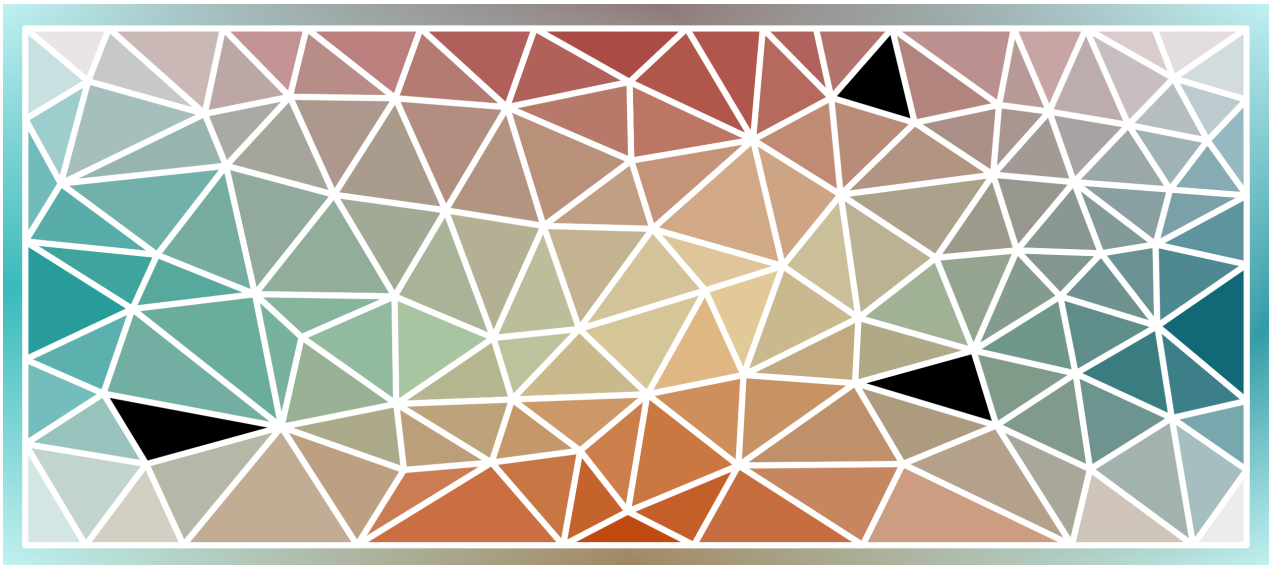
UNIVERSITY OF COPENHAGEN
FACULTY OF HEALTH AND MEDICAL SCIENCES

PhD thesis

Peter Christoffer Holm

Machine Learning-Based Precision Medicine in Ischemic Heart Disease



Primary Supervisor: Søren Brunak

Co-Supervisors: Henning Bundgaard and Karina Banasik

This thesis has been submitted to the
Graduate School of Health and Medical Sciences,
University of Copenhagen January 3, 2024

PETER CHRISTOFFER HOLM

Machine Learning-Based Precision Medicine in Ischemic Heart Disease

GRADUATE SCHOOL OF HEALTH AND MEDICAL SCIENCES
UNIVERSITY OF COPENHAGEN

CANDIDATE

Peter Christoffer Holm, MSc

Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Denmark

SUPERVISORS

Søren Brunak, PhD, Professor (principal supervisor)

Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Denmark

Henning Bundgaard, PhD, Dr.Med, Professor (co-supervisor)

Department of Cardiology, Copenhagen University Hospital, Denmark

Karina Banasik, PhD, Associate Professor (co-supervisor)

Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Denmark

PUBLISHED BY THE GRADUATE SCHOOL OF HEALTH AND MEDICAL SCIENCES UNIVERSITY OF COPENHAGEN

This document was created using the LaTeX typesetting software. The layout is based on the `tufte-latex` document class, and the main body of the text is set in ETbb and Libertine. Unless otherwise indicated, all figures and graphics in the main text are either the property of the author or are in the public domain.

Preface

This PhD thesis has been submitted to the Graduate School of Health and Medical Sciences, University of Copenhagen.

The work presented in this thesis was performed at the Novo Nordisk Foundation Center for Protein Research (CPR), University of Copenhagen, Denmark. This thesis project is part of a larger NordForsk-funded project 'Precision Diagnostics and Predictions in Ischemic Heart Disease Including Identification of Over-Treated Patients' (#90580), or 'PM-Heart', which is a interdisciplinary Nordic collaboration on precision medicine research in ischemic heart disease.

I declare no conflicts of interests.

Copenhagen, December 2023
Peter Christoffer Holm

Summary of Thesis

In modern medicine, ever increasing amounts of data is continuously being generated and recorded. Electronic health records, although not maintained for research purposes, stands as a unique and valuable source of real-world data with the potential to revolutionise precision medicine. This thesis explores the role of machine learning (ML) in extracting clinically meaningful insights from such large-scale heterogenous health data. With a primary focus on ischemic heart disease (IHD), a global leading cause of morbidity and mortality, the thesis presents three original studies under the framework of ML-based precision medicine.

In Study I, we used unsupervised clustering analysis to explore the comorbidity landscape of 72 249 patients with IHD. Using the time of the first coronary angiography or coronary computed tomography angiography as the index date, we defined multimorbidity from the entire spectrum of diagnosis codes in prior hospital records. By constructing a patient similarity network from this data, we applied the Markov cluster algorithm, a scalable graph-based clustering method, to identify 31 distinct clusters characterized by distinct patterns of multimorbidity and specific risks of subsequent outcomes. The study's findings can be used to identify knowledge gaps that exist for patient subgroups with specific patterns of multimorbidity, which are often excluded from clinical trials.

In Study II, we present the development and validation of a neural network-based survival model for prediction of all-cause mortality in patients with IHD. This model, `PMHnetV1`, was developed using a large and diverse dataset of 39 746 IHD patients from the Eastern Denmark Heart Registry (EDHR) and incorporates a comprehensive set of 584 features, including diagnosis history, procedural codes, laboratory test results, and clinical measurements. The model's performance was assessed using time-dependent area under the receiver receiver operating characteristic (td-AUC) and the Brier score, and was compared against the GRACE risk score 2.0. In the test set, `PMHnetV1` demonstrated a td-AUC of 0.88 at both six months and one year, 0.84 at three years, and 0.82 at five years, showing a notably higher performance than both GRACE2.0 and other simpler models. External validation on an independent Icelandic dataset of 8287 patients further showed that the model performance is generalizable. This study establishes `PMHnetV1` as a valuable tool for assessing risk of all-cause mortality in a real-world cohort of IHD patients, and can potentially aid clinicians in making informed decisions about patient care and interventions.

In Study III, we introduce `PMHnetV2`, an advanced iteration of our IHD prognostication algorithm. This updated version predicts three new outcomes, which includes cause-specific mortality, new ischemic events, and development of IHD complications, including heart failure and cardiac arrest. The study's key contributions are twofold: firstly, it presents a novel framework for neural network-based discrete-time models capable of modelling time-to-event data with competing risks. Secondly, it offers an updated version of our AI-driven prognostication tool, equipped to predict a broader range of disease-relevant outcomes beyond all-cause mortality. Our competing-risks framework, which can be viewed as an extension to discrete-time approach by Gensheimer and Narasimhan (2019), has been developed into the open-source Python package DiscoTime (available through PyPI or Github). The included manuscript, while still a work in progress, effectively showcases the potential and capabilities of our proposed methodology and refined models.

Dansk Resumé

I moderne medicin genereres og registreres der løbende store mængder data. Elektroniske patientjournaler er som sådan ikke designet til at understøtte forskning, men udgør ikke desto mindre en unik og særdeles værdifuld kilde til realtidsdata, der potentiel kan være med til at revolutionere præcisionsmedicin. I denne PhD-afhandling undersøger jeg hvordan maskinlæring (eng: machine learning) kan anvendes til at opnå klinisk relevant indsigt og mening fra store databaser indeholdende heterogene sundhedsdata. Med hovedfokus på iskæmisk hjertesygdom (IHD), en globalt ledende årsag til morbiditet og mortalitet, præsenterer afhandlingen tre originale studierne inden for rammerne af maskinlæringsbaseret præcisionsmedicin.

I Studie 1 anvendte vi uovervåget klyngeanalyse (eng: clustering) til at udforske sammensætningen af multimorbiditet hos 72 249 patienter med IHD. Ved at bruge tidspunktet for patienternes første koronarangiografi eller koronar-CT som indexdato, definerede vi multimorbiditet fra hele spektret af tidligere registrerede diagnosekoder fra patienternes sundhedshistorik. På baggrund af disse data, konstruerede vi et patientsimilaritetsnetværk og anvendte derefter Markov-klyngealgoritmen, en skalerbar klyngemetode til netværksdata, til at identificere 31 unikke patientundergrupper. Disse undergrupper var karakteriseret ved at have forskellige multimorbiditetsmønstre og tilhørende sygdomsrisici. Studiets resultater kan bruges til at kortlægge klinisk relevante multimorbiditetsmønstre, hvilket kan bruges til at identificere patientundergrupper, der lider under multimorbiditet, og som oftest udelades fra kliniske forsøg og derved reelt set ikke dækkes af eksisterende kliniske retningslinjer.

I Studie 2 præsenterer vi udviklingen og valideringen af en overlevelsesmodel baseret på neurale netværk til forudsigelse af overlevelsessandsynlighed hos patienter med IHD. Denne model, `PMHnetV1`, blev udviklet ved hjælp af et stort og mangfoldigt dataset af 39 746 IHD-patienter fra det Østdanske Hjerteregister og inkorporerer et omfattende sæt af 584 foreskellige variable, herunder diagnosekoder, procedurekoder, laboratorieresultater og øvrige kliniske markører. Modellens ydeevne blev vurderet ved hjælp af td-AUC og Brier-scoren og blev sammenlignet med version 2.0 af GRACE risikoalgoritment. På et internet test-datasæt demonstrerede `PMHnetV1` en td-AUC på 0.88 ved både 6 måneder og et år, 0.84 ved tre år og 0.82 ved fem år, hvilket viste en betydeligt højere præstation end både GRACE2.0 og andre mere simple modeller. Ekstern

validering på et uafhængigt Islandsk datasæt bestående af 8287 patienter viste yderligere, at modellen er generaliserbar. Dette studie underbygger `PMHnetV1` som et værdifuldt værktøj der kan anvendes til vurdere risikoen for død i hos patienter med IHD og kan potentielt på sigt hjælpe klinikere med at træffe informerede beslutninger om patientpleje og interventioner.

I Studie 3 introducerer vi `PMHnetV2`, en opdateret udgave af vores prognosticeringsalgoritme til patienter med IHD. Denne opdaterede version forudsiger tre nye endepunkter, herunder hjertesygdomsrelateret mortalitet, nye iskæmiske hændelser og udviklingen af IHD-komplikationer, herunder hjertesvigt og hjertestop. Studiets hovedbidrag er tofoldigt: for det første præsenterer det en ny metodemæssig tilgang til træning af neurale netværksbaserede overlevelsesmodeller som kan tage højde for konkurrerende endepunkter (eng: competing risks). Derudover præsenterer en opdateret version af vores prognosticeringsværktøj, som kan forudsige et bredere udvalg af sygdomsrelaterede endepunkter ud over generel mortalitet. Vores nye metodologi, som kan betragtes som en udvidelse af diskret-tidsmetoden fra Gensheimer og Narasimhan (2019), er derudover blevet til et 'open source' softwarebibliotek til programmeringssproget Python. Denne pakke, `DiscoTime` er tilgængelig via PyPI ('the Python Package-Index') og på Github. Det inkluderede manuskript, selvom det repræsenterer ikke endnu afsluttet forskningsarbejde, viser potentialet af, og mulighederne i, vores foreslåede metode og de opdaterede risikostratificeringsmodeller.

List of Manuscripts

Manuscripts included in this thesis

Manuscript for Study I

title:	'Subgrouping multimorbid patients with ischemic heart disease by means of unsupervised clustering: a cohort study of 72,249 patients defined comprehensively by diagnoses prior to presentation'
authors:	Amalie D. Haue*, Peter C. Holm*, Karina Banasik, Agnete T. Lundgaard, Victorine P. Muse, Timo Röder, David Westergaard, Piotr J. Chmura, Alex H. Christensen, Peter E. Weeke, Erik Sørensen, Ole B. V. Pedersen, Sisse R. Ostrowski, Kasper K. Iversen, Lars V. Køber, Henrik Ullum, Henning Bundgaard, and Søren Brunak
preprint:	medRxiv (2023): 10.1101/2023.03.31.23288006
status	Submitted (under revision)

An asterisk (*) denotes equal contribution. This manuscript was also included in the thesis of Amalie D. Haue.

Manuscript for Study II

title:	'Development and validation of a neural network-based survival model for mortality in ischemic heart disease'
authors:	Peter C. Holm, Amalie D. Haue, David Westergaard, Timo Röder, Karina Banasik, Vinicius Tragante, Alex H. Christensen, Laurent Thomas, Therese H. Nøst, Anne-Heidi Skogholt, Kasper K. Iversen, Frants Pedersen, Dan E. Høfsten, Ole B. Pedersen, Sisse Rye Ostrowski, Henrik Ullum, Mette N. Svendsen, Iben M. Gjødsbøl, Thorarinn Gudnason, Daníel F. Guðbjartsson, Anna Helgadottir, Kristian Hveem, Lars V. Køber, Hilma Holm, Kari Stefansson, Søren Brunak, and Henning Bundgaard:
preprint:	medRxiv (2023): 10.1101/2023.06.16.23291527v1
status	Submitted (under review, 2nd round)

An earlier version of this manuscript was also included in the thesis of Amalie D. Haue.

Manuscript for Study III

title:	'Development of a neural network-based competing risk model for individualized prognostication in ischemic heart disease using a large database of electronic health records and clinical registries'
authors:	Peter C. Holm, Søren Brunak, and Henning Bundgaard
preprint:	None
status	Work-in-progress

Manuscripts co-authored but not included in this thesis

1. Isa K. Kirk, …, Peter C. Holm, …, Søren Brunak **'Linking glycemic dysregulation in diabetes to symptoms, comorbidities, and genetics through EHR data mining '** in *eLife* (2019)

2. Ina H. Laursen, …, Peter C. Holm, …, Henrik Ullum **'Cohort profile: Copenhagen Hospital Biobank - Cardiovascular Disease Cohort (CHB-CVDC): Construction of a large-scale genetic cohort to facilitate a better understanding of heart diseases '** in *BMJ Open* (2021)

3. Amalie D. Haue, …, Peter C. Holm, …, Henning Bundgaard, and Søren Brunak **'Temporal patterns of multi-morbidity in 570157 ischemic heart disease patients: a nationwide cohort study '** in *Cardiovascular Diabetology* (2022)

4. Alex W. Jung, Peter C. Holm, …, Søren Brunak, and Moritz Gerstung **'Multi-cancer risk stratification based on national health data: a retrospective modelling and validation study '** preprint in *medRxiv* (2022)

5. Karina Banasik, …, Peter C. Holm, …, Thomas F. Hansen **'DanMAC5: a browser of aggregated sequence variants from 8,671 whole genome sequenced Danish individuals '** in *BMC Genomic Data* (2023)

6. David Westergaard, …, Peter C. Holm, …, Søren Brunak, and Henriette S. Nielsen **'Immune Changes in Pregnancy: Associations with Pre-existing Conditions and Obstetrical Complications at the 20th Gestational Week-A Prospective Cohort Study '** preprint in *medRxiv* (2023)

Contents

List of Figures

List of Tables

List of Acronyms

ACME	Automated Classification of Medical Entities
ACS	acute coronary syndromes
AI	artificial intelligence
AMI	acute myocardial infarction
ATC	Anatomical Therapeutic Chemical Classification System
AUC	area under the receiver receiver operating characteristic
BCC	Clinical Chemistry Laboratory System
BTH	BigTempHealth project
CABG	coronary artery bypass grafting
CAG	coronary angiography
CCS	Canadian Cardiovascular Society
CCTA	coronary computed tomography angiography
CDF	cumulative distribution function
CHF	cumulative hazard function
CIF	cumulative incidence function
CI	confidence interval
CPR	Danish Civil Registration System
DA	diffuse atheromatosis
DAR	Danish Register of Causes of Death
DL	deep learning
ECG	electrocardiogram
EDHR	Eastern Denmark Heart Registry
EHR	electronic health record
ESC	European Society for Cardiology
GDPR	General Data Protection Regulation
GRACE	Global Registry of Acute Coronary Events
HPO	hyperparameter optimization
HR	hazard ratio
ICD	International Classification of Diseases
ICD-10	10th revision of the International Classification of Diseases (ICD)
ICD-8	8th revision of the ICD
IFCC	International Federation of Clinical Chemistry and Laboratory Medicine
IHD	ischemic heart disease
IPA	index of prediction accuracy
IUPAC	International Union of Pure and Applied Chemistry
LABKA	Clinical Laboratory System
LSTM	long short-term memory
LOINC	Logical Observation Identifiers Names and Codes
LPR	Danish National Patient Register
LSR	Register of Pharmaceutical Sales
LVEF	left-ventricular ejection fraction
MACE	major adverse cardiovascular event
MCL	Markov clustering
MI	myocardial infarction
ML	machine learning
MLP	multilayer perceptron
NER	named entity recognition
NLP	natural language processing
NOMESCO	Nordic Medico-Statistical Committee Classification of Surgical Procedures
NPU	Nomenclature, Properties, and Units
NSTEMI	non-ST-elevation myocardial infarction
NYHA	New York Heart Association
O/E	observed over expected (i.e. O/E ratio)
OR	odds ratio
PCI	percutaneous coronary intervention

PDF	cumulative probability function
PK	pharmacokinetics
PRS	polygenic risk score
PyPI	Python Package Index
RCT	randomized clinical trial
ROC	receiver operating characteristic
RKKP	Danish Clinical Quality Program
ReLU	rectified linear unit
SGD	stochastic gradient descent
SHAP	Shapley additive explanations
SKS	Danish Medical Classification System
SNP	single-nucleotide polymorphism
STEMI	ST-elevation myocardial infarction
SVD	singular value decomposition
UA	unstable angina
XAI	explainable artificial intelligence
tanh	hyperbolic tangent
td-AUC	time-dependent area under the receiver receiver operating characteristic

Thesis Objectives and Structure

With the overall aim of furthering our knowledge on ischemic heart disease (IHD), a leading global cause of morbidity and mortality, this thesis explores the potential of machine learning (ML) in deriving clinically relevant insights from extensive electronic health data. The research primarily focuses on the development of ML-based methods for data-driven phenotyping and risk prediction, contributing to the advancement of precision medicine in secondary prevention of IHD.

Thesis Objectives

The primary objectives of the thesis are:

1. From a comprehensive database including hospital data on over 2.6 million individuals with data originating from electronic health records and national and clinical registries, extract and curate high-quality data and setup ML experiments and analyses. This includes:
   (a) Writing data-processing code to consolidate, clean, and organize heterogeneous data from various sources.
   (b) Ensuring the maintenance of robust scientific software engineering practices, including version control, workflow managers, and containerization for reproducibility.
   (c) Creating definition algorithms for the precise identification of patient populations, disease onset, and clinical outcomes.
2. Using unsupervised clustering, explore and characterise the comorbidity landscape in IHD, identifying distinct patternts of multimorbidity and their associated risk of disease progression and mortality.
3. Develop and validate clinically relevant risk-prediction algorithms for IHD using using real-world heterogenous healthcare data and right-censored time-to-event outcomes.
   (a) As a proof-of-concept, start by focusing on prediction models for all-cause mortality, and endpoint that is easy to define and where competing risks is of little concern.
   (b) Expand the prediction targets to include cardiovascular mortality and specific markers of disease progression, which requires accounting for competing risks.
4. Use explainable artificial intelligence (XAI) techniques, such as Shapley additive explanations (SHAP)-analysis, to deconvolve the ML prediction models such that it is possible to understand the decision-making process of these models and to identify the key factors influencing predictions. This will enhance the transparency and trustworthiness of the models for potential clinical application.

Structure and Scope

The thesis is written in the form of a synopsis and is as such based on three key manuscripts around which the content of thesis is centered. The following gives an overview of the structure and provides a high-level outline of the included chapters.

- In chapter 1, I provide some background on the pathophysiology and disease manifestations of IHD, motivating the role of precision medicine for improved secondary prevention in this disease.
- In chapter 2, I introduce central concepts of machine learning, with a specific focus on neural networks. While generally broad in scope, the chapter places a particular emphasis on the theory and methods used in the three main studies of the thesis.
- in chapter 3, I give a overview of survival analysis, which is an essential field of statistics utilised throughout all three studies. In addition, it introduces the relevant theory and approaches for modeling time-to-event data with neural network models. Specifically, it introduces the discrete-time logistic-hazard approach used in Study II, which we further extended upon in Study III.
- The final background chapter, chapter 4, offers a general description of the different databases and registries used in the studies, providing crucial context for understanding the data foundation of our research.
- In chapters 5 to 7, I summarise each of the three included manuscripts. The summaries outline important methodological details, highlight the main research findings, and present the main conclusions from these.
- In chapter 8, I summarise the principal findings of the thesis project, give an overview of both general and study-specific limitations, and conclude with some perspectives for future research.
- The appendices: appendix A, appendix B, and appendix C includes the three full-length scientific manuscripts that form the core of this thesis.

PART I

Background and Methods

Chapter 1

Precision Medicine in Ischemic Heart Disease

Ischemic heart disease is a term covering a variety of conditions, all caused by myocardial ischemia—an imbalance between the coronary blood supply and the oxygen requirements of the myocardium. In the overwhelming majority of cases, this imbalance can be attributed to obstructive atherosclerotic disease that limits coronary blood flow[1]. In these cases, ischemic heart disease is therefore synonymous with coronary artery disease.

The central etiological entity in ischemic heart disease is therefore the atherosclerotic plaque[2] An atherosclerotic plaque consists of blood cells, lipids, calcium and connective tissue that are gradually deposited in the arterial wall over a number of years.[3] The plaque can grow large enough to severely narrow the arterial lumen, or it can become unstable and as a consequence rupture or erode, leading to thrombosis.[4] Both of these scenarios can severely affect the perfusion of tissues and organs, and when coronary arteries are affected, it leads to ischemic heart disease.

[1] Vinay Kumar et al. *Robbins and Cotran Pathologic Basis of Disease*. Elsevier Health Sciences, 2014

[2] Kumar et al., see n. 1
[3] Peter Libby and Pierre Theroux. 'Pathophysiology of Coronary Artery Disease'. *Circulation* (2005)

[4] Valentin Fuster et al. 'The Pathogenesis of Coronary Artery Disease and the Acute Coronary Syndromes'. *New England Journal of Medicine* (1992)

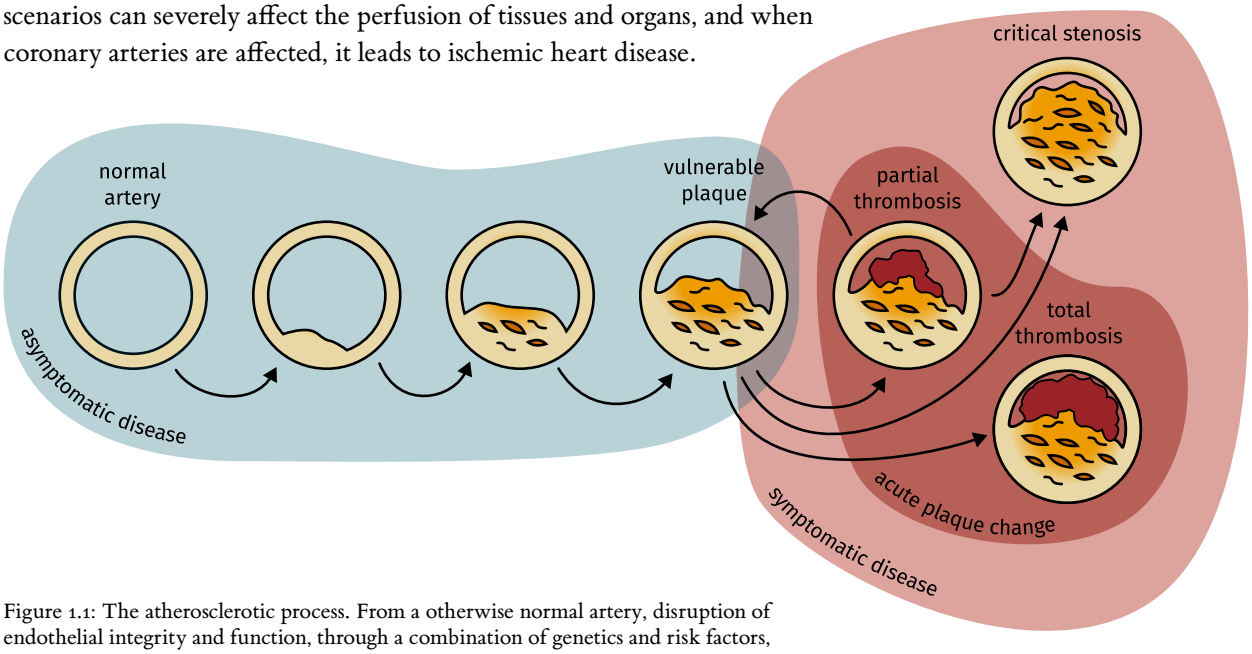


Figure 1.1: The atherosclerotic process. From a otherwise normal artery, disruption of endothelial integrity and function, through a combination of genetics and risk factors, leads to endothelial injury and low-grade inflammation. Over time, this results in plaque formation through accumulation of lipids, lipoproteins, calcium, and connective tissue. Eventually, the plaque may become 'vulnerable', making it susceptible to sudden rupture or erosion. Such an event can trigger thrombosis and acute changes in the plaque, which, depending on their severity, may either immediately obstruct the arterial lumen—leading to myocardial infarction or sudden cardiac death—or contribute to further calcification through remodeling. When the progressive narrowing of the arteries reaches a point where it causes symptoms, the stenosis is considered critical, and the myocardium experiences inadequate perfusion, manifesting as ischemic heart disease.

1.1 *Disease Manifestation*

The pathological process underlying ischemic heart disease is inherently chronic with atherosclerotic lesions gradually developing over time, but in the event of plaque rupture it can abruptly transition and manifest as an acute condition. The clinical presentation of ischemic heart disease is consequentially diverse and includes both acute[5] and chronic coronary syndromes.[6]

Acute coronary syndromes includes unstable angina (UA), ST-elevation myocardial infarction (STEMI), and non-ST-elevation myocardial infarction (NSTEMI), that collectively represents a spectrum of acute onset or progression of myocardial ischemia. If the ischemia is sufficient to cause myocardial necrosis, it is per definition called myocardial infarction (MI).[7] STEMI and NSTEMI are both forms of MI that are distinguised by a characteristic presence or absence of ST-segment elevation on a electrocardiogram (ECG). All of the acute coronary syndromes are typically associated with acute plaque change and atherothrombosis, and are medical emergencies that require immediate intervention to limit or prevent myocardial damage.[8]

Chronic coronary syndromes are more stable manifestations of the disease, and include stable angina and chronic ischemic heart disease. Stable angina, or *angina pectoris*, is characterised by episodes of crushing chest pain caused by myocardial ischemia, initially typically during exercise. By definition, the level of ischemia is not severe enough to lead to tissue necrosis.[9] Unlike UA, the symptoms of stable angina are often more predictable, reliably triggered by a specific level of physical exertion, and typically absent when the individual is at rest. Chronic ischemic heart disease can either be a long-term progression of stable ischemic heart disease or a late-stage stabilization following a MI that has undergone revascularization.[10] This condition represents the cumulative effects of prolonged myocardial ischemia and accrued myocardial damage, ultimately leading to progressive congestive heart failure.[11]

[5] Robert A Byrne et al. '2023 ESC Guidelines for the Management of Acute Coronary Syndromes'. *European Heart Journal* (2023).

[6] Juhani Knuuti et al. '2019 ESC Guidelines for the Diagnosis and Management of Chronic Coronary Syndromes'. *European Heart Journal* (2020).

[7] Kristian Thygesen et al. 'Fourth Universal Definition of Myocardial Infarction (2018)'. *European Heart Journal* (2019).

[8] Kumar et al., see n. 1.

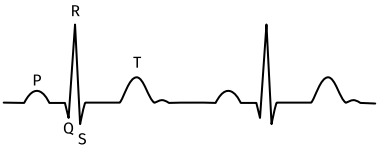


Figure 1.2: Schematic of a normal sinus rhythm as seen from an ECG. In STEMI, the ST-segment is found elevated.

[9] Knuuti et al., see n. 6.

[10] Knuuti et al., see n. 6.

[11] Kumar et al., see n. 1.

1.2 *Diagnosis and Treatment*

Since the 1950s, a series of groundbreaking scientific advances have improved our understanding and management of cardiovascular disease, leading to a drastic decline of mortality in ischemic heart disease.[12] These advances span from from diagnostic imaging techniques to pharmacological interventions and surgical procedures, each contributing to a more nuanced understanding of the disease and more effective treatment options.[13] To translate this constantly evolving body of knowledge into actionable medical practice, the European Society of Cardiology (ESC) annually releases comprehensive clinical practice guidelines that cover a wide array of cardiovascular conditions.

In line with ESC guidelines, invasive management is the recommended approach for immediate treatment of acute coronary syndromes. This includes primary percutaneous coronary intervention (PCI)[14] for STEMI and emergency angiography, potentially with concurrent PCI,

[12] Elizabeth G. Nabel and Eugene Braunwald. 'A Tale of Coronary Artery Disease and Myocardial Infarction'. *New England Journal of Medicine* (2012).

[13] Nabel and Braunwald, see n. 12.

[14] PCI is a minimally invasive procedure used for treatment of atherosclerosis. It involves the use of a small balloon catheter to widen flow-limiting stenoses and restore cardiac perfusion.

for patients with very high-risk NSTEMI and UA. For those with a more stable presentation, but still at high-risk, angiography within the first 24 hours is indicated to assess the need for revascularization.[15] Based on factors such as the coronary anatomy, number and grade of stenotic vessels, and the estimated risk of surgical complications coronary artery bypass grafting (CABG) is sometimes preferred over PCI.[16] The primary objective of these interventions is providing timely revascularization where needed, to restore coronary perfusion and limit myocardial damage.

Contrastingly, for chronic coronary syndromes, invasive coronary angiography is typically not the first-line diagnostic test.[17] However, for patients with a high clinical likelihood of ischemic heart disease, who present with easily induceable or refractory angina pectoris and a high event risk, coronary angiography is recommended for assessment of revascularization options.[18]

Concurrent with invasive treatment, the guideline gives a class I recommendation for initiation of antithrombotic therapy in patients with ischemic heart disease (IHD).[19] While the specifics of this therapy is beyond the scope of this thesis, it generally involves a combination of antiplatelet medications, aimed at preventing further thrombotic events.[20]

[15] Byrne et al., see n. 5.

[16] Franz-Josef Neumann et al. '2018 ESC/EACTS Guidelines on Myocardial Revascularization'. *European Heart Journal* (2019).

[17] Knuuti et al., see n. 6.

[18] Knuuti et al., see n. 6.

[19] Byrne et al., see n. 5.

[20] Nabel and Braunwald, see n. 12.

1.3 *Secondary Prevention*

Primary prevention aims to prevent disease onset, often through healthy lifestyle choices.
Secondary prevention focuses on managing an existing condition to prevent complications and progession.

Once the acute phase of the disease has been stabilized through revascularization or other treatments, the clinical focus shifts to long-term management and secondary prevention. Patients with established ischemic heart disease are generally of high risk of subsequent events, particularly if risk factors are not adequately managed.[21] Guidelines advocate for a multifaceted approach to secondary prevention, incorporating lifestyle changes like quitting smoking and starting exercise, as well as pharmacological interventions such as lipid-lowering therapies.[22]

The goal of long-term treatment is essentially twofold: to limit the progression of existing atherosclerotic plaque and to prevent and limit thrombus formation if plaques should rupture or erode.[23] The clinical trajectories of patients with chronic manifestations of ischemic heart disease can remain stable for several years, before unexpectedly deteriorating to major adverse cardiovascular events.[24] Consequently, continuous monitoring and risk factor control are recommended to guide secondary prevention therapy.

In this setting, risk stratification models that combine the clinical characteristics and risk factor profiles of the individual patient can be useful.[25] These models could help identify at-risk patients most likely to benefit from aggressive therapy. However, the real-world application of such models is not without challenges. There is a need for rigorous validation studies to establish their clinical utility, as well as the development of protocols for their integration into routine clinical practice.

Additionally, there is a gap in understanding how comorbidities—whether cardiovascular or non-cardiovascular—affect outcomes and

[21] Alexander M. Clark et al. 'Meta-Analysis: Secondary Prevention Programs for Patients with Coronary Artery Disease'. *Annals of Internal Medicine* (2005).

[22] Frank L J Visseren et al. '2021 ESC Guidelines on Cardiovascular Disease Prevention in Clinical Practice'. *European Heart Journal* (2021).

[23] Keith A. A. Fox et al. 'The Myth of 'Stable' Coronary Artery Disease'. *Nature Reviews Cardiology* (2020).

[24] Fox et al., see n. 23.

[25] Visseren et al., see n. 22.

Gene	Variant	Drug	Description	Reference
ABCG2	rs2231142	rosuvastatin	Genotypes GT and TT is associated with increased plasma concentrations of rosuvastatin compared to genotype GG.	1451666660
CYP2C9	*2 + *3	fluvastatin	Heterozygous or homozygous mutant allele carriers (*2 or *3) have an increased plasma concentration of fluvastatin.	1451666740
CYP2C9	*2 + *3	fluvastatin	Heterozygous or homozygous mutant allele carriers (*2 or *3) have an increased likelihood of adverse events when treated with fluvastatin compared to CYP2C9 *1/*1.	1451678600
SLC01B1	rs4149056	rosuvastatin, simvastatin, pravastatin, lovastatin, fluvastatin, atorvastatin	Genotypes CC and CT is associated with an increased risk of myopathy when treated with either of the listed statins compared to genotype GG.	1451357200, 1451244720, 1451244740, 1043880818, 655384011, 1451465324

Table 1.1: Examples of pharmacogenomics variants related to statin treatment from the PharmGKB database. The included variants are all Level 1A clinical annotations, which specify combinations of genetic variants and drugs for which there is targeted prescribing guidance available either in current clinical guidelines or in FDA-approved drug labels. Additionally, the Level 1A annotations are required to have a minimum of one supporting published article, in addition to the variant-specific recommendations.

should be factored into treatment planning and prognosis.[26] This is where the role of precision medicine becomes particularly salient. By leveraging advanced computational techniques and large-scale clinical data, precision medicine has the potential to fill these gaps.

1.4 *Perspectives of Precision Medicine*

A 2013 Cochrane review on 'Statins for the Primary Prevention of Cardiovascular Disease' concluded that statins effectively lower all-cause mortality and reduce the incidence of both fatal and non-fatal cardiovascular events without any serious adverse effects.[27] For instance, the relative risk of fatal cardiovascular events when using statins as opposed to placebo was estimated to 0.82 with a 95% confidence interval (CI) of 0.70 to 0.96.[28] This suggests those treated with statins, on average, are 18% less likely to die from cardiovascular causes. However, it is important to emphasize that this 18% reduction is an average effect for the 'average individual'. There might be groups of people for whom the reduction could be even higher, while others may see little to no benefit. Understanding and making use of such individual variability is the central objective of 'precision medicine'.

Precision medicine is broadly speaking an approach to healthcare that aims to tailor medical treatment and management to the individual patient. Instead of employing a 'one-size-fits-all' methodology, where treatment and preventive care is being developed to the average patient, precision medicine uses data-driven approaches to also account for the specific factors of the individual. As such, the underlying idea of precision medicine is not really new; tissue- and blood typing, for example, has been used to guide organ and blood donation for several decades. However, the prospective of leveraging large clinical databases and broad array of phenotypic information is adding renewed interest in the concept.[29]

In the context of statin treatment, this class of medication is generally both highly effective and very well-tolerated, with limited side effects.[30] Some individuals may experience mild side effects such as muscle aches, while more severe side effects like myopathy or rhabdomyolysis are

[26] Visseren et al., see n. 22

[27] Fiona Taylor et al. 'Statins for the Primary Prevention of Cardiovascular Disease'. *Cochrane Database of Systematic Reviews* (2013).

[28] Taylor et al., see n. 27.

[29] Francis S. Collins and Harold Varmus. 'A New Initiative on Precision Medicine'. *New England Journal of Medicine* (2015).

[30] Taylor et al., see n. 27.

exceedingly rare.[31] The underlying mechanism of statin-induced skeletal muscle side effects is not well defined, but appears to be connected with the drug's serum concentration.[32] Pharmacogenomics, a critical component of the pharmacological aspects of precision medicine, adds another layer to this understanding. It explores how genetic factors influence drug response, allowing for understanding etiology and more nuanced treatment plans. Genome-wide analyses have already identified single-nucleotide polymorphisms (SNPs) that are associated with an increased risk of statin-induced side effects (Table 1.1).[33,34] Through targeted screening of such genetic variants, and considering concurrent medication and experiential variability of the patients, it might be possible to refine medication regimens to maximize efficacy while minimizing side effects—a practical application of precision medicine.

Pharmacogenomics is an important step on the way, but as such it is only one of the 'building blocks' of precision medicine. To fully realize the promise of precision medicine, we need a more holistic approach that consider not only genetics but the multitude of other factors that can and will influence the ideal course of treatment. Figure 1.3 shows a schematic of how such a precision cardiology workflow could look like.

[31] Paul D. Thompson et al. 'Statin-Associated Myopathy'. *JAMA* (2003).

[32] Thompson et al., see n. 31.

[33] SEARCH Collaborative Group et al. 'SLCO1B1 Variants and Statin-Induced Myopathy—a Genomewide Study'. *The New England Journal of Medicine* (2008).

[34] Caroline F. Thorn et al. 'PharmGKB: The Pharmacogenomics Knowledge Base'. *Pharmacogenomics: Methods and Protocols*. Humana Press, 2013.

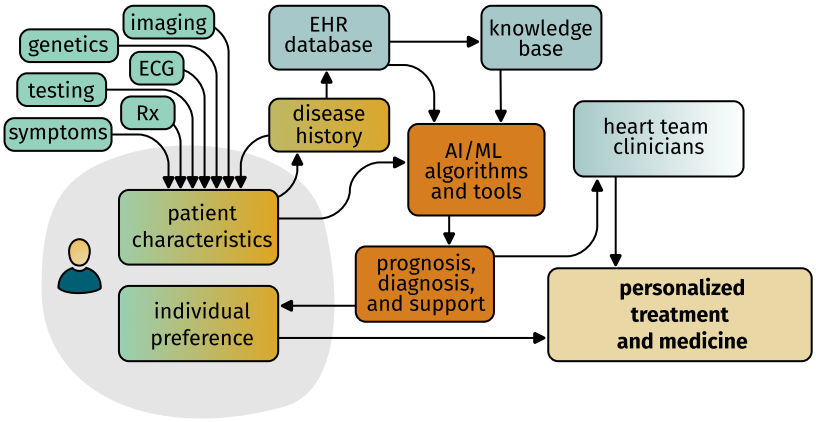


Figure 1.3: Schematic of a precision cardiology workflow. Patient characteristics encompass the full range of available data, including the individual's entire health and disease history. This information resides in an electronic health record (EHR) database, which also holds data from other patients. The EHR database is a valuable research resource, that can contribute to a curated clinical knowledge base. Patient data, the EHR database, and the knowledge base is used as input by computational algorithms and tools. These tools act as clinical decision support systems, providing prognostic and diagnostic models. The outputs of these models are shared with both the patient and the physician, facilitating personalized treatment through shared decision-making.

The central elements of this precision cardiology workflow is: (i) big data collection and organization; (ii) computer-based clinical tools and algorithms; and (iii) implementation of these tools in clinical decision making. In the research underlying this thesis, the primary emphasis have been on (i) and (ii), but without (iii) the advances of precision medicine will remain mostly academic in scope. The challenges and future perspectives of the clinical implementation will be discussed later in the thesis. Focusing on the first two elements, how do we derive tools and algorithms from big datasets?

Chapter 2

Fundamentals of Machine Learning and Neural Networks

In modern medicine, ever increasing amounts of data is continuously being generated and collected. Ranging from structured administrative data used primarily for billing purposes to advanced imaging and high-througput 'omics' analyses, the array of available data is as diverse as it is plentiful. Making sense and making use of such massive amounts of data necessitates automated methods for data analysis.

Artificial intelligence (AI), and specificically machine learning (ML) (fig. 2.1), seeks to address this need by development of methods and algorithms that allows computers to 'learn' from data to solve new problems instead of them being explicitly programmed.[1] The field of ML have progressed exponentially over the past couple of decades, and with the advent of generative AI models like GPT-4[2] and DALL-E[3] there is a growing public interest in the use of AI and ML. In the context of precision medicine, the promise of AI lies in the ability to integrate large amounts of data from huge data sets and register and highlight patterns with clinical importance. In his review on artificial intelligence in medicine,[4] Eric Topol expressed his view that in the not so distant future 'almost every type of clinician, ranging from specialty doctor to paramedic, will be using AI technology, and in particular deep learning'.

The underlying principle of ML is to formulate a learning problem with a well-defined objective and a quantifiable measure of performance.[5] Subsequently, a loosely defined computer program is established and fed with data representing said objective. Guided by the performance metric, ML algorithms iteratively refine the underlying computer program until the objective is optimally addressed. In his 1997 book *Machine Learning*, Mitchell defines this formally as:

> A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.[6]

Though this conceptual framework is common to all ML models, the diversity in ML arises from not only the specific ML algorithm, but also choices regarding objectives, performance metrics, and program attributes. These selections introduce the myriad variations and nuances within ML, which can be broadly categorized into two distinct approaches: supervised learning and unsupervised learning.[7]

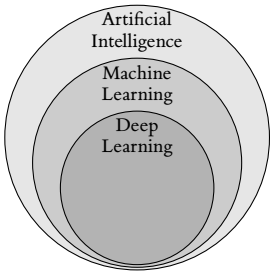


Figure 2.1: Artificial intelligence (AI) is an umbrella term for computer software that have some sort of 'intelligence' and machine learning (ML) is a collection of methods through which that can be achieved. Deep learning (DL), a specific method in ML, relies on neural networks with many layers to analyse complex data.

[1] Ian Goodfellow et al. *Deep Learning*. MIT press, 2016.

[2] OpenAI. 'GPT-4 Technical Report'. *arXiv* (2023). URL: arxiv.org/abs/2303.08774. preprint.

[3] Aditya Ramesh et al. 'Zero-Shot Text-to-Image Generation'. *arXiv* (2021). URL: arxiv.org/abs/2102.12092. preprint.

[4] Eric J. Topol. 'High-Performance Medicine: The Convergence of Human and Artificial Intelligence'. *Nature Medicine* (2019).

[5] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[6] Tom M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[7] Murphy, see n. 5.

2.1 *Supervised Learning*

In supervised learning, models are trained on labeled examples—a dataset $\mathfrak{D} = \{(\boldsymbol{x}_i, y_i) \mid i \in \{1, \ldots, N\}\}$ of size N that contains both input features $\boldsymbol{x}$ and corresponding correct output values y for all the examples in the dataset. Here $\mathfrak{D}$ is typically refered to as the training set.[8] The primary aim in supervised learning is to learn a function f that correctly maps input data $\boldsymbol{x}$ to output data y, i.e. correctly assigns output labels or values based on the features present in the input data. When the output is discrete labels or classes, the task is called a classification problem. Conversely, if the output is continuous values, it is called a regression problem.

Illustrating with a supervised learning task in the domain of classification, we can consider a database of coronary angiography images that have been manually annotated to indicate the presence or absence of critical stenosis in any of the coronary arteries. In this scenario, the output classes are denoted as $y \in \{\text{stenosis}, \text{no stenosis}\}$, and the objective is to classify the images into either of these categories based on their pixel values. The labeled data is being used to guide and 'supervise' the model in order for it accurately accomplish this categorisation.

Two other examples of supervised learning tasks is presented in Study II and Study III within this thesis. However, these tasks differ from classical supervised learning in that they involve time-to-event predictions and censored labels. This subject matter will be elaborated upon in chapter 3.

[8] Murphy, see n. 5.

2.2 *Unsupervised Learning*

In contrast to supervised learning, unsupervised learning is concerned with finding underlying patterns or structures within unlabeled datasets. In this paradigm, the algorithm operates without the aid of predetermined labels or categories, instead learning from the data itself. The aim is to discover intrinsic structure in the dataset, which can then be used for tasks such as dimensionality reduction, clustering, or anomaly detection.[9]

A canonical example of unsupervised learning is the problem of clustering, where the objective is to partition a set of objects into subgroups based on similarity.[10] Things that are similar should be grouped together and should be relatively dissimilar to things in other groups. Defining what constitutes 'similar' is therefore a central challenge in clustering; different measures of similarity often result in fundamentally different clusterings.[11] Clustering lies at the heart of precision medicine: By identifying distinct subgroups of patients with varying risk profiles, clinicians can tailor prevention and treatment strategies more effectively, optimizing healthcare outcomes as a result.

In Study I of this thesis, we present an example of clustering analysis of patients with ischemic heart disease by considering the patterns of comorbidity common in subgroups of patients. The specific methods used in this work is outlined in chapter 5.

[9] Murphy, see n. 5

[10] Murphy, see n. 5

[11] To illustrate, we can consider a set of playing cards, that for convenience is limited to aces, court cards, and tens. One possible clustering groups the cards by suit:

$$\begin{array}{lllll} \{\{10\heartsuit, & J\heartsuit, & Q\heartsuit, & K\heartsuit, & A\heartsuit\}, \\ \{10\spadesuit, & J\spadesuit, & Q\spadesuit, & K\spadesuit, & A\spadesuit\}, \\ \{10\diamondsuit, & J\diamondsuit, & Q\diamondsuit, & K\diamondsuit, & A\diamondsuit\}, \\ \{10\clubsuit, & J\clubsuit, & Q\clubsuit, & K\clubsuit, & A\clubsuit\}\} \end{array}$$

Another equally valid clustering groups them by rank:

$$\begin{array}{llll} \{\{10\heartsuit, & 10\spadesuit, & 10\diamondsuit, & 10\clubsuit\}, \\ \{J\heartsuit, & J\spadesuit, & J\diamondsuit, & J\clubsuit\}, \\ \{Q\heartsuit, & Q\spadesuit, & Q\diamondsuit, & Q\clubsuit\}, \\ \{K\heartsuit, & K\spadesuit, & K\diamondsuit, & K\clubsuit\}, \\ \{A\heartsuit, & A\spadesuit, & A\diamondsuit, & A\clubsuit\}\} \end{array}$$

The choice between these clusterings depends on whether suits or ranks are considered more important, which probably depends on the specific card game in question. This challenge applies to most clustering problems—the ideal clustering is usually highly context dependent.

2.3 *Generalization and Overfitting*

Returning to supervised learning, it is important to note that achieving good performance on the training set is not the sole objective. For a ML model to be of utility, it should maintain its accuracy when applied to unseen data. This concept is known as *generalization* and is a central problem in supervised learning—especially when dealing with highly flexible models such as neural networks.[12] We can keep track of a model's generalization error by introducing an additional dataset, that is kept separate from the training set. This additional set of labeled examples is refered to as the test set and is exclusively used for evaluation of model performance. Assuming that the test set is representative[13], the performance on the test set can be used as an estimate of the generalization error since it represents unseen data.

Figure 2.2 shows a theoretical training history of a neural network model where the performance is calculated on both a training and a test set after each iteration of the training. Here it is illustrated that the training loss is monotonically decreasing with increasing number of iterations. The validation loss is at first also decreasing, but if training continues for long enough, at some point it will start to increase instead. The divergence between training and test set performance indicates overfitting: instead of learning generalizable patterns representative of the underlying data-generating process, the model starts to learn or even memorise the noise and idiosyncracies of the data that, although characteristic in the narrow scope of the training set, would not be representative of neither biology nor disease etiology.[14] As a consequence, the test set performance is considerably worse than the training set performance.

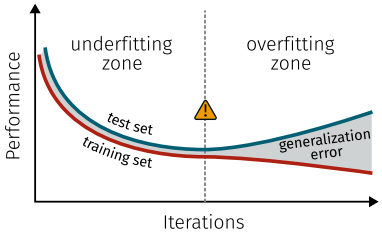


Figure 2.2: Training a neural network model for many iterations runs the risk of overfitting the model to the training data. Although the training error keeps decreasing, it happens at the expense of increased generalization error. Inspired by Goodfellow et al., see n. 1.

[12] Goodfellow et al., see n. 1.

[13] An underlying assumption is that the two datasets are independent and identically distributed (typically abbreviated as i.i.d.), and thus share the same underlying *data-generating process*. [Goodfellow et al., see n. 1]

[14] Murphy, see n. 5.

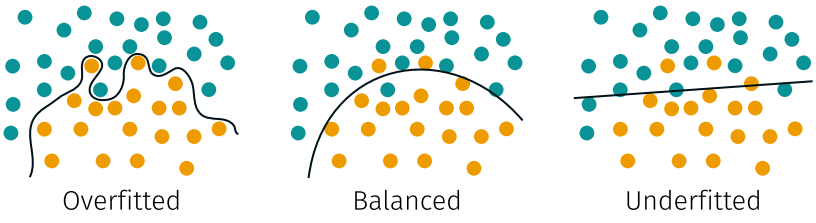


Figure 2.3: Illustration of overfitting and underfitting in a simple classification task. An overfitted model learns the training data too well, and may even remember noise and outliers, which makes it perform poorly on unseen data. Underfitted models, on the other hand, are to simple to capture meaningful patterns in the data.

Overfitting, and its counterpart, underfitting, are key considerations in training of ML models (fig. 2.3). Overfitting happens when a model learns the details of the training data too well, including the noise, which makes it perform poorly on unseen data. In constrast, underfitting occurs when the model is too simplistic to capture meaningful patterns in the data, resulting in poor performance on both the training and test sets. Both issues highlight the need for balancing the complexity of models, to ensure that they can effectively generalize to unobserved data. In the section Regularization, I will outline some of the specific methods used to balance the complexity of neural network models.

2.4 *Neural Networks*

In Study II and Study III, neural network models are utilized to develop risk prediction models for ischemic heart disease. Neural networks are a versatile class of machine learning models well-suited for handling large and heterogeneous datasets, and they currently represent the state-of-the-art in ML. The rest of this chapter will concentrate mainly on neural networks, outlining the methodological details and highlight relevant practical challenges and considerations in their implementation.

Historically, neural networks were designed using the archicteture of neurons in a human brain as inspiration.[15] The simplest model is that of a perceptron, which can be seen as a computational approximation of a real neuron or nerve cell.[16] A typical neuron has many dendrites, a cell body, and a single axon (fig. 2.4). The dendrites carries the input signal to a neuron, and if the cumulative signal is great enough[17], then the neuron will propagate an action potential down the axon.[18] In similar fashion, a perceptron may receive many different inputs and produces a single output (fig. 2.5). In the case of a neuron, the 'all-or-none' principle means that nerve cells either signals at full strength or not all. For a perceptron, this principle can be emulated with the followingly Heaviside step function:[19]

$$g_\phi(\boldsymbol{x}) = \begin{cases} 1 & \text{if } b + \boldsymbol{w} \cdot \boldsymbol{x} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

By combing several of such artificial neurons, in a multilayer-perceptron or feedforward neural network, we can create a model that, in theory, can learn even the most complex of patterns.[20] In this design (fig. 2.6), information travels in one direction, from the input layer through the hidden layers, and finally to the output layer. Each layer in a feedforward neural network is fully connected to the subsequent layer—every neuron in one layer is connected to every neuron in the next.

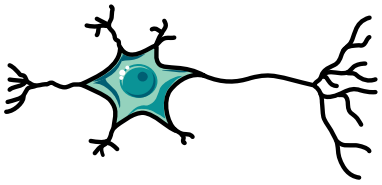


Figure 2.4: Schematic diagram of a neuron. A typical neuron has dendrites, a cell body, and a single axon; the dendrites receive input signals from other neurons, and propagates output signals along the axon.

[15] Goodfellow et al., see n. 1.

[16] Eugene Charniak. *Introduction to Deep Learning*. Illustrated. The MIT Press, 2019.

[17] This threshold is known as the *threshold potential*, and is typically between -50 and -55 mV.

[18] Julian Seifter et al. *Concepts in Medical Physiology*. Lippincott Williams & Wilkins, 2005.

[19] Charniak, see n. 16.

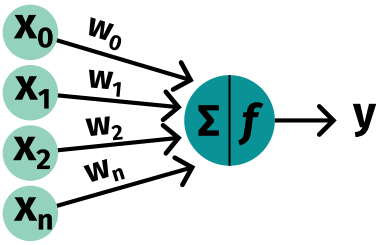


Figure 2.5: Schematic diagram of a perceptron

[20] Christopher M. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press, 1995.

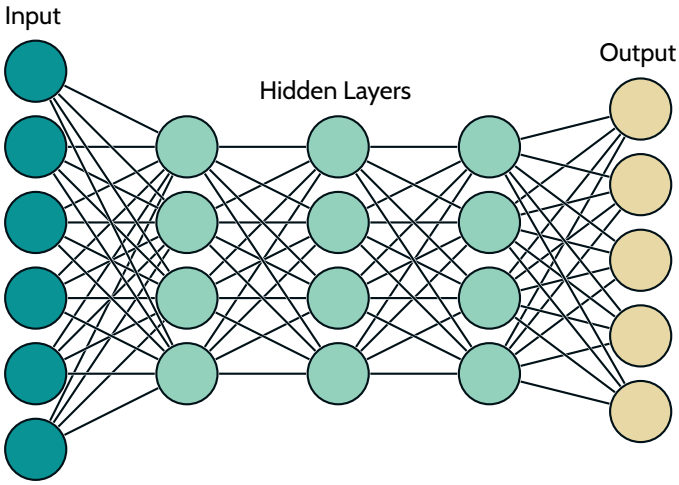


Figure 2.6: A schematic representation of a feedforward neural network, comprising an input layer, multiple hidden layers, and an output layer. Each circle denotes a neuron, and the connecting lines represent connections between neurons.

2.4.1 *Activation Functions*

In the context of modern neural networks, the simplistic step function in eq. (2.1) has certain limitations. In particular, it is non-differentiable at $x = 0$ and has a zero derivative elsewhere, rendering it incompatible with gradient-based optimization algorithms. To address these issues, other activation functions have been introduced, with the sigmoid, hyperbolic tangent (tanh), and rectified linear unit (ReLU) being popular choices (fig. 2.7), but many other variations exists.[21]

[21] Francois Chollet. *Deep Learning with Python*. 2nd ed. Manning, 2021.

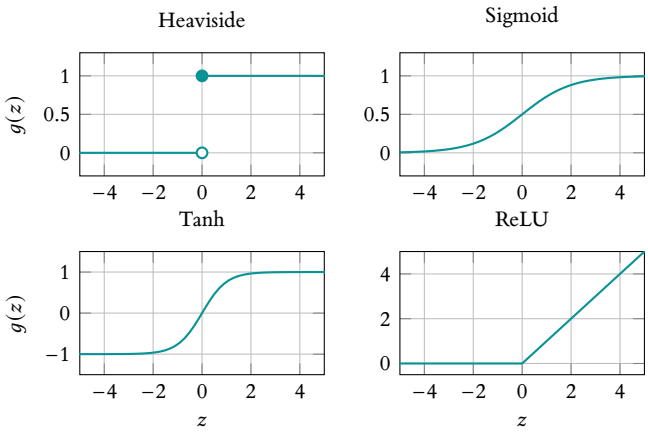


Figure 2.7: Plots of well-known activation functions used in neural networks. From top left to bottom right: The Heaviside step function, the sigmoid (or logistic) function, the hyperbolic tangent (tanh), and the rectified linear unit (ReLU). Each function is plotted against its input value z to show its respective output $g(z)$.

2.4.2 *Neural Network Architectures*

In addition to feed forward neural networks, ongoing research has developed other neural network architectures tailored to particular types of data. For example, convolutional neural networks were designed for processing image data[22] and have revolutionized the field of computer vision.[23] Similarly, recurrent neural networks, such as the long short-term memory network, can model sequential data and have had great impact for both time series analysis and natural language processing.[24] Another innovation, is the introduction of skip or residual connections in architectures such as residual networks (ResNet), which have enabled the training of exceptionally deep networks.[25]

[22] Yann LeCun et al. 'Handwritten Digit Recognition with a Back-Propagation Network'. *Advances in Neural Information Processing Systems*. 1989.

[23] Simon J.D. Prince. *Understanding Deep Learning*. MIT Press, 2023.

[24] Sepp Hochreiter and Jürgen Schmidhuber. 'Long Short-Term Memory'. *Neural Computation* (1997).

[25] Kaiming He et al. 'Deep Residual Learning for Image Recognition'. *arXiv* (2015). URL: arxiv.org/abs/1512.03385. preprint.

2.5 *Practical Implementation of Neural Networks*

In developing neural network models for ML objectives, several methodological considerations and decisions have to be made. It is not feasible to exhaustively cover every nuance within the scope of this thesis, instead the book *Deep Learning* serves as a comprehensive reference.[26] As a guide, the following list provides a high-level overview of the typical workflow:[27]

1. *Problem Definition:* Clearly describe the problem, and in the process identify if the objective belongs to classification, regression, or a third category. Attached to the objective should be a measure of performance, which in the neural network literature typically is known as loss function, which is used to direct training.
2. *Preparing the Data*:
   (a) *Data Collection:* Gather and organize data for both training and testing the model. Importantly, data should be of reasonable quality, pertinent to the problem at hand, and of sufficient quantity. If not, it can be a good idea to adjust item 1 to better reflect the available data.
   (b) *Data Preprocessing:* Prepare the data for model training, which may include cleaning, normalization, and transformation tasks. The classical saying 'garbage in, garbage out' is worth repeating here. Any data-informed tasks (e.g. normalization) should exclusively be setup using the training set to avoid leakage of data.[28]
3. *Network Structure:* Choose the architecture of the neural network, defining elements such as the number of layers, number of neurons within each layer, and activation functions. Certain architectures have shown to be useful for specific types of data, e.g. convolutional neural networks are typically the architecture of choice for computer vision tasks.[29]
4. *Configuring Model Training:*
   (a) *Optimization Algorithm:* Choose and configure an optimization algorithm. Stochastic gradient descent (SGD) and variants thereof are the most common algorithms for neural networks,[30] and all have different parameters that needs to be defined and possibly tuned (see item 5), such as e.g. the learning rate.
   (b) *Regularization:* To mitigate overfitting and ensure better generalization, consider using regularization techniques such as L1/L2-regularization or Dropout.[31]
5. *Hyperparameter Tuning:* Tweak and adjust the relevant hyperparameters specified in any of the previous steps, including learning rate (from item 4a) and specific details of the neural network architecture (item 3) to find the best configuration. This typically involves training many different intermediate versions of the model and evaluating their performance using a validation set.
6. *Model Training:* Train the final version of the model using the designated training set.
7. *Model Evaluation:* Assess the trained model using the test set, calculating metrics like accuracy, precision, and recall to gauge its effectiveness. Model explainability techniques can here be helpful in aiding the interpretation of the model.

While presented as sequential steps, the items in the list are almost all interrelated and can and should affect one another. As an example, it is especially evident that item 2a drastically influences the range of possibilities in item 1. In the remaining sections of this chapter, I will be highlighting select concepts integral to building neural network models which have specific relevance to the papers included in the thesis.

[26] Goodfellow et al., see n. 1

[27] The list draws inspiration from chapter 19.9 in Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. 3rd edition. Pearson, 2009 and chapter 4 in Chollet, see n. 21.

[28] Chollet, see n. 21.

[29] LeCun et al., see n. 22.

[30] Chollet, see n. 21.

[31] Goodfellow et al., see n. 1.

2.6 Model Selection

We can estimate the generalization error of a model by evaluating it on a test set. If we are only creating a single model, then this approach suffices. However, we might want to compare many different models, or slightly tweak an already existing model, such that we can select the best performing version. This is particularly relevant in the context of hyperparameter optimization, a topic that I will return to later in this chapter. If we select the final model based on the test set alone, we might inadvertently have biased the process, and could, in a sense, have overfitted to the test data.[32] To avoid this, we need to completely hide away the test data until we are done with training, experimenting, and model selection. To enable this, a common solution is to introduce a third dataset by splitting the training data into two sets of data: a training set and a validation set. The three sets of data used in the development process is then:

- a training set to train or develop candidate models
- a validation set to evaluate and select the best model
- a test set for the final evaluation of model performance

[32] Murphy, see n. 5.

2.7 Regularization

Regularization is a collection of strategies used to avoid overfitting by penalizing the complexity of ML models. Two classical examples are L2 and L1 regularization that adds an regularization term ω, on the model parameters ϕ, to the loss function $\mathcal{L}$.[33]

[33] Goodfellow et al., see n. 1.

$$\widetilde{\mathcal{L}}(\boldsymbol{\phi}, \mathbf{X}, \boldsymbol{y}) = \mathcal{L}(\boldsymbol{\phi}, \mathbf{X}, \boldsymbol{y}) + \underbrace{\omega(\boldsymbol{\phi})}_{\text{regularization term}} \tag{2.2}$$

In the case of L1 regularization, the regularization term consists of the sum of the absolute values of the model parameters, also known as the L1 norm. For L2 regularization, the term comprises the sum of the squares of the model parameters, otherwise known as the squared L2 norm.[34]

[34] Murphy, see n. 5.

$$L1: \quad \omega(\boldsymbol{\phi}) = \lambda_1 ||\boldsymbol{\phi}||_1 = \lambda_1 \sum_i |\phi_i| \tag{2.3}$$

$$L2: \quad \omega(\boldsymbol{\phi}) = \lambda_1 ||\boldsymbol{\phi}||_2^2 = \lambda_1 \sum_i \phi_i^2 \tag{2.4}$$

The regularization strength is controlled by a hyperparameter, λ; a value close to zero imposes minimal regularization, while larger values increases the amount. In the context of neural networks, another commonly used regularization method is *dropout*.[35] This method simply involves randomly dropping some of the output features of the hidden layers during each iteration of the training. This process in a sense creates a different architecture at every step, discouraging the model from becoming overly dependent on any single feature and thereby enhances generalization.[36] Empirically, dropout have been found to give

[35] Nitish Srivastava et al. 'Dropout: A Simple Way to Prevent Neural Networks from Overfitting'. *The journal of machine learning research* (2014).

[36] Charniak, see n. 16.

[37] Srivastava et al., see n. 35.

significant improvements across many different architectures[37] and is a consequence broadly utilized.[38] The droput rate, which controls the probability of dropping out each individual unit, is a hyperparameter that needs to be specified.

[38] Charniak, see n. 16.

2.8 *Hyperparameter Optimization*

Neural networks, as well as most other ML models, have parameters and settings that are not adjusted during training and therefore needs to pre-specified.[39] These parameters are known as hyperparameters, and common examples include aspects such as learning rate, number of layers in the neural network, number of nodes in each layer, Other specific examples have also been described previously in this chapter.

[39] Goodfellow et al., see n. 1.

Although it is possible to assign default values to hyperparameters based on prior experience and personal preference, it is imperative to acknowledge their possible impact on model performance. Consequently, it is common to explore different setting and combinations of hyperparameters in the model building process.[40] In the field of machine learning, this process is known as hyperparameter optimization (HPO).

[40] Goodfellow et al., see n. 1.

If the number of hyperparameters are sufficiently small, a commonly employed strategy for HPO is creating a range of possible candidate values for each hyperparameter and simply testing the entire space of possible combinations in what is known as a *grid search*. The disadvantage is, however, that the search space quickly explodes in size and this strategy may therefore not be feasible. An alternative strategy, which have been empririically and theoretically shown to outperform grid search, and therefore typically should be prefered, is *random search*.[41] In random search, the hyperparameters values are neither binned nor discretized and are instead sampled from a uniform distribution.[42] State-of-the-art HPO approaches includes Bayesian optimization models, multi-fidelity optimization, and metaheuristics algorithms.[43] Many of these algorithms are implemented in the open-source Python package *Optuna*, a very popular software framework for HPO in Python.[44]

[41] James Bergstra and Yoshua Bengio. 'Random Search for Hyper-Parameter Optimization'. *Journal of Machine Learning Research* (2012).

[42] Bergstra and Bengio, see n. 41.

[43] Li Yang and Abdallah Shami. 'On Hyperparameter Optimization of Machine Learning Algorithms: Theory and Practice'. *Neurocomputing* (2020).

[44] Takuya Akiba et al. 'Optuna: A Next-generation Hyperparameter Optimization Framework'. KDD '19: The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2019.

2.9 *Model Explainability*

As described above, the overall goal of ML is to make accurate predictions on unseen data, and the 'how' and 'why' of such predictions is, in the general ML paradigm, explicitly of little concern. Consequently, it is accepted that complex neural networks with deep architectures and many thousands of parameters are 'black box' models that can not be easily described nor understood.[45] For many applications, this lack of transparency can be accepted, but for other applications where trust is paramount, including precision medicine, ongoing efforts seeks to adress this inherent limitation.[46]

[45] Russell and Norvig, see n. 27.

[46] Bas H. M. van der Velden et al. 'Explainable Artificial Intelligence (XAI) in Deep Learning-Based Medical Image Analysis'. *Medical Image Analysis* (2022).

2.9.1 *Interpretability and Explanability*

In the discussion of explainable artificial intelligence (XAI), there is a meaningful distinction to be made between interpretability and ex-

plainability: A model is said to be interpretable if we can relatively easily understand the model through inspection of the model itself.[47] An explainable model, on the other hand, is a simplified external process that provides an interpretable approximation of the complex non-interpretable model.[48] For neural networks, model explainability techniques are generally the only available option.

[47] Russell and Norvig, see n. 27.

[48] Scott M Lundberg and Su-In Lee. 'A Unified Approach to Interpreting Model Predictions'. *Advances in Neural Information Processing Systems*. 2017.

2.9.2 *What is SHAP Analysis?*

A popular method for explainability analysis of neural network models is Shapley additive explanations (SHAP), first published by Lundberg and Lee in 2017.[49] SHAP is a model-agnostic method grounded in co-operative game theory that provides a unified measure of feature importance for ML models. It is based on the concept of Shapley values, an approach for fairly distributing payoffs in a coalitional game based on the individual contributions of the players.[50]

[49] Lundberg and Lee, see n. 48.

[50] Lloyd S Shapley. 'A Value for N-Person Games' (1953).

In this framework, the 'payoff' is the model output, and the 'players' are the features. For a specific feature vector $\boldsymbol{x}'$ we decompose the model output $f(\boldsymbol{x}')$ into the individual feature contributions v_j.[51]

$$f(\boldsymbol{x}') \approx v_0 + \sum_j v_j \tag{2.5}$$

[51] Kjersti Aas et al. 'Explaining Individual Predictions When Features Are Dependent: More Accurate Approximations to Shapley Values'. *Artificial Intelligence* (2021).

where v_0 is a shared baseline and $\check{v}_j$ is the marginal contribrution for the jth feature in $\boldsymbol{x}$.

The shared baseline is typically $\mathbb{E}[f(\boldsymbol{x})]$, either the average or median prediction from the training set.[52] With this formulation, the Shapley values represent the difference between the prediction $f(\boldsymbol{x}')$ and this reference prediction, and from eq. (2.5) it follows that these contributions are additive.[53] This enables visualizations such as fig. 2.8.

[52] Aas et al., see n. 51.

[53] Aas et al., see n. 51.

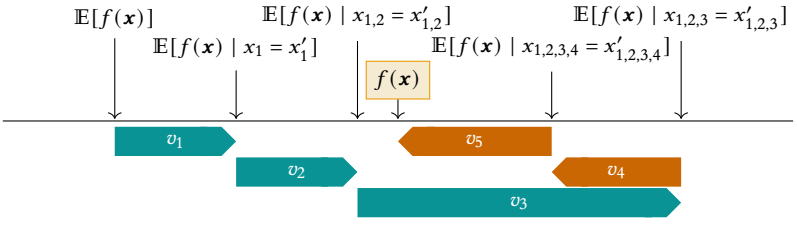


Figure 2.8: Shapley additive explanations (SHAP) values represents the difference between the specific model output $f(\boldsymbol{x}')$ and a baseline prediction $\mathbb{E}[\boldsymbol{x}]$. The SHAP values v_j are additive and by adding them to the baseline prediction, we effetively condition the expected model prediction on that feature.

It is beyond the scope of this thesis to detail how Shapley values are computed in the SHAP framework. However, it is important to note that their exact calculation is NP-complete and therefore computationally intractable.[54] Instead, the SHAP method have pioneered different clever approximations that can be computed in reasonable time.[55]

[54] Xiaotie Deng and Christos H. Papadimitriou. 'On the Complexity of Cooperative Solution Concepts'. *Mathematics of Operations Research* (1994).

[55] Lundberg and Lee, see n. 48.

Chapter 3

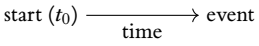
Time-to-Event Prediction with Neural Networks

In the prior chapter, I provided and overview of machine learning and neural networks, highlighting central ideas and concepts releveant to the research presented in this thesis. Specifically, neural networks were employed in were used in Study II and Study III to develop prediction models for ischemic heart disease. These models, however, diverge from classical neural network methods in that they include adaptions that that render them suitable for modelling and prediction of time-to-event data. This chapter delves into the fundamentals of survival analysis, subsequently detailing the theoretical approaches used for implementing survival analysis in neural network models.

What is survival analysis?

outcome time until an event occurs. Can be measured in seconds, days, months, etc.

event death, relapse, remission, engine failure, etc.



3.1 *Introduction to Survival Analysis*

Generally, survival analysis is the collection of statistical methods for the modelling and analysis of time-to-event data, which is a type of data where the outcome variable of interest is the time until 'something' happens.[1] This 'something' is a particular event of interest, which, depending on the type of analytical problem, could be cancer relapse, diabetes remission, or death.

[1] David G. Kleinbaum and Mitchel Klein. *Survival Analysis: A Self-Learning Text, Third Edition*. 3rd edition. Springer, 2011.

In cardiovascular research, common examples of time-to-event outcomes include (i) time to death attributed to any cause (all-cause mortality); (ii) time to death due to a specific cause (e.g. sudden cardiac arrest); and (iii) time to first occurence of a major adverse cardiovascular event (MACE).

To figure out what processes and characteristics that are associated with these events, in survival analysis, we try to model the relationship between explanatory variables and the number of weeks, months, or years until that particular event is likely to occur.

Survival analysis have applications outside biomedical research. In engineering, it is called *reliability analysis* and is used to model the time-to-failure of system-critical components such as e.g. bearings or valves.

Although this task can be daunting in its own right, an additional complication to survival analysis is the presence of observations that are subject to censoring. This concept, censoring, refers to cases where the event of interest has not been observed before the end of follow-up, e.g. when a study or experiment has to be stopped. In such cases, we would know that a given subject did not experience a relapse in the three months he or she was included in the study, but after the study period ends, we have no information on the status of the patient. Including and utilizing this partial information is a cornerstone in many survival analysis problems.

There exists different forms of censoring, such as right censoring, left censoring, and interval censoring. In the study designs used throughout this thesis we have only had to deal with right censoring, the most common form of censoring, so the two other types will not be described further. See instead the text book by Klein and Moeschberger for more details on this.[2]

[2] John P. Klein and Melvin L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, 2003.

3.2 Fundamentals of Survival Analysis

In survival analysis, the central outcome variable is survival time, a non-negative random variable denoted as T. When refering to specific values of T, a lower case t is typically used. A survival dataset $\mathfrak{D}$ of size N is given by

$$\mathfrak{D}_N = \{(t_i, \sigma_i, \boldsymbol{x}_i) \mid i = 1, \ldots, N\} \tag{3.1}$$

where $t_i = \min(T_i, C_i)$ is the survival time for the ith subject, with T_i denoting the survival time and C_i denoting the censoring time. Also, $\boldsymbol{x}_i = (x_1, x_2, \ldots, x_p)'$ is the covariate vector and σ_i is the event indicator, which is defined as

$$\sigma_i = \begin{cases} 0 & \text{if subject is censored } (T_i > C_i) \\ 1 & \text{if event is observed } (T_i \leq C_i) \end{cases} \tag{3.2}$$

In the following, I will initially be assuming that T is continuous and that there is an absence of competing risks, however both of these assumptions will later be relaxed in the discussion of competing risks and discrete-time survival analysis.

3.2.1 Basic Survival Quantities

In survival analysis, the central function of interest is the survival function $S(t)$, that represents the probability of an individual still being alive after some specified duration of time, we have that

$$S(t) = \Pr(T > t), \quad 0 < t < \infty. \tag{3.3}$$

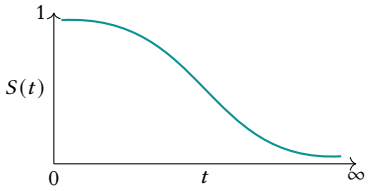


Figure 3.1: An illustration of a theoretical survival function $S(t)$. Per definition, a survival function starts at $S(0) = 1$ and is monotonically non-increasing. [Klein and Moeschberger, see n. 2]

The survival function is the integral of the probability density function, $f(t)$, and is the complement to the cumulative distribution function, $F(t)$, which means that[3]

$$S(t) = 1 - F(t) \quad \text{and} \quad S(t) = \int_t^\infty f(u)\,\mathrm{d}u \tag{3.4}$$

[3] Klein and Moeschberger, see n. 2.

Another fundamental quantity is the hazard function, or hazard rate, which represents the instantaneous failure rate at a given timepoint, and is defined as

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \tag{3.5}$$

from which it can be shown that[4]

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{\mathrm{d}}{\mathrm{d}t} \ln[S(t)]. \tag{3.6}$$

[4] Klein and Moeschberger, see n. 2.

and thus the hazard function completely describes the distribution of T, such that all the the other quantities can be obtained from it—as well as the other way around.

In terms of its interpretation, from eq. (3.5) it follows that $\lambda(t)\Delta t$ is a measure of the conditional probability of failure in a small time window, given that the individual is still alive at time t.[5]

[5] Klein and Moeschberger, see n. 2.

Analogous to the relation between $f(t)$ and $F(t)$, integrating $\lambda(t)$ with respect to t, we obtain cumulative hazard function, defined as

$$\Lambda(t) = \int_0^t \lambda(u)\,\mathrm{d}u = -\ln[S(t)]. \tag{3.7}$$

3.2.2 *The Kaplan-Meier Estimator*

The survival function of a population can be estimated using the Kaplan-Meier method, which is the standard estimator of the survival function.[6,7] In order to estimate the survival function, in the Kaplan-Meier method we first order the distinct failure times such that[8]

$$t_{(1)} < t_{(2)} < \ldots < t_{(j)},$$

and we introduce two quantities to keep track of the number of failures/events at each timepoint $\bar{D}(j)$, as well as the number of subjects still at risk at each timepoint $\bar{A}(j)$. They are defined as

$$\begin{aligned} \bar{D}(j) &= \#\{i \in \{1,\ldots,n\} \mid t_i = t_{(j)}, \sigma_i = 1\} \\ \bar{A}(j) &= \#\{i \in \{1,\ldots,n\} \mid t_i > t_{(j)}\}. \end{aligned} \tag{3.8}$$

With these two quantities in place, the Kaplan-Meier estimator can then be formulated as

$$\widehat{S}(t) = \prod_{j \mid t_{(j)} \le t} \frac{\bar{A}(j) - \bar{D}(j)}{\bar{A}(j)} = \prod_{j \mid t_{(j)} \le t} 1 - \frac{\bar{D}(j)}{\bar{A}(j)}. \tag{3.9}$$

While the Kaplan-Meier estimator is very useful for estimating the average survival of a population, it does not account for the effect of covariates. Instead, another approach is needed for regression analyses.

[6] Edward L Kaplan and Paul Meier. 'Nonparametric Estimation from Incomplete Observations'. *Journal of the American statistical association* (1958).

[7] Klein and Moeschberger, see n. 2.

[8] Following the example of [Kleinbaum and Klein, see n. 1], the t's denoted with subscripts within parentheses $t_{(j)}$ refers to the jth element of the ordered distinct failure times and are thus different from $t_1, t_2, \ldots, t_i$ that refers to the observed failure time of subject 1, 2, and i

3.2.3 *Cox's Proportional Hazards Model*

To describe and model the relationship between explanatory variables and time-to-event phenomenons, a widely used statistical model is the Cox proportional hazards model.[9] This model seeks to model the hazard function over time t, of an individual with a covariate vector $\boldsymbol{x} = (x_1, x_2, \ldots)'$, and assumes that it takes the form of

$$\widehat{\lambda}(t \mid \boldsymbol{x}) = \lambda_0(t)\exp[g(\boldsymbol{x})], \tag{3.10}$$

where $\lambda_0(t)$ is an unspecified baseline hazard function, and $g(\boldsymbol{x})$ is some parametric function. For this reason, the Cox model is referred to as a semi-parametric model. In its classical formulation, this function is a linear combination of parameters $\boldsymbol{\beta}$ and covariates $\boldsymbol{x}$, as given by

$$g(\boldsymbol{x}) = \boldsymbol{\beta}'\boldsymbol{x} = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \tag{3.11}$$

In estimation of the parameters $\boldsymbol{\beta}$, the baseline hazard $\lambda_0(t)$ is treated as a nuisance function and the coefficients are estimated by maximising a partial likelihood in which $\lambda_0(t)$ has been abstracted away.[10]

[9] D. R. Cox. 'Regression Models and Life-Tables'. *Journal of the Royal Statistical Society. Series B (Methodological)* (1972).

[10] Kalbfleisch. *The Statistical Analysis of Failure Time Data, 2nd Edition*. 2nd edition. Wiley-Interscience, 2002.

A central assumption in the Cox model, at least in the standard version with fixed covariates ($\boldsymbol{\beta}$ instead of $\boldsymbol{\beta}(t)$), is that of proportional hazards. Let $\boldsymbol{x}$ and $\boldsymbol{x}'$ be two different covariate vectors, now the ratio between their respective Cox-estimated hazards is

$$\begin{aligned}\frac{\widehat{\lambda}(t \mid \boldsymbol{x})}{\widehat{\lambda}(t \mid \boldsymbol{x}')} &= \frac{\lambda_0(t)\exp(\boldsymbol{\beta}\cdot\boldsymbol{x})}{\lambda_0(t)\exp(\boldsymbol{\beta}\cdot\boldsymbol{x}')} \\ &= \frac{\exp(\boldsymbol{\beta}\cdot\boldsymbol{x})}{\exp(\boldsymbol{\beta}\cdot\boldsymbol{x}')} \\ &= \exp(\boldsymbol{\beta}\cdot(\boldsymbol{x}-\boldsymbol{x}'))\end{aligned} \tag{3.12}$$

Since the right-hand side of the equation does not include a term for t, the hazard ratio between the two samples are constant and they are thus proportional to one another. This shows that by assuming the hazard takes the form of eq. (3.10), then it is also assumed that the hazards between two subjects are proportional. Although this assumption is a strong one, and the validity of the Cox model relies on it, the assumption makes interpretation of parameters easier.[11] For example, in an randomized clinical trial studying the survival effect of a new type of medication, we can let $x = 1$ represent the experimental treatment and $x' = 0$ represent standard of care, then the hazard ratio in eq. (3.12) takes the form of

$$\exp\left(\beta(x-x')\right) = \exp\left(\beta(1-0)\right) = \exp(\beta), \tag{3.13}$$

which means that if $\beta < 0$, then the hazard of the experimental treatment is $\exp(\beta)$ times lower than standard of care and should therefore be preferred.[12]

[11] Gerhard Tutz and Matthias Schmid. *Modeling Discrete Time-to-Event Data*. 1st ed. 2016 edition. Springer, 2016.

[12] This example is a slightly modified version of the one given in Tutz and Schmid, see n. 11, pp. 50

3.3 Time-to-Event Prediction

Up until now, I have outlined various concepts foundational to survival analysis, focusing primarily on quantities and statistics of time-to-event outcomes at a population level. These measures play an important role in understanding and interpretation of survival data.

In the context of precision medicine, however, the emphasis shifts towards making individualized predictions taking distinct patient-level characteristics into account. Consequently, as described in chapter 2, the primary concern lies in making accurate predictions on unseen data, rather than in the exploration of disease etiology and underlying mechanisms.

For prediction of time-to-event outcomes, classical approaches include models based on the previously presented semi-parametric Cox model as well as various parametric survival models, such as those based on exponential, Weibull, or log-normal distributions.[13] This thesis, however, explores the use of contemporary machine learning methods in time-to-event prediction, with a particular emphasis on the application of neural networks.

[13] Klein and Moeschberger, see n. 2.

3.3.1 Neural Networks and Time-to-Event Outcomes

The first application of neural networks for time-to-event prediction was demonstrated by Faraggi and Simon in 1995, and involves parameterising

the parametric part of the Cox model with a neural network, such that the $g(\boldsymbol{x})$ term in eq. (3.10) is a flexible neural network model instead of a simple linear function.[14]

use neural network

$$\widehat{\lambda}(t \mid \boldsymbol{x}) = \lambda_0(t) \exp[\, g(\boldsymbol{x}) \,]$$

[14] David Faraggi and Richard Simon. 'A Neural Network Model for Survival Data'. *Statistics in Medicine* (1995).

This approach was later further refined in the *DeepSurv* paper from 2018, in which modern neural network techniques were added to Faraggi-Simon framework, which markedly improved its usefulness.[15] Katzman et al. showed that the flexibility offered by neural networks led to increased performance in both synthetic and real-life time-to-event prediction applications compared to a standard Cox model. However, the *DeepSurv* approach is still limited by the assumption of proportional hazards.

[15] Jared L. Katzman et al. 'DeepSurv: Personalized Treatment Recommender System Using a Cox Proportional Hazards Deep Neural Network'. *BMC Medical Research Methodology* (2018).

3.3.2 *Overview of Approaches*

Recently, there have been considerable interest in neural network-based time-to-event prediction models, and as a consequence, many new methods have since been developed. For a thorough overview of the existing approaches, Wiegrebe et al.[16] and Kvamme and Borgan[17] provide valuable insights. Generally, two prevailing types of approaches exists: continuous-time methods based on the Cox model, which includes *DeepSurv*, and discrete-time methods as exemplified by Lee et al.[18] and Gensheimer and Narasimhan.[19]

The discrete-time approaches offer several advantages that make them particularly relevant for neural network. Furthermore, they have been shown to offer better predictive performance compared to the Cox-based methods.[20] Notably, *DeepHit*[21] and *Logistic-Hazard*[22] are the two most cited papers in this context as of the time of writing. Among these and other tested approaches, Kvamme and Borgan[23] found that DeepHit offers excellent discrimination but suffers from poor calibration. In contrast, the Logistic-Hazard model have nearly as good discrimination and also significantly better calibration. Consequently, the Logistic-Hazard model, and an extension hereof, was chosen for application in Study II and Study III.

In the following section, I will be giving a brief description of this discrete-time formulation of time-to-event analysis and elaborate on the Logistic-Hazard model in more detail.

[16] Simon Wiegrebe et al. 'Deep Learning for Survival Analysis: A Review'. *arXiv* (2023). URL: arxiv.org/abs/2305.14961. preprint.

[17] Håvard Kvamme and Ørnulf Borgan. 'Continuous and Discrete-Time Survival Prediction with Neural Networks'. *Lifetime Data Analysis* (2021).

[18] Changhee Lee et al. 'DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks'. *Proceedings of the AAAI Conference on Artificial Intelligence* (2018).

[19] Michael F. Gensheimer and Balasubramanian Narasimhan. 'A Scalable Discrete-Time Survival Model for Neural Networks'. *PeerJ* (2019).

[20] Kvamme and Borgan, see n. 17; Lee et al., see n. 18; Gensheimer and Narasimhan, see n. 19.

[21] Lee et al., see n. 18.

[22] Gensheimer and Narasimhan, see n. 19.

[23] Kvamme and Borgan, see n. 17.

3.4 *Discrete-Time Survival Analysis*

Most textbooks on survival analysis treats survival time as continuous, and that is also usually the case across the biomedical litterature. However, handling time as a something discrete can be advantegous. In practice, most measurements of time is inherently discrete with durations being recorded in, for example, days; months; and weeks. The continuous time approaches presented earlier in this chapter, are also applicable

to discrete time data, however, methods designed specifically for discrete time-to-event data have some advantages:[24]

- If observed event times are inherently discrete, then modelling them as such is arguably more appropriate.
- In the discrete-time setting, hazards can be formulated as conditional probabilities which are much more intuitive to both interpret and understand.
- Discrete time-to-event models are more easily transferred to other more general purpose modelling frameworks such as generalized linear models, random survival forests, neural networks.

The latter point is the main motivation behind both the *DeepHit* and *Logistic-Hazard* approach. For a complete overview of the theory enabling these two approaches, the book by Tutz and Schmid[25] is a valuable resource, and serves as the main source of reference for the following.

[24] Tutz and Schmid, see n. 11.

[25] Tutz and Schmid, see n. 11.

3.4.1 *Notation and Definitions*

In the discrete-time framework, continuous follow-up time T_c is divided into q contiguous intervals, that is

$$(0, a_1], (a_1, a_2], \ldots, (a_{q-1}, a_q]$$

and $T_d \in \{1, \ldots, q\}$ is a discrete random variable such that if $T_d = \tau$ is observed, then the event falls in the interval $(a_{\tau-1}, a_\tau]$. Similarly, the discretized censoring time is $C_d \in \{1, \ldots, q\}$.

With this discrete time scale, the distribution of T_d, given some vector of covariates $\boldsymbol{x}$, can be described using discrete equivalents of the previously outlined basic quantities of survival analysis, that is

$$\text{probability mass function:} \quad f(\tau \mid \boldsymbol{x}) = \Pr(T_d = \tau \mid \boldsymbol{x}) \tag{3.14}$$

$$\text{cumulative mass function:} \quad F(\tau \mid \boldsymbol{x}) = \Pr(T_d \leq \tau \mid \boldsymbol{x}) \tag{3.15}$$

$$\text{hazard function:} \quad \lambda(\tau \mid \boldsymbol{x}) = \Pr(T_d = \tau \mid T_d \geq \tau, \boldsymbol{x}) \tag{3.16}$$

$$\text{survival function:} \quad S(\tau \mid \boldsymbol{x}) = \Pr(T_d > \tau \mid \boldsymbol{x}) \tag{3.17}$$

3.4.2 *The Logistic-Hazard Model*

Gensheimer and Narasimhan's approach, which they refer to as 'Nnet-survival',[26] is more accurately characterized as the Logistic-Hazard method, as described in Kvamme and Borgan.[27] In the Logistic-Hazard method, the time-to-event data is described by modelling the effect of covariates on the discrete hazard function (eq. (3.16)) using a neural network. The concept is not novel, employing the discrete hazard for statistical modeling is a common method, as covered extensively in Tutz and Schmid.[28] In addition, a neural-network based model with the same general idea was presented by Brown et al. in 1997.[29] However, Gensheimer and Narasimhan were the first to adapt the approach to current neural network methodologies.

[26] Gensheimer and Narasimhan, see n. 19.

[27] Kvamme and Borgan, see n. 17.

[28] Tutz and Schmid, see n. 11.

[29] S.F. Brown et al. 'On the Use of Artificial Neural Networks for the Analysis of Survival Data'. *IEEE Transactions on Neural Networks* (1997).

3.4.3 *Log-Likelihood of the Discrete Hazard*

Let $\mathfrak{D}_\mathrm{d}$ be a discrete-time survival dataset of size N,

$$\mathfrak{D}_\mathrm{d} = \{(\tau_i, \sigma_i, \boldsymbol{x}_i) \mid i = 1, \ldots, N\}, \tag{3.18}$$

where τ_i is the discretized survival time, σ_i is the event indicator as defined in eq. (3.2), and $\boldsymbol{x}_i = (x_1, x_2, \ldots, x_p)'$ is the feature vector. With the assumption of *noninformative censoring*,[30] in the Logistic-Hazard model, the contribution of the ith individual to the likelihood function can be shown to be[31]

[30] Kalbfleisch, see n. 10.

[31] Tutz and Schmid, see n. 11.

$$\mathcal{L}_i = \begin{cases} \Pr(T_{\mathrm{d}i} = \tau_i) & \text{if non-censored} \\ \Pr(T_{\mathrm{d}i} > \tau_i) & \text{if censored.} \end{cases} \tag{3.19}$$

These two probabilities can be expressed using the discrete hazards, as it can be seen that

$$\begin{aligned} \Pr(T_\mathrm{d} = \tau) &= \Pr(T_\mathrm{d} = \tau \mid T_\mathrm{d} \geq \tau) \Pr(T_\mathrm{d} \geq \tau) \\ &= \Pr(T_\mathrm{d} = \tau \mid T_\mathrm{d} \geq \tau) \Pr(T_\mathrm{d} > \tau - 1) \\ &= \lambda(\tau) \prod_{s=1}^{\tau-1} (1 - \lambda(s)) \end{aligned} \tag{3.20}$$

and similarly

$$\begin{aligned} \Pr(T_\mathrm{d} > \tau) &= \Pr(T_\mathrm{d} > \tau \mid T_\mathrm{d} \geq \tau) \Pr(T_\mathrm{d} \geq \tau) \\ &= (1 - \Pr(T_\mathrm{d} = \tau \mid T_\mathrm{d} \geq \tau)) \Pr(T_\mathrm{d} \geq \tau) \\ &= (1 - \Pr(T_\mathrm{d} = \tau \mid T_\mathrm{d} \geq \tau)) \Pr(T_\mathrm{d} > \tau - 1) \\ &= \prod_{s=1}^{\tau} (1 - \lambda(i)). \end{aligned} \tag{3.21}$$

Now, by introducing an indicator function, defined according to Tutz and Schmid[32] as

[32] Tutz and Schmid, see n. 11.

$$\bar{y}_i(\tau) = \begin{cases} 1, & \text{if individual fails in } (a_{\tau-1}, a_\tau], \\ 0, & \text{if individual survives } (a_{\tau-1}, a_\tau], \end{cases} \tag{3.22}$$

and by including the discrete hazard function and the covariates, the likelihood contribution for the ith individual can be expressed as

$$\mathcal{L}_i = \prod_{s=1}^{\tau_i} \lambda(s \mid \boldsymbol{x}_i)^{\bar{y}_i(s)} (1 - \lambda(s \mid \boldsymbol{x}_i))^{1 - \bar{y}_i(s)}. \tag{3.23}$$

The total log-likelihood of all datapoints then gives the loss-function used in the Logistic-Hazard model,[33] which can be expressed

$$\ell = \sum_{i=1}^{N} \sum_{s=1}^{\tau_i} \bar{y}_i(s) \log(\lambda(s \mid \boldsymbol{x}_i)) + (1 - \bar{y}_i(s)) \log(1 - \lambda(s \mid \boldsymbol{x}_i)). \tag{3.24}$$

[33] Gensheimer and Narasimhan, see n. 19; Tutz and Schmid, see n. 11.

3.4.4 *Loss Function Explained*

As an example, the following set of observations with discretized time-to-event data with a single risk, e.g. all-cause mortality, and a follow-up time that have been discretized into seven contiguous intervals, constitutes a survival dataset.

subject i	1	2	3	4	5
time τ_i	5	7	4	5	3
event σ_i	1	1	0	0	1

(3.25)

In this setting, using the Logistic-Hazard model, the neural network output for this dataset is a 2-dimensional matrix with 5 rows (subjects) and 7 columns (time intervals), and each entry is the predicted conditional hazard for the specific subject at a specific timepoint. We can write this as

$$\hat{\mathbf{\Lambda}} = \begin{bmatrix} \hat{\lambda}_{11} & \hat{\lambda}_{12} & \hat{\lambda}_{13} & \hat{\lambda}_{14} & \hat{\lambda}_{15} & \hat{\lambda}_{16} & \hat{\lambda}_{17} \\ \hat{\lambda}_{21} & \hat{\lambda}_{22} & \hat{\lambda}_{23} & \hat{\lambda}_{24} & \hat{\lambda}_{25} & \hat{\lambda}_{26} & \hat{\lambda}_{27} \\ \hat{\lambda}_{31} & \hat{\lambda}_{32} & \hat{\lambda}_{33} & \hat{\lambda}_{34} & \hat{\lambda}_{35} & \hat{\lambda}_{36} & \hat{\lambda}_{37} \\ \hat{\lambda}_{41} & \hat{\lambda}_{42} & \hat{\lambda}_{43} & \hat{\lambda}_{44} & \hat{\lambda}_{45} & \hat{\lambda}_{46} & \hat{\lambda}_{47} \\ \hat{\lambda}_{51} & \hat{\lambda}_{52} & \hat{\lambda}_{53} & \hat{\lambda}_{54} & \hat{\lambda}_{55} & \hat{\lambda}_{56} & \hat{\lambda}_{57} \end{bmatrix} \begin{matrix} \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} \text{time} \rightarrow \\ \\ \\ \\ \\ \end{matrix} \quad \text{subject} \downarrow \tag{3.26}$$

Now, the indicator function can be computed using the defintion in eq. (3.22) and the observed data τ_i and σ_i. In matrix form, the output of this function is

$$\bar{Y} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & - & - \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & - & - & - \\ 0 & 0 & 0 & 0 & 0 & - & - \\ 0 & 0 & 1 & - & - & - & - \end{bmatrix} \tag{3.27}$$

Now, combining these two matrices according to the formula in eq. (3.23), we obtain the likelihood in matrix form as

$$\mathcal{L}(\hat{\mathbf{\Lambda}}, \bar{Y}) = \begin{bmatrix} 1-\hat{\lambda}_{11} & 1-\hat{\lambda}_{12} & 1-\hat{\lambda}_{13} & 1-\hat{\lambda}_{14} & \hat{\lambda}_{15} & - & - \\ 1-\hat{\lambda}_{21} & 1-\hat{\lambda}_{22} & 1-\hat{\lambda}_{23} & 1-\hat{\lambda}_{24} & 1-\hat{\lambda}_{25} & 1-\hat{\lambda}_{26} & \hat{\lambda}_{27} \\ 1-\hat{\lambda}_{31} & 1-\hat{\lambda}_{32} & 1-\hat{\lambda}_{33} & 1-\hat{\lambda}_{34} & - & - & - \\ 1-\hat{\lambda}_{41} & 1-\hat{\lambda}_{42} & 1-\hat{\lambda}_{43} & 1-\hat{\lambda}_{44} & 1-\hat{\lambda}_{45} & - & - \\ 1-\hat{\lambda}_{51} & 1-\hat{\lambda}_{52} & \hat{\lambda}_{53} & - & - & - & - \end{bmatrix} \tag{3.28}$$

from which the log-likelihood, eq. (3.24), is then

$$\begin{aligned} \ell(\hat{\mathbf{\Lambda}}, \bar{Y}) = {} & \log(1-\hat{\lambda}_{11}) + \log(1-\hat{\lambda}_{12}) + \log(1-\hat{\lambda}_{13}) + \log(1-\hat{\lambda}_{14}) + \log(\hat{\lambda}_{15}) \\ & + \log(1-\hat{\lambda}_{21}) + \log(1-\hat{\lambda}_{22}) + \log(1-\hat{\lambda}_{23}) + \log(1-\hat{\lambda}_{24}) + \log(1-\hat{\lambda}_{25}) + \log(1-\hat{\lambda}_{26}) + \log(\hat{\lambda}_{27}) \\ & + \log(1-\hat{\lambda}_{31}) + \log(1-\hat{\lambda}_{32}) + \log(1-\hat{\lambda}_{33}) + \log(1-\hat{\lambda}_{34}) \\ & + \log(1-\hat{\lambda}_{41}) + \log(1-\hat{\lambda}_{42}) + \log(1-\hat{\lambda}_{43}) + \log(1-\hat{\lambda}_{44}) + \log(1-\hat{\lambda}_{45}) \\ & + \log(1-\hat{\lambda}_{51}) + \log(1-\hat{\lambda}_{52}) + \log(\hat{\lambda}_{53}) \end{aligned}$$

3.5 *Survival Analysis with Competing Risks*

Up to this point, the description of concepts in survival analysis has assumed the presence of only a single event type, such as all-cause

mortality (fig. 3.2). In practice, particularly in clinical settings, this single-event model can be too restrictive, and instead one needs to consider competing risks (fig. 3.3). By definition, a competing risk is a secondary event whose occurence prevents the primary event from occuring. For example, in a study where the primary outcome is cardiovascular mortality, deaths from non-cardiovascular causes are a competing risk.

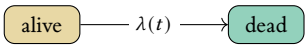


Figure 3.2: A simple survival analysis setup involves modelling a single transition between states 'alive' and 'dead'.

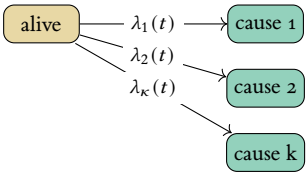


Figure 3.3: A survival analysis setup with competing risks involves modelling transitions between states 'alive' and k different absorbing states, 'cause 1' to 'cause κ'

3.5.1 *Cause-Specific Survival Quantities*

To describe time-to-event phenomena with competing risks, we introduce the cause-specific hazard function and cumulative-incidence function. With $R \in \{1, \dots, \kappa\}$ denoting the κ different competing risks, the continuous cause-specific hazard function is defined as

$$\lambda_r(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t, R = r \mid T \geq t)}{\Delta t} \tag{3.29}$$

where r refers to a specific value of R. The cause-specific cumulative incidence function is defined as[34]

$$F_r(t) = \Pr(T \leq t, R = r). \tag{3.30}$$

The overall hazard and cumulative incidence, which combines failures of any of the κ causes, correspond to the hazard function and the cumulative distribution function in the single-event setting, that is

$$\lambda(t) = \sum_{r=1}^{\kappa} \lambda_r(t) \quad \text{and} \quad F(t) = \sum_{r=1}^{\kappa} F_r(t). \tag{3.31}$$

[34] Kalbfleisch, see n. 10.

3.5.2 *Modelling the Cause-Specific Hazard*

In the competing risks setting, the typical approach is to treat competing risks as censored events and use Cox regression to estimate the cause-specific hazards. In this approach, the cause-specific hazard takes the form of

$$\hat{\lambda}_r(t \mid \boldsymbol{x}) = \lambda_{0r} \exp(\boldsymbol{\beta}_r \cdot \boldsymbol{x}) \tag{3.32}$$

from which one can obtain the cause-specific regression coefficients $\boldsymbol{\beta}_r$. In Study I, we use this to estimate the cause-specific hazards of ischemic heart disease (IHD) progression and non-IHD mortality associated with different clusters of IHD patients, as defined by their respective comorbidity profiles.

However, an important assumption required by all survival methods outlined so far, including the Cox proportional hazards model, is that of noninformative censoring.[35] This assumption states that the 'probability of being censored at time t does not depend on prognosis for failure at time t',[36] which in the context of competing risks can be especially problematic, since it also implies that competing failure types should be independent. It is difficult to ascertain if this is the case from observed data, however if we include risk-factors that are shared by competing events, it is possible to alleviate the bias related to this possibly erroneous assumption.[37]

[35] Kleinbaum and Klein, see n. 1.

[36] Kleinbaum and Klein, see n. 1.

[37] Kleinbaum and Klein, see n. 1.

3.5.3 *The Aalen-Johansen Estimator*

In estimation of the population-level cause-specific incidence, the approach of simply treating competing events as censored and applying the standard Kaplan-Meier estimator, is generally a bad idea, since it often leads to a very biased estimate of $F(t)$.[38] Instead, an alternative approach is the Aalen-Johansen estimator that allows estimation of the cause-specific cumulative incidence.[39] Of note, the Aalen-Johansen is a general method for estimating transition probabilities in state-transition models, and can be used to describe complex multi-state models, including those with repeated events and with non-terminal states.[40] However, we will be assuming a standard competing-risk setting with κ different terminal states, as depicted in fig. 3.3.

If we again order the distinct failure times, corresponding to any cause, such that $t_{(1)} < t_{(2)} < \ldots < t_{(j)}$, and update the definition of $\bar{D}(j)$ to keep track of cause-specific events, such that we have

$$\begin{aligned} \bar{D}(j,r) &= \#\{i \in \{1,\ldots,n\} \mid t_i = t_{(j)}, r_i = r\} \\ \bar{A}(j) &= \#\{i \in \{1,\ldots,n\} \mid t_i > t_{(j)}\}. \end{aligned} \tag{3.33}$$

Now, the Aalen-Johansen estimator of the cumulative incidence function can be defined as

$$\widehat{F}_r(t) = \sum_{j \mid t_{(j)} \leq t} \widehat{S}(t_{(j-1)}) \frac{\bar{D}(j,r)}{\bar{A}(j)} \tag{3.34}$$

[38] Margaret S. Pepe and Motomi Mori. 'Kaplan—Meier, Marginal or Conditional Probability Curves in Summarizing Competing Risks Failure Time Data?' *Statistics in Medicine* (1993).

[39] Odd O. Aalen and Søren Johansen. 'An Empirical Transition Matrix for Non-Homogeneous Markov Chains Based on Censored Observations'. *Scandinavian Journal of Statistics* (1978).

[40] Terry M Therneau. *A Package for Survival Analysis in R*. manual. 2023.

Chapter 4

Overview of Data Resources

Data stands as the cornerstone of both precision medicine and machine learning. Its availability is crucial; without it, research in these fields would be almost impossible. The emergence of high-throughput analyses and electronic health records (EHRs) has led to a Cambrian explosion of the volume of data being generated, which has the potential of revolutionizing the entire landscape of biomedical research. However, the sensitive nature of this data means that access is frequently a major challenge, often serving as a major bottleneck in many research endeavors.

In the context of the studies conducted for this thesis, I have been in the privileged position of working within a research group where permissions, data access, and the necessary infrastructure were already well-established. In terms of infrastructure, a key aspect of our data handling involved the use of a secure high-performance computing environment. This not only ensured the efficient processing of large datasets and training of large neural networks, but also maintained the highest standards of data security and confidentiality, which are paramount in dealing with sensitive health records. These aspects have been instrumental in enabling and driving the research and analyses presented in this thesis.

Focusing on the data itself, this chapter aims to provide a comprehensive overview of the various data sources utilized in the studies comprising this thesis. It details the databases and registries that were accessed and analyzed, highlighting how each contributed to the research.

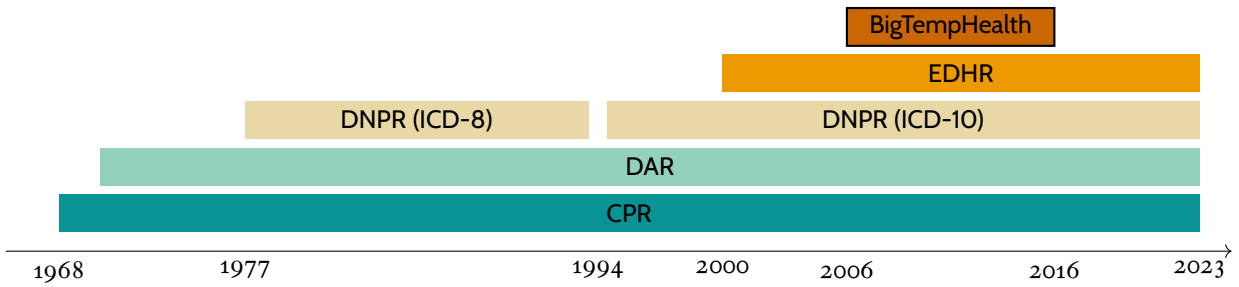


Figure 4.1: Overview of the different data resources used in the research projects presented in this thesis. This schematic shows the timeline of the different datasets.

4.1 *The Danish Civil Registration System*

Danish Civil Registration System (CPR) is the central administrative register in Denmark, and stores personal information on the entire Danish population, including birth date, sex, addresses, vital status, and, importantly, a unique personal identification number, known as 'CPR-nummer' or 'personnummer'.[1]

The CPR number is assigned at birth or upon obtaining Danish citizenship, and have been in use since 1968. As of January 2, 2014, 9 484 792 CPR numbers were assigned.[2] Of these 5 685 912 were 'active', and 3 798 880 were 'non-active', the latter primarily attributed to death and emigration.[3]

In Denmark, the CPR number is as essential as a bicycle, required for opening a bank account, borrowing a library book, getting treated for appendicitis, and everything in between. This widespread use, combined with the country's long-standing tradition of organising and keeping record of detailed data in administrative databases and registries, have enabled the construction of a large network of interlinkable epidemiological resources.[4] In this light, the whole nation can be utilized as a research cohort, as outlined by Frank[5] in a letter to *Science*.

Linkage between the different data sources and registries presented throughout this thesis relied on the use of the CPR number, specifically an encrypted form of it, ensuring that no actual CPR numbers were being handled.

[1] Morten Schmidt et al. 'The Danish Civil Registration System as a Tool in Epidemiology'. *European Journal of Epidemiology* (2014).

[2] Schmidt et al., see n. 1.

[3] Schmidt et al., see n. 1.

[4] Morten Schmidt et al. 'The Danish Health Care System and Epidemiological Research: From Health Care Contacts to Database Records'. *Clinical Epidemiology* (2019).

[5] Lone Frank. 'Epidemiology. When an Entire Country Is a Cohort'. *Science (New York, N.Y.)* (2000).

4.2 *The Danish National Patient Register*

Of the Danish registries, Danish National Patient Register (LPR) is arguably one of the most important. LPR, or 'Landspatientregisteret', is a comprehensive clinical register that has been instrumental for clinical research and administration in Denmark, and serves multiple critical functions.[6] Primarily, it underpins the Danish Health and Medicines Authority's hospital statistics and is a main foundation for health economic calculations. Additionally, the LPR is instrumental in monitoring the prevalence of various diseases and treatments. Furthermore, the registry plays a key role in facilitating quality assurance of Danish healthcare services and provides hospital physicians access to patients' hospitalization histories, enhancing patient care and treatment efficacy.[7] The register is updated monthly based on reports from the hospitals, and has been collecting data continuously since 1977.[8]

[6] Morten Schmidt et al. 'The Danish National Patient Registry: A Review of Content, Data Quality, and Research Potential'. *Clinical Epidemiology* (2015).

[7] Schmidt et al., see n. 6.

[8] Schmidt et al., see n. 6.

4.2.1 *Content and Structure*

The LPR encompasses a wide array of data on each individual, such as personal information, admission and discharge details, diagnoses, examinations, treatments including surgeries, information on accidents, and additional details concerning births.[9] The information in the register is organised in a structured format with different data types being stored in distinct tables that can be linked following a specified relational data model.[10]

[9] Schmidt et al., see n. 6.

[10] *Landspatientregisteret, Dokumentation*. eSundhed. URL: esundhed.dk/Dokumentation?rid=5 (visited on 25/11/2023).

This data model, referred to as LPR2, have remained largely unchanged since the release of the registry in 1977. However, in early 2019, it underwent a significant overhaul to a new and refined data model, LPR3.[11] The LPR3 model addresses certain limitations of its predecessor, notably enabling the creation of more fine-grained patient care timelines. While the specifics of these improvements are beyond the scope of this thesis, it is important to note that the LPR3 model is not entirely backwards compatible with the LPR2 data model. Nevertheless, depending on the specific use case, it remains possible to create data extracts that are compatible to one another.

[11] Lisbeth Nielsen. *LPR3 går snart i luften*. 2018. URL: sundhedsdatastyrelsen.dk.

4.2.2 *Classification of Diseases*

The highly structured data within the LPR is coded using the national SKS classification scheme ('Danish Medical Classification System'), a collection of Danish, international, and Nordic classification standards maintained by the Danish Health Data Authority.[12] These standards includes the *Nordic Medico-Statistical Committee Classification of Surgical Procedures* (NOMESCO) system for surgical procedures, the *Anatomical Therapeutic Chemical Classification System* (ATC) system for medication, and the *International Classification of Diseases* (ICD) system for diagnoses.[13]

[12] Schmidt et al., see n. 6.

[13] Schmidt et al., see n. 6.

Focusing on diagnosis codes, the LPR is currently using the 10th revision of the ICD (ICD-10), and have been doing so since the start of 1994 where it replaced the ICD-8.[14] This transition can complicate longitudinal studies of disease occurence, but prior efforts by the Brunak group have succesfully created a mapping between ICD-8 and ICD-10 that can be used to mitigate such challenges.[15] However, since the two systems are not directly compatible, this mapping only offers a partial solution and might not be universally applicable.

[14] Schmidt et al., see n. 6.

[15] Mette Krogh Pedersen et al. 'A Unidirectional Mapping of ICD-8 to ICD-10 Codes, for Harmonized Longitudinal Analysis of Diseases'. *European Journal of Epidemiology* (2023).

The ICD-10 coding system follows a hierarchical structure with every code beginning with a letter followed by two or more digits. Each code falls in one of 21 high-level categorisation of diagnoses that e.g. includes chapters (ii) *Neoplasms*; (iv) *Endocrine, Nutritional, and Metabolic Diseases*; and (ix) *Diseases of the Circulatory System*.

Using the latter as an example, fig. 4.2 shows the hierarchical structure of the ICD-10. This hierarchical structure can be utilized in feature engineering for machine learning applications, offering features of varying specificity. Following the example in fig. 4.2, a relatively specific 'level-4' code, such as e.g. *acute transmural myocardial infaction of the inferior wall* (I21.1), can also be represented as a 'level-3' code, a 'block' code, or a 'chapter' by stepping back through the hierarchy. This versatility aids in balancing between statistical power and specificity of the included code.

4.2.3 *Role in this Thesis*

The LPR has been an key data resource in all three studies of this thesis. Diagnosis codes and their associated admissions timestamps were used to features for clustering analysis in Study I and for machine learning models in Study II and Study III. Additionally, procedure, treatment, and examination codes played a vital role in defining cohorts in all studies,

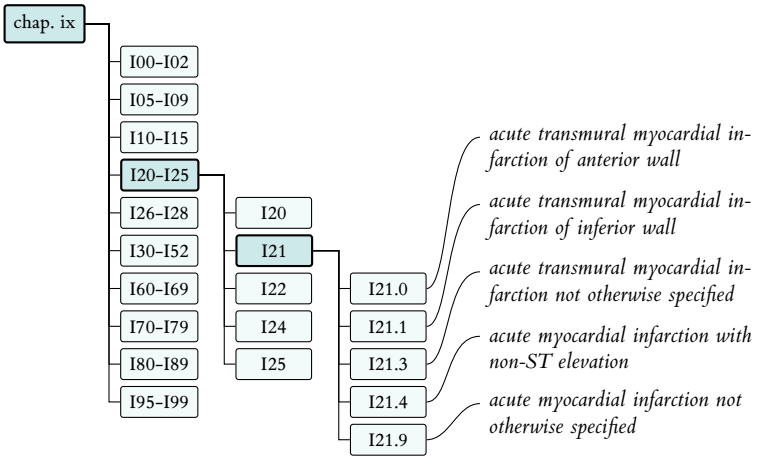


Figure 4.2: Schematic illustrating the hierarchical structure of the ICD-10 coding system. Using chapter ix—which covers diseases of the circulatory system (codes I00–I99) and comprises ten different 'blocks'—as an example, the diagram shows how these blocks are segmented into 'level 3' codes, such as I20 to I25, corresponding to ischemic heart disease. It provides an expanded view of the I21 category, containing specific subtypes, 'level 4' codes, of acute myocardial infarctions from I21.0 for the anterior wall to I21.9 for unspecified instances. Deeper levels of the hierarchy also exists, but is not included in this diagram.

served as features in Study II and Study III, and were crucial for defining outcomes in Study I and Study III.

4.3 *The Causes of Death Register*

When someone in Denmark dies, it has since 1871 been mandatory by law that a physician performs a post-mortem examination and fills in a death certificate.[16] Starting in 1970, these certificates have been stored electronically in Danish Register of Causes of Death (DAR) ('dødsårsagsregisteret'), which is the main source of data for Danish mortality statistics, analyses of medical treatment, and to define outcomes in various research projects.[17] The registry includes data such as the date of death, the primary and contributing causes of death, as well as demographic information about the deceased. Since 2002, the underlying cause of death have been coded using Automated Classification of Medical Entities (ACME), an international standard for automated selection of such.[18] In Study I and Study III of this thesis, the DAR were used to define cause-specific mortality.

[16] Karin Helweg-Larsen. 'The Danish Register of Causes of Death'. *Scandinavian Journal of Public Health* (2011).

[17] Helweg-Larsen, see n. 16.

[18] Helweg-Larsen, see n. 16.

4.4 *The Eastern Denmark Heart Registry*

The Eastern Denmark Heart Registry (EDHR), or 'Web-PATS', is a clinical database collecting information on all coronary angiographies (CAGs) and percutaneous coronary interventions (PCIs) performed in the Capital Region of Denmark and Region Zealand (see fig. 4.3). It collects a wide array of prognostic factors, demographics, administrative details, and procedure-related information and findings that are used for both research and clinical quality assessment.[19]

In relation to the latter, a central role of the EDHR is to deliver data to the national Danish Heart Registry, which is part of Danish Clinical Quality Program (RKKP). The RKKP is a national program set in place to support the management and infrastructure related to clinical quality databases. All RKKP registries undergoes revision by the National Health Authority every three years to continuously assess specific clinical quality

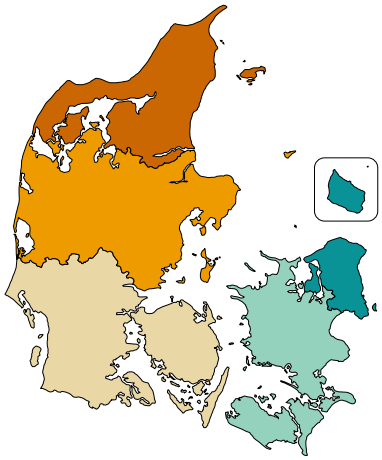


Figure 4.3: The five administrative regions of Denmark, each in charge of the regional hospitals, are: (●) the Capital Region of Denmark, (●) Region Zealand, (●) Region of Southern Denmark, (●) Central Denmark Region, and (●) Northern Denmark Region

[19] Cengiz Özcan et al. 'The Danish Heart Registry'. *Clinical Epidemiology* (2016).

measures for function and safety.[20]

The EDHR contains a number of key features. For Study II and Study III we used the EDHR to obtain information from the CAG procedures detailing the extent of stenosis in the coronary vasculature, both the amount of stenosis[21]aswell as the number of significant lesions. Additionally, the EDHR contains important prognostic information on the patients including tobacco usage, body-mass index, familial history, and various cardiological classifications—such as e.g. the Killip classification[22] or Canadian Cardiovascular Society (CCS) grading.[23] None of these factors are readily available in structured form from any of the other included registries.

[20] *The Danish Clinical Quality Program (RKKP)*. URL: rkkp.dk/in-english/ (visited on 28/11/2023).

[22] Thomas Killip and John T. Kimball. 'Treatment of Myocardial Infarction in a Coronary Care Unit: A Two Year Experience with 250 Patients'. *The American Journal of Cardiology* (1967)

[23] Todd J. Anderson et al. '2012 Update of the Canadian Cardiovascular Society Guidelines for the Diagnosis and Treatment of Dyslipidemia for the Prevention of Cardiovascular Disease in the Adult'. *Canadian Journal of Cardiology* (2013)

[23] Estimated visually by the physician using a well-known technique known as 'eye-balling'

4.5 *BigTempHealth: a Database of Electronic Health Records*

This thesis project relies on data from a unique resource: BigTempHealth project (BTH). This data comprises a complete extract from EHR systems used in the Capital Region of Denmark and Region Zealand (fig. 4.3) from 2006 to 2016. It is a diverse dataset including laboratory test results, administrative, medication, and image data, along with large collection of unstructured clinical notes.

Applications of this multifacted dataset are extensive. It has aided in text mining for extraction of adverse drug events,[24,25] contributed to a large-scale study on polypharmacy and drug–drug interactions,[26] and facilitated large-scale analyses of seasonal laboratory test variations,[27] among other uses.

In this thesis project, we incorporated laboratory results and data from clinical notes as key BTH-derived modalities in the included studies.

[24] Benjamin Skov Kaas-Hansen et al. 'Language-Agnostic Pharmacovigilant Text Mining to Elicit Side Effects from Clinical Notes and Hospital Medication Records'. *Basic & Clinical Pharmacology & Toxicology* (2022)

[25] Freja Karuna Hemmingsen Sørup et al. 'Sex Differences in Text-Mined Possible Adverse Drug Events Associated with Drugs for Psychosis'. *Journal of Psychopharmacology* (2020)

[26] Cristina Leal Rodríguez et al. 'Drug Dosage Modifications in 24 Million In-Patient Prescriptions Covering Eight Years: A Danish Population-Wide Study of Polypharmacy'. *PLOS Digital Health* (2023).

[27] Victorine P. Muse et al. 'Population-Wide Analysis of Hospital Laboratory Tests to Assess Seasonal Variation and Temporal Reference Interval Modification'. *Patterns* (2023).

[28] Muse et al., see n. 27.

4.5.1 *Laboratory Test Databases*

The laboratory or biochemical tests data originated from the two databases Clinical Laboratory System (LABKA) and Clinical Chemistry Laboratory System (BCC), used in the Capital Region and Region Zealand respectively.

A total of 339 609 717 records were obtained from the two systems and subsequently preprocessed and harmonised as detailed in Muse et al.[28] The LABKA system contains test results from 2009 to mid 2016, and BCC system from 2011 to mid 2016 as seen in fig. 4.4.

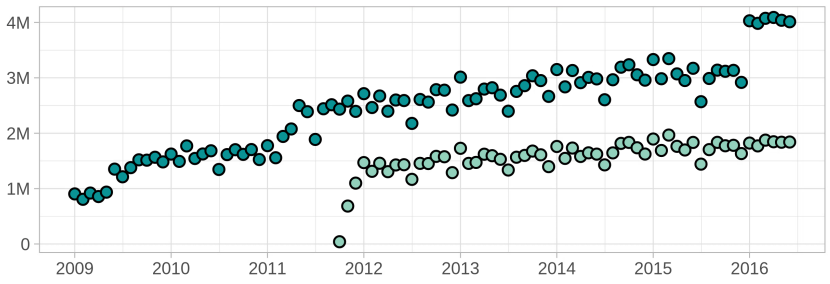


Figure 4.4: Monthly number of laboratory measurements in BigTempHealth project (BTH) originating from (●) the LABKA and (◯) the BCC laboratory systems used in the Capital Region and Region Zealand respectively.

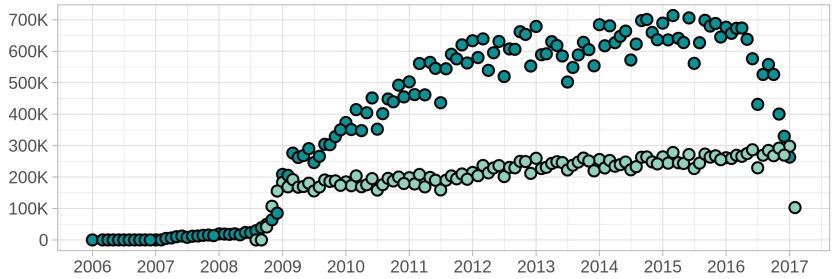


Figure 4.5: Monthly number of clinical notes in BigTempHealth project (BTH) originating from (●) the Capital Region hospitals and (●) the Region Zealand hospitals.

The laboratory tests results in the database is registered using the Nomenclature, Properties, and Units (NPU) system, which is an international terminology curated and maintained by a commitee operating under International Union of Pure and Applied Chemistry (IUPAC) and International Federation of Clinical Chemistry and Laboratory Medicine (IFCC).[29] The NPU terminology is analagous to the Logical Observation Identifiers Names and Codes (LOINC) which have been widely adopted in several countries, including the United States, Switzerland, Australia, and Germany.[30]

[29] Johan Frederik Håkonsen Arendt et al. 'Existing Data Sources in Clinical Epidemiology: Laboratory Information System Databases in Denmark'. *Clinical Epidemiology* (2020).

[30] Clement J McDonald et al. 'LOINC, a Universal Standard for Identifying Laboratory Observations: A 5-Year Update'. *Clinical Chemistry* (2003).

4.5.2 *Patient Notes*

An important hallmark of the BTH project is its inclusion of an extensive corpus of unstructured clinical text. This corpus comprises 74 336 119 unique entries, spanning from 2006 to 2017, with notably lower number of entries from 2006 to 2009, as illustrated in fig. 4.5.

Originally, these clinical notes serve to meticulously record the individual patient's healthcare trajectory, detailing symptoms, observations by medical professionals, test results, and treatment responses. While such data is not initially intended for research, a main objective of the BTH project was to harness these notes as a rich source of phenotypic data, crucial for the advancement of precision medicine.

As detailed in the review by Jensen et al.,[31] clinical text represents the most challenging EHR data modality for computational analysis, since it is inherently non-structured. Deriving structured information from written text generally involves the use of natural language processing (NLP) methods.[32]

One such method is named entity recognition (NER), a type of information extraction used in NLP that involves identification and classification of predefined entities, typically recorded in large dictionaries of known 'concepts'. In the context of clinical text, entities could include diagnoses, symptoms, and medication as illustrated in fig. 4.6. Current and prior members in the Brunak group have used NER in many applications, for example, Sørup et al.,[33] Hjaltelin et al.,[34] and Kirk et al.[35]

However, one important feature missing from these earlier applications is extraction of key–value pairs; as also highlighted in fig. 4.6, a typical clinical note can contain semi-structured stretches of text, with strings such as 'blood pressure is 149/91' or 'bmi=25.2'. To extract this

[31] Peter B. Jensen et al. 'Mining Electronic Health Records: Towards Better Research Applications and Clinical Care'. *Nature Reviews Genetics* (2012).

[32] Jensen et al., see n. 31.

[33] Sørup et al., see n. 25.

[34] Jessica Xin Hjaltelin et al. 'Pancreatic Cancer Symptom Trajectories from Danish Registry Data and Free Text in Electronic Health Records'. *eLife* (2023).

[35] Isa Kristina Kirk et al. 'Linking Glycemic Dysregulation in Diabetes to Symptoms, Comorbidities, and Genetics through EHR Data Mining'. *eLife* (2019).

clinically highly-relevant data, I have as side-project developed an approach based on regular expressions for such key–value extraction. Regular expressions, or regexes in short, is a type of concise metalanguage that specifies a search pattern for text.[36]

Using an iterative process, I constructed detailed regular expressions to extract blood pressure, height, weight, body-mass index, pack-years, pulse, and temperature from the large BTH corpus with high specificity.

We have not yet decided if this approach should be published, and no manuscript detailing its development is therefore included here. The underlying idea is neither novel, see e.g. Turchin et al.[37] for a very similar example, nor particularly interesting. However, we found it to be surprisingly useful as a data-resource for ongoing research projects. As such, the clinical information extracted using this approach have been utilized as prognostic factors in Study II and Study III.

[36] *Regular Expression*. *Wikipedia*. 2023. URL: en.wikipedia.org/w/index.php?title=Regular_expression&oldid=1186625751 (visited on 29/11/2023).

[37] Alexander Turchin et al. 'Using Regular Expressions to Abstract Blood Pressure and Treatment Intensification Information from the Text of Physician Notes'. *Journal of the American Medical Informatics Association : JAMIA* (2006).

```
68-year-old woman, no earlier known cardiovascular disease,
referred by gp for observation of angina pectoris.

Risk factors:
Hypertension: Yes, newly discovered, currently well-treated.
Hypercholesterolemia: Yes. Under treatment.
Family: No family history of ischemic heart disease.
Smoking: Quit 15 years ago, before that 20 pack-years.
Claudication: No.

Previously:
Known since 2001 with type II diabetes mellitus, treated
with Metformin.  Followed up with regular checks.
Complications with discrete neuropathy and beginning
macular degeneration.

Currently:
For about 3 months, intermittent pressure in the left side
of the chest. No radiation, independent of physical
activity. Daily attacks lasting a few seconds. During the
same period, has been diagnosed with severely elevated blood
pressure. Recently started anti-hypertensive treatment with
good effect.

Medication:
- Metformin 1000 mg x 2
- Simvastatin 20 mg x 1
...

Objective:
Normal general condition.
Nutritional status above average.
Blood pressure 149/91, pulse 76, height 177 cm, weight 96 kg
```

Figure 4.6: Artificial example of unstructured text in an EHR showcasing different types of information that can be extracted for machine learning (ML) applications.

PART II

Outline of Studies

Chapter 5

Study I: Comorbidity Clustering in Ischemic Heart Disease

In this chapter, I provide a summary of the work from Study I. I describe the background and rationale, outline essential methodological details, and discuss the main research findings.

The manuscript, titled 'Subgrouping multimorbid patients with ischemic heart disease by means of unsupervised clustering: A cohort study of 72,249 patients defined comprehensively by diagnoses prior to presentation', is currently under revision. An earlier version have been deposited on the medRxiv preprint server.[1] The revised, full-length manuscript is included in appendix A.

[1] Amalie D. Haue et al. 'Subgrouping Multimorbid Patients with Ischemic Heart Disease by Means of Unsupervised Clustering: A Cohort Study of 72,249 Patients Defined Comprehensively by Diagnoses Prior to Presentation'. *medRxiv* (2023). URL: medrxiv.org/content/10.1101/2023.03.31.23288006v2. preprint.

5.1 *Background and Rationale*

Ischemic heart disease (IHD) is highly heterogeneous in its onset, burden, and progression. As delineated in chapter 1, its manifestations range from acute myocardial infarction (AMI) to slowly progressing chronic coronary syndromes. This heterogeneity is partly explained, and further complicated, by the fact that most patients with IHD have one or more comorbidities. Current clinical practice is historically mainly based on a single-disease paradigm and thus, complexities imposed by concurrent comorbid diseases are therefore often overlooked.[2]

[2] Daniel E. Forman et al. 'Multimorbidity in Older Adults With Cardiovascular Disease'. *Journal of the American College of Cardiology* (2018).

In this study, we sought to characterise the spectrum of multimorbidity in IHD. We adopted a data-driven strategy, using unsupervised machine learning methods to identify and characterize subgroups with distinct comorbidity patterns. Our hypothesis centered on the notion that the variety and types of comorbidities, here classified according to ICD-10 codes, could facilitate the identification of distinct and clinically relevant patient clusters in IHD.

5.2 *Study Design and Outcomes*

We linked the BigTempHealth project (BTH) dataset to the Danish National Patient Register (LPR) and Danish Register of Causes of Death (DAR), and identified all patients with an ICD-10 code for IHD, who underwent coronary angiography (CAG) or coronary computed tomography angiography (CCTA) between 2004 and 2016 (72 249 patients in total). We used the date of the first CAG or CCTA as the index date. All ICD-10 codes before this date were collected for clustering analysis, excluding any IHD codes (ICD-10: I20-25).

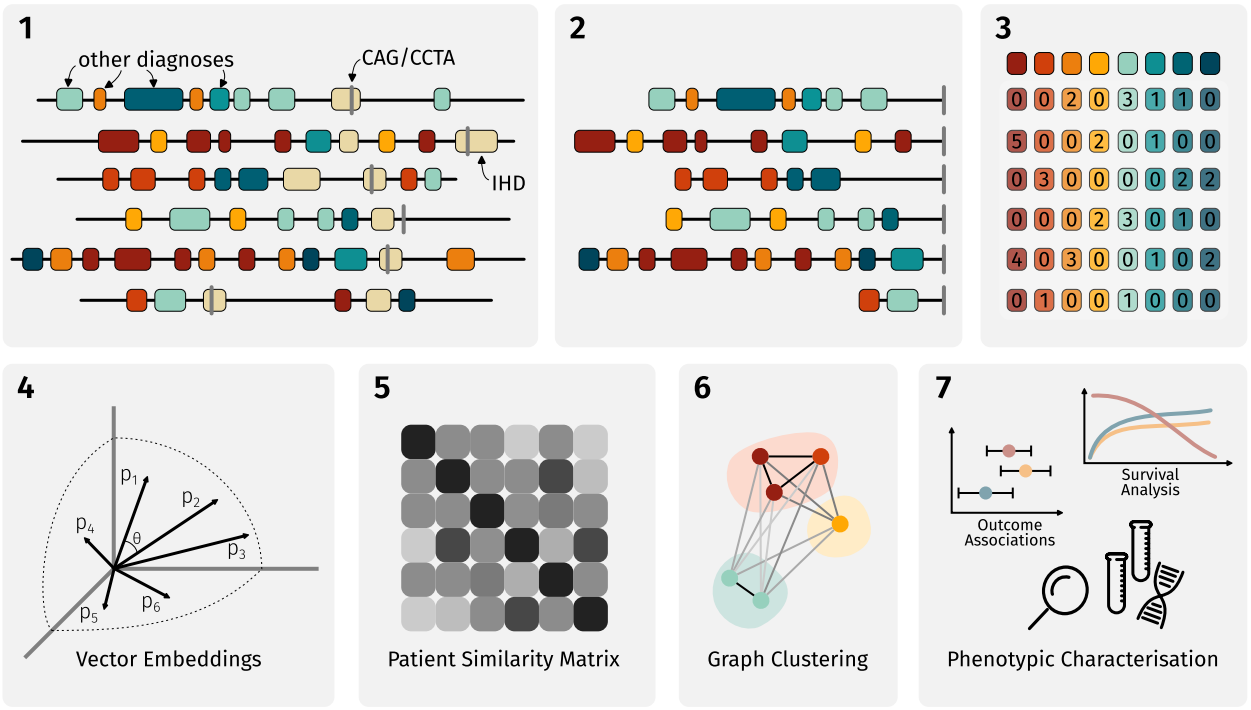


Figure 5.1: Overview of the comorbidity clustering approach in Study I, detailing the different steps going from LPR patient record data to identification and characterisation of distinct multimorbidity clusters.

In our study, we defined two main outcomes to assess the risk profiles of the different multimorbidity subgroups: (i) new ischemic events and (ii) mortality from non-IHD causes. New ischemic events, a composite outcome, included (a) hospital admission for AMI or unstable angina (UA) after 30 days of follow-up (b) revascularization procedures unrelated to the index CAG/CCTA, and (c) any deaths with IHD as the primary or secondary cause registered on the death certificate.

We used days since the index procedure as the time-scale and limited follow-up to at most five years. The two main outcomes were treated as competing risks.

5.3 *Methodology*

The overall methodology employed in this study is illustrated in fig. 5.1. In the following, I will be outlining the different steps of our comorbidity clustering approach.

In this study, we represented multimorbidity by constructing patient-level vectors that aggregated all diagnosis codes assigned up until the index date (as illustrated by steps 1 and 2 in fig. 5.1). We excluded IHD codes (ICD-10: I20-25) and codes belonging from chapters XV, XVI, XVII, XIX, XX, and XXI.[3] In addition, we removed rarely used codes assigned to less than five patients.

The remaining codes were then counted (step 3), and embedded in a vector space model[4] using singular value decomposition (SVD)[5] (step 4). Next, we used these embedded patient-level vectors to create a patient similarity matrix (step 5). For this matrix, we used cosine similarity

[3] These chapters are in the Danish ICD-10 version defined as

- Chapter XV (O00-O99): Pregnancy, childbirth and the puerperium
- Chapter XVI (P00-P99): Certain conditions originating in the perinatal period
- Chapter XVII (Q00-Q99): Congenital malformations, deformations and chromosomal abnormalities
- Chapter XIX (S00-T98): Injury, poisoning and certain other consequences of external causes
- Chapter XX (X60-Y09): External causes of morbidity and mortality
- Chapter XXI (Z00-Z99): Factors influencing health status and contact with health services

[4] G. Salton et al. 'A Vector Space Model for Automatic Indexing'. *Communications of the ACM* (1975).

[5] G. H. Golub and C. Reinsch. 'Singular Value Decomposition and Least Squares Solutions'. *Handbook for Automatic Computation: Volume II: Linear Algebra*. Springer, 1971.

as the similarity measure, which calculates the cosine of the angle θ between the embedded vectors.

From the similarity matrix we could then construct a patient similarity network, which is a weighted undirected graph with patients as vertices. The edges in the graph represents the connections between patients, which is weighted by the similarity of their respective diagnosis vectors. As this graph could contain a total of 2 609 922 876 edges, which is computationally intractable, we pruned the network by discarding low-similarity edges ($\cos\theta \leq 0.3$) and limited the number of edges connected to each vertex to the 8000 with greatest weight.

Subsequently, the patient similarity network, was then subject to cluster analysis, using the Markov clustering (MCL) algorithm[6] (step 6). The clusters obtained were then characterised (step 7). This characterisation involved four key aspects (i) estimation of hazard ratios (HRs) for cluster comparisons using Cox proportional hazards models, (ii) phenotypic enrichment analysis, (iii) examination of clusters based on laboratory test profiles, and (iv) testing for genetic associations through polygenic risk scores (PRSs). In the following I will limit the presentation of results to aspects (i) and (ii), as these are most integral to the study, and will otherwise refer to the full-length manuscript.

[6] Stijn Van Dongen. 'Graph Clustering Via a Discrete Uncoupling Process'. *SIAM Journal on Matrix Analysis and Applications* (2008).

5.4 *Main Findings*

5.4.1 *Cluster Analysis and Outcomes*

The clustering resulted in 31 distinct patient subgroups, each characterised by specific patterns of multimorbidity. We incorporated cluster membership in Cox regression models, which was further adjusted for sex and age. This was to obtain estimates for the risk associated with the comorbidity profiles in each cluster, beyond those patterns primarily dependent on age or sex. The HRs were estimated by contrasting each single cluster against all others. The size of clusters, mean age at index, and average propotion of males, and the adjusted HRs for the two outcomes, are depicted in fig. 5.2.

Our analysis revealed that certain clusters had significantly different risks compared to others. Comorbidity profiles within five clusters (C5, C9, C18, C23, C30) were associated with a significantly increased risk of new ischemic events. Conversely, profiles in two clusters (C2, C3) were associated with a significantly reduced risk of these events. Twelve clusters (C4, C5, C14, C16, C17, C18, C20, C23, C24, C25, C28, C30) had profiles associated with a significantly higher risk of non-IHD mortality. In contrast, six clusters (C2, C3, C6, C7, C11, C13) had profiles that were linked to a significantly lower risk of non-IHD mortality. Of note, four of the five cluster profiles with an increased risk of new ischemic events also exhibited an increased risk of non-IHD mortality.

5.4.2 *Phenotypic Characterisation of Clusters*

As a first step, we evaluated the intra-cluster prevalence of all unique diagnosis codes (3046 in total) within patient vectors. These prevalences

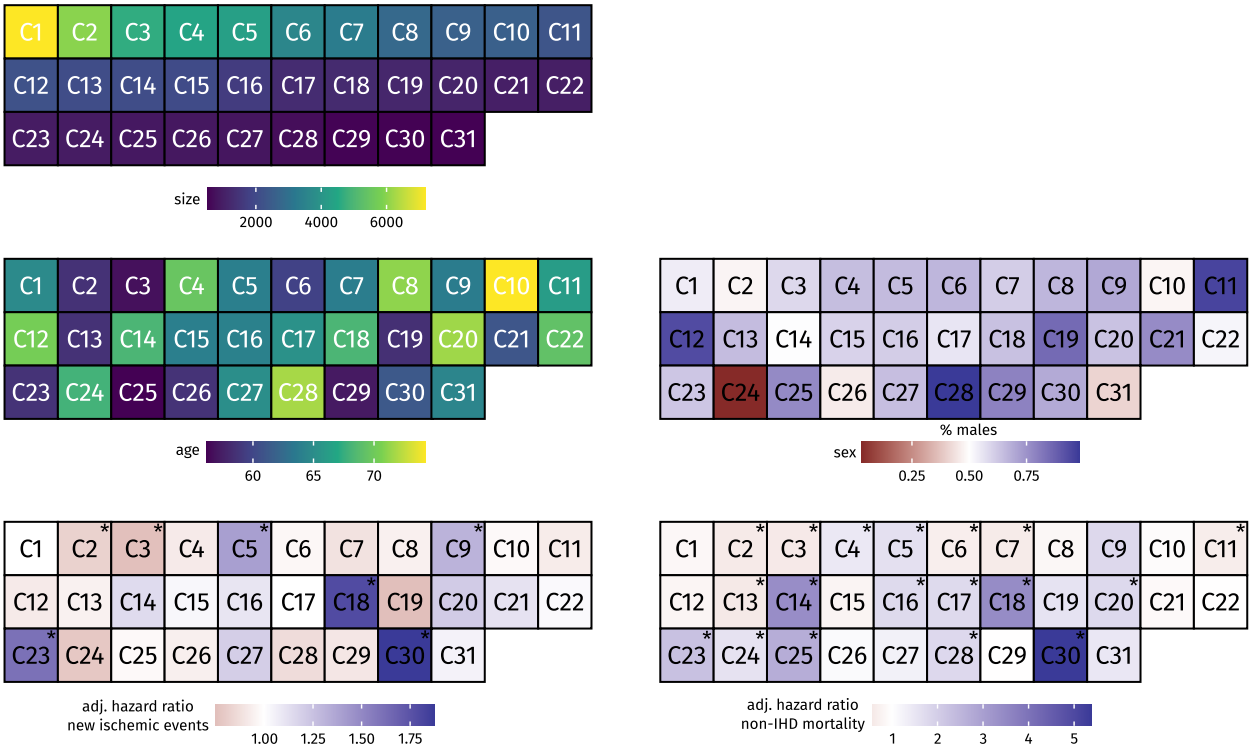


Figure 5.2: Cluster characteristics and adjusted hazard ratios. The first panel, from top-left to bottom-right, shows the number of patients in each cluster, arranged according to their respective sizes. This ordering is maintained in subsequent panels. The second panel shows the average age at the index CAG/CCTA, ranging from 56.2 years (C25) to 74.2 years (C10). The third panel shows the sex distribution in each cluster, using fill color to indicate the proportion of males; red for clusters with more than 50 % females, and blue for those with more than 50 % males. The fourth and fifth panels shows the adjusted hazard ratios for new ischemic events and non-IHD mortality, respectively. Here, clusters with a red fill indicate an HR below one, while those with a blue fill have an HR above one. Clusters with a hazard ratio significantly different from one are marked with an asterisk (*).

was then compared with the average prevalences across all clusters by calculating observed over expected (O/E)-ratios, to pinpoint ICD-10 codes that were disproportionately represented in various clusters. The top ten most overrepresented or underrepresented codes for each cluster are detailed in supplementary tables 5A and 5B of the manuscript (appendix A).

We conducted a manual review of these findings, assigning a label to each cluster based on the associated codes. Given that multiple codes could pertain to a single cluster, these labels are not definitive but offer a general overview, as depicted in fig. 5.3. Furthermore, in fig. 5.3, we organise the clusters by both their common traits and the hazard ratios derived from the Cox analyses.

5.5 *Interpretation*

From fig. 5.3, we see that the five clusters (C5, C9, C18, C23, C30) with a significantly increased risk of both new ischemic events and non-IHD mortality were characterised by the presence of important prognostic comorbidities in IHD: diabetes, peripheral atherosclerosis, heart failure, and chronic kidney disease.

Both diabetes and chronic kidney disease are highlighted as important non-cardiovascular comorbidities in the 2019 ESC guidelines for chronic coronary syndromes, and both peripheral artery disease and renal dysfunction is explicitly described as comorbidities that negatively impact prognosis.[7] Heart failure, also known to affect prognosis and included

[7] Juhani Knuuti et al. '2019 ESC Guidelines for the Diagnosis and Management of Chronic Coronary Syndromes'. *European Heart Journal* (2020)

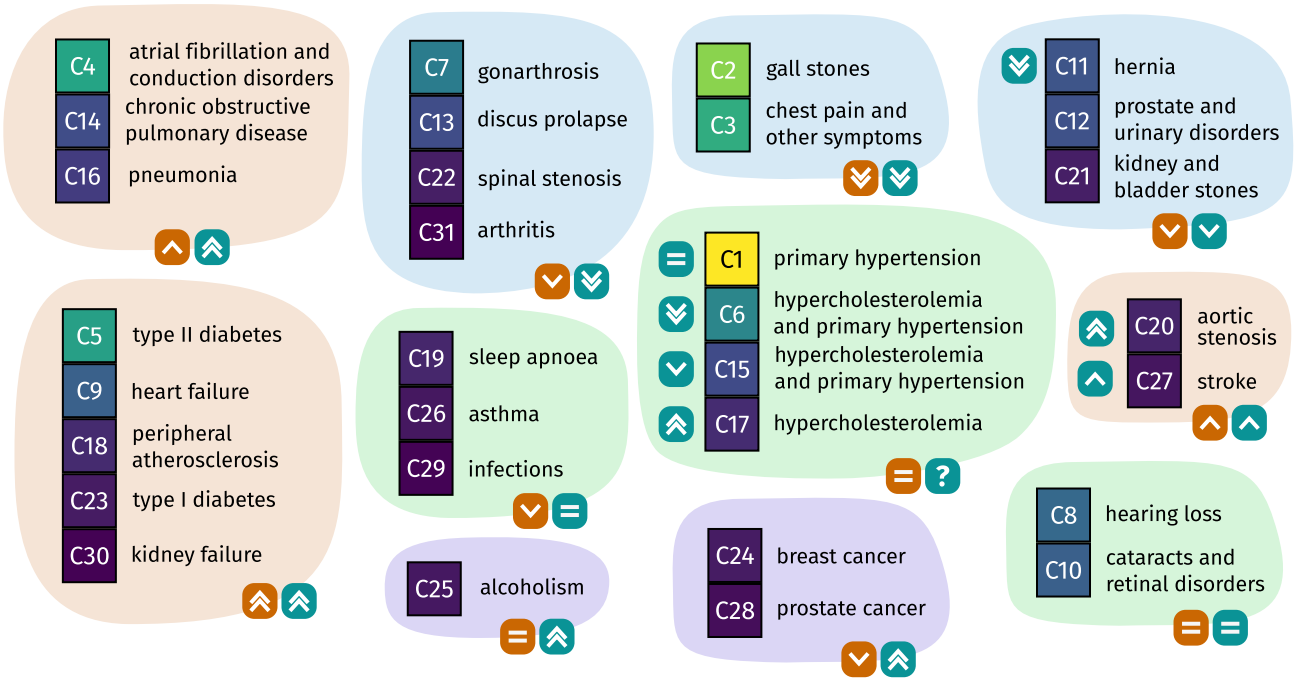


Figure 5.3: From manual review of the between-cluster enrichment of diagnosis codes, we assigned each cluster a label based on its most prevalent codes. To aid the intrepretation, the clusters were further organized by consideration of their estimated hazard ratios and shared diagnostic similarity. Clusters marked with ⏫ or ⏫ have a significantly increased risk of new ischemic events and non-IHD mortality, respectively. Conversely, ⏬ or ⏬ marks those with a significantly decreased risk. A single arrow indicates the direction of the effect and that the association was not found to be significant.

in clinical guidelines, is, for example, included as one of eight carefully selected predictors in the widely used GRACE risk and mortality calculator.[8] This showcases that our framework is able to identify known comorbidites that affect the clinical course of cardiovascular disease. In addition, it further emphasizes the significance of these comorbidities and the broader concept of considering comorbidity in clinical assessment of IHD.

We did not identify other clusters with comorbidity profiles that significicantly increased the risk of new ischemic events, however, both the second and third largest clusters (C2 and C3) were found to have comorbidities associated with a significant decreased risk of both new ischemic events and non-IHD mortality. Cluster C2 is characterised by the presence of codes for both gallstones (K80) and abdominal pain (R10). Cluster C3 relates to symptom codes (R00–R99), with high O/E-ratios for pain in throat and chest (R07.9), other chest pain (R07.3), and muscle strain (M62.6).

It is interesting that the second largest cluster (C2) is characterised by gallstones as a comorbid condition. Many studies have previously reported a link between gallstones and ischemic heart disease,[9,10,11] however, the underlying reasons for this association remain somewhat unclear and it is uncertain the link is causal. It is known that the two diseases shares pathogenicity factors, which include hypercholesterolemia, diabetes, and hypertension, so parts of the pathological mechanism is likely shared.[12]

In our study, focusing on patients with incident IHD, the presence of gallstones appear to positively affect prognosis. A possible explanation is that acute symptoms of gallstone disease can immitate heart disease symptoms, and thus, the cluster could be enriched for patients that may not have IHD to begin with. This area warrants further investigation to

[8] Keith A. A. Fox et al. 'Should Patients with Acute Coronary Disease Be Stratified for Management According to Their Risk? Derivation, External Validation and Outcomes Using the Updated GRACE Risk Score'. *BMJ Open* (2014)

[9] Yan Zheng et al. 'Gallstones and Risk of Coronary Heart Disease'. *Arteriosclerosis, Thrombosis, and Vascular Biology* (2016).

[10] S. Upala et al. 'Gallstone Disease and the Risk of Cardiovascular Disease: A Systematic Review and Meta-Analysis of Observational Studies'. *Scandinavian Journal of Surgery* (2017).

[11] Janine Wirth et al. 'Presence of Gallstones and the Risk of Cardiovascular Diseases: The EPIC-Germany Cohort Study'. *European Journal of Preventive Cardiology* (2015).

[12] Zheng et al., see n. 9.

better understand these connections and their implications for clinical practice.

Two clusters were characterised by the co-occurence of cancer co-morbidities. Clusters C24 and C28 were enriched for breast and prostate cancer, respectively. Our current methodology does not distinguish between active cancer or if the patient has been cured, which represents a key limitation of the approach. However, both clusters were found to have an increased risk of non-IHD mortality.

In patients with active cancer, the management of IHD, and specifically acute coronary syndromes (ACS), is challenged by increased risk of bleeding, low platelet count, and increased thrombotic risk.[13] Furthermore, many chemotherapeutic agents have cardiotoxic side effects, as discussed in a 2016 ESC position paper on the cardivascular toxicity of cancer treatment.[14] Moreover, studies indicate that radiotherapy breast cancer is associated with an increased risk of developing IHD, in a dose-dependent manner.[15] This underscores the complex balancing act involved in concurrently managing both conditions, where the treatment of one disease can potentially exacerbate the other.

[13] Robert A Byrne et al. '2023 ESC Guidelines for the Management of Acute Coronary Syndromes'. *European Heart Journal* (2023).

[14] Jose Luis Zamorano et al. '2016 ESC Position Paper on Cancer Treatments and Cardiovascular Toxicity Developed under the Auspices of the ESC Committee for Practice Guidelines: The Task Force for Cancer Treatments and Cardiovascular Toxicity of the European Society of Cardiology (ESC)'. *European Heart Journal* (2016).

[15] Sarah C. Darby et al. 'Risk of Ischemic Heart Disease in Women after Radiotherapy for Breast Cancer'. *New England Journal of Medicine* (2013).

5.6 *Conclusion*

In this study, we presented a large-scale data-driven approach for analysis of comorbidity patterns in more than 70 000 adult patients with incident IHD. We took a hypothesis-free approach and used a broad definition of multimorbidity, including more than 3000 different ICD-10 codes in the decription of prior and coexisting comorbidities in IHD. Using unsupervised clustering, we identified disctinct groups of patients each characterised by specific patterns of multimorbidity and associated risks of both disease progression and mortality from unrelated causes.

Using this approach, we were able to identify clusters characterised by the presence of well-established prognostic comorbidities, including diabetes, peripheral atherosclerosis, heart failure, and chronic kidney disease. All clusters associated with these diseases were all found to be significantly associated with an increased risk of adverse events. These findings thus represents a form of positive control which supports the validity of the described methodology.

It is important to emphasize that the presented clustering is not intended to provide the definitive or universally applicable multimorbidity subgroups in IHD. Instead, the purpose and implication is to provide a valuable tool for the data-driven exploration of real-world multimorbidity patterns in IHD. As such, it can be used for generating hypotheses and can likely inform and guide future research on multimorbidity-informed treatment and management of IHD.

Mapping out the landscape of multimorbidities in a real-world cohort of patients with IHD, could inform clinical managment and could serve as a tool for identification of comorbidity combinations for which current clinical knowledge is currently limited. Strict inclusion and exclusion criteria in many randomized clinical trials (RCTs) potentially limit the applicability of existing IHD management guidelines on patients

with pronounced multimorbidity.[16] To adress such limitations, an important first step is to obtain an overview of the specific patterns of multimorbidity associated with IHD.

[16] Michael W. Rich et al. 'Knowledge Gaps in Cardiovascular Care of the Older Adult Population'. *Circulation* (2016).

Chapter 6

Study II: Time-to-Event Prediction of All-Cause Mortality

In this chapter, I provide a summary of the work from Study II. I describe the background and rationale, outline essential methodological details, and discuss the main research findings.

The manuscript, titled 'Development and validation of a neural network-based survival model for mortality in ischemic heart disease', is currently under review (2nd round). An earlier version of the manuscript has been deposited on the medRxiv preprint server.[1] The revised full-length manuscript is included in appendix B.

[1] Peter C. Holm et al. 'Development and Validation of a Neural Network-Based Survival Model for Mortality in Ischemic Heart Disease'. *medRxiv* (2023). URL: medrxiv.org/content/10.1101/2023.06.16.23291527v1. preprint.

6.1 *Background and Objectives*

For patients with ischemic heart disease (IHD), the use of risk stratification algorithms, which assess the presence or absence of prognostic risk factors and disease indicators, can be crucial in tailoring patient-specific treatment approaches.[2,3,4] Notably, the revised GRACE risk score (GRACE 2.0),[5] which has been widely adopted in the clinical setting and received a class IIa recommendation for assessing and managing patients with non-ST-elevation myocardial infarction (NSTEMI) in recent ESC guidelines,[6] serves as a prime example of such an algorithm.

Presently, risk stratification algorithms for secondary prevention in IHD, including GRACE 2.0, are limited to a narrow range of established risk factors and are constructed using traditional statistical methods like the Cox proportional hazards model or logistic regression. These conventional models, while useful, have strong assumptions, lack flexibility, are imprecise, and are not well-suited for integration of large and heterogenous arrays of predictors—at least not without extensive feature engineering and selection.

In this study, we hypothesised that prognostication models in IHD would benefit from the inclusion of a much wider array of features, which is already available from electronic health record (EHR) systems in clinical use. To enable the inclusion of such features, and also allow the modelling of time-to-event outcomes with potential censoring, we set out use a neural network-based survival model. Specifically, the discrete time framework proposed by Gensheimer and Narasimhan,[7] which I have described in detail in chapter 3.

Since this framework does not allow inclusion of competing risks, we limited the scope of the study to only consider prediction of all-cause mortality.

[2] Juhani Knuuti et al. '2019 ESC Guidelines for the Diagnosis and Management of Chronic Coronary Syndromes'. *European Heart Journal* (2020).

[3] Jean-Philippe Collet et al. '2020 ESC Guidelines for the Management of Acute Coronary Syndromes in Patients Presenting without Persistent ST-segment Elevation'. *European Heart Journal* (2021).

[4] Ph. Gabriel Steg et al. 'ESC Guidelines for the Management of Acute Myocardial Infarction in Patients Presenting with ST-segment Elevation'. *European Heart Journal* (2012).

[5] Keith A. A. Fox et al. 'Should Patients with Acute Coronary Disease Be Stratified for Management According to Their Risk? Derivation, External Validation and Outcomes Using the Updated GRACE Risk Score'. *BMJ Open* (2014).

[6] Collet et al., see n. 3.

[7] Michael F. Gensheimer and Balasubramanian Narasimhan. 'A Scalable Discrete-Time Survival Model for Neural Networks'. *PeerJ* (2019).

6.2 Study Design

For model development, we defined a retrospetive cohort from the Danish National Patient Register (LPR) and Eastern Denmark Heart Registry (EDHR) which consisted of 39 746 adult patients with IHD.

All patients in the cohort had their first coronary angiography (CAG) performed between 1st of January 2006 and 7th of July 2016, which were required to have led to a diagnosis of one-, two-, or three-vessel disease (1-3VD) or diffuse atheromatosis (DA).

We defined the index date as the date of the inclusion CAG, and included five years of follow-up from the Danish Civil Registration System (CPR), which was used to define all-cause mortality.

The development cohort was randomly split into a training set and a hold-out test set consisting of 34 746 and 5000 patients, respectively.

For external validation, a cohort of 8287 Icelandic adults with IHD who had undergone CAG were similarly defined, as detailed in the methods section of the manuscript (see appendix B).

6.3 Methodology

Using the discrete time logistic-hazard approach described in chapter 3, we trained neural network models to predict the probability of all-cause mortality within five years after the index CAG. The model output are discrete time conditional hazards, and as such, can be used to construct a survival function that gives the estimated probability at any timepoint within the prediction horizon.

Models were trained using only the training set, and the hyperparameters were tuned using the hyperparameter optimization (HPO) library Optuna in Python.[8] For the HPO, we used 5-fold cross-validation to obtain a bias-corrected estimate of the model performance.

[8] Takuya Akiba et al. 'Optuna: A Next-generation Hyperparameter Optimization Framework'. KDD '19: The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2019.

Different intermediate models were trained using various combinations of the included features listed in table 6.1, to explore their respective impact on model performance. For the complete model, which we refer to as `PMHnetV1`, we included all available features.

	Category	Features
	ClinicalOne	age, pulse, systolic blood pressure, cardiac arrest (yes/no), abnormal cardiac enzymes (yes/no), killip class, serum creatinine, ST-segment deviation (yes/no)
	ClinicalTwo	abnormal ECG (yes/no), CCS, diastolic blood pressure, coronary artery dominance, familial IHD (yes/no), height, weight, ICD-device or pacemaker (yes/no), ischemia test, LVEF, NYHA class, sex (male/female), smoking status, coronary pathology (1-3VD or DA)
	Diagnoses	322 different level-3 ICD-10 diagnosis codes
	Procedures	154 NOMESCO procedure codes corresponding to different examinations and surgical procedures
	Biochemical	85 different lab tests with results categorized as below, within, or above the reference range

Table 6.1: Overview of `PMHnetV1` input features. The different features were grouped into five different categories according to their respective characteristics. The 'clinical' features were separated into two different subgroups, where 'ClinicalOne' consist of the exact same features as those used in the GRACE 2.0 score.

We evaluated model performance by assessing both model discrimination and calibration. Model discrimination was quantified by calculation of time-dependent area under the receiver receiver operating characteristic (td-AUC), which has the same interpretation as the commonly used *c*-index, but is instead appropriate to use for the evaluation of *t*-year predicted risk, as described in Blanche et al.[9] We assessed model calibration by constructing calibration curves, comparing model predictions with estimated actual risks, and by computing the Brier score. For these model evaluation metrics, we use the implementations provided by the R-package `riskRegression`.[10] All performance metrics were exclusively calculated using the internal or external test set.

Model performance was compared to that of the GRACE 2.0 risk score.[11] The GRACE 2.0 score was calculated by extracting the source code from the GRACE 2.0 webtool and deploying a local R-based equivalent.[12] Since GRACE 2.0 does not allow missing values, and all eight variables it uses was only available for 51.4 % of our cohort, we imputed missing values using the `missForest` method prior to calculation.[13] Additionally, since GRACE 2.0 is not explicitly developed to provide predictions at the index time used in our study, we also trained a neural network that was limited to consider only the GRACE 2.0 features (corresponding to the 'ClinicalOne' features in table 6.1), which I will refer to as GRACE 2.0 (re-fitted). In this model, as well as in `PMHnetV1`, missing features were left missing as the neural network was designed to handle such.

Lastly, to provide model explanations and further investigate individual features' impact on the model predictions, we performed SHAP analyses. SHAP is a model-agnostic approach that can provide measures of feature importance in otherwise 'black-box' machine learning (ML) models (see section 2.9.2).

[9] Paul Blanche et al. 'The C-Index Is Not Proper for the Evaluation of t-Year Predicted Risks'. *Biostatistics (Oxford, England)* (2019).

[10] Thomas A. Gerds and Michael W. Kattan. *Medical Risk Prediction Models: With Ties to Machine Learning*. Chapman and Hall/CRC, 2021.

[11] Fox et al., see n. 5.

[12] *GRACE Risk Score 2.0*. URL: `outcomes-umassmed.org/grace/acs_risk2/index.html` (visited on 12/12/2023).

[13] Daniel J. Stekhoven and Peter Bühlmann. 'MissForest—Non-Parametric Missing Value Imputation for Mixed-Type Data'. *Bioinformatics* (2012).

6.4 *Main Findings*

From the model evaluation, we found the `PMHnetV1` model to provide excellent model discrimination, outperforming both the standard GRACE 2.0 score and our neural network model using the GRACE 2.0 features (table 6.2). We also found the model to be well calibrated as exemplified by fig. 6.1, which shows the calibration curves for the 3-year predictions. The GRACE 2.0 score was miscalibrated, but our re-fitted version had comparable calibration to the `PMHnetV1` model.

To further assess the generalisability of the model, we performed

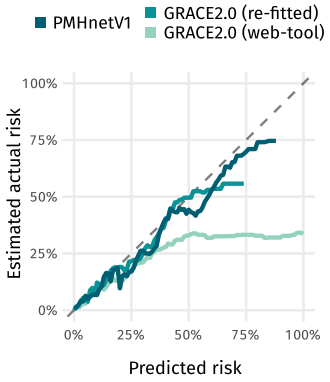


Figure 6.1: Calibration curve for the `PMHnetV1` model and the GRACE 2.0 reference models at a prediction horizon of three years.

	6 months		1 year		3 years		5 years	
	td-AUC	(95%CI)	td-AUC	(95%CI)	td-AUC	(95%CI)	td-AUC	(95%CI)
`PMHnetV1`	0.88	(0.86–0.90)	0.88	(0.86–0.90)	0.84	(0.82–0.86)	0.82	(0.80–0.84)
GRACE 2.0 (web tool)	0.77	(0.74–0.80)	0.77	(0.74–0.80)	0.73	(0.71–0.75)	–	–
GRACE 2.0 (re-fitted)	0.79	(0.76–0.83)	0.78	(0.75–0.81)	0.76	(0.74–0.78)	–	–

Table 6.2: td-AUC scores for `PMHnetV1`, GRACE 2.0 (web tool), and GRACE 2.0 (re-fitted) at four different prediction horizons. The standard GRACE 2.0 score does not predict 5-year survival and an td-AUC was therefore not calculated at that prediction horizon.

external validation on an Icelandic cohort of 8287 patients. Due to data availability, we used a slightly down-scaled version of the model that was limited to the 404 features that could be obtained from the Icelandic data. In the internal test set, the td-AUC of this model was 0.87 (0.85–0.90) for the 6 month prediction, 0.87 (0.85–0.89) for the 1 year prediction, and 0.82 (0.80–0.85) for the 3 year prediction. On the Icelandic data, the td-AUC was 0.87 (0.84–0.90) for the 6 month prediction, 0.84 (0.81–0.87) for the 1 year prediction, and 0.81 (0.79–0.85) for the 3 year prediction. We acknowledge that model performance can not directly be compared across cohorts, but this qualitatively shows that model discrimination remained high in an external dataset.

To examine the effect of including different types of features, we calculated td-AUC of different intermediate models limited to only consider certain combinations of feature modalities. These results are presented in fig. 6.2 and shows that performance increased with increasing number of features, however, with diminishing incremental gains.

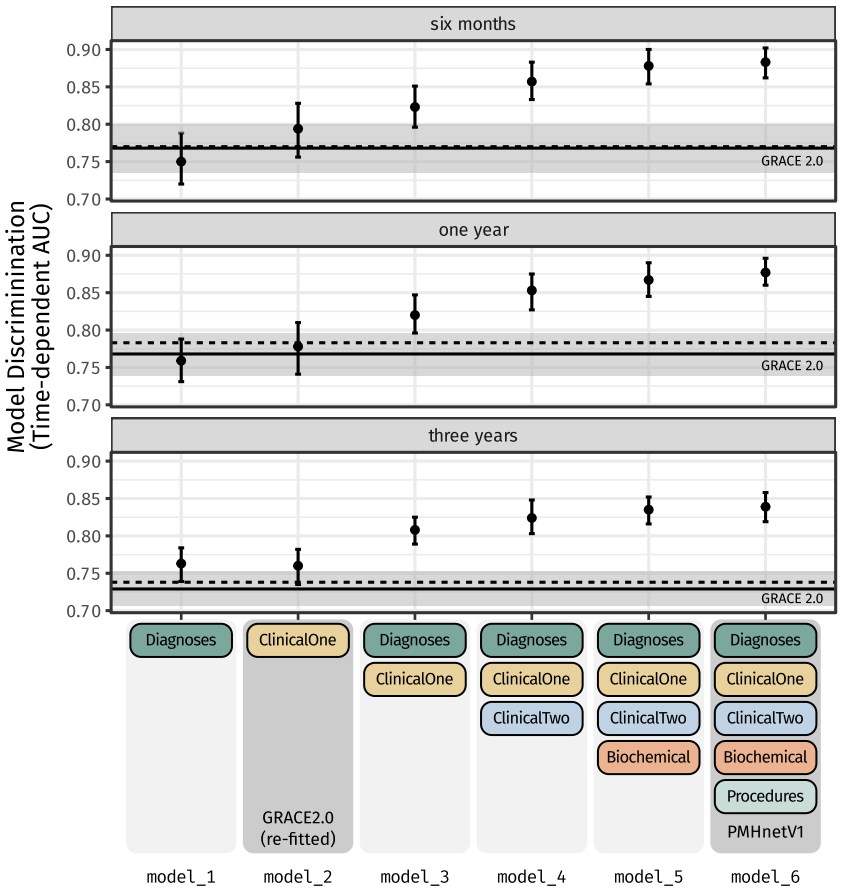


Figure 6.2: td-AUC scores of different intermediate `PMHnetV1` models each limited to specific combinations of the designated feature categories. The horizontal reference lines indicate the performance of the GRACE 2.0 score on the same data (the test set). The solid line is the td-AUC of GRACE 2.0 on all patients, and the dashed line is for the subset of patients where imputation was not required.

Lastly, we performed SHAP analyses to explain the impact of different features on model predictions. With SHAP, each feature for each patient is assigned a value measuring the impact on the model output, which enables the construction of model explanations at different levels of granularity. Figure 6.3, shows local patient-level explanations for three different representative examples, which in a clinical setting could be used to summarise the most impactful predictors for a single patient.

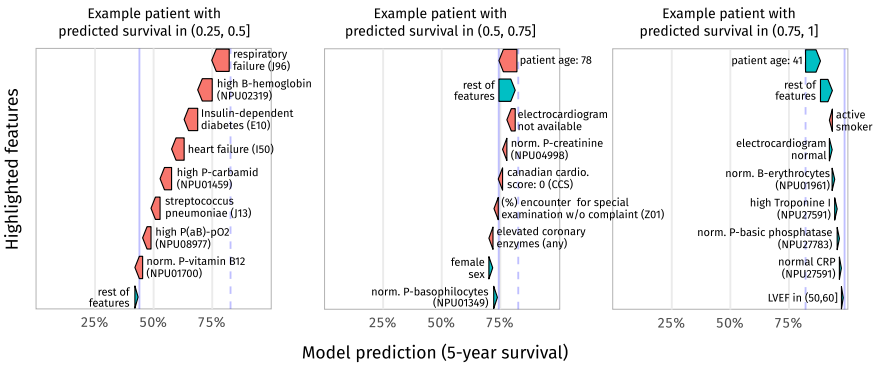


Figure 6.3: Local patient-level SHAP explanations of 5-year predictions from PMHnetV1. The three panels shows an explanation of the PMHnetV1 prediction for test-set patients with different predicted probability of survival, where each arrow shows the SHAP estimated impact of the labeled feature on the specific prediction. The included patient data have manually been adjusted as to make it non person-sensitive. The vertical dashed lines is the median prediction, and the solid line is the prediction for each patient, which one can obtain by adding all SHAP values to the median prediction.

Aggregating all SHAP values for each patient across features produces provides a more global view of feature importance (fig. 6.4), which shows that on average, the 'Biochemical' feature category has the highest impact on model output, and that 'age', expectedly, is the most impactful feature. Using the impact of 'age' as an example, we also explored the link between the specific feature values and the corresponding SHAP values (fig. 6.5).

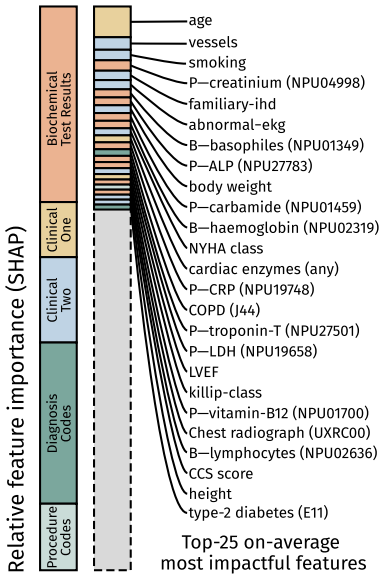


Figure 6.4: By summarising the magnitude of SHAP values, we obtained an overview of the relative impact of the different PMHnetV1 features, either aggregated across feature categories (left) or across each individual features (right).

6.5 *Conclusion*

In this study, we presented the development and validation of a neural network-based discrete time-to-event model for prediction of all-cause mortality in patients with IHD. Our ML model, PMHnetV1, was found to be well-calibrated and to have excellent model discrimination, also in a external cohort of IHD patients from another country. Compared to the well-established GRACE 2.0 risk-prediction algorithm, PMHnetV1 was found to be significantly better at prediction of all-cause mortality. Furthermore, we showed that by including and utilising a broad array of input features, we obtain risk-prediction models with better performance compared to models only considering a single feature modality and those limited to a selection of only well-known risk factors.

The precise identification of patients at either end of the risk-spectrum, might be used to select patients likely to benefit from more extensive clinical management aswell as patients for whom less treatment perhaps would be the better option. However, prospective studies are needed to determine its impact and clinical utility. Although not part of the included manuscript, I have been working together with the public company CIMT in the implementation of PMHnetV1 in 'Sundhedsplatformen', the EHR system used in the Capital Region and Region Zealand hospitals. We succeeded in implementing the model, and there is currently an ongoing randomized clinical trial (RCT) such that the model can be clinically tested.[14]

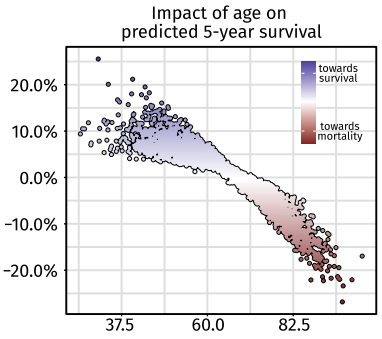


Figure 6.5: Example of a SHAP dependence plot showing the SHAP value of each 'age' feature across the entire test set.

[14] Henning Bundgaard. *Clinical Implementation of a Novel Decision Support Tool in Patients With Ischemic Heart Disease*. Clinical trial registration. clinicaltrials.gov, 2023. URL: clinicaltrials.gov/study/NCT06033014 (visited on 01/01/2023).

Chapter 7

Study III: Time-to-Event Prediction with Competing Risks

In this chapter, I provide an outline of our research in Study III. The manuscript, titled 'Development of a neural network-based competing risk model for long-term prognostication in ischemic heart disease from a large database of electronic health records and clinical registries', is currently work in progress, and thus the version included in appendix C is a draft manuscript.

7.1 *Background and Aims*

In our previous study, Study II, we demonstrated that a machine learning (ML) based time-to-event prediction algorithm can improve the prediction of all-cause mortality in patients with ischemic heart disease (IHD). While all-cause mortality is an important clinical outcome, a limitation of our previous work was the absence of more disease-specific outcomes such as cardiovascular mortality and disease progression events. The neural network-based Logistic-Hazard model employed in `PMHnetV1` is not able to model competing risks, which precluded the inclusion of such outcomes.

To address this shortcoming, and further expand on our prior work, the primary goals of this study are to develop and implement an extension to the discrete time Logistic-Hazard model from Gensheimer and Narasimhan[1] to enable joint-modelling of competing risks, and then use our novel framework in the creation of `PMHnetV2`, such that it is possible differentiate between deaths related to IHD and those arising from completely unrelated causes, in addition to predicting specific measures of disease progression.

[1] Michael F. Gensheimer and Balasubramanian Narasimhan. 'A Scalable Discrete-Time Survival Model for Neural Networks'. *PeerJ* (2019).

7.2 *The Logistic-Hazard Approach for Competing Risks*

In the following, I will outline how the discrete-time framework can be extended to allow for jointly modelling time-to-event data with competing risks. The theory underlying this approach is well-established in classical statistical literature, as exemplified by Tutz and Schmid,[2] but have to the best of our knowledge not yet been adapted to neural network models.

[2] Gerhard Tutz and Matthias Schmid. *Modeling Discrete Time-to-Event Data*. 1st ed. 2016 edition. Springer, 2016.

As delineated in section 3.4, in the discrete-time framework, continuous follow-up time T_c is divided into q contiguous intervals

$$(0, a_1], (a_1, a_2], \ldots, (a_{q-1}, a_q]$$

and $T_\mathrm{d} \in \{1, \dots, q\}$ is then a discrete random variable specifying the event time that refers to each interval $(a_{\tau-1}, a_\tau]$, and similarly, $C_\mathrm{d} \in \{1, \dots, q\}$ specifies the time of censoring.

In this framework, a right-censored survival dataset $\mathfrak{D}_\mathrm{d}$ with κ different competing risks is defined as

$$\mathfrak{D}_\mathrm{d} = \{(\tau_i, \sigma_i, \boldsymbol{x}_i) \mid i = 1, \dots, N\} \tag{7.1}$$

where $t_i = \min(T_i, C_i)$ is the observed follow-up time, $\sigma_i \in \{\varnothing, 1, \dots, \kappa\}$ is the event indicator (with $\varnothing$ specifying censored observations), and $\boldsymbol{x}_i \in \mathbb{R}^p$ is a feature vector of size p.

7.2.1 *Model Formulation*

For modelling this data, we use the discrete cause-specific hazard, which for cause r is defined as[3]

$$\lambda_r(t \mid \boldsymbol{x}) = \Pr(T_\mathrm{d} = \tau, R = r \mid T_\mathrm{d} \geq \tau, \boldsymbol{x}). \tag{7.2}$$

This hazard describes the conditional probability of experiencing event r in the interval $(a_{\tau-1}, a_\tau]$ given that the individual is still at risk at the beginning of the interval.

For κ competing risks, the survival data can be described with κ different hazard functions, $\lambda_1(\tau \mid \boldsymbol{x}), \dots, \lambda_\kappa(\tau \mid \boldsymbol{x})$. To describe the overall hazard $\lambda(\tau \mid \boldsymbol{x})$, these functions can be combined as[4]

$$\lambda(\tau \mid \boldsymbol{x}) = \sum_{r=1}^{\kappa} \lambda_r(\tau \mid \boldsymbol{x}) = \Pr(T_\mathrm{d} = \tau \mid T_\mathrm{d} \geq \tau, \boldsymbol{x}), \tag{7.3}$$

which describes the risk of experiencing any of the competing risks.

From eq. (7.3), we can obtain the survival function, which describes the probability of not experiencing any of the competing risks.

$$S(\tau \mid \boldsymbol{x}) = \Pr(T_\mathrm{d} > \tau \mid \boldsymbol{x}) = \prod_{s=1}^{\tau} (1 - \lambda(s \mid \boldsymbol{x})) \tag{7.4}$$

At each interval $(a_{\tau-1}, a_\tau]$, there are $\kappa + 1$ different possible outcomes, either one of the κ risks occurs or the individual survives and continues to the next interval, which means that the sum of these probabilities is 1.

$$\lambda_1(\tau \mid \boldsymbol{x}) + \cdots + \lambda_\kappa(\tau \mid \boldsymbol{x}) + (1 - \lambda(\tau \mid \boldsymbol{x})) = 1 \tag{7.5}$$

To model these $\kappa + 1$ events, we construct a neural network where the output is a $N \times q \times (\kappa + 1)$ matrix of 'logits'[5] as illustrated in fig. 7.1. To obtain outputs on the probability scale, the logits are passed through a Softmax activation function, such that the numbers across the dimension of the probability matrix sum to 1.

The $1 - \lambda(\tau \mid \boldsymbol{x})$ term is not strictly necessary to include, since it can be obtained from the others, however in the machine learning literature it is common practice to include all output classes in multinomial predictions. In the following, I will refer to this term as $\lambda_\varnothing(\tau \mid \boldsymbol{x})$.

[3] Tutz and Schmid, see n. 2.

[4] Tutz and Schmid, see n. 2.

[5] In the context of machine learning,the term 'logits' typically refers to the raw unnormalized output that can range from $-\infty$ to ∞. To obtain probabilities from logits, they are passed through an activation function such as the logistic or Softmax function

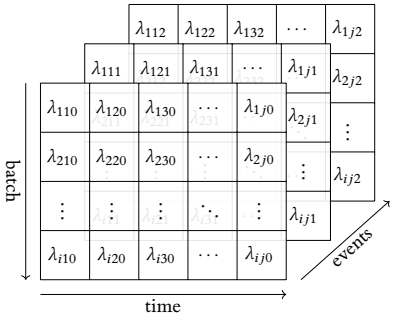


Figure 7.1: The output of the extended Logistic-Hazard model is a $N \times q \times (\kappa + 1)$ matrix of logits, which represents the cause-specific hazards.

[6] Tutz and Schmid, see n. 2.

7.2.2 *Derivation of Loss Function*

As detailed in Tutz and Schmid,[6] the contribution of the ith individual on the likelihood is

$$\mathcal{L}_i = \begin{cases} \Pr(T_\text{d} = \tau_i, R = \sigma_i \mid \boldsymbol{x}_i) \Pr(C_\text{d} \geq \tau \mid \boldsymbol{x}_i) & \text{if non-censored} \\ \Pr(T_\text{d} > \tau_i \mid \boldsymbol{x}_i) \Pr(C_\text{d} = \tau \mid \boldsymbol{x}_i) & \text{if censored.} \end{cases} \tag{7.6}$$

Assuming that censoring is non-informative, the probabilities involving the censoring time C_d can be omitted.[7] Further, we can rewrite the terms $\Pr(T_\text{d} = \tau_i, R = \sigma_i \mid \boldsymbol{x}_i)$ and $\Pr(T_\text{d} > \tau_i \mid \boldsymbol{x}_i)$ as a product of the conditional hazards

[7] Tutz and Schmid, see n. 2.

$$\begin{aligned} \Pr(T_\text{d} = \tau_i, R = \sigma_i \mid \boldsymbol{x}_i) &= \Pr(T_\text{d} = \tau_i, R = \sigma_i \mid T_\text{d} \geq \tau_i, \boldsymbol{x}_i) \Pr(T_\text{d} \geq \tau_i \mid \boldsymbol{x}_i) \\ &= \lambda_{\sigma_i}(\tau_i \mid \boldsymbol{x}_i) \Pr(T_\text{d} > \tau - 1 \mid \boldsymbol{x}_i) \\ &= \lambda_{\sigma_i}(\tau_i \mid \boldsymbol{x}_i) \textstyle\prod_{s=1}^{\tau_i - 1} (1 - \lambda(s \mid \boldsymbol{x}_i)) \\ &= \lambda_{\sigma_i}(\tau_i \mid \boldsymbol{x}_i) \textstyle\prod_{s=1}^{\tau_i - 1} \lambda_\varnothing(s \mid \boldsymbol{x}_i) \end{aligned} \tag{7.7}$$

$$\begin{aligned} \Pr(T_\text{d} > \tau_i \mid \boldsymbol{x}_i) &= \Pr(T_\text{d} > \tau_i \mid T_\text{d} \geq \tau_i, \boldsymbol{x}_i) \Pr(T_\text{d} \geq \tau_i \mid \boldsymbol{x}_i) \\ &= (1 - \Pr(T_\text{d} = \tau_i \mid T_\text{d} \geq \tau_i, \boldsymbol{x}_i)) \Pr(T_\text{d} > \tau_i - 1 \mid \boldsymbol{x}_i) \\ &= (1 - \lambda(\tau_i \mid \boldsymbol{x}_i)) \textstyle\prod_{s=1}^{\tau_i - 1} (1 - \lambda(s \mid \boldsymbol{x}_i)) \\ &= \textstyle\prod_{s=1}^{\tau_i} \lambda_\varnothing(s \mid \boldsymbol{x}_i), \end{aligned} \tag{7.8}$$

and the likelihood contribution is then

$$\mathcal{L}_i = \lambda_{\sigma_i}(\tau_i \mid \boldsymbol{x}_i) \prod_{s=1}^{\tau_i - 1} \lambda_\varnothing(s \mid \boldsymbol{x}_i) \tag{7.9}$$

To avoid computational issues with floating point precision, we use the log-likelihood instead, which becomes

$$\ell_i = \log[\lambda_{\sigma_i}(\tau_i \mid \boldsymbol{x}_i)] + \sum_{s=1}^{\tau_i - 1} \log[\lambda_\varnothing(s \mid \boldsymbol{x}_i)] \tag{7.10}$$

The total log-likelihood of all datapoints gives the loss-function used in the extended Logistic-Hazard model for competing risks, which is

$$\ell(\mathfrak{D}_\text{d}) = \sum_{i=1}^{N} \left(\log[\lambda_{\sigma_i}(\tau_i \mid \boldsymbol{x}_i)] + \sum_{s=1}^{\tau_i - 1} \log[\lambda_\varnothing(s \mid \boldsymbol{x}_i)] \right) \tag{7.11}$$

7.2.3 *Implementation*

This loss function, along with several useful classes and functions for discrete time-to-event neural networks, have been implemented in the python package `DiscoTime`.[8] This implementation relies on the `PyTorch` machine learning framework and is built to use the `PyTorch-Lightning` interface, to make protyping and experimentation as easy as possible. `DiscoTime` is available on the Python Package Index (PyPI) and on GitHub at 'peterchristofferholm/discotime'. Currently in early development, the documentation is rather limited and the code base is expected to undergo re-factorization, but we still expect that the framework in its current state is of general interest to other researchers in the field. We used version 0.1.0 of the package for this study.

[8] Peter Holm. *Discotime: Discrete-time Competing Risk Analysis with Neural Networks*. Version 0.1.0. URL: `github.com/peterchristofferholm/discotime` (visited on 25/12/2023).

7.3 Study Design and Methodology

To test the utility of the presented competing risk methodology, we set out to develop PMHnetV2, a collection of four different neural network-based time-to-event prediction models for cause-specific post-angiography prognostication in IHD.

7.3.1 Defining the Derivation Cohort

For development of the neural network models, we linked the BigTempHealth project (BTH) dataset to the Danish National Patient Register (LPR) and the Eastern Denmark Heart Registry (EDHR). From these, we identified patients who underwent a coronary angiography (CAG) which led to a diagnosis of one-, two-, or three-vessel disease or diffuse atheromatosis between January 1, 2006, and December 31, 2016. Patients were excluded if they lacked an ICD-10 code for IHD (I20-25), were under 18 years at the time of the CAG, or did not survive at least two days after the index procedure (fig. 7.2).

This cohort consists of 52 809 adults with IHD, closely resembling and considerably overlapping the one used in the PMHnetV1 study (Study II). However, unlike PMHnetV1, which only included patients undergoing their first CAG during the inclusion period, this study expanded the criteria to also include patients with a history of one or more CAGs. This approach likely provides a more comprehensive representation of the diverse manifestations of chronic coronary syndromes.

Prior to statistical analysis and model training, each patient in the development cohort was randomly allocated to the training or test split with probabilities of 80 % and 20 %, respectively. This process resulted in a training set of 42 048 patients and a test set of 10 761 patients (fig. 7.3).

Serving as model predictors,, we identified and created more than 2200 different features belonging to five different overall categories. We included 80 clinical features, 418 procedure and examination codes, 785 distinct medical prescriptions, 504 different diagnoses, and 475 unique laboratory test features. Further details on features and the preprocessing is included in the manuscript in appendix C.

7.3.2 Included Time-to-Event Endpoints

As the index date, we used the date of the inclusion CAG. For the PMHnetV2 time-to-event prediction models, follow-up was defined as the number of days between the index date and the onset of endpoints or censoring, whichever came first, with a maximum follow-up duration of five years. We defined four different primary outcomes, three of which included competing risks:

ACMO—all-cause mortality, was obtained from the Danish Civil Registration System (CPR) using the same definition as in Study II. ACMO did not have competing risks.

CVMO—cardiovascular mortality, was defined from the Danish Register of Causes of Death (DAR) as deaths where the underlying cause of death

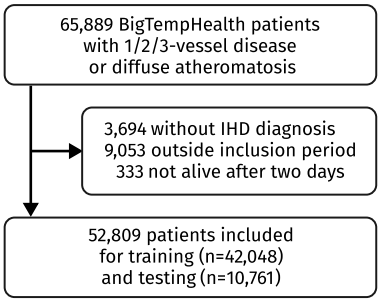


Figure 7.2: Flow diagram showing the inclusion and exclusion criteria for derivation of thè PMHnetV2 cohort. An IHD diagnosis was defined as any prior or concurrent hospital admission with a primary or secondary diagnosis code (ICD-10) of I20-25. The inclusion period ranged from 01.01.2006 to 31.12.2016, both dates inclusive.

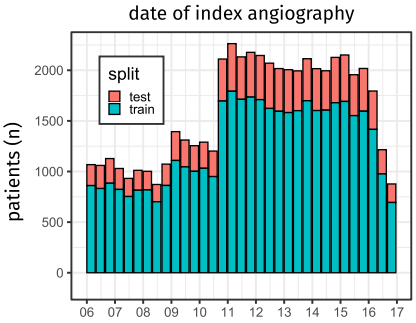


Figure 7.3: Histogram showing the temporal distribution of the PMHnetV2 index coronary angiography procedures. The bars represent the number of patients with an index event in each four-month period from 2006 to 2017, further segmented to display the proportions of train and test patients.

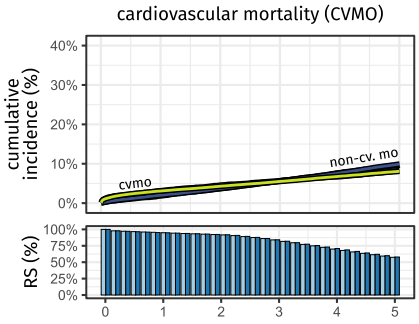


Figure 7.4: Cause-specific cumulative incidence functions for the CVMO outcome (cvmo) and the competing endpoint 'non-cardiovascular' mortality. (non-cv. mo). The bottom panels shows the proportion of train/test patients still at risk (i.e. non-censored and still alive) at each two-month period following the index CAG.

was assigned an ICD-10 code from the cardiovascular chapter (I00–99). Deaths attributed to different causes were treated as competing risks, as depicted in fig. 7.4

CVCO—cardiovascular complications, was defined from the LPR as a composite outcome covering hospital admissions with a primary diagnosis of 'heart failure' (ICD-10: I50), 'atrial fibrillation or flutter' (I48), 'cardiac arrest' (I46), and 'cerebrovascular accident' (I61, I63–64) and in-hospital procedures for 'implanation of pacemaker' (SKS: `"BFCA0*"`) and 'implanation of cardioverter-defibrillator' (`"BFCB0*"`). Events within four weeks after the index angiography was assumed to be unrelated to disease progression and was ignored. ACMO was treated as a competing risk.

MIEV—new myocardial ischemia, was defined from the LPR as unplanned percutaneous coronary interventions (PCIs), coronary artery bypass graftings (CABGs), or in-hospital admissions longer than 24 h with a primary diagnosis of IHD (I20-25). Events within the eight weeks after the index CAG were not considered 'new' events and were therefore excluded.

7.3.3 *Neural Network Architecture*

For the `PMHnetV2` models, we used a 'ResNet'-inspired neural network architecture adapted for tabular data, as advocated by Gorishniy et al.[9] This architecture consists of multiple so-called residual blocks, or 'ResBlocks', sequentially connected to one another. For `PMHnetV2`, these ResBlocks were defined as

$$\text{ResBlock}_h(z) = (\text{BN} \circ \text{FC}_{h,h} \circ \text{DO} \circ \text{SiLU} \circ \text{FC}_{h,h} \circ \text{DO})(z) + z \quad (7.12)$$

where BN is a batch normalization function, $\text{FC}_{h,h}$ is a fully-connected linear function with h inputs and h outputs, DO is a dropout function for regularization, and SiLU is the sigmoid-weighted linear unit (SiLU)—a non-linear activation function.[10] An important aspect of the ResBlock is the skip-connection, $f(z) + z$, where a learned representation is added on top of the untransformed input, which can enable training of very deep neural networks.[11] For the models tested, we constrained all ResBlocks to have the same number of hidden units h in each of the hidden layers.

By stacking together several of these building blocks, we can adjust the depth and complexity of the final architecture

$$\text{ResNet}(z) = (\text{FC}_{m,h} \circ \text{ResBlock}_h \circ \cdots \circ \text{ResBlock}_h \circ \text{FC}_{h,o})(z) \quad (7.13)$$

which depends on the number of input features m, the number of hidden units m, and the number of output logits o.[12]

7.4 *Model Training and Hyperparameter Tuning*

For training neural network models, we utilized the 'super-convergence' training protocol described by Smith and Topin,[13] a general methodology for fast and efficient training of neural network models. In this approach,

[9] Yury Gorishniy et al. 'Revisiting Deep Learning Models for Tabular Data'. *arXiv* (2023). URL: `arxiv.org/abs/2106.11959`. preprint.

[10] Stefan Elfwing et al. 'Sigmoid-Weighted Linear Units for Neural Network Function Approximation in Reinforcement Learning'. *arXiv* (2017). URL: `arxiv.org/abs/1702.03118`. preprint.

[11] A. Emin Orhan and Xaq Pitkow. 'Skip Connections Eliminate Singularities'. *arXiv* (2018). URL: `arxiv.org/abs/1701.09175`. preprint.

[12] Which for our discrete-time competing risk setup is $q \cdot (\kappa + 1)$, where q is the number of time bins and κ is the number of competing risks.

[13] Leslie N. Smith and Nicholay Topin. 'Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates'. *arXiv* (2018). URL: `arxiv.org/abs/1708.07120`. preprint.

models are trained for a pre-specified number of steps using the 'AdamW' stochastic optimization algorithm,[14] and the learning-rate is continuously adjusted during training following a one-cycle learning rate policy. We used the `OneCycleLR` implementation from the `PyTorch` library.[15]

Given the multitude of settings that needs to be specfied for configuration of both the neural network models and the training process itself, we conducted several hyperparameter optimization (HPO) experiments to explore various combinations of hyperparameters. For this purpose, we further subdivided the training data into a training and validation split. This validation split was used to assess model performance of the models constructed during the hyperparameters sweeps. The following gives an overview of the hyperparameters included in the HPO, and the range of possible values explored.

We included five parameters to adjust the architecture of the networks and the complexity of the discretization grid:

a) `n_timebins`: number of time bins in the discretization grid. Allowed values are 1 to 100.
b) `n_hidden`: number of hidden units in each hidden layer. Controls the width of the neural network. Allowed values are 10 to 100.
c) `n_blocks`: number of residual blocks. Controls the depth of the neural network. Allowed values are 1 to 20.
d) `enable_skipconn`: should the skip-connection part of the ResBlock (eq. (7.12)) be included? Toggle between true/false.
e) `enable_batchnorm`: should the batch-normalization layer of the ResBlock (eq. (7.12)) be included? Toggle between true/false.

Four parameters were used to configure the model training process and to regulate the amount and type of regularization.

f) `n_step`: number of training epochs. Range: 5 to 25.
g) `max_lr`: maximum learning rate in the one-cycle scheduler. Range: 1×10^{-3} to 1.
h) `weight_decay`: amount of weight decay used in the `AdamW` optimizer. Range: 1×10^{-5} to 1.
i) `dropout`: dropout rate. Range: 0 % to 90 %.

Finally, four parameters controlled the upper limit on the inclusion of retrospective data for various features categories, namely:

j) `cutoff_bioc`: inclusion window for laboratory tests results. Range: 0.5 to 5, stepsize: 0.5 .
k) `cutoff_diag`: inclusion window for diagnosis codes. Range: 0.5 to 10, stepsize: 0.5 .
l) `cutoff_proc`: inclusion window for procedure and examination codes. Range: 0.5 to 10, stepsize: 0.5 .
m) `cutoff_medi`: inclusion window for drug prescription features. Range: 0.5 to 10, stepsize: 0.5 .

The HPO process was automatized using the `Optuna` framework.[16] As the tuning parameter, we used the index of prediction accuracy (IPA) of the primary outcome (ACMO, CVMO, CVCO, and MIEV). Since the IPA is a time-dependent measure, we calculated the IPA at 50 evenly spaced timepoints from 0.5 to 5 years and numerically integrated the values to provide a single performance measure for the HPO.

After HPO, we used the best performing configuration for each outcome and trained the final models on the entire training data.

7.4.1 *Model Evaluation*

For evaluation of the `PMHnetV2` time-to-event prediction models, we assessed model discrimination and calibration. Similar to Study II, we used time-dependent area under the receiver receiver operating characteristic (td-AUC) to quantify the models' cause-specific ability to discriminate

[14] Ilya Loshchilov and Frank Hutter. 'Decoupled Weight Decay Regularization'. *arXiv* (2019). URL: arxiv.org/abs/1711.05101. preprint.

[15] Adam Paszke et al. 'PyTorch: An Imperative Style, High-Performance Deep Learning Library'. *arXiv* (2019). URL: arxiv.org/abs/1912.01703. preprint

[16] Takuya Akiba et al. 'Optuna: A Next-generation Hyperparameter Optimization Framework'. KDD '19: The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2019.

between cases and non-cases. For quantification of model calibration, we computed the Brier score and the IPA.

The IPA, which is an R^2-type measure of model accuracy, is obtained by scaling the Brier score of the model with that of a covariate-less null model based on the estimated incidences from the Kaplan-Meier or Aalen-Johansen estimator.[17] It is sometimes referred to as the 'scaled Brier score'. One advantage of the IPA metric is its easy interpretation: a perfect model has a score of 100 %, models with a positive IPA are potentially useful, and models with a negative IPA are useless or harmful.

For the three outcomes including competing risks—CVMO, CVCO, and MIEV—we also included a version of the model where competing events were treated as censored. To analyse the difference between the models with and without competing risks, we compared the differences in Brier score and td-AUC. Standard errors for the ΔBrier scores are obtained using an approach similar to that of one-sample t-tests, as described in Gerds and Kattan.[18] Correspondingly, standard errors for the ΔAUC were obtained using the Delong-Delong method,[19] also following Gerds and Kattan.[20]

For these model evaluation metrics and tests, we use the implementations provided by the R-package `riskRegression`.[21] All performance metrics were exclusively calculated using the test set, or in the case of HPO, a validation split of the training data.

7.5 *Main Findings*

We developed neural network models for time and cause-specific probabilistic predictions of experiencing each of the four primary endpoints ACMO, CVMO, CVCO, and MIEV. From the HPO experiments performed on each of these models, we found that several hyperparameter configurations were associated with a negative validation-set IPA and therefore resulted in decidedly inaccurate time-to-event models. On the other end of the spectrum, the best performing configurations for each of the four outcomes were all found to have useful IPA scores, with 22.2 % for ACMO, 12.6 % for CVMO, 23.5 % for CVCO, and 8.00 % for MIEV, highlighting the importance of the HPO process.

From analysis of the HPO sweeps, we found that `enable_batchnorm` and `enable_skipconn` considerably affected model accuracy. In every case tested, we concluded that both parameters should be set to `true` for optimal performance. For trials without skip-connections and batch-normalization, the model architecture resembles an multilayer perceptron (MLP) rather than a ResNet-like one, which negatively affected model performance. Gorishniy et al.[22] found that a ResNet-like architecture are well-suited for tabular data neural networks, consistently outperforming MLP-based models. Our experiments support this and extend the finding to Logistic-Hazard time-to-event models, both with and without competing risks.

The best-performing configurations, as determined by the HPO process, were then used in the setup and training of the final `PMHnetV2` models. These models were then evaluated on the previously unseen

[17] Michael W. Kattan and Thomas A. Gerds. 'The Index of Prediction Accuracy: An Intuitive Measure Useful for Evaluating Risk Prediction Models'. *Diagnostic and Prognostic Research* (2018).

[18] Thomas A. Gerds and Michael W. Kattan. *Medical Risk Prediction Models: With Ties to Machine Learning*. Chapman and Hall/CRC, 2021.

[19] E. R. DeLong et al. 'Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach'. *Biometrics* (1988).

[20] Gerds and Kattan, see n. 18.

[21] Thomas Alexander Gerds et al. *riskRegression: Risk Regression Models and Prediction Scores for Survival Analysis with Competing Risks*. Version 2023.12.21. 2023. URL: `cran.r-project.org/web/packages/riskRegression/index.html` (visited on 29/12/2023).

[22] Gorishniy et al., see n. 9.

hold-out test data. Figures 7.5 to 7.7 here summarise the main evaluation results, but additional details are included in the full-length manuscript in appendix C.

Importantly, models based on our competing risk Logistic-Hazard framework were all found to be well-calibrated, as exemplified by the calibration plots in fig. 7.5. These plots compare the 1- and 5-year predictions with the observed incidences obtained from the Kaplan-Meier and Aalen-Johansen estimates.

From visual inspection of the distribution of model predictions conditional on the observed outcomes (fig. 7.6), we found the predicted cause-specific probabilities to be consistently higher for patients that experience the primary outcome ('primary') compared to those that remained event-free ('event-free'). This effect, however, was less pronounced for the MIEV outcome. Interestingly, for the three outcomes with competing risks; CVMO, CVCO, and MIEV; the distribution of the primary-cause predictions between patients experiencing the primary outcome ('primary') and those experiencing the competing event ('competing'), were overlapping considerably. This suggests that it is difficult to distinguish between the competing risks, perhaps due to many shared risk factors, indicating that treating competing events as censoring would invalidate the assumption of uninformative censoring.

To further quantify the model discrimination and calibration, we calculacted the td-AUC and IPA across 100 evenly separated prediction horizons, ranging from 31 days to 5 years (fig. 7.7). From this, we observed good model performance across all tested prediction horizons. For the three outcomes including competing risks, we also included a naïve reference model where competing events were treated as censored. We consistently found the td-AUC and IPA of the naïve models to be worse than the competing-risk version.

To test the performance gain of jointly modelling the competing risks, we analysed the differences in td-AUC and Brier scores between the competing-risk and the naïve versions across the CVMO, CVCO, and MIEV outcomes. The results of these comparisons are summarised in table 7.1, which shows that in all cases, the competing risk models are better or comparable to models disregarding competing events. Specifically, model calibration were found to be the most impacted performance metric.

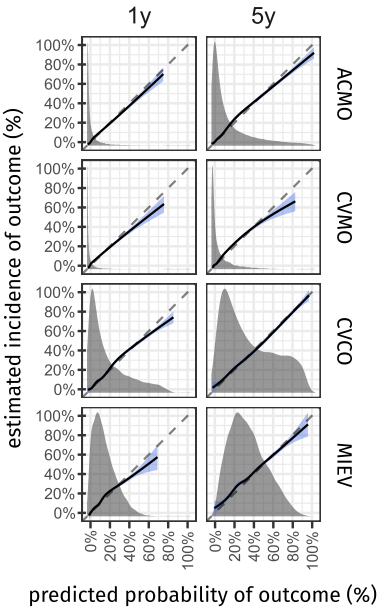


Figure 7.5: Calibration plots comparing the cause-specific t-year `PMHnetV2` model predictions with the estimated incidences at a 1 and 5 year prediction horizon. The included regression line is obtained by fitting a natural cubic spline ($d.o.f. = 6$) with pseudo-values obtained by jackknife resampling of the Kaplan-Meier (ACMO) or the Aalen-Johansen (CVMO, CVCO, and MIEV) estimates. The dashed 45-degree reference lines represent perfect calibration. The density curve in the background of each panel shows the distribution of the model estimates.

	ΔAUC			ΔBrier		
	$p < 0.05$		$p \geq 0.05$	$p < 0.05$		$p \geq 0.05$
	$\Delta > 0$	$\Delta < 0$	$\Delta \approx 0$	$\Delta > 0$	$\Delta < 0$	$\Delta \approx 0$
CVMO vs. naïve	56	0	44	97	0	3
CVCO vs. naïve	91	0	9	97	0	3
MIEV vs. naïve	87	0	13	99	0	1

Table 7.1: Differences in Brier score and td-AUC between the `PMHnetV2` competing risk models and naïve reference models were competing events were treated as censoring. The table shows the number of timepoints for which the competing risk model had significantly better ($\Delta > 0$), significantly worse ($\Delta < 0$), or comparable performance ($\Delta \approx 0$) to the naïve model.

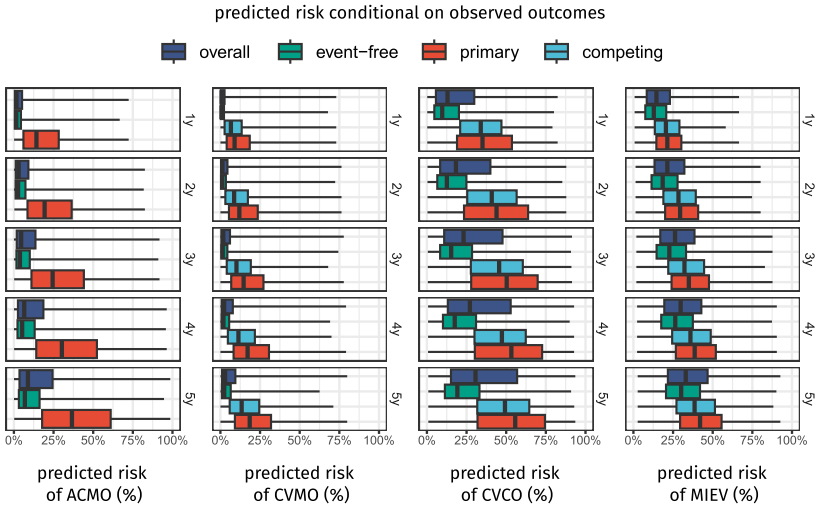


Figure 7.6: Visual assessment of test set discrimination. Boxplots display the predicted t-year risk quantiles for the primary endpoints, conditional on the observed t-year outcomes observed. The boxplot marked 'overall' show the observed quantiles of the entire test set. The quantiles included in the boxplots marked 'primary', 'competing', and 'event-free' have been estimated using inverse probability of censoring weighting.

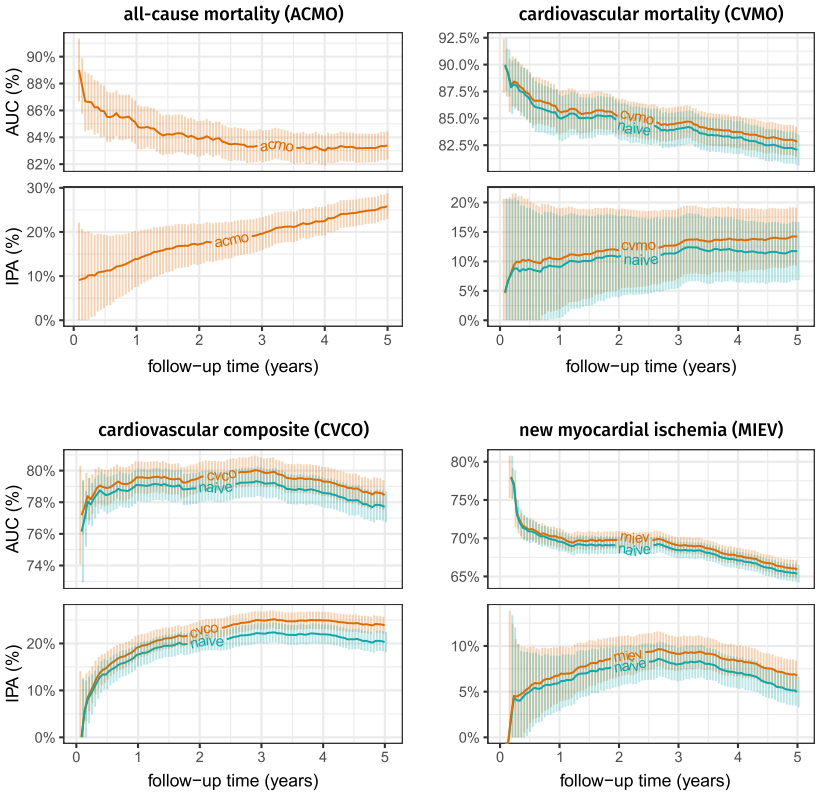


Figure 7.7: Time-dependent Area under the receiver receiver operating characteristic (td-AUC) and index of prediction accuracy (IPA) for prediction of the primary endpoint for each of the four `PMHnetV2` models. td-AUC and IPA are computed at 100 evenly spaced timepoints, ranging from 31 days to 5 years. For the three outcomes including competing risks—CVMO, CVCO, and MIEV—we also included a naïve single-risk model where competing events were treated as censored.

7.6 *Conclusions*

In this study, we presented the development of `PMHnetV2`, an ensemble of four distinct neural network models for time-to-event prediction of important clinical outcomes in patients with IHD. These models were developed to predict the onset of important clinical endpoints following coronary angiography: all-cause mortality (ACMO), cardiovascular mortality (CVMO), cardiovascular complications (CVCO), and new myocardial ischemia events (MIEV).

To enable the construction of these models, a central contribution of our research, is the development of a novel approach for neural network-based modelling of discrete time-to-event data with competing risks. This approach can be viewed as an extension to Logistic-Hazard model from Gensheimer and Narasimhan,[23] and is grounded in classical statistical analysis of discrete time-to-event data.[24] To the best of our knowledge, we are the first to utilize this methodology for neural network-based competing risk analysis.

Our approach has been implemented in the python package `DiscoTime`, a general-purpose software library that facilitates the application and further exploration of this methodology for time-to-event prediction with neural networks.[25]

In our experiments, `PMHnetV2` models, utilizing this novel approach, provided accurate and well-calibrated time- and cause-specific estimates for secondary prognostication in patients with IHD. Compared to the common practice of treating competing events as censored, our methodology demonstrated an increase in both model discrimination and, particularly, in model calibration. In the context of clinical research where competing risks are common, this finding underscores the importance of using methodologies, like ours, that properly handle competing risks. Our study suggests such approaches should be considered in the development of future neural network-based medical risk prediction models.

[23] Gensheimer and Narasimhan, see n. 1.

[24] Tutz and Schmid, see n. 2.

[25] Holm, see n. 8.

PART III

Concluding Remarks

Chapter 8

Principal Findings, Limitations, and Future Perspectives

In this thesis, I have explored the use of informatics-based approaches for addressing critical aspects pertinent to the understanding and management of ischemic heart disease (IHD). Central to this research was the application of advanced machine learning (ML) methods on large-scale electronic health data for development of precision medicine approaches for secondary prevention in IHD. This involved identifying and characterizing patterns of multimorbidity in IHD and developing feature-rich clinical prediction models for precision prognostication.

In this chapter, I will briefly reiterate the main findings of the included studies, adress some general and study-specific limitations of the work undertaken, and discuss perspectives for future research.

8.1 *Principal Findings*

Throughout the previous chapters, I have described three different scientific manuscripts detailing our research in the framework of ML-based precision medicine. The following provides a brief summary of the principal findings of each of the included studies.

8.1.1 *Comorbidity Clustering in Ischemic Heart Disease*

In Study I, we used unsupervised clustering analysis to explore the comorbidity landscape of 72 249 patients with IHD. We used the broadest possible definition of multimorbidity and defined comorbidity as the historical co-occurence of a broad array of diagnosis codes in the individual patient records. The accrued patient-specific comorbidity profiles, containing more than 3000 different diagnosis codes, led to the identification of 31 distinct patient subgroups. These clusters represent distinct patterns of multimorbidity linked to IHD, were found to be associated with specific risk of subsequent outcomes, and can be used to better understand the complex nature of multimorbidity in IHD.

8.1.2 *Time-to-Event Prediction of All-Cause Mortality*

In Study II, we presented the development and validation of a novel neural network-based prognostication model for prediction of all-cause mortality in patients with IHD. This model, `PMHnetV1`, utilises a discrete-time approach for modelling of time-to-event data with neural networks and can provide time-specific probability estimates of survival across a

five-year prediction horizon. The model was developed using a large and diverse dataset 39 746 IHD patients from the Eastern Denmark Heart Registry (EDHR) and incorporates a comprehensive set of 584 features, including diagnosis history, procedural codes, laboratory test results, and clinical measurements obtained from electronic health record (EHR) data and registry data. Compared to both the GRACE 2.0 score, and a neural network-based model limited to the GRACE features, the feature-rich `PMHnetV1` model provided a significant improvement in model performance. External validation on an independent Icelandic dataset of 8287 patients further showed that the model performance is generalizable. Furthermore, by including SHAP-analysis we were able to provide explanations of the model output and assess feature importance. The study established `PMHnetV1` as a valuable tool for post-angiography assessment of all-cause mortality risk in IHD patients, and can potentially aid clinicians in making informed decisions about treatment and management of IHD.

8.1.3 *Time-to-Event Prediction with Competing Risks*

In Study III, we introduced an new framework for construction of neural network-based competing risk models and presented the development `PMHnetV2`, an advanced iteration of our IHD prognostication algorithm. The updated `PMHnetV2` model provide cause- and time-specific risk estimates for all-cause mortality, cardiovascular mortality, cardiovascular complications, and new myorcardial ischemia events. From internal validation, we found the model estimates to be well-calibrated and to accurately predict patient at both high and low risk of the four different outcomes. Compared to the standard practice of treating competing events as censored, we showed that models capable of jointly modelling competing risks were associcated with a better model discrimination and calibration. While still a work in progress, the presented work establishes the usefullness of the updated methodology and presents `PMHnetV2` as a promising tool for prognostication in IHD.

8.2 *Limitations*

Despite their strengths, a number of limitations and constraints related to the presented studies, potentially affects the overall interpretation of the findings. In the following, I will address and discuss both study-specific and general limitations of our research.

8.2.1 *Definition of Comorbidities*

In Study I, a possible limitation relates to its exclusively data-driven definition of multimorbidity that included the historical co-occurence of a very broad array of diagnosis codes. This approach constrasts with that of similar studies in the domain. As an example, Forman et al.[1] defined multimorbidity as 'two or more medical diseases or conditions, each lasting more than one year'. Similarly, another study also limited their definition to only cover chronic conditions, specifically the 20 most

[1] Daniel E. Forman et al. 'Multimorbidity in Older Adults With Cardiovascular Disease'. *Journal of the American College of Cardiology* (2018).

common ones.[2] Unlike these studies that focused on chronic conditions, our study did not differentiate between chronic and acute diagnoses. As a result, our clustering could, for example, be influenced by a 3-year old pneumonia diagnosis. However, since we accrued the number of admissions for each diagnosis the chronic nature of certain conditions is likely implicitly accounted.

[2] Walter A. Rocca et al. 'Prevalence of Multimorbidity in a Geographically Defined American Population: Patterns by Age, Sex, and Race/Ethnicity'. *Mayo Clinic Proceedings* (2014).

8.2.2 *Lack of Temporal Resolution in Features*

In this thesis, a notable limitation is the absence of temporal resolution in the input features, affecting both the clustering in Study I and the prediction models in Study II and Study III. This lack of temporal granularity means that the models and analyses do not account for the timing and sequence of medical events or diagnoses.

In Study I, the clustering could have been enhanced by somehow incorporating the chronological order of the diagnoses in the patient vectors. This would allow for a more nuanced description of the co-morbidity burden of the individual patient, and could in addition help alleviate the limitation of chronic versus acute conditions described in section 8.2.1. Previous research within our group by Jensen et al.[3] illustrated a method to identify temporal disease trajectories from retrospective registry data. They also demonstrated the use of these trajectories in clustering applications. However, this method only captures temporal patterns with clear directionality, which could exclude many of the comorbidities we considered in our study. Thus, while it offers a possible avenue for future research, it also has its limitations in fully representing the range of comorbidities.

[3] Anders Boeck Jensen et al. 'Temporal Disease Trajectories Condensed from Population-Wide Registry Data Covering 6.2 Million Patients'. *Nature Communications* (2014).

In Study II and Study III, time resolved input features could enable the neural network models to learn from sequential patterns of medical events and diagnoses. Such information could provide valuable information for accurate prognostication. For instance, knowing the progression of IHD and comorbidities could potentially inform more timely and tailored interventions. Additionally, the study design used in the development of `PMHnetV1` and `PMHnetV2` was limited in scope to only provide predictions subsequent to an index coronary angiography. While these models might be applicable at other timepoints, it is not something that we have tested, and it would probably affect their performance.

Alternatives to address this limitation include the use of time-series data and longitudinal study designs such as those based on landmark analysis.[4] These approaches can facilitate the creation of dynamic risk prediction models.[5] In the context of neural networks, this would likely involve using architectures like long short-term memory (LSTM)[6] or Transformers,[7] which are designed to use sequential features.

[4] Urania Dafni. 'Landmark Analysis at the 25-Year Landmark Point'. *Circulation: Cardiovascular Quality and Outcomes* (2011).

[5] Hans C. Van Houwelingen. 'Dynamic Prediction by Landmarking in Event History Analysis'. *Scandinavian Journal of Statistics* (2007).

[6] Sepp Hochreiter and Jürgen Schmidhuber. 'Long Short-Term Memory'. *Neural Computation* (1997).

[7] Ashish Vaswani et al. 'Attention Is All You Need'. *Advances in Neural Information Processing Systems*. 2017.

8.2.3 *Generalisability of Clusterings*

For Study I, an inherent limitation of clustering applications is the lack of standardized techniques for external 'validation' compared to those in supervised learning. In supervised learning, evaluating the model generalizability is straightforward: apply the model to a test set, which

could be an internal hold-out set or an external dataset, and then directly measure its performance. However, this approach is not feasible in most unsupervised clustering applications due to the absence of predefined labels. Alternative strategies do exists, as outlined by Ullmann et al.,[8] and includes:

- Applying the clustering algorithm to a representative external dataset. Subsequently, examine if the cluster structure obtained on this external dataset shares internal and external characteristics with the original clustering.

- Transferring the original clustering to the external dataset by first using, for example, a supervised classifier.[9] This classifier is trained to predict the cluster labels derived from the original dataset and then applied to the external dataset. If clustering on the external data is consistent with the transferred labels, then it indicates that the cluster algorithm have captured patterns that are not just specific to the initial dataset.

Such approaches, while not direct validations in the traditional sense, could provide insights into the overall generalisability of the clustering outside the context of the original dataset.[10] Nonetheless, these approaches have not been implemented in our research. Consequently, we do not assert that the clustering presented is definitively the 'best' but rather utilize it as a method to condense the extensive array of diagnostic codes into interpretable subgroups.

[8] Theresa Ullmann et al. 'Validation of Cluster Analysis Results on Validation Data: A Systematic Framework'. *WIREs Data Mining and Knowledge Discovery* (2022).

[9] Ullmann et al., see n. 8.

[10] Ullmann et al., see n. 8.

8.2.4 *Choice of Clustering Algorithm*

A further potential limitation of Study I is that we did not compare or test other clustering methodologies besides the Markov clustering (MCL) algorithm. Although numerous different clustering algorithms exists, our choice of using MCL was motivated by a number of key aspects. Firstly, the MCL algorithm is fast and has been explicitly designed to handle very large networks with a substantial number of vertices and edges.[11] Secondly, in this algorithm, the number of clusters neither can nor should be pre-specified. Instead the issue of 'how many clusters' is handled by a strong internal logic, rather than being dealt with in an arbitrary manner as is common in other clustering algorithms.[12]

[11] Stijn Van Dongen. 'Graph Clustering Via a Discrete Uncoupling Process'. *SIAM Journal on Matrix Analysis and Applications* (2008).

[12] Van Dongen, see n. 11.

8.2.5 *Lack of Primary Care Data*

A fundamental limitation in our research stems from the nature of the data accessed, as all our studies primarily utilized hospital data. This reliance on hospital data is likely to lead to an underrepresentation of data related to conditions and diseases primarily managed in primary care settings, including hypertension, non-complex infections, and various soft-tissue disorders.[13]

In Study III, we attempted to mitigate this limitation by incorporating prescription data, which can serve as proxy for the conditions managed in primary care. However, it is important to note that prescription data

[13] Caitlin R. Finley et al. 'What Are the Most Common Conditions in Primary Care?' *Canadian Family Physician* (2018).

is only partly able to compensate for the lack of detailed primary care patient records.

8.2.6 *Limitations of Explainable AI*

The last limitation I would like to higlight are some general shortcomings of explainable artificial intelligence (XAI) that often are overlooked. Currently, XAI is only implemented in Study II, but our plan is include it in Study III as well, and for this future work, these limitations also apply.

As described earlier in this thesis, the goal of ML is to make accurate predictions on unseen data, and as consequence, the 'how' and 'why' of predictions is of less concern. However, for critical applications, including healthcare, it is generally agreed that transparency is important and that the 'black box' nature of ML needs to be adressed.[14] This is exemplifed by article 15, of the European Union's General Data Protection Regulation (GDPR),[15] which specifies an requirement for transparency that applies to algorithmic decision-making.[16]

For ML, including neural networks, XAI is a form of post-hoc analysis that seeks to provide the required transparency for otherwise complex and non-transparent models. In this domain, SHAP-analysis,[17] which we utilized in Study II for providing explanations of the `PMHnetV1` model, is arguably one of the most popular XAI approaches. SHAP, along with other similar approaches, relies on the usage of simpler surrogate model to estimate the expected marginal contribution of each feature to the model's output.[18] However, this approach requires certain assumptions, including the premise that the model can be locally approximated by a simpler model and that features are independent.[19] The independence assumptions is very strong and often very unrealistic, which is likely to bias the estimates of feature contributions—nevertheless, this approach is still in widespread use. Recent research has demonstrated that is is possible to partially mitigate this limitation, but at the expense of a significantly increased computational complexity, which can limit its practical usefulness.[20]

XAI algorithms such as SHAP are approximations of the complete model, therefore the fidelity is not perfect and as a consequence, neither are the explanations. Currently, there are no established standards for assessing the quality of these explanations. While it is possible the esimate the error of the approximations, this does not necessarily indicate whether the explanations are interpretable and understandable to end-users.

From a practical standpoint, the explanations offered by XAI can be subject to misinterpretation, particularly by users less familiar with technical details of the AI-model and the XAI method used. During the clinical implementation of the `PMHnetV1` model for a, currently ongoing, clinical trial,[21] we provide SHAP values alongside the model predictions to inform clinicians on the basis of the model predictions. However, a pilot experiment in which clinicians were asked to qualitatively evaluate the model's output and explanations revealed some challenges in the interpretation of these.

For example, one clinician found it counterintuitive that the model

[14] Eric J. Topol. 'High-Performance Medicine: The Convergence of Human and Artificial Intelligence'. *Nature Medicine* (2019).

[15] European Parliament and Council of the European Union. *Regulation (EU) 2016/679 of the European Parliament and of the Council. of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation)*. 2016. URL: `data.europa.eu/eli/reg/2016/679/oj` (visited on 13/04/2023).

[16] Item (h) in paragraph 1 of the GDPR article 15 states that 'the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.'

[17] Scott M Lundberg and Su-In Lee. 'A Unified Approach to Interpreting Model Predictions'. *Advances in Neural Information Processing Systems*. 2017.

[18] Vaishak Belle and Ioannis Papantonis. 'Principles and Practice of Explainable Machine Learning'. *Frontiers in Big Data* (2021).

[19] Lundberg and Lee, see n. 17.

[20] Kjersti Aas et al. 'Explaining Individual Predictions When Features Are Dependent: More Accurate Approximations to Shapley Values'. *Artificial Intelligence* (2021).

[21] Henning Bundgaard. *Clinical Implementation of a Novel Decision Support Tool in Patients With Ischemic Heart Disease*. Clinical trial registration. clinicaltrials.gov, 2023. URL: `clinicaltrials.gov/study/NCT06033014` (visited on 01/01/2023).

identified hyperlipidemia (ICD-10: E78) as a factor contributing to increased survival. While it is possible that this finding is caused by the aforementioned limitations, there are other plausible explanations as to why hyperlipidemia could be identified as a 'protective' feature. It is important to note that SHAP values are correlations and do not imply causation. Patients already known with hyperlipidemia prior to their index coronary angiography may represent a group of patients with non-acute manifestations of IHD, which relative to the median IHD patient could have improved survival. Additionally, these patients are likely to have initiated statin treatment before the time of prediction, which once again, could be associated with a improved prognosis relative to the median patient.

This example underscores the complexities inherent in interpreting XAI explanations, especially when they appear counterintuitive or misaligned with conventional medical knowledge. It also emphasizes the need for thorough education and effective communication with healthcare professionals regarding clinical decision support tools that incorporate these technologies.

8.3 *Future Perspectives*

Adressing the various just discussed limitations and constraints all represent topics for future research, some more important than others. To conclude this thesis, I would like to highlight two central challenges related to this thesis project of utmost importance for future data-driven research in precision medicine.

8.3.1 *Clinical Implementation of Machine Learning Models*

In this thesis, I have argued for the potential benefits of AI/ML methodologies in improving the clinical treatment and management of patients with IHD. We have developed two neural network-based IHD prognostication models, evaluated their performance using state-of-the-art statistical metrics, and concluded that high-dimensional ML-based models are superior to existing alternatives. However, the theoretical clinical impact is, as of now, only just that—theoretical.

It is generally accepted that the overwhelming majority of published medical prediction models are never implemented in clinical practice.[22] To increase the adoption of ML-based prognostic models, we need more prospective clinical studies to ascertain the impact and practical applicability of such models. This includes exploring how AI/ML can positively affect clinical decision-making, patient outcomes, and allocation of healthcare resources. Such research is crucial in realising the potential of AI/ML in the advancement of precision medicine.

As an extension of the work presented in Study II, although not included as a central part of this thesis, we have established a collaboration with the public company CIMT to implement `PMHnetV1` in 'Sundhedsplatformen', the EHR system used in the Capital Region and Region Zealand hospitals, for prospective clinical validation. We successfully integrated the model into a real-world clinical EHR setting and have initi-

[22] Ewout W. Steyerberg et al. 'Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research'. *PLOS Medicine* (2013).

ated a clinical trial.[23] This study is still in progress, but independently of its outcomes, the mere implementation of the model in a clinical setting is a significant achievement.

Looking ahead, depending on the findings of this trial, the focus needs to shift towards the implementation of processes for continuous monitoring of model performance, for regularly updating the model with new data, for refining models to include additional endpoints (`PMHnetV2`), and several other important aspects.

[23] Bundgaard, see n. 21.

8.3.2 *Sharing of Healthcare Data*

As outlined in chapter 4, the studies presented in this thesis draws on hospital data from more than 2.6 million individuals, which originates from combination of different data sources, including electronic health records, national registries, and clinical quality databases. In the context of this thesis project, a major challenge have been processing, combining, cleaning, and organizing these diverse sources of data into curated datasets appropriate for ML applications.

However, because such data, for good reason, is subject to ethical and privacy-protecting rules and regulations, it cannot realistically be shared with researchers outside our institution. This means that it is impossible for others to reproduce our findings, develop and benchmark rival models, and benefit from the data cleaning and curation that have already been done. It is a waste of resources and limits the development of the field as a whole.

It is evident that there exists an unmet need for regulations and approaches that enable combining and benefitting from otherwise siloed datasets. In this context, the concept of federated health data networks have been suggested as a possible solution to overcome existing barriers preventing sharing of data.[24] In the ongoing effort of establishing a European Health Data Space, the European Commission's proposal have also included procedures and regulations for secondary research use of health data. As suggested by Raab et al.,[25] this effort could be coupled with the establishment of pan-European federated health data network, which could break the many barriers limiting current big data-based clinical research.

Future research should focus on technical solutions for establishing such networks, development of algorithms for distributed machine learning, and construction of interoperability formats which could further cross-institutional and international collaboration.

[24] Harry Hallock et al. 'Federated Networks for Distributed Analysis of Health Data'. *Frontiers in Public Health* (2021).

[25] René Raab et al. 'Federated Electronic Health Records for the European Health Data Space'. *The Lancet Digital Health* (2023).

Bibliography

Aalen, Odd O. and Søren Johansen. 'An Empirical Transition Matrix for Non-Homogeneous Markov Chains Based on Censored Observations'. *Scandinavian Journal of Statistics* (1978).

Aas, Kjersti, Martin Jullum and Anders Løland. 'Explaining Individual Predictions When Features Are Dependent: More Accurate Approximations to Shapley Values'. *Artificial Intelligence* (2021).

Akiba, Takuya et al. 'Optuna: A Next-generation Hyperparameter Optimization Framework'. KDD '19: The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2019.

Anderson, Todd J. et al. '2012 Update of the Canadian Cardiovascular Society Guidelines for the Diagnosis and Treatment of Dyslipidemia for the Prevention of Cardiovascular Disease in the Adult'. *Canadian Journal of Cardiology* (2013).

Arendt, Johan Frederik Håkonsen et al. 'Existing Data Sources in Clinical Epidemiology: Laboratory Information System Databases in Denmark'. *Clinical Epidemiology* (2020).

Belle, Vaishak and Ioannis Papantonis. 'Principles and Practice of Explainable Machine Learning'. *Frontiers in Big Data* (2021).

Bergstra, James and Yoshua Bengio. 'Random Search for Hyper-Parameter Optimization'. *Journal of Machine Learning Research* (2012).

Bishop, Christopher M. *Neural Networks for Pattern Recognition*. Clarendon Press, 1995.

Blanche, Paul, Michael W. Kattan and Thomas A. Gerds. 'The C-Index Is Not Proper for the Evaluation of t-Year Predicted Risks'. *Biostatistics (Oxford, England)* (2019).

Brown, S.F., A.J. Branford and W. Moran. 'On the Use of Artificial Neural Networks for the Analysis of Survival Data'. *IEEE Transactions on Neural Networks* (1997).

Bundgaard, Henning. *Clinical Implementation of a Novel Decision Support Tool in Patients With Ischemic Heart Disease*. Clinical trial registration. clinicaltrials.gov, 2023. URL: clinicaltrials.gov/study/NCT06033014 (visited on 01/01/2023).

Byrne, Robert A et al. '2023 ESC Guidelines for the Management of Acute Coronary Syndromes'. *European Heart Journal* (2023).

Charniak, Eugene. *Introduction to Deep Learning*. Illustrated. The MIT Press, 2019.

Chollet, Francois. *Deep Learning with Python*. 2nd ed. Manning, 2021.

Clark, Alexander M., Lisa Hartling, Ben Vandermeer and Finlay A. McAlister. 'Meta-Analysis: Secondary Prevention Programs for

Patients with Coronary Artery Disease'. *Annals of Internal Medicine* (2005).
Collet, Jean-Philippe et al. '2020 ESC Guidelines for the Management of Acute Coronary Syndromes in Patients Presenting without Persistent ST-segment Elevation'. *European Heart Journal* (2021).
Collins, Francis S. and Harold Varmus. 'A New Initiative on Precision Medicine'. *New England Journal of Medicine* (2015).
Cox, D. R. 'Regression Models and Life-Tables'. *Journal of the Royal Statistical Society. Series B (Methodological)* (1972).
Dafni, Urania. 'Landmark Analysis at the 25-Year Landmark Point'. *Circulation: Cardiovascular Quality and Outcomes* (2011).
Darby, Sarah C. et al. 'Risk of Ischemic Heart Disease in Women after Radiotherapy for Breast Cancer'. *New England Journal of Medicine* (2013).
DeLong, E. R., D. M. DeLong and D. L. Clarke-Pearson. 'Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach'. *Biometrics* (1988).
Deng, Xiaotie and Christos H. Papadimitriou. 'On the Complexity of Cooperative Solution Concepts'. *Mathematics of Operations Research* (1994).
Elfwing, Stefan, Eiji Uchibe and Kenji Doya. 'Sigmoid-Weighted Linear Units for Neural Network Function Approximation in Reinforcement Learning'. *arXiv* (2017). URL: arxiv.org/abs/1702.03118. preprint.
European Parliament and Council of the European Union. *Regulation (EU) 2016/679 of the European Parliament and of the Council. of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation)*. 2016. URL: data.europa.eu/eli/reg/2016/679/oj (visited on 13/04/2023).
Faraggi, David and Richard Simon. 'A Neural Network Model for Survival Data'. *Statistics in Medicine* (1995).
Finley, Caitlin R. et al. 'What Are the Most Common Conditions in Primary Care?' *Canadian Family Physician* (2018).
Forman, Daniel E. et al. 'Multimorbidity in Older Adults With Cardiovascular Disease'. *Journal of the American College of Cardiology* (2018).
Fox, Keith A. A., Marco Metra, João Morais and Dan Atar. 'The Myth of 'Stable' Coronary Artery Disease'. *Nature Reviews Cardiology* (2020).
Fox, Keith A. A. et al. 'Should Patients with Acute Coronary Disease Be Stratified for Management According to Their Risk? Derivation, External Validation and Outcomes Using the Updated GRACE Risk Score'. *BMJ Open* (2014).
Frank, Lone. 'Epidemiology. When an Entire Country Is a Cohort'. *Science (New York, N.Y.)* (2000).
Fuster, Valentin, Lina Badimon, Juan J. Badimon and James H. Chesebro. 'The Pathogenesis of Coronary Artery Disease and the Acute Coronary Syndromes'. *New England Journal of Medicine* (1992).
Gensheimer, Michael F. and Balasubramanian Narasimhan. 'A Scalable Discrete-Time Survival Model for Neural Networks'. *PeerJ* (2019).

Gerds, Thomas A. and Michael W. Kattan. *Medical Risk Prediction Models: With Ties to Machine Learning*. Chapman and Hall/CRC, 2021.

Gerds, Thomas Alexander et al. *riskRegression: Risk Regression Models and Prediction Scores for Survival Analysis with Competing Risks*. Version 2023.12.21. 2023. URL: cran.r-project.org/web/packages/riskRegression/index.html (visited on 29/12/2023).

Golub, G. H. and C. Reinsch. 'Singular Value Decomposition and Least Squares Solutions'. *Handbook for Automatic Computation: Volume II: Linear Algebra*. Springer, 1971.

Goodfellow, Ian, Yoshua Bengio and Aaron Courville. *Deep Learning*. MIT press, 2016.

Gorishniy, Yury, Ivan Rubachev, Valentin Khrulkov and Artem Babenko. 'Revisiting Deep Learning Models for Tabular Data'. *arXiv* (2023). URL: arxiv.org/abs/2106.11959. preprint.

GRACE Risk Score 2.0. URL: outcomes-umassmed.org/grace/acs_risk2/index.html (visited on 12/12/2023).

Hallock, Harry et al. 'Federated Networks for Distributed Analysis of Health Data'. *Frontiers in Public Health* (2021).

Haue, Amalie D. et al. 'Subgrouping Multimorbid Patients with Ischemic Heart Disease by Means of Unsupervised Clustering: A Cohort Study of 72,249 Patients Defined Comprehensively by Diagnoses Prior to Presentation'. *medRxiv* (2023). URL: medrxiv.org/content/10.1101/2023.03.31.23288006v2. preprint.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren and Jian Sun. 'Deep Residual Learning for Image Recognition'. *arXiv* (2015). URL: arxiv.org/abs/1512.03385. preprint.

Helweg-Larsen, Karin. 'The Danish Register of Causes of Death'. *Scandinavian Journal of Public Health* (2011).

Hjaltelin, Jessica Xin et al. 'Pancreatic Cancer Symptom Trajectories from Danish Registry Data and Free Text in Electronic Health Records'. *eLife* (2023).

Hochreiter, Sepp and Jürgen Schmidhuber. 'Long Short-Term Memory'. *Neural Computation* (1997).

Holm, Peter. *Discotime: Discrete-time Competing Risk Analysis with Neural Networks*. Version 0.1.0. URL: github.com/peterchristofferholm/discotime (visited on 25/12/2023).

Holm, Peter C. et al. 'Development and Validation of a Neural Network-Based Survival Model for Mortality in Ischemic Heart Disease'. *medRxiv* (2023). URL: medrxiv.org/content/10.1101/2023.06.16.23291527v1. preprint.

Jensen, Anders Boeck et al. 'Temporal Disease Trajectories Condensed from Population-Wide Registry Data Covering 6.2 Million Patients'. *Nature Communications* (2014).

Jensen, Peter B., Lars J. Jensen and Søren Brunak. 'Mining Electronic Health Records: Towards Better Research Applications and Clinical Care'. *Nature Reviews Genetics* (2012).

Kaas-Hansen, Benjamin Skov et al. 'Language-Agnostic Pharmacovigilant Text Mining to Elicit Side Effects from Clinical Notes and Hos-

pital Medication Records'. *Basic & Clinical Pharmacology & Toxicology* (2022).

Kalbfleisch. *The Statistical Analysis of Failure Time Data, 2nd Edition*. 2nd edition. Wiley-Interscience, 2002.

Kaplan, Edward L and Paul Meier. 'Nonparametric Estimation from Incomplete Observations'. *Journal of the American statistical association* (1958).

Kattan, Michael W. and Thomas A. Gerds. 'The Index of Prediction Accuracy: An Intuitive Measure Useful for Evaluating Risk Prediction Models'. *Diagnostic and Prognostic Research* (2018).

Katzman, Jared L. et al. 'DeepSurv: Personalized Treatment Recommender System Using a Cox Proportional Hazards Deep Neural Network'. *BMC Medical Research Methodology* (2018).

Killip, Thomas and John T. Kimball. 'Treatment of Myocardial Infarction in a Coronary Care Unit: A Two Year Experience with 250 Patients'. *The American Journal of Cardiology* (1967).

Kirk, Isa Kristina et al. 'Linking Glycemic Dysregulation in Diabetes to Symptoms, Comorbidities, and Genetics through EHR Data Mining'. *eLife* (2019).

Klein, John P. and Melvin L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, 2003.

Kleinbaum, David G. and Mitchel Klein. *Survival Analysis: A Self-Learning Text, Third Edition*. 3rd edition. Springer, 2011.

Knuuti, Juhani et al. '2019 ESC Guidelines for the Diagnosis and Management of Chronic Coronary Syndromes'. *European Heart Journal* (2020).

Kumar, Vinay, Abul K. Abbas, Nelson Fausto and Jon C. Aster. *Robbins and Cotran Pathologic Basis of Disease*. Elsevier Health Sciences, 2014.

Kvamme, Håvard and Ørnulf Borgan. 'Continuous and Discrete-Time Survival Prediction with Neural Networks'. *Lifetime Data Analysis* (2021).

Landspatientregisteret, Dokumentation. eSundhed. URL: esundhed.dk/Dokumentation?rid=5 (visited on 25/11/2023).

LeCun, Yann et al. 'Handwritten Digit Recognition with a Back-Propagation Network'. *Advances in Neural Information Processing Systems*. 1989.

Lee, Changhee, William Zame, Jinsung Yoon and Mihaela van der Schaar. 'DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks'. *Proceedings of the AAAI Conference on Artificial Intelligence* (2018).

Libby, Peter and Pierre Theroux. 'Pathophysiology of Coronary Artery Disease'. *Circulation* (2005).

Loshchilov, Ilya and Frank Hutter. 'Decoupled Weight Decay Regularization'. *arXiv* (2019). URL: arxiv.org/abs/1711.05101. preprint.

Lundberg, Scott M and Su-In Lee. 'A Unified Approach to Interpreting Model Predictions'. *Advances in Neural Information Processing Systems*. 2017.

McDonald, Clement J et al. 'LOINC, a Universal Standard for Identifying Laboratory Observations: A 5-Year Update'. *Clinical Chemistry* (2003).

Mitchell, Tom M. *Machine Learning*. McGraw-Hill, 1997.

Murphy, Kevin P. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

Muse, Victorine P., Alejandro Aguayo-Orozco, Sedrah B. Balaganeshan and Søren Brunak. 'Population-Wide Analysis of Hospital Laboratory Tests to Assess Seasonal Variation and Temporal Reference Interval Modification'. *Patterns* (2023).

Nabel, Elizabeth G. and Eugene Braunwald. 'A Tale of Coronary Artery Disease and Myocardial Infarction'. *New England Journal of Medicine* (2012).

Neumann, Franz-Josef et al. '2018 ESC/EACTS Guidelines on Myocardial Revascularization'. *European Heart Journal* (2019).

Nielsen, Lisbeth. *LPR3 går snart i luften*. 2018. URL: sundhedsdatastyrelsen.dk.

OpenAI. 'GPT-4 Technical Report'. *arXiv* (2023). URL: arxiv.org/abs/2303.08774. preprint.

Orhan, A. Emin and Xaq Pitkow. 'Skip Connections Eliminate Singularities'. *arXiv* (2018). URL: arxiv.org/abs/1701.09175. preprint.

Özcan, Cengiz et al. 'The Danish Heart Registry'. *Clinical Epidemiology* (2016).

Paszke, Adam et al. 'PyTorch: An Imperative Style, High-Performance Deep Learning Library'. *arXiv* (2019). URL: arxiv.org/abs/1912.01703. preprint.

Pedersen, Mette Krogh et al. 'A Unidirectional Mapping of ICD-8 to ICD-10 Codes, for Harmonized Longitudinal Analysis of Diseases'. *European Journal of Epidemiology* (2023).

Pepe, Margaret S. and Motomi Mori. 'Kaplan—Meier, Marginal or Conditional Probability Curves in Summarizing Competing Risks Failure Time Data?' *Statistics in Medicine* (1993).

Prince, Simon J.D. *Understanding Deep Learning*. MIT Press, 2023.

Raab, René et al. 'Federated Electronic Health Records for the European Health Data Space'. *The Lancet Digital Health* (2023).

Ramesh, Aditya et al. 'Zero-Shot Text-to-Image Generation'. *arXiv* (2021). URL: arxiv.org/abs/2102.12092. preprint.

Regular Expression. *Wikipedia*. 2023. URL: en.wikipedia.org/w/index.php?title=Regular_expression&oldid=1186625751 (visited on 29/11/2023).

Rich, Michael W. et al. 'Knowledge Gaps in Cardiovascular Care of the Older Adult Population'. *Circulation* (2016).

Rocca, Walter A. et al. 'Prevalence of Multimorbidity in a Geographically Defined American Population: Patterns by Age, Sex, and Race/Ethnicity'. *Mayo Clinic Proceedings* (2014).

Rodríguez, Cristina Leal et al. 'Drug Dosage Modifications in 24 Million In-Patient Prescriptions Covering Eight Years: A Danish Population-Wide Study of Polypharmacy'. *PLOS Digital Health* (2023).

Russell, Stuart and Peter Norvig. *Artificial Intelligence: A Modern Approach*. 3rd edition. Pearson, 2009.

Salton, G., A. Wong and C. S. Yang. 'A Vector Space Model for Automatic Indexing'. *Communications of the ACM* (1975).

Schmidt, Morten, Lars Pedersen and Henrik Toft Sørensen. 'The Danish Civil Registration System as a Tool in Epidemiology'. *European Journal of Epidemiology* (2014).

Schmidt, Morten et al. 'The Danish Health Care System and Epidemiological Research: From Health Care Contacts to Database Records'. *Clinical Epidemiology* (2019).

Schmidt, Morten et al. 'The Danish National Patient Registry: A Review of Content, Data Quality, and Research Potential'. *Clinical Epidemiology* (2015).

SEARCH Collaborative Group et al. 'SLCO1B1 Variants and Statin-Induced Myopathy–a Genomewide Study'. *The New England Journal of Medicine* (2008).

Seifter, Julian, David Sloane and Austin Ratner. *Concepts in Medical Physiology*. Lippincott Williams & Wilkins, 2005.

Shapley, Lloyd S. 'A Value for N-Person Games' (1953).

Smith, Leslie N. and Nicholay Topin. 'Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates'. *arXiv* (2018). URL: arxiv.org/abs/1708.07120. preprint.

Sørup, Freja Karuna Hemmingsen et al. 'Sex Differences in Text-Mined Possible Adverse Drug Events Associated with Drugs for Psychosis'. *Journal of Psychopharmacology* (2020).

Srivastava, Nitish et al. 'Dropout: A Simple Way to Prevent Neural Networks from Overfitting'. *The journal of machine learning research* (2014).

Steg, Ph. Gabriel et al. 'ESC Guidelines for the Management of Acute Myocardial Infarction in Patients Presenting with ST-segment Elevation'. *European Heart Journal* (2012).

Stekhoven, Daniel J. and Peter Bühlmann. 'MissForest—Non-Parametric Missing Value Imputation for Mixed-Type Data'. *Bioinformatics* (2012).

Steyerberg, Ewout W. et al. 'Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research'. *PLOS Medicine* (2013).

Taylor, Fiona et al. 'Statins for the Primary Prevention of Cardiovascular Disease'. *Cochrane Database of Systematic Reviews* (2013).

The Danish Clinical Quality Program (RKKP). URL: rkkp.dk/in-english/ (visited on 28/11/2023).

Therneau, Terry M. *A Package for Survival Analysis in R*. manual. 2023.

Thompson, Paul D., Priscilla Clarkson and Richard H. Karas. 'Statin-Associated Myopathy'. *JAMA* (2003).

Thorn, Caroline F., Teri E. Klein and Russ B. Altman. 'PharmGKB: The Pharmacogenomics Knowledge Base'. *Pharmacogenomics: Methods and Protocols*. Humana Press, 2013.

Thygesen, Kristian et al. 'Fourth Universal Definition of Myocardial Infarction (2018)'. *European Heart Journal* (2019).

Topol, Eric J. 'High-Performance Medicine: The Convergence of Human and Artificial Intelligence'. *Nature Medicine* (2019).

Turchin, Alexander et al. 'Using Regular Expressions to Abstract Blood Pressure and Treatment Intensification Information from the Text of Physician Notes'. *Journal of the American Medical Informatics Association : JAMIA* (2006).

Tutz, Gerhard and Matthias Schmid. *Modeling Discrete Time-to-Event Data*. 1st ed. 2016 edition. Springer, 2016.

Ullmann, Theresa, Christian Hennig and Anne-Laure Boulesteix. 'Validation of Cluster Analysis Results on Validation Data: A Systematic Framework'. *WIREs Data Mining and Knowledge Discovery* (2022).

Upala, S., A. Sanguankeo and V. Jaruvongvanich. 'Gallstone Disease and the Risk of Cardiovascular Disease: A Systematic Review and Meta-Analysis of Observational Studies'. *Scandinavian Journal of Surgery* (2017).

Van der Velden, Bas H. M., Hugo J. Kuijf, Kenneth G. A. Gilhuijs and Max A. Viergever. 'Explainable Artificial Intelligence (XAI) in Deep Learning-Based Medical Image Analysis'. *Medical Image Analysis* (2022).

Van Dongen, Stijn. 'Graph Clustering Via a Discrete Uncoupling Process'. *SIAM Journal on Matrix Analysis and Applications* (2008).

Van Houwelingen, Hans C. 'Dynamic Prediction by Landmarking in Event History Analysis'. *Scandinavian Journal of Statistics* (2007).

Vaswani, Ashish et al. 'Attention Is All You Need'. *Advances in Neural Information Processing Systems*. 2017.

Visseren, Frank L J et al. '2021 ESC Guidelines on Cardiovascular Disease Prevention in Clinical Practice'. *European Heart Journal* (2021).

Wiegrebe, Simon et al. 'Deep Learning for Survival Analysis: A Review'. *arXiv* (2023). URL: arxiv.org/abs/2305.14961. preprint.

Wirth, Janine et al. 'Presence of Gallstones and the Risk of Cardiovascular Diseases: The EPIC-Germany Cohort Study'. *European Journal of Preventive Cardiology* (2015).

Yang, Li and Abdallah Shami. 'On Hyperparameter Optimization of Machine Learning Algorithms: Theory and Practice'. *Neurocomputing* (2020).

Zamorano, Jose Luis et al. '2016 ESC Position Paper on Cancer Treatments and Cardiovascular Toxicity Developed under the Auspices of the ESC Committee for Practice Guidelines: The Task Force for Cancer Treatments and Cardiovascular Toxicity of the European Society of Cardiology (ESC)'. *European Heart Journal* (2016).

Zheng, Yan et al. 'Gallstones and Risk of Coronary Heart Disease'. *Arteriosclerosis, Thrombosis, and Vascular Biology* (2016).

Appendices

Appendix A

Manuscript for Study I

Subgrouping multimorbid patients with ischemic heart disease by means of unsupervised clustering: A cohort study of 72,249 patients defined comprehensively by diagnoses prior to presentation

Short title: Unsupervised clustering of patients with ischemic heart disease

Amalie D. Haue, PhD[1,2,¶], Peter C. Holm, MSc[1,¶], Karina Banasik, PhD[1], Agnete T. Lundgaard, PhD[1], Victorine P. Muse, MEng[1], Timo Röder, MSc[1], David Westergaard, PhD[1], Piotr J. Chmura, MSc[1], Alex H. Christensen, PhD[2,3], Peter E. Weeke, PhD[2], Erik Sørensen, PhD[4], Ole B. V. Pedersen, PhD[4,5], Sisse R. Ostrowski, DMSc[4,6], Kasper K. Iversen, DMSc[3], Lars V. Køber, DMSc[2,6], Henrik Ullum, DMSc[7], Henning Bundgaard, DMSc[2,5*], Søren Brunak, PhD[1,8*]

[1]Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark
[2]Department of Cardiology, The Heart Center, Rigshospitalet, Copenhagen, Denmark
[3]Department of Cardiology, Copenhagen University Hospital, Herlev, Denmark.
[4]Department of Clinical Immunology, Copenhagen University Hospital, Copenhagen, Denmark
[4]Department of Clinical Immunology, Copenhagen University Hospital, Copenhagen, Denmark
[5]Department of Clinical Immunology, Zealand University Hospital, Køge, Denmark
[6]Department of Clinical Medicine, University of Copenhagen, Rigshospitalet, Copenhagen, Denmark
[7]Statens Serum Institut, Copenhagen, Denmark
[8]Copenhagen University Hospital, Rigshospitalet, Copenhagen, Denmark

[*]E-mail: soren.brunak@cpr.ku.dk (SB)

¶These authors contributed equally to this work.

Total word count (including Title Page, Abstract, Text, References, Tables and Figures Legends: 7,589

Abstract

Background: There are no methods for classifying patients with ischemic heart disease (IHD) based on the entire spectrum of pre-existing diseases. Such methods might be clinically useful due to the marked differences in presentation and course of disease.

Methods: A population-based cohort study from a Danish secondary care setting of patients with IHD (2004-2016) and subjected to a coronary angiography (CAG) or coronary computed tomography angiography (CCTA). Data sources were The Danish National Patient Registry, in-hospital laboratory data, and genetic data from Copenhagen Hospital Biobank. Comorbidities included diagnoses assigned prior to presentation of IHD. Patients were clustered by means of the Markov Clustering Algorithm using the entire spectrum of registered multimorbidity. The two prespecified outcomes were: New ischemic events (including death from IHD causes) and death from non-IHD causes. Patients were followed from date of CAG/CCTA until one of the two outcomes occurred or end of follow-up, whichever came first. Biological and clinical appropriateness of clusters was assessed by comparing risks (estimated from Cox proportional hazard models) in clusters and by phenotypic and genetic enrichment analyses, respectively.

Findings: In a cohort of 72,249 patients with IHD (mean age 63.9 years, 63.1% males), 31 distinct clusters (C1-31, 67,136 patients) were identified. Comparing each cluster to the 30 others, seven clusters (9,590 patients) had statistically significantly higher or lower risk of new ischemic events (five and two clusters, respectively). 18 clusters (35,982 patients) had a higher or lower risk of death from non-IHD causes (12 and six clusters, respectively). All clusters at increased risk of new ischemic events, associated with risk of death from non-IHD causes as well. Cardiovascular or inflammatory diseases were commonly enriched in clusters (13), and distributions for 24 laboratory test results differed significantly across clusters.

Clusters enriched for cerebrovascular diseases were generally not at increased risk of the two outcomes. Polygenic risk scores were increased in a total of 15 clusters (48.4%).

Conclusions: Clustering of patients with IHD based on pre-existing comorbidities identified subgroups of patients with significantly different clinical outcomes and presented a tool to rank pre-existing comorbidities based on their association with clinical outcomes. This novel method may support better classification of patients and thereby differentiation of treatment intensity depending on expected outcomes in subgroups.

Non-standard abbreviations

CAG: Coronary arteriography

CCTA: Coronary computed tomography angiography

ICD-10: International Statistical Classification of Diseases and Related Health Problems 10th Revision

IHD: Ischemic heart disease

MCL: Markov clustering

NPR: Danish National Patient Registry

O/E-ratio: Observed-expected-ratio

PRS: Polygenic risk score

Introduction

Ischemic heart disease (IHD) is a common, chronic, and complex disease and mode of onset, disease burden and disease progression vary considerably between patients[1–3]. This heterogeneity relates to several factors, but a major contribution is multimorbidity as more than 85% of IHD patients have been diagnosed with other chronic diseases; a phenomenon coined cardiometabolic multimorbidity[4,5]. The increased mortality in patients with cardiometabolic multimorbidity is generally only related to single disease states, such as obstructive lung disease, diabetes, or stroke, although it is known that the risk of cardiovascular diseases is increased in many chronic, inflammatory disorders[6,7]. As more patients at older age and with more and more co-morbidities are seen, new methods for characterizing and studying cardiometabolic multimorbidity are needed[8–12].

Unsupervised clustering algorithms and other network-based methods can systematically reveal structures in large, feature-rich datasets and may be used to identify distinct patient subgroups within a heterogenous population[13,14]. Proof-of-concept analyses of cardiovascular phenotypes, including IHD, heart failure, diabetes, and atrial fibrillation have already been performed[15–21]. While these studies successfully identify subgroups resembling those from traditional analyses, they often fail to demonstrate that clustering analysis leads to novel understanding of a given dataset. Rather, they are typically restricted to characterize high-, medium-, and low-risk subgroups which by and large resemble more conservative approaches from an earlier, less data-rich, epoch[22].

For decades, Danish healthcare registries have had a strong position within epidemiological research[22–24]. Given the opportunities for using clinical data more extensively, we carried out an unsupervised clustering analysis of 72,249 patients with IHD based on their entire disease

history until IHD onset. Explicitly, we wanted to classify IHD based on the entire spectrum of multimorbidity. We identified distinct patient subgroups derived from a pool of 3,046 different diagnoses assigned prior to IHD onset. The biological and clinical factors characteristic to distinct patient subgroups identified by unsupervised clustering analysis were asserted by assessments of their associations with clinical outcomes and clinical characteristics, laboratory data, and genetics (Figure 1).

Methods

Data sources, study population, and outcomes

Data from the Danish National Patient Registry (NPR) and the Danish Registry for Causes of Death were linked to in-hospital electronic health data covering the two Danish healthcare regions in Eastern Denmark (~2.9 mil inhabitants), and the Copenhagen Hospital Biobank Cardiovascular Disease Cohort[23,25,26]. Linkage of different healthcare data sources was obtained via the personal identification number and only patients admitted to a hospital in Eastern Denmark in years 2004 to 2016 were considered[27]. We identified all patients in NPR who were assigned an ICD-10 code for IHD[28]. To increase the positive predictive value of IHD diagnoses and align included patients in time, we further required that patients had been subjected to coronary arteriography (CAG) or coronary computed tomography angiography (CCTA). To qualify that CAG/CCTAs were conclusive for IHD, patients were only included if the CAG/CCTA was performed during a contact where patients were assigned an ICD-10 code for IHD. We set the earliest CAG/CCTA fulfilling this criterium as the index date and excluded patients with an index date before year 2004 or after 2016 (Fig 2).

There were two predefined outcomes: 1) New ischemic events and 2) Death from other causes than IHD (non-IHD causes). The outcome “new ischemic event” was a composite outcome of a) hospitalization minimum 30 days after index for myocardial infarction or

unstable angina pectoris (i.e., hospitalization with myocardial infarction or unstable angina pectoris as the primary diagnosis), b) revascularization not related to the index date, and c) any death where IHD was listed as the primary or secondary cause. Outcomes were obtained from NPR and Danish Registry for Causes of Death. Eligible codes for inclusion, outcomes and specific cutoffs are available in S1 Fig and S1 Table.

Data preprocessing and application of the Markov cluster algorithm

We performed a clustering analysis of included patients based on their multimorbidity prior to their IHD diagnosis (index) using the Markov cluster (MCL) algorithm[29]. Multimorbidity was represented as patient-specific vectors using diagnoses assigned prior to or at index. ICD-10 codes assigned to less than five patients (n=1,673) were excluded from the analysis. As we focused the studies on multimorbidity in IHD, ICD-10 codes for IHD (I20-I25) were excluded from patients-specific vectors. Thus, a total of 3,046 ICD-10 codes were the basis for constructing a patient similarity network that was used as MCL algorithm input. Patient-specific vectors of length 3,046 with integers indicating the number of times a patient had been assigned a particular ICD-10 code. The length of the vectors corresponded to the number of input features (ICD-10 codes). By combining the patient-specific vectors from all included patients, a matrix of size n x m was constructed, where n indicates the number of included patients and m indicates the number of input features (ICD-10 codes). Following a series of preprocessing steps described in S1 Appendix, a patient similarity network was created based on the n x m matrix and used as input for the MCL algorithm[30]. Resulting clusters were denoted *C* followed by an integer indicating the rank of the clusters with respect to cluster size (number of patients in that cluster). Thus, C1 denotes the largest cluster and cluster-membership was used to denote a cluster as a covariate in subsequent analyses. Robustness of clustering was assessed by generating a series of diluted and shuffled versions

of the resulting clusters (reference clustering), and their similarity was quantified using the variance of the information measure as previously described[31]. Explicitly, a series of diluted and shuffled versions of the input graph were generated[32]. In total, 20 variations of the input graph were constructed by shuffling and deleting edges, respectively. The variation in the graphs was then quantified by means of variation of the information meassure. Details regarding the MCL settings and a description of cluster robustness assessment are available in the S1 Appendix.

Preprocessing of laboratory and genetic data

Clusters were characterized by laboratory and genetic data based on the subset of patients where these data types were available. A panel of 25 different lab parameters was included in the analyses. Only tests taken up to 90 days before index or at the day of index were included. Included lab tests were plasma levels of potassium, sodium, hemoglobin, estimated glomerular filtration rate (eGFR), creatinine, carbamide, glucose, troponin (I/T), HDL cholesterol, LDL cholesterol, total cholesterol, leukocytes, C-reactive protein, lymphocytes, monocytes, neutrophils, basophiles, platelets, INR, alanine transaminase, albumin, alkaline phosphatase, bilirubin, and triglyceride. For every cluster, a *score* was computed based on the number of patients with a lab test below, within, or above the standard reference value, indicated by -1, 0 and 1, respectively. The *score* was defined as the mean of the summarized values per cluster.

Autosomal genotype data were obtained by identifying included patients who were also among the study participants in the Copenhagen Hospital Biobank – Cardiovascular Disease Cohort[26]. For included patients with genetic data available, we calculated polygenic risk scores (PRSs) for 14 traits, obtained from nine GWAS meta-analyses (atrial fibrillation, BMI-

adjusted non-insulin diabetes, chronic kidney disease, HDL cholesterol levels, heart failure, LDL cholesterol levels, stroke, total cholesterol levels, triglyceride levels) and five GWAS (acute myocardial infarction, coronary artery disease, diastolic blood pressure, non-alcoholic fatty liver disease, systolic blood pressure)[39–42]. PRSs were calculated using the "LDpred2-auto" algorithm, implemented in the R package "bigsnpr" (version 1.11.6) with R version 4.0.0 and the workflow management system Snakemake[43–45]. Each trait's PRS distribution was scaled to a mean of zero and a standard deviation of one.

Statistical analyses of clusters identified by the MCL algorithm

As the study was designed to identify patient subgroups and not individual variation, clusters of size < 500 were excluded from the remaining analyses. Mean age at IHD onset in each cluster was compared to the mean age at onset in all the other clusters using Tukey's Honest Significant Difference (HSD) method. Significance level was set to 0.05 and P-values were adjusted using the Holm method assuming 465 tests (adj. P-val.).

To investigate the association between cluster-membership and the competing risks of new ischemic events and death from non-IHD causes, we used Cox proportional-hazards models (Cox models). Patients were followed from index until occurrence of either of the two outcomes, or end of follow-up (year 2018), whichever came first. The dependent variable was either risk of new ischemic events or death from non-IHD causes, and the independent variables were cluster, sex, and age at index. To age-adjust the models, analyses were performed using restricted cubic spline with three knots for age at index. Follow-up time was truncated to a maximum of five years. For each cluster, hazard ratios (HRs) and 95% confidence intervals (CIs) were estimated by comparing HRs for the members of the cluster with the HRs with that of non-members.

Further characterization of clusters consisted of: (1) phenotypic enrichment analysis, (2) characterization of clusters with respect to their laboratory profiles and (3) a test for genetic enrichment. The phenotypic enrichment analysis was carried out based on ratios between observed (O) and expected (E) frequencies of diagnoses in the clusters (O/E-ratios). That is, ratios between the frequencies of ICD-10 codes in each cluster (observed frequencies) and the frequencies of ICD-10 codes in the entire population (expected frequencies) were calculated and expressed as O/E-ratios[46]. In subsequent characterization of clusters, enrichment denoted O/E-ratios > 2, and clusters were characterized as having little enrichment if the sum of the ten largest O/E-ratios < 50. Inverse changes were used to denote O/E-ratios between 0 and 1.

Hierarchical clustering was applied to estimate the cluster similarity with respect to the laboratory tests using the Euclidean distance between the *score* of each cluster for each test.

For each of the fourteen traits we calculated PRSs for, we used Wilcoxon rank-sum tests to compare the PRS distribution of each cluster to the combined PRS distribution of PRSs in all other clusters. Resulting P-values were converted to the false discovery rate (FDR) to account for multiple testing, with a total of 434 tests. We report effect sizes as calculated by the "wilcox.test" function built into R version 4.0.0. Level of significance was set to FDR < 0.05, assuming 434 tests.

Further details regarding preprocessing and analyses of laboratory and genetic data are available in the S2 Appendix.

Ethics approvals and data access

The study was approved by The National Ethics Committee (1708829, 'Genetics of CVD'—a genome-wide association study on repository samples from Copenhagen Hospital Biobank), The Danish Data Protection Agency (ref: 514-0255/18-3000, 514-0254/18-3000, SUND-2016-50), The Danish Health Data Authority (ref: FSEID-00003724 and FSEID-00003092), and The Danish Patient Safety Authority (3-3013-1731/1/). Danish personal identification numbers were pseudonymized prior to any analysis. Study design, methods and results were reported in agreement with the STROBE statement[47].

Results

Cohort demographics and co-morbidities

A total of 72,249 patients (63.1% males, mean age 63.9 years) were included (Table 1). Angina pectoris (I20) was the most common IHD diagnosis (38,239 patients, 52.9%), followed by acute myocardial infarction (I21) (33,229 patients, 46.0 %) and chronic IHD (I25) (22,750 patients, 31.5%). The most common co-morbidity prior to the IHD index was hypertension (I10.9) (24,818 patients, 34.4%) followed by dyslipidemia (E78.0) (12,780 patients, 17.7%) and non-insulin dependent diabetes (E11.9) (7,551 patients, 10.5%). Prior to index, the mean number of diagnoses per patient was 8.1. A total of 68,103 patients (94.3%) had co-morbidities registered prior to index. The overall incidence (new ischemic events and death from non-IHD causes) was 94 events per 1000 person-years (Table 1).

Unsupervised clustering of multimorbid patients with IHD

In the cohort, the MCL algorithm identified 36 distinct clusters based on the set of 3,046 ICD-10 codes assigned to the patients prior to or at index. The 36 clusters contained a total of 68,084 patients. Expectedly, the remaining 4,365 patients (6.0% of included patients) that did not cluster were primarily patients with no diagnoses prior to index (>99%). Further, cluster

247 robustness was assessed as described in Methods, where the variation of information measure
248 less than 2 if 25% of the edges in the input graph were deleted or shuffled (S4 Figure). Next,
249 the 31 of the 36 clusters with >500 patients (67,136 patients) were characterized (Table 2).
250 Using Tukey's HSD to compare the age at index between all 31 clusters (a total of 466
251 combinations), we found significant differences in 391 comparisons (84.1%, S3 Table). For
252 demographics of patients that did not cluster or were in clusters of size < 500, see S4 Table.

Clusters, clinical outcomes, and phenotypic enrichment

To assess if the unsupervised clustering identified patient subgroups at different risks of disease progression, we used cluster-membership (C1-C31) as a covariate in a series of Cox models. A total of 14,679 patients experienced a new ischemic event during follow-up and 10,684 patients died from other causes than IHD. Mean follow-up time was 3.72 years (Table 1). Risks for new ischemic events and death from non-IHD causes in each cluster were compared to the pooled risk for patients in the remaining 30 clusters. The survival analysis demonstrated that the MCL algorithm stratified patients according to risk of new ischemic events and death from non-IHD causes (Fig 3). Comparing each cluster (n=1) to all the others (n=30), a total of seven clusters (20,221 patients) had a statistically significantly higher or lower risk of new ischemic events (Adj. P-val. < 0.05). Five clusters (9,590 patients) and two clusters (10,631 patients) were at increased and decreased risk of new ischemic events, respectively. Similarly, a total of 18 clusters (43,173 patients) had a statistically significantly higher or lower risk of death from non-IHD causes (Adj. P-val. < 0.05); where 14 clusters (21,282 patients) and four clusters (21,891 patients) were at increased or decreased risk of death from non-IHD causes. All clusters at increased risk of new ischemic events, associated with risk of death from non-IHD causes as well. The same was true for the two clusters at decreased risk of new ischemic events, i.e., these clusters were at decreased risk of death from non-IHD causes as well. A total of 13 clusters, (23,963 patients) were not have altered risk of the two outcomes, when compared to the other clusters (Table 2).

The distribution of O/E-ratios was heavily left-skewed as less than 99% (n=101) of all O/E-ratios were >10 and roughly 7% (n=887) of all O/E-ratios were >2. About 60% of all O/E-ratios (n=8,056) were in the range of 0 and 1 corresponding to inverse changes. Generally, clusters that had high risk of new ischemic events or death from non-IHD causes were also

characterized by large, summarized O/E-values corresponding to a high degree of multimorbidity (S5 Table 5). The results of the enrichment analysis were summarized according to nine different disease categories: (1) diabetes mellitus, (2) cardiac diseases, (3) diseases affecting the upper airways, (4) cerebrovascular diseases, (5) infections and other acquired diseases, (6) gynecologic diseases, (7) Inflammatory and degenerative of the musculoskeletal system, (8) diseases of the urinary system, and (9) hypertension (Fig. 4).

An in-depth characterization of clusters enriched for cardiometabolic or -vascular diseases, degenerative or inflammatory diseases and clusters characterized by little enrichment and inverse changes is provided in the following paragraphs.

Clusters enriched for cardiometabolic and -vascular diseases

Four of the five clusters at increased risk of new ischemic events (and death from non-IHD causes) were enriched for diabetes (C5, C18, C23, and C30). In these four clusters, HRs ranged from 1.40 (C5, 95%CI: 1.30;1.50, adj. P-val. < 0.001) to 1.88 (C30, 95%CI: 1.60;2.00, adj. P-val. < 0.001) with a significant difference in age at index (C5: 63.9 years, C30: 61.2 years, Adj. P-val. < 0.001, TukeyHSD). C18 and C23 were only enriched for insulin-dependent diabetes, but differed in that C18 was also enriched for insulin-dependent diabetes with vascular complications and periphery atherosclerosis. In contrast, C5 was only enriched for non-insulin dependent diabetes and included diabetes with as well as without complications. Lastly, C30 was only enriched for diabetes with complications (insulin and non-insulin dependent) and was the diabetes cluster enriched for chronic kidney disease and bacterial infections, as well (S5 Table 5).

Other cardiac diseases that displayed enrichment were supraventricular arrythmias (C4), cardiomyopathies (C9), and valve diseases (C20). Of the three clusters, only C9 had increased risk of new ischemic events (HR: 1.31 (C9, 95%CI: 1.20;1.44, Adj. P-val: < 0.001). Risk of death from non-IHD causes was 1.79 (95%CI: 1.60;2.00, adj. P-val. < 0.001). In contrast, C4 and C20 only had increased risk of death from non-IHD causes with HRs of 1.49 (C4, 95%CI: 1.34;1.59, adj. P-val. < 0.001) and 1.78 (C20, 95%CI: 1.54;2.04, adj. P-val. < 0.001). Interestingly, the cluster enriched for cerebrovascular diseases (C27) did not have altered risk of any of the two outcomes. In sum, all clusters that had increased risk of new ischemic events were enriched for cardiometabolic diseases, albeit not all clusters enriched for cardiometabolic and -vascular diseases had increased risk of new ischemic events (Table 2 and S5 Table 5).

Clusters enriched for degenerative or inflammatory diseases

Six clusters (C7, C13, C14, C22, C26, and C31) were enriched for diagnoses describing degenerative or inflammatory diseases, i.e., osteoarthritis (C7), degenerative spine disease (C13 and C22), chronic obstructive pulmonary disease (C14), asthma (C26), and rheumatoid arthritis (C31). Remarkably, none of the four clusters had increased risk of new ischemic events and only one cluster (C14) had increased risk of death from non-IHD causes (HR: 3.39, 95%CI: 3.09;3.71, adj. P-val. < 0.001). Conversely, C7 and C13 had reduced risk of death from non-IHD causes (C7, HR: 0.61, 95%CI: 0.52;0.72, adj. P-val. < 0.001 and C13, HR: 0.58, 95%CI: 0.45;0.74, adj. P-val. < 0.001). Age at index for the clusters enriched for degenerative or inflammatory diseases range between 58.6 years (C13) and 69.2 years (C22) (Table 2). Taken together, these findings hint to the dual nature of inflammation as a potential disease modifier as well as a risk factor.

Clusters characterized by little enrichment and inverse changes

Six clusters (C1, C2, C3, C6, C15, and C17) were characterized by little enrichment, which included the two clusters with reduced risk of new ischemic events (C2, HR: 0.82, 95%CI: 0.76;0.89, adj. P-val. < 0.001 and C3, HR: 0.76, 95%CI: 0.52;0.69, adj. P-val. < 0.001). Not surprisingly, none of these six clusters had increased risk of either of the two outcomes, but three clusters (C2, C3, and C6) had reduced risk of death from non-IHD causes (C2, HR: 0.60, 95%CI: 0.52;0.69, adj. P-val. < 0.001, C3, HR: 0.59, 95%CI: 0.59;0.69, adj. P-val. < 0.001 and C6, HR: 0.68, 95%CI: 0.57;0.79, adj. P-val. < 0.001) (Table 2). It was a common attribute of the clusters without altered risk of any of the two outcomes that O/E-ratios for hypertension and dyslipidemia were among the largest. In contrast, diabetes, heart failure, and chronic obstructive pulmonary disease frequently displayed inverse changes (O/E-ratios < 1) in these clusters (S5 Table). Taken together, these observations indicate that risk of disease progression in this populations necessitates a more sophisticated analysis of multimorbidity.

For a list with results of the enrichment analysis for all clusters, including the 13 clusters not described above, S5 Table 5.

Clusters and their association with laboratory measurements and genetic data

Clusters were also characterized by means of datatypes not included among the MCL algorithm input features. For patients in the 31 clusters, we had laboratory measurements on 30,755 (49.5%) and genetic data on 19,422 (31.3%). To assess if the phenotypic differences captured by the MCL algorithm were also reflected in laboratory measurements, we tested if the distributions of test results within and out of reference ranges differed significantly. There were significantly different distributions of tests within and out of reference ranges in clusters

for the 24 most frequent tests. Overall, this indicates that the phenotypic patterns within the entire spectrum of cardiovascular multimorbidity registered before index correlate with results of clinical laboratory tests (S6 Table). Thus, these findings are a strong indicator that the patterns captured by the MCL algorithm are biologically relevant. For a graphical summary of the laboratory scores in each cluster, see S5 Figure.

Finally, we identified 41 cases (out of 434 tests) where the PRS distribution for a specific trait in a cluster was significantly different from that trait's combined PRS distribution of the other 30 clusters. Among these cases, we found the largest effects size to be a higher genetic risk for atrial fibrillation in cluster C4 (0.57, FDR < 0.001) as well as a higher genetic risk for non-insulin dependent diabetes in cluster C5 (0.55, FDR < 0.001). These findings are congruent with the results of the enrichment analysis for C4 and C5, respectively. In contrast, C1 (phenotypically characterized by inverse changes) had relatively large, positive effect sizes for systolic as well as diastolic blood pressure (0.20 and 0.16, FDR < 0.001). Similarly, there were positive effect sizes for total cholesterol and triglycerides in C6, which was also characterized by little phenotypic enrichment as well as a high degree of inverse changes. A list of significant effect sizes for the 41 significant cases, see S7 Table.

Discussion

In this study, we developed a novel, data-driven method for structuring the entire spectrum of multimorbidity by means of an unsupervised clustering analysis. In a cohort of 72,249 patients with IHD patients, we identified 31 distinct clusters (67,136 patients) based on 3,046 diagnoses assigned prior to or at index. By comparing risk of new ischemic events and death from non-IHD causes across clusters and then performing an enrichment analysis, we found that clusters at increased risk of new ischemic events were enriched for diabetes (four clusters) or cardiomyopathies (one cluster). Neither the cluster enriched for supraventricular

arrythmias, nor valve diseases had increased risk of new ischemic events. Degenerative and inflammatory diseases were enriched in a total of six clusters and displayed no clear trend in their relation to the outcomes. The results of the enrichment analysis were supported by trends in laboratory test results and clusters enriched for supraventricular arrythmias and non-insulin diabetes also had congruently, higher genetic risks.

The results of the study agree with common knowledge on risk of IHD, while also adding insights to the disease-diseases associations, which are currently underappreciated in the literature. The fact that clusters enriched for diabetes were generally the most high-risk clusters serves as a methodological reality check[6]. Added value of the study lies in the fact that the method allows for a more sophisticated description of such associations, as the method allows to study the entire spectrum of multimorbidity. For example, four clusters were enriched for diabetes, which is in line with the current paradigm that a single term is insufficient to describe a multifactorial disease, such as diabetes[18,31]. By integrating different data types, the findings indicate how phenotypic and genetic data complement each other, by exemplifying (1) that clustering analysis facilitates stronger genetic signals in patient subgroups and (2) that genetic data may unveil patterns not captured by phenotypic data alone.

In addition, the method developed in this study and subsequent findings add perspective to the relatively limited body of literature regarding associations between chronic inflammatory and cardiovascular diseases[7]. While previous studies have concluded that the risk of cardiovascular diseases is increased in most chronic inflammatory disorders, the results of our study indicate that pre-existing degenerative or inflammatory disorders in patients with IHD do not increase the risk of new ischemic events.

The pre-selected outcomes in the present study are also a unique aspect of the study, as previous clustering analyses within the cardiovascular domain studies have mainly analyzed all-cause mortality[19,20]. This aspect of the study allows to distinguish between risk of progression related to IHD and risk of progression that is related to comorbidity drawing attention to important aspects of multimorbidity in this domain. For example, clusters enriched for supraventricular arrythmias and chronic obstructive pulmonary disease, respectively, only had increased risk of death from non-IHD causes. The study design, including the enrichment analysis, also revealed that classical risk factors for IHD (e.g., hypertension and dyslipidemia) did not drive the clustering. This finding agrees with previously published comorbidity phenotypes in patients with IHD[20]. We argue that the present study displays that continuous exploration and characterization of multimorbidity in IHD are key elements in optimizing the exploit the full potential of continuously developing treatment strategies.

Previous clustering analyses within the cardiovascular domain have typically included either thousands of patients or hundreds of input features, but not both[16,17]. For example, Hall et al. defined multimorbidity using only eight different chronic conditions, whereas Crowe et al. defined multimorbidity with reference to 20 predefined conditions[19,20]. Thus, the scale of our study exceeds that of previous work, as it includes more than 70,000 patients and more than 3,000 input features. And further, we limited the risk of introducing bias by not exerting feature selection prior to clustering.

The two main limitations with respect to the data foundation are that (1) owing to the novelty of the method, there were no standardized way of assessing the representation of multimorbidity and (2) it was only a subset for which laboratory and genetic data were

available. These challenges are naturally overcome in clustering analyses based on data from randomized controlled trials, such as the studies by Inohara et al, and Karwath et al.[17,21] However, in the present, data-rich era, we argue that it is highly important to develop methods for structuring and studying other data than what is being collected for trials. Ideally, the two approaches, based on nationwide data and randomized controlled trials, respectively, will complement each other; and will facilitate more precise identification of patients who are likely to benefit from different treatment options as well as guide optimized selection of patients for randomized controlled trials.

In conclusion, the study further showcases the strengths of a more fine-grained analysis of patient subgroups, which, in turn, may pave the way for successful implementation of precision medicine. Owing to its flexibility, the comprehensive, data-driven analysis of cardiovascular multimorbidity represents a novel method for characterizing multimorbidity in IHD with great potential of applying it to other diseases of interest or other clinical data. Such trends may guide clinical decision making in cases, where for example it is not obvious how to manage the angiographic findings or the combination of drugs that a specific patient will benefit most from.

In conclusion, the present study cements the complexity of multimorbid patients with IHD and exemplifies the clinical relevance of a more fine-grained patient subgrouping by carrying out a cluster-based risk-stratifying the cohort. Further, owing to its flexibility, the comprehensive, data-driven method of cardiovascular multimorbidity presented here represents a novel method for characterizing multimorbidity in IHD with great potential. Improved patient subgrouping may be critical guide future clinical decision making in cases,

where it is non-trivial how to manage the angiographic findings or to find the optimal combination of drugs for a given patient.

Funding

This work was financially supported by Novo Nordisk Foundation (Grants NNF17OC0027594 and NNF14CC0001) and the Innovation Fund Denmark via the NordForsk project PM Heart (5184-00102B).

Acknowledgement

The authors would like to thank (1) research programmer, Troels Siggaard, Novo Nordisk Foundation Center for Research, University of Copenhagen, Denmark for continuous and reliable infrastructure support, and (2) Head of Cardiovascular Research, Hilma Hólm, deCODE genetics, Iceland for insightful comments

Data access

Application for registry data access can be made to the Danish Health Data Authority (contact: servicedesk@sundhedsdata.dk). Anyone wishing access to the data and use them for research will be required to meet research credentialing requirements as outlined at the authority's web site: sundhedsdatastyrelsen.dk/da/english/health_data_and_registers/research_services. Requests are normally processed within three to six months.

Code availability statement

The code used to generate the results including the clustering pipeline will be made publicly available upon publication.

References

1. Antman, E. M. & Braunwald, E. Managing Stable Ischemic Heart Disease. *N. Engl. J. Med.* **382**, 1468–1470 (2020).
2. Ferraro, R. *et al.* Evaluation and Management of Patients With Stable Angina: Beyond the Ischemia Paradigm. *J. Am. Coll. Cardiol.* **76**, 2252–2266 (2020).
3. Nabel, E. G. & Braunwald, E. A tale of coronary artery disease and myocardial infarction. *N. Engl. J. Med.* **366**, 54–63 (2012).
4. Forman, D. E. *et al.* Multimorbidity in Older Adults With Cardiovascular Disease. *J. Am. Coll. Cardiol.* **71**, 2149–2161 (2018).
5. Afilalo, J. *et al.* Frailty Assessment in the Cardiovascular Care of Older Adults. *J. Am. Coll. Cardiol.* **63**, 747–762 (2014).
6. The Emerging Risk Factors Collaboration. Association of Cardiometabolic Multimorbidity With Mortality. *JAMA* **314**, 52–60 (2015).
7. Dregan, A., Charlton, J., Chowienczyk, P. & Gulliford, M. C. Chronic Inflammatory Disorders and Risk of Type 2 Diabetes Mellitus, Coronary Heart Disease, and Stroke. *Circulation* **130**, 837–844 (2014).
8. Glynn, L. G. Multimorbidity: another key issue for cardiovascular medicine. *The Lancet* **374**, 1421–1422 (2009).
9. Joshi, A., Rienks, M., Theofilatos, K. & Mayr, M. Systems biology in cardiovascular disease: a multiomics approach. *Nat. Rev. Cardiol.* **18**, 313–330 (2021).
10. Khera Amit V. & Kathiresan Sekar. Is Coronary Atherosclerosis One Disease or Many? *Circulation* **135**, 1005–1007 (2017).
11. Rahimi, K., Lam, C. S. P. & Steinhubl, S. Cardiovascular disease and multimorbidity: A call for interdisciplinary research and personalized cardiovascular care. *PLOS Med.* **15**, e1002545 (2018).

12. Haue, A. D. *et al.* Temporal patterns of multi-morbidity in 570157 ischemic heart disease patients: a nationwide cohort study. *Cardiovasc. Diabetol.* **21**, 87 (2022).
13. Kaufman, L. & Rousseeuw, P. J. *Finding Groups in Data: An Introduction to Cluster Analysis*. (John Wiley & Sons, 2009).
14. Lee, L. Y., Pandey, A. K., Maron, B. A. & Loscalzo, J. Network medicine in Cardiovascular Research. *Cardiovasc. Res.* **117**, 2186–2202 (2021).
15. Shah, R. V. *et al.* Association of Multiorgan Computed Tomographic Phenomap With Adverse Cardiovascular Health Outcomes: The Framingham Heart Study. *JAMA Cardiol.* **2**, 1236–1246 (2017).
16. Ahmad, T. *et al.* Clinical implications of chronic heart failure phenotypes defined by cluster analysis. *J. Am. Coll. Cardiol.* **64**, 1765–1774 (2014).
17. Inohara, T. *et al.* Association of of Atrial Fibrillation Clinical Phenotypes With Treatment Patterns and Outcomes: A Multicenter Registry Study. *JAMA Cardiol.* **3**, 54–63 (2018).
18. Ahlqvist, E. *et al.* Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. *Lancet Diabetes Endocrinol.* **0**, (2018).
19. Hall, M. *et al.* Multimorbidity and survival for patients with acute myocardial infarction in England and Wales: Latent class analysis of a nationwide population-based cohort. *PLOS Med.* **15**, e1002501 (2018).
20. Crowe, F. *et al.* Comorbidity phenotypes and risk of mortality in patients with ischaemic heart disease in the UK. *Heart* **106**, 810–816 (2020).
21. Karwath, A. *et al.* Redefining β-blocker response in heart failure patients with sinus rhythm and atrial fibrillation: a machine learning cluster analysis. *The Lancet* **398**, 1427–1435 (2021).

22. Bowman, L. *et al.* Understanding the use of observational and randomized data in cardiovascular medicine. *Eur. Heart J.* **41**, 2571–2578 (2020).
23. Schmidt, M. *et al.* The Danish health care system and epidemiological research: from health care contacts to database records. *Clin. Epidemiol.* **11**, 563–591 (2019).
24. Hemingway, H. *et al.* Big data from electronic health records for early and late translational cardiovascular research: challenges and potential. *Eur. Heart J.* **39**, 1481–1495 (2018).
25. Helweg-Larsen, K. The Danish Register of Causes of Death. *Scand. J. Public Health* **39**, 26–29 (2011).
26. Sørensen, E. *et al.* Data Resource Profile: The Copenhagen Hospital Biobank (CHB). *Int. J. Epidemiol.* (2020) doi:10.1093/ije/dyaa157.
27. Schmidt, M., Pedersen, L. & Sørensen, H. T. The Danish Civil Registration System as a tool in epidemiology. *Eur. J. Epidemiol.* **29**, 541–549 (2014).
28. Sundbøll, J. *et al.* Positive predictive value of cardiovascular diagnoses in the Danish National Patient Registry: a validation study. *BMJ Open* **6**, (2016).
29. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–84 (2002).
30. MCL - a cluster algorithm for graphs. http://micans.org/mcl/.
31. Kirk, I. K. *et al.* Linking glycemic dysregulation in diabetes to symptoms, comorbidities, and genetics through EHR data mining. *eLife* **8**, e44941 (2019).
32. Meilă, M. Comparing clusterings—an information based distance. *J. Multivar. Anal.* **98**, 873–895 (2007).
33. Nielsen, J. B. *et al.* Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. *Nat. Genet.* **50**, 1234–1239 (2018).

547 34. Mahajan, A. *et al.* Fine-mapping type 2 diabetes loci to single-variant resolution using
548 high-density imputation and islet-specific epigenome maps. *Nat. Genet.* **50**, 1505–1513
549 (2018).
550 35. Wuttke, M. *et al.* A catalog of genetic loci associated with kidney function from analyses
551 of a million individuals. *Nat. Genet.* **51**, 957–972 (2019).
552 36. Surakka, I. *et al.* The impact of low-frequency and rare variants on lipid levels. *Nat.*
553 *Genet.* **47**, 589–597 (2015).
554 37. Shah, S. *et al.* Genome-wide association and Mendelian randomisation analysis provide
555 insights into the pathogenesis of heart failure. *Nat. Commun.* **11**, 163 (2020).
556 38. Malik, R. *et al.* Multiancestry genome-wide association study of 520,000 subjects
557 identifies 32 loci associated with stroke and stroke subtypes. *Nat. Genet.* **50**, 524–537
558 (2018).
559 39. Jiang, L., Zheng, Z., Fang, H. & Yang, J. A generalized linear mixed model association
560 tool for biobank-scale data. *Nat. Genet.* **53**, 1616–1621 (2021).
561 40. van der Harst, P. & Verweij, N. Identification of 64 Novel Genetic Loci Provides an
562 Expanded View on the Genetic Architecture of Coronary Artery Disease. *Circ. Res.* **122**,
563 433–443 (2018).
564 41. Hoffmann, T. J. *et al.* Genome-wide association analyses using electronic health records
565 identify new loci influencing blood pressure variation. *Nat. Genet.* **49**, 54–64 (2017).
566 42. Anstee, Q. M. *et al.* Genome-wide association study of non-alcoholic fatty liver and
567 steatohepatitis in a histologically characterised cohort☆. *J. Hepatol.* **73**, 505–515 (2020).
568 43. Privé, F., Arbel, J. & Vilhjálmsson, B. J. LDpred2: better, faster, stronger. *Bioinformatics*
569 **36**, 5424–5431 (2020).
570 44. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation
571 for Statistical Computing, 2020).

45. Mölder, F. *et al.* Sustainable data analysis with Snakemake. Preprint at https://doi.org/10.12688/f1000research.29032.2 (2021).

46. Violán, C. *et al.* Multimorbidity patterns with K-means nonhierarchical cluster analysis. *BMC Fam. Pract.* **19**, 108 (2018).

47. Elm, E. von *et al.* The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *J. Clin. Epidemiol.* **61**, 344–349 (2008).

Figure legends

Fig 1: Graphical overview of study. Conceptual figure displaying the study design. A: Assemblage of patient-specific vectors that were the basis for construction of a matrix and an *n* x *m* matrix, where *n* corresponds to the number of included patients and *m* corresponds to the number of diagnoses. B: Unsupervised clustering of IHD patients using the MCL algorithm, which was the basis for performing unsupervised clustering to identify distinct clusters and associating them with clinical outcomes. C: Risk of disease progression (new ischemic events or death from non-IHD causes) in clusters. Color bar indicates increased, not altered, or decreased risk for patients in one cluster relative to the patients not in that cluster. D: Phenotypic and genetic characterization of clusters. Red: Increased risk of both outcomes. IHD: Ischemic heart disease. MCL: Markov Clustering.

Fig 2: Flowchart: Data sources, study population, and outcomes. Gray: Identification. Blue: Screening. Red: Eligibility. Green: Inclusion and outcomes. AMI: Acute myocardial infarction. UAP: Unstable angina pectoris. NPR: The Danish National Patient Registry. IHD: ischemic heart disease (ICD-10 codes I20-I25). CAG: Coronary arteriography. CCTA: Coronary computed tomography angiography. ICD-10: International Statistical Classification of Diseases and Related Health Problems 10th Revision. SKS: Sundhedsvæsenets Klassifikationssystem (The Danish Health Authority Classification System).

Fig 3: Risk of new ischemic events and non-IHD causes stratified by cluster. Forest plots where clusters are shown against HR for new ischemic events (left) and death from non-IHD causes (right). X-axis: HR for a single cluster relative to mean HR of the 30 other clusters. Y-axis: Clusters arranged by risk of new ischemic events, increasing risk from top to bottom. Colors indicating significance. Dark green: Reduced risk of new ischemic events and

death from non-IHD causes. Lighter green: Reduced risk of death from non-IHD causes. Yellow: No significance. Orange: Increased risk of death from non-IHD causes. Red: Increased risk of new ischemic events and increased risk of death from non-IHD causes. IHD: Ischemic heart disease. HR: Hazzard ratio.

Fig 4: Infographic summarizing the results of the study. Center: Study cohort. Periphery: Graphical overview of results from clustering analysis, survival analysis and characterization of clusters. Arrows indicate disease categories (for details, see text). 1: Diabetes mellitus. 2: Cardiac diseases. 3: Diseases affecting the upper airways. 4: Cerebrovascular diseases. 5: Infections and other acquired diseases. 6: Gynecologic diseases. 7: Inflammatory and degenerative of the musculoskeletal system. 8: Diseases of the urinary system. 9: Hypertension. C1-31: Clusters. “Underline” indicates little enrichment. “*” indicates genetic enrichment. For underlying data, see S5 and S7 Tables.

Table 1: Patient demographics, co-morbidities, and outcomes

Cohort demographics	**Total**	**Males**	**Females**
Number of patients (%)	72,249	45,576 (63.1)	26,673 (36,1)
Mean age at index (SD)	63.9 (11.9)	62.9 (11.6)	65.6 (12.1)
IHD manifestations (ICD-10)	**Total**	**Males**	**Females**
Angina pectoris (I20)	38,239	22,628	15,611
Acute myocardial infarction (I21)	33,299	27,720	10,579
Subsequent myocardial infarction (I22)	61	34	27
Certain current complications following acute myocardial infarction (I23)	138	92	46
Other acute ischemic heart diseases (I24)	1,341	814	527
Chronic ischemic heart disease (I25)	22,750	14,589	8,152
Common comorbidities (ICD-10)	**Total**	**Males**	**Females**
Primary (essential) hypertension (I10.9)	24,818	14,508	10,310
Hypercholesterolemia (E78.0)	12,780	7,842	4,938
Non-insulin dependent diabetes (E11.9)	7,551	4,891	2,660
Atrial fibrillation and atrial flutter, unspecified (I48.9)	7,075	4,509	2,566
Heart failure, unspecified (I50.9)	6,160	4,059	2,101
Chest pain, unspecified (R07.9)	5,863	3,441	2,422
Senile cataract, unspecified (H25.9)	5,764	2,795	2,969
Pneumonia, unspecified (J18.9)	5,469	3,236	2,260
Hyperlipidaemia, unspecified (E78.5)	5,002	3,306	1,696
Chronic obstructive pulmonary disease (J44.9)	4,621	2,449	2,172
Outcomes, number of cases	**Total**	**Males**	**Females**
New ischemic events (%)	14,679	10,152	4,527
■ *Myocardial infarction*	*5,833*	*3,709*	*2,124*
■ *Revascularization*	*6,282*	*4.718*	*2,124*
■ *Death caused by IHD*	*2,563*	*1.724*	*839*
Death from non-IHD causes (%)	10,684	6,710	3,974
Censored (%)	46,886	28,713	18,172
Outcomes, time to event	**Mean time to event in years (SD)**		
	Total	**Males**	**Females**
New ischemic events	1.48 (1.40)	1.49 (1.41)	1.48 (1.40)
■ Myocardial infarction	2.40 (1.87)	2.41 (1.89)	2.38 (1.85)
■ Revascularization	2.25 (1.88)	2.28 (1.89)	2.16 (1.84)
■ Death caused by IHD	1.92 (1.13)	1.95 (2.02)	1.88 (2.05)
Death from non-IHD causes	2.16 (1.50)	2.14 (1.49)	2.20 (1.51)
Censored	4.37 (1.08)	4.36 (1.09)	4.39 (1.06)
Total	3.72 (1.64)	3.67 (1.67)	3.81 (1.60)

618

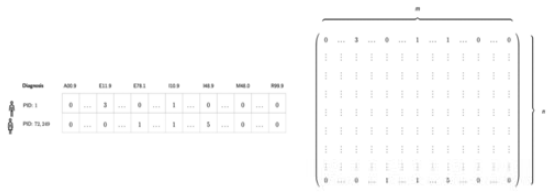
Table 2: Cluster demographics, characteristics, and associations with outcomes

Cluster	Size	Mean age at index in years (SD)	Males	Females	New ischemic events		Death from non-IHD causes	
					HR	Adj. P-val.	HR	Adj. P-val.
C1	7,191	64.8 (11.3)	3,897	3,294	1.000	> 0.050	0.856	> 0.050
C2	5,990	58.6 (11.5)	2,862	3,127	0.825	**< 0.001**	0.600	**< 0.001**
C3	4,641	56.8 (11.4)	2,727	1,914	0.757	**< 0.001**	0.586	**< 0.001**
C4	4,401	69.6 (10.2)	2,853	1,548	0.920	> 0.050	1.461	**< 0.001**
C5	4,290	63.9 (10.7)	2,803	1,487	1.402	**< 0.001**	1.629	**< 0.001**
C6	3,589	59.7 (10.9)	2.388	1,201	0.969	> 0.050	0.675	**< 0.001**
C7	3,309	63.8 (11.0)	2,025	1,284	0.889	> 0.050	0.611	**< 0.001**
C8	2,802	71.1 (10.9)	1,867	935	0.943	> 0.050	0.842	> 0.050
C9	2,581	63.7 (11.8)	1,803	778	1.314	**< 0.001**	1.789	> 0.050
C10	2,562	74.2 (9.6)	1,225	1,337	0.978	> 0.050	0.928	> 0.050
C11	2,292	66.1 (11.0)	2,186	106	0.926	> 0.050	0.650	**< 0.001**
C12	2,213	70.3 (10.2)	2,068	145	0.920	> 0.050	0.805	> 0.050
C13	2,070	58.6 (10.2)	1,348	722	0.946	> 0.050	0.577	**< 0.050**
C14	2,070	68.2 (9.6)	1,030	1,010	1.146	> 0.050	3.390	**< 0.001**
C15	2,040	63.9 (10.1)	1,208	805	1.031	> 0.050	0.784	> 0.050
C16	1,654	64.1 (12.1)	1,013	641	1.107	> 0.050	1.761	**< 0.001**
C17	1,281	65.3 (9.9)	714	567	1.001	> 0.050	1.761	**< 0.001**
C18	1,251	68.2 (9.8)	802	449	1.790	**< 0.001**	3.421	**< 0.001**
C19	1,168	58.5 (9.7)	995	173	0.752	> 0.050	1.571	> 0.050
C20	1,119	71.5 (11.3)	713	406	1.213	> 0.050	1.782	**< 0.001**
C21	1,000	61.0 (11.0)	769	231	1.116	> 0.050	0.890	> 0.050
C22	988	69.2 (10.4)	516	472	1.023	> 0.050	0.978	> 0.050
C23	935	58.7 (12.2)	588	347	1.609	**< 0.001**	2.275	**< 0.001**
C24	932	67.9 (10.1)	28	904	0.787	> 0.050	1.589	**< 0.001**
C25	860	56.2 (9.9)	664	196	0.978	> 0.050	2.691	**< 0.001**
C26	852	58.7 (12.1)	391	461	0.939	> 0.050	1.108	> 0.050
C27	823	65.1 (10.9)	532	291	1.201	> 0.050	1.289	> 0.050
C28	686	71.7 (8.0)	673	13	0.866	> 0.050	1.786	**< 0.001**
C29	550	57.2 (11.1)	435	115	0.906	> 0.050	0.985	> 0.050
C30	533	61.2 (11.7)	391	172	1.874	**< 0.001**	5.364	**< 0.001**
C31	520	64.4 (11.2)	213	307	1.052	> 0.050	1.484	> 0.050
NA*	5,113	60.1 (11.1)	3,878	1,235	NA	NA	NA	NA

*Patients that did not cluster or were in clusters of size < 500

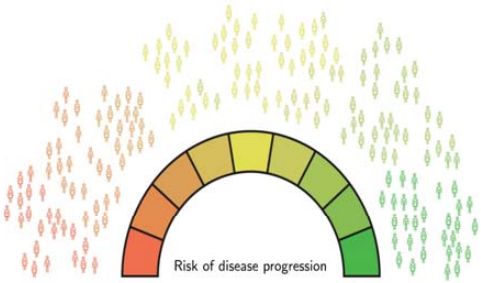
Fig 1

A. The assemblage of patient-specific vectors



B. Unsupervised clustering based on patient-specific vectors

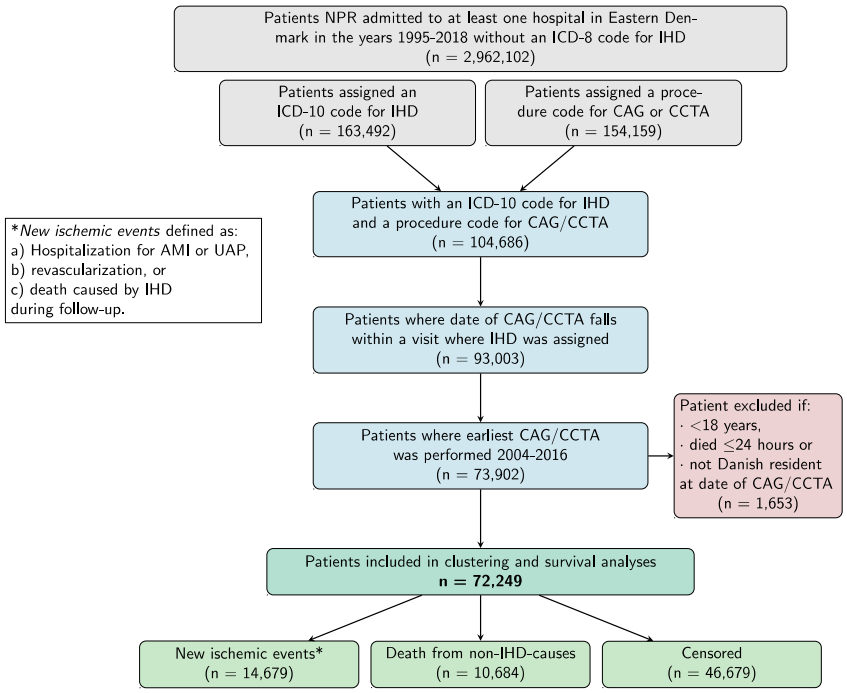
C. Association between clusters and clinical outcomes



D. Phenotypic and genetic characterization of clusters

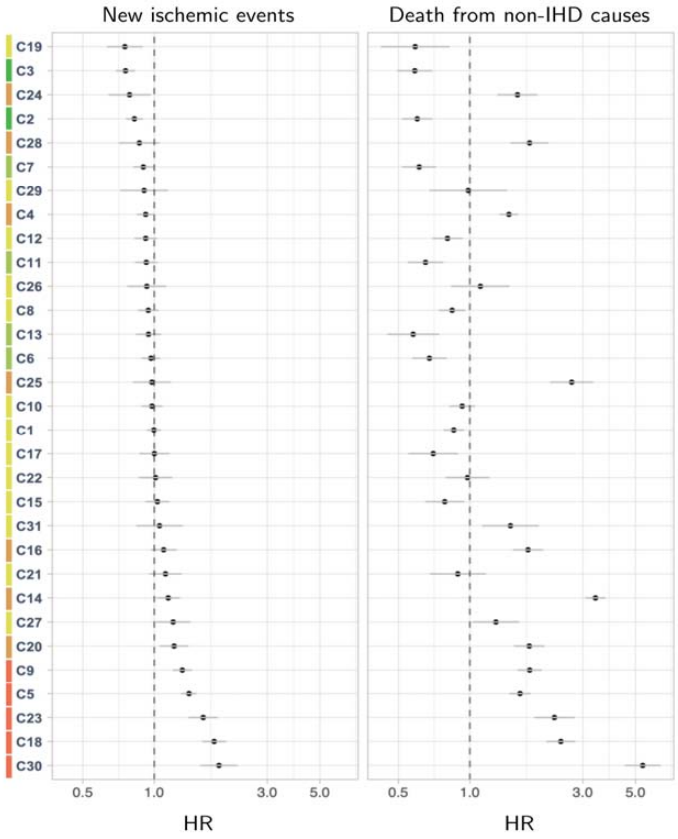
620

Fig 2


Patients NPR admitted to at least one hospital in Eastern Denmark in the years 1995-2018 without an ICD-8 code for IHD (n = 2,962,102)
Patients assigned an ICD-10 code for IHD (n = 163,492)
Patients assigned a procedure code for CAG or CCTA (n = 154,159)
Patients with an ICD-10 code for IHD and a procedure code for CAG/CCTA (n = 104,686)
*New ischemic events defined as:
a) Hospitalization for AMI or UAP,
b) revascularization, or
c) death caused by IHD
during follow-up.
Patients where date of CAG/CCTA falls within a visit where IHD was assigned (n = 93,003)
Patients where earliest CAG/CCTA was performed 2004-2016 (n = 73,902)
Patient excluded if:
· <18 years,
· died ≤24 hours or
· not Danish resident
at date of CAG/CCTA
(n = 1,653)
Patients included in clustering and survival analyses
n = 72,249
New ischemic events* (n = 14,679)
Death from non-IHD-causes (n = 10,684)
Censored (n = 46,679)

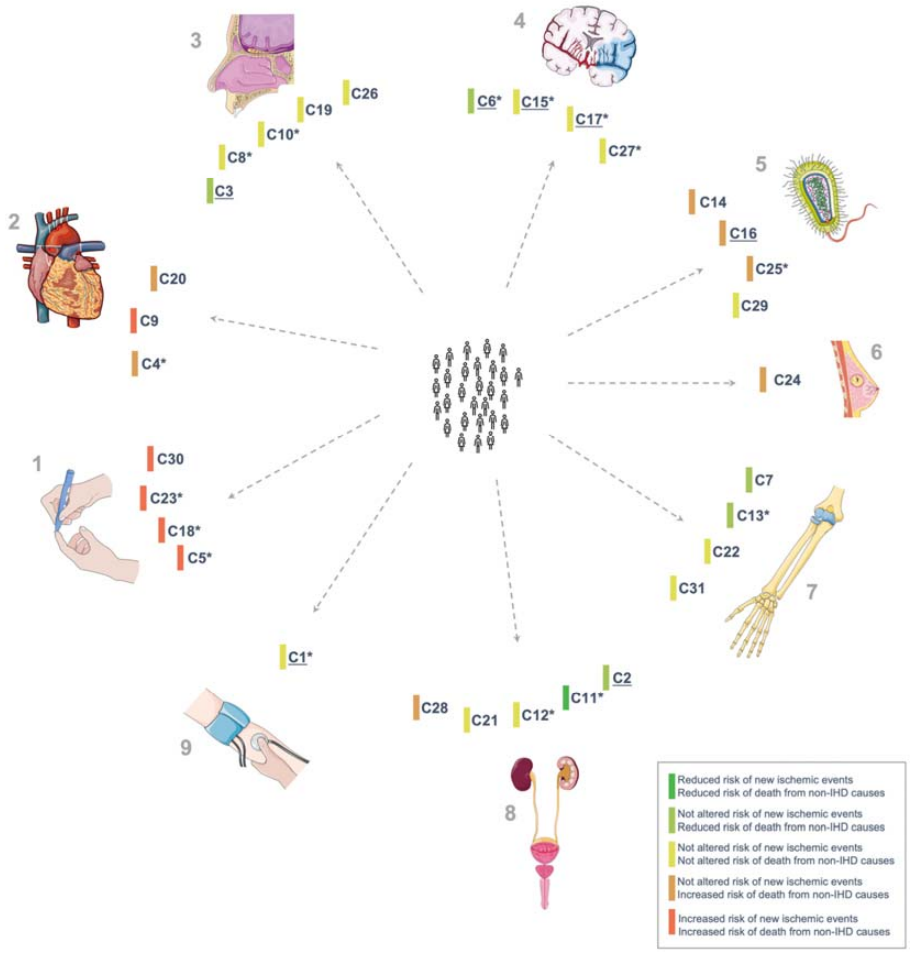
621

Fig 3


New ischemic events
Death from non-IHD causes
C19
C3
C24
C2
C28
C7
C29
C4
C12
C11
C26
C8
C13
C6
C25
C10
C1
C17
C22
C15
C31
C16
C21
C14
C27
C20
C9
C5
C23
C18
C30
0.5
1.0
3.0
5.0
HR
HR

622

Fig 4


3
C26
C19
C10*
C8*
C3
4
C6*
C15*
C17*
C27*
5
C14
C16
C25*
C29
2
C20
C9
C4*
6
C24
1
C30
C23*
C18*
C5*
7
C7
C13*
C22
C31
C1*
C2
C28
C21
C12*
C11*
9
8
Reduced risk of new ischemic events
Reduced risk of death from non-IHD causes
Not altered risk of new ischemic events
Reduced risk of death from non-IHD causes
Not altered risk of new ischemic events
Not altered risk of death from non-IHD causes
Not altered risk of new ischemic events
Increased risk of death from non-IHD causes
Increased risk of new ischemic events
Increased risk of death from non-IHD causes

623

Supplemental material

S1 Fig: Classification of new ischemic events.

S1 Table: Eligible codes for inclusion and outcomes

S1 Appendix: Construction of patient similarity network, MCL algorithm settings and assessment of cluster robustness

- **S2 Fig: Selection of number of components.**
- **S3 Fig: Limiting edge-density and average node degree in sex-specific similarity networks**.

S2 Appendix: Preprocessing of laboratory data

- **S2 Table: Laboratory codes included in assessment of data quality and completeness**

S3 Appendix: Calculation of polygenetic risk scores for 14 traits

S4 Fig: Results of robustness analysis.

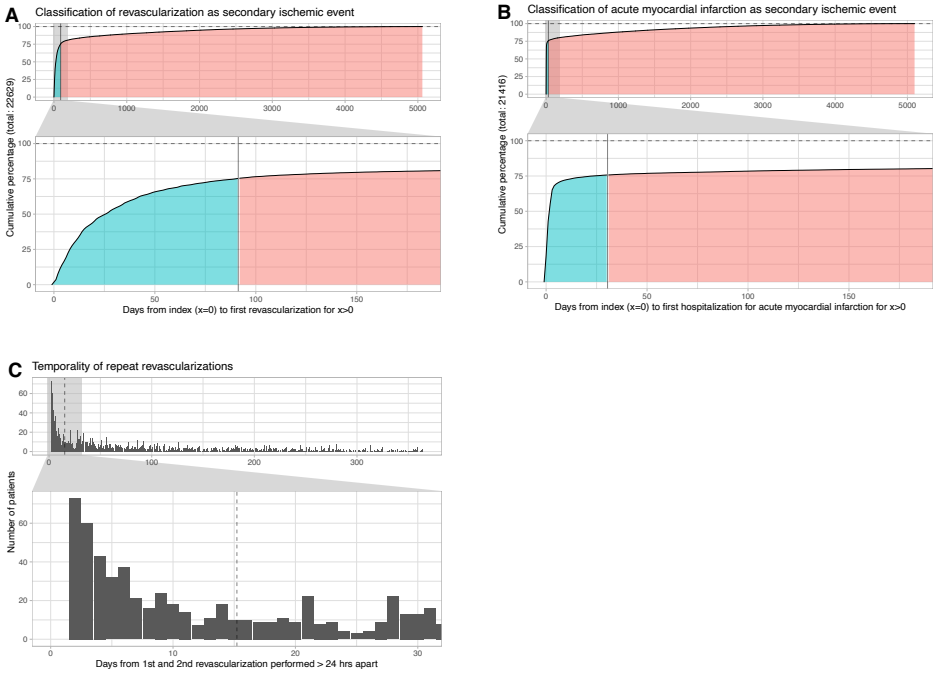
S3 Table: Comparison of mean age at index in 31 cluster using Tukey's HSD

S4 Table: Demographics for patients not cluster or were in clusters of size < 500

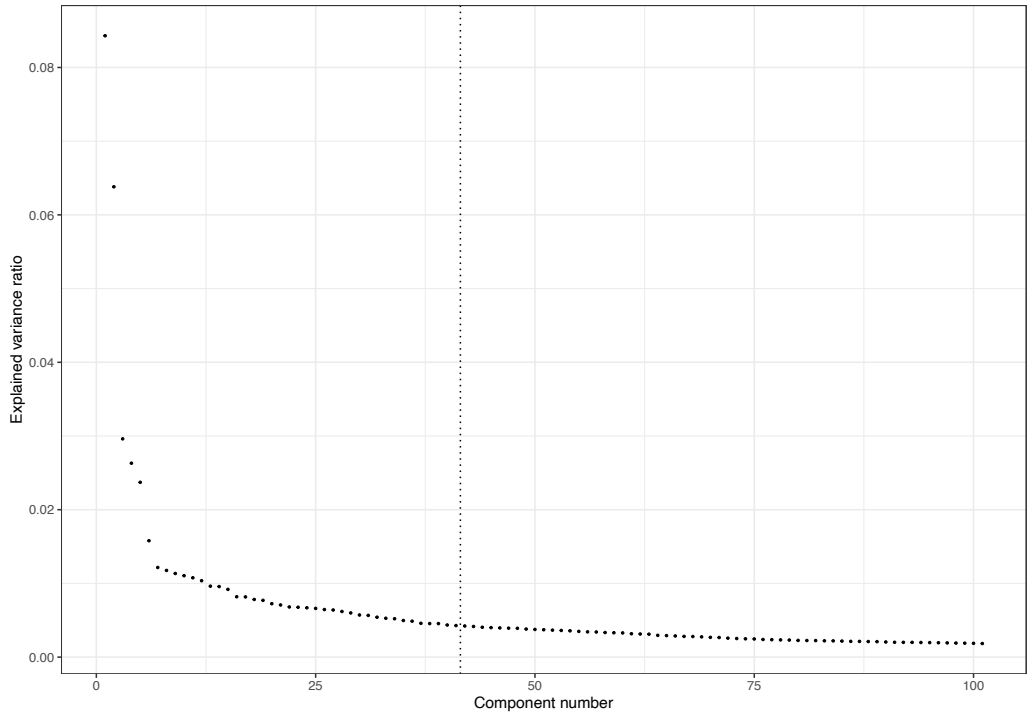
S5A-B Table: Cluster-wise summarized O/E-ratios, 10 largest O/E-ratios and 10 lowest O/E-ratios.

S6 Table: Chi-squared test for distribution laboratory values in clusters

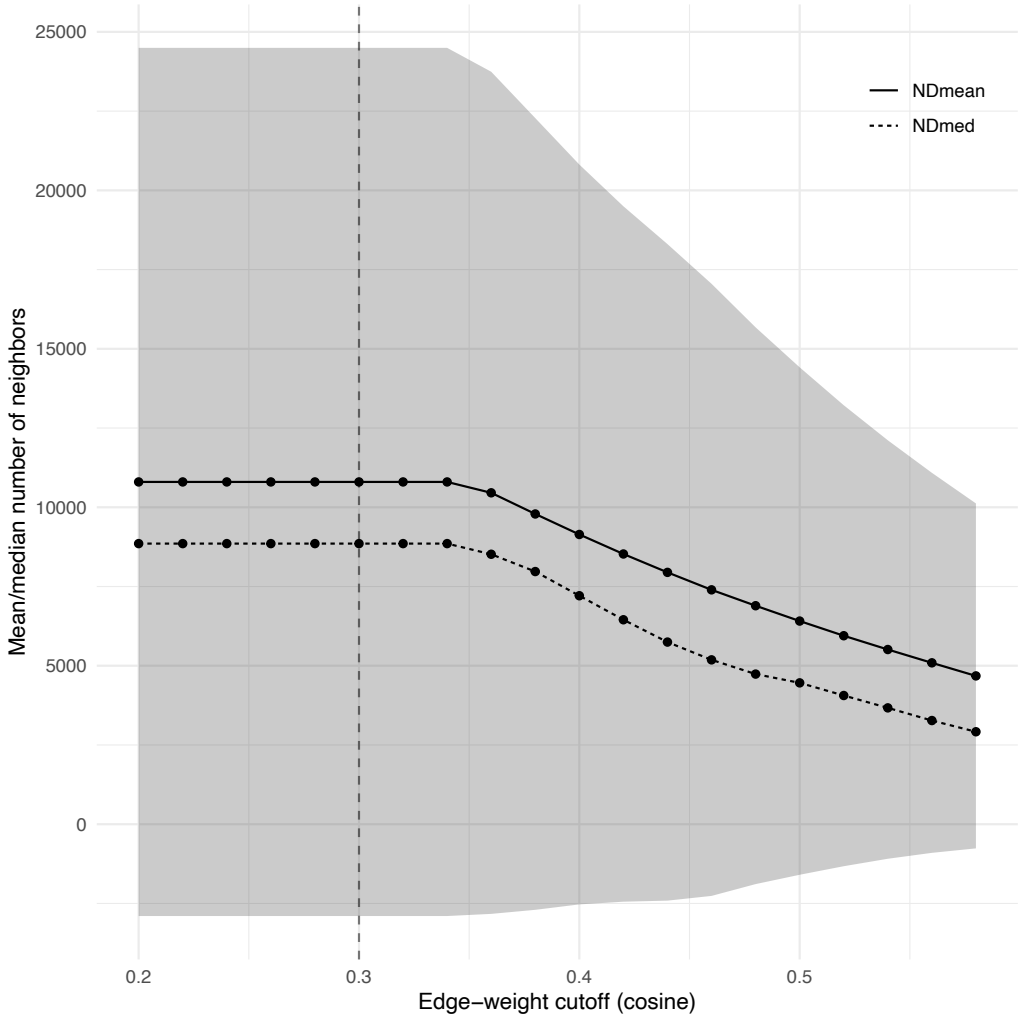
S7 Table: Traits with significantly different PGS distributions in clusters



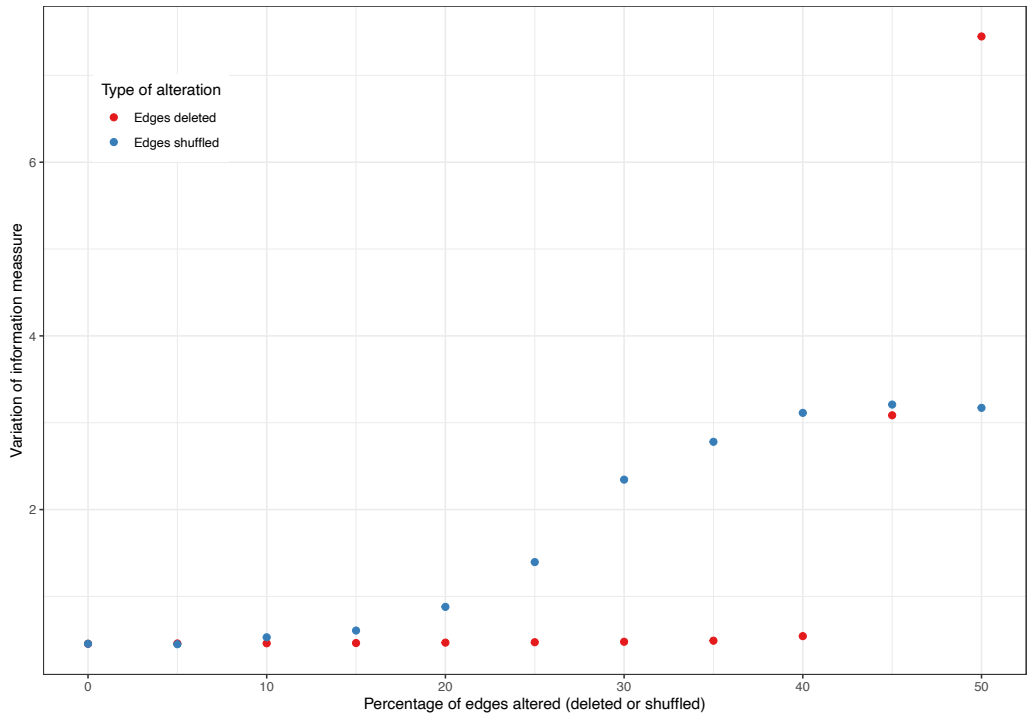
S1 Fig: Classification of new ischemic events. A: Time from index to first revascularization vs. percentage of patients revascularized. Blue corresponds to events related to establishment of IHD. Red corresponds to events considered new ischemic events. B: Time from index to first hospitalization for acute myocardial infarction vs. percentage of patients hospitalized. Blue corresponds to events related to index. Red corresponds to events considered new ischemic events. C: Distribution of days between revascularization for patients subjected to >1 performed >24 hours apart. Revascularizations performed <2 weeks apart were analyzed as a single event performed at date of the earliest revascularization. Marked by dashed line. IHD: Ischemic heart disease.



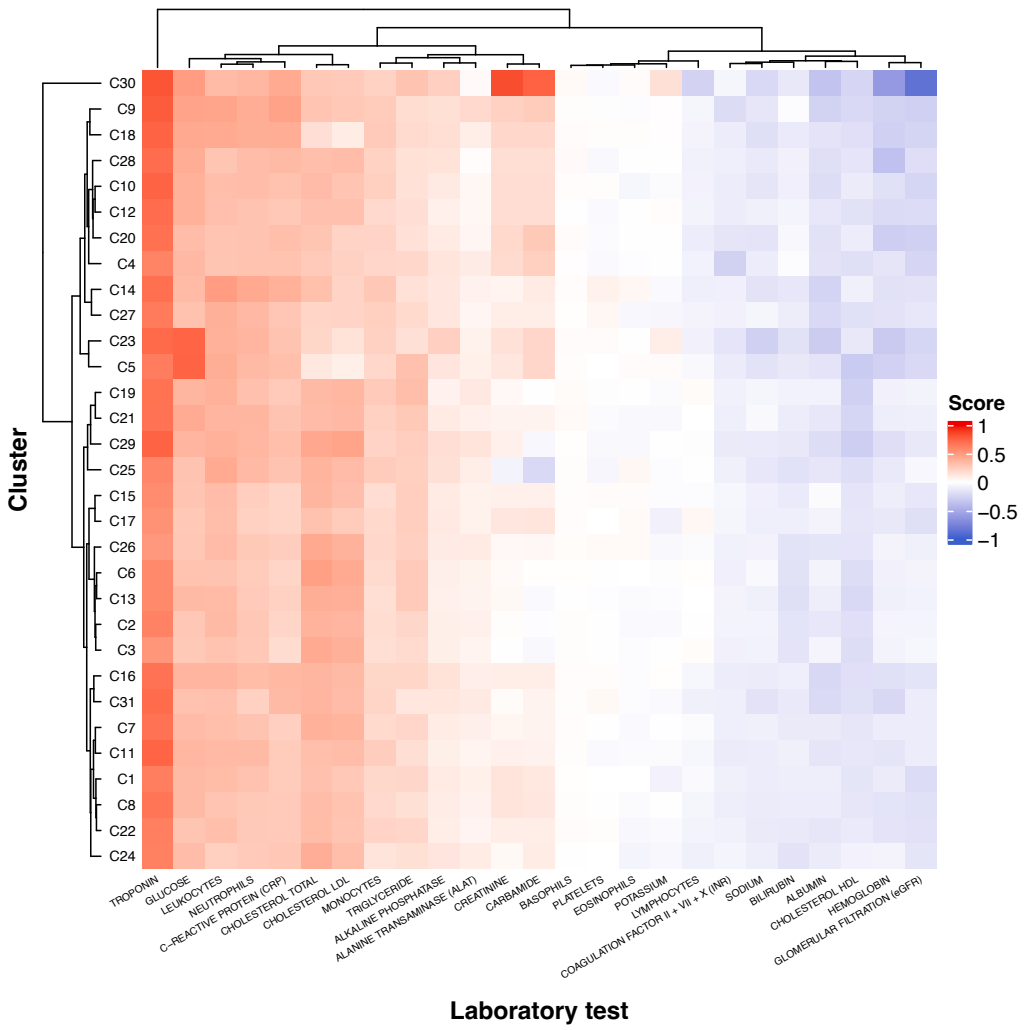
S2 Fig: Selection of number of components. X-axis: Component number ranked by explained variance ratio. Y-axis: Explained variance ratio. Dashed horizontal line indicates the cutoff.



S3 Fig: Limiting edge-density and average node degree in sex-specific similarity networks. Mean/median number of neighbors against edge-weight cutoff in the patient similarity network. Only edges with a weight higher than 0.3 (as indicated by the vertical, dashed bar) were retained.



S4 Fig: Results of robustness analysis. X-axis: Percentage of altered edges (deleted or removed). Y-axis: Variation of information measure compared to the reference graph. Legend: Type of alteration, with 10 mutations of the reference graph for each type.



S5 Fig: Heatmap of clusters based on laboratory profiles. Summary of results from the phenotypic characterization of clusters based on laboratory data. *Score* refers to the mean summarized values per cluster, where values were assigned based on the results of the laboratory test per patient. Values of -1, 0, and 1, indicates below, within or above reference range, respectively. For details, see Methods. X-axis: Laboratory test. Y-axis: Cluster.

S1 Table: Eligible codes for inclusion and outcomes

ICD-10[1] chapter IX			**Definition, level 3**
Block	**Level 3**	**Level 4**	
R94	I20	I20.0*, I20.1, I20.8, I20.9	Angina pectoris
	I21*	I21.0, I21.1, I21.2, I21.3, I21.4, I21.9	Acute myocardial infarction
	I23	I25.0, I25.1, I25.2, I25.23, I25.4, I25.5, I25.6, I25.8, I25.9	Certain current complications following acute myocardial infarction
	I24	I24.0, I24.1, I24.8, I24.9	Certain current complications following acute myocardial infarction
	I25	I25.0, I25.1, I25.2, I25.23, I25.4, I25.5, I25.6, I25.8, I25.9	Chronic ischemic heart disease
Nomesco[2] code		**Procedure**	
FNA*		Connection to coronary artery from internal mammary artery	
FNB*		Connection to coronary artery from gastroepiploic artery	
FNC*		Aorto-coronary venous bypass	
FND*		Aorto-coronary bypass using prosthetic graft	
FNE*		Coronary bypass using free arterial graft	
FNF*		Coronary thrombendarterectomy	
FNG*		Expansion and recanalisation of coronary artery	
SKS[3] code		**Procedure**	
UXAC85[A-D]		Coronary arteriography	
UXCC00A		Coronary computed tomography angiography	
SHAK[4] code		**Hospital**	
1301		Rigshospitalet	
1309		Bispebjerg og Frederiksberg Hospitaler	
1330		Amager og Hvidovre Hospital	
1351		Amager Hospital	
1401		Frederiksberg Hospital	
1501		Gentofte Hospital	
1502		Glostrup Hospital	
1516		Herlev og Gentofte Hospital	
2000		Hospitalerne i Nordsjælland	
2501		Amtssygehuset i Roskilde	
3800		**Region Sjællands Sygehusvæsen**	
4001		Bornholms Hospital	

[1] ICD-10 = WHO International classification of diseases and health related problems 10th edition. Danish version where code types A, B and G included in our definition of primary and secondary codes.

[2] NOMESCO = Nordic Medico-Statistical Committee

[3] SKS = Sundhedsstyrelses klassifikationsstystem [Danish]

[4] SHAK = Sygehus- og afdelingsklassifikation [Danish]

[*] Included in the composite outcome new ischemic events. For ICD-10 codes only code types A (primary) and in-hospital patients.

S2 Table: Laboratory codes included in assessment of data quality and completeness

Blood analyte	NPU codes and local systems
Sodium	NPU03429, GEN00992, NPU03796, POC00022, 240, POC00021, POC00023, GEN00990
Potassium	NPU03230, GEN00995, POC00019, POC00018, POC00020, GEN00993
Hemoglobin	NPU02319, GEN00989, NPU02321, NPU02320, NPU02322, NPU17007, POC00013, NPU04208, NPU01393, POC00012, POC00014, NPU29057, GEN00987
Creatinine / EGFR	NPU04998, NPU03918, NPU09102, NPU19661, NPU14048, NPU03800, HLL00037, DNK35131, POC00109, RHB00941, NPU28842

S3 Table: Comparison of mean age at index in 31 cluster using Tukey's HSD

contrast	null.value	estimate	conf.low	conf.high	adj.p.value
C2-C1	0	-6.18	-6.89	-5.47	0
C3-C1	0	-7.96	-8.73	-7.20	0
C4-C1	0	0,221527778	04.02	05.57	0
C5-C1	0	-0.920	-1.70	-0.139	3.60e- 3
C6-C1	0	-5.12	-5.95	-4.29	0
C7-C1	0	-0.992	-1.84	-0.141	4.40e- 3
C8-C1	0	06.29	05.39	07.20	0
C9-C1	0	-0.0638	-0.993	0,600694444	1 e+ 0
C10-C1	0	09.43	08.50	10.04	0
C11-C1	0	01.32	0,240972222	02.29	1.42e- 4
C12-C1	0	05.48	04.50	06.46	0
C13-C1	0	-6.13	-7.14	-5.12	0
C14-C1	0	03.47	02.45	04.48	0
C15-C1	0	-0.908	-1.93	0,078472222	1.79e- 1
C16-C1	0	-0.761	-1.86	0,238194444	7.29e- 1
C17-C1	0	0,355555556	-0.715	0,093055556	1.00e+ 0
C18-C1	0	03.39	02.15	0,210416667	3.19e-13
C19-C1	0	-6.27	-7.55	-4.99	0
C20-C1	0	0,301388889	05.44	08.04	0
C21-C1	0	-3.78	-5.15	-2.41	3.24e-13
C22-C1	0	04.38	03.01	0,261111111	0
C23-C1	0	-6.09	-7.50	-4.68	0
C24-C1	0	03.10	0,089583333	04.51	3.94e-13
C25-C1	0	-8.56	-10.0	-7.10	0
C26-C1	0	-5.98	-7.45	-4.52	0
C27-C1	0	0,239583333	-1.14	0,099305556	1.00e+ 0
C28-C1	0	0,313194444	05.29	08.53	0
C29-C1	0	-7.55	-9.34	-5.76	0
C30-C1	0	-3.55	-5.37	-1.73	9.14e-11
C31-C1	0	-0.335	-2.17	01.50	1.00e+ 0
C3-C2	0	-1.78	-2.58	-0.993	3.02e-13
C4-C2	0	11.00	10.02	11.08	0
C5-C2	0	05.26	04.45	06.07	0
C6-C2	0	01.06	0,140972222	0,104861111	1.32e- 3
C7-C2	0	05.19	04.31	06.06	0
C8-C2	0	12.05	11.05	13.04	0
C9-C2	0	06.11	05.16	07.07	0
C10-C2	0	15.06	14.07	16.06	0
C11-C2	0	07.50	06.50	08.49	0
C12-C2	0	11.07	10.07	12.07	0
C13-C2	0	0,356944444	-0.981	01.08	1 e+ 0
C14-C2	0	0,420138889	0,375694444	10.07	0
C15-C2	0	05.27	04.23	06.31	0
C16-C2	0	05.42	04.29	06.54	0
C17-C2	0	0,297916667	05.44	0,356944444	0
C18-C2	0	09.57	08.31	10.08	0
C19-C2	0	-0.0893	-1.38	01.21	1 e+ 0
C20-C2	0	12.09	11.06	14.02	0
C21-C2	0	02.40	01.02	0,179166667	3.07e- 8
C22-C2	0	10.06	09.17	12.00	0

contrast	null.value	estimate	conf.low	conf.high	adj.p.value
C23-C2	0	0,607638889	-1.34	01.51	1 e+ 0
C24-C2	0	09.28	0,351388889	10.07	0
C25-C2	0	-2.38	-3.86	-0.908	5.53e- 7
C26-C2	0	0,135416667	-1.29	0,088888889	1 e+ 0
C27-C2	0	06.52	05.02	08.03	0
C28-C2	0	13.01	11.05	14.07	0
C29-C2	0	-1.38	-3.18	0,297222222	5.03e- 1
C30-C2	0	0,127083333	0,555555556	04.46	2.82e- 5
C31-C2	0	0,266666667	0,19375	0,339583333	0
C4-C3	0	12.08	11.09	13.06	0
C5-C3	0	07.04	06.19	0,354166667	0
C6-C3	0	0,141666667	0,106944444	0,176388889	0
C7-C3	0	0,317361111	06.05	0,353472222	0
C8-C3	0	14.03	13.03	15.02	0
C9-C3	0	0,354166667	0,313194444	0,395138889	0
C10-C3	0	17.04	16.04	18.04	0
C11-C3	0	09.28	08.25	10.03	0
C12-C3	0	13.04	12.04	14.05	0
C13-C3	0	0,1	0,531944444	0,146527778	4.82e- 8
C14-C3	0	11.04	10.04	12.05	0
C15-C3	0	07.06	0,276388889	08.14	0
C16-C3	0	07.20	06.04	08.36	0
C17-C3	0	08.48	07.20	0,427083333	0
C18-C3	0	11.04	10.01	12.06	0
C19-C3	0	0,090277778	0,256944444	03.02	6.19e- 4
C20-C3	0	14.07	13.04	16.01	0
C21-C3	0	04.18	0,136805556	05.59	9.64e-14
C22-C3	0	12.03	10.09	13.08	0
C23-C3	0	0,102083333	0,292361111	03.32	5.07e- 4
C24-C3	0	11.01	0,417361111	12.05	0
C25-C3	0	-0.599	-2.10	0,627777778	1.00e+ 0
C26-C3	0	0,109722222	0,327083333	03.49	3.37e- 4
C27-C3	0	08.31	0,304166667	0,433333333	0
C28-C3	0	14.09	13.02	16.05	0
C29-C3	0	0,284722222	-1.42	02.24	1.00e+ 0
C30-C3	0	04.41	02.56	06.27	2.13e-13
C31-C3	0	0,335416667	0,261111111	09.50	0
C5-C4	0	-5.71	-6.58	-4.84	0
C6-C4	0	-9.91	-10.8	-9.00	0
C7-C4	0	-5.78	-6.71	-4.85	0
C8-C4	0	01.50	0,363888889	02.48	3.38e- 6
C9-C4	0	-4.85	-5.86	-3.85	0
C10-C4	0	0,211111111	0,169444444	0,253472222	0
C11-C4	0	-3.47	-4.52	-2.43	0
C12-C4	0	0,478472222	-0.366	0,093055556	8.21e- 1
C13-C4	0	-10.9	-12.0	-9.84	0
C14-C4	0	-1.32	-2.41	-0.239	1.76e- 3
C15-C4	0	-5.70	-6.79	-4.61	0
C16-C4	0	-5.55	-6.72	-4.38	0
C17-C4	0	-4.28	-5.56	-2.99	0
C18-C4	0	-1.40	-2.70	-0.102	1.70e- 2
C19-C4	0	-11.1	-12.4	-9.73	0

contrast	null.value	estimate	conf.low	conf.high	adj.p.value
C20-C4	0	0,107638889	0,411111111	03.30	2.86e- 5
C21-C4	0	-8.57	-9.99	-7.15	0
C22-C4	0	-0.409	-1.83	01.02	1.00e+ 0
C23-C4	0	-10.9	-12.3	-9.42	0
C24-C4	0	-1.69	-3.15	-0.228	5.18e- 3
C25-C4	0	-13.4	-14.9	-11.8	0
C26-C4	0	-10.8	-12.3	-9.26	0
C27-C4	0	-4.45	-5.98	-2.91	2.72e-13
C28-C4	0	02.12	0,317361111	0,179166667	6.71e- 4
C29-C4	0	-12.3	-14.2	-10.5	0
C30-C4	0	-8.34	-10.2	-6.48	0
C31-C4	0	-5.13	-7.00	-3.25	3.35e-13
C6-C5	0	-4.20	-5.12	-3.29	0
C7-C5	0	-0.0714	-1.01	0,600694444	1 e+ 0
C8-C5	0	07.21	06.23	08.20	0
C9-C5	0	0,594444444	-0.152	0,101388889	2.61e- 1
C10-C5	0	10.04	09.34	11.04	0
C11-C5	0	02.24	01.19	03.29	8.06e-13
C12-C5	0	06.40	05.34	07.46	0
C13-C5	0	-5.21	-6.29	-4.12	0
C14-C5	0	04.39	03.30	05.48	0
C15-C5	0	0,0875	-1.08	01.11	1 e+ 0
C16-C5	0	0,111111111	-1.01	01.33	1 e+ 0
C17-C5	0	01.43	0,099305556	0,133333333	1.04e- 2
C18-C5	0	04.31	03.01	0,250694444	0
C19-C5	0	-5.35	-6.68	-4.01	0
C20-C5	0	0,3375	06.30	09.02	0
C21-C5	0	-2.86	-4.28	-1.44	1.70e-11
C22-C5	0	05.30	0,185416667	0,300694444	0
C23-C5	0	-5.17	-6.63	-3.71	0
C24-C5	0	04.02	02.56	05.49	2.89e-13
C25-C5	0	-7.64	-9.16	-6.13	0
C26-C5	0	-5.06	-6.58	-3.55	0
C27-C5	0	01.27	-0.275	0,139583333	3.30e- 1
C28-C5	0	0,349305556	06.17	09.49	0
C29-C5	0	-6.63	-8.47	-4.80	0
C30-C5	0	-2.63	-4.49	-0.769	4.58e- 5
C31-C5	0	0,40625	-1.29	02.47	1.00e+ 0
C7-C6	0	04.13	03.15	05.11	0
C8-C6	0	11.04	10.04	12.04	0
C9-C6	0	05.06	04.01	06.10	0
C10-C6	0	14.06	13.05	15.06	0
C11-C6	0	06.44	05.36	07.52	0
C12-C6	0	10.06	09.51	11.07	0
C13-C6	0	-1.01	-2.12	0,077083333	1.59e- 1
C14-C6	0	08.59	07.47	0,424305556	0
C15-C6	0	04.21	03.09	05.34	0
C16-C6	0	04.36	03.16	05.56	0
C17-C6	0	0,252083333	04.32	0,315972222	0
C18-C6	0	08.51	07.18	0,433333333	0
C19-C6	0	-1.15	-2.51	0,150694444	2.81e- 1
C20-C6	0	11.09	10.05	13.02	0

contrast	null.value	estimate	conf.low	conf.high	adj.p.value
C21-C6	0	01.34	-0.107	0,138194444	1.21e- 1
C22-C6	0	09.50	08.05	11.00	0
C23-C6	0	-0.970	-2.46	0,359027778	8.23e- 1
C24-C6	0	08.22	0,301388889	0,424305556	0
C25-C6	0	-3.44	-4.98	-1.90	3.25e-13
C26-C6	0	-0.862	-2.41	0,472222222	9.65e- 1
C27-C6	0	05.47	0,1875	07.03	0
C28-C6	0	12.00	10.03	13.07	0
C29-C6	0	-2.43	-4.29	-0.579	3.36e- 4
C30-C6	0	01.57	-0.307	03.45	2.91e- 1
C31-C6	0	0,221527778	0,145138889	0,297916667	2.61e-13
C8-C7	0	07.29	06.25	08.32	0
C9-C7	0	0,644444444	-0.135	0,110416667	2.10e- 1
C10-C7	0	10.04	09.36	11.05	0
C11-C7	0	02.31	01.21	03.41	1.70e-12
C12-C7	0	06.47	05.36	07.58	0
C13-C7	0	-5.14	-6.27	-4.00	0
C14-C7	0	04.46	03.32	0,25	0
C15-C7	0	0,583333333	-1.06	01.23	1 e+ 0
C16-C7	0	0,160416667	-0.988	01.45	1.00e+ 0
C17-C7	0	01.50	0,119444444	0,141666667	7.96e- 3
C18-C7	0	04.38	03.04	0,259027778	0
C19-C7	0	-5.28	-6.65	-3.90	0
C20-C7	0	0,342361111	06.33	09.13	0
C21-C7	0	-2.79	-4.25	-1.33	3.09e-10
C22-C7	0	05.37	0,188194444	0,308333333	0
C23-C7	0	-5.10	-6.60	-3.60	0
C24-C7	0	04.10	02.59	0,25	3.44e-13
C25-C7	0	-7.57	-9.12	-6.02	0
C26-C7	0	-4.99	-6.55	-3.44	0
C27-C7	0	01.34	-0.240	0,146527778	2.64e- 1
C28-C7	0	0,354166667	06.20	0,416666667	0
C29-C7	0	-6.56	-8.43	-4.70	0
C30-C7	0	-2.56	-4.45	-0.668	1.53e- 4
C31-C7	0	0,45625	-1.25	02.57	1.00e+ 0
C9-C8	0	-6.36	-7.46	-5.25	0
C10-C8	0	03.14	02.03	04.25	3.03e-13
C11-C8	0	-4.98	-6.12	-3.84	0
C12-C8	0	-0.813	-1.96	0,234722222	6.78e- 1
C13-C8	0	-12.4	-13.6	-11.2	0
C14-C8	0	-2.83	-4.00	-1.65	2.02e-13
C15-C8	0	-7.20	-8.38	-6.02	0
C16-C8	0	-7.05	-8.31	-5.80	0
C17-C8	0	-5.78	-7.15	-4.42	0
C18-C8	0	-2.90	-4.28	-1.52	1.41e-12
C19-C8	0	-12.6	-14.0	-11.2	0
C20-C8	0	0,309027778	-0.987	0,102777778	1.00e+ 0
C21-C8	0	-10.1	-11.6	-8.58	0
C22-C8	0	-1.91	-3.41	-0.414	6.53e- 4
C23-C8	0	-12.4	-13.9	-10.9	0
C24-C8	0	-3.19	-4.72	-1.66	2.46e-12
C25-C8	0	-14.9	-16.4	-13.3	0

contrast	null.value	estimate	conf.low	conf.high	adj.p.value
C26-C8	0	-12.3	-13.9	-10.7	0
C27-C8	0	-5.95	-7.55	-4.34	0
C28-C8	0	0,427777778	-1.11	02.34	1.00e+ 0
C29-C8	0	-13.8	-15.7	-12.0	0
C30-C8	0	-9.84	-11.8	-7.93	0
C31-C8	0	-6.63	-8.56	-4.70	0
C10-C9	0	09.50	08.37	10.06	0
C11-C9	0	01.38	0,152777778	02.54	3.01e- 3
C12-C9	0	05.54	04.37	0,3	0
C13-C9	0	-6.06	-7.26	-4.87	0
C14-C9	0	03.53	02.33	0,217361111	1.45e-13
C15-C9	0	-0.844	-2.05	0,25	6.94e- 1
C16-C9	0	-0.697	-1.97	0,401388889	9.74e- 1
C17-C9	0	0,4	-0.808	0,108333333	1.00e+ 0
C18-C9	0	03.46	02.06	0,225694444	3.13e-13
C19-C9	0	-6.20	-7.63	-4.78	0
C20-C9	0	0,305555556	05.35	08.25	0
C21-C9	0	-3.72	-5.22	-2.21	3.27e-13
C22-C9	0	04.45	0,147916667	0,275	1.72e-13
C23-C9	0	-6.03	-7.57	-4.48	0
C24-C9	0	03.17	0,084722222	0,215972222	6.14e-12
C25-C9	0	-8.50	-10.1	-6.91	0
C26-C9	0	-5.92	-7.52	-4.32	0
C27-C9	0	0,284027778	-1.21	02.03	1.00e+ 0
C28-C9	0	0,317361111	05.23	0,382638889	0
C29-C9	0	-7.49	-9.39	-5.59	0
C30-C9	0	-3.48	-5.41	-1.56	4.48e- 9
C31-C9	0	-0.271	-2.22	0,088888889	1 e+ 0
C11-C10	0	-8.12	-9.28	-6.95	0
C12-C10	0	-3.95	-5.13	-2.78	0
C13-C10	0	-15.6	-16.8	-14.4	0
C14-C10	0	-5.97	-7.17	-4.76	0
C15-C10	0	-10.3	-11.5	-9.13	0
C16-C10	0	-10.2	-11.5	-8.92	0
C17-C10	0	-8.92	-10.3	-7.54	0
C18-C10	0	-6.04	-7.44	-4.64	0
C19-C10	0	-15.7	-17.1	-14.3	0
C20-C10	0	-2.69	-4.15	-1.24	1.25e- 9
C21-C10	0	-13.2	-14.7	-11.7	0
C22-C10	0	-5.05	-6.57	-3.54	0
C23-C10	0	-15.5	-17.1	-14.0	0
C24-C10	0	-6.33	-7.88	-4.78	0
C25-C10	0	-18.0	-19.6	-16.4	0
C26-C10	0	-15.4	-17.0	-13.8	0
C27-C10	0	-9.09	-10.7	-7.47	0
C28-C10	0	-2.52	-4.26	-0.783	2.17e- 5
C29-C10	0	-17.0	-18.9	-15.1	0
C30-C10	0	-13.0	-14.9	-11.1	0
C31-C10	0	-9.77	-11.7	-7.82	0
C12-C11	0	04.16	0,15	05.37	0
C13-C11	0	-7.45	-8.67	-6.22	0
C14-C11	0	02.15	0,636805556	03.38	2.38e- 8

contrast	null.value	estimate	conf.low	conf.high	adj.p.value
C15-C11	0	-2.23	-3.46	-0.989	5.77e- 9
C16-C11	0	-2.08	-3.38	-0.772	9.60e- 7
C17-C11	0	-0.806	-2.22	0,421527778	9.55e- 1
C18-C11	0	02.07	0,452083333	03.50	1.84e- 5
C19-C11	0	-7.59	-9.04	-6.13	0
C20-C11	0	05.42	0,190277778	0,3125	0
C21-C11	0	-5.10	-6.63	-3.56	0
C22-C11	0	03.06	01.52	0,208333333	3.32e-11
C23-C11	0	-7.41	-8.98	-5.84	0
C24-C11	0	0,096527778	0,147916667	03.36	7.19e- 3
C25-C11	0	-9.88	-11.5	-8.26	0
C26-C11	0	-7.30	-8.93	-5.68	0
C27-C11	0	-0.973	-2.62	0,466666667	9.33e- 1
C28-C11	0	05.59	0,182638889	07.35	0
C29-C11	0	-8.87	-10.8	-6.95	0
C30-C11	0	-4.87	-6.81	-2.92	2.88e-13
C31-C11	0	-1.65	-3.62	0,218055556	2.82e- 1
C13-C12	0	-11.6	-12.8	-10.4	0
C14-C12	0	-2.01	-3.26	-0.770	4.92e- 7
C15-C12	0	-6.39	-7.63	-5.14	0
C16-C12	0	-6.24	-7.56	-4.93	0
C17-C12	0	-4.97	-6.39	-3.55	0
C18-C12	0	-2.09	-3.52	-0.656	1.82e- 5
C19-C12	0	-11.7	-13.2	-10.3	0
C20-C12	0	01.26	-0.227	0,134722222	2.66e- 1
C21-C12	0	-9.26	-10.8	-7.72	0
C22-C12	0	-1.10	-2.65	0,3125	6.69e- 1
C23-C12	0	-11.6	-13.2	-9.99	0
C24-C12	0	-2.38	-3.96	-0.796	6.89e- 6
C25-C12	0	-14.0	-15.7	-12.4	0
C26-C12	0	-11.5	-13.1	-9.83	0
C27-C12	0	-5.14	-6.79	-3.48	0
C28-C12	0	01.43	-0.340	03.20	3.67e- 1
C29-C12	0	-13.0	-15.0	-11.1	0
C30-C12	0	-9.03	-11.0	-7.08	0
C31-C12	0	-5.82	-7.79	-3.84	1.31e-13
C14-C13	0	09.59	08.33	10.09	0
C15-C13	0	05.22	0,190972222	06.49	0
C16-C13	0	05.37	04.03	0,298611111	0
C17-C13	0	0,294444444	05.20	08.08	0
C18-C13	0	09.52	08.07	11.00	0
C19-C13	0	-0.141	-1.62	01.34	1 e+ 0
C20-C13	0	12.09	11.04	14.04	0
C21-C13	0	02.35	0,547222222	0,188194444	6.61e- 6
C22-C13	0	10.05	0,398611111	12.01	0
C23-C13	0	0,250694444	-1.56	0,085416667	1 e+ 0
C24-C13	0	09.23	0,335416667	10.08	0
C25-C13	0	-2.44	-4.08	-0.793	1.07e- 5
C26-C13	0	0,1	-1.50	0,096527778	1 e+ 0
C27-C13	0	06.47	0,222222222	08.14	0
C28-C13	0	13.00	11.03	14.08	0
C29-C13	0	-1.43	-3.37	0,357638889	5.91e- 1

contrast	null.value	estimate	conf.low	conf.high	adj.p.value
C30-C13	0	02.58	0,425	04.54	3.42e- 4
C31-C13	0	0,263194444	0,18125	0,345833333	2.55e-13
C15-C14	0	-4.37	-5.65	-3.10	0
C16-C14	0	-4.23	-5.57	-2.89	0
C17-C14	0	-2.96	-4.40	-1.51	6.09e-12
C18-C14	0	-0.0752	-1.53	01.38	1 e+ 0
C19-C14	0	-9.74	-11.2	-8.25	0
C20-C14	0	03.27	0,094444444	0,220833333	5.00e-13
C21-C14	0	-7.25	-8.81	-5.68	0
C22-C14	0	0,634722222	-0.655	02.48	9.43e- 1
C23-C14	0	-9.56	-11.2	-7.96	0
C24-C14	0	-0.364	-1.96	01.24	1.00e+ 0
C25-C14	0	-12.0	-13.7	-10.4	0
C26-C14	0	-9.45	-11.1	-7.80	0
C27-C14	0	-3.12	-4.79	-1.45	9.54e-10
C28-C14	0	03.44	0,0875	05.23	1.89e-10
C29-C14	0	-11.0	-13.0	-9.08	0
C30-C14	0	-7.02	-8.99	-5.05	0
C31-C14	0	-3.80	-5.79	-1.81	2.86e-10
C16-C15	0	0,102083333	-1.20	01.49	1 e+ 0
C17-C15	0	01.42	-0.0271	0,14375	6.33e- 2
C18-C15	0	04.30	0,141666667	0,261111111	1.24e-13
C19-C15	0	-5.36	-6.85	-3.87	0
C20-C15	0	0,336805556	06.14	09.16	0
C21-C15	0	-2.87	-4.44	-1.31	2.33e- 9
C22-C15	0	05.29	0,175	0,309722222	0
C23-C15	0	-5.18	-6.79	-3.58	0
C24-C15	0	04.01	02.41	0,251388889	2.86e-13
C25-C15	0	-7.66	-9.30	-6.01	0
C26-C15	0	-5.08	-6.73	-3.42	0
C27-C15	0	01.25	-0.422	0,147916667	5.49e- 1
C28-C15	0	0,348611111	06.03	0,417361111	0
C29-C15	0	-6.65	-8.59	-4.70	0
C30-C15	0	-2.64	-4.61	-0.669	2.01e- 4
C31-C15	0	0,397916667	-1.42	02.56	1.00e+ 0
C17-C16	0	01.27	-0.234	0,1375	2.71e- 1
C18-C16	0	04.15	0,127777778	0,254861111	3.14e-13
C19-C16	0	-5.51	-7.05	-3.96	0
C20-C16	0	07.50	0,272916667	09.07	0
C21-C16	0	-3.02	-4.64	-1.40	1.10e- 9
C22-C16	0	05.14	03.51	0,303472222	0
C23-C16	0	-5.33	-6.99	-3.67	0
C24-C16	0	0,184722222	02.21	05.52	2.42e-13
C25-C16	0	-7.80	-9.50	-6.10	0
C26-C16	0	-5.22	-6.93	-3.52	0
C27-C16	0	01.11	-0.621	0,140972222	8.50e- 1
C28-C16	0	0,338194444	0,265972222	09.51	0
C29-C16	0	-6.79	-8.79	-4.80	0
C30-C16	0	-2.79	-4.80	-0.772	8.57e- 5
C31-C16	0	0,295833333	-1.61	02.46	1.00e+ 0
C18-C17	0	0,144444444	01.27	04.49	7.47e- 9
C19-C17	0	-6.78	-8.42	-5.14	0

contrast	null.value	estimate	conf.low	conf.high	adj.p.value
C20-C17	0	06.23	04.57	0,352777778	0
C21-C17	0	-4.29	-6.00	-2.58	2.71e-13
C22-C17	0	0,185416667	02.16	05.58	2.99e-13
C23-C17	0	-6.60	-8.34	-4.86	0
C24-C17	0	02.59	0,588888889	04.33	9.90e- 6
C25-C17	0	-9.08	-10.9	-7.29	0
C26-C17	0	-6.50	-8.29	-4.71	0
C27-C17	0	-0.167	-1.98	0,086111111	1 e+ 0
C28-C17	0	06.40	04.48	08.31	0
C29-C17	0	-8.07	-10.1	-6.00	0
C30-C17	0	-4.06	-6.15	-1.97	1.10e-10
C31-C17	0	-0.847	-2.95	01.26	1.00e+ 0
C19-C18	0	-9.66	-11.3	-8.01	0
C20-C18	0	03.35	0,088888889	05.01	1.86e-11
C21-C18	0	-7.17	-8.89	-5.46	0
C22-C18	0	0,686805556	-0.734	0,132638889	9.52e- 1
C23-C18	0	-9.48	-11.2	-7.73	0
C24-C18	0	-0.289	-2.04	01.46	1 e+ 0
C25-C18	0	-12.0	-13.7	-10.2	0
C26-C18	0	-9.38	-11.2	-7.58	0
C27-C18	0	-3.05	-4.86	-1.23	1.26e- 7
C28-C18	0	03.52	01.59	05.44	2.68e- 9
C29-C18	0	-10.9	-13.0	-8.87	0
C30-C18	0	-6.94	-9.03	-4.85	0
C31-C18	0	-3.73	-5.84	-1.61	1.43e- 8
C20-C19	0	13.00	11.03	14.07	0
C21-C19	0	02.49	0,516666667	04.23	3.54e- 5
C22-C19	0	10.06	0,395833333	12.04	0
C23-C19	0	0,122916667	-1.60	0,107638889	1 e+ 0
C24-C19	0	09.37	07.59	11.01	0
C25-C19	0	-2.29	-4.11	-0.476	8.54e- 4
C26-C19	0	0,197222222	-1.54	02.11	1 e+ 0
C27-C19	0	0,292361111	0,220138889	08.46	0
C28-C19	0	13.02	11.02	15.01	0
C29-C19	0	-1.29	-3.38	0,561111111	8.99e- 1
C30-C19	0	0,133333333	0,41875	0,225	5.58e- 4
C31-C19	0	0,272916667	0,180555556	08.07	2.91e-13
C21-C20	0	-10.5	-12.3	-8.76	0
C22-C20	0	-2.36	-4.12	-0.589	2.25e- 4
C23-C20	0	-12.8	-14.6	-11.0	0
C24-C20	0	-3.63	-5.43	-1.84	1.18e-11
C25-C20	0	-15.3	-17.1	-13.5	0
C26-C20	0	-12.7	-14.6	-10.9	0
C27-C20	0	-6.39	-8.25	-4.53	0
C28-C20	0	0,11875	-1.79	02.13	1 e+ 0
C29-C20	0	-14.3	-16.4	-12.2	0
C30-C20	0	-10.3	-12.4	-8.16	0
C31-C20	0	-7.07	-9.22	-4.92	0
C22-C21	0	08.16	06.35	0,443055556	0
C23-C21	0	-2.31	-4.15	-0.469	9.66e- 4
C24-C21	0	0,311111111	05.04	0,384027778	0
C25-C21	0	-4.78	-6.67	-2.90	2.43e-13

contrast	null.value	estimate	conf.low	conf.high	adj.p.value
C26-C21	0	-2.20	-4.09	-0.316	4.30e- 3
C27-C21	0	04.13	02.22	06.03	5.42e-13
C28-C21	0	10.07	0,380555556	12.07	0
C29-C21	0	-3.77	-5.92	-1.62	1.78e- 8
C30-C21	0	0,161111111	-1.94	02.40	1 e+ 0
C31-C21	0	03.45	01.26	0,252083333	1.42e- 6
C23-C22	0	-10.5	-12.3	-8.63	0
C24-C22	0	-1.28	-3.13	0,396527778	7.22e- 1
C25-C22	0	-12.9	-14.8	-11.1	0
C26-C22	0	-10.4	-12.3	-8.47	0
C27-C22	0	-4.04	-5.95	-2.13	1.26e-12
C28-C22	0	02.53	0,358333333	04.54	9.36e- 4
C29-C22	0	-11.9	-14.1	-9.78	0
C30-C22	0	-7.93	-10.1	-5.75	0
C31-C22	0	-4.72	-6.91	-2.52	6.49e-13
C24-C23	0	09.19	07.32	11.01	0
C25-C23	0	-2.47	-4.38	-0.559	4.88e- 4
C26-C23	0	0,074305556	-1.81	02.02	1 e+ 0
C27-C23	0	06.44	04.50	08.37	0
C28-C23	0	13.00	11.00	15.00	0
C29-C23	0	-1.46	-3.64	0,495138889	7.74e- 1
C30-C23	0	02.54	0,239583333	0,218055556	5.09e- 3
C31-C23	0	0,261111111	03.54	0,359027778	3.28e-13
C25-C24	0	-11.7	-13.6	-9.75	0
C26-C24	0	-9.09	-11.0	-7.17	0
C27-C24	0	-2.76	-4.69	-0.822	3.68e- 5
C28-C24	0	0,18125	0,095138889	0,266666667	9.20e-10
C29-C24	0	-10.7	-12.8	-8.48	0
C30-C24	0	-6.65	-8.85	-4.45	0
C31-C24	0	-3.44	-5.65	-1.22	2.37e- 6
C26-C25	0	02.58	0,431944444	04.54	3.02e- 4
C27-C25	0	0,396527778	0,314583333	10.09	0
C28-C25	0	15.05	13.04	17.05	0
C29-C25	0	01.01	-1.20	03.22	9.98e- 1
C30-C25	0	05.01	0,1375	07.25	3.14e-13
C31-C25	0	08.23	0,276388889	10.05	0
C27-C26	0	06.33	04.35	08.31	0
C28-C26	0	12.09	10.08	15.00	0
C29-C26	0	-1.57	-3.78	0,447222222	6.70e- 1
C30-C26	0	02.44	0,138194444	0,213194444	1.45e- 2
C31-C26	0	0,253472222	03.40	0,354166667	2.79e-13
C28-C27	0	06.56	04.47	0,379166667	0
C29-C27	0	-7.90	-10.1	-5.67	0
C30-C27	0	-3.89	-6.14	-1.64	3.41e- 8
C31-C27	0	-0.680	-2.95	01.59	1.00e+ 0
C29-C28	0	-14.5	-16.8	-12.1	0
C30-C28	0	-10.5	-12.8	-8.12	0
C31-C28	0	-7.24	-9.60	-4.89	0
C30-C29	0	04.01	01.54	06.47	4.11e- 7
C31-C29	0	07.22	0,218055556	0,423611111	2.62e-13
C31-C30	0	03.21	0,499305556	0,257638889	5.28e- 4
M-F	0	-2.75	-2.91	-2.58	0

S4 Table: Demographics for patients that did not cluster or were in clusters of size < 500

Cohort demographics	**Total**	**Males**	**Females**
Number of patients	5,113	3,878	1,235
Mean age at index (SD)	60.7	60.0	63.0
Outcomes, number of cases	**Total**	**Males**	**Females**
New ischemic events	995	780	175
Death from non-IHD causes	352	274	78
Censored	3,624	2,707	917
Outcomes, time to event	**Mean time to event in years (SD)**		
	Total	**Males**	**Females**
New ischemic events	1.55 (1.41)	1.59 (1.43)	1.39 (1.32)
Death from non-IHD causes	2.25 (1.50)	2.18 (1.47)	2.5 (1.49)
Censored	4.54 (0.95)	4.52 (0.96)	2.47 (1.49)
Total	4.02 (1.52)	3.98 (1.54)	4.17 (1.44)

S5A Table: Degree of enrichment (sum) and top-10 O/E-ratios per cluster

Clst	Sum	ICD-10	# code	# clst	Obs	Exp	O/E-ratio	description
C1	345	I109	24818	7191	1.000	0.294	3.40	Essential (primary) hypertension
C1	345	H350	548	7191	0.016	0.007	2.24	Background retinopathy and retinal vascular changes
C1	345	E871	393	7191	0.012	0.005	2.23	Hypo-osmolality and hyponatraemia
C1	345	I159	414	7191	0.012	0.005	2.15	Secondary hypertension, unspecified
C1	345	I959	283	7191	0.008	0.004	2.15	Hypotension, unspecified
C1	345	E789	417	7191	0.012	0.006	2.10	Disorder of lipoprotein metabolism, unspecified
C1	345	E785	5002	7191	0.133	0.067	1.98	Hyperlipidaemia, unspecified
C1	345	I119	534	7191	0.013	0.007	1.85	Hypertensive heart disease without (congestive) heart failure
C1	345	D251	256	7191	0.006	0.004	1.83	Intramural leiomyoma of uterus
C1	345	N811	974	7191	0.024	0.013	1.77	Cystocele
C2	372	K802	2849	5990	0.176	0.029	6.01	Calculus of gallbladder without cholecystitis
C2	372	K801	411	5990	0.023	0.004	5.22	Calculus of gallbladder with other cholecystitis
C2	372	K805	1027	5990	0.054	0.011	4.70	Calculus of bile duct without cholangitis or cholecystitis
C2	372	K800	468	5990	0.022	0.006	3.88	Calculus of gallbladder with acute cholecystitis
C2	372	R100	2884	5990	0.123	0.035	3.51	Acute abdomen
C2	372	R108	3700	5990	0.158	0.045	3.51	Abdominal and pelvic pain
C2	372	R103	488	5990	0.020	0.006	3.29	Pain localized to other parts of lower abdomen
C2	372	R102	566	5990	0.023	0.007	3.29	Pelvic and perineal pain
C2	372	N832	489	5990	0.020	0.006	3.28	Other and unspecified ovarian cysts
C2	372	D251	256	5990	0.010	0.003	3.19	Intramural leiomyoma of uterus
C3	268	R079	5863	4641	0.363	0.067	5.43	Pain in throat and chest
C3	268	G409	841	4641	0.043	0.010	4.23	Epilepsy, unspecified
C3	268	I309	297	4641	0.014	0.004	3.70	Acute pericarditis, unspecified
C3	268	M626	4440	4641	0.187	0.057	3.28	Muscle strain
C3	268	G430	351	4641	0.013	0.005	2.83	Migraine without aura [common migraine]
C3	268	R073	2009	4641	0.072	0.027	2.67	Other chest pain
C3	268	R002	691	4641	0.025	0.009	2.66	Palpitations
C3	268	R519	1667	4641	0.058	0.022	2.61	Headache
C3	268	R064	542	4641	0.019	0.007	2.54	Hyperventilation
C3	268	R072	367	4641	0.012	0.005	2.48	Precordial pain
C4	520	I489	7075	4401	0.995	0.043	23.14	Atrial fibrillation and atrial flutter, unspecified
C4	520	I495	482	4401	0.055	0.004	14.14	Sick sinus syndrome
C4	520	I480	364	4401	0.039	0.003	12.49	Paroxysmal atrial fibrillation
C4	520	I471	1920	4401	0.183	0.018	10.27	Supraventricular tachycardia
C4	520	I499	436	4401	0.035	0.005	7.71	Cardiac arrhythmia, unspecified
C4	520	I479	614	4401	0.045	0.007	6.73	Paroxysmal tachycardia, unspecified
C4	520	R001	394	4401	0.026	0.004	5.80	Bradycardia, unspecified
C4	520	R000	391	4401	0.025	0.004	5.51	Tachycardia, unspecified
C4	520	I340	971	4401	0.051	0.012	4.32	Mitral (valve) insufficiency
C4	520	I491	273	4401	0.013	0.003	3.76	Atrial premature depolarization
C5	596	E119	7551	4290	0.947	0.056	17.06	Type 2 diabetes mellitus: Without complications
C5	596	E113	720	4290	0.081	0.006	13.55	Type 2 diabetes mellitus: With ophthalmic complications
C5	596	E114	881	4290	0.088	0.008	11.06	Type 2 diabetes mellitus: With neurological complications
C5	596	E149	958	4290	0.095	0.009	10.82	Unspecified diabetes mellitus: Without complications

Clst	Sum	ICD-10	# code	# clst	Obs	Exp	O/E-ratio	description
C5	596	E117	980	4290	0.096	0.009	10.58	Type 2 diabetes mellitus: With multiple complications
C5	596	E118	2669	4290	0.251	0.025	9.93	Type 2 diabetes mellitus: With unspecified complications
C5	596	H360	1467	4290	0.137	0.014	9.80	Diabetic retinopathy
C5	596	E112	761	4290	0.070	0.007	9.53	Type 2 diabetes mellitus: With renal complications
C5	596	E148	483	4290	0.044	0.005	9.50	Unspecified diabetes mellitus: With unspecified complications
C5	596	E115	551	4290	0.048	0.005	8.88	Type 2 diabetes mellitus: With peripheral circulatory complications
C6	239	E780	12780	3589	0.867	0.152	5.70	Pure hypercholesterolaemia
C6	239	E785	5002	3589	0.177	0.069	2.57	Hyperlipidaemia, unspecified
C6	239	I999	383	3589	0.010	0.005	1.89	Other and unspecified disorders of circulatory system
C6	239	I639	1989	3589	0.044	0.029	1.53	Cerebral infarction, unspecified
C6	239	I652	563	3589	0.011	0.008	1.39	Occlusion and stenosis of carotid artery
C6	239	G459	2066	3589	0.042	0.030	1.39	Transient cerebral ischaemic attack, unspecified
C6	239	R670	750	3589	0.014	0.011	1.24	NA
C6	239	E113	720	3589	0.013	0.011	1.21	Type 2 diabetes mellitus: With ophthalmic complications
C6	239	M100	691	3589	0.012	0.010	1.20	Idiopathic gout
C6	239	N434	294	3589	0.005	0.004	1.16	Spermatocele
C7	374	M171	2940	3309	0.359	0.027	13.06	Other primary gonarthrosis
C7	374	M179	2242	3309	0.257	0.022	11.80	Gonarthrosis, unspecified
C7	374	M170	2145	3309	0.232	0.022	10.74	Primary gonarthrosis, bilateral
C7	374	M234	258	3309	0.027	0.003	10.33	Loose body in knee
C7	374	M235	269	3309	0.028	0.003	10.19	Chronic instability of knee
C7	374	M232	2404	3309	0.238	0.025	9.42	Derangement of meniscus due to old tear or injury
C7	374	M238	532	3309	0.042	0.006	6.89	Other internal derangements of knee
C7	374	M239	1105	3309	0.081	0.013	6.21	Internal derangement of knee, unspecified
C7	374	M712	363	3309	0.025	0.004	5.63	Synovial cyst of popliteal space [Baker]
C7	374	M169	1539	3309	0.088	0.020	4.48	Coxarthrosis, unspecified
C8	384	H919	4610	2802	0.664	0.043	15.54	Hearing loss, unspecified
C8	384	H911	3527	2802	0.495	0.033	14.90	Presbycusis
C8	384	H905	1160	2802	0.151	0.011	13.13	Sensorineural hearing loss, unspecified
C8	384	H833	1412	2802	0.180	0.014	12.70	Noise effects on inner ear
C8	384	H838	254	2802	0.032	0.003	12.38	Other specified diseases of inner ear
C8	384	H938	1116	2802	0.132	0.012	11.34	Other specified disorders of ear
C8	384	H908	631	2802	0.071	0.007	10.50	Mixed conductive and sensorineural hearing loss, unspecified
C8	384	H931	1228	2802	0.123	0.014	9.01	Tinnitus
C8	384	H810	300	2802	0.029	0.003	8.64	Ménière disease
C8	384	H809	374	2802	0.034	0.004	7.82	Otosclerosis, unspecified
C9	323	I420	706	2581	0.095	0.007	13.21	Dilated cardiomyopathy
C9	323	I509	6160	2581	0.783	0.064	12.21	Heart failure, unspecified
C9	323	I429	479	2581	0.056	0.005	10.75	Cardiomyopathy, unspecified
C9	323	I501	1502	2581	0.124	0.018	6.74	Left ventricular failure
C9	323	I500	2327	2581	0.188	0.029	6.57	Congestive heart failure
C9	323	R570	320	2581	0.025	0.004	6.38	Cardiogenic shock
C9	323	I460	1028	2581	0.075	0.013	5.78	Cardiac arrest with successful resuscitation
C9	323	I472	752	2581	0.053	0.010	5.52	Ventricular tachycardia

Clst	Sum	ICD-10	# code	# clst	Obs	Exp	O/E-ratio	description
C9	323	I110	273	2581	0.018	0.004	5.20	Hypertensive heart disease with (congestive) heart failure
C9	323	I340	971	2581	0.059	0.013	4.68	Mitral (valve) insufficiency
C10	361	H259	5764	2562	0.866	0.055	15.77	Senile cataract, unspecified
C10	361	H264	1015	2562	0.113	0.011	10.03	After-cataract
C10	361	H330	418	2562	0.045	0.005	9.68	Retinal detachment with retinal break
C10	361	H353	1743	2562	0.174	0.020	8.67	Degeneration of macula and posterior pole
C10	361	H401	388	2562	0.032	0.005	6.86	Primary open-angle glaucoma
C10	361	H438	291	2562	0.022	0.004	6.14	Other disorders of vitreous body
C10	361	H333	264	2562	0.019	0.003	5.74	Retinal breaks without detachment
C10	361	H521	366	2562	0.023	0.005	4.84	Myopia
C10	361	H348	337	2562	0.017	0.005	3.79	Other retinal vascular occlusions
C10	361	H260	486	2562	0.023	0.007	3.55	Infantile, juvenile and presenile cataract
C11	328	K409	3787	2292	0.984	0.024	41.64	Unilateral or unspecified inguinal hernia, without obstruction or gangrene
C11	328	K402	280	2292	0.042	0.003	14.76	Bilateral inguinal hernia, without obstruction or gangrene
C11	328	N433	367	2292	0.023	0.005	4.67	Hydrocele, unspecified
C11	328	N434	294	2292	0.010	0.004	2.40	Spermatocele
C11	328	K429	896	2292	0.030	0.013	2.32	Umbilical hernia without obstruction or gangrene
C11	328	N484	456	2292	0.011	0.007	1.71	Impotence of organic origin
C11	328	N508	300	2292	0.007	0.004	1.70	Other specified disorders of male genital organs
C11	328	M720	1004	2292	0.024	0.015	1.67	Palmar fascial fibromatosis [Dupuytren]
C11	328	D179	257	2292	0.006	0.004	1.63	Benign lipomatous neoplasm, unspecified
C11	328	I714	517	2292	0.012	0.008	1.56	Abdominal aortic aneurysm, without mention of rupture
C12	383	N409	3319	2213	0.701	0.027	25.77	Hyperplasia of prostate
C12	383	R339	1530	2213	0.239	0.015	15.55	Retention of urine
C12	383	R391	2230	2213	0.241	0.026	9.24	Other difficulties with micturition
C12	383	N359	253	2213	0.027	0.003	9.12	Urethral stricture, unspecified
C12	383	R319	3787	2213	0.323	0.047	6.82	Unspecified haematuria
C12	383	N459	612	2213	0.052	0.008	6.72	Orchitis, epididymitis and epididymo-orchitis without abscess
C12	383	M720	1004	2213	0.059	0.013	4.36	Palmar fascial fibromatosis [Dupuytren]
C12	383	N434	294	2213	0.015	0.004	3.71	Spermatocele
C12	383	N309	834	2213	0.042	0.011	3.64	Cystitis, unspecified
C12	383	N319	265	2213	0.013	0.004	3.60	Neuromuscular dysfunction of bladder, unspecified
C13	522	M511	3357	2070	0.937	0.022	42.98	Lumbar and other intervertebral disc disorders with radiculopathy
C13	522	M519	604	2070	0.120	0.005	22.05	Intervertebral disc disorder, unspecified
C13	522	M512	366	2070	0.071	0.003	21.34	Other specified intervertebral disc displacement
C13	522	M544	1209	2070	0.144	0.014	10.28	Lumbago with sciatica
C13	522	M543	491	2070	0.057	0.006	9.94	Sciatica
C13	522	M513	1585	2070	0.178	0.019	9.54	Other specified intervertebral disc degeneration
C13	522	M539	274	2070	0.026	0.003	7.54	Dorsopathy, unspecified
C13	522	M472	897	2070	0.082	0.011	7.35	Other spondylosis with radiculopathy
C13	522	M501	910	2070	0.080	0.011	7.01	Cervical disc disorder with radiculopathy
C13	522	M549	1419	2070	0.103	0.019	5.58	Dorsalgia, unspecified
C14	676	J440	743	2040	0.191	0.005	35.25	Chronic obstructive pulmonary disease with acute lower respiratory infection

Clst	Sum	ICD-10	# code	# clst	Obs	Exp	O/E-ratio	description
C14	676	J441	1678	2040	0.423	0.013	33.71	Chronic obstructive pulmonary disease with acute exacerbation, unspecified
C14	676	J449	4621	2040	0.948	0.041	22.97	Chronic obstructive pulmonary disease, unspecified
C14	676	J439	317	2040	0.064	0.003	22.18	Emphysema, unspecified
C14	676	J448	396	2040	0.075	0.004	20.31	Other specified chronic obstructive pulmonary disease
C14	676	J429	714	2040	0.105	0.008	13.66	Unspecified chronic bronchitis
C14	676	J960	837	2040	0.120	0.009	13.13	Acute respiratory failure
C14	676	J969	681	2040	0.072	0.008	8.78	Respiratory failure, unspecified
C14	676	J159	1222	2040	0.103	0.016	6.62	Bacterial pneumonia, unspecified
C14	676	J209	527	2040	0.041	0.007	6.05	Acute bronchitis, unspecified
C15	296	E780	12780	2013	1.000	0.165	6.05	Pure hypercholesterolaemia
C15	296	I109	24818	2013	1.000	0.350	2.86	Essential (primary) hypertension
C15	296	D629	333	2013	0.009	0.005	1.96	Acute posthaemorrhagic anaemia
C15	296	I639	1989	2013	0.055	0.029	1.91	Cerebral infarction, unspecified
C15	296	I693	467	2013	0.013	0.007	1.91	Sequelae of cerebral infarction
C15	296	R670	750	2013	0.020	0.011	1.87	NA
C15	296	E114	881	2013	0.021	0.013	1.66	Type 2 diabetes mellitus: With neurological complications
C15	296	G459	2066	2013	0.049	0.030	1.61	Transient cerebral ischaemic attack, unspecified
C15	296	E118	2669	2013	0.063	0.039	1.60	Type 2 diabetes mellitus: With unspecified complications
C15	296	R072	367	2013	0.008	0.005	1.57	Precordial pain
C16	365	J189	5496	1654	0.743	0.065	11.40	Pneumonia, unspecified
C16	365	J849	265	1654	0.027	0.003	8.10	Interstitial pulmonary disease, unspecified
C16	365	C349	284	1654	0.026	0.004	7.06	Malignant neoplasm: Bronchus or lung, unspecified
C16	365	R919	1471	1654	0.103	0.020	5.17	Abnormal findings on diagnostic imaging of lung
C16	365	J909	291	1654	0.019	0.004	4.72	Pleural effusion, not elsewhere classified
C16	365	J181	298	1654	0.018	0.004	4.43	Lobar pneumonia, unspecified
C16	365	J159	1222	1654	0.073	0.017	4.35	Bacterial pneumonia, unspecified
C16	365	R091	418	1654	0.023	0.006	3.96	Pleurisy
C16	365	J180	257	1654	0.013	0.004	3.71	Bronchopneumonia, unspecified
C16	365	J969	681	1654	0.031	0.010	3.27	Respiratory failure, unspecified
C17	375	E780	12780	1281	1.000	0.175	5.73	Pure hypercholesterolaemia
C17	375	E789	417	1281	0.018	0.006	3.00	Disorder of lipoprotein metabolism, unspecified
C17	375	H350	548	1281	0.023	0.008	2.98	Background retinopathy and retinal vascular changes
C17	375	G459	2066	1281	0.087	0.030	2.95	Transient cerebral ischaemic attack, unspecified
C17	375	I639	1989	1281	0.084	0.029	2.92	Cerebral infarction, unspecified
C17	375	I109	24818	1281	1.000	0.357	2.80	Essential (primary) hypertension
C17	375	I652	563	1281	0.022	0.008	2.69	Occlusion and stenosis of carotid artery
C17	375	I693	467	1281	0.017	0.007	2.54	Sequelae of cerebral infarction
C17	375	M109	547	1281	0.019	0.008	2.36	Gout, unspecified
C17	375	I694	1733	1281	0.059	0.025	2.36	Sequelae of stroke, not specified as haemorrhage or infarction
C18	535	I702	2251	1251	0.812	0.019	43.33	Atherosclerosis of arteries of extremities
C18	535	I739	2027	1251	0.544	0.020	26.65	Peripheral vascular disease, unspecified
C18	535	I709	433	1251	0.081	0.005	16.02	Generalized and unspecified atherosclerosis
C18	535	L979	605	1251	0.091	0.007	12.23	Ulcer of lower limb, not elsewhere classified
C18	535	E105	416	1251	0.060	0.005	11.58	Type 1 diabetes mellitus: With peripheral circulatory complications
C18	535	E115	551	1251	0.066	0.007	9.34	Type 2 diabetes mellitus: With peripheral circulatory complications

Clst	Sum	ICD-10	# code	# clst	Obs	Exp	O/E-ratio	description
C18	535	E148	483	1251	0.042	0.007	6.35	Unspecified diabetes mellitus: With unspecified complications
C18	535	I714	517	1251	0.043	0.007	6.14	Abdominal aortic aneurysm, without mention of rupture
C18	535	L984	299	1251	0.021	0.004	5.02	Chronic ulcer of skin, not elsewhere classified
C18	535	I999	383	1251	0.023	0.005	4.31	Other and unspecified disorders of circulatory system
C19	374	G473	1897	1168	0.712	0.016	44.12	Sleep apnoea
C19	374	R065	983	1168	0.271	0.010	26.88	Mouth breathing
C19	374	J342	1097	1168	0.218	0.013	17.10	Deviated nasal septum
C19	374	J330	351	1168	0.038	0.005	8.10	Polyp of nasal cavity
C19	374	J320	513	1168	0.033	0.007	4.65	Chronic maxillary sinusitis
C19	374	J350	315	1168	0.015	0.005	3.22	Chronic tonsillitis
C19	374	N508	300	1168	0.014	0.004	3.18	Other specified disorders of male genital organs
C19	374	E669	2114	1168	0.086	0.031	2.80	Obesity, unspecified
C19	374	M766	339	1168	0.013	0.005	2.62	Achilles tendinitis
C19	374	G510	365	1168	0.013	0.005	2.42	Bell's palsy
C20	337	I350	2664	1119	0.856	0.026	33.13	Aortic (valve) stenosis
C20	337	I351	696	1119	0.090	0.009	10.02	Aortic (valve) insufficiency
C20	337	K053	360	1119	0.023	0.005	4.59	Chronic periodontitis
C20	337	K045	333	1119	0.021	0.005	4.58	Chronic apical periodontitis
C20	337	R040	1768	1119	0.100	0.025	3.99	Epistaxis
C20	337	D649	1637	1119	0.084	0.023	3.59	Anaemia, unspecified
C20	337	I340	971	1119	0.045	0.014	3.20	Mitral (valve) insufficiency
C20	337	M353	627	1119	0.028	0.009	3.07	Polymyalgia rheumatica
C20	337	D509	459	1119	0.020	0.007	2.97	Iron deficiency anaemia, unspecified
C20	337	K921	287	1119	0.012	0.004	2.80	Melaena
C21	532	N200	1391	1000	0.765	0.009	80.82	Calculus of kidney
C21	532	N201	1381	1000	0.605	0.012	51.56	Calculus of ureter
C21	532	N209	520	1000	0.202	0.005	42.01	Urinary calculus, unspecified
C21	532	N133	267	1000	0.045	0.003	13.41	Other and unspecified hydronephrosis
C21	532	N109	555	1000	0.047	0.008	6.12	Acute tubulo-interstitial nephritis
C21	532	R319	3787	1000	0.197	0.054	3.63	Unspecified haematuria
C21	532	N359	253	1000	0.009	0.004	2.44	Urethral stricture, unspecified
C21	532	N308	384	1000	0.013	0.006	2.32	Other cystitis
C21	532	R100	2884	1000	0.096	0.042	2.28	Acute abdomen
C21	532	A419	968	1000	0.032	0.014	2.26	Sepsis, unspecified
C22	592	M480	2424	988	0.932	0.023	41.03	Spinal stenosis
C22	592	M431	583	988	0.135	0.007	19.79	Spondylolisthesis
C22	592	M472	897	988	0.159	0.011	14.20	Other spondylosis with radiculopathy
C22	592	M513	1585	988	0.200	0.021	9.56	Other specified intervertebral disc degeneration
C22	592	M539	274	988	0.031	0.004	8.54	Dorsopathy, unspecified
C22	592	M478	630	988	0.072	0.008	8.50	Other spondylosis
C22	592	M479	531	988	0.059	0.007	8.21	Spondylosis, unspecified
C22	592	M543	491	988	0.046	0.007	6.76	Sciatica
C22	592	M503	520	988	0.039	0.007	5.43	Other cervical disc degeneration
C22	592	R522	1177	988	0.088	0.016	5.34	Other chronic pain
C23	968	E103	576	935	0.339	0.004	86.66	Type 1 diabetes mellitus: With ophthalmic complications
C23	968	E107	583	935	0.303	0.005	66.79	Type 1 diabetes mellitus: With multiple complications

Clst	Sum	ICD-10	# code	# clst	Obs	Exp	O/E-ratio	description
C23	968	E104	434	935	0.214	0.004	60.52	Type 1 diabetes mellitus: With neurological complications
C23	968	E102	377	935	0.165	0.003	48.90	Type 1 diabetes mellitus: With renal complications
C23	968	E162	510	935	0.199	0.005	40.65	Hypoglycaemia, unspecified
C23	968	E108	1073	935	0.363	0.011	32.70	Type 1 diabetes mellitus: With unspecified complications
C23	968	E109	2680	935	0.887	0.028	31.71	Type 1 diabetes mellitus: Without complications
C23	968	E105	416	935	0.137	0.004	31.47	Type 1 diabetes mellitus: With peripheral circulatory complications
C23	968	H360	1467	935	0.398	0.017	24.05	Diabetic retinopathy
C23	968	H431	334	935	0.061	0.004	14.57	Vitreous haemorrhage
C24	559	C509	1103	932	0.835	0.005	170.04	Malignant neoplasm: Breast, unspecified
C24	559	N639	773	932	0.153	0.010	16.12	Unspecified lump in breast
C24	559	D249	758	932	0.097	0.010	9.57	Benign neoplasm of breast
C24	559	N602	712	932	0.076	0.010	7.87	Fibroadenosis of breast
C24	559	N950	699	932	0.059	0.010	6.07	Postmenopausal bleeding
C24	559	L905	295	932	0.020	0.004	4.89	Scar conditions and fibrosis of skin
C24	559	M819	1292	932	0.083	0.018	4.50	Osteoporosis, unspecified
C24	559	N629	571	932	0.036	0.008	4.50	Hypertrophy of breast
C24	559	N840	546	932	0.030	0.008	3.84	Polyp of corpus uteri
C24	559	E052	322	932	0.014	0.005	2.99	Thyrotoxicosis with toxic multinodular goitre
C25	681	F103	518	860	0.331	0.004	94.26	Mental and behavioural disorders due to use of alcohol: Withdrawal state
C25	681	F102	1189	860	0.641	0.010	66.56	Mental and behavioural disorders due to use of alcohol: Dependence syndrome
C25	681	F100	1212	860	0.535	0.011	47.14	Mental and behavioural disorders due to use of alcohol: Acute intoxication
C25	681	F101	1115	860	0.430	0.011	38.27	Mental and behavioural disorders due to use of alcohol: Harmful use
C25	681	F172	349	860	0.090	0.004	21.82	Mental and behavioural disorders due to use of tobacco: Dependence syndrome
C25	681	F339	299	860	0.053	0.004	14.01	Recurrent depressive disorder, unspecified
C25	681	R568	477	860	0.067	0.006	10.67	Other and unspecified convulsions
C25	681	K920	428	860	0.057	0.006	9.96	Haematemesis
C25	681	F329	816	860	0.093	0.011	8.38	Depressive episode, unspecified
C25	681	F419	335	860	0.036	0.005	7.86	Anxiety disorder, unspecified
C26	540	J459	2487	852	0.927	0.026	36.22	Asthma, unspecified
C26	540	J451	417	852	0.133	0.005	28.92	Nonallergic asthma
C26	540	J450	401	852	0.127	0.004	28.68	Predominantly allergic asthma
C26	540	J448	396	852	0.040	0.005	7.31	Other specified chronic obstructive pulmonary disease
C26	540	J209	527	852	0.046	0.007	6.22	Acute bronchitis, unspecified
C26	540	J330	351	852	0.029	0.005	5.97	Polyp of nasal cavity
C26	540	J370	343	852	0.028	0.005	5.85	Chronic laryngitis
C26	540	R490	415	852	0.023	0.006	3.94	Dysphonia
C26	540	J429	714	852	0.039	0.010	3.77	Unspecified chronic bronchitis
C26	540	J441	1678	852	0.083	0.024	3.44	Chronic obstructive pulmonary disease with acute exacerbation, unspecified
C27	440	I649	2991	823	0.666	0.037	18.07	Stroke, not specified as haemorrhage or infarction
C27	440	I694	1733	823	0.360	0.022	16.60	Sequelae of stroke, not specified as haemorrhage or infarction
C27	440	I693	467	823	0.079	0.006	13.03	Sequelae of cerebral infarction

Clst	Sum	ICD-10	# code	# clst	Obs	Exp	O/E-ratio	description
C27	440	I639	1989	823	0.332	0.026	12.82	Cerebral infarction, unspecified
C27	440	I652	563	823	0.085	0.007	11.44	Occlusion and stenosis of carotid artery
C27	440	G459	2066	823	0.270	0.028	9.70	Transient cerebral ischaemic attack, unspecified
C27	440	R670	750	823	0.077	0.010	7.39	NA
C27	440	G409	841	823	0.080	0.012	6.86	Epilepsy, unspecified
C27	440	G969	710	823	0.055	0.010	5.45	Disorder of central nervous system, unspecified
C27	440	R298	1366	823	0.104	0.019	5.41	Other and unspecified symptoms and signs involving the nervous and musculoskeletal systems
C28	483	C619	1173	686	0.943	0.008	119.15	Malignant neoplasm of prostate
C28	483	R339	1530	686	0.131	0.022	6.05	Retention of urine
C28	483	N484	456	686	0.031	0.007	4.68	Impotence of organic origin
C28	483	C443	522	686	0.026	0.008	3.46	Malignant neoplasm: Skin of other and unspecified parts of face
C28	483	N133	267	686	0.013	0.004	3.38	Other and unspecified hydronephrosis
C28	483	N479	462	686	0.022	0.007	3.25	Redundant prepuce, phimosis and paraphimosis
C28	483	N409	3319	686	0.147	0.048	3.04	Hyperplasia of prostate
C28	483	R391	2230	686	0.096	0.033	2.95	Other difficulties with micturition
C28	483	A419	968	686	0.039	0.014	2.78	Sepsis, unspecified
C28	483	N359	253	686	0.010	0.004	2.76	Urethral stricture, unspecified
C29	264	A630	320	550	0.082	0.004	19.81	Anogenital (venereal) warts
C29	264	L022	506	550	0.076	0.007	10.96	Cutaneous abscess, furuncle and carbuncle of trunk
C29	264	L024	953	550	0.135	0.013	10.19	Cutaneous abscess, furuncle and carbuncle of limb
C29	264	L089	1001	550	0.113	0.014	7.99	Local infection of skin and subcutaneous tissue, unspecified
C29	264	L029	622	550	0.069	0.009	7.88	Cutaneous abscess, furuncle and carbuncle, unspecified
C29	264	I803	1509	550	0.142	0.021	6.60	Phlebitis and thrombophlebitis of lower extremities, unspecified
C29	264	A499	492	550	0.044	0.007	6.21	Bacterial infection, unspecified
C29	264	A469	1352	550	0.109	0.019	5.62	Erysipelas
C29	264	I829	297	550	0.016	0.004	3.78	Embolism and thrombosis of unspecified vein
C29	264	R509	907	550	0.047	0.013	3.57	Fever, unspecified
C30	1434	N180	454	533	0.610	0.002	314.82	Chronic kidney disease
C30	1434	N199	711	533	0.585	0.006	97.71	Unspecified kidney failure
C30	1434	N189	1094	533	0.848	0.010	87.98	Chronic kidney disease, unspecified
C30	1434	K650	264	533	0.148	0.003	53.36	Acute peritonitis
C30	1434	E102	377	533	0.135	0.005	29.50	Type 1 diabetes mellitus: With renal complications
C30	1434	N179	377	533	0.086	0.005	17.37	Acute renal failure, unspecified
C30	1434	E112	761	533	0.167	0.010	16.55	Type 2 diabetes mellitus: With renal complications
C30	1434	K053	360	533	0.077	0.005	16.06	Chronic periodontitis
C30	1434	E107	583	533	0.120	0.008	15.41	Type 1 diabetes mellitus: With multiple complications
C30	1434	N133	267	533	0.045	0.004	12.34	Other and unspecified hydronephrosis
C31	727	M059	682	520	0.679	0.005	137.45	Seropositive rheumatoid arthritis, unspecified
C31	727	M069	662	520	0.567	0.006	102.97	Rheumatoid arthritis, unspecified
C31	727	M060	398	520	0.246	0.004	60.73	Seronegative rheumatoid arthritis
C31	727	M204	350	520	0.063	0.005	13.34	Other hammer toe(s) (acquired)
C31	727	M139	488	520	0.067	0.007	9.90	Arthritis, unspecified
C31	727	M029	304	520	0.031	0.004	7.12	Reactive arthropathy, unspecified
C31	727	J849	265	520	0.019	0.004	5.02	Interstitial pulmonary disease, unspecified
C31	727	M201	858	520	0.060	0.012	4.80	Hallux valgus (acquired)

Clst	Sum	ICD-10	# code	# clst	Obs	Exp	O/E-ratio	description
C31	727	M255	842	520	0.050	0.012	4.08	Pain in joint
C31	727	M190	837	520	0.048	0.012	3.94	Primary arthrosis of other joints

S5B, Table: Bottom-10 O/E-ratios < 1 per cluster

Clst	ICD-10	# code	# clst	Obs	Exp	O/E-ratio	description
C1	E107	583	7191	0.001	0.010	0.12	Type 1 diabetes mellitus: With multiple complications
C1	D303	600	7191	0.002	0.010	0.20	Benign neoplasm: Bladder
C1	E105	416	7191	0.001	0.007	0.20	Type 1 diabetes mellitus: With peripheral circulatory complications
C1	J441	1678	7191	0.006	0.027	0.21	Chronic obstructive pulmonary disease with acute exacerbation, unspecified
C1	E780	12780	7191	0.046	0.208	0.22	Pure hypercholesterolaemia
C1	E103	576	7191	0.002	0.009	0.22	Type 1 diabetes mellitus: With ophthalmic complications
C1	C619	1173	7191	0.004	0.019	0.23	Malignant neoplasm of prostate
C1	E102	377	7191	0.001	0.006	0.23	Type 1 diabetes mellitus: With renal complications
C1	H360	1467	7191	0.006	0.024	0.23	Diabetic retinopathy
C1	E104	434	7191	0.002	0.007	0.24	Type 1 diabetes mellitus: With neurological complications
C2	D303	600	5990	0.000	0.010	0.02	Benign neoplasm: Bladder
C2	C619	1173	5990	0.000	0.019	0.02	Malignant neoplasm of prostate
C2	E107	583	5990	0.000	0.010	0.02	Type 1 diabetes mellitus: With multiple complications
C2	E103	576	5990	0.000	0.009	0.02	Type 1 diabetes mellitus: With ophthalmic complications
C2	N180	454	5990	0.000	0.007	0.02	Chronic kidney disease
C2	H360	1467	5990	0.001	0.024	0.03	Diabetic retinopathy
C2	E108	1073	5990	0.001	0.017	0.03	Type 1 diabetes mellitus: With unspecified complications
C2	N409	3319	5990	0.002	0.054	0.03	Hyperplasia of prostate
C2	E104	434	5990	0.000	0.007	0.05	Type 1 diabetes mellitus: With neurological complications
C2	E148	483	5990	0.001	0.008	0.06	Unspecified diabetes mellitus: With unspecified complications
C3	E112	761	4641	0.000	0.012	0.02	Type 2 diabetes mellitus: With renal complications
C3	E113	720	4641	0.000	0.012	0.02	Type 2 diabetes mellitus: With ophthalmic complications
C3	D303	600	4641	0.000	0.010	0.02	Benign neoplasm: Bladder
C3	C619	1173	4641	0.000	0.019	0.02	Malignant neoplasm of prostate
C3	I350	2664	4641	0.001	0.043	0.03	Aortic (valve) stenosis
C3	M171	2940	4641	0.002	0.047	0.03	Other primary gonarthrosis
C3	J441	1678	4641	0.001	0.027	0.03	Chronic obstructive pulmonary disease with acute exacerbation, unspecified
C3	N409	3319	4641	0.002	0.053	0.04	Hyperplasia of prostate
C3	I509	6160	4641	0.004	0.098	0.04	Heart failure, unspecified
C3	M059	682	4641	0.000	0.011	0.04	Seropositive rheumatoid arthritis, unspecified
C4	N180	454	4401	0.002	0.007	0.26	Chronic kidney disease
C4	J350	315	4401	0.001	0.005	0.28	Chronic tonsillitis
C4	E105	416	4401	0.002	0.007	0.28	Type 1 diabetes mellitus: With peripheral circulatory complications
C4	E103	576	4401	0.003	0.009	0.30	Type 1 diabetes mellitus: With ophthalmic complications
C4	K801	411	4401	0.002	0.006	0.36	Calculus of gallbladder with other cholecystitis
C4	E107	583	4401	0.003	0.009	0.38	Type 1 diabetes mellitus: With multiple complications
C4	M235	269	4401	0.002	0.004	0.38	Chronic instability of knee

Clst	ICD-10	# code	# clst	Obs	Exp	O/E-ratio	description
C4	G439	463	4401	0.003	0.007	0.41	Migraine, unspecified
C4	H360	1467	4401	0.010	0.023	0.43	Diabetic retinopathy
C4	J039	402	4401	0.003	0.006	0.44	Acute tonsillitis, unspecified
C5	D303	600	4290	0.003	0.009	0.27	Benign neoplasm: Bladder
C5	I495	482	4290	0.002	0.008	0.31	Sick sinus syndrome
C5	K402	280	4290	0.001	0.004	0.32	Bilateral inguinal hernia, without obstruction or gangrene
C5	I491	273	4290	0.001	0.004	0.33	Atrial premature depolarization
C5	K409	3787	4290	0.021	0.059	0.35	Unilateral or unspecified inguinal hernia, without obstruction or gangrene
C5	M235	269	4290	0.002	0.004	0.39	Chronic instability of knee
C5	M539	274	4290	0.002	0.004	0.44	Dorsopathy, unspecified
C5	I479	614	4290	0.004	0.009	0.44	Paroxysmal tachycardia, unspecified
C5	C509	1103	4290	0.008	0.017	0.45	Malignant neoplasm: Breast, unspecified
C5	N508	300	4290	0.002	0.005	0.45	Other specified disorders of male genital organs
C6	N180	454	3589	0.001	0.007	0.12	Chronic kidney disease
C6	C679	292	3589	0.001	0.005	0.12	Malignant neoplasm: Bladder, unspecified
C6	N199	711	3589	0.001	0.011	0.12	Unspecified kidney failure
C6	I489	7075	3589	0.015	0.110	0.14	Atrial fibrillation and atrial flutter, unspecified
C6	J441	1678	3589	0.004	0.026	0.14	Chronic obstructive pulmonary disease with acute exacerbation, unspecified
C6	J180	257	3589	0.001	0.004	0.14	Bronchopneumonia, unspecified
C6	C509	1103	3589	0.003	0.017	0.15	Malignant neoplasm: Breast, unspecified
C6	J960	837	3589	0.002	0.013	0.15	Acute respiratory failure
C6	J440	743	3589	0.002	0.012	0.17	Chronic obstructive pulmonary disease with acute lower respiratory infection
C6	K409	3787	3589	0.010	0.059	0.17	Unilateral or unspecified inguinal hernia, without obstruction or gangrene
C7	E107	583	3309	0.000	0.009	0.03	Type 1 diabetes mellitus: With multiple complications
C7	E103	576	3309	0.000	0.009	0.03	Type 1 diabetes mellitus: With ophthalmic complications
C7	N180	454	3309	0.000	0.007	0.04	Chronic kidney disease
C7	E104	434	3309	0.000	0.007	0.04	Type 1 diabetes mellitus: With neurological complications
C7	I709	433	3309	0.000	0.007	0.04	Generalized and unspecified atherosclerosis
C7	E105	416	3309	0.000	0.007	0.05	Type 1 diabetes mellitus: With peripheral circulatory complications
C7	E102	377	3309	0.000	0.006	0.05	Type 1 diabetes mellitus: With renal complications
C7	N189	1094	3309	0.001	0.017	0.05	Chronic kidney disease, unspecified
C7	J180	257	3309	0.000	0.004	0.07	Bronchopneumonia, unspecified
C7	E115	551	3309	0.001	0.009	0.11	Type 2 diabetes mellitus: With peripheral circulatory complications
C8	E103	576	2802	0.000	0.009	0.04	Type 1 diabetes mellitus: With ophthalmic complications
C8	E108	1073	2802	0.001	0.017	0.09	Type 1 diabetes mellitus: With unspecified complications
C8	E104	434	2802	0.001	0.007	0.11	Type 1 diabetes mellitus: With neurological complications
C8	E105	416	2802	0.001	0.006	0.11	Type 1 diabetes mellitus: With peripheral circulatory complications

Clst	ICD-10	# code	# clst	Obs	Exp	O/E-ratio	description
C8	E115	551	2802	0.001	0.009	0.13	Type 2 diabetes mellitus: With peripheral circulatory complications
C8	M512	366	2802	0.001	0.006	0.13	Other specified intervertebral disc displacement
C8	E162	510	2802	0.001	0.008	0.14	Hypoglycaemia, unspecified
C8	E109	2680	2802	0.006	0.041	0.15	Type 1 diabetes mellitus: Without complications
C8	N180	454	2802	0.001	0.007	0.15	Chronic kidney disease
C8	E107	583	2802	0.001	0.009	0.16	Type 1 diabetes mellitus: With multiple complications
C9	M539	274	2581	0.000	0.004	0.09	Dorsopathy, unspecified
C9	D179	257	2581	0.000	0.004	0.10	Benign lipomatous neoplasm, unspecified
C9	D251	256	2581	0.000	0.004	0.10	Intramural leiomyoma of uterus
C9	D303	600	2581	0.001	0.009	0.13	Benign neoplasm: Bladder
C9	R072	367	2581	0.001	0.006	0.14	Precordial pain
C9	M512	366	2581	0.001	0.006	0.14	Other specified intervertebral disc displacement
C9	G439	463	2581	0.001	0.007	0.16	Migraine, unspecified
C9	M653	738	2581	0.002	0.011	0.17	Trigger finger
C9	M511	3357	2581	0.009	0.052	0.17	Lumbar and other intervertebral disc disorders with radiculopathy
C9	M431	583	2581	0.002	0.009	0.17	Spondylolisthesis
C10	A630	320	2562	0.000	0.005	0.08	Anogenital (venereal) warts
C10	K298	256	2562	0.000	0.004	0.10	Duodenitis
C10	I999	383	2562	0.001	0.006	0.13	Other and unspecified disorders of circulatory system
C10	F172	349	2562	0.001	0.005	0.14	Mental and behavioural disorders due to use of tobacco: Dependence syndrome
C10	M771	492	2562	0.001	0.008	0.16	Lateral epicondylitis
C10	N479	462	2562	0.001	0.007	0.16	Redundant prepuce, phimosis and paraphimosis
C10	N180	454	2562	0.001	0.007	0.17	Chronic kidney disease
C10	I309	297	2562	0.001	0.005	0.17	Acute pericarditis, unspecified
C10	I830	293	2562	0.001	0.005	0.17	Varicose veins of lower extremities with ulcer
C10	I802	280	2562	0.001	0.004	0.18	Phlebitis and thrombophlebitis of other deep vessels of lower extremities
C11	C509	1103	2292	0.000	0.017	0.03	Malignant neoplasm: Breast, unspecified
C11	H360	1467	2292	0.001	0.023	0.04	Diabetic retinopathy
C11	E113	720	2292	0.000	0.011	0.04	Type 2 diabetes mellitus: With ophthalmic complications
C11	E107	583	2292	0.000	0.009	0.05	Type 1 diabetes mellitus: With multiple complications
C11	E103	576	2292	0.000	0.009	0.05	Type 1 diabetes mellitus: With ophthalmic complications
C11	E104	434	2292	0.000	0.007	0.06	Type 1 diabetes mellitus: With neurological complications
C11	E105	416	2292	0.000	0.006	0.07	Type 1 diabetes mellitus: With peripheral circulatory complications
C11	E112	761	2292	0.001	0.012	0.07	Type 2 diabetes mellitus: With renal complications
C11	D249	758	2292	0.001	0.012	0.07	Benign neoplasm of breast
C11	E108	1073	2292	0.001	0.017	0.08	Type 1 diabetes mellitus: With unspecified complications
C12	C509	1103	2213	0.000	0.017	0.03	Malignant neoplasm: Breast, unspecified
C12	N921	875	2213	0.000	0.013	0.03	Excessive and frequent menstruation with irregular cycle
C12	E103	576	2213	0.000	0.009	0.05	Type 1 diabetes mellitus: With ophthalmic complications
C12	N924	561	2213	0.000	0.009	0.05	Excessive bleeding in the premenopausal period

Clst	ICD-10	# code	# clst	Obs	Exp	O/E-ratio	description
C12	E148	483	2213	0.000	0.007	0.06	Unspecified diabetes mellitus: With unspecified complications
C12	N920	948	2213	0.001	0.015	0.06	Excessive and frequent menstruation with regular cycle
C12	F419	335	2213	0.000	0.005	0.09	Anxiety disorder, unspecified
C12	R064	542	2213	0.001	0.008	0.11	Hyperventilation
C12	D251	256	2213	0.000	0.004	0.12	Intramural leiomyoma of uterus
C12	N832	489	2213	0.001	0.008	0.12	Other and unspecified ovarian cysts
C13	E112	761	2070	0.000	0.012	0.04	Type 2 diabetes mellitus: With renal complications
C13	E102	377	2070	0.000	0.006	0.08	Type 1 diabetes mellitus: With renal complications
C13	N179	377	2070	0.000	0.006	0.08	Acute renal failure, unspecified
C13	D509	459	2070	0.001	0.007	0.14	Iron deficiency anaemia, unspecified
C13	H330	418	2070	0.001	0.006	0.15	Retinal detachment with retinal break
C13	J819	545	2070	0.001	0.008	0.17	NA
C13	E162	510	2070	0.001	0.008	0.19	Hypoglycaemia, unspecified
C13	E117	980	2070	0.003	0.015	0.19	Type 2 diabetes mellitus: With multiple complications
C13	I501	1502	2070	0.005	0.023	0.21	Left ventricular failure
C13	C679	292	2070	0.001	0.004	0.22	Malignant neoplasm: Bladder, unspecified
C14	H360	1467	2040	0.002	0.022	0.11	Diabetic retinopathy
C14	E103	576	2040	0.001	0.009	0.17	Type 1 diabetes mellitus: With ophthalmic complications
C14	E105	416	2040	0.001	0.006	0.23	Type 1 diabetes mellitus: With peripheral circulatory complications
C14	G510	365	2040	0.001	0.006	0.26	Bell's palsy
C14	H431	334	2040	0.001	0.005	0.29	Vitreous haemorrhage
C14	E113	720	2040	0.003	0.011	0.31	Type 2 diabetes mellitus: With ophthalmic complications
C14	M519	604	2040	0.003	0.009	0.32	Intervertebral disc disorder, unspecified
C14	L905	295	2040	0.001	0.004	0.33	Scar conditions and fibrosis of skin
C14	E107	583	2040	0.003	0.009	0.33	Type 1 diabetes mellitus: With multiple complications
C14	H438	291	2040	0.001	0.004	0.33	Other disorders of vitreous body
C15	J439	317	2013	0.000	0.005	0.10	Emphysema, unspecified
C15	C679	292	2013	0.000	0.004	0.11	Malignant neoplasm: Bladder, unspecified
C15	K510	283	2013	0.000	0.004	0.12	Ulcerative (chronic) pancolitis
C15	I495	482	2013	0.001	0.007	0.14	Sick sinus syndrome
C15	K045	333	2013	0.001	0.005	0.20	Chronic apical periodontitis
C15	R570	320	2013	0.001	0.005	0.20	Cardiogenic shock
C15	R590	305	2013	0.001	0.005	0.21	Localized enlarged lymph nodes
C15	N180	454	2013	0.001	0.007	0.22	Chronic kidney disease
C15	D303	600	2013	0.002	0.009	0.22	Benign neoplasm: Bladder
C15	N133	267	2013	0.001	0.004	0.24	Other and unspecified hydronephrosis
C16	E149	958	1654	0.001	0.015	0.04	Unspecified diabetes mellitus: Without complications
C16	E114	881	1654	0.001	0.013	0.04	Type 2 diabetes mellitus: With neurological complications
C16	E103	576	1654	0.001	0.009	0.07	Type 1 diabetes mellitus: With ophthalmic complications
C16	H360	1467	1654	0.002	0.022	0.08	Diabetic retinopathy
C16	E105	416	1654	0.001	0.006	0.10	Type 1 diabetes mellitus: With peripheral circulatory complications
C16	E102	377	1654	0.001	0.006	0.10	Type 1 diabetes mellitus: With renal complications

Clst	ICD-10	# code	# clst	Obs	Exp	O/E-ratio	description
C16	E108	1073	1654	0.002	0.016	0.11	Type 1 diabetes mellitus: With unspecified complications
C16	H431	334	1654	0.001	0.005	0.12	Vitreous haemorrhage
C16	E117	980	1654	0.002	0.015	0.12	Type 2 diabetes mellitus: With multiple complications
C16	E107	583	1654	0.001	0.009	0.14	Type 1 diabetes mellitus: With multiple complications
C17	N199	711	1281	0.001	0.011	0.07	Unspecified kidney failure
C17	E162	510	1281	0.001	0.008	0.10	Hypoglycaemia, unspecified
C17	E105	416	1281	0.001	0.006	0.12	Type 1 diabetes mellitus: With peripheral circulatory complications
C17	I469	372	1281	0.001	0.006	0.14	Cardiac arrest, unspecified
C17	D303	600	1281	0.002	0.009	0.17	Benign neoplasm: Bladder
C17	E107	583	1281	0.002	0.009	0.18	Type 1 diabetes mellitus: With multiple complications
C17	E103	576	1281	0.002	0.009	0.18	Type 1 diabetes mellitus: With ophthalmic complications
C17	K580	576	1281	0.002	0.009	0.18	Irritable bowel syndrome with diarrhoea
C17	K650	264	1281	0.001	0.004	0.20	Acute peritonitis
C17	D179	257	1281	0.001	0.004	0.20	Benign lipomatous neoplasm, unspecified
C18	J450	401	1251	0.001	0.006	0.13	Predominantly allergic asthma
C18	H521	366	1251	0.001	0.006	0.14	Myopia
C18	J350	315	1251	0.001	0.005	0.17	Chronic tonsillitis
C18	N924	561	1251	0.002	0.008	0.19	Excessive bleeding in the premenopausal period
C18	K402	280	1251	0.001	0.004	0.19	Bilateral inguinal hernia, without obstruction or gangrene
C18	N840	546	1251	0.002	0.008	0.19	Polyp of corpus uteri
C18	M235	269	1251	0.001	0.004	0.20	Chronic instability of knee
C18	M234	258	1251	0.001	0.004	0.20	Loose body in knee
C18	N832	489	1251	0.002	0.007	0.22	Other and unspecified ovarian cysts
C18	K800	468	1251	0.002	0.007	0.23	Calculus of gallbladder with acute cholecystitis
C19	C509	1103	1168	0.002	0.017	0.10	Malignant neoplasm: Breast, unspecified
C19	D649	1637	1168	0.003	0.025	0.10	Anaemia, unspecified
C19	F103	518	1168	0.001	0.008	0.11	Mental and behavioural disorders due to use of alcohol: Withdrawal state
C19	K810	505	1168	0.001	0.008	0.11	Acute cholecystitis
C19	N811	974	1168	0.002	0.015	0.12	Cystocele
C19	I429	479	1168	0.001	0.007	0.12	Cardiomyopathy, unspecified
C19	I709	433	1168	0.001	0.007	0.13	Generalized and unspecified atherosclerosis
C19	N390	1219	1168	0.003	0.018	0.14	Urinary tract infection, site not specified
C19	R001	394	1168	0.001	0.006	0.14	Bradycardia, unspecified
C19	E871	393	1168	0.001	0.006	0.14	Hypo-osmolality and hyponatraemia
C20	N920	948	1119	0.001	0.014	0.06	Excessive and frequent menstruation with regular cycle
C20	N921	875	1119	0.001	0.013	0.07	Excessive and frequent menstruation with irregular cycle
C20	G442	667	1119	0.001	0.010	0.09	Tension-type headache
C20	N924	561	1119	0.001	0.008	0.10	Excessive bleeding in the premenopausal period
C20	D279	435	1119	0.001	0.007	0.14	NA
C20	E104	434	1119	0.001	0.007	0.14	Type 1 diabetes mellitus: With neurological complications
C20	M224	423	1119	0.001	0.006	0.14	Chondromalacia patellae
C20	R104	391	1119	0.001	0.006	0.15	Other and unspecified abdominal pain
C20	E041	369	1119	0.001	0.006	0.16	Nontoxic single thyroid nodule

Clst	ICD-10	# code	# clst	Obs	Exp	O/E-ratio	description
C20	M512	366	1119	0.001	0.006	0.16	Other specified intervertebral disc displacement
C21	J960	837	1000	0.001	0.013	0.08	Acute respiratory failure
C21	E107	583	1000	0.001	0.009	0.11	Type 1 diabetes mellitus: With multiple complications
C21	E103	576	1000	0.001	0.009	0.12	Type 1 diabetes mellitus: With ophthalmic complications
C21	I493	575	1000	0.001	0.009	0.12	Ventricular premature depolarization
C21	I495	482	1000	0.001	0.007	0.14	Sick sinus syndrome
C21	N180	454	1000	0.001	0.007	0.15	Chronic kidney disease
C21	R091	418	1000	0.001	0.006	0.16	Pleurisy
C21	E789	417	1000	0.001	0.006	0.16	Disorder of lipoprotein metabolism, unspecified
C21	E871	393	1000	0.001	0.006	0.17	Hypo-osmolality and hyponatraemia
C21	E102	377	1000	0.001	0.006	0.18	Type 1 diabetes mellitus: With renal complications
C22	F100	1212	988	0.001	0.018	0.06	Mental and behavioural disorders due to use of alcohol: Acute intoxication
C22	L979	605	988	0.001	0.009	0.11	Ulcer of lower limb, not elsewhere classified
C22	I429	479	988	0.001	0.007	0.14	Cardiomyopathy, unspecified
C22	N180	454	988	0.001	0.007	0.15	Chronic kidney disease
C22	M224	423	988	0.001	0.006	0.16	Chondromalacia patellae
C22	J448	396	988	0.001	0.006	0.17	Other specified chronic obstructive pulmonary disease
C22	E102	377	988	0.001	0.006	0.18	Type 1 diabetes mellitus: With renal complications
C22	I469	372	988	0.001	0.006	0.18	Cardiac arrest, unspecified
C22	R570	320	988	0.001	0.005	0.21	Cardiogenic shock
C22	L400	319	988	0.001	0.005	0.21	Psoriasis vulgaris
C23	F103	518	935	0.001	0.008	0.14	Mental and behavioural disorders due to use of alcohol: Withdrawal state
C23	M191	474	935	0.001	0.007	0.15	Post-traumatic arthrosis of other joints
C23	R490	415	935	0.001	0.006	0.17	Dysphonia
C23	C619	1173	935	0.003	0.018	0.18	Malignant neoplasm of prostate
C23	I999	383	935	0.001	0.006	0.18	Other and unspecified disorders of circulatory system
C23	E041	369	935	0.001	0.006	0.19	Nontoxic single thyroid nodule
C23	H659	368	935	0.001	0.006	0.19	Nonsuppurative otitis media, unspecified
C23	I480	364	935	0.001	0.005	0.20	Paroxysmal atrial fibrillation
C23	H908	631	935	0.002	0.010	0.22	Mixed conductive and sensorineural hearing loss, unspecified
C23	I839	1522	935	0.005	0.023	0.23	Varicose veins of lower extremities without ulcer or inflammation
C24	N409	3319	932	0.002	0.050	0.04	Hyperplasia of prostate
C24	C619	1173	932	0.001	0.018	0.06	Malignant neoplasm of prostate
C24	E108	1073	932	0.001	0.016	0.07	Type 1 diabetes mellitus: With unspecified complications
C24	H833	1412	932	0.002	0.021	0.10	Noise effects on inner ear
C24	E107	583	932	0.001	0.009	0.12	Type 1 diabetes mellitus: With multiple complications
C24	M109	547	932	0.001	0.008	0.13	Gout, unspecified
C24	E116	477	932	0.001	0.007	0.15	Type 2 diabetes mellitus: With other specified complications
C24	M191	474	932	0.001	0.007	0.15	Post-traumatic arthrosis of other joints
C24	H109	465	932	0.001	0.007	0.15	Conjunctivitis, unspecified
C24	N484	456	932	0.001	0.007	0.16	Impotence of organic origin
C25	H938	1116	860	0.001	0.017	0.07	Other specified disorders of ear
C25	H360	1467	860	0.002	0.022	0.10	Diabetic retinopathy
C25	N200	1391	860	0.002	0.021	0.11	Calculus of kidney

Clst	ICD-10	# code	# clst	Obs	Exp	O/E-ratio	description
C25	H911	3527	860	0.007	0.053	0.13	Presbycusis
C25	C619	1173	860	0.002	0.018	0.13	Malignant neoplasm of prostate
C25	E107	583	860	0.001	0.009	0.13	Type 1 diabetes mellitus: With multiple complications
C25	E103	576	860	0.001	0.009	0.13	Type 1 diabetes mellitus: With ophthalmic complications
C25	I493	575	860	0.001	0.009	0.13	Ventricular premature depolarization
C25	H919	4610	860	0.010	0.069	0.15	Hearing loss, unspecified
C25	H833	1412	860	0.003	0.021	0.16	Noise effects on inner ear
C26	E112	761	852	0.001	0.011	0.10	Type 2 diabetes mellitus: With renal complications
C26	N409	3319	852	0.006	0.050	0.12	Hyperplasia of prostate
C26	C619	1173	852	0.002	0.018	0.13	Malignant neoplasm of prostate
C26	I350	2664	852	0.006	0.040	0.15	Aortic (valve) stenosis
C26	I714	517	852	0.001	0.008	0.15	Abdominal aortic aneurysm, without mention of rupture
C26	E162	510	852	0.001	0.008	0.15	Hypoglycaemia, unspecified
C26	H360	1467	852	0.004	0.022	0.16	Diabetic retinopathy
C26	I495	482	852	0.001	0.007	0.16	Sick sinus syndrome
C26	K800	468	852	0.001	0.007	0.17	Calculus of gallbladder with acute cholecystitis
C26	N180	454	852	0.001	0.007	0.17	Chronic kidney disease
C27	H360	1467	823	0.001	0.022	0.06	Diabetic retinopathy
C27	M170	2145	823	0.004	0.032	0.11	Primary gonarthrosis, bilateral
C27	D303	600	823	0.001	0.009	0.14	Benign neoplasm: Bladder
C27	C509	1103	823	0.002	0.017	0.15	Malignant neoplasm: Breast, unspecified
C27	J209	527	823	0.001	0.008	0.15	Acute bronchitis, unspecified
C27	J320	513	823	0.001	0.008	0.16	Chronic maxillary sinusitis
C27	E162	510	823	0.001	0.008	0.16	Hypoglycaemia, unspecified
C27	L022	506	823	0.001	0.008	0.16	Cutaneous abscess, furuncle and carbuncle of trunk
C27	K810	505	823	0.001	0.008	0.16	Acute cholecystitis
C27	M179	2242	823	0.006	0.034	0.18	Gonarthrosis, unspecified
C28	M546	977	686	0.001	0.015	0.10	Pain in thoracic spine
C28	D249	758	686	0.001	0.011	0.13	Benign neoplasm of breast
C28	E113	720	686	0.001	0.011	0.14	Type 2 diabetes mellitus: With ophthalmic complications
C28	E669	2114	686	0.004	0.032	0.14	Obesity, unspecified
C28	R002	691	686	0.001	0.010	0.14	Palpitations
C28	K359	678	686	0.001	0.010	0.14	NA
C28	H350	548	686	0.001	0.008	0.18	Background retinopathy and retinal vascular changes
C28	I119	534	686	0.001	0.008	0.18	Hypertensive heart disease without (congestive) heart failure
C28	J209	527	686	0.001	0.008	0.18	Acute bronchitis, unspecified
C28	E162	510	686	0.001	0.008	0.19	Hypoglycaemia, unspecified
C29	M171	2940	550	0.002	0.044	0.04	Other primary gonarthrosis
C29	I489	7075	550	0.005	0.106	0.05	Atrial fibrillation and atrial flutter, unspecified
C29	H919	4610	550	0.004	0.069	0.05	Hearing loss, unspecified
C29	I702	2251	550	0.002	0.034	0.05	Atherosclerosis of arteries of extremities
C29	M179	2242	550	0.002	0.034	0.05	Gonarthrosis, unspecified
C29	M170	2145	550	0.002	0.032	0.06	Primary gonarthrosis, bilateral
C29	G459	2066	550	0.002	0.031	0.06	Transient cerebral ischaemic attack, unspecified
C29	I639	1989	550	0.002	0.030	0.06	Cerebral infarction, unspecified
C29	G473	1897	550	0.002	0.028	0.06	Sleep apnoea
C29	N409	3319	550	0.004	0.050	0.07	Hyperplasia of prostate

Clst	ICD-10	# code	# clst	Obs	Exp	O/E-ratio	description
C30	F100	1212	533	0.002	0.018	0.10	Mental and behavioural disorders due to use of alcohol: Acute intoxication
C30	N393	827	533	0.002	0.012	0.15	Stress incontinence
C30	M503	520	533	0.002	0.008	0.24	Other cervical disc degeneration
C30	R065	983	533	0.004	0.015	0.26	Mouth breathing
C30	M546	977	533	0.004	0.015	0.26	Pain in thoracic spine
C30	R490	415	533	0.002	0.006	0.30	Dysphonia
C30	M060	398	533	0.002	0.006	0.32	Seronegative rheumatoid arthritis
C30	F102	1189	533	0.006	0.018	0.32	Mental and behavioural disorders due to use of alcohol: Dependence syndrome
C30	J448	396	533	0.002	0.006	0.32	Other specified chronic obstructive pulmonary disease
C30	C509	1103	533	0.006	0.017	0.34	Malignant neoplasm: Breast, unspecified
C31	F102	1189	520	0.002	0.018	0.11	Mental and behavioural disorders due to use of alcohol: Dependence syndrome
C31	C619	1173	520	0.002	0.018	0.11	Malignant neoplasm of prostate
C31	K439	981	520	0.002	0.015	0.13	Other and unspecified ventral hernia without obstruction or gangrene
C31	E117	980	520	0.002	0.015	0.13	Type 2 diabetes mellitus: With multiple complications
C31	E112	761	520	0.002	0.011	0.17	Type 2 diabetes mellitus: With renal complications
C31	I351	696	520	0.002	0.010	0.18	Aortic (valve) insufficiency
C31	N200	1391	520	0.004	0.021	0.18	Calculus of kidney
C31	E118	2669	520	0.008	0.040	0.19	Type 2 diabetes mellitus: With unspecified complications
C31	M519	604	520	0.002	0.009	0.21	Intervertebral disc disorder, unspecified
C31	E107	583	520	0.002	0.009	0.22	Type 1 diabetes mellitus: With multiple complications

S6 Table: Chi-squared test for distribution laboratory values in clusters

Component	P-val.	Adj. P-val.
Alanine transaminase (ALAT)	4.78 e-22	1.15e-20
Albumin	4.81e-22	1.15e-20
Alkaline phosphatase	2.01e-22	4.82e-21
Bilirubin	1.09e-13	2.60e-12
C-reactive protein (CRP)	1.65e-96	3.95e-95
Carbamide	5.49-e200	1.32e-198
Cholesterol HDL	1.99e-66	4.77e-65
Cholesterol LDL	4.86e-53	1.17e-51
Cholesterol total	2.64e-58	6.34 e-57
Coagulation factor II + VII + X	7.96e-280	1.91e-278
Creatinine	9.28e-302	2.23e-300
Eosinophils	4.43e-6	1.06e-4
Estimated glomerular filtration rate (eGFR)	0	0
Glucose	0	0
Hemoglobin	2.77e-218	6.65e-217
Leukocytes	1.42e-39	3.41e-38
Lymphocytes	1.54e-17	3.69e-16
Monocytes	1.06e-11	2.55e-10
Neutrophils	5.69e-20	1.36e-18
Platelets	2.39e-23	5.73e-22
Potassium	9.03e-32	2.17e-30
Sodium	2.24e-74	5.38e-74
Triglyceride	2.10e-60	5.04 e-59
Troponin	7.10e-73	1.70e-71

S7 Table: Traits with significantly different PGS distributions in clusters

Cluster	n	trait	effect	effect size	FDR
C1	2,025	Systolic Blood Pressure	+	0.20	<0.0005
		Diastolic Blood Pressure	+	0.16	<0.0005
		Total Cholesterol	-	-0.08	0.026
C4	1,532	Atrial Fibrillation	+	0.57	<0.0005
		Heart Failure	+	0.08	0.031
		Coronary Artery Disease	-	-0.12	0.001
		T2D (BMI-adj.)	-	-0.11	0.001
		Acute Myocardial Infarction	-	-0.08	0.031
		Triglyceride	-	-0.08	0.044
		Total Cholesterol	-	-0.08	0.046
C5	1,136	T2D (BMI-adj.)	+	0.55	<0.0005
		NAFLD	+	0.11	0.021
C6	860	Total Cholesterol	+	0.21	<0.0005
		Triglyceride	+	0.20	<0.0005
		LDL Cholesterol	+	0.15	0.001
		Coronary Artery Disease	+	0.15	0.001
		Diastolic Blood Pressure	-	-0.13	0.015
		Systolic Blood Pressure	-	-0.11	0.040
C8	817	Systolic Blood Pressure	-	-0.16	0.001
		Stroke	-	-0.12	0.023
		Coronary Artery Disease	-	-0.12	0.028
		Diastolic Blood Pressure	-	-0.11	0.031
		LDL Cholesterol	-	-0.10	0.047
C10	744	Coronary Artery Disease	-	-0.13	0.017
		Acute Myocardial Infarction	-	-0.13	0.021
C11	718	Stroke	-	-0.11	0.040
		Heart Failure	-	-0.11	0.040
C12	649	Coronary Artery Disease	-	-0.14	0.013
		Acute Myocardial Infarction	-	-0.12	0.033
C13	606	Diastolic Blood Pressure	-	-0.12	0.040
C15	588	LDL Cholesterol	+	0.14	0.020
		Total Cholesterol	+	0.14	0.026
C17	348	Systolic Blood Pressure	+	0.16	0.040
C18	481	T2D (BMI-adj.)	+	0.15	0.023
		Acute Myocardial Infarction	+	0.15	0.028
		Atrial Fibrillation	-	-0.13	0.049
C23	290	T2D (BMI-adj.)	+	0.27	0.001
C25	297	Coronary Artery Disease	+	0.24	0.002
		Acute Myocardial Infarction	+	0.18	0.032
		Heart Failure	+	0.18	0.031
C27	231	Stroke	+	0.24	0.015

Supplementary Material

Manuscript title: Subgrouping multimorbid patients with ischemic heart disease by means of unsupervised clustering: A cohort study of 72,249 patients defined comprehensively by diagnoses prior to presentation

Authors: Haue AD, Holm PC, et al.

S1 Appendix: Construction of patient similarity network, MCL algorithm settings and assessment of cluster robustness

To define the patient similarity network, a lower rank approximation of the *n* x *m* matrix was created using the "truncatedSVD" implecpmentation of SVD from the python package scikit-learn with 41 components, 10 iterations, and fixed random seed of 42 to ensure reproducible results. Thus, the 3,046 diagnoses were represented in 41 components based on this lower rank approximation. By selecting 41 components, the accumulated explained variance ratio was 0.50 (S2 Fig). To reduce the density of the patient similarity network, while still retaining an informative topology, all edges with an edge weight less than 0.3 were removed and the number of edges connected to each node were limited using the "#ceilnb" transformation from "mcl-edge"; a maximum of 8000 was used (S3 Fig). The weights of the remaining edges in the network were shifted such that the lowest weight was 0.0 as recommended in the MCL manual. The final pre-processed networks were then used as input for the "mcl" implementation of the MCL algorithm [1]. We used a pruning scheme of -P 7000 -S 800 -R 900 -pct 90, and a pre-inflation factor of 0.5 to make edge-weights more homogenous. For the MCL clustering, we selected a pre-inflation parameter of 2.0 corresponding to the default in the MCL manual [2].

For cluster robustness assessment, diluted versions of the reference clustering were generated by deleting edges with a probability of α, where α would range between 0 and 50% [3]. An α of 0 would leave the network unchanged. In contrast, the shuffled versions of the network had the same number of nodes and edges as the reference clustering. The shuffled networks were generated as described by Karrer et al. [4]. Finally, the generated clusters were compared to the reference clustering and the variances were quantified with reference to the so-called variation of information measure (VI) [5].

S2 Appendix: Preprocessing of laboratory data

The laboratory test results from the EHR data were originally archived in the administrative biochemical databases Labka and BCC [6]. In relation to the EHR data used in this study, Labka covers the hospitals with SHAK codes 1301, 1309, 1330, 1351, 1401, 1501, 1516 and 4001 in the Capital Region of Denmark and BCC covers the hospitals with SHAK codes 2000 and 2501 in Region Zealand for the periods 2009-2016 and 2012-2016, respectively (S2 Table). Biochemical laboratory tests were either classified in accordance with the Nomenclature, Properties and Units (NPU) or local systems [7]. Reference intervals were provided from the laboratories that analyzed the blood tests. Biochemical data was expected to be available for the patients where the index procedure was performed at a hospital located in either the Capital Region or Region Zealand at a time that was covered by the two databases.

A total of 48,957 patients (30,736 males and 18,221 females) were included from a hospital where biochemical data was available (67.8% of the entire cohort). As an indicator for data completeness and quality, the number of patients where available biochemical data at time of index agreed with the clinical standard of care was assessed. This implied that patients had sodium, potassium, hemoglobin, and creatinine (or estimated glomerular filtration rate) measured maximum 90 days before or at the day of index. The 31,224 patients who fulfilled this requirement were included in the biochemical analysis. Laboratory measurements available for at least 50% of these patients were included in the analysis. In cases where patients had more than one test available in the period up from 90 days before to index, the test closest to index was used. As listed in the main text included samples were plasma levels of potassium, sodium, hemoglobin, estimated glomerular filtration (eGFR), creatinine, carbamide, glucose, troponin (I/T), HDL cholesterol, LDL cholesterol, total cholesterol, leukocytes, C-reactive protein, lymphocytes, monocytes, neutrophils, basophils,

platelets, INR, alanine transaminase, albumin, alkaline phosphatase, bilirubin, and triglyceride. All analyses of biochemical data were performed in R 3.6.2 using the “ComplexHeatmap” and “circlize” packages .

S3 Appendix: Calculation of polygenetic risk scores for 14 traits

Polygenic risk scores were calculated using the LDpred2 framework, implemented in the R package bigsnpr (v1.11.6) with R version 4.0.0 and the workflow management system Snakemake [11–13]. In preparation for PGS calculations, autosomal genotype data from 242,644 individuals in the Copenhagen Hospital Biobank – Cardiovascular Disease Cohort (CHB-CVDC) [14] was filtered to only include variants present in LDpred2's recommended set of 1,054,330 reference variants. This recommended set is based on the reference set HapMap3 from the International HapMap project, which was established by genotyping 1.6 million single nucleotide polymorphisms (SNPs) in 1,184 individuals from 11 global populations [15]. Any missing genotype information was assumed to be the affected locus' reference allele.

We matched the remaining set of 994,643 genotyped variants with variants found in summary statistics data corresponding to 14 traits, obtained from nine GWAS meta-analyses (atrial fibrillation [16], BMI-adjusted type 2 diabetes [17], chronic kidney disease [18], HDL cholesterol levels [19], heart failure [20], LDL cholesterol levels [19], stroke [21], total cholesterol levels [19], triglyceride levels [19]) and five GWAS (acute myocardial infarction [22], coronary artery disease [23], diastolic blood pressure [24], non-alcoholic fatty liver disease [25], systolic blood pressure [24]). Variants present in both genotype and summary statistics data were then subject to LDpred2's recommended standard deviation quality control. After variant matching and quality control, a mean of 963,354 (S.D. 87,774) variants remained for subsequent per-chromosome risk score calculation for each of the 14 traits. We used the LDpred2-auto algorithm with 30 Gibbs sampling chains, 1,000 burn-in iterations and 500 iterations after burn-in. The initial values for the 30 sampling chains were a) the LDSC regression estimate for heritability h^2 (same for all chains); b) one of 30 initial values for the proportion of causal variants p, evenly spaced on a logarithmic scale from 10^{-4} to 0.5.

Variant effect sizes were calculated from each set of 30 sampling chains (per trait and chromosome) through a three-step process, which serves to ensure that the model (spanning 30 chains) successfully converged: 1) computing the standard deviations of each chains' predicted scores, 2) keeping only the chains within three median absolute deviations from the median standard deviation, 3) averaging the effect sizes of the remaining chains. Across the 308 per-chromosome models (14 traits times 22 chromosomes), 28.9 chains were included in the final score on average. For each individual, we calculated per-chromosome risk scores by multiplying the average variant effect sizes with the individual's corresponding genotype, and then added the per-chromosome risk scores up into one genome-wide PGS. To ease comparisons across traits, each trait's PGS distribution was scaled to a mean of zero and a standard deviation of one.

References, S1-3 Appendices

1. Van Dongen, S. M. Graph clustering by flow simulation. (2000).
2. MCL - a cluster algorithm for graphs. http://micans.org/mcl/.
3. Kirk, I. K. *et al.* Linking glycemic dysregulation in diabetes to symptoms, comorbidities, and genetics through EHR data mining. *eLife* **8**, e44941 (2019).
4. Karrer, B., Levina, E. & Newman, M. E. J. Robustness of community structure in networks. *Phys. Rev. E* **77**, 046119 (2008).
5. Meilă, M. Comparing clusterings—an information based distance. *J. Multivar. Anal.* **98**, 873–895 (2007).
6. Grann, A. F., Erichsen, R., Nielsen, A. G., Frøslev, T. & Thomsen, R. W. Existing data sources for clinical epidemiology: The clinical laboratory information system (LABKA) research database at Aarhus University, Denmark. *Clin. Epidemiol.* **3**, 133–138 (2011).
7. Petersen, U. M., Dybkaer, R. & Olesen, H. Properties and units in the clinical laboratory sciences. Part XXIII. The NPU terminology, principles, and implementation: A user's guide (IUPAC Technical Report)*. *Pure Appl Chem* **84**, 137–165 (2012).
8. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2019).
9. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811–2812 (2014).
10. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).
11. Mölder, F. *et al.* Sustainable data analysis with Snakemake. Preprint at https://doi.org/10.12688/f1000research.29032.2 (2021).

12. Privé, F., Arbel, J. & Vilhjálmsson, B. J. LDpred2: better, faster, stronger. *Bioinformatics* **36**, 5424–5431 (2020).
13. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2020).
14. Sørensen, E. *et al.* Data Resource Profile: The Copenhagen Hospital Biobank (CHB). *Int. J. Epidemiol.* **50**, 719–720e (2021).
15. Altshuler, D. M. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
16. Nielsen, J. B. *et al.* Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. *Nat. Genet.* **50**, 1234–1239 (2018).
17. Mahajan, A. *et al.* Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.* **50**, 1505–1513 (2018).
18. Wuttke, M. *et al.* A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nat. Genet.* **51**, 957–972 (2019).
19. Surakka, I. *et al.* The impact of low-frequency and rare variants on lipid levels. *Nat. Genet.* **47**, 589–597 (2015).
20. Shah, S. *et al.* Genome-wide association and Mendelian randomisation analysis provide insights into the pathogenesis of heart failure. *Nat. Commun.* **11**, 163 (2020).
21. Malik, R. *et al.* Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. *Nat. Genet.* **50**, 524–537 (2018).
22. Jiang, L., Zheng, Z., Fang, H. & Yang, J. A generalized linear mixed model association tool for biobank-scale data. *Nat. Genet.* **53**, 1616–1621 (2021).
23. van der Harst, P. & Verweij, N. Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. *Circ. Res.* **122**, 433–443 (2018).

24. Hoffmann, T. J. *et al.* Genome-wide association analyses using electronic health records identify new loci influencing blood pressure variation. *Nat. Genet.* **49**, 54–64 (2017).

25. Anstee, Q. M. *et al.* Genome-wide association study of non-alcoholic fatty liver and steatohepatitis in a histologically characterised cohort☆. *J. Hepatol.* **73**, 505–515 (2020).

Appendix B

Manuscript for Study II

Development and validation of a neural network-based survival model for mortality in ischemic heart disease

Peter C. Holm[a], Amalie D. Haue[a,b], David Westergaard[a, c, d], Timo Röder[a], Karina Banasik[a,c], Vinicius Tragante[e], Alex H. Christensen[b,f,g], Laurent Thomas[h,i], Therese H. Nøst[i], Anne-Heidi Skogholt[i], Kasper K. Iversen[f,g], Frants Pedersen[b,f], Dan E. Høfsten[b,f], Ole B. Pedersen[f,j], Sisse Rye Ostrowski[f,k], Henrik Ullum[f,k,l], Mette N. Svendsen[m], Iben M. Gjødsbøl[m] , Thorarinn Gudnason[n], Daníel F. Guðbjartsson[e], Anna Helgadottir[e], Kristian Hveem[i], Lars V. Køber[b,f], Hilma Holm[e], Kari Stefansson[e,o], Søren Brunak[a,p,✉], and Henning Bundgaard[b,f]

[a] Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Blegdamsvej 3B, DK-2200 Copenhagen, Denmark

[b] Department of Cardiology, The Heart Center, Copenhagen University Hospital, Rigshospitalet, Blegdamsvej 9, DK-2100 Copenhagen, Denmark

[c] Department Obstetrics and Gynecology, Copenhagen University Hospital, Kettegård Alle 30, DK-2650 Hvidovre, Denmark

[d] Methods and Analysis, Statistics Denmark, Sejrøgade 11, DK-2100 Copenhagen, Denmark

[e] deCODE genetics, Sturlugata 8, 102 Reykjavik, Iceland

[f] Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Blegdamsvej 3B, DK-2200 Copenhagen, Denmark

[g] Department of Cardiology, Copenhagen University Hospital, Herlev-Gentofte Hospital, Borgmester Ib Juuls Vej 1, DK-2730 Herlev, Denmark

[h] Department of Clinical and Molecular Medicine, Norwegian University of Science and Technology, 7491 Trondheim, Norway

[i] K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health, Norwegian University of Science and Technology, 7491 Trondheim, Norway

[j] Department of Clinical Immunology, Zealand University Hospital, DK-4600 Køge, Denmark

[k] Department of Clinical Immunology, Copenhagen University Hospital, Rigshospitalet, Blegdamsvej 9, DK-2100 Copenhagen, Denmark.

[l] Statens Serum Institut, Artillerivej 5, DK-2300 Copenhagen, Denmark

[m] Department of Public Health, University of Copenhagen, Øster Farimagsgade 5, DK-1353 Copenhagen, Denmark

[n] Laeknasetrid Cardiology Clinic, Thonglabakka 1, 109 Reykjavík, Iceland

[o] Faculty of Medicine, University of Iceland, Vatnsmyrarvegur 16, Reykjavik 101, Iceland

[p] Copenhagen University Hospital, Rigshospitalet, Blegdamsvej 9, DK-2100 Copenhagen, Denmark

✉ Correspondence: Søren Brunak (soren.brunak@cpr.ku.dk)

Abstract

Background: Current risk prediction models for ischemic heart disease (IHD) use a limited set of established risk factors and are based on classical statistical techniques. Using machine-learning techniques and including a broader panel of features from electronic health records (EHRs) may improve prognostication.

Objectives: Developing and externally validating a neural network-based time-to-event model (PMHnet) for prediction of all-cause mortality in IHD.

Methods: We included 39,746 patients (training: 34,746, test: 5,000) with IHD from the Eastern Danish Heart Registry, who underwent coronary angiography (CAG) between 2006-2016. Clinical and genetic features were extracted from national registries, EHRs, and biobanks. The feature-selection process identified 584 features, including prior diagnosis and procedure codes, laboratory test results, and clinical measurements. Model performance was evaluated using time-dependent AUC (tdAUC) and the Brier score. PMHnet was benchmarked against GRACE Risk Score 2.0 (GRACE2.0), and externally validated using data from Iceland (n=8,287). Feature importance and model explainability were assessed using SHAP analysis.

Findings: On the test set, the tdAUC was 0.88 (95% CI 0.86-0.90, case count, cc=196) at six months, 0.88(0.86-0.90, cc=261) at one year, 0.84(0.82-0.86, cc=395) at three years, and 0.82(0.80-0.84, cc=763) at five years. On the same data, GRACE2.0 had a lower performance: 0.77 (0.73-0.80) at six months, 0.77(0.74-0.80) at one year, and 0.73(0.70-0.75) at three years. PMHnet showed similar performance in the Icelandic data.

Conclusion: PMHnet significantly improved survival prediction in patients with IHD compared to GRACE2.0. Our findings support the use of deep phenotypic data as precision medicine tools in modern healthcare systems.

Keywords: ischemic heart disease, prediction models, survival analysis, artificial intelligence, neural networks, GRACE

Introduction

In patients with ischemic heart disease (IHD), improved clinical application of the wide array of prognostic risk factors and disease markers may inform treatment options for the individual patient[1–3]. For example, the updated version of Global Registry of Acute Coronary Events (GRACE) score (GRACE2.0) received a class IIa recommendation for assessing risk and management of patients with non–ST-elevation myocardial infarction (nSTEMI) in the 2020 European Society of Cardiology (ESC) guidelines[1]. However, GRACE2.0 and other traditional risk scoring schemes in IHD such as Framingham and Thrombolysis in Myocardial Infarction (TIMI) use a limited set of input features (<10) and likely underutilize most data available in modern electronic health records (EHRs)[4–7].

Integrating a richer set of input features could be overcome with machine learning (ML) models such as neural networks. These can capture non-linear interactions without the need for imputation of missing data or expert feature engineering[8,9], leveraging the multitude of heterogeneous healthcare data stored in modern EHRs and national registries in the development of clinical decision support tools.

ML-based approaches have shown promising results for risk-estimation in cardiology with better performance than traditional models[10–13]. In patients with stable IHD, Motwani et al. showed that an ML algorithm combining clinical variables with imaging variables from coronary CT angiography predicted 5-year all-cause mortality better than models using clinical metrics alone[11]. Similarly, Mohammad and colleagues developed and validated a neural network to predict 1-year mortality and re-admission for heart failure after incident myocardial infarction with greater discrimination than the GRACE2.0 score[12]. However, the majority of the secondary risk prediction models based on ML have not used time-to-event analysis. One notable exception is the model

presented by Steele et al, which however have not been externally validated[10]. By not using survival analysis, previous ML-based analyses omit data points with incomplete follow-up (censoring) and thereby effectively prevents a model from distinguishing between "died after a week" and "died after 10 months" which we believe is of obvious clinical interest.

To overcome these limitations, we describe the development and validation of a neural network-based survival model, PMHnet, for predicting all-cause mortality in patients with IHD using 584 different features extracted from population-wide healthcare registries and complete EHRs. We identified several influential features which previously have been omitted from risk prediction of patients with IHD.

Methods

Data foundation

The algorithm PMHnet was developed using a cohort constructed from the Danish National Patient Registry (NPR) and the Eastern Danish Heart Registry (EDHR)[14]. The EDHR contains structured information on all coronary artery angiographies (CAGs) performed in the Capital Region of Denmark and Region Zealand. The cohort was linked to a population-wide EHR database that covers Eastern Denmark from 1st of January 2006 to the 7th of July 2016 (BTH), and genotype data from the Copenhagen Hospital Biobank Cardiovascular Diseases study (CHB-CVD)[15–17]. The BTH dataset fully covered Eastern Denmark (2.6 million patients). Outcomes were obtained from the Central Person Registry and the Danish Register for Causes of Death[18,19]. Data sources were linked using encrypted Danish personal identification numbers[19].

Selection criteria and model development

First, we identified all adult Danish citizens (>18 years of age) in NPR with an ICD-10 code for IHD (I20-I25) who had undergone their first CAG between Jan 1, 2006, and Jun 1, 2016, demonstrating one-, two-, or three-vessel disease (1–3VD) or diffuse atheromatosis (DIF). Vascular disease is here defined as stenosis above 50%[20]. For patients fulfilling these criteria (n=39,746), we used the date of the CAG as the index date and included five years of follow-up. Patients were followed until either death or censoring, whichever came first. Using the hold-out method, the derivation data was randomly divided into a training set (n=34,746) and a test set (n=5,000) used for model development and independent assessment of performance, respectively[21] (Figure 1, Table 1).

For each of the 39,746 patients, we reduced the available features prior to index event (i.e. first CAG between Jan 1, 2006, and Jun 1, 2016) to a smaller set based on prevalence such that e.g., a diagnosis code could be found in at least 5% of the training set. The final set of 584 features was separated into five different categories: *ClinicalOne* (8 features), *ClinicalTwo* (15 features), *Diagnoses (322 features)*, *Procedures (154 features)*, and *Biochemical* (85 features) (Table 2). *ClinicalOne* included the same eight input features as used by GRACE2.0 and *ClinicalTwo* had 14 additional clinical features (Table 2) that were selected based on availability. Features were defined using data recorded prior to the index date, except for creatinine, cardiac biomarkers for ischemic heart disease, and blood pressure where high missingness led us to allow measurements obtained after the CAG in cases of missingness (7-day threshold for cardiac biomarkers, and 21-day threshold for the others). *Diagnoses* included ICD-10 codes registered in NPR and similarly, *Procedures* consisted of procedure codes (surgery and examinations such as X-rays) registered in NPR. *Biochemical* contained results of in-hospital blood tests. Additional details on feature extraction, missingness, pre-processing, encoding, and categories can be found in supplementary methods. The amounts of missingness across the training and test set have been tabulated in Table S1. Feature categories and encodings are available in the appendices. Missing values were left missing and encoded as such in PMHnet; meaning that for categorial variables all values were zero in case of missingness and for continuous variables missing values were assigned the mean variable (details are available in the Supplementary Material).

External validation data from Iceland

For the external validation cohort, we identified Icelandic adults who had undergone CAG at the only interventional cardiology center in Iceland, Landspitali– The National University Hospital in Reykjavík[22]. We obtained data collected prospectively between January 1, 2007, and December

31, 2017. Information on ICD-10 diagnoses and procedure codes were aggregated from the Landspitali, from registers kept by the Directorate of Health: The Register of Primary Health Contacts, the Register of Contacts with Medical Specialists in Private Practice and the Causes of Death Register, as well as at recruitment for deCODE studies. Biochemical assay measurements were obtained from the three largest clinical laboratories in Iceland, with measurements performed at: (i) Landspitali; (ii) The Laboratory in Mjodd, Reykjavik, Iceland; and (iii) Akureyri Hospital, the regional hospital in North Iceland.

Polygenic risk scores

Polygenic risk scores (PRSs) for patients with genotypes available through the CHB-CVD[17,23] (37.4% of the cohort) were calculated using the LDpred2 framework, implemented in the R package bigsnpr (v1.5.2) with R version 3.5.052[24]. PRSs were calculated based on GWAS summary statistics data from 19 traits relevant for cardiometabolic health, obtained from 17 GWAS meta-analyses. List of meta-analyses and details on the PRS calculations is included in the Supplementary Material.

Machine learning model architecture and development

To model time-to-event data and allow for censoring, we used the generic discrete-time survival model for neural networks described by Gensheimer and Narasimhan[25]. In this model, follow-up time is divided into a fixed number of intervals and the model estimates a conditional hazard for each interval, i.e., the probability of dying in that time interval given that the patient is still alive at the end of the preceding interval. PMHnet uses 30 intervals separated in time such that event times in the training data are evenly distributed across all intervals. To obtain predictions between breakpoints in the discretization grid, we assumed that the probability density function was

constant in each time interval, and we thus interpolated using a piecewise linear function[26]. The implementation applied the PyTorch machine-learning framework using the authors' Keras version as a reference[25].

We used a feed-forward neural network and tested various hyperparameters. The output layer was a fully connected sigmoid activated layer that outputs conditional hazards for each of the 30 different time points. We added dropout to each of the hidden layers to regularize the network and prevent over-fitting. The number of layers, neurons, learning-rate, and dropout rate for each layer were fine-tuned through hyperparameter optimization using the Optuna optimization framework, with a five-fold cross validation[27]. The hyperparameter search space and the best trial for the complete model is included in table S3. The neural networks were trained using stochastic gradient descent, with a constant learning rate, to minimize the negative log-likelihood.

Model evaluation and validation

Using the hold-out test set and the external validation data from Iceland, performance of PMHnet was evaluated through assessment of both model discrimination and calibration. We used time-dependent area under the receiver operating characteristic curve (tdAUC) as the main measure of discrimination, but also calculated the Brier score that can be used to assess both discrimination and calibration[28–30]. Calibration was also analyzed graphically by comparing the predicted risks with the estimated actual risks[28]. The `Score` function from the riskRegression R package was used to compute performance measures and compare models. For comparisons between two competing models, the `Score` function gives p-values that correspond to Wald tests on the standard errors obtained using an estimate of the influence functions following Blanche et al. [30].

To benchmark PMHnet we calculated the GRACE2.0 for all patients in the hold-out test set, using the GRACE2.0 webtool. We extracted the javascript source code for the GRACE2.0 webtool using the developer tools in Google Chrome. The javascript code was then manually converted to an R package for automatized computation of GRACE2.0 on the entire cohort. The eight variables used in GRACE2.0 were available for 51.4% of the cohort. Since GRACE2.0 does not allow for missing features, we imputed missing variables using the `missForest` R package[31]. Imputed values were only used for calculating the GRACE2.0 score. In addition to the conventional GRACE2.0 score, we re-fitted the GRACE2.0 score to our training data using the PMHnet architecture. This corresponds to `model_2` in Figure 4 that only uses the features from *ClinicalOne* as its input.

For independent validation of PMHnet we used, as described above, Icelandic EHR data from 8,287 patients. Of the 584 features identified in the Danish derivation cohort, we found matching data for 404 features in the Icelandic data. A down-scaled model was re-trained on the Danish training set to make comparisons.

Explainability and effect of missing features

To investigate the impact of different features on model predictions and to provide model explanations, we calculated Shapley additive explanation (SHAP) values for all features and patients in the training set using the `SHAP` python package[32]. In the model explanations, a negative SHAP value for a given feature means that the feature pulls the prediction towards mortality and vice versa for positive SHAP values relative to the median prediction. The magnitude of the SHAP value is percentage points.

To assess how resilient the model was in the event of missing data, missingness was introduced in the test data by replacing all values of a given feature with the median value of that feature in the

training data. The predictions were then compared with the predictions of PMHnet. Resiliency was then quantified using change in tdAUC (discrimination) and Brier scores (calibration), where the predictions of PMHnet were compared to that of model with artificially introduced values (a total of 584 comparisons).

As the network was not trained on the Icelandic data before external validation, no SHAP analysis was performed on the predictions on the Icelandic data.

Statistical analysis

Categorical features are reported as counts (%) and continuous features as mean [95% CI], 95% CIs are obtained from standard deviations or through bootstrapping. Time-dependent AUCs and Brier scores were calculated using the `riskRegression` R-package[28]. Likewise, model comparisons were obtained from the same package. All statistical analyses and visualizations were performed using R version 4.1.

Data access and ethics approvals

The study was approved by The National Ethics Committee (1708829, 'Genetics of CVD'—a genome-wide association study on repository samples from CHB), The Danish Data Protection Agency (ref: 514-0255/18-3000, 514-0254/18-3000, SUND-2016-50), The Danish Health Data Authority (ref: FSEID-00003724 and FSEID-00003092), and The Danish Patient Safety Authority (3-3013-1731/1/). Danish personal identifiers were pseudonymised prior to any analysis.

The study was approved by the Data Protection Authority of Iceland and the National Bioethics Committee of Iceland (VSN-15-114). Icelandic participants that donated biological samples

provided informed consent. Personal identities of the participants were encrypted with a third-party system provided by the Data Protection Authority of Iceland.

Study design, methods, and results were reported in agreement with the TRIPOD statement[33,34] and following the STROBE recommendations[35].

Funding

Novo Nordisk Foundation (grant agreements: NNF14CC0001 and NNF17OC0027594) – Hellerup, Denmark; NordForsk (*PM Heart*; grant agreement: 90580) – Oslo, Norge; and the Innovation Foundation (*BigTempHealth*; grant agreement: 5153-00002B) – Aarhus, Denmark.

Results

The derivation cohort of 39,746 Danish patients with IHD were randomly subdivided into a training (N=34,746) and a test set (N=5,000) (Table 1). At inclusion the patients mean age (95%-CIs) was 66.0 years [65.7; 66.4] (67.3% males) in the training set and 66.2 years [66.0;66.3] (68.2% males) in the test set. The distribution of the degree of coronary artery disease was similar in the two groups (distributions of patients presenting with one-, two-, or three-vessel disease or diffuse atherosclerosis, respectively).

The Kaplan-Meier estimate of five-year survival (all-cause) was 81.8% [81.4; 82.2] for the training set and 82.5 [81.3; 83.6] for the test set (Figure S1, Table S2). The restricted mean follow-up time was 1,635 days (±2.58) for the training set and 1,635 days (±6.49) for the test set.

PMHnet model predictions

In the internal validation using the hold-out test set, the complete PMHnet model had tdAUCs of 0.88 [0.86; 0.90] at six months (case count, cc=196): , 0.88 [0.86; 0.90] at one year (cc=261), 0.84 [0.82; 0.86] at three years (cc=395) (Figure 2), and 0.82 [0.80; 0.84] at five years (cc=763). In comparison, the corresponding values for the conventional GRACE2.0 score on the same dataset were 0.77 [0.74; 0.80] at six months, 0.77 [0.74; 0.80] at one year, and 0.73 [0.71; 0.75] at three years. For the re-fitted GRACE2.0 score, tdAUCs were 0.79 [0.76; 0.83] at six months, 0.78 [0.75; 0.81] at one year, and 0.76 [0.74; 0.78] at three years. Since GRACE2.0 features had to be imputed for 48.6% of the population, we also evaluated the performance on the subset without missingness, which were largely the same (Figure 3, dashed line). GRACE2.0 is not designed for providing predictions after three years and was therefore not evaluated beyond that time-point. The

difference in tdAUCs between PMHnet and either of the two GRACE2.0 models was significant at each of the three prediction horizons (Table 3).

As an additional visual test of discrimination, we constructed five different risk strata using the 5-year predicted survival (defined as 90%, 75%, 50%, and 25%) and examined the observed survival (Figure 4), which showed good separation between the five strata. PMHnet was found to be well-calibrated as seen from Figure 2B and the calculated Brier scores of 3.2% [2.8; 3.6] at six months, 4.1% [3.7; 4.5] at one year, 7.6% [7.1; 8.1] at three years, and 11.3% [10.5; 12.0] at five years.

In comparison, GRACE2.0 had Brier scores of 3.7% [3.3; 4.1] at six months, 4.9% [4.4; 5.4] at one year, 9.9% [9.3; 10.5] at three years. Re-fitting GRACE2.0 with the PMHnet architecture considerably improved the calibration as evident from the calibration curve which was comparable to that of PMHnet (Figure 2B). The differences in three-year predicted risk between PMHnet and the two GRACE2.0 scores for all patients in the test set are shown in Figure S2.

External validation of PMHnet

For external validation the cohort of 8,287 patients from Iceland was used (Table 1, *validation set*). Due to data availability a modified, and down-scaled version of PMHnet was used (404 features) on the Icelandic data. The tdAUCs were 0.87 [0.84;0.90] at six months, 0.84 [0.81;0.87] at one year, and 0.81 [0.79;0.83] at three years. In comparison, the performance of the down-scaled version on the Danish (internal, hold-out) test set was 0.87 [0.85;0.90] at six months, 0.87 [0.85;0.89] at one year, and 0.82 [0.80;0.85] at three years. The predictive performances were highly concordant in the Icelandic data, and the model maintained good calibration for the six months and one year predictions, but was found to be miscalibrated in the three year predictions (Figure S3).

Influence of the different input feature categories on the PMHnet predictionTo assess the importance of the different input feature categories systematically, we trained five intermediate survival models using the PMHnet architecture (Figure 3). For example, *model-1* used diagnosis codes as input features only, *model-2* used the clinical variables known from GRACE only. The other intermediate versions were trained using different combinations of the feature categories. Figure 3 shows the model discrimination at the three different prediction horizons for the six different models fitted using the PMHnet architecture. The performance of the intermediate model based on diagnosis codes only (*model-1*) was similar to the performance of the re-fitted GRACE2.0 model (*model-2*). Interestingly, combining diagnosis codes with the GRACE2.0 clinical features (*Diagnoses* + *ClinicalOne*) in *model-3*, there was an overall increase in AUC at three years, which would suggest a synergistic effect. With the addition of the *ClinicalTwo*-data, the ΔtdAUC between *model-3* and *model-4* was 3.4e-2 [1.8e-2; 5.0e-2] at six months, 3.2e-2 [1.8e-2; 4.7e-2] at one year, and 1.3e-2 [0.3e-2; 2.3e-2] at three years. Similarly, adding *Biochemical* to the input features in *model-5* associated with significant ΔtdAUCs of 2.1e-2 [0.2e-2; 3.6e-2] at six months, 1.4e-2 [0.2e-2; 2.7e-2] at one year, and a non-significant ΔtdAUC of 0.9e-2 [-0.1e-2; 1.9e-2] at three years. The gain in discrimination from *model-5* to the complete *model-6* (PMHnet) by adding *Procedures* to the input features was only significant at the one-year prediction horizon.

Testing inclusion of polygenic risk scores in PMHnet

For 37.4% of the cohort, we had genotype information available and used that to calculate 19 different polygenic risk scores (PRS) that all related to different cardiometabolic traits. Limiting both the training set and the test set to only individuals with genotype data (37.4%), we tested adding the PRS scores to *model-1* (*Diagnoses*), *model-2* (*Diagnoses* + *ClinicalOne*), and model-6

(*All Features*) and evaluated the model performance (Figure S4). Addition of the PRS did not significantly improve either model discrimination at any time-point (Figure S4A).

Explainability analysis using SHAP

We performed SHAP-analyses of the five-year model predictions to quantify the impact of the different features on PMHnet predictions. SHAP is a technique rooted in cooperative game theory that provides an estimate of feature impact on the model output. It quantifies the contribution of each feature to the prediction outcome, allowing for a better understanding of feature importance and model behavior. By considering the interactions and dependencies between features, SHAP analysis provides insights into the specific factors influencing the model's decision-making process and aids in identifying key drivers and relationships within the model[36]. Across the five different feature categories, biochemical test results and diagnosis codes were most impactful, while the category *ClinicalOne* was least impactful (Figure 5A). The most impactful diagnosis code was chronic obstructive pulmonary disease (ICD10: J44) and thus ranked higher than classical risk factors for IHD, such as type 2 diabetes. At the feature level, age, the number of affected vessels (1–3VD, or DIF), and smoking were on-average the most predictive features. The top-25 features in terms of average model impact (average magnitude of SHAP-value), included ten features from *Biochemical*, nine from *ClinicalTwo*, three from *ClinicalOne*, two from *Diagnoses*, and one from *Procedures* (Figure 5A). To further examine the impact on model prediction for the two most impactful features, number of affected *vessels* and *age*, we constructed SHAP dependence plots (Figure 5B)[37]. For *age* we observed non surprisingly that higher age pulled the prediction towards non-survival, and that age was estimated to add/remove anywhere between -25 and 20 percent points to the predicted 5-year survival. Similarly for *vessels*, more affected vessels impacted the predicted survival negatively. For both features, we noted that identical feature values not always

impacted the survival to the same extent. This vertical dispersion in both plots represent interaction effects with the other included features[37]. Although our SHAP analyses reveal that many features have lower impact relative to the most impactful features, the aggregated sum of the many low-impact features (560 lowest) outweighs many of the well-known risk-factors (25 highest). For this reason, we did not attempt to limit the number of included features. Extending the analysis of feature-level model impact to the rest of the top-25 features, we constructed a summary plot of the SHAP-values that shows the distribution of model impacts across feature values (Figure 6). Interestingly, we see that for several features the knowledge that the feature is missing has a pronounced impact on the model predictions. As an example, we see that if an electrocardiogram has not been obtained (or recorded in the database) then the model predictions are pulled towards non-survival.

Patient-level feature importance

Finally, we also generated individual explanations for three example patients in the test set and show the SHAP values for the nine most impactful features (Figure 7). The patients were randomly selected from the subsets of patients with a prediction in the intervals (0.25; 0.5], (0.5, 0.75], and (0.75; 1]. For *patient 1*, with the worst 5-year prognosis, *age* was not among the nine most impactful features. Instead, the explainability algorithm highlights diagnosis codes and biochemical values. For *patient 2* with a 5-year predicted risk of 73%, the most impactful feature was age (78 years), which pulled the prediction towards mortality. The impact of age was however largely cancelled by *rest*, which constitutes the aggregated sum of all the features not among the nine most predictive. For *patient 3*, the feature *age* was again highlighted as the most impactful, but in this case, it impacted the prognosis positively. Apart for a history of cigarette smoking, none of the highlighted features had negative impact on the prognosis.

Discussion

In this study, we developed a feature-rich neural network-based survival algorithm, PMHnet, for prediction of all-cause mortality in patients with IHD using data from 34,746 Danish patients. With the aim of providing predictions that can be used to guide treatment and care, the model was developed to operate with an index date immediately after the diagnosis-confirming coronary angiography and with a prediction horizon of five-years. The model was tested using data from 5,000 Danish patients and externally validated using data from 8,288 Icelandic patients. We found that PMHnet had excellent discrimination with tdAUCs ranging from 0.88 at six months to 0.82 at five years. Similar results were found on the external Icelandic data, which confirms that the model and its deep feature foundation generalized well to novel patients and a different healthcare setting. Evaluated on the Danish data, we found the model to be well-calibrated with predicted probabilities accurately reflecting the observed proportions, also in different risk strata.

To aid the clinical interpretation of model predictions, we used SHAP-values to highlight the most impactful features and to explain how the different features affect the prediction for the individual patient. Model explainability is important for evaluating the model output and is paramount for the clinical adaption of any ML-model[38], including focus on features that are clinically actionable, i.e. modifiable versus non-modifiable factors.

Compared to the GRACE2.0 score, which is widely considered the gold-standard risk-stratification tool in current clinical use for predicting mortality after acute coronary syndrome (ACS)[39], PMHnet had superior discrimination and calibration. However, it is important to note that there are differences between the intended patient populations for the two models. The GRACE2.0 score has been developed using a derivation cohort of patients with STEMI, n-STEMI, and unstable angina with time of initial admission as time-zero[5]. In contrast, our study used time at coronary

angiography as its baseline and was applied to all patients with IHD and coronary artery pathology ranging from diffuse atheromatosis to three-vessel disease. The validity of GRACE for a cohort with such characteristics has not been established. To provide a direct comparison of the two algorithms, we used the GRACE2.0 features and re-fitted the GRACE2.0 using our training data. The re-fitted GRACE2.0 score had the same model discrimination and better model calibration than the original version (Fig. 2), but still had inferior prediction compared to PMHnet. Moreover, PMHnet includes assessment of personalized risk predictions (Figure 7), meaning that the most impactful prognostic features might vary from patient to patient. It remains important to stress, that the model displays correlations only, but we argue that individualized risk predictions are of increasing importance in healthcare systems of increasing complexity with more and more patients surviving many years with a burden of multiple chronic diseases. To support safe and effective treatment is such healthcare systems, knowledge of modifiable as well as unmodifiable risk factors for disease progression is paramount to decide the better treatment option on a case to case basis.

The above observations are in good agreement with the current literature on ML-based secondary risk-stratification models, which have found machine learning models to offer better performance than simpler, existing scores in current clinical use[10–13,40,41]. A transition towards feature-rich models that utilize more of the available data can therefore be an advantage. The 584 features used in our final model are neither bespoke nor specifically collected for this decision-support application, and instead represent clinical information gathered during routine work-up, management and treatment of patients. Using models that rely on several hundred features means that the current practice of manually entering data into a webtool becomes very impractical[42]. Instead, novel risk-prediction models need to be fully integrated in the EHR systems such that data can be automatically pulled and integrated. In the present study, we demonstrate that PMHnet can

be scaled and used in a different healthcare system and provide evidence that the results are generalizable. However, clinical implementation of PMHnet is beyond the scope of the present study.

Among previously published machine-learning models for secondary prediction in IHD, our study has several strengths. Firstly, whereas almost all the existing literature uses binary classification, the use of survival or time-to-event models is less explored. One notable exception is the model reported by Steele et al.[10] which employed random survival forests and elastic net Cox regression to predict mortality in patients (n=80,000) with a history of coronary artery disease. One of the defining characteristics of survival models is the ability to handle censored data[43], which in other types of models would have been left out. In addition, survival models can distinguish between "died after a week" and "died after 10 months" which e.g., would be identical in a 1-year binary classification model. To the best of our knowledge, we are the first to use neural network-based survival models in this context. Secondly, we externally validated our model using data from a different country. Most models in the literature were not suboptimal externally validated[10,11,40,41], and among those that have been, only two used data from a different country[12,13]. Demonstrating that a given model can accurately predict beyond borders is crucial for generalizability. Thirdly, our model can predict all-cause mortality up to five years after the index date. Probably owing to the fact that survival models have not been used, most models in the domain exclusively operate with a prediction horizon of one or two years[12,13,40,41]. Longer prediction horizons might be necessary for long-term disease management.

Although the features we have used represent data collected from a typical clinical workflow and therefore should be generally applicable, inter-regional and inter-national differences in clinical practice may affect what data is available and when. This is for instance exemplified by differences

in diagnostic work-up, or timely access to coronary angiography, but may also relate to differences in access to previous medical data. Such aspects could affect the generalizability of our included features, but with internationally accepted treatment guidelines the differences should be minimal. Possible solutions are to reduce slightly the complexity of models to better match the intersection of features available across countries/regions and/or re-fitting and possibly retraining the model each time it is deployed in a new setting. In our external validation we used data from Iceland, which in an international perspective is very similar to Denmark when comparing healthcare systems[44,45]. For that reason, we downscaled the model slightly to account for data availability, but re-training on Icelandic data was not deemed necessary. Using the Icelandic data, we found the model on average to be well-calibrated, but from visual inspection of the calibration curve found evidence of miscalibration as the model was found to overestimate the risk for some patients which would suggest that further adjustment of the model is needed were it to be deployed in Iceland. However, since miscalibration does not affect the accurate risk-stratification of patients, we did not pursue that issue further in this study. For future studies, we note that techniques such as Platt-scaling or isotonic regression could be used to remedy calibration-issues[46]. Overall, we argue that it is an inherent strength of the study that no strict feature selection criteria were applied. First, neural networks are by design capable of handling correlated data and second, an increasing data-rich healthcare system demands a better usage of the data being generated. By means of the explainability analysis, the study succeeds in showcasing that dependencies are present. We acknowledge that an inherent limitation of the study is that it only unmasks correlations, and that the explainability step indeed is influenced by the dependencies.

The dynamic nature of clinical environments can lead to deterioration of model performance[47,48]. Advances in treatment and diagnosis mean that the baseline risk of patients with ischemic heart

disease could change over time, which in turn would lead to a drift of model calibration. As an example, the logistic EuroSCORE[49], a pan-European risk-stratification model for cardiac surgery published in 2003 was since its inception gradually found to overestimate mortality[50]. The EuroSCORE has since been replaced by EuroSCORE II which for the time being has remedied these issues[51]. This type of systematic decline of model performance has important implications for our model as well and necessitates that the model performance is continuously monitored to ensure acceptable up-to-date performance. The need for real-time monitoring of model performance is another strong argument for risk-stratification tools to be tightly integrated within EHR systems.

As a secondary analysis in this study, we tested adding a panel of polygenic risk scores to the input features in PMHnet and assessed how they might impact model performance. The 19 different genetic risk scores were included based on being related to cardiometabolic health and covered traits such as *blood pressure*, and *total cholesterol*, but also included *heart failure* and *acute myocardial infarction*. Where for example the *acute myocardial infarction* risk score is developed for primary risk prediction, i.e., disease development, our model is concerned with secondary risk prediction, i.e., modelling risk for those who already have the disease. It is known that CAD PRS associate with events in both primary and secondary event populations, however, these PRSs are PRSs of prevalent diseases, not mortality. Whether or not a primary risk score is useful in a secondary risk context is clear upfront, but as parts of the underlying disease process are known to be shared between the two, we hypothesized that the score could be used in our context. Focusing the analysis only on the subset of patients for which we could obtain genotypes (n=13,449, 37.4%), we found no significant difference in performance after adding the PRS scores to either *model-1* (*Diagnoses*) and *model-2* (*Diagnoses + ClinicalOne*). As the 585 features include the prior disease

history recorded over more than 25 years, our interpretation is that a "realized" life-course disease trajectory is more informative than the germline risk that can be calculated from the genotype. Moreover, the disease trajectory also holds information on life-course exposures and may therefore also include exposure information that quite implicitly is related to genetic data only. The version of PMHnet trained on the disease history only, with and without genetics, indicates that there is little gain from the PRS tested in this case. As we did not have genetic data for the full cohort, we cannot exclude that our interpretation is affected by this aspect.

Prospective studies are needed to ascertain how feature-rich risk-stratification methods can be used to alter, guide, and hopefully improve treatment. The ability to accurately predict high-risk is useful for identifying patients that may benefit from more extensive treatment and more frequent visits at the hospital. In contrast, accurate identification of low-risk patients may potentially be used to limit work-up, extent of pharmacological treatment and follow-up content and intensity and thereby prevent potential harmful overtreatment. Striking the correct balance between over- and undertreatment can contribute to the advancement of precision medicine, and here a well-calibrated and highly discriminative risk-prediction model can serve as an important tool.

In routine cardiology, a multitude of diagnostic, prognostic, and treatment-related scores are applied. However, all the presently applied scores are based on a very limited number of features. The present findings indicate a significantly added value of applying far more features and of introducing machine-learning. Thus, precision treatments in cardiology may benefit from using more features and machine-learning to replace the present scores.

Data availability statement

Due to national and EU regulations, the datasets used for model development and validation cannot be made publicly available. Research groups with access to secure and dedicated computing environments can request access to the source data registries via application to the Danish Health Data Authority.

Conflicts of interest

Søren Brunak reports ownerships in Intomics, Hoba Therapeutics, Novo Nordisk, Lundbeck, and ALK; and managing board memberships in Proscion and Intomics. Henning Bundgaard reports ownership in Novo Nordisk and has received lecture fees from Amgen, BMS, MSD and Sanofi. The following co-authors are employed by deCODE genetics/Amgen, Inc: Vinicius Tragante, Daníel F. Guðbjartsson, Anna Helgadottir, Hilma Holm, and Kari Stefansson.

Acknowledgements

We acknowledge Mette Hartlev, Franziska Walder, Mette Gørtz, and Katharina Ó Cathaoir for helpful comments and discussions in the writing of this manuscript.

References

1. Collet, J.-P. *et al.* 2020 ESC Guidelines for the management of acute coronary syndromes in patients presenting without persistent ST-segment elevation: The Task Force for the management of acute coronary syndromes in patients presenting without persistent ST-segment elevation of the European Society of Cardiology (ESC). *Eur. Heart J.* **42**, 1289–1367 (2021).
2. Knuuti, J. *et al.* 2019 ESC Guidelines for the diagnosis and management of chronic coronary syndromes: The Task Force for the diagnosis and management of chronic coronary syndromes of the European Society of Cardiology (ESC). *Eur. Heart J.* **41**, 407–477 (2020).
3. Steg, Ph. G. *et al.* ESC Guidelines for the management of acute myocardial infarction in patients presenting with ST-segment elevation. *Eur. Heart J.* **33**, 2569–2619 (2012).
4. Wilson, P. W. F. *et al.* Prediction of Coronary Heart Disease Using Risk Factor Categories. *Circulation* **97**, 1837–1847 (1998).
5. Fox, K. A. A. *et al.* Should patients with acute coronary disease be stratified for management according to their risk? Derivation, external validation and outcomes using the updated GRACE risk score. *BMJ Open* **4**, e004425 (2014).
6. Hung, J. *et al.* Performance of the GRACE 2.0 score in patients with type 1 and type 2 myocardial infarction. *Eur. Heart J.* **42**, 2552–2561 (2020).
7. Antman, E. M. *et al.* The TIMI Risk Score for Unstable Angina/Non–ST Elevation MI. *JAMA* **284**, 835 (2000).
8. Rajkomar, A., Dean, J. & Kohane, I. Machine Learning in Medicine. *N. Engl. J. Med.* **380**, 1347–1358 (2019).

9. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56 (2019).
10. Steele, A. J., Denaxas, S. C., Shah, A. D., Hemingway, H. & Luscombe, N. M. Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. *PLOS ONE* **13**, e0202344 (2018).
11. Motwani, M. *et al.* Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis. *Eur. Heart J.* ehw188 (2016) doi:10.1093/eurheartj/ehw188.
12. Mohammad, M. A. *et al.* Development and validation of an artificial neural network algorithm to predict mortality and admission to hospital for heart failure after myocardial infarction: a nationwide population-based study. *Lancet Digit. Health* **4**, e37–e45 (2022).
13. D'Ascenzo, F. *et al.* Machine learning-based prediction of adverse events following an acute coronary syndrome (PRAISE): a modelling study of pooled datasets. *The Lancet* **397**, 199–207 (2021).
14. Özcan, C. *et al.* The Danish Heart Registry. *Clin. Epidemiol.* **8**, 503–508 (2016).
15. Schmidt, M. *et al.* The Danish National Patient Registry: a review of content, data quality, and research potential. *Clin. Epidemiol.* 449 (2015) doi:10.2147/clep.s91125.
16. Nielsen, A. B. *et al.* Survival prediction in intensive-care units based on aggregation of long-term disease history and acute physiology: a retrospective study of the Danish National Patient Registry and electronic patient records. *Lancet Digit. Health* **1**, e78–e89 (2019).
17. Sørensen, E. *et al.* Data Resource Profile: The Copenhagen Hospital Biobank (CHB). *Int. J. Epidemiol.* **50**, 719–720e (2020).

18. Helweg-Larsen, K. The Danish Register of Causes of Death. *Scand. J. Public Health* **39**, 26–29 (2011).
19. Schmidt, M., Pedersen, L. & Sørensen, H. T. The Danish Civil Registration System as a tool in epidemiology. *Eur. J. Epidemiol.* **29**, 541–549 (2014).
20. Harris, P. J. *et al.* The prognostic significance of 50% coronary stenosis in medically treated patients with coronary artery disease. *Circulation* **62**, 240–248 (1980).
21. Arlot, S. & Celisse, A. A survey of cross-validation procedures for model selection. *Stat. Surv.* **4**, (2010).
22. Björnsson, E. *et al.* Association of Genetically Predicted Lipid Levels With the Extent of Coronary Atherosclerosis in Icelandic Adults. *JAMA Cardiol.* **5**, 13–20 (2020).
23. Laursen, I. H. *et al.* Cohort profile: Copenhagen Hospital Biobank - Cardiovascular Disease Cohort (CHB-CVDC): Construction of a large-scale genetic cohort to facilitate a better understanding of heart diseases. *BMJ Open* **11**, e049709 (2021).
24. Privé, F., Arbel, J. & Vilhjálmsson, B. J. LDpred2: better, faster, stronger. *Bioinforma. Oxf. Engl.* **36**, 5424–5431 (2020).
25. Gensheimer, M. F. & Narasimhan, B. A scalable discrete-time survival model for neural networks. *PeerJ* **7**, e6257 (2019).
26. Kvamme, H. & Borgan, Ø. Continuous and discrete-time survival prediction with neural networks. *Lifetime Data Anal.* **27**, 710–736 (2021).
27. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. *Optuna: A Next-generation Hyperparameter Optimization Framework*. https://arxiv.org/abs/1907.10902 (2019).
28. Gerds, T. A. & Kattan, M. W. *Medical risk prediction models: with ties to machine learning*. (CRC Press, 2021).

29. Schumacher, M., Graf, E. & Gerds, T. How to Assess Prognostic Models for Survival Data: A Case Study in Oncology. *Methods Inf. Med.* **42**, 564–571 (2003).
30. Blanche, P. *et al.* Quantifying and comparing dynamic predictive accuracy of joint models for longitudinal marker and time-to-event in presence of censoring and competing risks. *Biometrics* **71**, 102–113 (2015).
31. Stekhoven, D. J. & Buhlmann, P. MissForest--non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28**, 112–118 (2011).
32. Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, 56–67 (2020).
33. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. M. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD). *Circulation* **131**, 211–219 (2015).
34. Collins, G. S. & Moons, K. G. M. Reporting of artificial intelligence prediction models. *The Lancet* **393**, 1577–1579 (2019).
35. von Elm, E. *et al.* The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *J. Clin. Epidemiol.* **61**, 344–349 (2008).
36. Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. in *Advances in Neural Information Processing Systems 30* (eds. Guyon, I. et al.) 4765–4774 (Curran Associates, Inc., 2017).
37. Lundberg, S. M., Erion, G. G. & Lee, S.-I. Consistent Individualized Feature Attribution for Tree Ensembles. Preprint at https://doi.org/10.48550/arXiv.1802.03888 (2019).
38. Kundu, S. AI in medicine must be explainable. *Nat. Med.* **27**, 1328–1328 (2021).

39. D'Ascenzo, F. *et al.* TIMI, GRACE and alternative risk scores in Acute Coronary Syndromes: A meta-analysis of 40 derivation studies on 216,552 patients and of 42 validation studies on 31,625 patients. *Contemp. Clin. Trials* **33**, 507–514 (2012).
40. Kwon, J. *et al.* Deep-learning-based risk stratification for mortality of patients with acute myocardial infarction. *PLOS ONE* **14**, e0224502 (2019).
41. Wallert, J., Tomasoni, M., Madison, G. & Held, C. Predicting two-year survival versus non-survival after first myocardial infarction using machine learning and Swedish national register data. *BMC Med. Inform. Decis. Mak.* **17**, 99 (2017).
42. Sharma, V. *et al.* Adoption of clinical risk prediction tools is limited by a lack of integration with electronic health records. *BMJ Health Care Inform.* **28**, e100253 (2021).
43. George, B., Seals, S. & Aban, I. Survival analysis and regression models. *J. Nucl. Cardiol. Off. Publ. Am. Soc. Nucl. Cardiol.* **21**, 686–694 (2014).
44. Einhorn, E. S. Nordic Health Care Systems: Recent Reforms and Current Policy Challenges. *Scand. Stud.* **84**, 106–108 (2012).
45. Kristiansen, I. S. & Pedersen, K. M. [Health care systems in the Nordic countries--more similarities than differences?]. *Tidsskr. Den Nor. Laegeforening Tidsskr. Prakt. Med. Ny Raekke* **120**, 2023–2029 (2000).
46. Niculescu-Mizil, A. & Caruana, R. Predicting good probabilities with supervised learning. in *Proceedings of the 22nd international conference on Machine learning* 625–632 (Association for Computing Machinery, 2005). doi:10.1145/1102351.1102430.
47. Davis, S. E., Lasko, T. A., Chen, G. & Matheny, M. E. Calibration Drift Among Regression and Machine Learning Models for Hospital Mortality. *AMIA. Annu. Symp. Proc.* **2017**, 625–634 (2018).

48. Jenkins, D. A., Sperrin, M., Martin, G. P. & Peek, N. Dynamic models to predict health outcomes: current status and methodological challenges. *Diagn. Progn. Res.* **2**, 23 (2018).
49. Roques, F., Michel, P., Goldstone, A. R. & Nashef, S. a. M. The logistic EuroSCORE. *Eur. Heart J.* **24**, 882–883 (2003).
50. Hickey, G. L. *et al.* Dynamic trends in cardiac surgery: why the logistic EuroSCORE is no longer suitable for contemporary cardiac surgery and implications for future risk models. *Eur. J. Cardiothorac. Surg.* **43**, 1146–1152 (2013).
51. Nashef, S. A. M. *et al.* EuroSCORE II. *Eur. J. Cardiothorac. Surg.* **41**, 734–745 (2012).

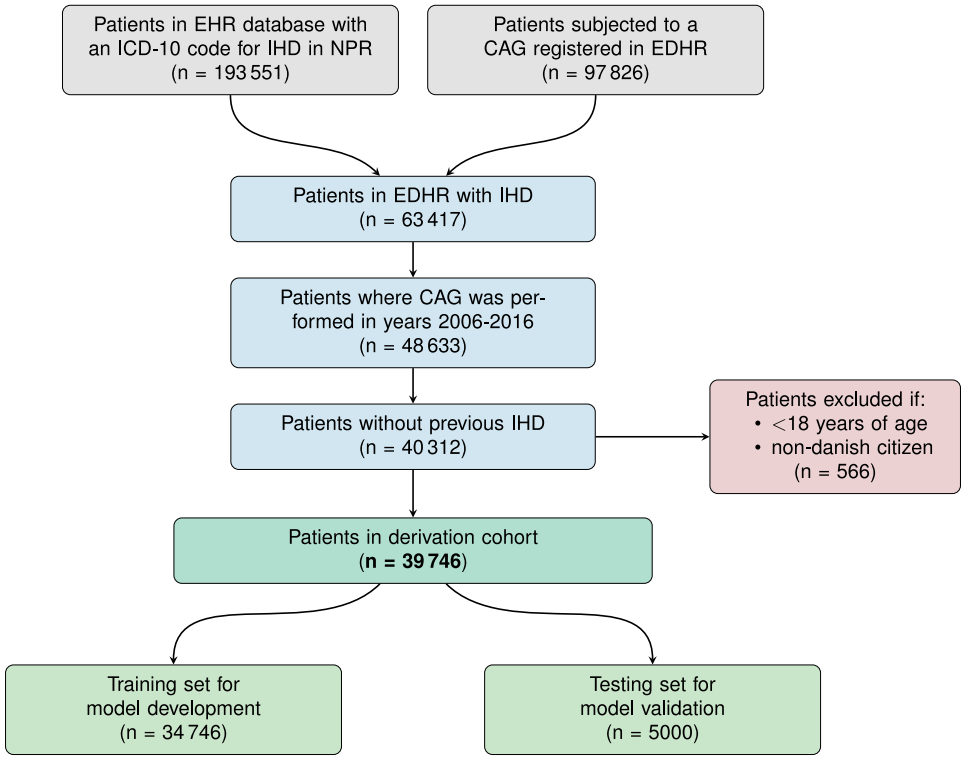


Figure 1: *Flowchart of inclusion of patient with ischemic heart disease.* Flowchart showing how the derivation cohort was identified based on data from NPR and EDHR. CAG: Coronary arteriography. EHRs: Electronic health records. EDHR: Eastern Danish Heart Registry. ICD-10: International classification of diseases, 10th revision. IHD: Ischemic heart disease. NPR: Danish National patient registry.

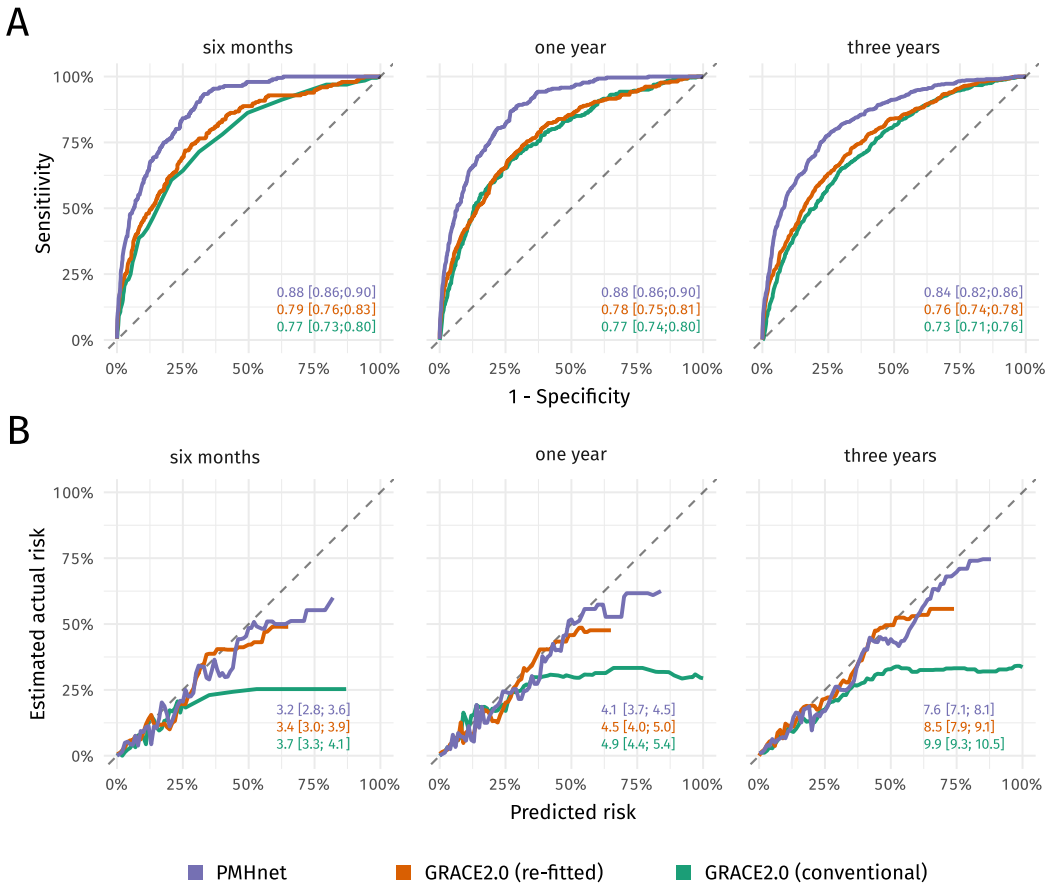


Figure 2: *Model performance of PMHnet and the GRACE2.0 score.* A) Time-dependent receiver operating characteristics (ROC) curves at three different prediction horizons for PMHnet, GRACE2.0 (re-fitted), and GRACE2.0 (conventional). Labels show the time-dependent area under the ROC curves (AUC). GRACE2.0 (re-fitted) is a model that uses the GRACE2.0 input features but uses the PMHnet architecture and is trained using our training data. B) Calibration curves showing the relation between predicted risk and the estimated actual risk. Labels show the Brier score for each of the three models. Lower scores are associated with better calibration and discrimination of predictions.

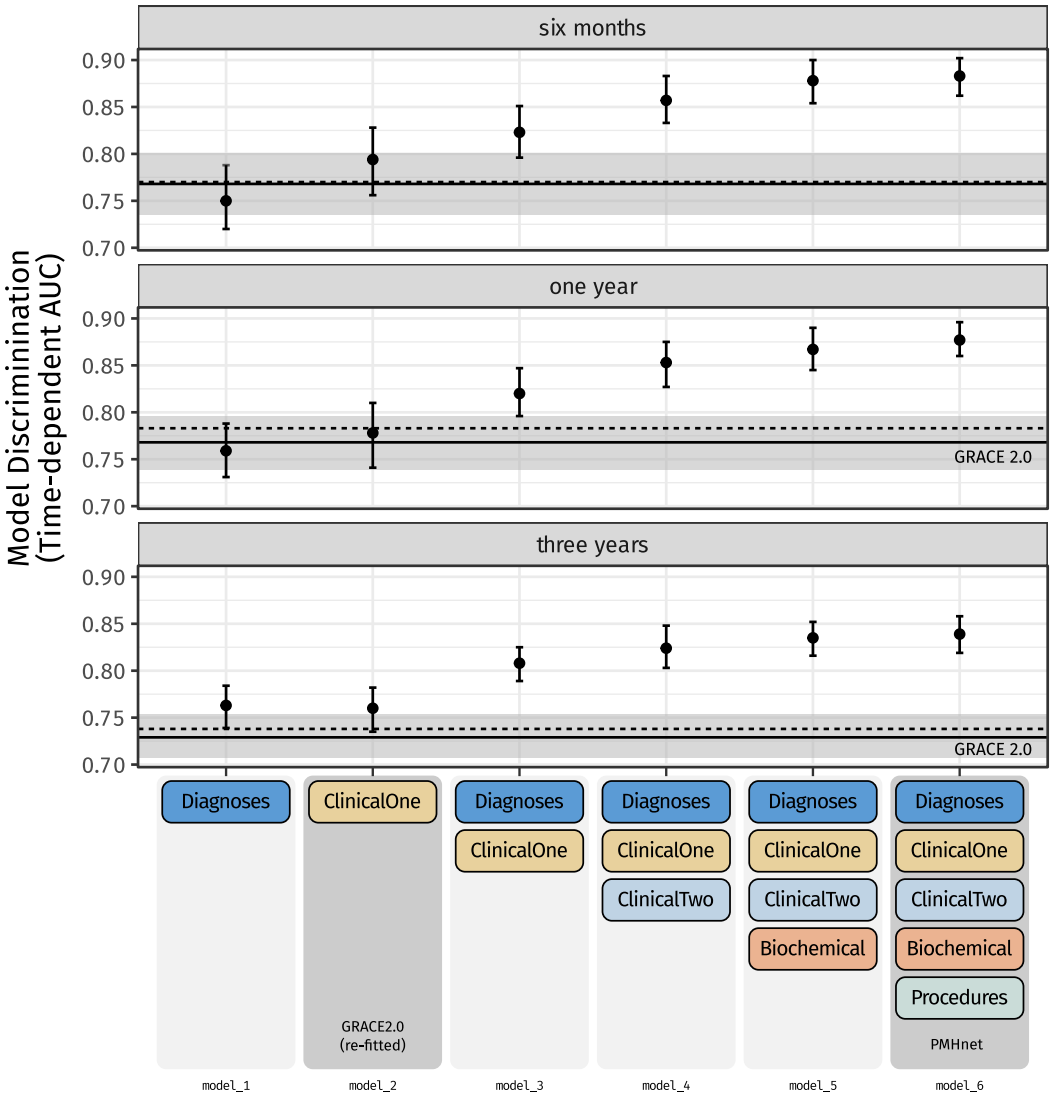


Figure 3: *Model discrimination with increasing number of feature categories.* Time-dependent area under the curve (tdAUC) for various intermediate PMHnet models (and the final one (model 6)) at six months, one year, and three years after the index coronary angiography. Discrimination was evaluated using the hold-out test set. The colored boxes represent the different feature categories that were used as model input in the different models. Horizontal reference lines show the model discrimination of the GRACE2.0 score on the same data. The solid line is the tdAUC of GRACE2.0 on all patients and the dotted line is the tdAUC on the subset of patients where none of the GRACE2.0 input features were missing.

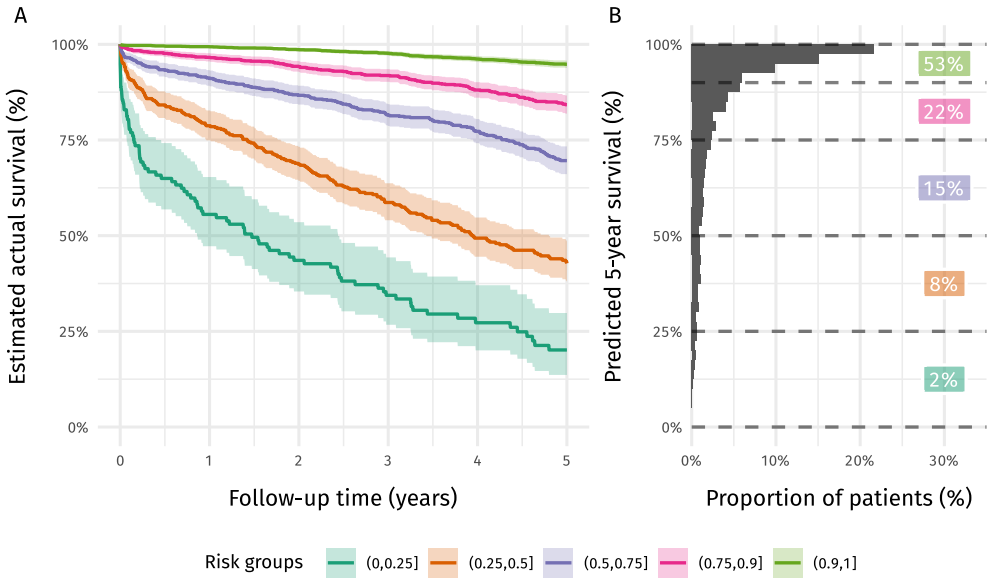


Figure 4: *Observed survival across 5-year predicted risk groups.* A) Estimated actual survival (Kaplan-Meier estimates) for patients in the test set manually stratified into five different risk-groups depending on the predicted survival at five-years by PMHnet. B) Distribution of PMHnet 5-year predicted risk. Vertical lines show the cut-offs that are used to define the risk strata used in A).

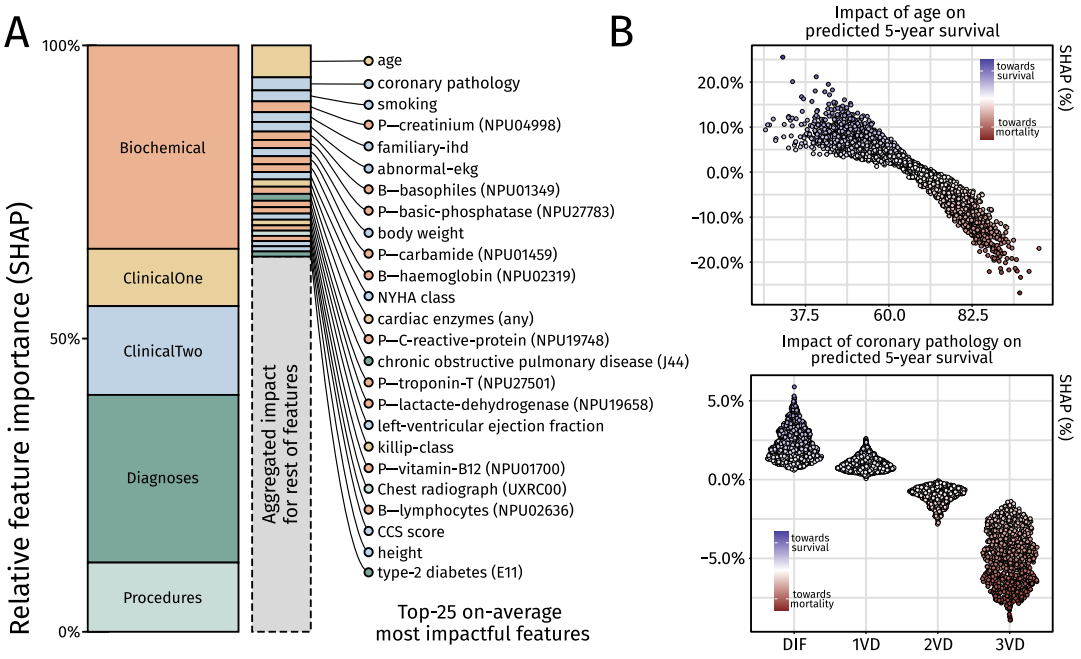


Figure 5: *Overview of feature importance.* Summary of the results from SHAP analysis on the model predictions at five years on patients in the Danish test set. A) Left: Relative feature importance aggregated across the five different feature categories. Biochemical test results and diagnoses were found to affect the model prediction the most. Right: Relative feature importance for all singular features included in the model. Features arranged according to SHAP-values and labels are included for the top 25-most impactful features. Color of features correspond to the feature category in which they belong. SHAP: SHapley Additive exPlanations. B) Relationship between age and SHAP-value with each point showing the SHAP-value for age for a patient in the test-set, and relationship between coronary pathology (e.g. vessel status) and impact on model prediction.

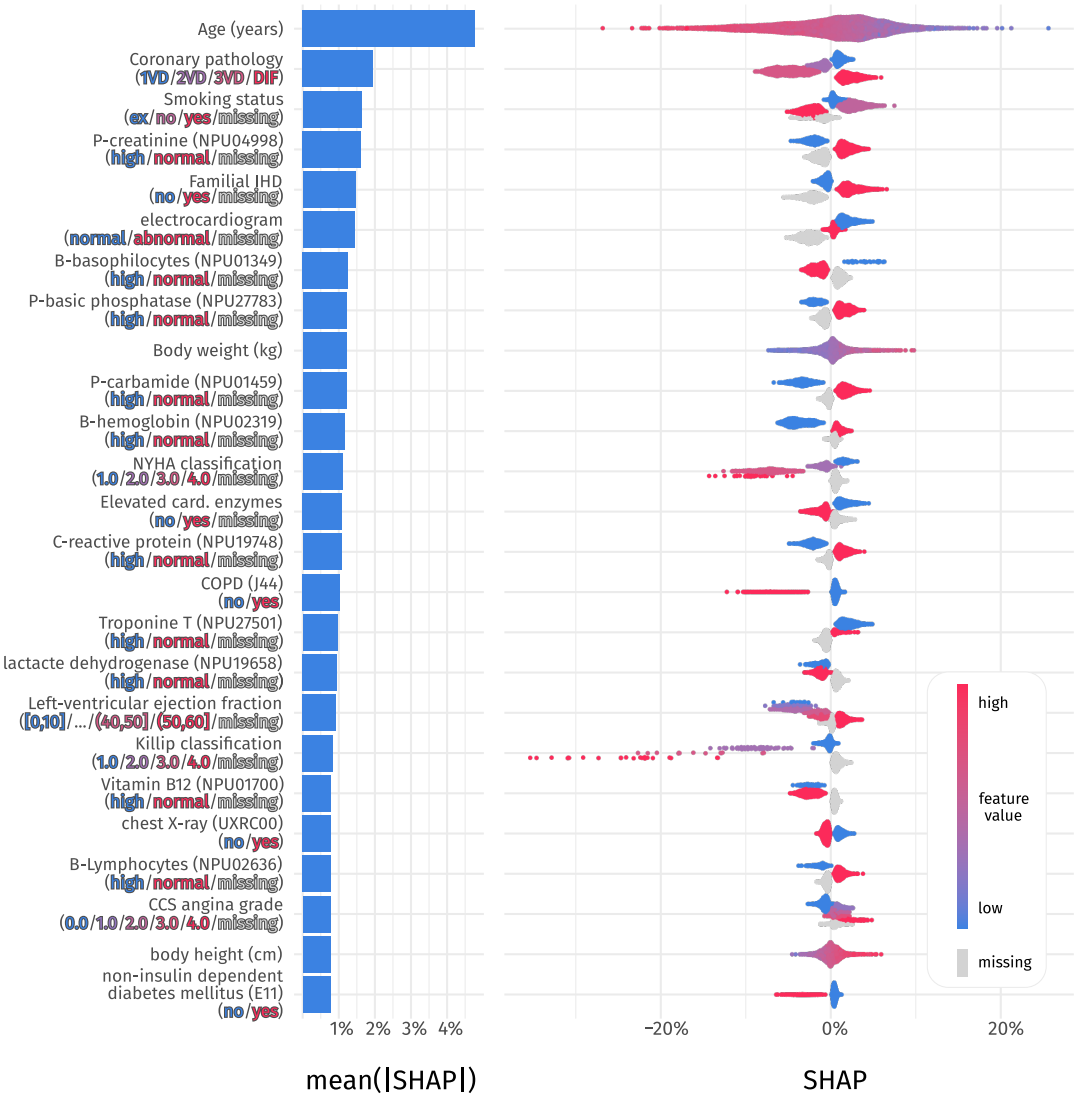


Figure 6: *SHAP summary plot for top 25 most impactful features*. Left: average magnitude of model impact. Y-axis labels specifies the feature name and inside parentheses is shown the unit or the factor levels, for continuous and categorial features, respectively. Right: Distribution of feature impacts across the test set. For continuous features, the colors correspond to the feature value ranging from blue (smallest) to red (largest). For categorical features, the factor levels are colored from blue to red. Grey indicates missingness.

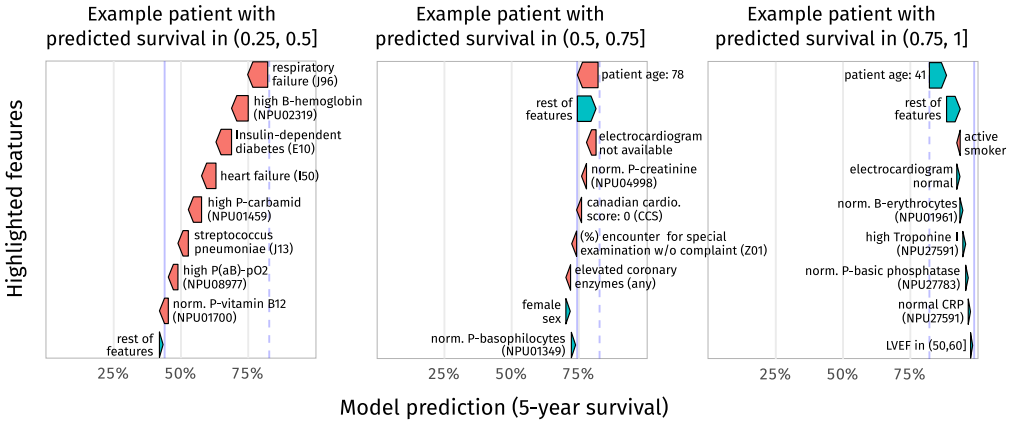


Figure 7: *Patient-level model explanations with SHAP.* Model predictions for three different representative patients with a predicted 5-year survival of 42%, 73%, and 98% with SHAP-explanations showing the estimated impact on model prediction. The patient data have been slightly adjusted to make them non-identifiable. Blue represents features that contribute positively to the prediction. Red represents features that contribute negatively to the prediction. SHAP: SHapley Additive exPlanations.The dashed line is the median prediction from the training set patient and the solid line is the prediction for each of the example patients.

	Training (n=34,746)	Test (n=5,000)	Validation (n=8,288)
Male Sex	23413/67.3%	3410/68.2	5876/70.9%
Age (year)[95%CI]	66.0 [65.7;66.4]	66.2 [66.0;66.3]	66.0 [65.8; 66.2]
Height (cm) [95%CI] (NA)	172.8 [172.7;172.9] (6.5%)	173 [172.7;173.2] (6.3%)	174.3 [174.1; 174.5]
Weight (kg) [95%CI] (NA)	81.3 [81.1;81.5] (4.3%)	81.7 [81.1;82.1] (4.1%)	87.2 [86.8; 87.5]
Comorbidities (ICD10)			
Diabetes	4635/13.3%	694/13.9	1035/12.5%
Hypertension	788/2.27%	105/2.1%	216/2.6%
COPD	2586/7.44%	389/7.78%	715/8.6%
Dyslipidemia	4291/12.3%	641/12.8%	381/4.6%
Medication (ATC)			
lipid-lowering drugs	18188/52.3%	2679/53.6%	6592/79.5%
anti-hypertensive drugs	25684/73.9%	3679/73.6%	7494/90.4%
type-2 diabetes drugs	5575/16.0%	808/16.2%	1048/12.6%
insulin	2088/6.01%	320/6.4%	158/1.9%
P-creatinine, mmol/l [95%CI] (NA) (*)	84.8 [84.5;85.2] (10.1%)	84.2 [83.4;85.1] (10.8%)	91.7 [90.6; 92.9] (8.3%)
P-creatinine NPU04998 (disc.)			
Within ref. range	15,490/44.6%	2,188/43.8%	-
Above ref. range	4,437/12.8%	659/13.2%	
Data missing	14,819/42.6%	2,153/43.1%	
B-basophiles NPU01349 (disc.)			
Within ref. range	12,982/37.4%	1,816/36.3%	-
Above ref. range	192/0.56%	28/0.56%	
Data missing	21,572/62.1%%	3,156/63.1%	
P-basic phosphatase NPU27783 (disc.):			
Within ref. range	12,597 / 36.3%	1,737/34.7%	-
Above ref. range	1,783/5.13%	280/5.6%	
Data missing	20,366/58.6%	2,983/59.7%	
Elevated cardiac biomarkers (†)	15531/44.7% (38.7%)	2158/43.2% (40%)	1599/19.3% (39.6%)
Abnormal EKG:			
Yes	14,679/42.2%	2,043/40.9%	-
No	10,64/30.6%	1,554/31.1%	
Data missing	9,426/27.1%	1,403/28.1%	
Familial IHD:			
Known family history	10,558/30.4%	1,495/29.9%	-
No known family history	18,067/52.0%	2,655 /53.1%	
Data missing	6,121/17.6%	850 /17%	
Heart rate, bpm [95%CI] (NA)	74.7 [74.5;74.9] (20.2%)	74.2 [73.7;74.8] (21.1%)	69.6 [69.3; 70.0] (23.7%)
Blood pressure			
Systolic, mmHg *[95%CI] (NA)*	139.0 [138.7;139.2] (17.6%)	138.9 [138.2;139.7] (17.9%)	148.1 [147.6; 148.6] (14.0%)
Diastolic, mmHg *[95%CI] (NA)*	77.9 [77.7;78.1] (33.5%)	77.7 [77.2;78.2] (34%)	85.5 [85.2; 85.8] (14.0%)
Cardiac arrest at admission?	513/1.5%	69/1.4%	130/1.6%
ICD or PM	557/1.6%	88/1.8%	188/2.3%
Killip class			
Killip: 1	14297/41.1%	2085/41.7%	3469/41.9%
Killip: 2	455/1.3%	62/1.2%	1028/12.4%
Killip: 3	138/0.4%	13/0.3%	191/2.3%
Killip: 4	170/0.5%	22/0.4%	135/1.6%
Data missing	19686/56.7%	2818/56.4%	3465/41.8%
Left-ventricular ejection fraction (%)			
LVEF: [0,10]	100/0.3%	16/0.3%	–
LVEF: (10,20]	675/1.9%	78/1.6%	43/0.5%
LVEF: (20,30]	1305/3.8%	171/3.4%	166/2.0%
LVEF: (30,40]	1760/5.1%	258/5.2%	308/3.7%
LVEF: (40,50]	3915/11.3%	530/10.6%	758/9.1%
LVEF: (50,60]	11068/31.9%	1576/31.5%	2175/26.3%
Data missing	15923/45.8%	2371/47.4%	4838/58.4%
Smoking status			
Active smoker	10237/29.5%	1497/29.9%	1551/18.8%
Former smoker	11961/34.4%	1758/35.2%	4331/52.2%
Never smoked	9347/26.9%	1296/25.9%	2316/28.0%
Data missing	3201/9.2%	449/9.0%	101/1.2%
Coronary pathology			
Diffuse atheromatosis	9288/26.7%	1331/26.6%	652/7.9%
1 vessel disease	13310/38.3%	1936/38.7%	2740/33.0%
2 vessel disease	6469/18.6%	945/18.9%	1424/17.2%
3 vessel disease	5679/16.3%	788/15.8%	1133/13.6%
Data missing	-	-	28.37% (#)

Table 1: *Cohort characteristics for training, test, and external validation set.* The comorbidities are defined from the ICD-10 codes that had been assigned to a given patient prior to the index date. (Continued next page)

Table 1: (Continued) Medication is defined from prescriptions prior to index date. Lipid-lowering drugs: C10, anti-hypertensive drugs: C02, C03, C07, C08 and C09, type-2 diabetes drugs: A10B, insulin: A10A. 95%CI: 95% confidence intervals. ATC: Anatomical Therapeutic Chemical Code. ICD-10: International classification of Disease, 10th Revision. NA: Not applicable. COPD: Chronic obstructive pulmonary disease, ICD: Implantable cardioverter-defibrillator, LVEF: Left-ventricular ejection fraction. PM: Permanent pacemaker.
(*): See supplementary methods for details on differences between continuous and discrete classification of creatinine. (#): Diffuse atheromatosis could not be defined with complete certainty for 28.4% of the Icelandic data, and coronary pathology was therefore set as NA in such cases.

Category	**Features**
ClinicalOne (CRACE2.0)	Age, pulse, systolic blood pressure, cardiac arrest at presentation (yes/no), abnormal cardiac enzymes (yes/no), Killip-class, creatinine, ST-segment deviation (yes/no)
ClinicalTwo	Abnormal ECG (yes/no), CCS class, diastolic blood pressure, coronary artery dominance (R/L/B), familial IHD (yes/no), height, weight, ICD-device or PM (yes/no), ischemia test, LVEF, NYHA class, sex, smoking status, coronary pathology,
Diagnoses	322 different level-3 ICD-10 diagnosis codes.
Procedures	154 different NOMESCO procedure codes corresponding to various radiological examinations and surgical procedures
Biochemical	85 different lab tests with results categorized as *below*, *within*, or *above* the reference range

Table 2: *Input features used for model development.* The different features were organized in five different categories each representing different domains. The clinical characteristics were divided into two subgroups where ClinicalOne contains the features used in the GRACE2.0 score.

	6 months		1 year		3 years	
Comparison	ΔAUC (%)	p-value	ΔAUC (%)	p-value	ΔAUC (%)	p-value
PMHnet vs. GRACE2.0 (conventional)	11.5 [9.0; 14.0]	1e-9	10.9 [8.6; 13.2]	6e-21	10.3 [8.3; 12.4]	1e-22
PMHnet vs. GRACE2.0 (re-fitted)	8.9 [6.5; 11.2]	2e-13	8.9 [6.5; 11.2]	2e-18	7.6 [5.8; 9.4]	1e-6

Table 3: *Difference in discrimination between PMHnet and the GRACE2.0 score.* For δAUC, we obtain 95% CIs (in brackets) and p-values from the Score function in the R package riskRegression38.

Supplementary material

Supplementary methods

Feature inclusion and pre-processing of the Danish training and test set

For each patient we extracted the following data: all diagnosis codes (ICD-10) and procedure codes (SKS/NOMESCO) assigned to at least 1% of the cohort between 1st of January 1994 and the time of the coronary angiography (from NPR); all results from laboratory tests taken on at least 5% of the cohort between 5 years prior-to and up until the time of the coronary angiography (from BTH); 23 other clinical features such as sex, age, smoking status, coronary pathology, etc. (from EDHR + BTH). Moreover, for 37.4% of the cohort a panel of 19 different PRSs was included (see below). In case of repeated tests/measurements/assignments the one closest to the time of the coronary angiography was used. No data originating after the index coronary angiography were used. Laboratory test results and sex- and age adjusted (if applicable) reference ranges were originally stored in the regional laboratory information management systems "Labka" and "BCC" and in this study obtained through BTH[43]. Tests were either annotated in accordance with the Nomenclature, Properties and Units ontology (NPU) or various local coding systems[44]. Using the sex- and age adjusted reference ranges, results of the biochemical tests were discretized to the categories *below*, *within*, *above*, and if a given test had not been taken a category called *missing*. Discretization of biochemical tests using the adjusted reference ranges was a pragmatic alternative to normalizing the many different biochemical tests in a manner that adequately takes both intra- and interdepartmental variance into account,

which both might be related to for instance differences in patient population and/or equipment and machinery.

Consequences of the choices above, including discretization were alluded to in Table 1. Here it is evident that the missingness of the continuous creatine feature creatinine was less than that of the discretized version because we allow measurements 21-days in the "future" relative to time zero. For the discretized version, we did not include those measurements, and consequently so missingness consequently higher. All continuous features were Z-score normalized and missing values were encoded with a value of zero. All categorical features were one-hot encoded and an additional category for missing values was added if applicable[25].

During model development, we tested and optimized various feature specific hyperparameters. We tested including diagnosis codes as either level-4, level-3, block, or chapter codes representing different steps in the ICD-10 hierarchy[45]. With the hypothesis that codes and results assigned many years prior to time zero time might carry less information, we introduced three "shelf-life" hyperparameters and filtered out diagnosis codes, procedure codes, and biochemical tests assigned/taken more than *n* years prior to the coronary angiography.

Cohort characteristics

For the cohort characteristics presented in Table 1, comorbidities were defined from the ICD-10 codes assigned before or at the index date: *diabetes* was defined as E10 or E11; *hypertension* was defined as I10, I11, I12, I13, I14, or I15; *COPD* was defined as J44; and *dyslipidemia* was defined as E78.0, E78.1, E78.2, E78.3, E78.4, E78.5, and E78.9. The

medication use was defined from prescriptions given prior to or at the index-date, using the same definition as used in Kiiskinen et al. (1), that is: *lipid-lowering drugs* is ATC class C10; *anti-hypertensive drugs* is C02, C03, C07, C08, and C09; *type-2 diabetes drugs* is A10B; and *insulin* is A10A.

For continuous features (age, height, etc.) the mean is given along with 95% bootstrap confidence intervals (CI) and, if applicable, the amount of missingness in parentheses – mean [low, high] (missingness). The mean and CI are calculated using only the non-missing features. For categorical features the raw counts and relative frequencies are both specified.

Downscaled version for external validation

Due to data availability, the model was downscaled for the external validation on Icelandic data. The features that were left out of the model in the down-scaling was all procedure codes (n=154) since they were encoded using another non-compatible coding scheme; five biochemical test which was not used in the Icelandic system; and seven different clinical features for which data could not be obtained: “arrest”, “enzymes”, “abnormal-ekg”, “abnormal-qrs-st”, “coronary artery dominance”, “ischemia test”, and “NYHA-class”.

Polygenic risk scores

PRSs were calculated based on GWAS summary statistics data from 19 traits relevant for cardiometabolic health, obtained from 17 GWAS meta-analyses: acute myocardial infarction(2), atrial fibrillation(3), coronary artery disease(4,5), heart failure(4), dilated cardiomyopathy(6), hypertrophic cardiomyopathy(6), systolic and diastolic blood

pressure(7), stroke(8), total cholesterol and triglyceride levels(9), non-alcoholic fatty liver disease(10), immune-mediated inflammatory diseases(11), chronic kidney disease(12), body mass index(13), type 2 diabetes with adjustment for BMI(14), frailty index(15).

Autosomal genotypes from 188,462 individuals in the CHB Cardiovascular Disease Cohort were filtered to only include variants present in the HapMap3 set of 1,120,696 reference variants. Any missing genotype information was conservatively imputed to be the affected locus' reference allele. We identified a set of 978,246 genotyped variants present in both genotype and GWAS summary statistics data that we subsequently subjected to Ldpred2's recommended standard deviation quality control. After variant matching and quality control, a mean of 969,607 (S.D. 8,777) variants remained for per-chromosome risk score calculation for each of the 19 traits. We used the Ldpred2-auto algorithm with 30 Gibbs sampling chains, 1,000 burn-in iterations and 500 iterations after burn-in. The initial values for the 30 sampling chains were a) the LDSC regression estimate for heritability h^2 (same for all chains); b) one of 30 initial values for the proportion of causal variants p, evenly spaced on a logarithmic scale from 10^{-4} to 0.9. Final per-chromosome effect sizes were calculated from each set of 30 sampling chains (per trait and chromosome) through a three-step process, which serves to ensure that the model (spanning 30 chains) converged: 1) computing the standard deviations of each chains' predicted scores, 2) keeping only the chains within three median absolute deviations from the median standard deviation, 3) averaging the effect sizes of the remaining chains. Across the 418 per-chromosome models (19 traits times 22 chromosomes), 26.89 chains were included in the final score on average. The lowest number of included chains was 18. Finally, the resulting per-chromosome risk scores were added together into genome-wide polygenic scores.

Neural network architecture

The neural network architecture is a relatively simple feed-forward neural network with between one and three densely connected hidden layers with rectified linear units as the activation function and with a dropout layer after each hidden layer. The output layer has 30 outputs – one for each time bin – and is being fed through a softmax activation function to produce probabilities. The neural network architecture is sketched out in figure S5.

Hyperparameter optimization

For hyperparameter optimization we used the Optuna hyperparameter optimization package for python (16). For all hyperparameter sweeps, we ran 2500 trials. For each trial, hyperparameters were sampled using the TPE (Tree-structured Parzen Estimator) algorithm as implemented in Optuna using 400 startup trials and otherwise all other arguments left on default settings(17). To prune unpromising trials, we used the hyperband pruner, with a minimum resource of 20, and a reduction factor of 3.

Resiliency of PMHnet to missingness

Finally, the resiliency of the model was assessed as described in Methods. Across the 584 features, age was the only single feature, where the change in tdAUC changed significantly when artificially removed. In the case of model calibration, missingness of six individual features significantly affected the performance. These features were troponins, LVEF, age, smoking status, abnormal ECG, and coronary pathology. Taken together, these results are evidence that PMHnet is resilient to missing data.

Supplementary figures

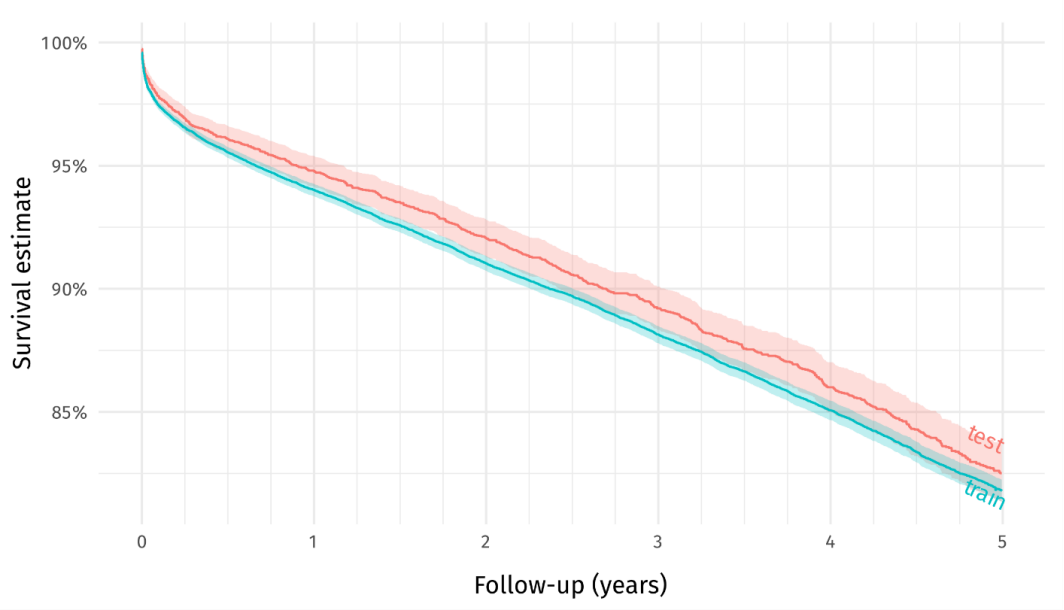


Figure S1: ***Kaplan-Meier estimates for the PMHnet derivation cohort.*** *Kaplan-Meier estimates for the training (blue) and testing (red) set of 34,749 and 5,000 patients, respectively.*

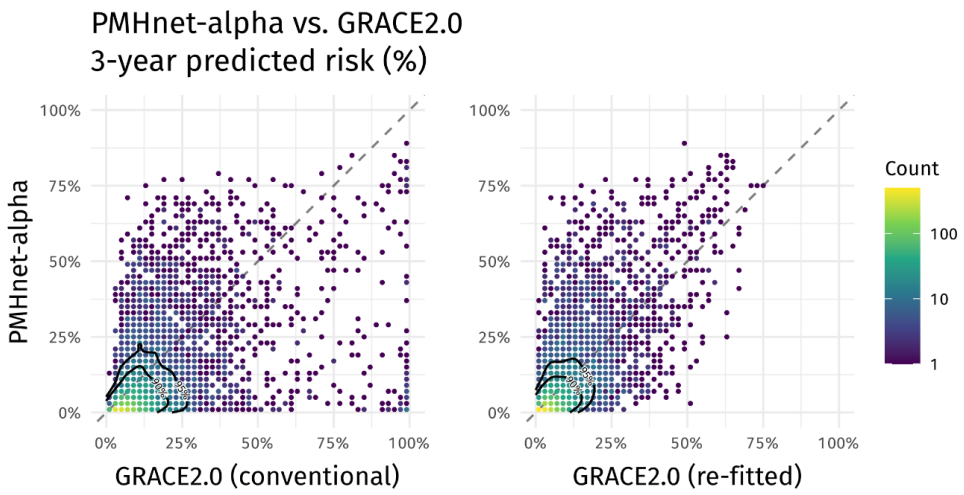


Figure S2: ***Difference in predicted risk between PMHnet and GRACE2.0.*** *Binned scatterplot showing the difference between the 3-year predicted risk by PMHnet (y-axis) and GRACE2.0 (conventional) [left-panel, x-axis] or GRACE2.0 (re-fitted) [right-panel, x-axis]. Points are colored according to how many patients in the test set fall in that bin. Contour lines indicate 90%- and 95%-point densities.*

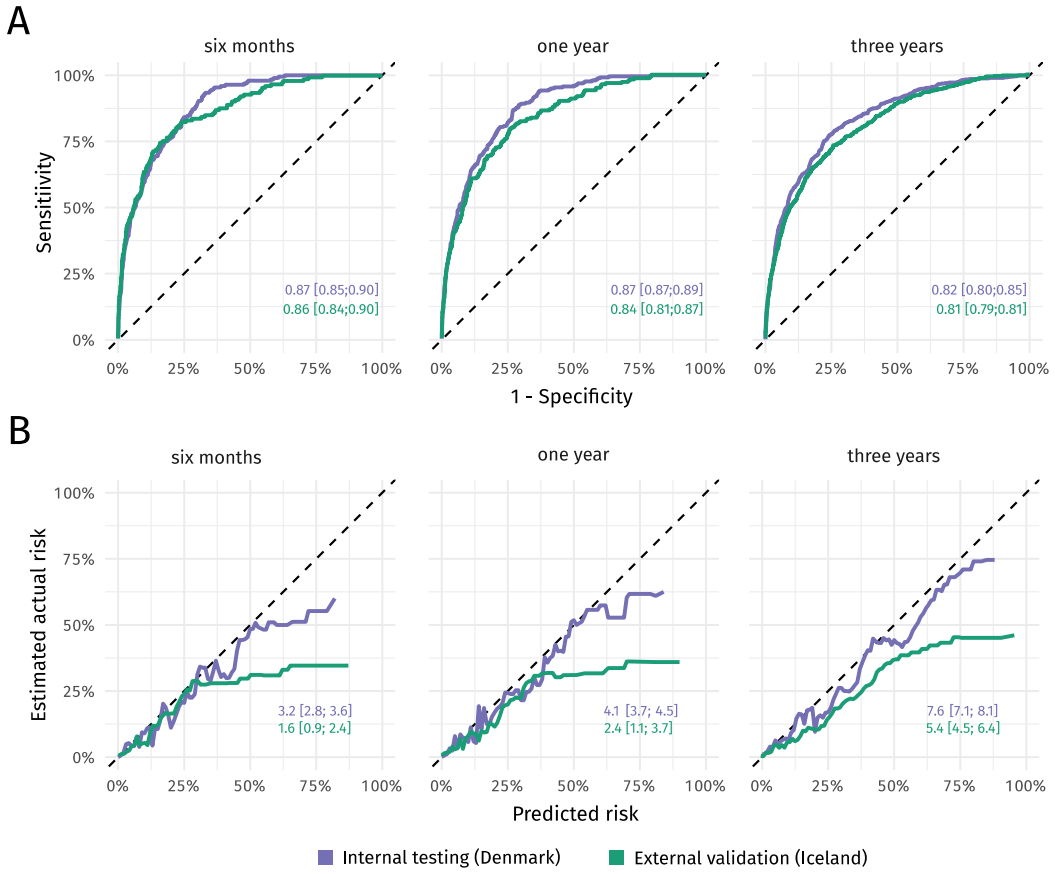


Figure S3: ***External validation of PMHnet on Icelandic data.*** *A) Time-dependent ROC curves at three different prediction horizons for PMHnet evaluated on the Danish test set and the Icelandic external validation set, respectively. Text labels show the corresponding tdAUC scores. B) Calibration curves showing the relation between predicted risk and the estimated actual risk. Labels show the Brier score for each of datasets. Lower scores are associated with better calibration and discrimination of predictions. Data for both type of plots and scores were generated using riskRegression and visualized using ggplot2.*

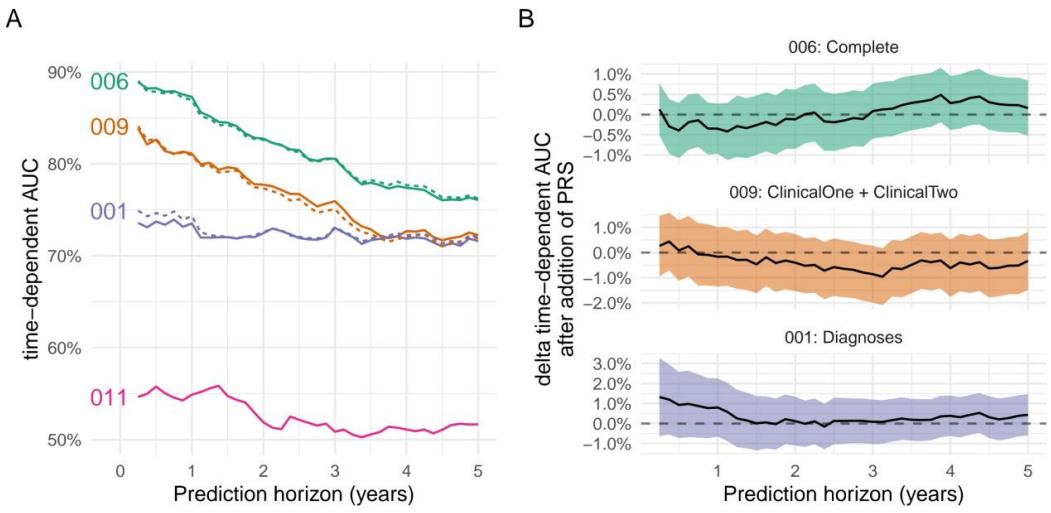


Figure S4: ***Effect of including polygenic risk scores to clinical features.***

[A] The time-dependent AUC values for different models evaluated on the subset of the hold-out test set where genetic information could be obtained. From top to bottom, the first two lines shows the performance of the complete model ("006") with and without PRS features added. The next sets of two correspond to a model only using clinical features ("009", ClinicalOne + ClinicalTwo) and a model only using diagnosis codes ("001"). Finally, the last line is the performance of a model that is using PRSs as the sole predictors. [B] A formal test of ΔtdAUC obtained by adding PRS information to each of the models in panel A)._Ribbon shows confidence interval obtained through the "Score" function from the riskRegression R-package.

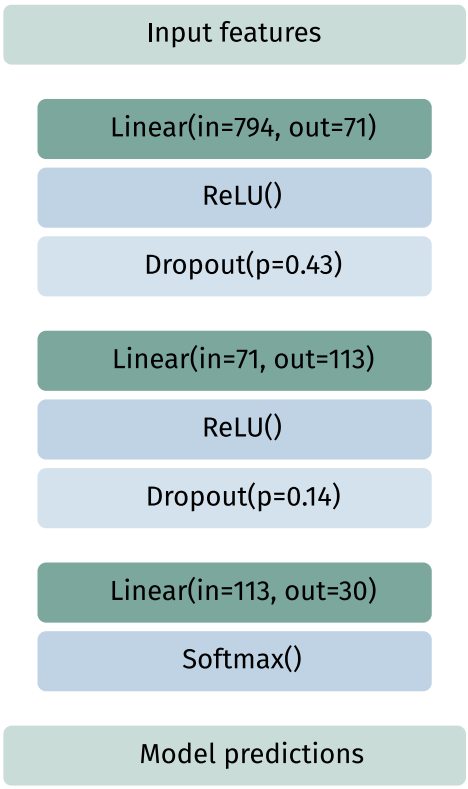


Figure S5: ***Neural network architecture of PMHnet (full version).***

Using the PyTorch machine learning framework for python, the illustrated neural network architecture was implemented. The architecture hyperparameters, number of layers; number of units; and droprate, were determined from the hyperparameter search performed using the Optuna hyperparameter optimization framework as detailed in the methods section.

1 Supplementary tables

2 *Table S1:* ***Missingness across clinical features for the Danish derivation data.***

***ClinicalOne* features**	**training**	**testing**
Systolic blood pressure	18%	18%
Elevated cardiac enzymes	39%	40%
Cardiac arrest	0%	0%
Age	0%	0%
STEMI	0%	0%
Heart rate	20%	21%
Creatinine	10%	11%
Killip-class	57%	56%
***ClinicalTwo* features**	**training**	**testing**
Abnormal ECG	0%	0%
Abnormal QRS-ST	0%	0%
Canadian cardiovascular score	8.6%	8.8%
Diastolic blood pressure	33.5%	34%
Dominance	3.7%	3.8%
Familial IHD	0%	0%
Height	6.4%	6.4%
ICD-or-PM	0%	0%
Ischemia test	0%	0%
LVEF	0%	0%
NYHA	75.9%	76.8%
Sex	0%	0%
Smoking	9.2%	9.0%
Vessel status	0%	0%
Weight	4.3%	4.1%

Table S2: Outcomes and Kaplan-Meier estimates across training, test, and external validation data.

Outcomes	Training set	Test set	External validation set
RMST(1825), days	1635, SE: 2.58	1651, SE: 6.49	1697, SE: 3.82
KM-estimate, 6 months	95.5% [95.3; 95.8]	96.1% [95.5; 96.6]	98.1% [97.9; 98.4]97.3%
KM-estimate, 1 year	94.0% [93.8; 94.3]	94.8% [94.2; 95.4]	[97.0; 97.7]93.7% [93.2;
KM-estimate, 3 years	88.1% [87.8; 88.5]	89.2% [88.3; 90.1]	94.2]89.6% [88.9; 90.2]
KM-estimate, 5 years	81.8% [81.4; 82.2]	82.5% [81.3; 83.6]	

Table S3: Hyperparameters search space and final configuration

Hyperparameter	*Search space*	*Best trial*
Biochemical inclusion window		*2.5 years*
Diagnosis inclusion window		*13.5 years*
Procedures inclusion window		*3.5 years*
Number of hidden layers		*2*
Number of units per hidden layer		*Layer 1: 71,* *Layer 2: 113*
Droprate per hidden layer		*Layer 1: 42.8%* *Layer 2: 14.0%*
Learning rate		
Momentum		

References

1. Kiiskinen T, Helkkula P, Krebs K, Karjalainen J, Saarentaus E, Mars N, et al. Genetic predictors of lifelong medication-use patterns in cardiometabolic diseases. Nat Med. 2023 Jan;29(1):209–18.

2. Jiang L, Zheng Z, Fang H, Yang J. A generalized linear mixed model association tool for biobank-scale data. Nat Genet. 2021 Nov;53(11):1616–21.

3. Nielsen JB, Thorolfsdottir RB, Fritsche LG, Zhou W, Skov MW, Graham SE, et al. Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. Nat Genet. 2018 Sep;50(9):1234–9.

4. Nikpay M, Goel A, Won HH, Hall LM, Willenborg C, Kanoni S, et al. A comprehensive 1000 Genomes–based genome-wide association meta-analysis of coronary artery disease. Nat Genet. 2015 Oct;47(10):1121–30.

5. van der Harst P, Verweij N. Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. Circ Res. 2018 Feb 2;122(3):433–43.

6. Sakaue S, Kanai M, Tanigawa Y, Karjalainen J, Kurki M, Koshiba S, et al. A cross-population atlas of genetic associations for 220 human phenotypes. Nat Genet. 2021 Oct;53(10):1415–24.

7. Hoffmann TJ, Ehret GB, Nandakumar P, Ranatunga D, Schaefer C, Kwok PY, et al. Genome-wide association analyses using electronic health records identify new loci influencing blood pressure variation. Nat Genet. 2017 Jan;49(1):54–64.

8. Malik R, Chauhan G, Traylor M, Sargurupremraj M, Okada Y, Mishra A, et al. Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. Nat Genet. 2018 Apr;50(4):524–37.

9. Surakka I, Horikoshi M, Mägi R, Sarin AP, Mahajan A, Lagou V, et al. The impact of low-frequency and rare variants on lipid levels. Nat Genet. 2015 Jun;47(6):589–97.

10. Anstee QM, Darlay R, Cockell S, Meroni M, Govaere O, Tiniakos D, et al. Genome-wide association study of non-alcoholic fatty liver and steatohepatitis in a histologically characterised cohort☆. J Hepatol. 2020 Sep 1;73(3):505–15.

11. Acosta-Herrera M, Kerick M, González-Serna D, Consortium MG, Consortium SG, Wijmenga C, et al. Genome-wide meta-analysis reveals shared new loci in systemic seropositive rheumatic diseases. Ann Rheum Dis. 2019 Mar 1;78(3):311–9.

12. Wuttke M, Li Y, Li M, Sieber KB, Feitosa MF, Gorski M, et al. A catalog of genetic loci associated with kidney function from analyses of a million individuals. Nat Genet. 2019 Jun;51(6):957–72.

13. Yengo L, Sidorenko J, Kemper KE, Zheng Z, Wood AR, Weedon MN, et al. Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. Hum Mol Genet. 2018 Oct 15;27(20):3641–9.

14. Mahajan A, Taliun D, Thurner M, Robertson NR, Torres JM, Rayner NW, et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. Nat Genet. 2018 Nov;50(11):1505–13.

15. Atkins JL, Jylhävä J, Pedersen NL, Magnusson PK, Lu Y, Wang Y, et al. A genome-wide association study of the frailty index highlights brain pathways in ageing. Aging Cell. 2021;20(9):e13459.

16. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A Next-generation Hyperparameter Optimization Framework [Internet]. arXiv; 2019 Jul. Report No.: 1907.10902. Available from: https://arxiv.org/abs/1907.10902

17. Bergstra J, Bardenet R, Bengio Y, Kégl B. Algorithms for Hyper-Parameter Optimization. In: Advances in Neural Information Processing Systems [Internet]. Curran Associates, Inc.; 2011 [cited 2023 Jun 14]. Available from: https://proceedings.neurips.cc/paper_files/paper/2011/hash/86e8f7ab32cfd12577bc2619bc635690-Abstract.html

Appendix C

Manuscript for Study III

Development of a neural network-based competing risk model for long-term prognostication in ischemic heart disease from a large database of electronic health records and clinical registries

Peter C. Holm (1), Amalie D. Haue (1, 3), Alex H. Christensen (3, 4), Henning Bundgaard (3), and Søren Brunak (1, 5)

1. Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Blegdamsvej 3B, DK-2200 Copenhagen, Denmark
2. Department Obstetrics and Gynecology, Copenhagen University Hospital, Kettegård Alle 30, DK-2650 Hvidovre, Denmark
3. Department of Cardiology, Copenhagen University Hospital, Rigshospitalet, Blegdamsvej 9, DK-2100 Copenhagen, Denmark
4. Department of Cardiology, Copenhagen University Hospital, Herlev-Gentofte Hospital, Borgmester Ib Juuls Vej 1, DK-2730 Herlev, Denmark
5. Copenhagen University Hospital, Rigshospitalet, Blegdamsvej 9, DK-2100 Copenhagen, Denmark

Abstract

Background: Machine-learning based risk-stratification of patients with ischemic heart disease (IHD) can improve prognostication through identification of both high and low risk patients. However, most models are unable to handle competing risks which can affect usability and scope.

Methods: We present PMHnetV2, designed for joint modelling of competing risks in the presences of censoring. PMHnetV2 was developed using health data from 52,787 patients (train: 42,065, test: 10,722) with a diagnosis of IHD who underwent coronary angiography between 2006 and 2016. Follow-up information could be collected up until 2019. Model performance was evaluated using the index of prediction accuracy and time-dependent AUC.

Results: PMHnetV2 accurately predicts all-cause mortality, cardiovascular mortality, recurrent ischemic events, and hospital admissions with ischemic heart disease complications and sequela. The models are well-calibrated and improve on the baseline models.

Conclusion: modeling competing risks jointly can improve risk prediction. Utilizing this approach could offer patients and clinicians a post-angiography assessment that might aid in the decision-making process.

Funding: Novo Nordisk Foundation, NordForsk, and the Innovation Fund Denmark.

Introduction

In recent years, machine learning models, particularly neural networks, have emerged as pivotal tools in precision medicine (1). These models excel at identifying complex patterns in large datasets, a characteristic particularly helpful for analysis of large-scale modern electronic health records. This capability has paved the way for creation of new and advanced clinical decision support tools (2). In the field of cardiology, machine learning models are showing immense potential, often surpassing the performance of traditional methods. These models are especially leading in the area of prediction and prognostication, where time-to-event outcomes and censoring are prevalent (3–7). Here, neural network-based survival models represent the current state of the art (6,8–12). Our previous contribution to this evolving landscape is PMHnetV1, a neural network-based survival model developed to predict all-cause mortality in patients with ischemic heart disease (13).

A notable shortcoming in our previous work was the inability to differentiate between deaths related to complications of ischemic heart disease and those arising from completely unrelated causes, alongside predicting specific measures of disease progression. To address this, we need methods capable of handling competing risks (14), an attribute rarely found in existing neural network-based models. The most prominent exception is the "DeepHit" approach presented by Lee et al. (6) for survival analysis with competing risks. In the cardiological domain, Pieszko et al. leveraged this approach in developing a prediction model for personalized risk assessment following myocardial perfusion imaging (5). Although they did not present calibration metrics, their unified model surpassed traditional perfusion abnormality measures in discrimination, offering event-specific risk estimates for outcomes such as all-cause mortality, acute coronary syndrome, and revascularization. Yet, as Kvamme et al. highlighted, DeepHit-based models, while presenting excellent discrimination, tend to lag in terms of calibration compared to alternatives (11).

Recognizing both advancements and existing gaps in the current literature, this study seeks to build on our previous work by introducing PMHnetV2. This enhanced version extends the number of included predictors and incorporates prediction of cardiovascular mortality, recurrent myocardial ischemia, and future hospital admissions with ischemic heart disease complications and sequelae. To this end, we implemented and utilized a discrete-time logistic hazard model that allows jointly modelling competing risks. We compare our competing risk models that use this strategy against the simplistic strategy of treating such risks as censoring.

Materials and Methods

Sources of data

The data used for model development originates from a combination of national registries (CPR, DAR, LMSR, LPR), Eastern Danish Heart Registry (PATS), and an electronic health record database (BTH). Data sources were linked using the Danish unique national personal identification number ("CPR number"), which is assigned to all residents of the country.

	Description	Ref.
The Danish Central Person Register (CPR)	Contains date of birth, sex, person status (alive/dead/emigrated), time of status change, etc. on all persons living in Denmark.	(15)
The Causes of Death Register (DAR)	Contains date of death and information on direct and contributory causes of deaths coded using the ICD-10 classification system. Based on the physician-filled death certificates, that have been legally required since 1871.	(16)
The Register of Pharmaceutical Sales (LMSR)	Contains data on all prescription drugs sold in Danish pharmacies since 1994.	(17)
The Danish National Patient Registry (LPR)	Nationwide longitudinal registry with data on all admissions, examinations, treatments, procedures, diagnoses, etc. in Danish public hospitals since 1977. Have been using the ICD10 coding system for diagnosis codes since 1994.	(18)
The east-Danish heart registry (PATS)	Clinical quality database with data from coronary angiographies and percutaneous coronary interventions performed at specialized departments in the Capital Region of Denmark and Region Zealand.	(19,20)
BTH	Database with data from population-wide electronic health record system (2.6 million patients) from 1st of January 2006 to 31st of July 2016. Contains highly heterogenous data ranging from administrative data to laboratory test results and unstructured journal text.	-

Inclusion criteria

The target population is individuals with a diagnosis of ischemic heart disease. Patients were included if they had been subjected to a coronary angiography demonstrating 1/2/3-vessels disease or diffuse atheromatosis in the period 1st of January 2006 and 31st of December 2016 (both inclusive). We excluded patients without a prior ICD-10 code for ischemic heart disease (I20-25), below 18 years of age at time of the coronary angiography, and those not alive two days after the procedure (Figure 1A). For additional details, see *Supplementary Methods*.

Follow-up and endpoints

The time origin was the index coronary angiography, which was defined as the first coronary angiography for each patient that satisfied the other inclusion criteria. We defined four different major outcomes:

1. All-cause mortality (ACMO), which was identified from the central person register.
2. Cardiovascular mortality (CVMO), which was defined from the cause of death register as any death with an ICD-10 code of I00-99 registered as the underlying cause of death. Deaths from other causes were competing risks.
3. Cardiovascular complications (CVCO), which was defined as a composite outcome from the national patient register diagnosis or procedure codes. CVCO was defined as hospital admission with either “heart failure” (ICD-10: I50), atrial fibrillation or flutter (I48), “cardiac arrest” (I46), or “cerebrovascular accident” (I61, I63-64) as the primary diagnosis code; or implantation of pacemaker (SKS: BFCA0*) or cardioverter-defibrillator (BFCB0*). Admissions within four weeks after time origin were ignored (Fig S1 and *Supplementary Methods*). ACMO was treated as a competing risk.
4. Myocardial ischemic events (MIEV), which was identified from the national patient register as in-patients hospitalized for at least 24 hours with the primary diagnosis code ICD-10: I20-I25, or a procedure code for PCI or CABG. In both cases, events within eight weeks after time of origin were ignored (Fig S2 and *Supplementary Methods*). ACMO was treated as a competing risk.

Variables and features

In the following, we distinguish variables from features: we use variables when referring to the raw input data from the various data sources and features to describe the pre-processed inputs to the machine learning model. To illustrate, the categorical variable "tobacco usage" is transformed into four separate binary features through one-hot encoding: "former", "current", "never", and "missing".

In total, we used 1,860 different variables, belonging to five different categories, which after preprocessing amounted to 2,262 distinct features. The five categories, and the number of features belonging in each, were clinical variables: 80, procedure codes: 418, prescription data: 785, diagnoses: 504, and laboratory tests: 475. The full set of features can be seen in *Supplementary File 1* and additional details are found in *Supplementary Methods*.

Numerical variables included in the clinical category are site-specific (n=17) stenosis amount and the baseline values of systolic and diastolic blood pressures, pulse, age, height, and weight. Stenosis amount and age were non-missing and only needed scaling. The others were processed with median imputation before scaling, and we added an extra feature to indicate any missing values. As an example, the numerical variable "systolic blood pressure" is represented by two features: one indicating if data is missing and another containing the median-imputed and scaled blood pressure values. We used one-hot encoding for all categorical variables, designating a "missing" category where necessary to retain potentially useful information in the patterns of missingness. We set up the preprocessing pipeline using only the training data to prevent data leakage between the training and testing datasets.

Discrete-time survival analysis with competing risks

To model time-to-event data with censoring and competing risks, we implemented a discrete-time logistic hazard model that allows jointly modelling competing risks. This approach can be viewed as an extension of the methodology proposed by Gensheimer and Narasimhan (8). Our updated framework is a fully parametric discrete time survival model parameterized by a neural network which enables jointly modelling discrete-time survival data with competing risks.

Briefly, and largely following the book by Gerhard Tutz (21), we discretize the follow-up time to estimate the conditional hazard for individual subjects. This hazard represents the probability that the event of interest occurs at a specific time point, given that the subject is still at risk at the start of the interval. To facilitate this, we introduce the indicator δ_{ijr} defined as:

$$\delta_{ijr} = \begin{cases} \mathbb{1}\{r = 0\} & \text{if } t_{(j)} < t_i \\ \mathbb{1}\{r = r_i, t_{(j)} = t_i\} & \text{otherwise} \end{cases}$$

where t_i denotes the observed discrete survival time of subject i, r_i represents the observed event (with 0 indicating censoring), and $t_{(j)}$ refers to the time bin j. With this setup, we can write the negative log-likelihood as

$$loss = -\sum_{i=1}^{N}\sum_{s=1}^{t_i}\left(\sum_{r \in \kappa}\delta_{ijr}\log\lambda_r(s|x_i) + \delta_{ij0}\log\left(1 - \sum_{r \in \kappa}\lambda_r(s|x_i)\right)\right)$$

The model outputs conditional hazards as logits with an estimate for each risk (including survival) at each time interval.

Model development

We used a ResNet-like architecture for tabular data similar to the one described in (22). The architecture depends on the number of input features m, the number of hidden neurons in each hidden layer h, and the number of output logits o, which is $|\text{time bins}| \times (|\text{risks}| + 1)$.

$$ResBlock(x) = x + \big(BatchNorm \circ Linear_{h,h} \circ Dropout \circ SiLU \circ Linear_{h,h} \circ Dropout\big)(x)$$

$$ResNet_{i,o}(x) = (Linear_{i,h} \circ ResBlock_h \circ \ldots \circ Resblock_h \circ Linear_{h,o})(x)$$

Predictions are obtained by passing the output logits through a SoftMax activation function, such that the conditional probability of each possible event (including survival) sums to 100%.

We used the holdout method and randomly split our derivation dataset into a training set (80%) and a test set (20%) for validation (Figure 1B). The training set was further subdivided into training (80%) and validation (20%) splits to have an unbiased estimate of model performance during hyperparameter optimization. All models were trained using the "AdamW" stochastic optimization algorithm (23) following the "super-convergence" training regimen previously described (24), which enables fast, accurate, and resource-efficient training of neural network models (25). We used a fixed batch size of 1024 and trained for a pre-specified number of epochs using the one-cycle learning rate scheduler ("OneCycleLR" in PyTorch).

Hyperparameters; which consisted of a) number of epochs b) max learning rate (upper boundary for the learning rate policy), c) dropout rate, d) weight decay, e) number of residual blocks, f) number of hidden units in each residual block, g) number of time bins in the discretization grid, upper limits on the retrospective inclusion windows for h) biochemical, i) diagnosis, j) procedure, and k) medication data, and whether to use l) skip-connections and m) batch-normalization; were optimized using Optuna (26). We used the index of prediction accuracy (IPA) of the primary outcome on the validation split as the tuning parameter, which is an R-squared type measure that reflects both calibration and discrimination (27). The IPA is time-dependent, so we calculated IPA at 50 evenly spaced timepoints from 0.5 to 5 years and numerically integrated the values to provide a single measure of performance. See *Supplementary Methods* for additional details.

Serving as benchmark, we trained naïve single-risk models where competing events was treated as censoring.

Statistics

For evaluation of the time-to-event prediction models, we used the time-dependent area under the receiver operating characteristic curve (AUC) and the IPA, a scaled version of the Brier score, as the main measures of performance (27–29). Model calibration was assessed graphically by comparing model estimates with pseudovalues of the actual outcomes (30,31). See *Supplementary Methods* for additional details.

We derived estimates for the incidence of the different endpoints using the Kaplan-Meier and the Aalen-Johansen estimators for endpoints without or with competing risks, respectively.

Software

We used PyTorch 1.13.0 (32), Lightning 2.0 (33), Scikit-Learn 1.3 (34), and Optuna 3.1 (26) for Python 3.10 for developing and training the neural network models. Our general purpose discrete-time competing risk framework that we developed for this application has been released on the python package index (PyPI) under the name “Discotime” and can also be found on GithHub at “peterchristofferholm/discotime”. We used version 0.1.0 of Discotime for this study.

Statistical analysis, data wrangling, and visualization was primarily performed using R (v. 4.1) (35) using the “riskregression” package for evaluation of prediction models (29), “survival“ for survival analysis (36), “ggplot2” for visualization (37), and the various packages in the “tidyverse” for ad-hoc data analysis (38).

The source code for the analyses in this study will be made available on GitHub after publication.

Conflicts of interest

Søren Brunak reports ownerships in Intomics, Hoba Therapeutics, Novo Nordisk, Lundbeck, and ALK; and managing board memberships in Proscion and Intomics. Henning Bundgaard reports ownership in Novo Nordisk and has received lecture fees from Amgen, BMS, MSD and Sanofi.

Data access and ethics

The study was approved by the Danish Data Protection Agency (ref: 514-0255/18-3000, 514-0254/18-3000, SUND-2016-50), The Danish Health Data Authority (ref: FSEID-00003724 and FSEID-00003092), and The Danish Patient Safety Authority (3-3013-1731/1/). Danish personal identifiers were pseudonymized prior to any analysis.

Study design, methods, and results were reported in agreement with the TRIPOD statement (39).

Data availability statement

Due to national and EU regulations, the source data used in this study cannot be made publicly available. Research groups with access to secure and dedicated computing environments can request access to the source data registries via application to the Danish Health Data Authority.

Funding

This study was funded by Novo Nordisk Foundation (grant agreements: NNF14CC0001 and NNF17OC0027594) – Hellerup, Denmark; NordForsk (PM Heart; grant agreement: 90580) – Oslo, Norge; and the Innovation Foundation (BigTempHealth; grant agreement: 5153-00002B) – Aarhus, Denmark.

Results

Table 1 shows the baseline characteristics of the 52,809 ischemic heart disease patients in our derivation cohort, which were randomly allocated into a training set (n = 42,048) and a test set (n = 10,761). Figures 1A and 1B illustrate patient inclusion and the temporal distribution of the index procedures, respectively. Both sets demonstrated considerable similarity in baseline characteristics. The baseline characteristics have been further stratified by sex, to enable comparison between men and women which are known to have different disease manifestations (40). In the training set, women comprised 31.1%, and men 68.9%; in the test set, the distribution was 30.8% women and 69.2% men. The median age was consistent at 66 (IQR: 15) years for men and 70 (IQR: 15) years for women in the training set and 70 (IQR: 16) years for women in the test set.

At the time of the index coronary angiography, women had a higher prevalence of diffuse coronary atherosclerosis compared to men, with 39.9% in the training set and 39.6% in the test set for women, versus 23.3% for men in the training set and 22.8% in the test set. Conversely, men exhibited a higher incidence of two or three-vessel disease, with 41.0% in the training set and 41.4% in the test set, compared to 27.1% in women for the training set and 27.3% for the test set. Tobacco usage was higher among men, with 28.5% of men in the training set and 28.7% in the test set identified as active smokers, in contrast to 25.1% of women in the training set and 26.3% in the test set.

Medication usage was prominent, with 61.8% of men and 64.6% of women in the training set taking lipid-lowering medications, and similar trends observed in the test set (61.9% of men and 65.1% of women). Anti-hypertensive medication usage was particularly substantial, with 76.6% of men and 86.1% of women in the training set, and 76.4% of men and 86.3% of women in the test set on these drugs. Hypertension was a common diagnosis recorded in 35.2% of men and 44.3% of women in the training set, and 34.7% of men and 44.3% of women in the test set. The hypertension diagnosis in this context refers to earlier hospital admissions where an I10 diagnosis was given, explaining the discrepancy with the proportion of patients on anti-hypertensive medication.

We defined four endpoints for our time-to-event prediction models: a) All-Cause Mortality (ACMO), b) Cardiovascular Mortality (CVMO), c) Other Cardiovascular Complications (CVCO), and d) New Myocardial Ischemia Events (MIEV). For endpoints other than ACMO, we also delineated the competing risks that would prevent patients from reaching the primary endpoint of interest. With follow-up data extending up to February 2019, we developed prediction models aimed at long-term risk stratification over a five-year prediction horizon.

Figure 1C showcases Kaplan-Meier estimates of survival probability for ACMO and Aalen-Johansen estimates for the cause-specific cumulative incidences of CVMO, MIEV, and CVCO. For ACMO, the 5-year survival probability estimate is 81.7% (CI: 81.3—82.1%) for the training set and 81.6% (CI: 80.9—82.4%) for the test set. In the case of CVMO, the estimated 5-year incidence of cardiovascular death is 8.1% (CI: 7.8—8.4%) in the training set and 8.3% (CI: 7.8—8.8%) in the test set. For MIEV, we estimated a 5-year incidence of new ischemic events to be 36.8% (CI: 36.3—37.2%) in the training set and 37.3% (CI: 36.3—38.2%) in the test set. Lastly, for CVCO, the 5-year incidence of complications was found to be 37.3% (CI: 36.9—37.8%) in the training set and 37.4% (CI: 36.5—38.4%) in the test set.

We set up neural network models for prediction of each of the four endpoints. To finetune key parameters of the models, we performed hyperparameter optimization. Figure S3 shows the overview of the hyperparameter sweeps. The best performing trial, based on the integrated cause-specific IPA, yielded scores of 22.2% for ACMO, 12.6% for CVMO, 23.5% for CVCO, and 8.0% for MIEV. These scores were all calculated using a validation split of the training set. For the IPA metric, a higher score is better: a perfect model has a score of 100%, a useful model has a score above 0%, and models with an IPA $\leq 0\%$ are considered useless or harmful (27). In each of the sweeps, there were several trials with a performance in this "useless" range, which showcases the importance of careful parameter tuning. Figure S4 shows the relationship between the individual hyperparameter values and the average performance of trials using those specific values. Some hyperparameters were found to impact performance more than others. Enabling skip connections and batch normalization were in all cases associated with a better model performance. From visual inspection, learning rate, number of hidden units, number of blocks, and weight decay appears to be the parameters most critical to fine-tune for optimal performance. Table S1 shows the overview of the hyperparameters, specifying the search space used during optimization and the specific hyperparameter configuration for the best-performing trials. These final configurations were used in the training of the final models. Figure S5 shows the training history of the final models, tracking the negative log-likelihood for both the training and validation splits, as well as the IPA scores on the validation split at each training step. Across all models, we observe a concurrent increase in IPA as the negative log-likelihood decreases, with no signs of model overfitting.

The models were subsequently evaluated using the hitherto unseen test data. The prediction performance was assessed both qualitatively, through graphical representations (Figure 2), and quantitatively, using performance metrics AUC, Brier score, and IPA (Figure 3). The performance metrics are all time-dependent and were thus computed at 100 evenly spaced prediction horizons

ranging from 31 days to 5 years. For ease of presentation, we will primarily highlight the 1, 3, and 5-years predictions in the text.

Comparing the predicted t-year ACMO estimates across the observed outcomes (Figure 2), we observe the quantiles of the predicted mortality risk to be consistently higher for patients that experience the primary outcome (labeled “primary”) compared to those that do not (labeled “event-free”). This suggests good model discrimination, which is corroborated quantitatively by the AUC values. Specifically, the AUC of the ACMO model was 84.9% (CI: 83.3 - 86.6) at 1 year, 83.4% (CI: 82.3 - 84.6) at 3 years, and 83.4% (CI: 82.8 - 84.4) at 5 years (Figure 4). The IPA, which measures both discrimination and calibration, was 13.7% (CI: 7.4 - 20.1) at 1 year, 19.7% (CI: 16.0 - 23.5) at 3 years, and 25.8% (CI: 22.0 - 28.8) at 5 years. The IPA is a measure obtained by scaling the Brier score of the model with that of a null model based on observed incidence (27), and reflects both model discrimination and calibration. The Brier score for the ACMO model was significantly better (i.e., lower) than that of the null model (Kaplan-Meier) across all 100 evaluated timepoints (*Supplementary File 2*). To further assess calibration, we examined a calibration plot and found the model to be well-calibrated, with the calibration regression curve closely aligning with the 45-degree reference line (Figure 3) across all evaluated timepoints.

Examining the cause-specific CVMO predictions, we noted that the predicted risk was higher for patients experiencing cardiovascular mortality ("primary") and those with non-cardiovascular mortality ("competing") compared to those who remained event-free ("event-free") (Figure 2). Additionally, the risk quantiles for the "primary" group were consistently higher than for the "competing" group, although the difference was less pronounced. In terms of calibration, the model generally performed well, but there was some evidence of overestimation among those with the highest predicted risks (Figure 3). The number of patients at the highest end of predicted risks is however very low, so the calibration estimate there carries considerable uncertainty.

Quantifying the model performance, AUC for the CVMO model was 85.7% (CI: 83.7 - 87.7) at 1 year, 84.6% (CI: 83.1 - 86.1) at 3 years, and 82.8% (CI: 81.4 – 84.2) at 5 years. The IPA was 10.4% (CI: 1.7 – 19.0) at 1 year, 13.0% (CI: 7.1 – 19.0) at 3 years, and 14.2% (CI: 9.3 - 19.0) at 5 years. When comparing the Brier score of our PMHnet model to a null model based on observed incidence (Aalen-Johansen), our model outperformed the null model across all 100 evaluated timepoints.

We also included a single-risk naïve version of the CVMO model in the analysis, treating competing risks as censored events. This naïve model had worse AUC values at 56 of the 100 timepoints but was otherwise comparable to the competing risks model in terms of discrimination. However, the Brier score was worse at 97 out of the 100 evaluated timepoints.

For the MIEV predictions, the boxplots conditional on the t-year outcome reveal that cases undergoing with new ischemic events ("primary") generally have higher model predictions compared to those in the "competing" or "event-free" categories. However, the difference between the medians is relatively modest (Figure 2). This observation is supported by the AUC values, which are 70.1% (CI: 68.9 – 71.3) at 1 year, 69.1% (CI: 68.0 - 70.2) at 3 years, and 65.9% (CI: 64.8 - 67.1) at 5 years (Figure 4). The IPA further quantifies this with scores of 6.8% (CI: 3.8 - 9.0) at 1 year, 9.2% (CI: 7.3 - 11.1) at 3 years, and 6.8% (CI: 5.2 – 10.0) at 5 years.

The model's predictions for the test set are generally well-calibrated, except at the most extreme end of the predicted risks (Figure 3). Compared to a reference "null" model, the Brier score of the MIEV model was superior at 98 of the 100 evaluated timepoints. The same was observed for the “naïve” single-risk version of the model. However, compared to the competing risk model, this single-risk model had a worse AUC and Brier at 87 and 99 of the timepoints, respectively.

The CVCO model was also found to be well-calibrated, as seen from the calibration curve (Figure 3), and the Brier score (Figure 4). At 99 of the 100 timepoints the MACO model had significantly better brier score than the null model and 97 of 100 had better Brier score than the “naïve” single risk model. In terms of discrimination, the AUC was 79.6 (CI: 78.6-80.6) at 1 year, 79.9 (CI: 79.0-80.9) at 3 years, and 78.5 (CI: 77.5-79.4) at 5 years. Additionally, the IPA was 19.2% (CI: 16.7-21.7) at 1 year, 24.9% (CI: 23.0-26.8) at 3 years, and 23.9 (CI: 22.0-25.7) at 5 years.

Discussion

In this study, we presented the development of a collection of four different neural network-based time-to-event models for risk-prediction in ischemic heart disease. These models were developed to predict four different key outcomes following coronary angiography: all-cause mortality (ACMO), cardiovascular mortality (CVMO), cardiovascular complications (CVCO), and recurrent myocardial ischemic events (MIEV). We utilized a comprehensive dataset combining electronic health records and clinical registries, including more than 52 thousand Danish patients with ischemic heart disease and 2262 different features. Models were trained using a subset of 42 thousand patients, while validation was carried out on a separate subset of more than 10 thousand patients. From the evaluation of the model performances, we found that the cause and time-specific estimates of all models were well-calibrated and can be used to discriminate between patients experiencing the different endpoints and those who remain event-free, which points to the ability of the models to be of clinical utility.

This collection of models, which we refer to as PMHnetV2, represents an update to our previous contribution to field of machine learning-based prognostication in ischemic heart disease, PMHnetV1 (13), which was limited to the prediction of all-cause mortality. Enabling the development of the PMHnetV2 models, a key contribution of our study is the introduction of a new approach for construction of competing risk time-to-event models with neural networks. Our discrete-time approach can be viewed as an extension to the methodology first proposed by Gensheimer and Narasimhan (8) which enables jointly modelling competing risk data. The theoretical foundations of this extension is described in detail by Tutz and Schmid in the context of classical statistical analysis of discrete failure times (21), but we are, to the best of our knowledge, the first to apply this methodology to neural networks.

In the model evaluation, we compared the PMHnetV2 models based on this methodology to single-risk models that treat competing events as censoring. The single-risk version of the models was found to have good model discrimination and calibration, however in almost all cases, the models that are able account for competing risks significantly improved on both discrimination and calibration. In general, we found that the effect on calibration was the most pronounced. From this, we conclude that in prediction of endpoints with competing risks, it is more effective to adopt a methodology that allows for joint modelling of competing events, rather than the common practice of treating competing events as censoring.

In addition to our novel competing risk neural network-approach, a major strength of this study is the size of the dataset and its diversity, which enhances the applicability of our models to a broad patient population with ischemic heart disease. Where we in PMHnetV1 limited inclusion to only cover patients subject to their first ever coronary angiography, we in this study also included patients with one or more angiographies recorded prior to the inclusion period. Consequently, the patient population now include cases with more chronic manifestations of ischemic heart disease, as exemplified by the fact that 5,917 of patients had a history of one or more PCI/CABG procedures at the time of their index coronary angiography.

Despite its strengths, our study has several limitations worth highlighting. Presently, our model utilizes more than 2200 different features, many of which may be colinear and possibly redundant. There is a possibility that the model could be equally effective with a reduced number of features, however this is not an aspect that we have explored.

Moreover, most of our features are categorical and not continuous. We chose to discretize laboratory values according to their reference ranges, a decision aimed at enhancing the generalizability of feature processing and more closely aligning with clinical practice. However, an inherent limitation of our current framework is the lack of temporal resolution of features. For example, our models are unable discern whether an above-reference serum creatinine test occurred one week or one year prior to the assessment. To address this, we included hyperparameters defining the time-window for inclusion of features, specifying how far back in time data is considered. While this approach provides a partial solution, it is not ideal. Future studies should instead explore the integration of neural network architectures such as LSTM (41) or Transformers (42) for inclusion of sequential and time-resolved features.

Utilizing retrospective clinical data presented the challenge of missing variables, a phenomenon that can occur for several reasons. It could be a deliberate omission, such as the decision to forgo a specific blood test irrelevant to the diagnostic process. Alternatively, a clinician might have considered it unnecessary to document certain details, like a normal blood pressure measurement, at a given contact. Furthermore, there could be inadvertent gaps over time related to systematic issues with data storage, registration, or processing that often cannot be avoided in a dynamic clinical environment. Rather than presuming the missingness to be random and hence suitable for imputation through sophisticated methods, we chose to explicitly encode where data was missing. This approach preserves the integrity of the existing data and leverages the missing information to enhance the predictive capabilities of the model, resulting in a model that is both robust and reflective of real-world clinical scenarios.

Lastly, we found it very difficult to create the perfect algorithmic definition for determining the timing of disease recurrence and progression from large retrospective clinical. As a pragmatic work around, the endpoints cardiovascular complications (CVCO) and recurrent myocardial ischemic events (MIEV) were defined from surrogate markers, such as revascularization (PCI/CABG) and hospital admission with critical diagnoses as the pragmatic prediction target. These events, reliably identified from registry data, serve as indicators of development or worsening of symptoms in such a degree that it warranted clinical intervention.

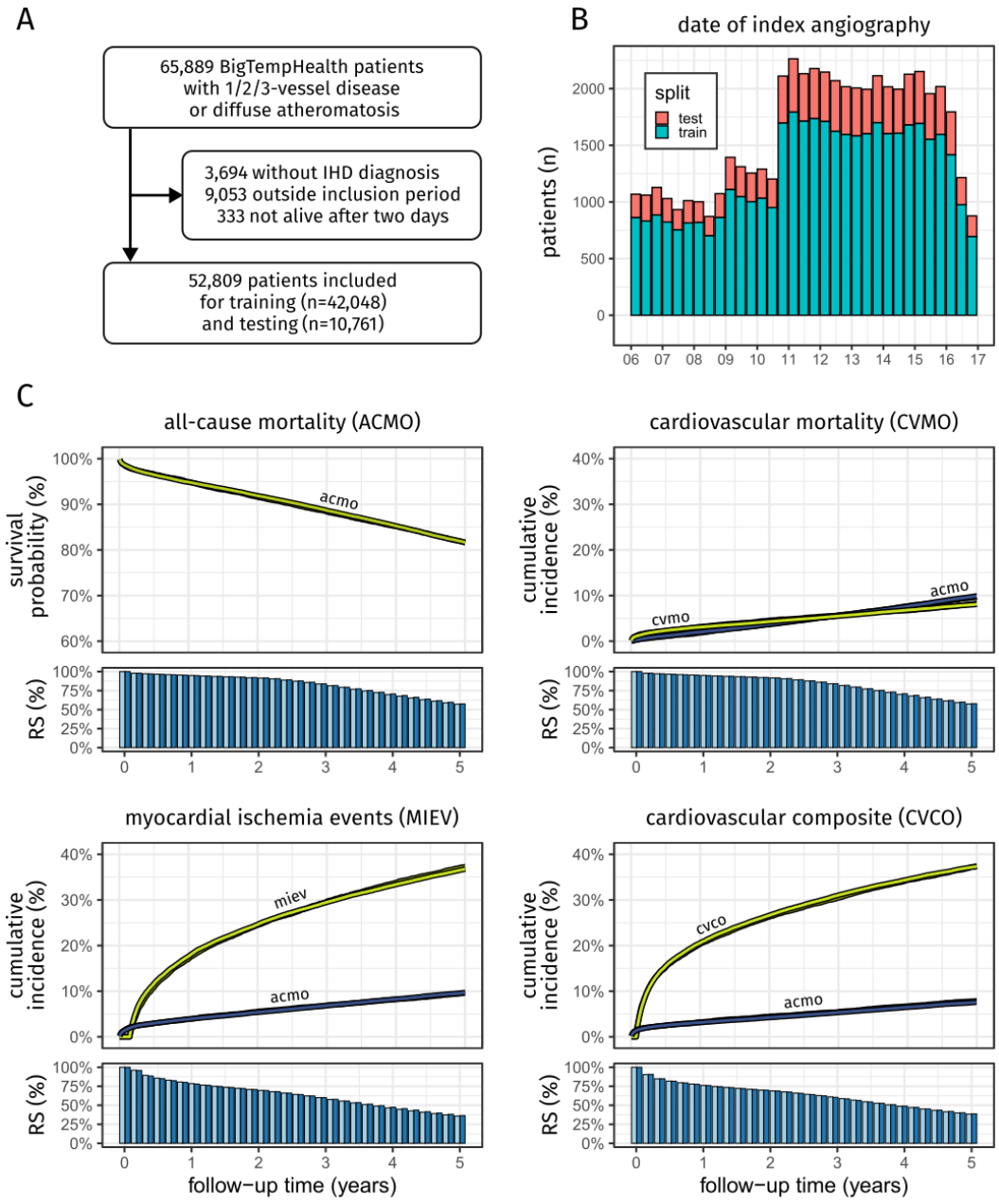
In conclusion, our study makes a significant contribution to the field by implementing and testing a novel neural network approach for modeling survival data with competing risks in patients with ischemic heart disease.

Tables

		Training (n = 42,048)		Test (n = 10,761)	
Variable	Value	Male (n = 28,980)	Female (n = 13,068)	Male (n = 7,442)	Female (n = 3,319)
Age (years)	Median (IQR)	66 (15)	70 (15)	66 (15)	70 (16)
Height (cm)	Median (IQR) [NA%]	177 (9) [8.0%]	164 (8) [7.6%]	177 (9) [8.1%]	164 (8) [7.4%]
Weight (kg)	Median (IQR) [NA%]	85 (19) [4.7%]	70 (20) [4.9%]	85 (19) [5.1%]	70 (20) [4.9%]
Systolic blood pressure (mmHg)	Median (IQR) [NA%]	138 (32) [24.2%]	140 (32) [23.9%]	138 (31) [24.0%]	140 (33) [22.8%]
Diastolic blood pressure (mmHg)	Median (IQR) [NA%]	80 (18) [24.2%]	76 (18) [23.9%]	80 (19) [24.0%]	76 (19) [22.8%]
Coronary vessel pathology	Diffuse coronary atherosclerosis	6,760 (23.3%)	5,213 (39.9%)	1,696 (22.8%)	1,314 (39.6%)
	Single-vessel disease	10,318 (35.6%)	4,314 (33.0%)	2,669 (35.9%)	1,101 (33.2%)
	Two-vessel disease	5,808 (20.0%)	1,917 (14.7%)	1,471 (19.8%)	490 (14.8%)
	Three-vessel disease	6,094 (21.0%)	1,624 (12.4%)	1,606 (21.6%)	414 (12.5%)
Tobacco usage	Active	8,272 (28.5%)	3,283 (25.1%)	2,136 (28.7%)	873 (26.3%)
	Former	11,771 (40.6%)	4,204 (32.2%)	3,058 (41.1%)	1,035 (31.2%)
	Never	6,361 (21.9%)	4,505 (34.5%)	1,637 (22.0%)	1,103 (33.2%)
	Missing	2,576 (8.9%)	1,076 (8.2%)	611 (8.2%)	308 (9.2%)
Same day PCI		11,503 (39.7%)	4,340 (33.2%)	3,011 (40.5%)	1,061 (32.0%)
Previous PCI/CABG	0	25,488 (88.0%)	11,825 (90.5%)	6,582 (88.4%)	2,997 (90.3%)
	1	2,677 (9.2%)	986 (7.6%)	673 (9.0%)	255 (7.7%)
	2+	815 (2.8%)	257 (2.0%)	187 (2.5%)	67 (2.0%)
Medication	Lipid-lowering	17,908 (61.8%)	8,438 (64.6%)	4,610 (61.9%)	2,160 (65.1%)
	Anti-hypertensive	22,189 (76.6%)	11,252 (86.1%)	5,684 (76.4%)	2,863 (86.3%)
	Non-insulin glucose lowering medication	5,304 (18.3%)	2,219 (17.0%)	1,360 (18.3%)	587 (17.7%)
	Insulin	1,909 (6.6%)	994 (7.6%)	475 (6.4%)	239 (7.2%)
Diagnoses (ICD-10)	Diabetes (E10, E11)	4,650 (16.0%)	2,144 (16.4%)	1,153 (15.5%)	543 (16.4%)
	Hypertension (I10)	10,211 (35.2%)	5,794 (44.3%)	2,579 (34.7%)	1,470 (44.3%)
	COPD (J44)	2,017 (7.0%)	1,431 (11.0%)	512 (6.9%)	358 (10.8%)
	Heart failure (I50)	4,107 (14.2%)	1,611 (12.3%)	1,024 (13.8%)	440 (13.3%)
	Atrial fibrillation or atrial flutter (I44)	3,523 (12.2%)	1,556 (11.8%)	880 (11.8%)	378 (11.4%)

Table 1: Baseline statistics of the study population stratified by dataset split (training/test) and sex (male/female). Numbers in this table exclusively represent data available at the time origin. Medication was defined from the prescription database (LMSR) as a filled prescription of a drug belonging to the class of *lipid-lowering medication* (ATC: C10), *anti-hypertensive medication* (C02, C03, C07, C08, C09), *type-2 diabetes medicine* (A10B), or *insulin* (A10A). Numbers are reported as "median (IQR)", "median (IQR) [NA%]", or "count (%)". CABG: coronary artery bypass grafting, IQR: Interquartile range. PCI: percutaneous coronary intervention.

396 **Figures**



397

398 ***Figure 1:*** *Inclusion diagram (A), overview of inclusion dates (B), and observed survival and cumulative*
399 *incidence curves across the four primary endpoints (C).*

400

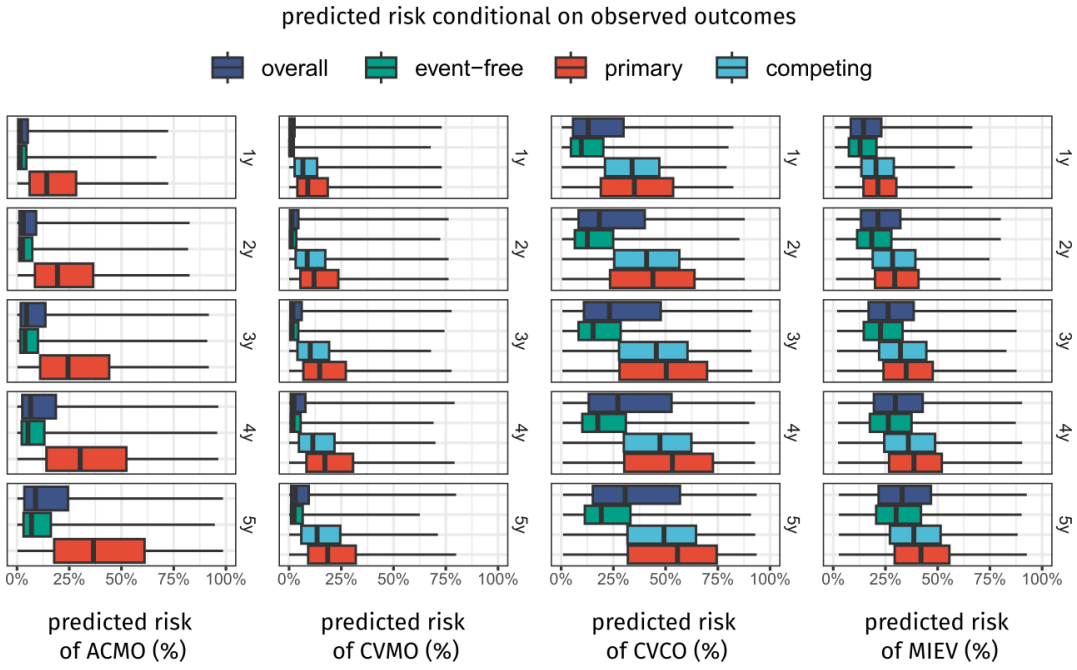
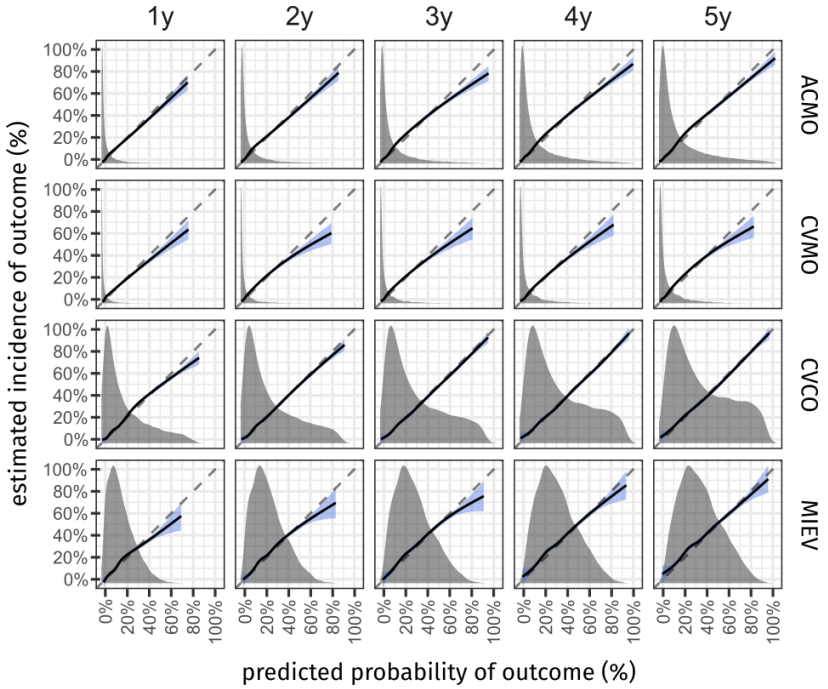


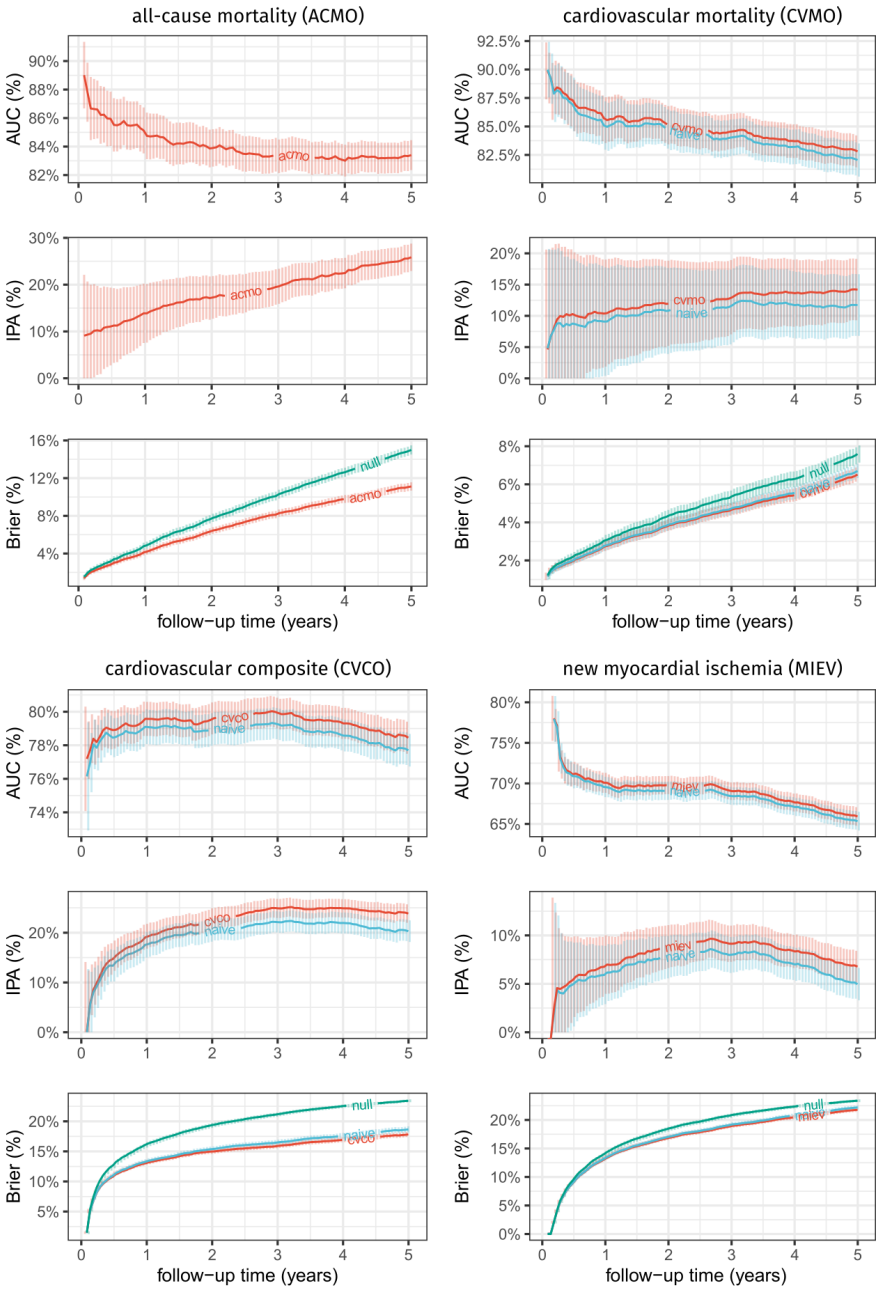
Figure 2: *Visual assessment of test-set discrimination. Boxplots showing the quantiles of predicted t-year risks of the primary endpoints conditional on the observed t-year outcome. The "overall" boxplot shows the observed quantiles in the test set, and boxplots labelled "primary", "competing", and "event-free" are estimated using inverse probability of censoring weighting.*



406

407 ***Figure 3:*** *Visual assessment of test-set calibration. Calibration plots showing t-year risk estimates*
408 *incidences observed in the test set. The regression line is obtained by fitting a natural cubic spline*
409 *with 6 degrees of freedom on pseudo-values obtained through jackknife resampling of the Kaplan-*
410 *Meier (ACMO) or Aalen-Johansen (CVMO, RVSC, and MACO) estimates. The dashed 45-degree*
411 *reference lines indicate perfect calibration. The density curve in the background of each panel shows*
412 *the distribution of the model estimates.*

413



414

415 ***Figure 4**: Evaluation of test-set performance across the four outcomes. Panels in A, B, C, and D display*
416 *performance metrics for the all-cause mortality model (ACMO), cardiovascular mortality model*
417 *(CVMO), de-novo revascularization model (RVSC), and the major complications model (MACO),*
418 *respectively. Metrics include the time-dependent AUC (AUC), the index of prediction accuracy (IPA),*
419 *and the Brier score. These metrics are calculated at 100 evenly spaced intervals ranging from 31 days*
420 *to 10 years. In the Brier score panel, the "null" curve represents the Brier score based on either the*
421 *Kaplan-Meier estimate for ACMO or the Aalen-Johansen estimate for the primary outcomes in CVMO,*
422 *RVSC, and MACO. For models with competing risks (CVMO, RVSC, MACO), an additional curve shows*
423 *the performance of a naïve model that treats competing events as censored.*

References

1. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019 Jan;25(1):44–56.

2. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. N Engl J Med. 2019 Apr 4;380(14):1347–58.

3. Steinfeldt J, Buergel T, Loock L, Kittner P, Ruyoga G, Belzen JU zu, et al. Neural network-based integration of polygenic and clinical information: development and validation of a prediction model for 10-year risk of major adverse cardiac events in the UK Biobank cohort. Lancet Digit Health. 2022 Feb 1;4(2):e84–94.

4. D'Ascenzo F, Filippo OD, Gallone G, Mittone G, Deriu MA, Iannaccone M, et al. Machine learning-based prediction of adverse events following an acute coronary syndrome (PRAISE): a modelling study of pooled datasets. The Lancet. 2021 Jan 16;397(10270):199–207.

5. Pieszko K, Shanbhag AD, Singh A, Hauser MT, Miller RJH, Liang JX, et al. Time and event-specific deep learning for personalized risk assessment after cardiac perfusion imaging. Npj Digit Med. 2023 May 1;6(1):1–11.

6. Lee C, Zame W, Yoon J, Schaar M van der. DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks. Proc AAAI Conf Artif Intell [Internet]. 2018 Apr 26 [cited 2023 Sep 8];32(1). Available from: https://ojs.aaai.org/index.php/AAAI/article/view/11842

7. Steele AJ, Denaxas SC, Shah AD, Hemingway H, Luscombe NM. Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. Singh TR, editor. PLOS ONE. 2018 Aug 31;13(8):e0202344.

8. Gensheimer MF, Narasimhan B. A scalable discrete-time survival model for neural networks. PeerJ. 2019 Jan 25;7:e6257.

9. Nielsen AB, Thorsen-Meyer HC, Belling K, Nielsen AP, Thomas CE, Chmura PJ, et al. Survival prediction in intensive-care units based on aggregation of long-term disease history and acute physiology: a retrospective study of the Danish National Patient Registry and electronic patient records. Lancet Digit Health. 2019 Jun 1;1(2):e78–89.

10. Zhao L, Feng D. Deep Neural Networks for Survival Analysis Using Pseudo Values. IEEE J Biomed Health Inform. 2020 Nov;24(11):3308–14.

11. Kvamme H, Borgan Ø. Continuous and discrete-time survival prediction with neural networks. Lifetime Data Anal. 2021 Oct 1;27(4):710–36.

12. Kvamme H, Borgan Ø, Scheel I. Time-to-Event Prediction with Neural Networks and Cox Regression [Internet]. arXiv; 2019 [cited 2023 Sep 8]. Available from: http://arxiv.org/abs/1907.00825

13. Holm PC, Haue AD, Westergaard D, Röder T, Banasik K, Tragante V, et al. Development and validation of a neural network-based survival model for mortality in ischemic heart disease [Internet]. medRxiv; 2023 [cited 2023 Sep 8]. p. 2023.06.16.23291527. Available from: https://www.medrxiv.org/content/10.1101/2023.06.16.23291527v1

14. Kleinbaum DG, Klein M. Survival Analysis: A Self-Learning Text, Third Edition. 3rd edition. New York, NY: Springer; 2011. 715 p.

15. Schmidt M, Pedersen L, Sørensen HT. The Danish Civil Registration System as a tool in epidemiology. Eur J Epidemiol. 2014 Aug 1;29(8):541–9.

16. Helweg-Larsen K. The Danish Register of Causes of Death. Scand J Public Health. 2011 Jul;39(7 Suppl):26–9.

17. Kildemoes HW, Sørensen HT, Hallas J. The Danish National Prescription Registry. Scand J Public Health. 2011 Jul;39(7 Suppl):38–41.

18. Schmidt M, Schmidt SAJ, Sandegaard JL, Ehrenstein V, Pedersen L, Sørensen HT. The Danish National Patient Registry: a review of content, data quality, and research potential. Clin Epidemiol. 2015 Nov;449.

19. Özcan C, Juel K, Flensted Lassen J, von Kappelgaard LM, Mortensen PE, Gislason G. The Danish Heart Registry. Clin Epidemiol. 2016 Oct 25;8:503–8.

20. Schmidt M, Andersen LV, Friis S, Juel K, Gislason G. Data Resource Profile: Danish Heart Statistics. Int J Epidemiol. 2017 Oct 1;46(5):1368–1369g.

21. Tutz G, Schmid M. Modeling Discrete Time-to-Event Data. 1st ed. 2016 edition. New York, NY: Springer; 2016. 257 p.

22. Gorishniy Y, Rubachev I, Khrulkov V, Babenko A. Revisiting Deep Learning Models for Tabular Data [Internet]. arXiv; 2023 [cited 2023 Sep 14]. Available from: http://arxiv.org/abs/2106.11959

23. Loshchilov I, Hutter F. Decoupled Weight Decay Regularization [Internet]. arXiv; 2019 [cited 2023 Sep 14]. Available from: http://arxiv.org/abs/1711.05101

24. Smith LN, Topin N. Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates [Internet]. arXiv; 2018 [cited 2023 Sep 14]. Available from: http://arxiv.org/abs/1708.07120

25. Coleman C, Narayanan D, Kang D, Zhao T, Zhang J, Nardi L, et al. DAWNBench: An End-to-End Deep Learning Benchmark and Competition.

26. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A Next-generation Hyperparameter Optimization Framework [Internet]. arXiv; 2019 Jul. Report No.: 1907.10902. Available from: https://arxiv.org/abs/1907.10902

27. Kattan MW, Gerds TA. The index of prediction accuracy: an intuitive measure useful for evaluating risk prediction models. Diagn Progn Res. 2018 May 4;2(1):7.

28. Schumacher M, Graf E, Gerds T. How to Assess Prognostic Models for Survival Data: A Case Study in Oncology. Methods Inf Med. 2003;42(5):564–71.

29. Gerds TA, Kattan MW. Medical risk prediction models: with ties to machine learning. 1st ed. Boca Raton: CRC Press; 2021. (Chapman & hall/crc biostatistics series).

30. Andersen PK, Perme MP. Pseudo-observations in survival analysis. Stat Methods Med Res. 2010 Feb;19(1):71–99.

31. Geloven N van, Giardiello D, Bonneville EF, Teece L, Ramspek CL, Smeden M van, et al. Validation of prediction models in the presence of competing risks: a guide through modern methods. BMJ. 2022 May 24;377:e069249.

32. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library [Internet]. arXiv; 2019 [cited 2023 Sep 18]. Available from: http://arxiv.org/abs/1912.01703

33. Falcon W, The PyTorch Lightning team. PyTorch Lightning [Internet]. 2019 [cited 2023 Sep 18]. Available from: https://github.com/Lightning-AI/lightning

34. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12(85):2825–30.

35. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2019. Available from: https://www.R-project.org/

36. Therneau TM. A Package for Survival Analysis in S.

37. Wickham H. ggplot2. WIREs Comput Stat. 2011;3(2):180–5.

38. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, et al. Welcome to the Tidyverse. J Open Source Softw. 2019 Nov 21;4(43):1686.

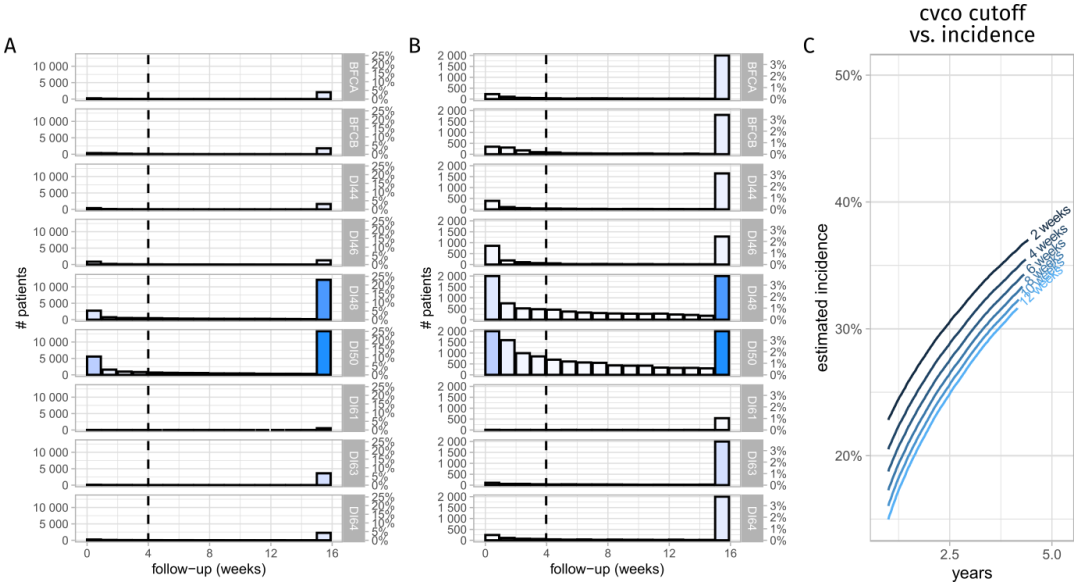
39. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD). Circulation. 2015 Jan 13;131(2):211–9.

40. Aggarwal NR, Patel HN, Mehta LS, Sanghani RM, Lundberg GP, Lewis SJ, et al. Sex Differences in Ischemic Heart Disease. Circ Cardiovasc Qual Outcomes. 2018 Feb;11(2):e004437.

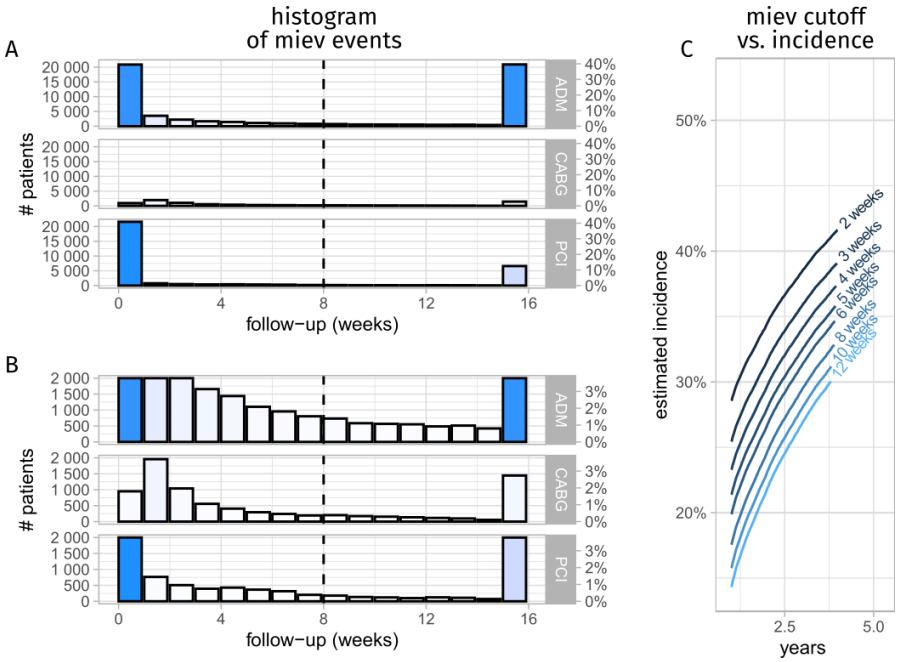
41. Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Comput. 1997 Nov;9(8):1735–80.

42. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is All you Need. In: Advances in Neural Information Processing Systems [Internet]. Curran Associates, Inc.; 2017 [cited 2023 Dec 22]. Available from: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

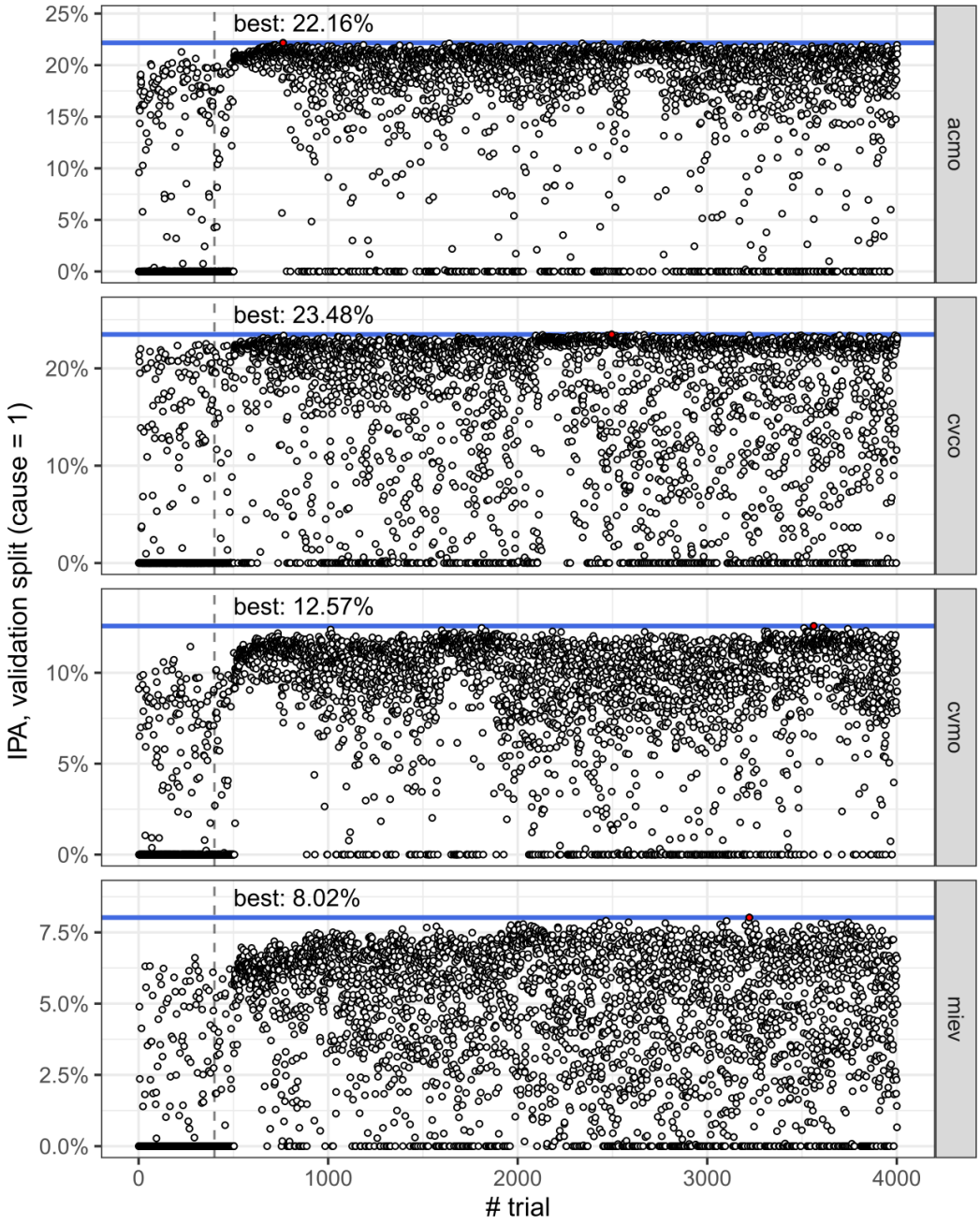
Supplementary figures



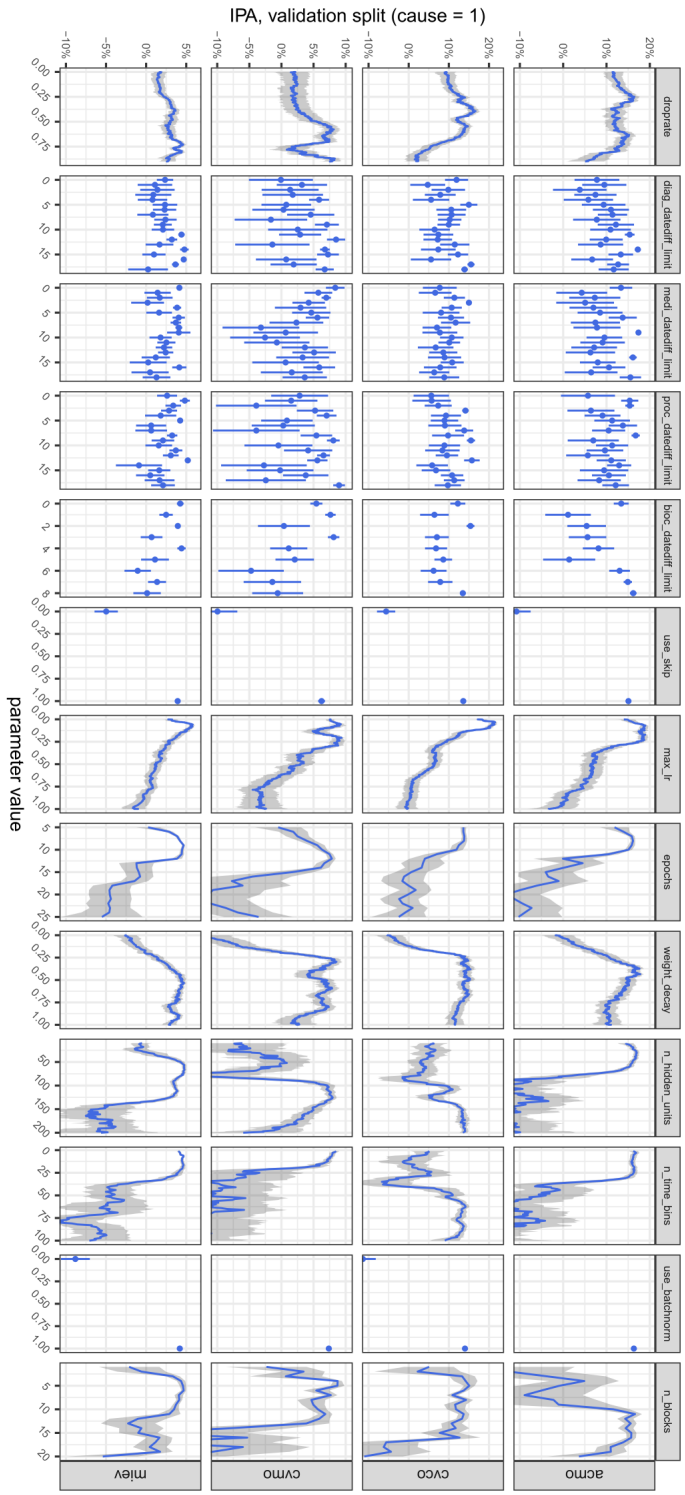
Supp. Figure 1: Time to any of the cardiovascular composite (CVCO) endpoints after index coronary angiography. A) shows the overall counts, and B) provides a zoomed in view of the same data. Bars show the number of unique patients with a CVCO procedure (BFCA, BFCB) or admission (I44, I46, I48, I50, I61, I63, and I64) in each week following the coronary angiography (same day events also included). X-axis have been limited to 16 weeks, and events after that have been aggregated. The dashed vertical lines show the 4-week blanking window used to separate index-related events from progression-related events (the CVCO endpoint). C) Show the effect of the varying the duration of the blanking window.



Supp. Figure 2: Time to recurrent myocardial ischemic events (MIEV), A) shows the overall counts, and B) provides a zoomed in view of the same data. Bars shows the number of unique patients with an MIEV admission (I20-I25) or revascularization procedure (PCI/CABG) in each week following the index coronary angiography. X-axis have been limited to 16 weeks, and events after that have been aggregated. The last bar thus shows how many distinct patients have a future admission with each of the complications. The dashed vertical lines show the 8-week cutoff we use to separate index-related admissions from disease progression complications (the MIEV endpoint). C) Shows the effect of varying the duration of the washout/cutoff period.



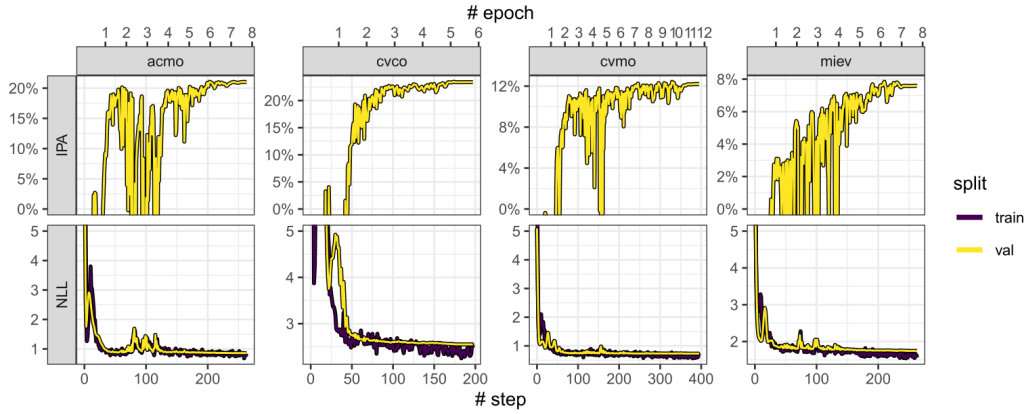
Supp. Figure 3: performance of trials in hyperparameter sweeps for the four different models. The optimization metric is the integrated IPA calculated on a validation split of the training set. Trials with a negative IPA have been clipped to 0%. The blue line shows the performance of the best performing trials. The dashed vertical line shows the transition from random sampler to TPE sampler after the 400 startup trials.



Supp. Figure 4: relationship between hyperparameter values and objective values of trials in the hyperparameter sweeps. Each panel shows the rolling average IPA ($\pm SD$) (y-axis) with a sliding window covering one decile of the hyperparameter values (x-axis).

546

547 .



Supp. Figure 5: Training history for the final models using the fine-tuned hyperparameters. Plot shows the negative log-likelihood (NLL) and cause-specific integrated IPA (IPA) calculated after each batch (step).

Supplementary tables

Hyperparameter	Search space	ACMO	CVMO	CVCO	MIEV
Enable skip connections	True or False	True	True	True	True
Enable batch normalization	True or False	True	True	True	True
Maximum Learning Rate	1e-3 to 1	0.204	0.072	0.064	0.076
Number of Epochs	5 to 25	8	12	6	8
Weight Decay	1e-5 to 1	0.3774	0.5269	0.2412	0.7359
Dropout rate	0% to 90%	28.4%	86.9%	25.9%	78.7%
Number of ResBlocks	1 to 20	15	5	6	7
Number of hidden units in ResBlocks	10 to 100	30	105	147	104
Number of time bins	1 to 60	23	6	84	9
Biochemical cutoff	0.5 to 5 years (stepsize: 0.5)	4.5	1	4.5	1.5
Diagnosis cutoff	0.5 to 10 years (stepsize: 0.5)	7.5	7.5	9.5	6.5
Medication cutoff	0.5 to 10 years (stepsize: 0.5)	5	1.5	2	4
Procedure cutoff	0.5 to 10 years (stepsize: 0.5)	7.5	5	2	4.5

Supp. Table 1: Overview of hyperparameters, the search space explored, and the best performing configurations found following hyperparameter optimization.

Supplementary Methods

Inclusion criteria

From the Eastern Danish Heart Registry (PATS), we identified all coronary angiographies where a coronary pathology of diffuse atheromatosis or 1/2/3-vessel disease was diagnosed, and then ascertained that patients had a prior diagnosis code (ICD10) of I20-25 (ischemic heart disease), R07.4 (chest pain), Z03.4 (suspected myocardial infarction), or Z03.5C (suspected stable angina) using the Danish National Patient Registry (DNPR). We then limited the coronary angiographies to those performed between 1st of January 2006 and 31st of December 2016. All but the first of the remaining procedures for each patient were then discarded and we limited the study population to individuals above 18 years of age. Finally, we discarded patients that were not alive two days after the procedure.

Endpoint definition

All-cause mortality (ACMO) was defined from the "t_person" table in the CPR where it is recorded as "C_STATUS = 90" and used "D_STATUS_HEN_START" as the event date. For patients recorded as being alive, "C_STATUS = 0", we used the last update date of the CPR as the censoring date.

Cardiovascular mortality (CVMO) was defined from the Cause of Death Register (DAR) as any death with an ICD-10 code of I00-99 registered as the underlying cause of death. We used the date from "D_STATDATO" to determine the date. We used the latest update date of the DAR, 2019-02-15, as the censoring date. Deaths from other causes were treated as competing risks.

Cardiovascular complications (CVCO) were defined from the National Patient Register (DNPR) and included hospital admissions and procedures. The hospital admissions) of interest were here any in-patient admission ("C_PATTYPE = 0") a primary diagnosis of "heart failure" (ICD-10: I50), "atrial fibrillation or flutter" (I48), "cardiac arrest" (I46), or "cerebrovascular accident" (I61, I63-64). In addition, procedure codes corresponding to implantation of pacemaker (SKS: BFCA0*) or cardioverter-defibrillator (BFCB0*) was also included. To differentiate between admissions and procedures related to the index coronary angiography and unplanned admissions and procedures that represent disease progression, we introduced a wash-out or blanking period, which we set to 4-weeks (Fig S1).

Recurrent myocardial ischemia events (MIEV) were likewise defined from in-hospital admissions registered in the DNPR. We defined recurrent events as a) hospitalizations with a duration longer than 24 hours with a primary diagnosis of ischemic heart disease (ICD-10: I20-25) and b) unplanned

percutaneous coronary intervention (PCI) and coronary bypass grafting (CABG) procedures (PCI: "^KFNG.*" and CABG: "^KFN[A-E].*") as recorded in the "t_sksopr" table. For a) we used admission date + 1 day as the event time and for b) we used the procedure date. Similar to the CVCO outcome, we wanted to differentiate between admissions directly related to the initial coronary angiography and unplanned ones representing disease progression. As a pragmatic approach, we decided on using a blanking-window here used an 8-week cutoff (Fig S3). All events prior to that cutoff were therefore ignored, and the earliest admission with any of the MIEV events was used as the primary endpoint.

Hyperparameter optimization

We performed hyperparameter optimization using the Optuna hyperparameter optimization package for python (26). For hyperparameter sweeps, we ran 4,000 different trials using the Tree-structured Parzen Estimator sampling algorithm included in Optuna. We used 400 startup trials and otherwise relied on the default settings of Optuna. Since we constrained our models to only train for a maximum of 25 epochs and training thus did not take long, we did not set up pruning in any of our sweeps.

Statistics

Calibration plots were generated by fitting natural cubic splines with six degrees of freedom to the model's estimates and pseudo-values, which were derived via jackknife estimation of the marginal cumulative incidence. This methodology is described in BMJ 2022;377:e069249 and DOI: 10.1177/0962280209105020). To aid interpretation, we overlaid the regression curve on a kernel density estimate of the model predictions at the specific prediction horizons.