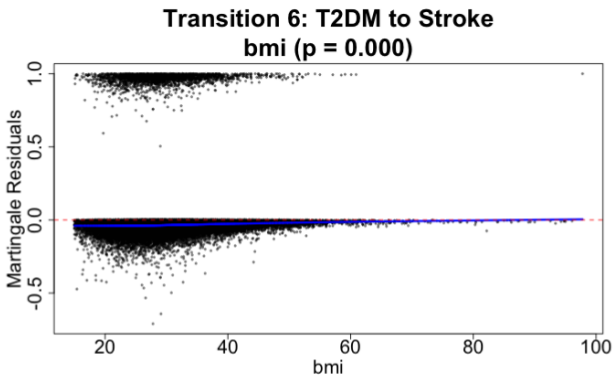

Transition 5: T2DM to MI
cholesterol (p = 0.000)
Martingale Residuals
cholesterol

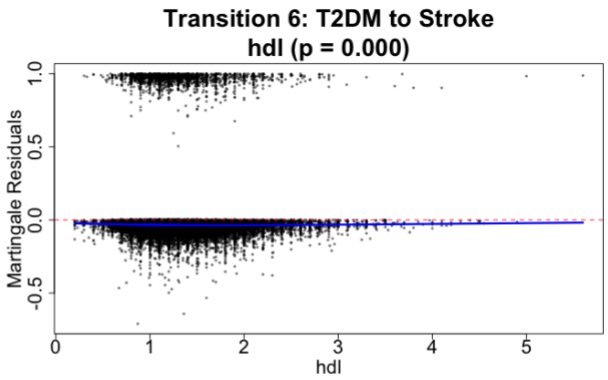

Transition 5: T2DM to MI
glucose (p = 0.633)
Martingale Residuals
glucose

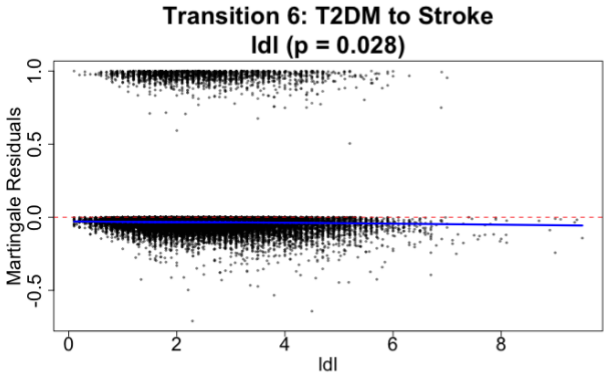

Transition 5: T2DM to MI
sbp (p = 0.000)
Martingale Residuals
sbp

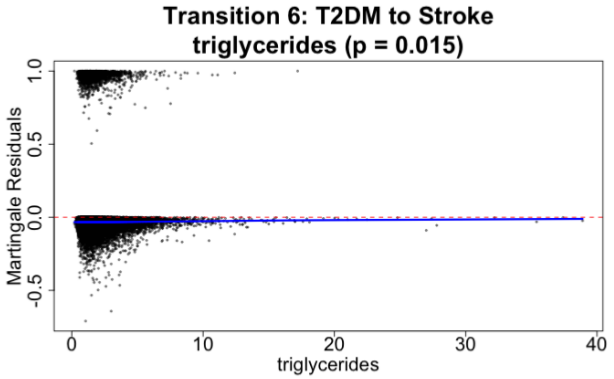

Transition 5: T2DM to MI
dbp (p = 0.000)
Martingale Residuals
dbp

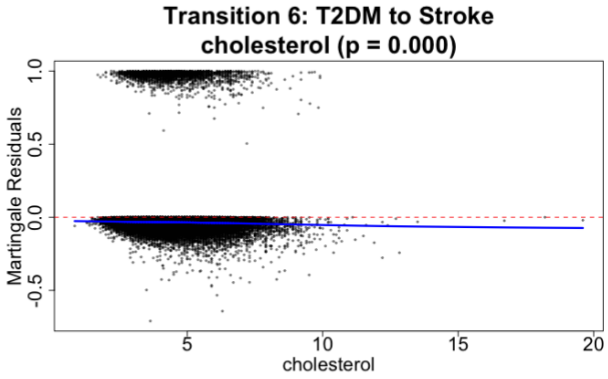

Transition 6: T2DM to Stroke
age (p = 0.000)
Martingale Residuals
age

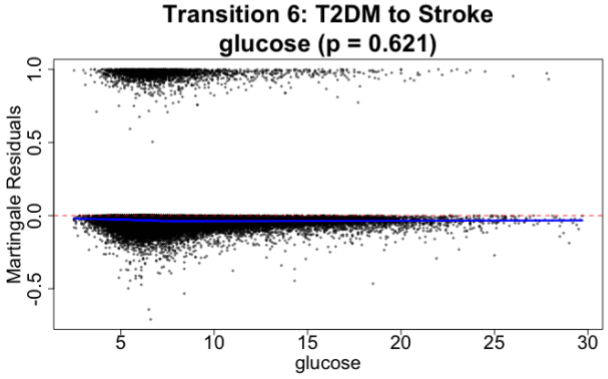

Transition 6: T2DM to Stroke
bmi (p = 0.000)
Martingale Residuals
bmi

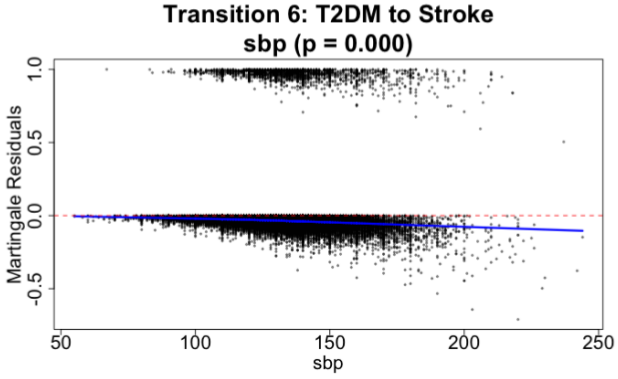

Transition 6: T2DM to Stroke
hdl (p = 0.000)
Martingale Residuals
hdl

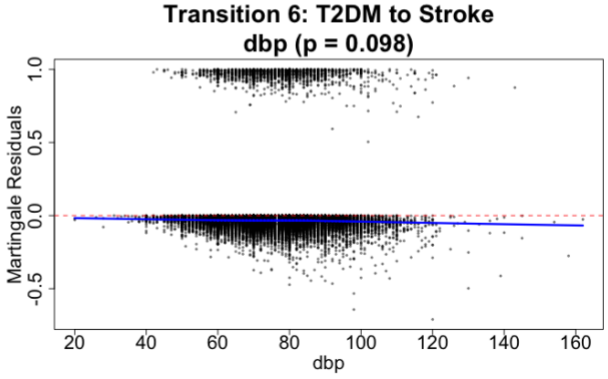

Transition 6: T2DM to Stroke
ldl (p = 0.028)
Martingale Residuals
ldl

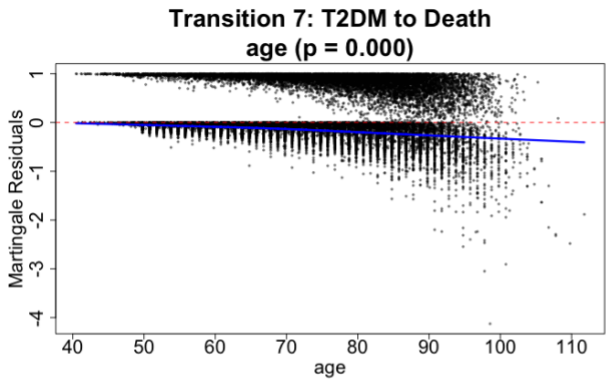

Transition 6: T2DM to Stroke
triglycerides (p = 0.015)
Martingale Residuals
triglycerides

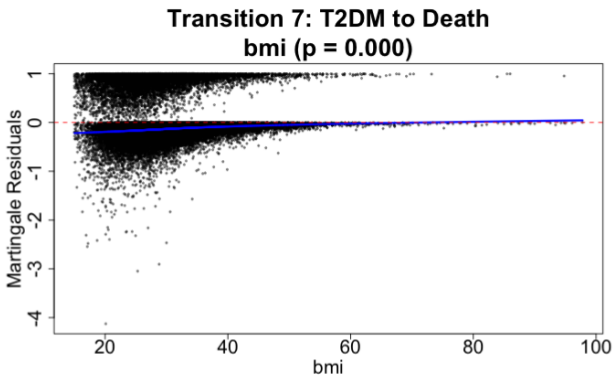

Transition 6: T2DM to Stroke
cholesterol (p = 0.000)
Martingale Residuals
cholesterol

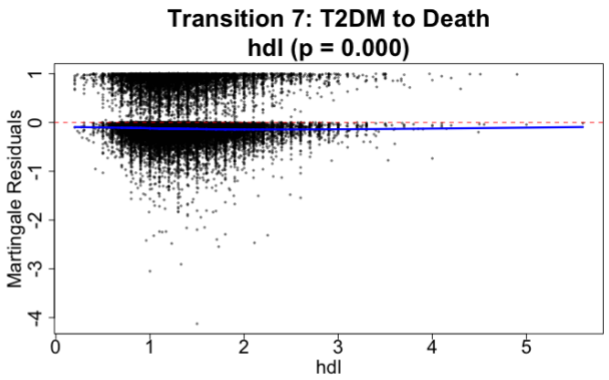

Transition 6: T2DM to Stroke
glucose (p = 0.621)
Martingale Residuals
glucose

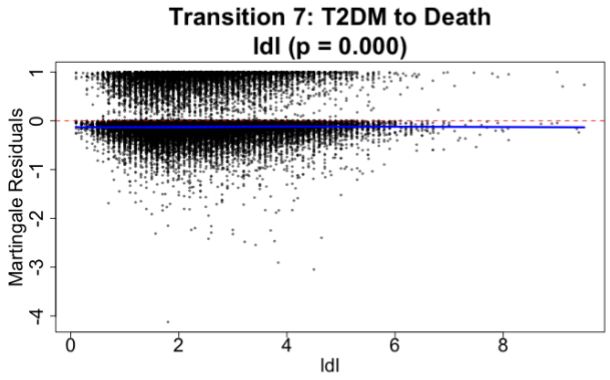

Transition 6: T2DM to Stroke
sbp (p = 0.000)
Martingale Residuals
sbp

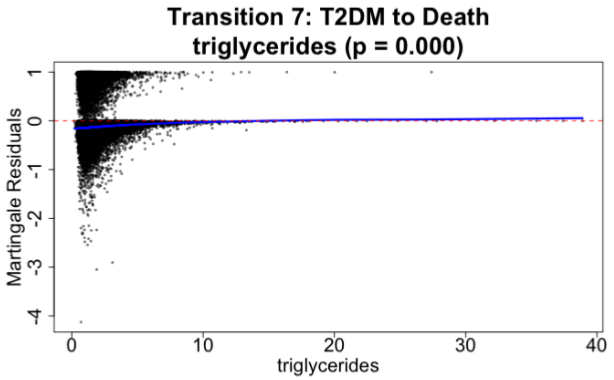

Transition 6: T2DM to Stroke
dbp (p = 0.098)
Martingale Residuals
dbp

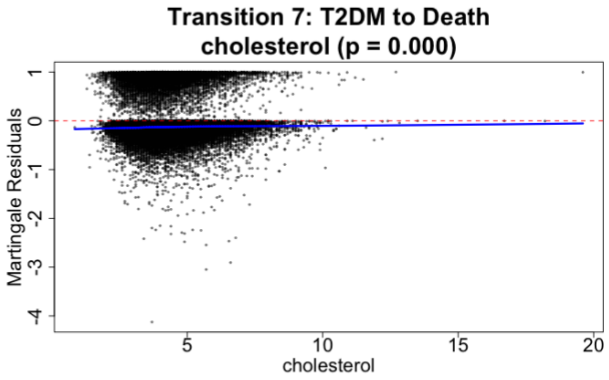

Transition 7: T2DM to Death
age (p = 0.000)
Martingale Residuals
age

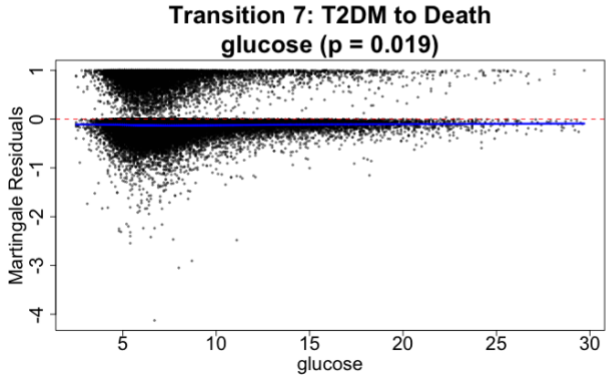

Transition 7: T2DM to Death
bmi (p = 0.000)
Martingale Residuals
bmi

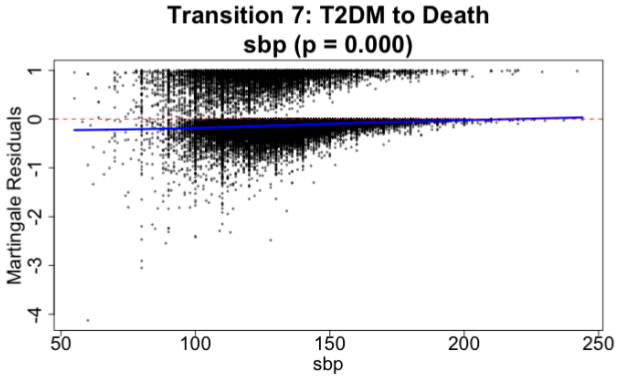

Transition 7: T2DM to Death
hdl (p = 0.000)
Martingale Residuals
hdl

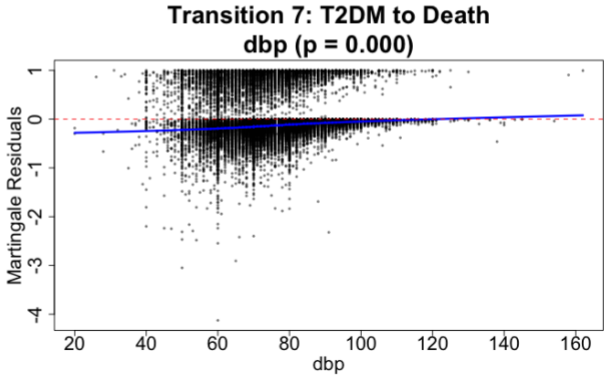

Transition 7: T2DM to Death
ldl (p = 0.000)
Martingale Residuals
ldl

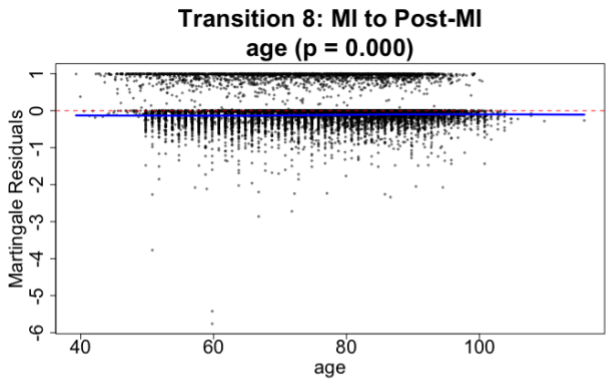

Transition 7: T2DM to Death
triglycerides (p = 0.000)
Martingale Residuals
triglycerides

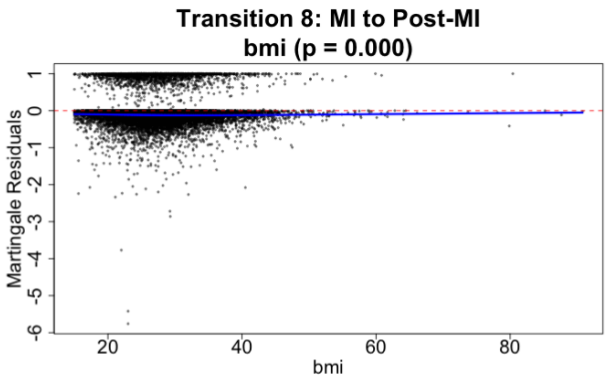

Transition 7: T2DM to Death
cholesterol (p = 0.000)
Martingale Residuals
cholesterol

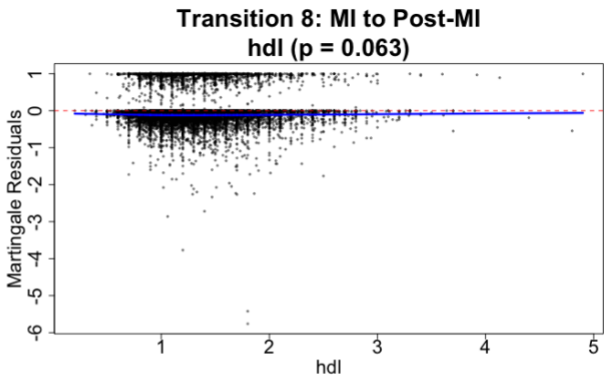

Transition 7: T2DM to Death
glucose (p = 0.019)
Martingale Residuals
glucose

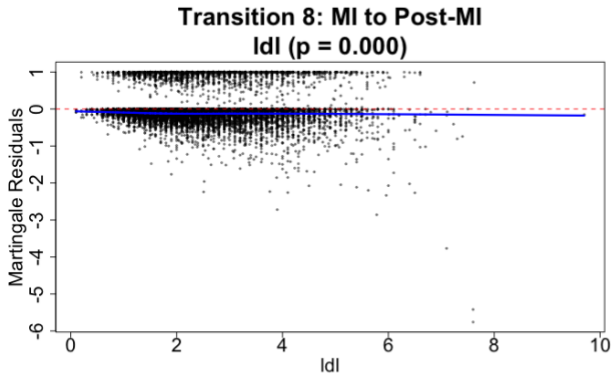

Transition 7: T2DM to Death
sbp (p = 0.000)
Martingale Residuals
sbp

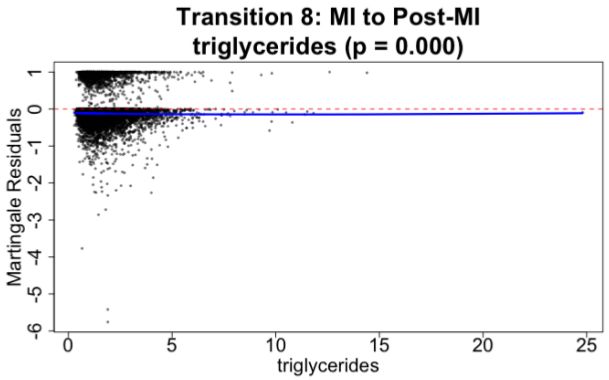

Transition 7: T2DM to Death
dbp (p = 0.000)
Martingale Residuals
dbp

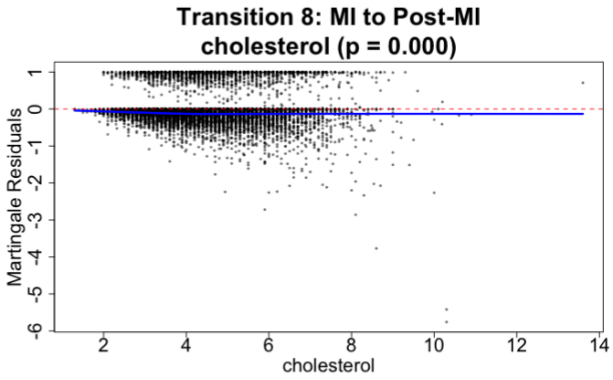

Transition 8: MI to Post-MI
age (p = 0.000)
Martingale Residuals
age

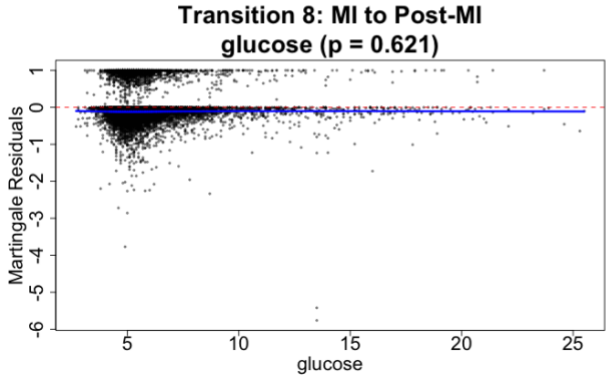

Transition 8: MI to Post-MI
bmi (p = 0.000)
Martingale Residuals
bmi

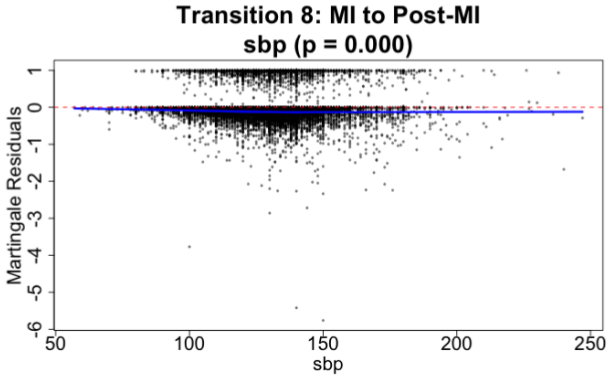

Transition 8: MI to Post-MI
hdl (p = 0.063)
Martingale Residuals
hdl

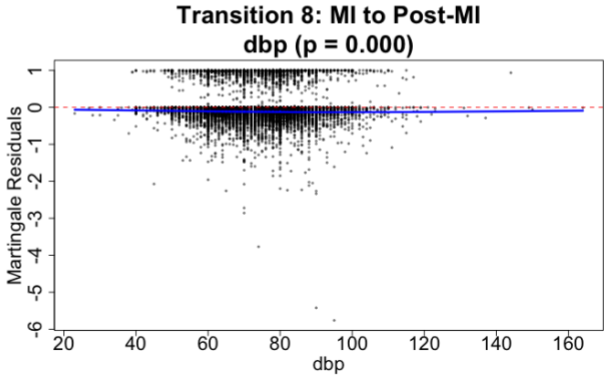

Transition 8: MI to Post-MI
ldl (p = 0.000)
Martingale Residuals
ldl

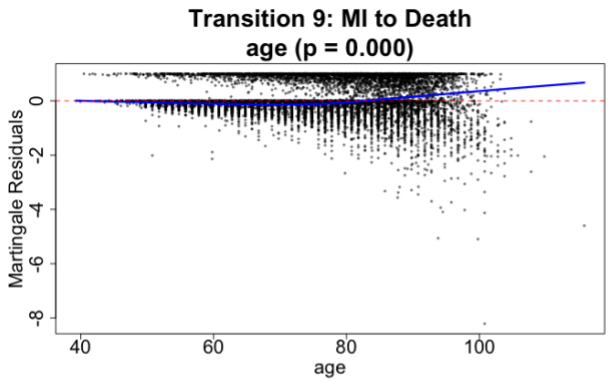

Transition 8: MI to Post-MI
triglycerides (p = 0.000)
Martingale Residuals
triglycerides

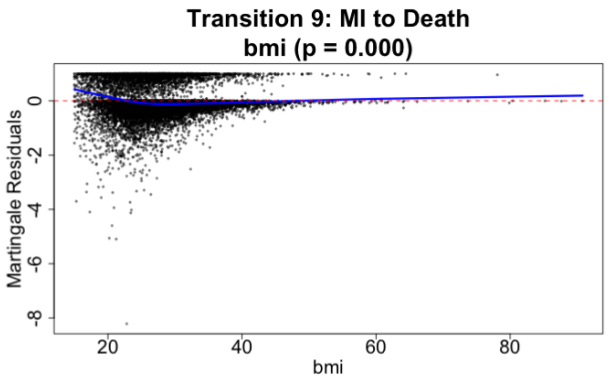

Transition 8: MI to Post-MI
cholesterol (p = 0.000)
Martingale Residuals
cholesterol

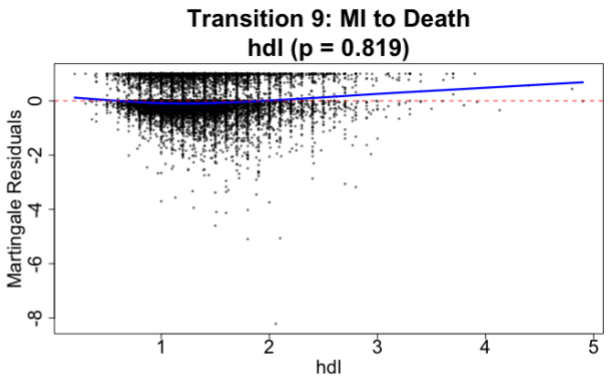

Transition 8: MI to Post-MI
glucose (p = 0.621)
Martingale Residuals
glucose

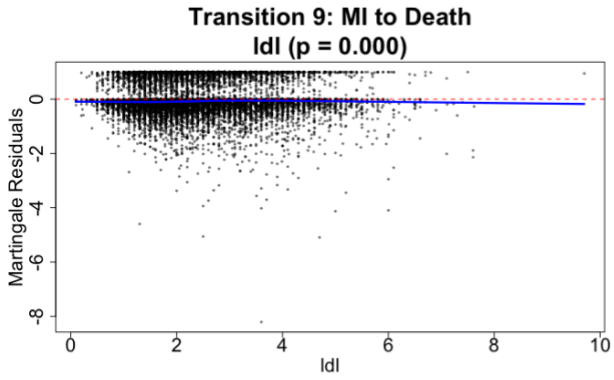

Transition 8: MI to Post-MI
sbp (p = 0.000)
Martingale Residuals
sbp

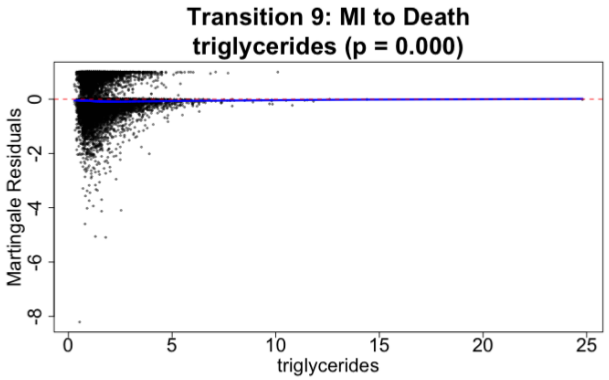

Transition 8: MI to Post-MI
dbp (p = 0.000)
Martingale Residuals
dbp

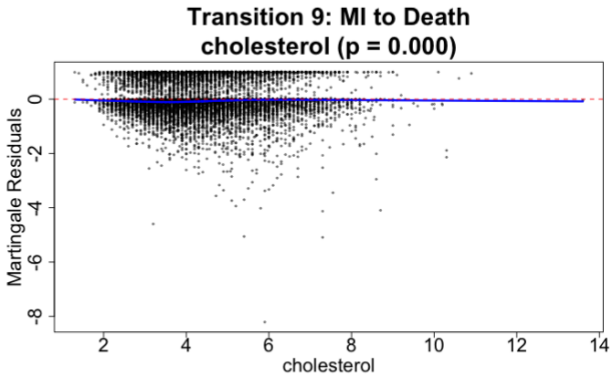

Transition 9: MI to Death
age (p = 0.000)
Martingale Residuals
age

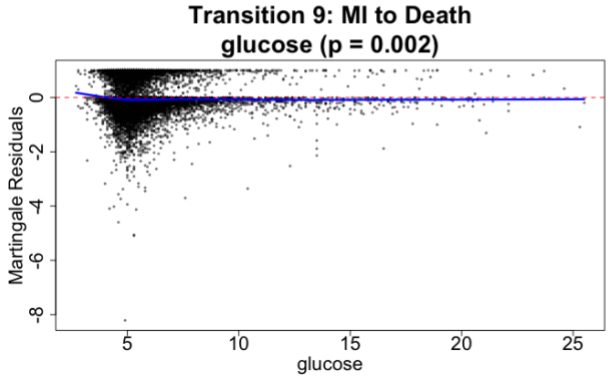

Transition 9: MI to Death
bmi (p = 0.000)
Martingale Residuals
bmi

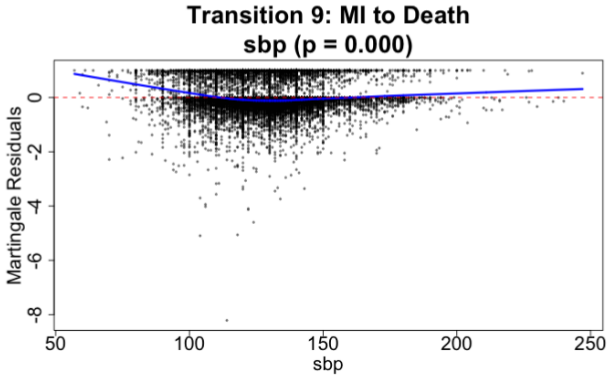

Transition 9: MI to Death
hdl (p = 0.819)
Martingale Residuals
hdl

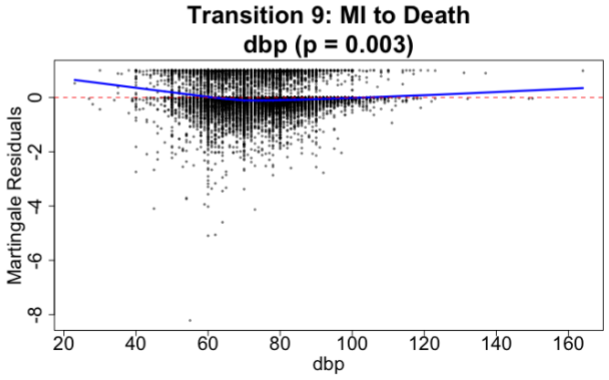

Transition 9: MI to Death
ldl (p = 0.000)
Martingale Residuals
ldl

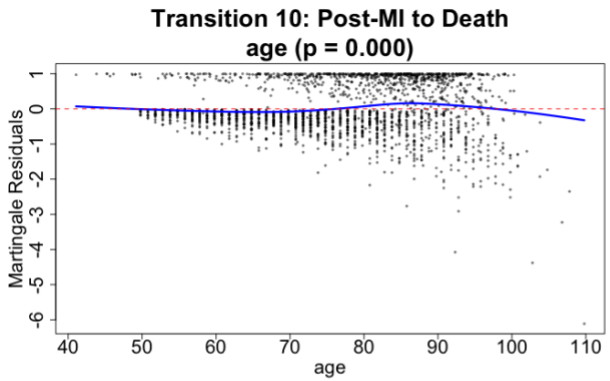

Transition 9: MI to Death
triglycerides (p = 0.000)
Martingale Residuals
triglycerides

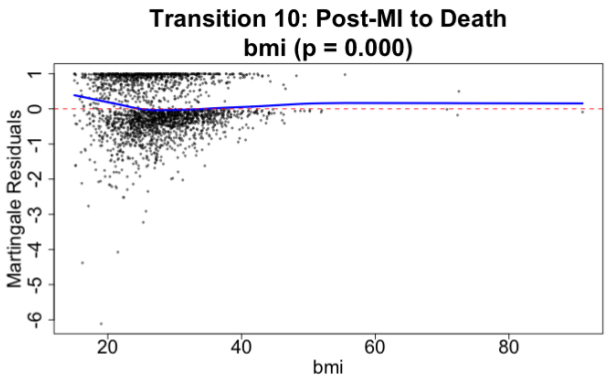

Transition 9: MI to Death
cholesterol (p = 0.000)
Martingale Residuals
cholesterol

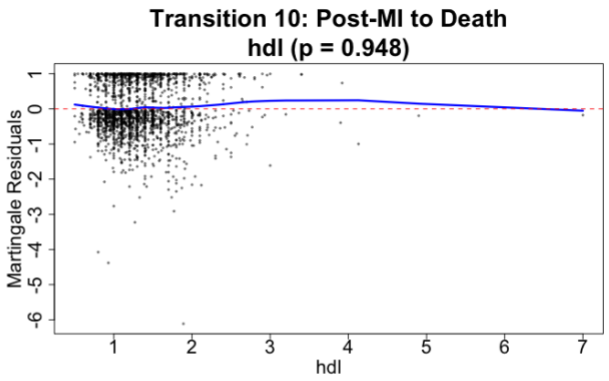

Transition 9: MI to Death
glucose (p = 0.002)
Martingale Residuals
glucose

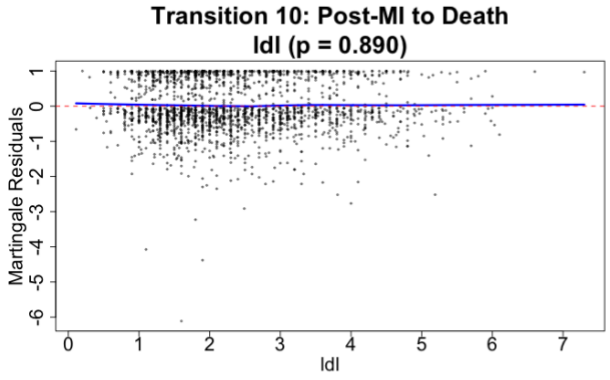

Transition 9: MI to Death
sbp (p = 0.000)
Martingale Residuals
sbp

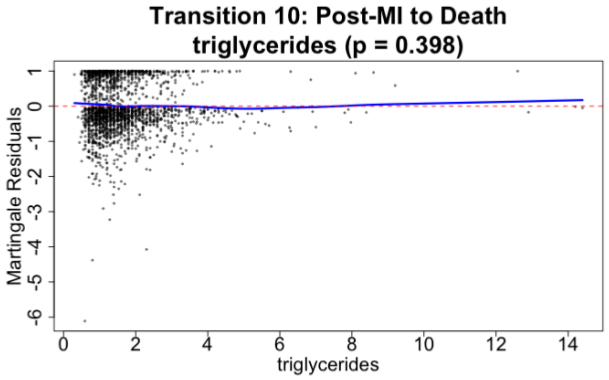

Transition 9: MI to Death
dbp (p = 0.003)
Martingale Residuals
dbp

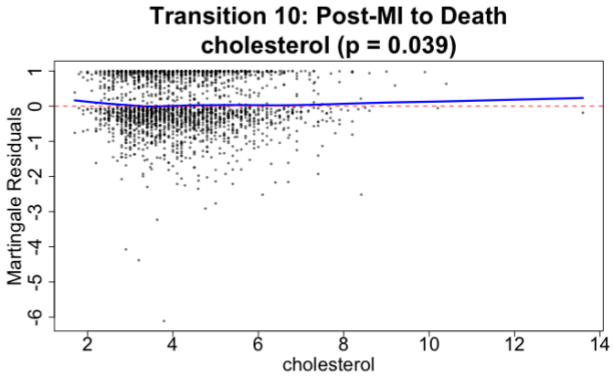

Transition 10: Post-MI to Death
age (p = 0.000)
Martingale Residuals
age

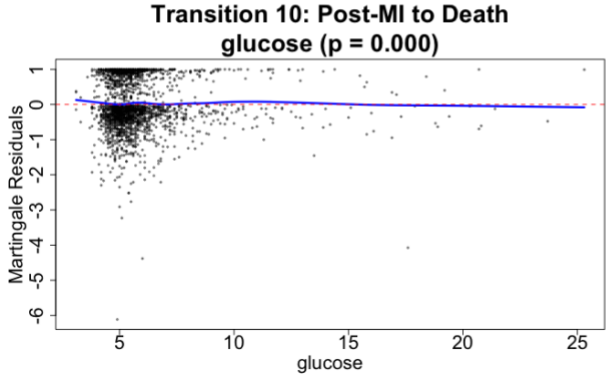

Transition 10: Post-MI to Death
bmi (p = 0.000)
Martingale Residuals
bmi

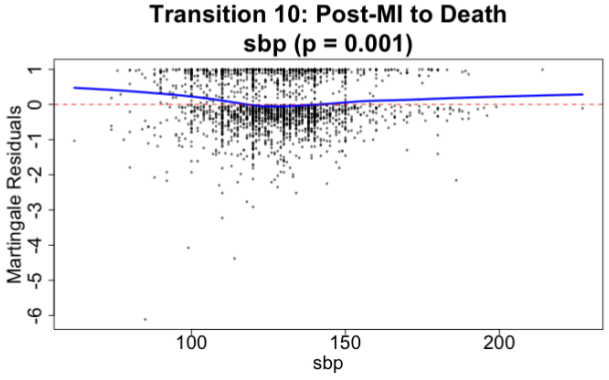

Transition 10: Post-MI to Death
hdl (p = 0.948)
Martingale Residuals
hdl

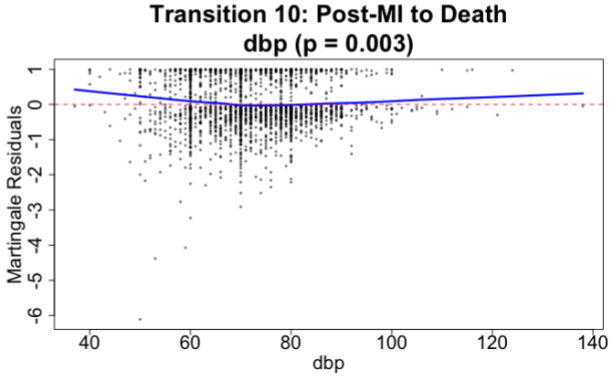

Transition 10: Post-MI to Death
ldl (p = 0.890)
Martingale Residuals
ldl

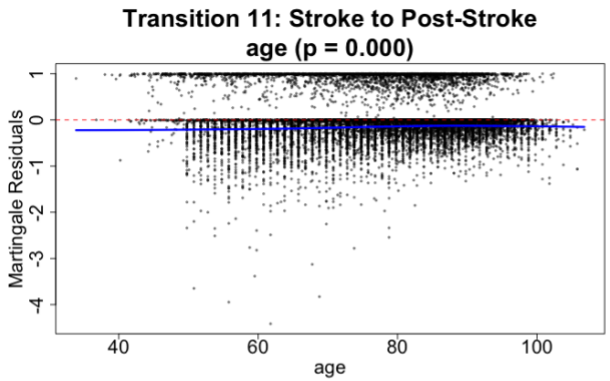

Transition 10: Post-MI to Death
triglycerides (p = 0.398)
Martingale Residuals
triglycerides

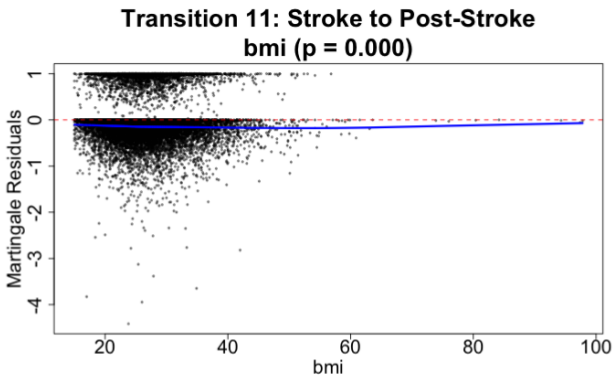

Transition 10: Post-MI to Death
cholesterol (p = 0.039)
Martingale Residuals
cholesterol

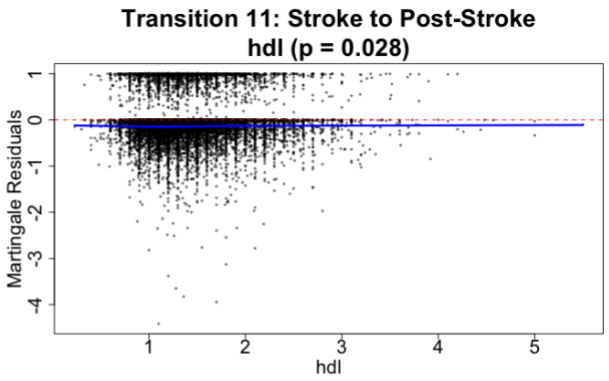

Transition 10: Post-MI to Death
glucose (p = 0.000)
Martingale Residuals
glucose

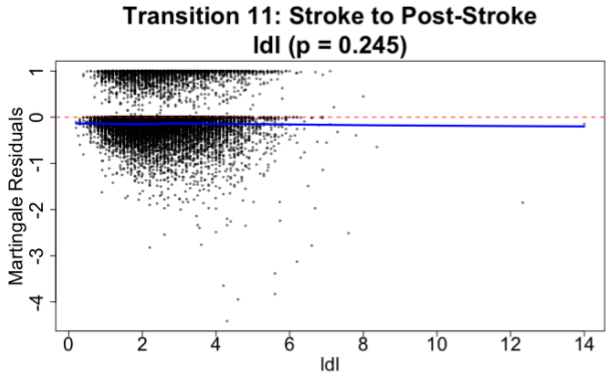

Transition 10: Post-MI to Death
sbp (p = 0.001)
Martingale Residuals
sbp

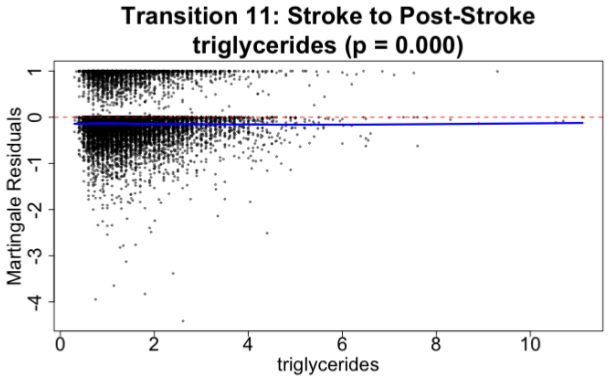

Transition 10: Post-MI to Death
dbp (p = 0.003)
Martingale Residuals
dbp

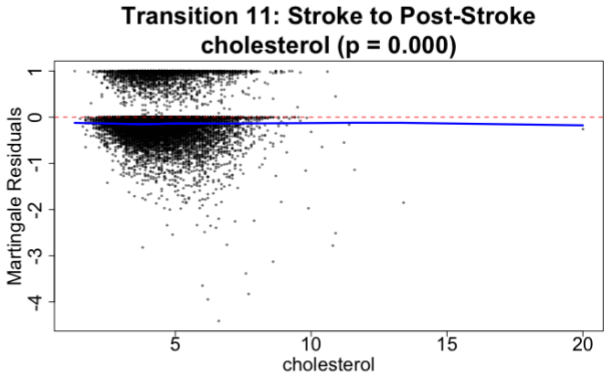

Transition 11: Stroke to Post-Stroke
age (p = 0.000)
Martingale Residuals
age

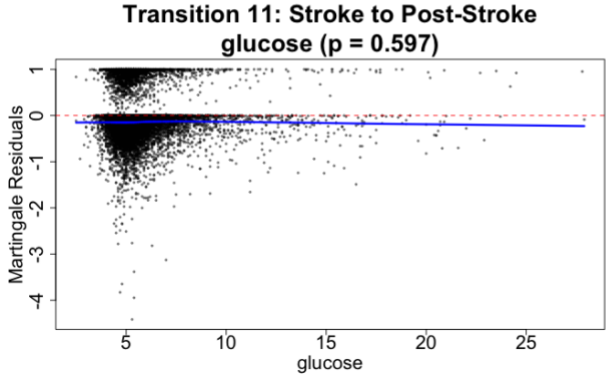

Transition 11: Stroke to Post-Stroke
bmi (p = 0.000)
Martingale Residuals
bmi

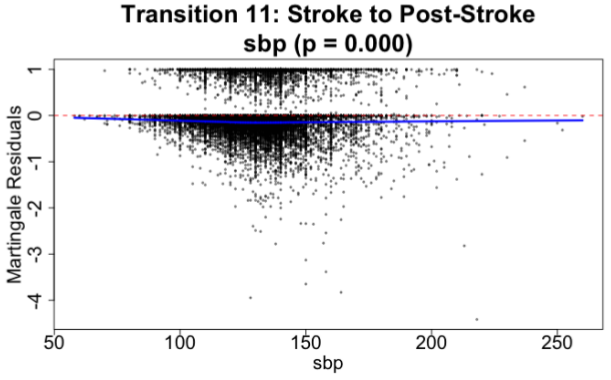

Transition 11: Stroke to Post-Stroke
hdl (p = 0.028)
Martingale Residuals
hdl

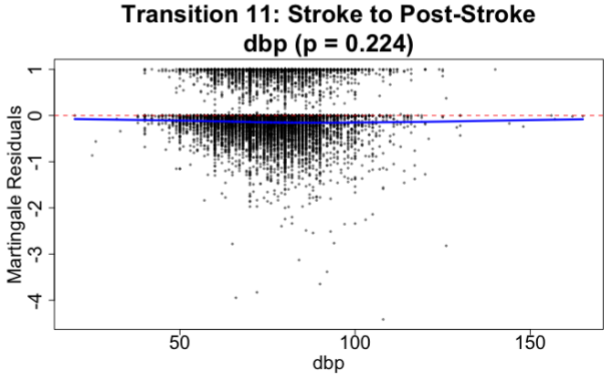

Transition 11: Stroke to Post-Stroke
ldl (p = 0.245)
Martingale Residuals
ldl

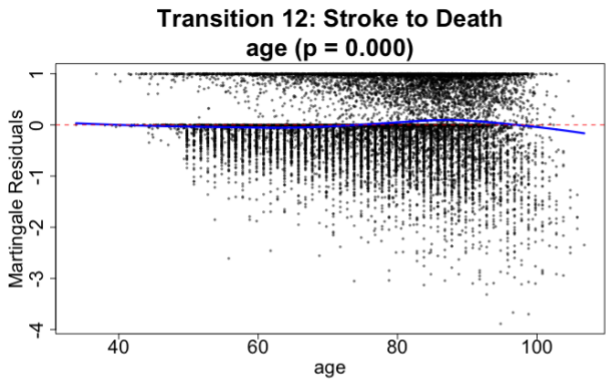

Transition 11: Stroke to Post-Stroke
triglycerides (p = 0.000)
Martingale Residuals
triglycerides

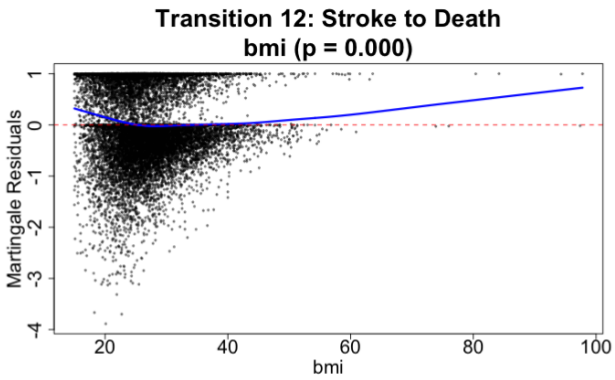

Transition 11: Stroke to Post-Stroke
cholesterol (p = 0.000)
Martingale Residuals
cholesterol

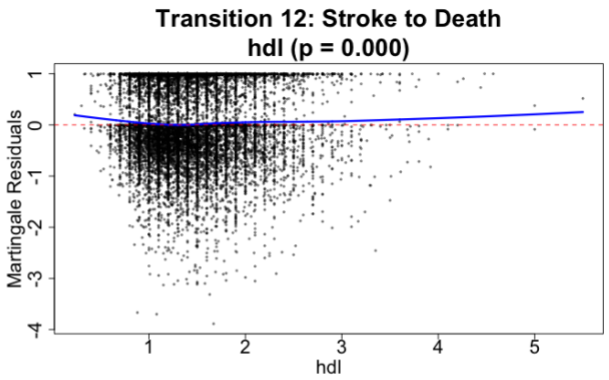

Transition 11: Stroke to Post-Stroke
glucose (p = 0.597)
Martingale Residuals
glucose

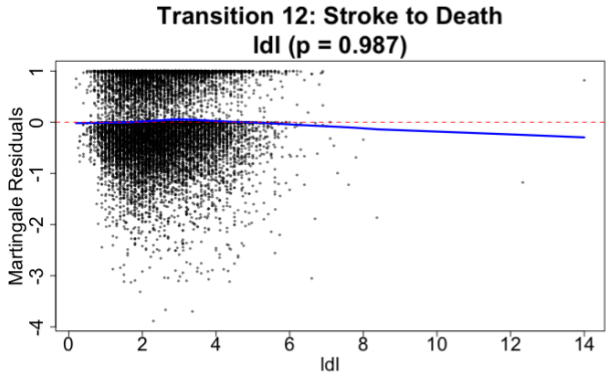

Transition 11: Stroke to Post-Stroke
sbp (p = 0.000)
Martingale Residuals
sbp

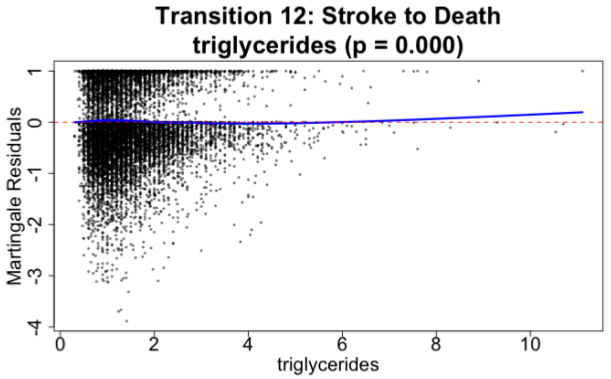

Transition 11: Stroke to Post-Stroke
dbp (p = 0.224)
Martingale Residuals
dbp

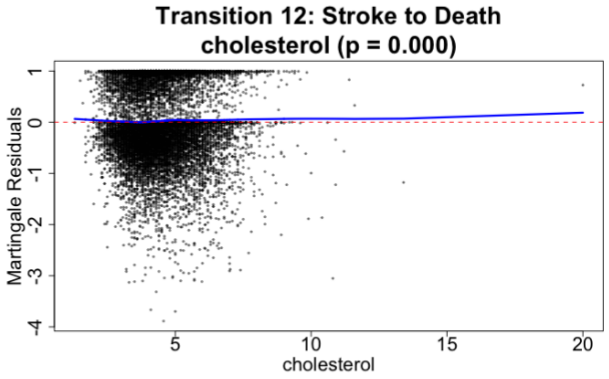

Transition 12: Stroke to Death
age (p = 0.000)
Martingale Residuals
age

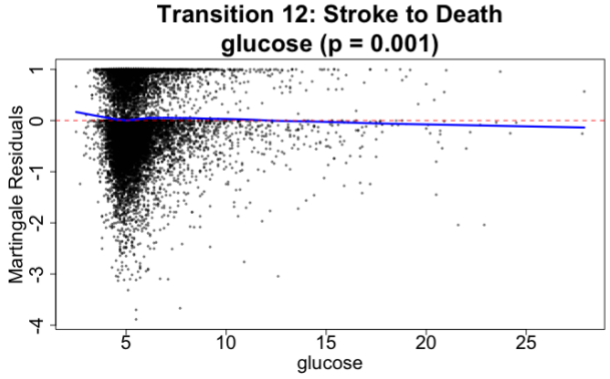

Transition 12: Stroke to Death
bmi (p = 0.000)
Martingale Residuals
bmi

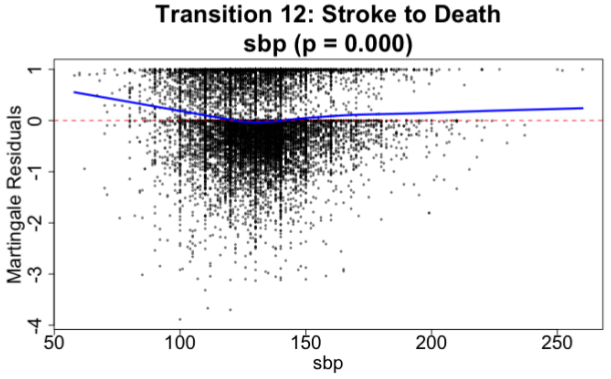

Transition 12: Stroke to Death
hdl (p = 0.000)
Martingale Residuals
hdl

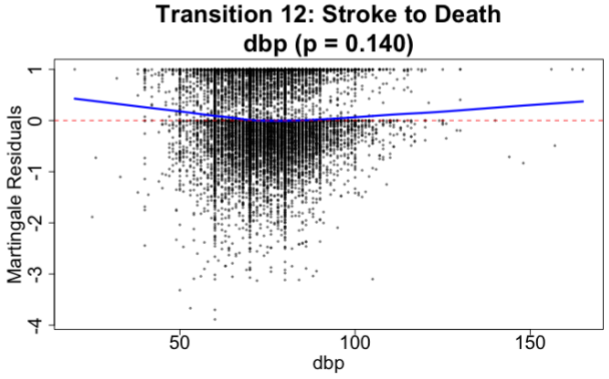

Transition 12: Stroke to Death
ldl (p = 0.987)
Martingale Residuals
ldl

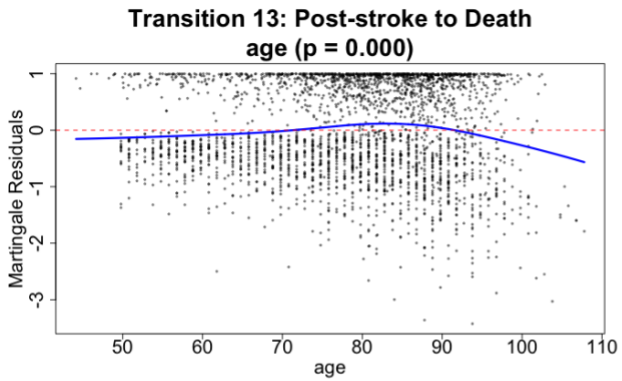

Transition 12: Stroke to Death
triglycerides (p = 0.000)
Martingale Residuals
triglycerides

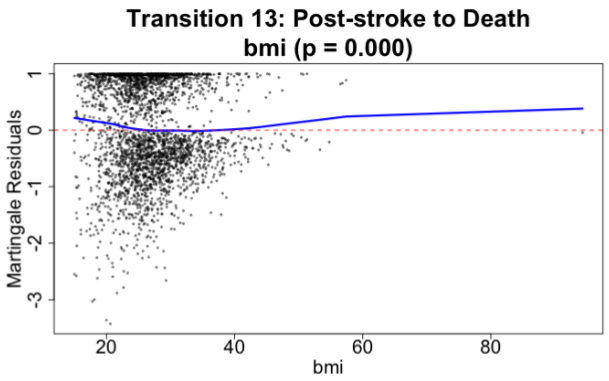

Transition 12: Stroke to Death
cholesterol (p = 0.000)
Martingale Residuals
cholesterol

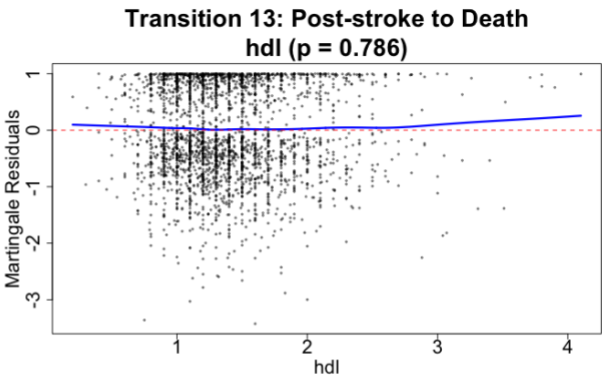

Transition 12: Stroke to Death
glucose (p = 0.001)
Martingale Residuals
glucose

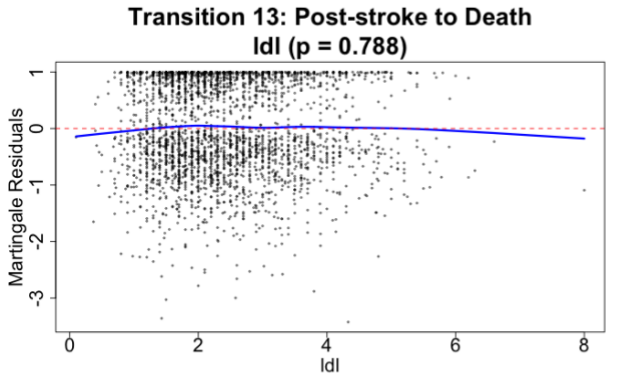

Transition 12: Stroke to Death
sbp (p = 0.000)
Martingale Residuals
sbp

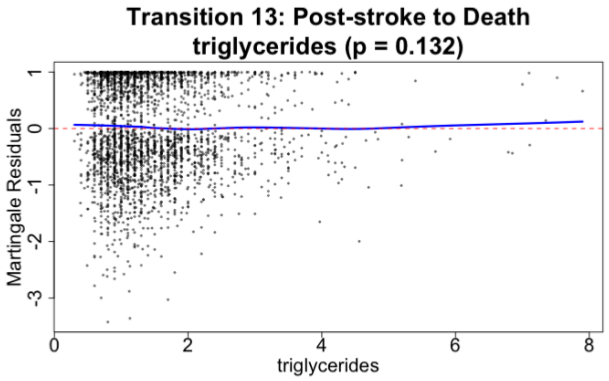

Transition 12: Stroke to Death
dbp (p = 0.140)
Martingale Residuals
dbp

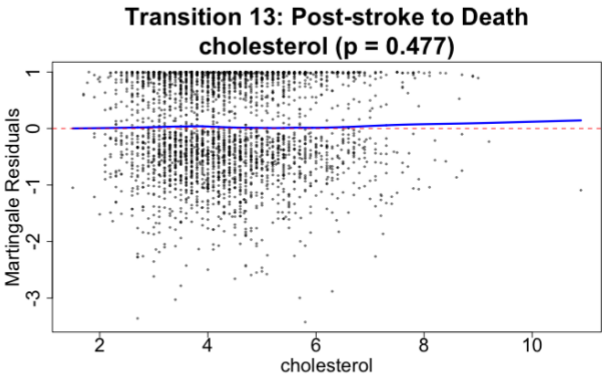

Transition 13: Post-stroke to Death
age (p = 0.000)
Martingale Residuals
age

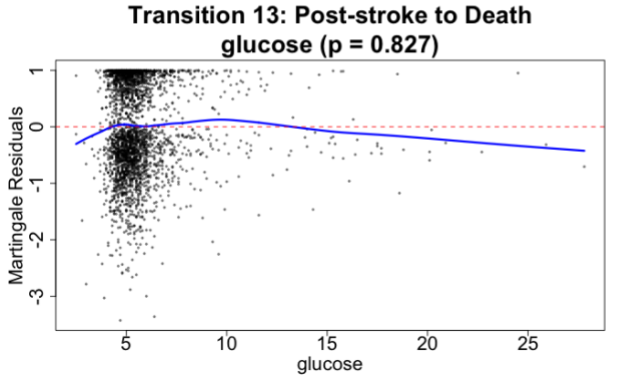

Transition 13: Post-stroke to Death
bmi (p = 0.000)
Martingale Residuals
bmi

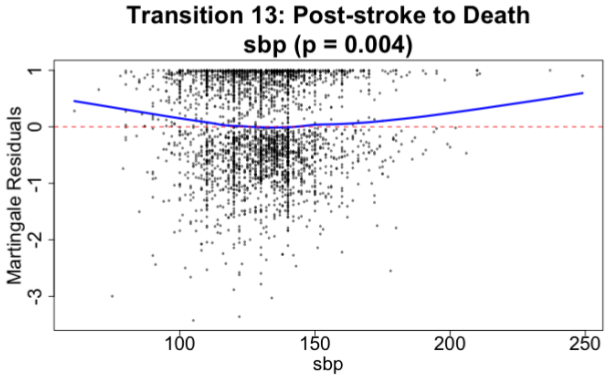

Transition 13: Post-stroke to Death
hdl (p = 0.786)
Martingale Residuals
hdl

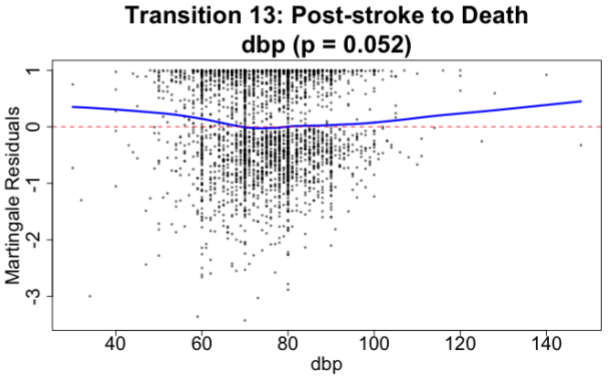

Transition 13: Post-stroke to Death
ldl (p = 0.788)
Martingale Residuals
ldl


Transition 13: Post-stroke to Death
triglycerides (p = 0.132)
Martingale Residuals
triglycerides


Transition 13: Post-stroke to Death
cholesterol (p = 0.477)
Martingale Residuals
cholesterol


Transition 13: Post-stroke to Death
glucose (p = 0.827)
Martingale Residuals
glucose


Transition 13: Post-stroke to Death
sbp (p = 0.004)
Martingale Residuals
sbp


Transition 13: Post-stroke to Death
dbp (p = 0.052)
Martingale Residuals
dbp

Appendix 8. R code for 1st case study

```
#Initial covariates values
initialCovariateValues <- c(
    startingAge = 18,
    gender = unname(genderMap["Female"]),
    ethnicity = unname(ethnicityMap["Black"]),
    deprivationIndex = 5, #most-deprived
    cvdFH = 0,
    diabetesFH = 0,
    atrialFib = 0,
    hypertension = 0,
    hyperlipidaemia = 1,
    latestSmokingStatus = unname(smokingMap["Non smoker"]),
    alcoholStatus = unname(alcoholMap["Safe alcohol"]),
    bmi = 32,
    hdl = 1.2,
    ldl = 3.5,
    triglycerides = 2.0,
    cholesterol = 7,
    glucose = 7,
    sbp = 125,
    dbp = 85
  )

  interventionCovariateValues <- initialCovariateValues
  interventionCovariateValues <-
  unlist(modifyList(as.list(initialCovariateValues), list(
  bmi = 24,
  hdl = 1.6,
  ldl = 2.5
  triglycerides = 1.4
  cholesterol = 4.8,
  glucose = 4.8,
  sbp = 115,
  dbp = 75,
  )))


#Intervention effect
covEval <- rbind(
      initialCovariateValues, # Disease-free -> Diabetes
      initialCovariateValues, # Disease-free -> MI
      initialCovariateValues, # Disease-free -> Stroke
      initialCovariateValues, # Disease-free -> Death
      interventionCovariateValues, # Diabetes -> MI
      interventionCovariateValues, # Diabetes -> Stroke
      interventionCovariateValues, # Diabetes -> Death
      interventionCovariateValues, # MI -> Post-MI
      interventionCovariateValues, # MI -> Death
```

```
      interventionCovariateValues, # Post-MI -> Death
      interventionCovariateValues, # Stroke -> Post-Stroke
      interventionCovariateValues, # Stroke -> Death
      interventionCovariateValues # Post-Stroke -> Death



#fitting model (standard vs felxible)
#parametric
dist <- cbind("logn", "logn", "logn",   "wei", "gom", "gom",
              "gom", "gam", "gam","gam", "gam", "gam", "gam")


#flexible parametric
dist <- cbind(
    "rps3", "rps1", "rps1", " rps3", "rps1", "rps2", "rps3",
    "rps3", "rps3", "rps3", "rps1", "rps2", "rps3")
```

Appendix 9. State occupancy probabilities and transition probabilities

State occupancy probabilities

age	time	Disease-free	T2DM	MI	Stroke	Post-MI	Post-Stroke	Death	sum
18	0	1.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000
18	1	1.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000
18	2	1.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000
18	3	1.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000
18	4	1.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000
18	5	1.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000
18	6	1.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000
18	7	1.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000
18	8	1.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000
18	9	1.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000
…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…
100	22259	0.000	0.001	0.002	0.001	0.001	0.000	0.995	1.000

Transition probabilities

time	T 1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13
0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.002	0.001	0.002	0.002	0.003	0.001
2	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.005	0.002	0.003	0.006	0.005	0.002
3	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.008	0.003	0.004	0.010	0.008	0.004
4	0.000	0.000	0.000	0.000	0.001	0.001	0.000	0.011	0.004	0.004	0.014	0.010	0.005
5	0.000	0.000	0.000	0.000	0.001	0.001	0.000	0.015	0.005	0.005	0.019	0.012	0.006
6	0.000	0.000	0.000	0.000	0.001	0.001	0.000	0.018	0.006	0.006	0.023	0.013	0.007
7	0.000	0.000	0.000	0.000	0.002	0.001	0.000	0.021	0.007	0.006	0.027	0.015	0.007
8	0.000	0.000	0.000	0.000	0.002	0.001	0.000	0.024	0.007	0.007	0.030	0.017	0.008
9	0.000	0.000	0.000	0.000	0.002	0.001	0.000	0.027	0.008	0.008	0.034	0.018	0.009
.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.
36498	0.9969	0.2410	0.1386	1.0000	0.3481	0.3001	0.8885	0.6047	0.5231	0.6144	0.6935	0.6418	0.5958
36499	0.9969	0.2410	0.1386	1.0000	0.3481	0.3001	0.8885	0.6047	0.5232	0.6144	0.6936	0.6418	0.5958
36500	0.9969	0.2410	0.1386	1.0000	0.3481	0.3001	0.8885	0.6047	0.5232	0.6144	0.6936	0.6418	0.5958

Appendix 10. CHEERS checklist

CHEERS 2022 Checklist

	Item	Guidance for Reporting	Reported insection
TITLE			
Title	1	Identify the study as an economic evaluation and specify theinterventions being compared.	Page 194
ABSTRACT			
Abstract	2	Provide a structured summary that highlights context, key methods,results and alternative analyses.	NA
INTRODUCTION			
Background and objectives	3	Give the context for the study, the study question and its practicalrelevance for decision making in policy or practice.	Page 194-196
METHODS			
Health economic analysis plan	4	Indicate whether a health economic analysis plan was developed and where available.	Page 196
Study population	5	Describe characteristics of the study population (such as age range, demographics, socioeconomic, or clinical characteristics).	Page 198
Setting and location	6	Provide relevant contextual information that may influence findings.	Page 198
Comparators	7	Describe the interventions or strategies being compared and why chosen.	Page 199
Perspective	8	State the perspective(s) adopted by the study and why chosen.	Page 199
Time horizon	9	State the time horizon for the study and why appropriate.	Page 199
Discount rate	10	Report the discount rate(s) and reason chosen.	Page 201
Selection of outcomes	11	Describe what outcomes were used as the measure(s) of benefit(s)and harm(s).	Page 202
Measurement of outcomes	12	Describe how outcomes used to capture benefit(s) and harm(s)were measured.	Page 202
Valuation of outcomes	13	Describe the population and methods used to measure and value outcomes.	Page 202
Measurement and valuation of resources and costs	14	Describe how costs were valued.	Page 202
Currency, price date, and conversion	15	Report the dates of the estimated resource quantities and unit costs, plus thecurrency and year of conversion.	Page 199
Rationale and description of model	16	If modelling is used, describe in detail and why used. Report if the model is publicly available and where it can be accessed.	Page 198
Analytics and assumptions	17	Describe any methods for analysing or statistically transforming data, anyextrapolation methods, and approaches for validating any model used.	NA
Characterizing heterogeneity	18	Describe any methods used for estimating how the results of the studyvary for sub-groups.	NA
Characterizing distributional effects	19	Describe how impacts are distributed across different individuals or adjustments made to reflect priority populations.	NA
Characterizing uncertainty	20	Describe methods to characterize any sources of uncertainty in the analysis.	Page 8-9
Approach to engagement with patients and others affected by the study	21	Describe any approaches to engage patients or service recipients, the general public, communities, or stakeholders (e.g., clinicians or payers) inthe design of the study.	NA
RESULTS			
Study parameters	22	Report all analytic inputs (e.g., values, ranges, references) includinguncertainty or distributional assumptions.	Page 200
Summary of main results	23	Report the mean values for the main categories of costs and outcomes ofinterest and summarise them in the most appropriate overall measure.	Page 204
Effect of uncertainty	24	Describe how uncertainty about analytic judgments, inputs, or projections affect findings. Report the effect of choice of discount rate and time horizon,if applicable.	Page 206-207
Effect of engagement with patients and othersaffected by the study	25	Report on any difference patient/service recipient, general public, community,or stakeholder involvement made to the approach or findings of the study	NA

DISCUSSION			
Study findings, limitations, generalizability, and current knowledge	26	Report key findings, limitations, ethical or equity considerations not captured,and how these could impact patients, policy, or practice.	Page 209-210
OTHER RELEVANT INFORMATION			
Source of funding	27	Describe how the study was funded and any role of the funder inthe identification, design, conduct, and reporting of the analysis	NA
Conflicts of interest	28	Report authors conflicts of interest according to journal or International Committee of Medical Journal Editors requirements.	NA

Husereau D, Drummond M, Augustovski F, de Bekker-Grob E, Briggs AH, Carswell C, Caulley L, Chaiyakunapruk N, Greenberg D, Loder E, Mauskopf J, Mullins CD, Petrou S, Pwu RF, Staniszewska S; CHEERS 2022 ISPOR Good Research Practices Task Force. Consolidated Health Economic Evaluation Reporting Standards 2022 (CHEERS 2022) Statement: Updated Reporting Guidance for HealthEconomic Evaluations. BMJ. 2022;376:e067975.