DEEP LEARNING FOR CAUSAL INFERENCE ON ELECTRONIC HEALTH RECORDS

Shishir Rao

St Catherine's College
Nuffield Department of Women's and Reproductive Health
University of Oxford

Thesis submitted for the degree of Doctor of Philosophy (DPhil)
March 2023

Supervised by:
Professor Kazem Rahimi

Dedicated to my late Thātha, Ajji, and Rocky

“Absence of evidence is not evidence of absence.”

-Carl Sagan

DECLARATION

This dissertation is the result of my own work and includes nothing, which is the outcome of work done in collaboration except where specifically indicated in the text. It has not been previously submitted, in part or whole, to any university of institution for any degree, diploma, or other qualification.

In accordance with the University of Oxford's Medical Science Division guidelines, this thesis does not exceed 50,000 words (exclusive of bibliography, appendices, diagrams, and tables).

Signed:

Date: 19 March 2023

Shishir Rao, MSc (Oxon)

ABSTRACT

Cardiovascular diseases (CVD) are the leading causes of mortality around the world and disentangling cause and effect is central to better understanding and treating these diseases. While randomised clinical trials are the "gold standard" of assessing the effect of an intervention, some hypotheses cannot be feasibly tested in the randomised setting. In these cases, observational studies with appropriate methods of confounding adjustment can deliver reliable evidence concerning the association between an exposure and outcome. Indeed, trusted conventional statistical models guided by subject area experts for confounder selection have been used to estimate associations in many observational studies; however, in the observational studies for which confounding is unknown and/or the population suffers from complex illness, the conventional approaches render insufficiently adjusted estimates. In parallel, recently, there has been unprecedented access to nationally representative multimodal electronic health record (EHR) datasets and advances in statistical learning including "deep" machine learning, a form of machine learning that relies on automatic feature capture dissolving the need for expert-driven feature engineering.

In this doctoral research, the aim was to develop a deep learning approach for causal inference on EHR. To do so in a structured way, the research was split into three investigations:

1) The development of a deep learning model for EHR data and assessment of risk prediction performance
2) Given the "black box" nature of deep learning modelling, the development of methods to explain the proposed model.
3) The derivation of a model for causal inference, and application of the models for association estimation in elderly/at-risk patient subgroups.

The model, Bidirectional EHR Transformer (BEHRT) was created for EHR representation learning and risk prediction. The model outperformed several benchmarks for risk prediction on a variety of tasks including incident heart failure prediction. Furthermore, in the second work, explainability investigations yielded that the model captured validated factors of risk (e.g., hypertension, diabetes, and other diseases) and offered several more factors that could be potentially preventative of incident heart failure. Lastly, a derivation of BEHRT was developed for association estimation, Targeted-BEHRT, that fused advances in deep learning and semi-parametric statistics. The model demonstrated superior estimation abilities on several simulated data experiments, and was applied to better understand the effects of antihypertensives, blood pressure, and paracetamol on cardiovascular endpoints, mortality, and other outcomes in at-risk patients.

Overall, the doctoral research has made advances in both methodological and clinical cardiovascular research. While the research focuses on developing methods for the study of cardiovascular diseases, the methods developed and tested have several important implications for epidemiological research in the observational setting at large. Especially in patient groups with pre-existing health issues, the causal models developed can be a more appropriate approach for association analysis than conventional statistical ones. In terms of clinical impact, the research has progressed our understanding of risk and protection in the context of CVD.

ACKNOWLEDGEMENTS

To begin with, I am very thankful to my primary supervisor, Professor Kazem Rahimi. With little exaggeration, I truly mean it when I say: without you, my doctoral research would not have been possible. In South Asian culture, a *Guru* is one who is not just a master of a discipline but also, a mentor and guide in addition to being an expert; you have indeed been a *Guru* to me, Kazem. It is well appreciated that you are a master of multiple disciplines but perhaps as importantly, your zeal to learn and grow is boundless and your work ethic, humbling. I thank you for supervising me on the doctoral journey, mentoring me in all things relevant to science, professional/academic growth, and life itself, and for being a guide when things were darkest. It is a privilege to be your student.

I thank my secondary supervisors: Professor Thomas Lukasiewicz for his valuable supervision on the deep learning components of the research and Drs. Mohammadhossein Mamouei and Jose Roberto Ayala Solares for their guidance on methodological aspects of the causal inference and risk prediction projects respectively.

I thank my colleagues and fellow doctoral students within the Deep Medicine research group for their collaboration, friendship, and guidance: Yikuan Li, Dr. Reza Khorshidi, Dr. Dexter Canoy, Milad Nazarzadeh Dr. Abdelaali Hassaine, Dr. Rema Ramakrishnan, Dr. Jing Huang, Zeinab Bidel, Emma Copland, and last but definitely not least, Naseem Akhtar. I have made some lifelong friends during my time in the Deep Medicine.

I thank Professor Krina Zondervan and the Nuffield Department of Women's and Reproductive Health (NDWRH) for your constant support of our group and my research. Specifically, I thank the Director of Graduate Studies, Professor Karl Morten and Delphine Vaneecke for offering me valuable pastoral support in addition to advising me throughout the various doctoral program checkpoints.

In terms of funding, I thank Professor Kazem Rahimi, the NDWRH, and the Alan Turing Institute for funding various parts of my doctoral research. I am grateful for the generous funding and stipend I have received over the past several years.

I would like to thank all of you who have provided me friendship, guidance, and comfort even though you were not directly involved in my academic journey. I am grateful to my parents and grandparents for sacrificing so very much to help me achieve my dreams. There is so much that you have done for me that is truly incalculable; I am floored by the love and support you have given me, and it is deeply touching. Thank you.

I am very grateful to Dr. Aashika Sekar, my partner, who has been my rock over the past four years: a chief cheerleader in the good times and a source of inspiration and energy in the poorer times. From sharing late-night instant noodles, to watching dog videos, to helping me with "cosmetic surgery" on various figures across papers, I am thankful to have you as a ray of light in my life in every way imaginable. Also, I thank various friends and well-wishers in Oxford, the Bay Area, New York, and abroad (in no particular order): Maddie Mitchell, Ameya Rao, Shawn Esmaili, Mark Rapaport, Olina Stathopoulou, Sucheta Korwar, Nikhil Gowd, Rizwan uncle, Rahina auntie, Peter Minary, Sunny Dasgupta, Rob Whitehurst, Ryan Harding, Amar Ghag, Tom Revell, Joana Bessa, Gerardo Garcia, Kat Friege, Greta Galeotti, Prachi Prasad, Amr Tamimi, Neha Venkatesh, Sahana Venkatesh, and many others who have been cheering me on all the way to the finish line. In terms of those elements that are inanimate but have provided me great comfort: Workhouse coffee, Currydor biryani, Sankethi Adukale meals, Maggi instant noodles, the falafel wrap place in the Gloucester Green square, Pockets falafel wraps (and Gloucester Green market food in general with the DM family), and Gail's chocolate chip cookies.

Last but definitely not lowest in priority, my musical family has been a core force of growth and definition for over two decades. My first love has and will always be music, and I would not have found the path of research if it were not for my education in Indian classical music. For this reason, I am indelibly grateful to my first *Gurus*, Sri. Nachiketa Yakkundi, Sri. Vivek Datar, and Sri. Ragavan Manian. In addition to being guides in life itself, you have given me sight to allow me to appreciate the beauty that lies in depth and to meticulously and incrementally build on work of the masters of the past – all integral to the scientific process and science itself. Thank you.

CONTENTS

LIST OF PUBLICATIONS

Publications contributing to DPhil:

Li Y*, **Rao S***, et al. BEHRT: Transformer for Electronic Health Records. *Scientific Reports* 2020.

Rao S*, Li Y*, et al. An explainable Transformer-based deep learning model for the prediction of incident heart failure. *IEEE Journal of Biomedical and Health Informatics* 2022.

Rao S, et al. Targeted-BEHRT: Deep Learning for Observational Causal Inference on Longitudinal Electronic Health Records. *IEEE Transactions on Neural Networks and Learning Systems* 2022.

Rao S, et al. Systolic blood pressure and cardiovascular risk in patients with diabetes: a prospective cohort study. Accepted for publication in *Hypertension* 2023.

Rao S, et al. Systolic blood pressure, chronic obstructive pulmonary disease, and cardiovascular risk: a prospective cohort study, Unpublished (Under review at *Heart*) 2023.

Rao S, et al. Association of sodium-based paracetamol with changes in systolic blood pressure, cardiovascular events, and all-cause mortality: a cohort study. Unpublished (Under review) 2023.

Other relevant publications:

Li Y, Mamouei M, Salimi-Khorshidi G, **Rao, S**, et al. Hi-BEHRT: Hierarchical Transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records. *IEEE Journal of Biomedical and Health Informatics* 2022

Canoy D, Tran J, Zottoli M, Ramakrishnan R, Hassaine A, **Rao S**, et al. Association between cardiometabolic disease multimorbidity and all-cause mortality in 2 million women and men registered in UK general practices. *BMC Medicine* 2022

Li Y, **Rao S**, et al. Deep Bayesian Gaussian processes for uncertainty estimation in electronic health records. *Scientific Reports* 2021.

Ramakrishnan R, **Rao S**, and He J. R. Perinatal health predictors using artificial intelligence: A review. *Women's Health* 2021.

Hassaine A, Canoy D, Solares J. R A, Zhu Y, **Rao S**, et al. Learning multimorbidity patterns from electronic health records using Non-negative Matrix Factorisation. *Journal of Biomedical Informatics* 2020.

*equal contribution.

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviation	Definition
ACEI	Angiotensin Converting Enzyme Inhibitor
AI	Artificial Intelligence
APC	Admitted Patient Care
ARB	Angiotensin Receptor Blocker
AUPRC	Area Under the Precision Recall Curve
AUROC	Area Under the Receiver Operator Characteristic
BART	Bayesian Additive Regression Tree
BB	Beta Blocker
BP	Blood Pressure
BEHRT	Bidirectional Electronic Health Records Transformer
BERT	Bidirectional Encoder Representations from Transformers
BMI	Body Mass Index
BNF	British National Formulary
CCB	Calcium Channel Blocker
CEVAE	Causal Effect Variational Autoencoder
CKD	Chronic Kidney Disease
CNN	Convolutional Neural Network
COPD	Chronic Obstructive Pulmonary Disease
COVID-19	Coronavirus disease 2019
CPRD	Clinical Practice Research Datalink
CV-TMLE	Cross Validated Targeted Maximum Likelihood Estimation
CVD	Cardiovascular Disease
DAG	Directed Acyclic Graph
DBP	Diastolic blood pressure
Deepr	Deep record
ECG	Electrocardiogram
EHR	Electronic Health Records
ELU	Exponential Linear Unit
ESC	European Society of Cardiology
GP	General Practice
GPU	Graphical Processing Unit
GRU	Gated Recurrent Unit
HDL	High Density Lipoprotein

HES	Hospital Episode Statistics
HF	Heart Failure
ICD-10	10th revision of the International Statistical Classification of Diseases and Related Health Problems
IHD	Ischaemic Heart Disease
IMD	Index of Multiple Deprivation
IQR	Interquartile Range
IPTW	Inverse Probability Treatment Weighting
ISAC	Independent Scientific Advisory Committee
JBHI	Journal of Biomedical Health Informatics
LABA	Long-Acting Beta Agonist
LDL	Low-density lipoprotein
LR	Logistic Regression
LSTM	Long short-term memory
LVEF	Left Ventricular Ejection Fraction
MD	Mean Difference
MI	Myocardial Infarction
MLM	Masked Language Modelling
MLP	Multi-Layer Perceptron
Mm Hg	Millimetre of Mercury
NHS	National Health Service
NP	Natriuretic Peptides
ONS	Office of National Statistics
PPV	Positive Predictive Value
R^2	Coefficient of Determination
RCT	Randomised Clinical Trial
RECORD	REporting of studies Conducted using Observational Routinely-collected Data
ReLU	Rectified Linear Unit
RETAIN	REverse Time AttentIoN model
RNN	Recurrent Neural Network
RR	Risk Ratio
SAE	Sum Absolute Error
SAS	Sufficient Adjustment Set
SBP	Systolic Blood Pressure
SE	Standard Error

STROBE	Strengthening the Reporting of Observational Studies in Epidemiology
Tanh	Hyperbolic Tangent
TARNET	Treatment Agnostic Response NETwork
TC	Total Cholesterol
TMLE	Targeted Maximum Likelihood Estimation
TG	Triglycerides
UK	United Kingdom
UTS	Up To Standard
VAE	Variational Autoencoder

1 INTRODUCTION

1.1 Motivation

Understanding cause and effect is central to cardiovascular disease research. While randomised clinical trials (RCT) are considered the "gold standard" of causality, trials are, at times, unethical or infeasible to conduct. While observational studies can be an appropriate alternative to answers questions about associations between hypothetical exposures and outcome, these non-randomised studies have limitations. Since exposures are not randomised, one must take appropriate measures to account for confounding in the observational data[1]. Several methods in statistics that adjust for confounding and estimate effect size have found acceptance in epidemiological observational research. However, these methods suffer from known limitations: (1) the confounders need to be conventionally selected for modelling, (2) the models are linear models and interactions must be manually engineered, (3) the models have difficulty addressing other biases that distort findings (e.g., selection, finite sample estimation, and other biases).

Moreover, cardiovascular diseases are the leading causes of mortality across the world with cardiometabolic multimorbidity heavily prevalent in today's populations. With predictions that more than 50% of individuals between 65 and 74 will be multimorbid by 2035, research of cardiovascular disease risk in sicker populations is becoming increasingly important[2]. However, there is currently a limited understanding of

multimorbidity, directly implying a limited understanding of confounding factors in observational cardiovascular research focused on individuals with poorer health[3].

Recently, access to comprehensive electronic health records (EHR) and developments in statistical methodologies such as deep learning give rise to opportunities in causal inference[4]. However, deep learning has been insufficiently explored on comprehensive EHR and models remain uninterpretable.

1.2 Aims of the thesis

This doctoral thesis aims to ultimately develop accurate deep learning models for causal inference for United Kingdom (UK) EHR data with a focus on cardiovascular disease research.

To this end, the research addresses the following three objectives:

(1) To develop deep learning models that can efficiently incorporate rich multimodal EHR and learn representations that are useful with respect to benchmarks.
(2) To develop auxiliary methods to understand the "black-box" decision making processes of proposed models.
(3) To tailor the model to conduct causal inference in the observational setting.

1.3 Structure of the thesis

In order to meet the overarching goal of developing deep learning methods for causal inference on administrative EHR, significant foundational work was conducted to appropriately facilitate the proposed research. In chapter 2, I present a literature review of concepts and past investigations relevant to my doctoral research. In chapter 3, I describe the routine clinical EHR dataset that I have utilised for the doctoral research, the

Clinical Practice Research Datalink (CPRD) dataset. This dataset contains anonymised, linked UK EHR data from primary care, secondary care, and the death registry that is nationally representative of the UK[5].

Building on these foundations, I begin the presentation of the work to meet the outlined objectives. In chapter 4, I describe the development of the Transformer-based deep learning architecture that can handle minimally processed longitudinal EHR called Bidirectional EHR Transformer (BEHRT). In order to test the utility of both, features automatically captured by the model and generally, the deep learning architecture prior to conducting causal investigations, the proposed model was applied to a variety of clinical prediction tasks (e.g., subsequent disease occurrence prediction) and the incident heart failure prediction task. Since deep learning models are known to be "black-box" models, and hence, "unexplainable", it was imperative to unmask the BEHRT model to better understand its processes given the safety-critical setting of healthcare research. Hence, after establishing that the model was able to conduct accurate prediction on some conventional clinical prediction tasks, in chapter 5, I describe methods developed to better understand the prediction processes of BEHRT. For this objective, BEHRT was firstly used to predict incident heart failure (HF) and secondly, auxiliary methods were developed in order to better understand the risk prediction processes of the BEHRT model. The research yielded that BEHRT can not only capture medically validated risk and protective factors of incident heart failure, thereby validating the feature extraction processes of the "black-box" deep learning model, but also, BEHRT generated novel hypotheses.

Upon establishing that BEHRT can (1) capture useful representations from routine clinical EHR demonstrated by superior predictive performance and (2) be trusted for EHR research, I finally proceed to present the development of deep learning models

for observational causal inference in chapter 6. Combining advances in semi-parametric statistics and deep learning for causal modelling, I developed a derivation of BEHRT for causal inference, Targeted-BEHRT. The proposed model was tested in a host of semi-synthetic data experiments and demonstrated more accurate estimation of causal effect than both statistical and deep learning benchmark models.

Following evaluation of the model in semi-synthetic data settings, I present in the second half of chapter 6 and the entirety of chapter 7, four applications of Targeted-BEHRT to better understand blood pressure, cardiovascular diseases (CVD), and relevant outcomes such as mortality:

(1) Association of various drug classes of antihypertensives and cancer outcomes.
(2) Association of systolic blood pressure and cardiovascular endpoints in patients with diabetes.
(3) Association of systolic blood pressure and cardiovascular endpoints in patients with chronic obstructive pulmonary disease.
(4) Association of sodium-based paracetamol and incident CVD, all-cause mortality, and systolic blood pressure as a continuous outcome with respect to non-sodium formulations.

In chapter 8, I present a summary of the findings, the strengths and limitations of the doctoral research as a whole, implications for research and clinical practice, future directions, and concluding remarks. Finally, chapter 9 presents the appendices containing supplementary material relevant to the four research chapters.

2 BACKGROUND

In this section, key concepts in the interdisciplinary doctoral research will be briefly introduced. First, cardiovascular diseases and some select risk and associated factors will be discussed. Second, EHR data will be briefly introduced. Third, risk prediction, important theoretical considerations, and established conventional methods for assessing risk of diseases will be discussed. Fourth, both the theoretical foundations and established approaches for causal inference methods will be discussed. Lastly, deep learning theory and modern neural networks will be discussed. The following material serve to gently motivate and introduce ideas central to the doctoral research. While breadth is the goal of these following sections, certain context-dependent concepts will be elaborated in greater depth in following chapters.

2.1 Cardiovascular diseases and related conditions

CVD is one of the leading causes of mortality and multimorbidity worldwide[6]. Despite progress in prognostication, prevention, and treatment of CVD, the disease still poses a serious global burden on mortality[6]. In the UK exclusively, cardiovascular mortality causes 30% of all mortality costing the National Health Services (NHS) over 15 billion dollars annually[7]. In addition, those with CVD suffer from other associated conditions with over two-thirds of all CVD patients with multimorbidity.

Much of the increased risk of CVD-related mortality manifests from four core risk factors. Smoking, hypertension (defined as systolic blood pressure ≥ 140 mm Hg and

diastolic blood pressure of ≥ 90 mm Hg), high cholesterol, and obesity are the four modifiable risk factors responsible for approximately 60% of all cardiovascular-related mortality[8]. Of the four, hypertension presents itself as the strongest modifiable factor of risk – prevalent in a fifth of all individuals globally and in over 30% of individuals in England[8,9]. The presence of hypertension poses even greater risk of CVD when considering that more than two-thirds of hypertensive individuals also have multimorbidity[9]. Obesity measured often via body-mass index (BMI) is another key risk factor of the four core factors exacerbating CVD risk. While the relationship between obesity and CVD endpoints (e.g., myocardial infarction) has been found to be log-linear in healthy individuals, there is still much contention surrounding the shape of the association of BMI and all-cause mortality[10,11]. The so-called "obesity paradox" has been found; while overweight individuals have higher risk of CVD, the same individuals are at lower risk of mortality than those at normal or even low weight range [10]. Researchers speculate perhaps this paradoxical relationship is a result of poor confounding adjustment and/or the limitations of BMI itself, which can't inherently differentiate between body fat and lean mass[10].

Other risk (and associated) factors of CVD include atrial fibrillation, chronic kidney disease (CKD), diabetes, and chronic obstructive pulmonary disease (COPD). In fact, in the UK, guidelines for CVD care recommend preventative statin and blood pressure lowering therapy for an individual comorbid with any of these conditions[12]. This automatic recommendation for treatment is in part due to a lack understanding of risk and protection in those with pre-existing comorbidities (in addition to lower costs of manufacturing preventative pharmacological therapies); trial evidence is limited and observational studies are conflicting ultimately leading to cautious recommendations concerning treatment. For example, the association between systolic blood pressure

(SBP) and cardiovascular outcomes in those with risk factors has not been explored comprehensively in both observational and randomised studies. While some studies find a discontinuous relationship between SBP and cardiovascular endpoints, several others find this to be a result of uncontrolled confounding and alternatively hypothesize a continuous one[13–16]. Evidence in either randomised or observational form is lacking and needed to more accurately understand risk and protection in these subgroups with pre-existing conditions.

The following subsections will explore major conditions constituting CVD in addition to a few diseases associated with CVD: heart failure, ischaemic heart disease, stroke, hypertension, diabetes mellitus, and COPD.

2.1.1 Heart failure

HF, simply defined, is a condition that indicates that the heart is unable to pump blood around the body to sufficiently meet the needs of the various organs of the body[17–19]. HF can manifest due to multiple causes such as hypertension, valvular heart disease, or prolonged, untreated damage to heart tissue[20,21]. HF is generally subcategorised based on measurements of the left ventricular ejection fraction (LVEF), which is a percentage quantification of how much blood is pumped out of the left ventricular in relation to the amount of blood entering said ventricle[19,20].

Due to the complex nature of the condition, HF is diagnosed in a multi-factorial way[22]. Initially, an initial clinical examination accounting for various risk factors, immediate symptoms, and history of other cardiovascular events is pursued[22]. Further examinations vis-à-vis electrocardiogram (ECG) can be recommended to comprehensively complete initial assessment[22]. Most commonly however, cardiac imaging via echocardiography is utilised for assessing incident HF since it captures data

on ejection volume in real time analyses of the chambers of the heart[22]. Alternatively, an investigation of the plasma concentration of natriuretic peptides (NPs) can be conducted[23]. Elevated levels of NPs can help inform an initial conception and screen for HF and can filter patients for further examination[23].

Furthermore, while HF may be the first of many cardiometabolic disorders, often arises in the presence of other comorbidities. In general, incident HF is a common condition following a long history of cardiac defect especially in elderly patients[20]. Thus, HF generally does not present itself in isolation but in the midst of other conditions such as COPD, diabetes, hypertension, and other risk factors[19,20]. In addition, treatment of various conditions such as corticosteroids have been found to cause iatrogenic risk of HF[24].

In general, HF incidence rates have declined in the 60 to 79 years age group implying that significant progress has been made in preventative therapy[19,22]. In comparison to incidence of cancer, HF incidence is now commensurate with the four most common cancers: lung, breast, bowel and the prostrate[19]. On the other hand, the absolute increase of heart failure around the world exerts great stress on health care systems worldwide[19].

2.1.2 Ischaemic heart disease

Ischaemic heart disease (IHD) or coronary artery disease is another such cardiovascular condition associated with substantial global mortality[25,26]. IHD is caused by blockages to the heart's blood supply or by build-up of fatty substances in the coronary arteries[25,26]. Specifically, the blockages can be caused clots, but more often, the blockage is due to an accumulation of plaque, termed artherosclerosis[25]. When the flow of blood is blocked to the heart muscle, the heart cells may die; this phenomenon is more

commonly known as a heart attack or a myocardial infarction (MI)[25]. IHD can be identified with a variety of tests including treadmill tests, radionuclide scans, a magnetic resonance imaging scan, or coronary angiography[25].

Additionally, while IHD may manifest as a sole condition and the first of many, the condition often occurs in the presence of multimorbidity similarly to HF[9]. Many IHD patients have concomitant diabetes, COPD, and other conditions; of the many risk factors, hypertension specifically is the most common risk factor in those with IHD [9,25].

Even though age standardised mortality rate of IHD has decreased, IHD still remains a substantial contributor of mortality and burden of disease globally[9,27]. Absolute incidence has remained constant (between 3 and 4 percent) ultimately leading to increased health care service utilisation across UK[27].

2.1.3 Stroke

Stroke as a cardiovascular condition has repercussions on both vascular and cerebral structures[28]. Stroke occurs when there is a disruption of blood to the brain, which can often result in brain damage and mortality[29]. Alternatively, when blood flow is only temporarily disrupted, a transient ischaemic attack or "mini stroke" may occur[29]. The condition of stroke accounts for over 6 million deaths worldwide and especially effects those who are 40 years of age and older[30].

Much like heart failure and IHD, stroke, while occurring as a preliminary or sole condition, often presents itself in the midst of multimorbidity; age standardised incidence and prevalence rate has attenuated over the recent years, but the overall absolute burden of this cerebrovascular condition has exerted great pressures on the health care systems[9,30]. In addition, elevated blood pressure is a leading risk factor for stroke[31].

2.1.4 Hypertension

As discussed, one of the greatest individual contributors of CVD risk is hypertension[8,32]. Hypertension is a condition that is marked with consistently raised pressure in the blood vessels. Blood or arterial pressure is measured by two numbers; the first number, the systolic blood pressure (SBP), measures the pressure when the heart beats and the second measures the diastolic blood pressure (DBP), the pressure in the vessels when the heart is resting or between beats. Normal blood pressure lies below 120 and 80 mm Hg of systolic and diastolic blood pressure respectively[33].

Elevated blood pressure (i.e., systolic and diastolic blood pressure measured higher than the aforementioned measurements) is responsible for a third of all annual deaths globally[32]. Hypertension has even been labelled as the "silent killer" due to its prolonged sub-clinical period. Presentation of condition via symptoms is limited in these sub-clinical stages of the disease. In large part due to this reason, the disease of hypertension often goes undiagnosed and remains a global burden[32]. However, if diagnosed early, controlling elevated blood pressure can significantly protect against downstream CVD[29,32].

The optimal threshold of blood pressure has caused significant controversy, especially in those with pre-existing conditions[16,34,35]. In those free of cardiovascular risk factors, the recommendation for SBP has generally been, "the lower, the better" – i.e., the lower the blood pressure, the lower the risk of a variety of cardiovascular disease and mortality endpoints[36,37]. However, in those with pre-existing conditions, the research into optimal threshold for SBP has generated much controversy in recent years. Some research has defended the paradoxical, "J-curve" association between SBP and various cardiovascular outcomes – an optimal SBP such that below and above this optimum, risk increases[13,16,35]. In large part due to these studies advocating an optimal SBP threshold,

hypertension guidelines have indeed recommended a recommendation of blood pressure lowering treatment goal of SBP <130 mm Hg (140 in elderly patients)[38]. However, as discussed, the degree to which residual confounding and reverse causation have roles to play in these conclusions is yet to be addressed with empirical evidence[15].

2.1.5 Diabetes

Diabetes is a metabolic disease afflicting one in ten people in the UK and worldwide, the cause of 1.5 million deaths directly as of 2019[39–41]. The condition manifests when blood glucose levels are elevated[41]. Furthermore, the condition, a known risk factor of CVD leads to multiple issues including HF, stroke, IHD and peripheral arterial diseases[29]. Additionally, as a direct consequence, it is also known to cause other downstream conditions such as diabetic neuropathy and diseases of the eye[42]. The disease is known to contribute to all-cause and specifically, cardiovascular-related mortality as well[9,42].

In previous research, the relationship between blood pressure and diabetes was captured to be log-linear much like established knowledge of the association between blood pressure and cardiovascular outcomes in those without prior risk[43]. Furthermore, meta-analysis of randomised trials has also demonstrated that blood pressure lowering indeed reduces risk of incident diabetes regardless of baseline blood pressure[44]. However, the relationship between blood pressure and CVD in patients with diabetes is lesser understood[16].

2.1.6 Chronic obstructive pulmonary disease

COPD is a respiratory disease consisting of emphysema and chronic bronchitis[45]. This disease has affected 1.2 million as of 2020[45] . Those who are elderly or smoke are at high risk of contracting this respiratory condition[45]. Furthermore, the condition shares many

risk factors with CVD and has been associated with increased hospitalisation, all-cause mortality, and specifically CVD (and cardiovascular death)[46]. In fact, IHD, HF, and cardiac arrhythmias are the most frequently observed outcomes in those with COPD[47]. Estimates of prevalence of IHD in those with COPD range between 20 to 60% across various, diverse study populations[46].

Despite the established increased risk of CVD in those with COPD, the association between SBP and CVD in those with COPD remains controversial. As discussed previously, past research advocates the "j-curve" association between the two and guidelines recommend the optimal blood pressure lowering treatment goal to be <130 mm Hg in those with COPD[35,48].

2.2 Electronic health records

Electronic health records (EHR) are a collection of patient health information recorded over health encounters stored electronically in a digital format. The data within EHR can range from unstructured (e.g., handwritten notes) to structured (e.g., diagnoses, measurements of biomarkers, prescriptions). EHR systems were originally devised for billing and other administrative purposes (e.g., quality of care assurances)[4]. For administration, the patient health data remains digital and private, and with appropriate permissions, can be easily shared between clinical care providers[4]. Additionally, recordings do not need to be replicated; instead, a single file can be copied and shared seamlessly allowing for easy access and low-cost transfer of healthcare date[4]. In large part due to these administrative benefits, there is now widespread adoption of EHR systems around the world[4]. For example, in the United States, as of 2019, about 75% of office-based physicians and 96% of acute care hospitals have adopted administrative EHR systems[49].

EHR databases furthermore present an incredibly rich source of data for downstream clinical and epidemiological research[4]. Partly due to the fact that the databases offer access to multitype, high-dimensional health variables at low-cost, the EHR sources of observational data offer ample opportunity for observational analyses[4,50]. Many impactful projects answering questions about immediate epidemics, rare diseases, generalisations of findings, and many other critical research topics in epidemiology have utilised data collected from EHR systems [4,50]. In fact, much of our understanding of the disease trajectories of the COVID-19 pandemic caused by the coronavirus comes from high impact observational studies utilising administrated EHR databases[51].

For research purposes, large-scale EHR databases provide some vital components: large-sample size, rich patient-level health data, and long follow-up[4]. Typically, observational studies are conducted by pooling many individual patients' EHR in the form of diagnostic, prescription, measurements, and other health annotations across time and by analysing them[50]. For research into, for example, subgroups of patients (those who satisfy a particular condition), for which data are usually scarce, large-scale EHR databases provide the opportunity for well-powered epidemiological research. Furthermore, the opportunity to follow patients for longer directly benefits our understanding of diseases that have lengthy sub-clinical periods (e.g., cancer)[4,50]. We can additionally better understand chronic conditions (e.g., cardiovascular conditions such as ischaemia) and chronic disease progression with longer follow-up periods. Overall, access to large samples of individuals, rich multidimensional health data, and longer follow-up offers fertile grounds for impactful research of complex conditions such as CVD [4,50] (Table 2-1).

Table 2-1: Comparison between conventional and electronic health records studies

Feature	*Conventional study*	*Electronic health records study*
Purpose of data collection	Research	Administrative use or clinical care; research is not the primary objective
Costs	Generally expensive; mostly funded by large institutions (e.g., governmentally funded)	Generally, less expensive; usually funded by the healthcare provisioner (e.g., National Healthcare Services)
Population definition	Recruitment based	Based on how much a patient utilises the health care system. Thus, large number of eligible patients for inclusion in cohort study
Follow-up	Dependent on how long funding is available. Thus, generally visits are defined in terms of fixed intervals	Dependent on encounters with provider. Also, usually timing between visit is not unform.
Data capture	Dependent on protocol defined prior to data collection. All data that is pre-specified will be collected - genetic sample, covariates, health variables, etc	Data are recorded by clinical staff (E.g., doctor, nurse) and organised into diagnoses, medications, laboratory tests, clinical notes (unstructured)
Variable capture	All covariates and outcome variables prespecified to be collected at beginning of study will be collected.	The variables and outcome variables that are a result of clinical encounter will be recorded. Health variables that have not manifested as diagnosis, medication, laboratory test, etc will not be recorded since those variables have not been assessed as a product of clinical care. Absent clinical states often not noted.
Socioeconomic data	Geographical information systems data are captured if prespecified. Also, directly measured if sample size is small.	Patient-level socioeconomic factors are sometimes linked to geographical datasets (i.e., with patient home address).

Adapted and restructured from tabular data from Casey et al[4].

Furthermore, the linkage of data across many healthcare datasets allows for more robust accounting of patient health[4]. Those between various healthcare settings – primary and secondary for example – afford researchers greater breadth of healthcare variables. Specifically, some aspects of health not covered in one dataset may be provided in a linked dataset ultimately facilitating more comprehensive accounting of patient health[50,52]. For example, while primary care may capture more chronic or long-term conditions and related biomarkers, acute conditions may be captured in the secondary care setting [4,50,52]. Furthermore, linkage to geographical and socioeconomical health

indicator datasets help capture orthogonal health variables, which may serve as proxy for other factors of health not fully accounted by clinical care datasets (e.g., primary care data) within the EHR database[4,50]. Lastly, access to mortality-related data and cause-specific mortality data specifically offer researchers the opportunity to conduct high-quality analyses concerning mortality and causes of death [50,52]. In these ways, EHR databases with rich linkages can capture the patient health timeline from "cradle to grave" [50,52].

Additionally, utilising the longitudinal aspect of records in EHR databases offers the promise of more nuanced data-analyses [4,50]. For instance, instead of just noting a condition at a fixed point of time, the date of diagnoses annotated by administrative EHR databases can give us a more sensitive understanding of disease occurrence and perhaps even, severity[4,50]. Naturally, for the researcher, the opportunities include richer downstream analyses (e.g., stratified analyses by age of diagnosis of incident condition) and more flexible cohort designs (e.g., setting random index date) with provisioning of temporal information[4]. The recording of time generally allows for more nuanced design and analysis of the medical history (adjustment) period, the baseline date, and the follow-up time window. Incorporating this information into observational studies can be crucial in the case of chronic conditions.

While there are issues of missingness, incorrectness, and heterogeneity in recording practices over time, there are many practical ways to attend to some of these matters[4,50]. For example, advanced statistical techniques such as imputation of EHR data can be used to directly address issues of missingness[53]. For the cases of continuous measurements (e.g., cholesterol and blood pressure) and categorical variables (e.g., deprivation or smoking status) as examples, imputation strategies have been demonstrated to be useful strategies for overcoming missingness in the observational

setting[43]. While incorrectness on the other hand is an issue, many administrative EHR databases provide algorithms to check the validity of the particular record, sometimes with validity measured across a spectrum[54]. With this measure, individual researchers can choose an appropriate threshold to filter "valid" records for downstream analyses. In this way, the validity of records can be more objectively measured and filtered thereby directly addressing concerns of the validity of individual records. Of course, EHR databases can still contain noise and contain erroneity; however, EHR data has been found to be more reliable for measuring health state than data points collected as a product of self-reporting[4]. In order to account for changes in practice, detailed attention must be given to research spanning several decades and the results stemming from said research. It is vital that sensitivity analyses are conducted to investigate the effects of temporal systemic changes in clinical care[4].

2.3 Risk prediction

Assessment of the risk of a particular condition (e.g., CVD) is incredibly important for questions in medicine and epidemiology. Predicting risk of a condition like CVD is vital for raising awareness about the burden of CVD to multiple stakeholders: patients themselves, clinical practitioners, and public health policy makers[12]. In the UK, risk assessment tools have been used to help identify those at high risk of CVD[12]. While some individuals are at lower predicted risk for CVD and can be encouraged to make changes to lifestyle choices (e.g., smoking and alcohol), others at higher risk require preventative pharmacological therapy in order to avoid developing CVD[12,55]. In this way, utilising risk assessment methods to identify those who need preventative therapy can ultimately be more efficient and lower costs for healthcare providers in the long run[12].

The following sections explores the motivations for risk prediction modelling, the development of risk prediction tools, and the limitations of risk prediction modelling with conventional statistical methods.

2.3.1 Motivating context

Both diagnosis and prognosis have an overlapping challenge: diagnosis is the capture of disease in the present, and prognosis is the capture of the disease state in the future. While diagnosis involves a diagnostic test or a biomarker capture, statistical models utilise several predictors captured in the present to capture disease state in the future[56]. For many epidemiological and clinical questions, the outcome of interest is usually the presence of a disease, meaning a binary variable. The prediction of the presence of said disease comes in the form of a predicted probability.

2.3.2 Development of models

In order to develop a discriminative statistical model that can conduct risk prediction, outcome and predictors must be defined. For binary outcomes, the outcome can be defined as an event that occurs within a finite period of time from the start of the study (study entry or index date). The predictors chosen can be diseases, medications, measurements, lab tests or any other health data points recorded at or before the start of the study. These predictors are conventionally chosen based on expert knowledge[55,57] and extracted from raw healthcare datasets (e.g., administrative EHR datasets), and appropriately transformed and normalised to be considered as input for modelling. Similarly, presence of disease in the follow-up period following start of the study must be extracted as a binary variable for each patient.

Following model development, the predictive performance of the model must be measured in order to inform the worth of the predictions. The conventional approach to

quantify performance is by simply measuring accuracy of the predictions: measure how close are the predicted values to the actual outcome[56,57]. Also, R^2 statistics and Brier score are additionally used to quantify performance[56,57].

Another method of quantifying predictive performance is by answering the question: do those who develop the disease outcome have higher predicted risk than those who do not? This is the central question that metrics of discrimination directly address. The receiver operator characteristic curve is one such metric of discrimination quantification; the curve plots sensitivity (i.e., true positive rate or recall) against 1.0-specificity (i.e., false positive rate)[56,57]. Quantifications of the area under the receiver operator characteristic curve (AUROC) can provide a summary statistic of the trade-off between true and false positive rate. While an AUROC of 0.5 represents a random model – a model that is predicting incidence of disease with random probability (i.e., coin flip prediction), one of 0.0 represents the poorest model (and one that performs worse than a random model), and one of 1.0 demonstrates that the model is predicting the outcome perfectly[57]. Other metrics utilising precision (i.e., number of correctly predicted positives out of all positive predictions) along with recall such as the area under the precision-recall curve (AUPRC) and F-measure combine and balance multiple measures into one omnibus statistic[57].

Before clinical acceptance and utilisation of these risk assessment models in the healthcare setting, the models need to be internally and often, externally validated using some of the aforementioned techniques[57]. Internal validation consists of assessing model for predictive performance on the same data used for development. Some common techniques to conduct internal validation is bootstrapping and k-fold cross validation[57,58]. While there are variations to the process, generally, bootstrapping involves repetitive (1) sampling of the dataset with replacement, (2) training on one part of the sample, and (3)

testing on the remaining part of the sample. Predictive performance is computed as a metric (e.g., AUROC) averaged across the repetitions with the additional option to compute standard deviation of the same metric. K-fold cross validation is an alternate method for internal validation, for which the dataset is split into a fixed number (i.e., k) of mutually exclusive partitions or "folds", and all the folds but one is used to train, and the remaining is used to test the model. This is repeated iteratively until every single fold has been used exactly once as a test set. Model performance on the dataset can be computed as an average over the test folds; similarly standard deviation can be calculated as well[57,58]. Both methods are commonly accepted methods to measure predictive performance of developed models. External validation is used to assess the generalisability of the model. In gist, the model is trained on an internal dataset but tested on a cut of data from another dataset[57]. Sometimes, data from a different geographical area or time-period are used to ensure that the model is generalisable[57]. If the model demonstrates acceptable discrimination, then the model has generalised acceptably to other data. However, if the model poorly discriminates, this means that likely, a new model is needed to perform adequately on the external data[57].

2.3.3 Limitations of conventional risk prediction modelling

Risk prediction models currently face many limitations especially for prediction of cardiovascular events. Risk prediction models currently require conventional predictor selection. In situations, in which risk factors are comprehensively understood, predictors can be easily extracted from data and statistical models can be efficiently developed. However, in situations, in which risk factors are lesser known, development of robust risk assessment solutions becomes infeasible[59]. For example, in the case of incident heart failure predictions, many proposed solutions for risk assessment demonstrate poor

predictive performance[60]. This is in part due to the limited understanding of factors of risk of heart failure[60].

Furthermore, even if the condition to be predicted is widely understood, conventional models such as logistic regression or log-binomial modelling for binary outcomes, are inherently linear models and have to undergo manual feature engineering. Complex non-linearities and interaction variables must be manually engineered. If the precise functional form of the interaction is unknown, the interaction variable may be poorly defined and thus provide limited utility for predictive performance. Thus, modelling of complex phenomenon is difficult in the conventional modelling paradigm.

2.4 Causal inference and association analyses

Central to research in medicine and epidemiology, is the determination of the cause of a particular disease or the effect of a particular intervention[61]. When an association such as the association of antihypertensives and CVD, is able to be investigated in a randomised setting, such as the setting of the RCT, the drug classes are randomly assigned to the patients (e.g., some receive beta-blockers, one class of antihypertensives while others, calcium channel blockers)[62]. In the randomised setting, all other variables contributing to risk or protection of the outcome are randomly distributed in both groups and hence controlled. Thus, the effect of the variable of choice on the outcome, the intervention can be directly assessed. For this reason, the RCT is considered the "gold standard" of assessing causality[62].

However, there are several hypotheses that cannot be tested in the standard RCT setting. Some hypotheses cannot be tested in the randomised setting due to ethical concerns[61,62]. For example, the effect of smoking on cancer cannot be ethically tested in an RCT; there are ethical issues with randomising smoking, the intervention in question

in this experiment. Furthermore, sometimes the hypothesis in question does not attract enough participants in order to have sufficient sample size. If sample size is lacking, for a complex question regarding cause and effect, even with randomisation, pre-intervention variables may not be balanced amongst the intervention and control groups[62]. Additionally, implementing, recruiting, and analysing an RCT takes time and resources. As of 2016, estimates of costs to run a trial for clinical interventions can be as expensive as $1.4 billion[63]. With these high costs, not all hypotheses can be allocated resources for testing[63]. Lastly, while RCTs are useful for estimating the association, the results are not generalisable. While some RCTs may claim that a hypothetical association is negative, others may defend that the association is positive. Often, a meta-analysis of multiple RCTs must be conducted to identify the generalisability of findings[64].

Causal inference methods indeed seek to emulate the trials whilst directly addressing the aforementioned limitations of the conventional randomised investigation. Observational studies utilising non-randomised data, if conducted appropriately can seek to find answers to questions relating to causation.

In the following sections, a hypothetical investigation of the aforementioned effect of antihypertensives on CVD will motivate and facilitate the discussion of causality. First, naïve methods of estimating causal effect will be introduced. Second, formalised frameworks of capturing causal effect will be presented. Third, a survey of methods for estimating causal effect will be presented.

2.4.1 Motivating causal inference

Motivating the discussion about causality with the aforementioned research of the effect of antihypertensives on CVD, the question is asked: how can this association be modelled using methods for the observational setting? As an initial causal analysis, a

researcher might conduct a cohort study of the association of beta-blockers on CVD with respect to calcium channel blockers.

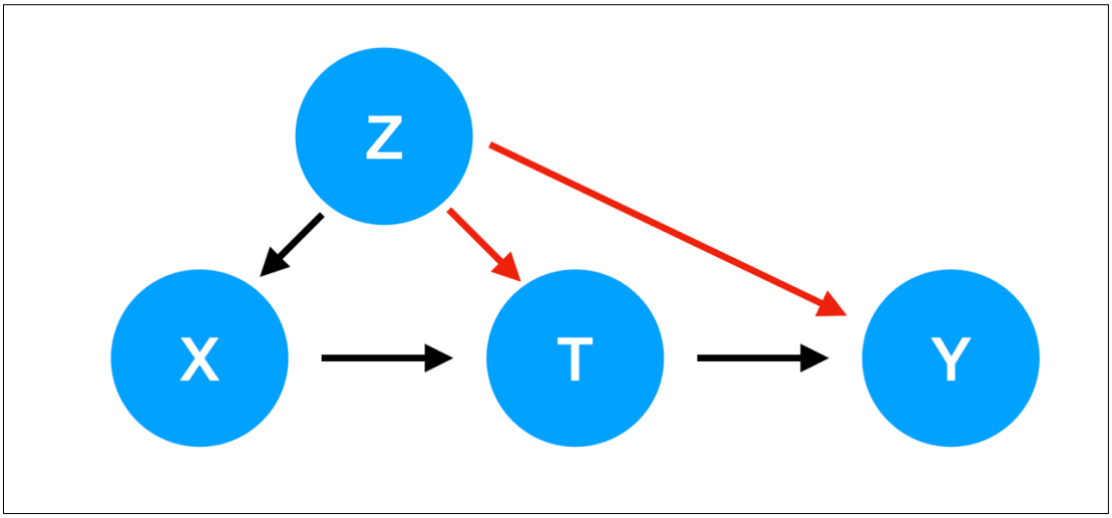
The study design will be the following: for a patient, the first date of the initiation of either of the two drugs would be considered the date the patient entered the study (i.e., study entry or "baseline" date). The outcome of CVD would be assessed within a follow-up time window; perhaps this window could be set as the 10-year window of time following study entry date.

Following this design, a cohort would be extracted using an appropriate data source, and for each patient, the information concerning exposure group assignment and the presence of the outcome will be extracted. With this data, the researcher might estimate risk ratio (RR), a widely accepted method of estimating strength of the association:

$$\widehat{\boldsymbol{\psi}} = \frac{\mathbb{E}[\boldsymbol{Y}|\boldsymbol{T}=\mathbf{1}]}{\mathbb{E}[\boldsymbol{Y}|\boldsymbol{T}=\mathbf{0}]} \tag{2-1}$$

With Y, as the outcome of CVD, and $T = 1$ as the indication of beta-blockers and $T = 0$ as the indication of calcium channel blockers, the $\hat{\psi}$ is the crude RR estimate. $\hat{\psi} > 1$ implies that beta-blockers cause CVD more so than the calcium channel blockers, $\hat{\psi} < 1$ implies that the beta-blockers cause CVD less so than the calcium channel blockers, and the $\hat{\psi} = 1$ implies that the effect of beta-blockers on CVD is commensurate with that of calcium channel blockers.

Figure 2-1: Illustration of confounding



This figure illustrates the confounding factors present in a hypothetical observational setting. X denotes some pre-treatment covariates solely associated to the exposure (T), and Z is a confounding variable associated to all variables: X, exposure and outcome (Y). Red arrows denote associations necessary for Z to be confounding variable.

However, given that the true association is $\hat{\psi} = 1$, if there are factors associated to both, use of beta-blockers and the outcome, CVD, and the association captured is $\hat{\psi} \neq 1$, it is said that these common causes of both intervention and outcome are "confounders" and responsible for distorting the effect measure (Figure 2-1 and discussion of confounding below in section 2.4.2). As a concrete example of confounding, if certain patients have COPD, they are less likely to be prescribed beta-blockers[65]. Furthermore, COPD is a condition that exacerbates risk of CVD[46]. With association to both exposure and outcome, COPD is a confounding variable[61]. With the naïve method of estimating RR shown in Equation (2-1), $\hat{\psi}$, the confounder of COPD is not accounted, and thus conclusions drawn from the effect measure, $\hat{\psi}$, may be biased and incorrectly capture of the true relationship between the classes of antihypertensives and CVD.

Conventionally in epidemiology, in order to more accurately capture causal effect in the presence of confounding, a statistical model would adjust for confounders.

However, this may not be possible if the confounders are not measured. If the investigation is poorly defined or if the observational setting is heavily confounded, identifying causal effects may not be possible[61]. In these cases, a formal framework is beneficial in these circumstances in order to delineate if (1) association analysis can be pursued and (2) if causal effect is identifiable.

2.4.2 Formal framework of causal inference

With many open questions about causal inference, the pursuit of formalising the framework of causal inference was undertaken by Donald Rubin in the 1970s when he extended the initial conception of the "potential outcomes framework" from Jerzy Neyman, who initially introduced the concept in his master's dissertation in 1923[66–68].

The potential outcomes framework explains cause and effect in the following way. With reference to the motivating example, if a patient given beta-blockers was offered calcium channel blockers, then the probability of the outcome of CVD would be different than if the patient had been given what the patient was assigned[67].

More formally, we let T_u and Y_u be the functional form of exposure and outcome respectively for patient u. Now to understand the effect of exposure on outcome, T is considered to be a function that can represent effect of intervening on the treatment in question. Y can now be functionally represented as $Y_u(t)$ for a patient u and exposure t. Of course, T_u is hypothetical as opposed to set, and the variable $Y(t)$, is a counterfactual or potential outcome variable.

In the antihypertensive study, the exposure is binary implying the variable T_u, may take on two states: 0 and 1. In notational form, the two states are $Y_u(0)$ and $Y_u(1)$. Thus, the RR can be defined as:

$$\boldsymbol{RR} = \frac{\mathbb{E}[\boldsymbol{Y_u(1)}]}{\mathbb{E}[\boldsymbol{Y_u(0)}]} \quad (2\text{-}2)$$

Equation (2-2) explains the causal effect in the potential outcomes or counterfactual framework. Philosophically, all units are the same in every way except the intervention, which is the variable of interest. So, under this framework, the outcome in those with the intervention status can be compared to those with the control status[67].

2.4.2.1 Assumptions for causal identifiability

Under some assumptions, the average causal effect can indeed be identifiable. Specifically, there are a number of variations of the assumptions themselves, but generally, the core sufficient assumptions for identifiability are ignorability, positivity, and consistency[1].

First, under the assumption of ignorability, sometimes clarified as strong ignorability, there are "no hidden confounders" and only if this condition is met, the average causal effect can be identifiable[1]. In notational, the assumption of strong ignorability is formalised as $\{Y(1), Y(0)\} \perp T, X$ [21]. More informally, this conditional independence relationship means that the potential outcomes are independent of the treatment given all confounders, X. In randomised and synthetic data situations, the assumption is more easily met. Randomised scenarios directly implies that the confounders in both exposure groups are likely balanced by randomisation of the exposure (given sufficient sample size), and in the synthetic data situations, in which confounding factors are generated and thus all known and identifiable, complete adjustment is possible. On the other hand, in observational settings utilising non-randomised, routine clinical data, the scope of known and unknown confounding is not measurable[1]. Thus, evaluating if all confounders are accounted is infeasible[1,61]. With access to richer observational data such as multimodal, longitudinal EHR with a host of

health indicators variables, more comprehensive confounding adjustment is possible; of course, uncontrolled confounding cannot be discounted. Thus, while strong ignorability may be theoretically an absolute assumption, in reality, the assumption functions on a spectrum from total disregard of ignorability (e.g., naïve or crude risk ratio estimation) to strong ignorability; while most observational studies cannot truly guarantee strong ignorability, weaker forms of ignorability can more feasibly be met[69].

Second, the assumption of positivity means there is a non-zero probability of being assigned either treatment for any given patient. To explicate upon positivity in terms of the motivating example, each patient needs to have a non-zero probability of being assigned either beta-blockers or calcium channel blockers. Since both drugs are antihypertensives, there is a realistic chance that patients in this hypothetical observational study are eligible of being assigned either drug class. However, taking another hypothetical observational study of the effect of a novel cancer drug on tumour size with respect to a placebo/control intervention, a patient without cancer at study entry has 0.0 probability of being assigned the cancer drug simply because the drug is not prescribed to those without cancer. Hence, in this hypothetical situation, the assumption of positivity would have failed to have been met. Failing to meet the assumption of positivity directly implies that those in the control group are being inappropriately compared to those in the intervention group[61]. Perhaps, some individuals selected for participation in the observational study do *not* have a positive probability of being assigned either exposure, and those individuals should not be included in the study.

Lastly, consistency is the assumption that for every patient, the potential outcome for the actual treatment assigned to the patient is equal to the factual outcome[61]. For example, if a hypothetical patient in the running example observational study was assigned beta-blockers, the potential outcome for the exposure, beta-blockers, would be

the same as what the patient actually experienced as the factual outcome. While this may be a natural statement to assume whilst conducting an observational study, often times, the assumption is not met if the exposure is not properly defined. For example, given a hypothetical study investigating the effect of body mass index (BMI) on CVD, a patient with a particular BMI may have the same BMI as another patient, but the outcome of CVD may differ between the two patients. This is because while one patient has a particular weight and height, the components of BMI, the other, might have a different set of the same two components. Due to poorly formulated exposure, the outcome of CVD is different for a particular exposure state. In this observational study, the exposure status would be considered "inconsistent"[61].

2.4.3 Methods in causal inference

While James Robins extended Rubin's potential outcomes framework in the 1980s, Judea Pearl coalesced and generalised these frameworks utilising directed acyclic graphs (DAGs) in order to appropriately illuminate the variables potentially distorting and/or biasing the causal effect[1]. Fundamentally, given modelling of all elements – exposure, outcome, and confounding variables – researchers can appropriately adjust for confounding variables and thus directly satisfy the assumption of ignorability in downstream analyses utilising stratification, regression modelling, or propensity-based methods.

2.4.3.1 Directed acyclic graphs

Mathematically, DAGs are inspired by causal graph theory. Variables are represented by nodes and relationships/association are represented by edges between nodes[1]. Paths are formed by a sequence of edges. Furthermore, the edges of DAGs are directed, meaning that a particular edge describes a directional relationship between two nodes[1]. Also, the graphs are specifically acyclic meaning that no variables can cause itself

either by self-loop or by path[1]. In some ways, DAGs have temporality in-built; since the DAGs use directed, acyclic edges between nodes, this naturally means that some variables manifested before others[1].

Importantly, the DAGs enable identification of conditional independence of variables defined on the causal graph[1]. If a dataset were to be generated from a particular DAG with formalised independence relationships, the generative process would preserve the same relationships between the same variables. To better formalise independence, the language of d-separation is introduced[1]. Two nodes on the DAG are d-separated from one another if all paths between the two nodes are blocked. A path, *p*, is defined to be blocked by a set of nodes, *V*, if and only if: (1) *p* contains a chain from $i \rightarrow j \rightarrow k$ or $i \leftarrow j \rightarrow k$ such that the middle node, $j \in V$ or (2) *p* contains a *"V-structure"*: $i \rightarrow j \leftarrow k$, such that the middle node(s), *j*, and any descendants of *j* are not in V[1]. The latter *V-structure* is also known as a collider structure; colliders naturally can block paths, and conditioning on these colliders can open paths[1]. The chain, $i \rightarrow j \rightarrow k$ functions with *j* as mediator, and $i \leftarrow j \rightarrow k$ functions with *j* as a confounder variable[1]. Conditioning on *j* in these two cases blocks the path since the chain is not a collider structure. Alternatively, given a set $\{j\}$, that contains nodes that blocks all paths between *i* and *k* (given there is only the one discussed path between *i* and *k*), this means that *i* is independent from *k* conditioned on $\{j\}$[1].

With this language of conditional independence, confounder variables can be described by DAGs. Using DAGs simultaneously with established frameworks such as the potential outcomes framework, causal effect can be identifiable. Specifically, using DAGs to capture variables to be conditioned upon (e.g., confounder variables) enables identification of the sufficient adjustment set (SAS) of confounders[1]. Defining the SAS using causal graphs and the mathematical framework underpinning the causal graph,

renders the exposure and outcome variables conditionally independent [1]. In addition to meeting assumption of strong ignorability with SAS of confounding variables, if observational studies are well designed and appropriately adhere to assumptions of positivity and consistency and any other study-specific assumptions, causal identification is directly enabled.

DAGs can be a very useful tool in expert-driven studies in guiding identification of confounding and mediating variables in an observational study. Furthermore, for didactic and explanatory purposes, DAGs can efficiently communicate the general relationship between variables. However, DAGs are non-parametric and thus cannot be used to infer magnitude of association between two variables or nodes on the graph[1].

2.4.3.2 Methods for confounding adjustment

Once the DAG has been identified for a particular observational study, in order to estimate causal effect, the identified confounders must be conditioned upon. Confounding adjustment can be conducted in many ways including but not limited to, mainly stratification, regression, and propensity score-based methods.

Stratification methods have a long history of use in epidemiological studies with first formalisation of the method by Cochran in the 1960s[70,71]. It has been used often when the SAS is limited in cardinality (i.e., perhaps simply age or sex as confounders). In this case, the data are split up into a fixed number of subgroups for the particular confounder; for example, with sex as a confounder, the number of subgroups would be two. The stratum specific estimand (e.g., RR) is calculated and then for an omnibus static, a weighted mean can be taken over the multiple strata; for the example of sex as a confounder, the final RR would be a weighted mean of the RR for male and RR for female patients[70,72]. The estimand is often a crude effect measure since confounding adjustment via stratification conditions upon the confounding variable(s) rendering the association

between exposure and outcome to be independent given stratification over confounders. The main drawback of stratification is that many confounders directly imply many stratifications of the data[72]. If the number of confounder combinations are far greater than the cardinality of the data, then sample size becomes insufficient in any given stratum. Rothman and Greenland eloquently refer to this as "when stratification exceeds the limits of the data" and the resulting estimates of causal effect over stratum can be highly volatile and susceptible to bias[71].

Regression indirectly addresses these issues and is another way to adjust for confounding variables[71]. Once confounders are selected perhaps through DAG based visualisation or expert-based selection, the outcome variable is regressed upon the exposure variable and the multiple confounders[71]. Depending on the model and given that the modelling assumptions are met, a regression coefficient for a variable in the model estimates the effect of the increase of one unit of that particular variable given the assumption that all other variables modelled remain the same. Thus, the exposure variable's respective coefficient can be extracted from a fit model and with minimal transformation (e.g., exponentiation), can provide an effect measure that could be interpreted as an appropriate estimand of association strength (e.g., risk ratio, hazard ratio)[71]. Alternatively, counterfactual estimation may be possible with regression models. With a fit model, the risk can be estimated as if all patients were given the exposure and compared with the risk as if all patients were given the non-exposure. In regression-based confounding adjustment, poor measurement of confounders may result in distorted coefficient for the exposure variable. For example, binarizing a more complex variable such as BMI status, originally a continuous variable, may be a form of insufficient adjustment of confounding, leaving the study susceptible to many issues of residual confounding[71]. With respect to stratification-based methods, while stratification-based

confounding adjustment fails to handle many confounding variables, the regression model can theoretically handle many confounding variables. Thus, regression-based confounding adjustment is preferred in observational studies, in which there are many confounding factors[71].

Lastly, propensity-score based methods are a more recent class of methods of confounding adjustment and have slowly begun to find acceptance in mainstream epidemiological research[73,74]. Propensity-score based modelling is based on two-stage modelling defined first by Rosenbaum and Rubin in 1983[74]. In the first stage, a regression model is implemented such that the treatment variable is regressed upon any and all pre-treatment variables. With this model, each patient in the study is given a propensity score – a probability-based score that informs the probability the patient will be given the exposure with 1.0 implying that the patient has a 100% chance of being in the exposure group and 0.0, a 100% chance of being in non-exposure group[73].

In the second stage, matching, stratification, weighting, and regression-based methods can be used, in gist, to "adjust" for the propensity score and estimate effect on the outcome variable[73]. In the case of matching, a variety of techniques can be used to match patients in one exposure group who have the same or similar (up to a caliper distance) propensity score in another. However, a common drawback of this procedure is that many individuals are not matched or inappropriately matched resulting in issues with samples and biasing effect measures downstream. Rosenbaum and Rubin extended the idea of stratification formalised by Cochran to the propensity score[75]. Specifically, the two demonstrated that implementing stratification on propensity score eliminates 90% of bias due to measured confounding variables[73]. Alternatively, weighting-based methods can be used. In the case of weights, the propensity score estimates are transformed into inverse probability of treatment weights (IPTW) defined as $\boldsymbol{w_i = \frac{T_i}{P(Z_i)} + \frac{1-T_i}{1-P(Z_i)}}$ with

T_i, the assigned treatment for patient *i*, and the $P(Z_i)$, the propensity score of patient *i*[76]. With this measure, the average treatment effect of RR can be directly estimated[76]. However, this method leads to instability in the estimation of treatment effect for subjects in the study with very low probability of receiving the treatment actually received. Specifically, the denominator vanishes, rendering large weights for a particular subject[73]. Lastly, one can adjust for confounding and estimate effect size by implementing covariate adjustment utilising propensity score[73,74]. The outcome variable is regressed upon the exposure indicator variable and the propensity score, and the association strength between exposure and outcome is determined from the fit model's exposure regression coefficient[73,74]. Since this method directly models the relationship between the propensity score and the outcome, this regression method requires that the modelling is correctly specified.

While these methods are most often used, this survey is by no means an exhaustive list. Other methods include doubly-robust estimators (introduced in Chapter 6) and other approaches that incorporate modern machine learning advances and address issues such as finite sample estimation and selection biases prevalent in the observational setting[77,78].

2.5 Deep learning

In order to introduce the concept of deep learning, the fundamental linear regression is important to facilitate discussion. For linear regression, a set of *N* input-output pairs $\{(\boldsymbol{x_1}, \boldsymbol{y_1}), \dots, (\boldsymbol{x_n}, \boldsymbol{y_n})\}$ are given. For example, input can be health status at study entry and output can be the outcome of BMI. Assuming that there exists a linear function that can map the input $\boldsymbol{x_i}$ to the output $\boldsymbol{y_i}$, the linear model can simply function as a linear transformation of the input. $\boldsymbol{f(x) = Wx + b}$, with $\boldsymbol{W}$ as a real number weight

matrix and $\boldsymbol{b}$ a real number weight vector[79]. Varying the two weights would yield different transformations of the input space, $\boldsymbol{x}$. The aim of the linear regression is to minimise some objective function – perhaps, the average squared error over the data and simultaneously best predict the outcome space, $\boldsymbol{Y}$ [79].

While linear transformations are useful for learning some relationships between input and output, many relationships are, in fact, not linear in nature. Instead, non-linear functions must be used to map the input to the output more accurately than simpler linear maps. For this purpose, linear basis function regression can be used[80]. In this case, a feature vector is created $\Phi(\boldsymbol{x}) = [\phi_1(\boldsymbol{x}), \dots, \phi_k(\boldsymbol{x})]$, and this feature vector can be used to perform linear regression instead of solely using the input, $\boldsymbol{x}$. These transformations are the basis functions and with the scalar input, *x*, these transformations can be non-linear in nature (e.g., polynomial, sinusoidal, and other non-linear functions). However, the basis functions are often just assumed to be fixed and orthogonal implying that the optimal basis functions are unknown and requires deduction. In order to alleviate the issues of these fixed functions, parametrised basis functions instead can be implemented[80]. As an example, the basis function, $\phi_k^{\boldsymbol{w_k}, b_k}$ is the aforementioned linear basis function $\phi_k$but applied with inner product to the input $\boldsymbol{x}$, yielding, $\langle \boldsymbol{w_k}, \boldsymbol{x} \rangle + \boldsymbol{b_k}$. With this inner product, the linear regression model can be notated as $f(x) = \Phi^{\boldsymbol{W_1}, \boldsymbol{b_1}}(\boldsymbol{x})\boldsymbol{W_2} + \boldsymbol{b_2}$, with $\Phi^{\boldsymbol{W_1}, \boldsymbol{b_1}} = \phi(\boldsymbol{W_1 x} + \boldsymbol{b_1})$. Not only is the linear model parametrised by weights, $\boldsymbol{W_2}$ and $\boldsymbol{b_2}$, also the basis itself is parametrised with weights to optimise, $\boldsymbol{W_1}$ and $\boldsymbol{b_1}$. With this construction, the task is to find the optimal four weights that yields the minimal average squared error.

In gist, the most fundamental architecture in deep learning can be described with parametrised basis functions. The parametrised basis functions demonstrate that the weights are functional compositions or more informally, nested functions; while the linear

regression weights itself need optimisation, the weights that directly transform the input require optimisation as well. This nested or hierarchical structure can be referred to as a neural network, with each level of feature vector constituting a layer in the network. Each layer is an iterative block in the hierarchy and with more layers comes "deeper" networks with more complex, non-linear transformations.

In the following sections, common neural network models will be introduced briefly followed by a brief comment on the "black box" nature of deep learning models and interpretability of these models.

2.5.1 Feed-forward neural networks

First, a fundamental neural network model with one hidden layer will be introduced[81]. Again, input is $\boldsymbol{x}$ and the first linear map is denoted by $\boldsymbol{W_1}$ and the bias term, $\boldsymbol{b}$ will be the translation. The resulting linear map will be $\boldsymbol{W_1 x + b.}$ A non-linear function, $\boldsymbol{\delta}$ is used to transform the output of the linear map. Some common functions are the rectified linear unit (ReLU) or the hyperbolic tangent function (TanH). A second linear map, $\boldsymbol{W_2}$ maps the output of the first linear map to the output layer (perhaps the output is a real number). The total neural network can be functionally expressed $\boldsymbol{\hat{y}} = \delta(\boldsymbol{W_1 x + b})\boldsymbol{W_2}$.

To optimise the weights such that neural network estimates the true outcome with minimal loss, a loss function must be chosen. The loss function can be the Euclidean loss function, $E(X,Y)_{\boldsymbol{W_1},\boldsymbol{W_2},\boldsymbol{b}} = \frac{1}{2N}\sum_{i=1}^{N} \|y_i - \hat{y}_i\|^2$. Minimising this loss with respect to the weights, $\boldsymbol{W_1}, \boldsymbol{W_2}$, and $\boldsymbol{b}$ would yield a model that perhaps generalises appropriately to test data (evaluation data that is not seen). However, one frequent problem with neural networks is the issue of overfitting on seen training data. To ameliorate this issue, the technique of regularisation is implemented. Regularisation in the form of weight decay is

added to the loss function on each of the parametrised weights. The full loss function is the following function:

$$\mathcal{L}(\boldsymbol{W_1}, \boldsymbol{W_2}, \boldsymbol{b}) = \boldsymbol{E}(\boldsymbol{X}, \boldsymbol{Y})_{\boldsymbol{W_1}, \boldsymbol{W_2}, \boldsymbol{b}} + \boldsymbol{\lambda_1}\|\boldsymbol{W_1}\| + \boldsymbol{\lambda_2}\|\boldsymbol{W_2}\| + \boldsymbol{\lambda_3}\|\boldsymbol{b}\| \quad (2\text{-}3)$$

With this function, one can see how the loss function is simply an addition of loss terms. If the neural network gets larger in number of layers and thus, weights, the loss can easily scale allowing more expressive modelling of the data.

2.5.2 Recurrent neural networks

Recurrent neural networks (RNN) are important methods to handle sequential data[82]. The model can take as input, sequential data, such as language or time-series data, and output both sequential outputs (e.g., next word prediction) and static outputs (e.g., text sentiment classification)[83,84]. The RNN model, at every time step, can take as input a vector of input. At the time step, an RNN cell processes the input in addition to the output from previous time steps. RNNs have demonstrated superior performance on a host of tasks. There have been many expansions and variations of the original conception of the RNN model including the Long Short-Term Memory (LSTM) and the more lightweight, Gated Recurrent Unit (GRU) networks[82,85]. With additions to the architecture such as provisioning of attention mechanisms, the model has demonstrated impressive predictive performance across many tasks including language translation[83]. In a way, a generalisation of the recurrent neural network with attention is the Transformer neural network. Instead of sequential processing of the sequential data, the Transformer model processes all inputs of the sequence at once and instead exclusively utilises self-attention to attend to all elements at once[86].

2.5.3 Convolutional neural networks

Convolutional neural networks (CNN) are special architectures of the neural network family of models that can attend to image processing tasks better than simpler statistical and neural network models themselves[81,87]. The CNN family of models uses convolution layers, a unique layer of linear transformations that attends to patches of a given image. In the layer, a kernel, a weight matrix that is smaller than the image size usually, attends to each image position one at a time and maps a small group of pixels to an output space. Following the linear map, a pooling layer maps the output space to a smaller output space. In this way, spatial context is preserved in the linear map. Furthermore, the kernel is a shared parameter that is used across the entire image, so the parametrised variables are less and lighter than fully connected neural networks.

2.5.4 Interpretability of neural network models

Unlike neural networks themselves, there is much contention about a mathematically valid definition of interpretability. Biran and Cotton define interpretability as: "the degree to which an observer can understand the cause of a decision"[88]. While more of a heuristic measure as opposed to a formal mathematical one, this definition is more or less accepted in the statistical and machine learning community. Perhaps proportionally, the "more" interpretable a model is, the "more" an observer or scientist can understand the pathways of model decision making. Ease of understanding is also important; if tools are developed to more efficiently and simply communicate the decision to the observer, this is also an important facet of a highly interpretable model. In this thesis, explainability and interpretability is used interchangeably. Although it is noted that some specialists in the field consider the two as different facets of models.

Neural networks are mathematically precise models however, the sequence of weights and non-linear activations makes it hard to understand the decisions. Rocher calls

this the “lack of transparency” in deep learning modelling[89]. In sensitive areas of research such as medicine, there is a proper need for models that are trustworthy. Without trust in these solutions, these models cannot be implemented in clinical practice. Thus, many secondary methods have been created to interpret these “black box” models.

Some of the methods offer more interpretability than others and some are more appropriate for certain types of networks (e.g., feed forward as opposed to recurrent). One of the fundamental ways is to directly inspect the weights. However, with many neural network models having millions if not billions of parametrized weights, the complexity of inspecting the neural network is replaced with another complex task: distilling the results of the direct weight analyses in a palatable way. With weight-sharing models such as CNN, this is easier, but only the first layer can be easily understood. With multiple non-linearities and pooling operations in subsequent layers, the deeper features extracted are not made transparent; ultimately, the issue is still a lack of transparency. There are many activation methods as well such as DeepLift that present the output of the neuron in a particular layer for a given input image/matrix[90]. Variations involving the gradient have been proposed as well; however, these are mostly used in the image setting with CNNs[90–92]. For sequence-based models such as RNNs, attention has helped interpret the decision making for these difficult models. However, there are concerns that attention is actually not a good measure of the processes of a neural network; rather attention is a noisy predictor of output at best, and is not completely trustable indicator of signal[93].

3 DATA: CLINICAL PRACTICE RESEARCH DATALINK

In this section, the EHR dataset, the CPRD database will be introduced. First, background information concerning the CPRD dataset will be presented. Second, the organisation of the dataset will be presented. Third, given the organisation of the database, the nuances of the linkages to other datasets will be presented. Fourth, the degree of validity and completeness of the records will be discussed. Fifth, the ethical approval required for undertaking research will be summarised and presented. Lastly, the data cut and high-level details of the dataset used in these doctoral research projects will be presented.

3.1 Background

The CPRD organisation is a service intended to support epidemiology and clinical studies that utilise EHR for research purposes. The CPRD is partly sponsored by the National Institute for Health Research, and collects anonymised patient health data from a network of primary care practices around the UK[54]. In addition, the CPRD database offers provisions to link to other healthcare datasets in the UK[54].

Historically, the CPRD database has progressed from a general practice information recording system to a database that collects and organises a full range of

healthcare variables and in addition, allows linkage to various other healthcare datasets. An organisation that has been conducting data collection for over 30 years, the database has evolved into one of the most trusted and utilised primary care databases in the world[54].

As of November 2021, the CPRD database holds patient data from over 60 million individuals registered at over 2,000 practices, of whom, approximately 16 million are currently registered. Furthermore, over 25% of patients have over 20 years of follow-up allowing for high-quality study of long-term, chronic conditions[54,94]. The database contains coded EHR that captures information on demographic characteristics, diagnoses, medications, vaccinations, laboratory tests, and referrals to secondary hospital and specialist care. All forms of data are prospectively collected are derived from routine clinical care. Hence, observational studies utilising such data are free of biases such as recall bias and non-response bias, both forms of biases prevalent in traditional observational studies[95]. The dataset used in this doctoral research is the "CPRD-GOLD" database henceforth referred to as CPRD. Specifically, the "GOLD" version of this dataset uses the Vision EHR software for recording and organization of the records[94].

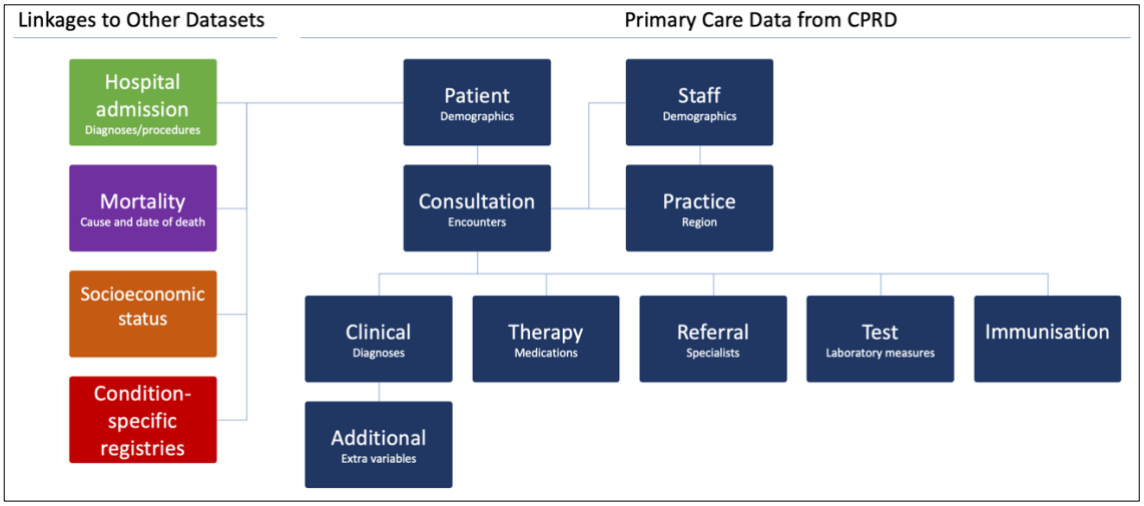
CPRD has been used as data source in over 3,000 publications including several epidemiological observational studies investigating conditions such as cardiovascular diseases, blood pressure, diabetes[43,96,97]. Furthermore, more recent publications have investigated multimorbidity and blood pressure trajectories in UK patients[98–100]. In large part due to the data richness and linkage capabilities that CPRD offers, the database has been vital for observational research.

3.2 Organization

CPRD, first and foremost, is a dataset that contains health variables collected during primary care visits. Multiple modalities (e.g., diagnosis, medications,

measurements) are offered by the dataset; however, the patient identification information that can assist in the de-anonymisation of the patient (e.g., patient name) is naturally excluded from the dataset. Linkage, furthermore is not guaranteed across the dataset. Only data from primary care practices that have agreed to allow patient-level linkage have the potential for linkage to secondary care, specialist care, and death registries datasets[94].

Figure 3-1: Clinical Practice Research Datalink (CPRD) organisation



This is a figure of the organisation of the CPRD database and relevant linkages derived from Herrett et al[5].

CPRD records are organised in a relational organisation; a hypothetical identifier (e.g., practice identifier) from one particular dataset has the potential to be linked to other datasets via the same identifier. An illustration of the linkage potential of CPRD is offered in Figure 3-1. In future chapters, the data processing for a particular investigation will offer more details into exact data modalities utilised.

In terms of representativeness of the CPRD data, the CPRD patient cohort used in this doctoral research has been demonstrated to be broadly representative of the UK population in terms of age, sex, region of birth, and body mass index[101–103].

3.3 Linkage

The datasets offered in CPRD, as discussed, offers access to a host of health variables across multiple datasets providing a comprehensive view of health for a hypothetical patient. First and foremost, the CPRD primary care database offers the potential to link eligible patients (i.e., those patients from practices that have agreed to CPRD linkage) to hospital episode data (Hospital Episode Statistics (HES)). Second, the CPRD importantly offers linkage to mortality data from the national death registry (Office of National Statistics (ONS)) enabling downstream research to make use of time and cause of death information. These linkages are only available for patients in England as opposed to all UK patients; due to the differences in NHS data collection in Wales, Northern Ireland and Scotland, linkage across the UK is not offered. With the data provided by CPRD for this doctoral research, linkage was offered for approximately 75% of the English practices accounting for approximately 50% of patients in CPRD.

The HES dataset covers data on all hospital admission to English NHS hospitals including hospital, primary care, and mental health trusts. In this doctoral research, HES data considered primarily stems from the Admitted Patient Care (APC) relevant health records. HES APC data consist of the health variables recorded during a hospital admission, in which a hospital bed was required (including day cases). In the dataset used for this doctoral research, accident and emergency-based hospital encounters were not given and hence not considered in downstream research[104,105]. HES data notably provide information at the patient-level, with diagnoses recorded during hospital admissions. First diagnoses in a single hospital admission are considered as primary diagnoses. The coding dictionary used to identify the hospital admissions are the International Classification of Diseases and Health-Related Problems, 10th edition, henceforth referred to as ICD-10 codes. Lastly, whilst procedures are theoretically provided as a part of the hospital

admission package, the data on hospital procedures were not provided for the data cut utilised for this doctoral research.

Secondly, the death registry data from ONS provides health information concerning mortality and relevant details concerning death of patients. The death registry provides data for all deaths in England and Wales. The dataset provides date of death, primary cause of death, and 15 secondary causes of death. The diagnostic coding system to identify cause of death is the ICD-10 system[106].

3.4 Data validity

Downstream research utilising EHR is predicated upon the fact the data analysed is valid and provides some minimal assurance of completeness and quality. To meet the demands of high-quality research, the CPRD organisation has taken efforts to ensure that the data provisioned meets high standards for downstream research.

Various contributors of data to the CPRD database are required to meet criteria for data submission. Primary care practices contributing to CPRD data are required to (1) record episode of illness or new occurrence of symptom or (2) all notable morbidity encounters (e.g., notable clinical contacts, diagnoses, abnormal laboratory test outputs, referrals to secondary care)[107]. In the case of diagnoses specifically, diagnoses must be manually recorded in the computers provided implying incompleteness and accidental errors in coding/phenotyping. Naturally, in terms of recording errors, there might be errors in diagnoses itself that may be recorded[107].

Therapy data points in CPRD is directly an output of administrative recording of the prescription in NHS computers. Specifically, minimal effort is required on the side of the clinical care practitioner; once the prescription has been initially noted, the dosage, renewals, and other information is automatically inputted. Ultimately, this implies that

the prescription recordings in CPRD are quite complete save medications taken over the counter or those prescribed in secondary care practices.

In terms of methods to validate quality of a particular record, the CPRD organisation conducts a variety of checks to validate the data points and make sure that the records provided are "up to standard" (UTS). The UTS checks are an assessment of individual patient data (e.g., checks of age, sex, registration, etc) in addition to completeness checks at the population level. As an example of the latter form of checking, the CPRD organisation ensures that a certain percentage of deaths are recorded at the least thereby ensuring a minimal measure of completeness[107]. Additionally, for contributing practices, all data points provided may not be considered as UTS; specifically, time periods, in which data points remains UTS and high-quality are offered by the evaluation services provided by the CPRD organisation. At the level of individual records, the records are labelled as "acceptable" or unacceptable as a function of an algorithm that excludes patients (and thus records) based on algorithmically determined suspicious recording patterns (e.g., discontinued follow-up, unknown general practice registration date)[54]. The dataset used in this doctoral research has exclusively been limited to records that have met quality standard checks; records that fall within UTS time periods provided and those, which are given the "acceptable" label.

3.5 Ethical approval for research

For this doctoral research, scientific approval was given by the CPRD Independent Scientific Advisory Committee (ISAC). The protocol number is 16_049R. The CPRD organisation provides anonymised data for research purposes under approval given the National Research Ethics Service Committee. Other than the approval by the CPRD ISAC, no other approval was sought or required for this doctoral research[94].

Whilst conducting research, safeguards have been taken to ensure that the CPRD data was handled with care. Data storage and processing occurred on secure computing environments – namely, servers on the university network and university provided personal computing devices that are provisioned with data encryption and password protection. As stated in the ISAC protocol, data in this doctoral research were exclusively utilised for the purpose of research.

3.6 Data used in doctoral research

In this dissertation, the CPRD cut used for analyses covers records from January 1 1985 to September 30 2015. This data cut contains primary care longitudinal records from a network of 674 general practices (GP) in the UK, linked to secondary care APC HES data and covers approximately 7% of the UK population. Furthermore, the data cut provides linkage to mortality records from the ONS providing causes and date of death.

4 MODEL DEVELOPMENT AND RISK PREDICTION

4.1 Introduction: From inference to causal inference

Conventionally, when using adjusted statistical models for association studies, model fitting in some form is conducted. For example, with the regression modelling approach, the outcome variable is regressed on a host of adjustment variables in addition to the exposure variable of interest. Specifically, the fit of the model is tested, and the study of the exposure variable does not proceed without ensuring adequate model fit. If the fit and generally speaking, predictive accuracy, is acceptable, analysis of the direction and strength of the association between the exposure and outcome is conducted.

In the same vein, the modelling paradigm for statistical models is extended to the deep learning framework of modelling. By testing proposed deep learning architectures in risk prediction experiments, the model is vetted for fit and predictive accuracy prior to exploring the potential of the model for causal inference and association analyses.

To this end, in these following sections, the various risk prediction investigations conducted for ascertaining sufficient predictive accuracy (i.e., the first objective – see section 1.2) will be presented. On the path to developing deep learning models for

observational causal inference utilising rich EHR, first, I present development of deep learning models that can appropriately handle multimodal, longitudinal records and capture important features in EHR shown via risk prediction investigations. Specifically, in section 0, I present development of the model, BEHRT, for handling multimodal EHR. Furthermore, I present how the model is utilised for prediction of subsequent occurrence of diseases and compared to benchmark deep learning models. Second in section 4.3, I present how the model was utilised for incident HF risk prediction. Investigations on these two tasks are conducted in order to ascertain the predictive utility (i.e., inference capability) of the proposed architecture prior to conducting causal inference downstream.

The following sections are published works: *"BEHRT: Transformer for Electronic Health Records"* in *Scientific Reports* (doi.org/10.1038/s41598-020-62922-y) and *"An Explainable Transformer-Based Deep Learning Model for the Prediction of Incident Heart Failure"* in *Journal of Biomedical and Health Informatics (JBHI)* (doi.org/10.1109/JBHI.2022.3148820). These publications are products of work by multiple authors. As a co-first author, for both publications, my role consisted in designing the study, conducting literature review, processing data, conducting statistical/deep learning analyses, and writing the first draft of the manuscripts. Material from the publications (including figures, tables, and text) have been amended for presentation in the following sections.

4.2 Deep learning modelling for electronic health records: model development and subsequent disease prediction

4.2.1 Introduction

Prognostication of future disease is key for preventative care. While simpler models indeed show that, for example, CVD risk is elevated generally in the population in the presence of some select risk factors (e.g., diabetes), the end goal is to create models that predict accurately in a personalised fashion. While the means to conduct such nuanced prediction requires further thought and research, the ultimate aim is to improve preventative medicine with precision and nuance[108].

Deep learning modelling has made some initial progress in cardiovascular medicine, radiology, and other subfields of medicine. For example, recent research has developed and validated high performing deep learning models for the prediction of atrial fibrillation utilising electrocardiograph data[109]. In the area of general cardiovascular disease research too, recent research has demonstrated that deep learning can perform cardiovascular disease-related prediction quite well utilising photographs of the retinal fundus[110]. In terms of understanding trends in clinical data, a study by Liu et al demonstrated how deep learning can predict paediatric "no-shows," or scheduled but unattended appointments[111].

With respect to accessibility of EHR data specifically, adoption of EHR has increased over the past several years. EHR administrative databases are now capturing millions of patient lives with recordings of health variables spanning over many decades. Furthermore, with linkage to multiple other healthcare databases, researchers using EHR data can get a comprehensive view of the medical health timeline of a hypothetical

patient. With this, large-scale EHR is a rich source of untapped research potential that is fertile grounds for interdisciplinary research concerning deep learning and health.

In traditional EHR data research, including ones using conventional machine learning techniques (e.g., tree-based regression methods), health for an individual is represented as input predictors. These predictors can be normalised and transformed and represented as a vector for each individual. Of course, as previously discussed, many a time, conventional predictor selection is predicated upon experts selecting some subset of important variables to be considered for modelling. On the other hand, recent deep learning approaches capture useful representations of baseline health using raw or minimally processed health variables.

Due to this reason, deep learning models have found initial success in modelling of EHR data as well. For example, Liang et al demonstrated that deep neural network models can outperform other conventional machine learning solutions (i.e., support vector machines, decision trees) on a number of different EHR related tasks and datasets[112]. In more recent research, Tran et al proposed the use of Restricted Boltzmann Machines for EHR modelling for suicide prediction outperforming models that utilised expert curation of predictors[113]. While these works applying deep learning on EHR failed to take into account the subtleties of EHR data (e.g., ordering of events, irregularity of the time interval), Nguyen et al developed a deep convolutional neural network model, Deepr, to address this limitation of past methods. Deepr utilised different concepts including diagnoses and medications to predict hospital readmission. In addition, the model sensitively accounted for time between visit by adding a token that represents the time difference[114]. Deepr was developed for the Australian EHR data setting; the model was developed, trained, and internally validated on data from a private Australian hospital offering diagnosis and procedure records for modelling[114]. Similarly, Choi et al, utilised

a recurrent neural network, Doctor AI, for predicting of health outcomes[115]. Unique to both Deepr and Doctor AI modelling, high-dimensional embeddings were used to represent clinical concepts[114,115].

In parallel, in natural language processing research, there was a landmark development in methods to process sequential data leading to the "attention" class of methods[83]. In the case of long sequences, a frequent problem manifested with RNN modelling: the issue of the fixed-length vector information bottleneck. As sequences get longer, more information must be represented by a summarisation vector that takes into account data in the long sequences. Alternatively, for both long and short sequences, the dimensionality of the downstream high-dimensional vector representation of health remains constant implying degradation in information retention in these vectors. The attention mechanism was developed to take into account only the relevant parts of data in the long sequence; relevance, naturally, is a function of the task at hand[83]. Features considered relevant for the task of language translation might be irrelevant for language sentiment classification. Incorporating this attention mechanism in traditional RNN architectures, such as LSTM yielded great improvements in predictive performance over a diverse set of tasks[83].

Given the success of the attention mechanism is language-based tasks, Choi et al proposed a LSTM model called RETAIN utilising reverse-time attention to focus on past notable visits for incident heart failure prediction[116]. RETAIN and more complex varieties such as RETAIN-EX successfully outperformed benchmark deep learning models and demonstrated the power of incorporating minimally processed EHR data[116,117]. RETAIN and RETAIN-EX models were developed on American and Korean datasets respectively. RETAIN was developed on a private Sutter Health dataset, consisting of patients between 50 and 80 years of age chosen for a HF prediction study[116].

The Sutter Health dataset offered access to demographic information (e.g., sex, ethnicity) in addition to disease, medication, and procedure records. On the other hand, the RETAIN-EX model was developed on the Health Insurance Review and Assessment Service Dataset, the national insurance services dataset for Korea[118]. This dataset consisted of 1.4 million Korean patients and provided demographic information in addition to disease, medication, interventional data records.

As the attention mechanism found success across language and EHR-based prediction tasks in recurrent neural network models, the same mechanism was generalised into a model that solely utilises attention without recurrent neural networks: the Transformer model. The Transformer demonstrated superior performance on several language translation tasks[86]. Furthermore, the encoder module of the Transformer was developed into a stand-alone model called Bidirectional Encoder Representations from Transformers for language classification tasks[119].

4.2.2 Aims

In this work, I aim to develop a Transformer-based model for EHR data: BEHRT, named after Bidirectional Encoder Representations from Transformers (BERT), from which the model derives inspiration and more concretely, a large part of its neural network architecture. I aim to evaluate this model on a host of EHR-related prediction tasks and compare predictive performance to that of benchmark models. In addition, I aim to investigate elements of the model's architecture, such as the embedding structure, for clinical relevance.

4.2.3 Methods

4.2.3.1 Data

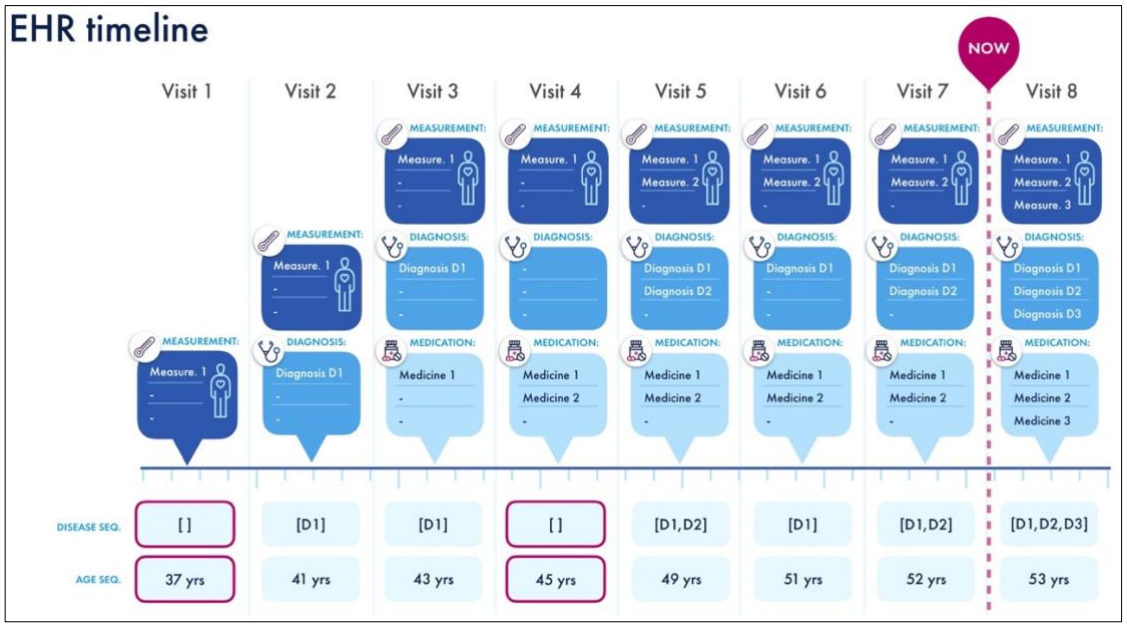
In this study, we used the described cut of CPRD (see section 3.6) covering records from January 1 1985 to September 30 2015 [5]. Initially with access to data from 8 million patients, in this study, we only considered data from GP that allowed for data linkage with HES in order to only consider patients with documented comprehensive medical health history. Furthermore, we limited the analyses to records that have met the quality checks that CPRD conducts. Additionally, to only keep those patients that have enough records for prediction, we kept those patients with at least 5 visits in the medical history – specifically, a visit counting as an encounter with GP or HES staff, in which diagnoses for conditions were recorded.

In the CPRD dataset, the disease coding system for GP diagnosis records are different than that for the APC HES linked diagnosis records. While the diseases in GP records are classified using Read codes, the diseases recorded in the HES dataset is encoded in 10th Revision of the International Statistical Classification of Diseases and Related Health Problems (1CD-10)[120,121]. ICD-10 system forms a clear hierarchical formulation of categorizing diseases; diseases systems form chapter while disease groups form sections in individual disease chapters and so on. In this way, while there exist 22 chapter-level ICD-10 codes, at the sub-chapter level, the coding system yields approximately 1,900 codes[121].

Since both these coding systems are heterogenous from one another, for machine readability, a harmonised coding system would be more appropriate. To this end, we used the Caliber phenotyping system to map codes from Read system and ICD-10 system to 301 Caliber disease codes[122]. In this way we denote $D = \{d_i\}_{i=1}^{G}$ where d_i denotes the ith disease code. Given a patient $p \in \{1, 2, \dots, P\}$, the medical history is made up of various

visits to the GP and hospital. Each visit contains encounters such as diagnosis, medications, and laboratory tests. In this study, however, only diagnoses records were considered.

Figure 4-1: Medical history of a hypothetical patient



The raw medical history of a hypothetical patient visualised. This patient's medical history consists of 8 total visits. In each of the visits, the records are diagnoses, medications, and measurements. For this work, we only take into account the diagnosis and the age. At the bottom of the image, the disease and age sequences are demonstrated for clarity. Some visits in the medical history will not be represented in this investigation due to the lack of diagnoses in those particular visits (purple boxes in the lower part of the figure). This figure was adapted from Li et al[123].

A hypothetical patient's EHR is denoted as $V_p = \{v_p^1, v_p^2, v_p^3, \dots, v_p^{n_p}\}$, for which the value, n_p is the number of visits in the medical history of the patient p. v_p^j is the collection of diagnoses in a particular visit – in this case, the jth visit. Since there may be more than one diagnosis in this particular visit, we generalise the notation of a visit to contain a list of m_p^j diagnoses (i.e., can be written out as $v_p^j = \{d_1, \dots, d_{m_p^j}\}$). Additionally, since sequential data are unsuitable for proposed deep learning architecture, BEHRT, we (1) order the data sequentially as written out in notation above, (2) introduce

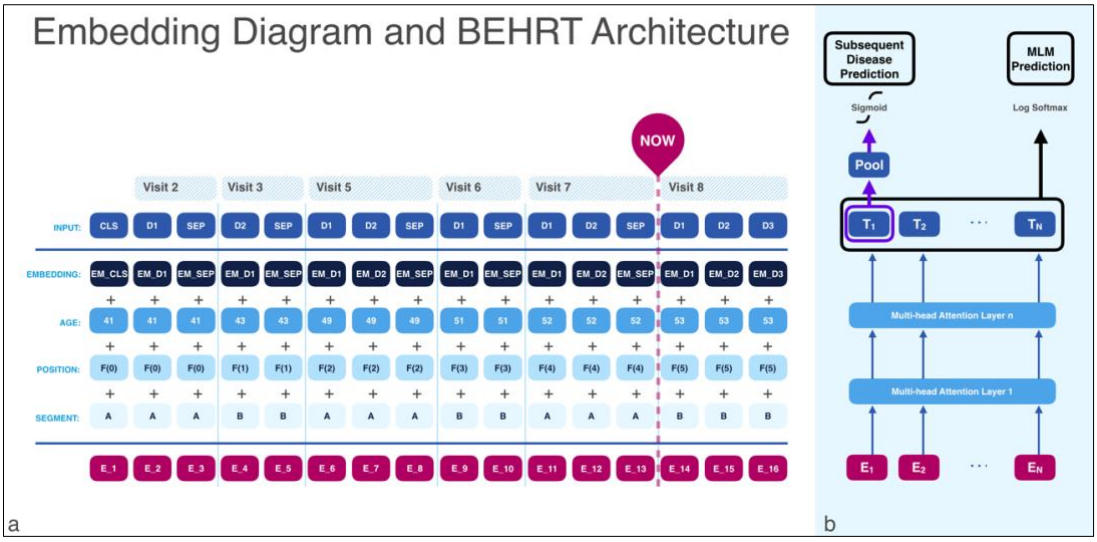
a new token to separate the visits ("SEP"), and (3) introduce a new token to denote a classification token or alternatively a token to signal the start of the medical history ("CLS")[119]. Ultimately, the sequence of medical history is amended to become: $V_p = \{CLS, v_p^1, SEP, v_p^2, SEP, \dots, v_p^{n_p}, SEP\}$. Henceforth, the medical history of a hypothetical patient will be represented in this way (shown in Figure 4-1).

4.2.3.2 Model development

In the conventional Transformer model, language is modelled at the word, sentence, and the paragraph level[86]. Words are encoded as individual tokens represented by high-dimensional embeddings, sentences are collections of tokens organised using sequential cues given to the model, and paragraphs are ordered sequences of sentences with separation tokens indicating the end of one sentence and the beginning of another.

While the Transformer model addresses some of the challenges posed by sequential data, the modelling of EHR specifically needs to account for four necessities of modelling: (1) necessity of capturing complex, non-linear interactions between encounters/visits, (2) necessity of representing heterogeneous encounters in visits of various cardinalities, (3) necessity of modelling irregularity of time between visits, and (4) necessity of modelling associations across time.

Figure 4-2: BEHRT embedding structure and overall architecture



Utilising the hypothetical simulated data in Figure 4-1, the figure illustrates the BEHRT model handling multimodal EHR. (a) demonstrates how BEHRT handles medical history sequence of variables. With disease, age, position, and visit segment embeddings, the model represents each encounter with rich attributions of relative time and visit number. The summation of these individual embeddings serves as a rich, time-aware way of embedding encounters. (b) presents the BEHRT architecture and the Transformer architecture in general. The summed embeddings are inputted and stacked layers of multi-head attention extract associations. Masked Language Modelling (MLM) is used for pre-training the model weights and the subsequent disease prediction task is the main supervised prediction task. This figure was adapted from Li et al[123].

BEHRT, with modification of the original Transformer chassis can directly account for these four requirements. First off, BEHRT is a Transformer derivation implying that certain elements of the model are defined a priori. The BEHRT model makes use of the native elements of Transformer model: the contextualised embedding layer, multi-head self-attention, and stacked Transformer encoder layers[86].

With contextualised embedding layers, the richness of raw EHR can be flexibly handled (Figure 4-2). Embeddings simply are high-dimensional vectors that can be trained with respect to a prediction task. The embedding layer consists of four individual embedding structures: the encounter (i.e., in this case, disease embedding), the position, the age, and the visit segment embeddings. Diseases are represented by the encounter

embeddings layer; the Caliber codes as described earlier, is mapped to high-dimensional vectors. In this way, the multiple disease codes can be represented by these vectors stacked on top of one another to form medical history. The same disease in medical history perhaps repeated is mapped to the same vector. The position encoding is directly borrowed from the original Transformer architecture; in this case, this encoding is a pre-determined function that maps visit number (position) to a vector[86]. The pre-defined function (given by Vaswani et al) was used in order to mitigate issues in imbalanced learning of variable sequence lengths in EHR data. In recurrent networks, the position may be implicitly given to the model via the sequential ordering of the data points. Transformers, on the other hand, do not use a recurrent architecture[86]. Thus, sequential ordering must be given to the model in an alternate way and positional encodings have demonstrated utility in informing sequence to the model (Figure 4-2). Age embedding is an embedding formulation created exclusively in order to handle the age component within EHR data[86]. Age at the time of the record of the encounter is mapped to a trainable high-dimensional vector. Much like position, this is done for every encounter (Figure 4-2). While language data do not have to handle this problem since age is not a construct native to language, age must be sensitively handled in EHR data since age is an important risk factor for many conditions. Lastly, the visit segment, denoted as either A or B, are additional symbolic information provided to the model to denote the separation between visits (Figure 4-2).

In total, the four embeddings, encounter, position, age, and visit segment embeddings form a sequence by themselves of the same cardinality for each patient (Figure 4-2). For “SEP” tokens (represented as encounter embeddings), age, visit segment, and position embeddings are supplemented from previous visit. For “CLS”, the other three embeddings take the values of the same embeddings from the first encounter

of the first visit. There is no prescribed order for the many diagnoses within a particular visit. Since position, age, visit segment is the same across all the diagnoses within a visit, the ordering of the individual diagnoses within a visit is irrelevant. Hence, an issue in some RNN architectures is alleviated; the sequence of individual diseases does not have any effect on the predictive process, but rather, the annotations of position, age, and visit segment of the visits are important[82]. This makes the BEHRT model order-invariant for the encounters within a visit. The total embedding for a particular encounter is the sum of the embeddings of the disease, the age, the position, and the visit segment.

Next, another important component of the BEHRT model is the multi-head self-attention module[86]. This module is a generalised attention mechanism which conducts multiple attention mechanism operations in parallel with the use of "heads"[86]. The independent outputs of each of the heads are concatenated and transformed into the required dimensionality. With multiple heads, each head can conduct a particular niche form of attention (e.g., attention on short-term dependencies as opposed to longer term or vice-versa).

Lastly, in addition to the embeddings and multi-head self-attention innovations for sequential learning, the Transformer architecture has a flexibly modular design[86]. Specifically, after the initial transformation using the embedding layer, BEHRT can stack multi-head self-attention layers to create a deeper model. With more layers, the network can naturally extract more complex longitudinal associations hidden in rich EHR. Thus, the modular design implemented on a slew of graphical processing units (GPUs) allows for high-throughput, deep network modelling of EHR.

Overall, BEHRT has many advantages with respect to previous approaches to model EHR. Firstly, we use feed-forward-like neural network models instead of the sequential variety thereby handling EHR encounters across time in parallel with one

another as opposed to sequentially. In addition to modelling advantages of the multi-head self-attention, the run time for training is cut down. Furthermore, RNN's suffer from the issue of exploding/vanishing gradient – meaning at some weights the gradient either explodes meaning the model weight updates involve large updates on the weights and training becomes unstable or the gradient vanishes, implying that the weights are failing to be updated efficiently, implying that the model has effectively stopped training[124]. Both issues hurt the models' ability to train effectively and thus hurt predictive performance. Convolutional neural networks, alternatively, suffer from issues in predictive power due to limitations in the receptive field. BEHRT's structure addresses both problems methodologically. First, often, the recurrent structure is one of the root causes of gradient-based issues, and BEHRT adapts the attention mechanism to function effectively without the RNN framework. Second, the limitations of a narrow receptive field are circumvented by multi-head self-attention attending to the entire sequence at once – effectively implying a receptive field that encompasses the length of the medical history. In sum, the model theoretically addresses many limitations of previous forms of deep learning models.

4.2.3.3 Pre-training

In order to allow the weights to capture latent longitudinal associations in the data (in both forward and reverse time fashion), the model was pre-trained on a unsupervised learning task. The task, masked language modelling (MLM) was derived and amended from the original BERT paper[119]. In gist, the model weights are first initialised using random initialisation. Then, the model is fed input of disease encounters and accompanying, age, position, and visit segment information. However, the disease encounters are not given as per the original data; instead, only 86.5% of the disease encounters are given as original encounters, 12% of the encounters are masked and

replaced with a "MASK" token, and the remaining 1.5% of the encounters are replaced with randomly-chosen diseases[119]. With this input, the model, BEHRT, is asked to predict the masked diseases correctly. Since the model is formulated as an end-to-end model, BEHRT predicts the masked disease and the error in prediction is calculated across a mini-batch of patient medical histories and gradients are computed with backpropagation and model weights are updated. Furthermore, with more modalities such as medications or laboratory measurements, this process can easily adapt to different vocabularies for encounters. Simply, the disease, medication, or test will be masked, and the model will be tasked with prediction of this generic health encounter. The classifier for the MLM task takes the output encounter states, shown as $T_1 \dots T_N$ in Figure 4-2.

In this task, the BEHRT model does not actually know which diseases are masked, so a contextual representation of all diseases is kept in latent space. Furthermore, the deviation from ground truth records only makes up 13.5%, so the model still learns from EHR data that has undergone minimal corruption due to masking and relabelling. Interestingly, of the 13.5% of corruption, the 1.5% due to the relabelling of disease alternatively acts as noise injection, a common method to build autoencoder models with better generalisation properties[125]. Alternatively, another interpretation of the 13.5% manipulation of encounter data is that the model is being trained using data augmentation strategies via masking and noise injection. Data augmentation strategies are also useful for improved generalisation and robustness of the model[126].

4.2.3.4 Subsequent disease prediction

The BEHRT model, following pre-training was trained and evaluated on three subsequent disease prediction tasks: (T1) prediction of disease in the next visit, (T2) prediction of diseases in the next 6 months, and (T3) prediction of diseases in the next 12 months.

For data processing, the patients in our dataset following initial CPRD data cleaning and filtering for patients with adequate number of visits, were split into train and test (80% and 20% of the data respectively) sets.

The examples in terms of input and output were defined in the following way for T1. First, randomly a visit index was chosen, j (such that $3 < j < n_p$) for each patient. And then, with this index, j, the input derived was $x_p = \{v_p^1, \dots, v_p^j\}$ and the output was $y_p = w_{j+1}$, with w_{j+1} being a multi-hot vector with cardinality G, with indicators of *1* denoting that indeed, the disease has occurred in the next visit. In this setup, for each task, each patient can maximally only contribute to one input and output pair.

For T2 and T3, the input and output have slightly amended processing steps. For patients without 6 or 12 months of history respectively for the two tasks after v_p^4 (i.e., those without 6/12 months of history following the fourth visit) will be excluded from the experiments. Additionally, the j is chosen randomly for these patients from $(3, n_*)$, for which the n_* denotes the greatest visit index such that after this visit there exists at least 6 or 12 months of records for T2 and T3 respectively. Lastly, we clearly identify the outputs for T2 and T3 as $y_p = w_{6m}$ and $y_p = w_{12m}$ respectively. These two vectors are again similar multi-hot vectors with cardinality as defined above for T1.

As a note, the BEHRT model is by design, "forced" to predict diseases in the patient's medical history, and only patients with at least 1 diagnosis in the next visit, next 6 months, and next 12 months for T1, T2, and T3 respectively will be included in the dataset as eligible patients.

These inputs (defined for each task) are inputted into BEHRT as medical history with appropriate age, position, and visit segment annotations. The output multi-hot prediction is created using a pooling layer implemented on the output state for the "CLS"

encounter token (see Figure 4-2). In this way the model is trained and tested on each of the tasks, with the predictions denoted as y_p^*, in which the *i*th entry of the prediction corresponds to the prediction of the patient having disease, d_i.

For evaluating the predictive performance, two metrics were used: area under the receiver operator characteristics (AUROC) and the AUPRC. AUPRC is a weighted mean of the precision and recall as a function of threshold. The AUROC and AUPRC was calculated for each patient first, and then averaged over all patients for a summary statistic.

In order to evaluate the proposed model's predictive performance with respect to known benchmarks in deep learning, we also implemented the Deepr and RETAIN model for model comparison[114,116]. While the implementation of Deepr was identical to the one used in the original publication, to boost predictive power, we amended RETAIN to include sex as a predictor, encoded timing into the visit level data, and additionally included bidirectionality on the foundational RNN framework[127]. These suggestions were offered by the original authors in the open-source code repository containing code material from the original publication[127].

For all three tasks, all three models (BEHRT, Deepr, and RETAIN) were evaluated on identical training and testing data (i.e., total of nine implementations). Naturally, while modifications of raw EHR (e.g., "SEP" and "CLS" inclusion) were conducted for inputting data into BEHRT, appropriate modifications of raw EHR were made for inputting data into Deepr and RETAIN.

In order to further investigate the worth of the BEHRT predictive performance, two secondary analyses were conducted: (1) we investigated if BEHRT can capture sex implicitly (since sex is not explicitly modelled by BEHRT embeddings), and if the

knowledge captured is useful for downstream prediction, and (2) we investigated the model's performance on the prediction of incidence of diseases (i.e., predicting diseases that had not previously occurred in the medical history prior to prediction period).

4.2.3.5 Embedding analyses

Furthermore, we investigated the embeddings after pre-training to investigate disease in a qualitative and quantitative manner. While clinical members of our team probed the clinical validity of the clusters derived in embedding latent space in a qualitative manner, a quantitative investigation was additionally conducted in-house within the Deep Medicine research group. For each disease that occurred in at least 1% of the population (87 Caliber phenotype diseases), the top 10 neighbouring diseases were first captured (using cosine similarity measures). We compared these derived neighbourhoods to 10 diseases provided as "neighbouring" conditions by a clinical researcher in a quantitative analysis. Specifically, for a given disease, as a fraction, we quantified how many of the 10 neighbouring diseases found by the model were found by the clinical researcher (i.e., $\frac{x}{10}$). We calculated the mean of this measure calculated for all 87 diseases in order to present on average, the overlap between model-derived sets of 10 closest neighbours for diseases and those found by the clinical researcher.

4.2.3.6 Implementation

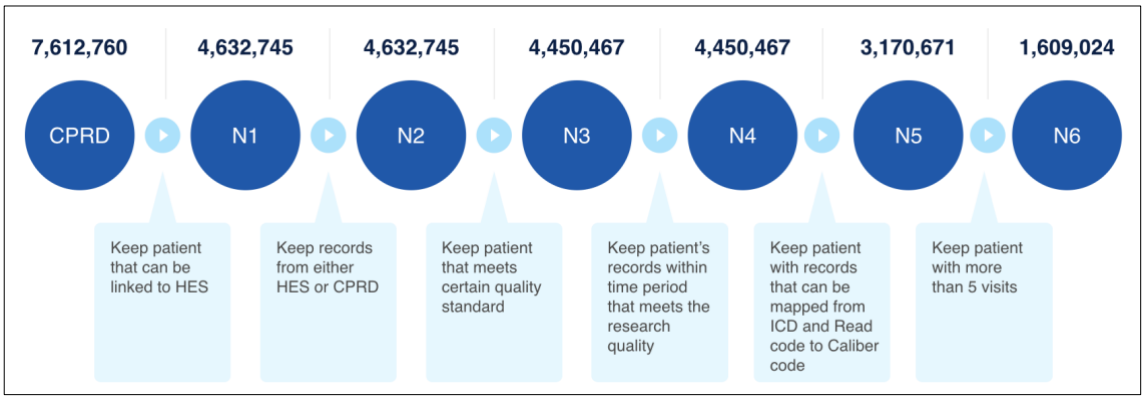
All data processing and modelling was conducted using the language, Python. For deep learning analyses, the NVIDIA Titan Xp Graphical Processing Units (GPU) were utilised for pre-training, training, and testing of the BEHRT model. Furthermore, we used Bayesian optimisation methods to find the optimal hyperparameters at the MLM pre-training stage[128]. The main hyperparameters investigated were number of layers, number of attention heads, hidden size, and intermediate size. These hyperparameters are also the same hyperparameters in the original, BERT architecture as well. Naturally,

optimal hyperparameters captured in this stage were used in the three subsequent disease prediction tasks. Also, for RETAIN and Deepr, Bayesian optimisation was used to find the optimal hyperparameters for the models[128]. The hyperparameter details can be found in Supplementary Table S1, Table S2, and Table S3.

4.2.4 Results

4.2.4.1 Population characteristics

Figure 4-3: Data processing flowchart



This is the flowchart for data processing on Clinical Practice Research Datalink (CPRD) data. Linkage is conducted between general practice and Hospital Episode Statistics dataset (HES). This figure was adapted from Li et al[123].

After data-processing, 1.6 million individuals were eligible for downstream MLM pre-training. For risk prediction analyses, after filtering processes on 1.6 million patients to ensure sufficient number of visits for T1, T2, and T3, we had 699, 391, and 342 thousand patients respectively (Figure 4-3). Full patient characteristics for each of the three tasks are shown in Table 4-1.

Table 4-1: Characteristics for patients eligible for subsequent disease prediction

Characteristic		*Next visit*	*Next 6m*	*Next 12m*
Gender	Male	41.80%	42.30%	41.70%
	Female	58.20%	57.70%	58.30%
Ethnicity	White	46.40%	48.30%	47.40%
	Unknown	43.80%	44.00%	44.50%
	Indian	0.40%	0.50%	0.50%
	Other	0.30%	0.30%	0.30%
	Pakistani	0.20%	0.30%	0.20%
	BI Carib.	0.20%	0.30%	0.20%
	other Asian	0.10%	0.10%	0.10%
	BI Afric.	0.10%	0.10%	0.10%
	Mixed	0.10%	0.10%	0.10%
	Bangladeshi	0.08%	0.07%	0.07%
	BI Other	0.07%	0.06%	0.06%
	Chinese	0.06%	0.06%	0.05%
Age Start	0.25 Quantile	45	46	46
	0.5 Quantile	58	60	59
	0.75 Quantile	70	71	70
Age End	0.25 Quantile	56	58	58
	0.5 Quantile	70	71	71
	0.75 Quantile	81	82	82
Unique Codes	0.25 Quantile	7	8	8
	0.5 Quantile	9	10	11
	0.75 Quantile	12	14	14
Number of Visits	0.25 Quantile	10	14	14
	0.5 Quantile	15	20	20
	0.75 Quantile	24	30	30

Age start and age end correspond to the age at first and last visit respectively. Table adapted from Li et al[123].

4.2.4.2 Risk prediction model comparison

The results from the risk prediction task can be seen in Table 4-2. Across the three tasks, we see that BEHRT clearly demonstrates superior predictive performance as compared to the benchmark model, RETAIN and Deepr on both metrics for evaluating performance (AUROC and AUPRC).

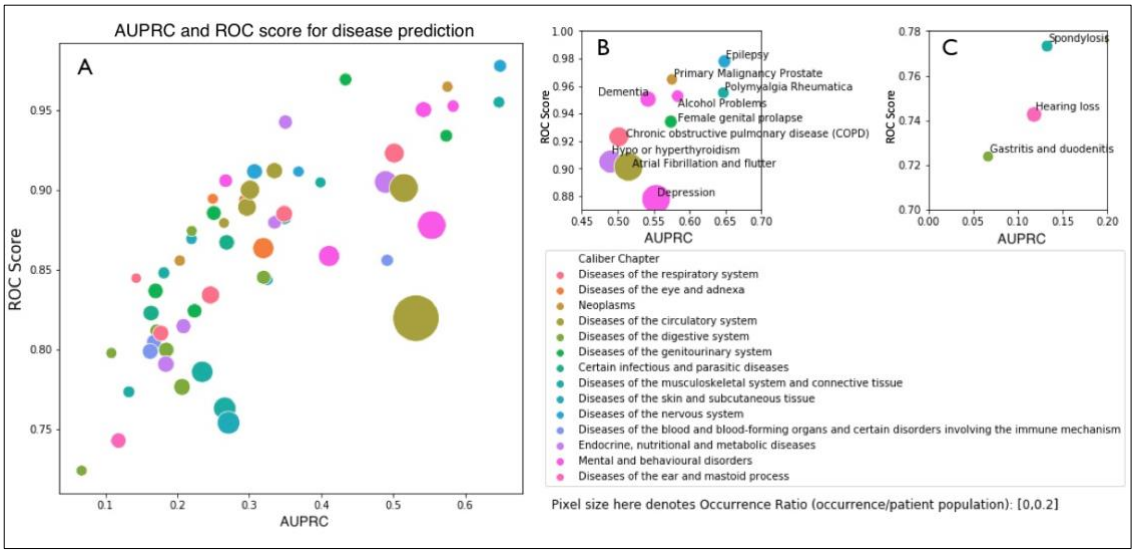
Table 4-2: Predictive performance of deep learning models for subsequent disease prediction

Model	*Next Visit (AUPRC|AUROC)*	*Next 6M (AUPRC|AUROC)*	*Next 12M (AUPRC|AUROC)*
BEHRT	0.462 | 0.954	0.525 | 0.958	0.506 | 0.955
Deepr	0.360 | 0.942	0.393 | 0.943	0.393 | 0.943
RETAIN	0.382 | 0.921	0.417 | 0.927	0.413 | 0.928

Performance in terms of Area under receiver operator characteristic (AUROC) and Area under the precision-recall curve (AUPRC). Table adapted from Li et al[123].

In addition to comparing AUROC and AUPRC on the population level, we also assessed BEHRT performance for each disease. For this analysis, we analyse disease by disease (i.e., for disease *i*, the *i*th entry on a prediction y_p^* is analysed) prediction under the second task (6-month prediction) and compare AUROC and AUPRC across diseases. The analysis is shown in form of a graph in Figure 4-4. We see that BEHRT can make predictions with high AUPRC for diseases such as epilepsy, prostate cancer, and depression. A full summary can be found in Supplementary Table S4.

Figure 4-4: Predictive performance of BEHRT for prediction of individual diseases



Receiver Operator Characteristic (ROC) score and Area under the precision-recall curve (AUPRC) for individual disease prediction for task 2, 6-month prediction. (A) is the complete figure. The right figures (B and C) are subset plots of A. This figure was adapted from Li et al[123].

Furthermore, we investigated the model's ability to capture sex accurately without being given the sex variable explicitly also conducted on the model trained for T2. The gender analyses investigated the predictions for sex-specific diseases; how many times did the BEHRT model predict that a female sex-specific disease happened to a male patient and vice-versa. The analysis (Supplementary Table S5) shows that, in general, BEHRT has correctly identified sex-specific diseases to the correct sex. In gist, those diseases that are female diseases are mostly being predicted in female patients, and in those diseases that are male sex-specific, in male patients. However, for male infertility, we see that some female patients have been predicted with this condition. For this condition, we investigated in our dataset and find that 365 male patients are diagnosed with this condition while 1,734 female patients are diagnosed with condition. Due to the sex-agnostic Read codes ("K26y300", "K26y400", and others) mapped to both male and female infertility with the Caliber phenotyping algorithms. Other than these issues with "male" and "female" infertility, BEHRT demonstrated that the model can implicitly capture sex albeit not explicitly fed the variable as a predictor in modelling.

For the diseases that have occurred as incident diseases in the dataset, we have analysed the results as well (Table 4-3). The results were immaterially different in ranking of models than the analyses of predictive performance across T1, T2, and T3. In absolute terms, the AUROC and AUPRC of the models are much lower than those for subsequent disease prediction across all models.

Table 4-3: Predictive performance of deep learning models for incident disease prediction

Model	*Next Visit (AUPRC |AUROC)*	*Next 6 M (AUPRC |AUROC)*	*Next 12 M (AUPRC |AUROC)*
BEHRT	0.216|0.904	0.228|0.907	0.226|0.905
Deepr	0.095|0.800	0.104|0.814	0.098|0.805
RETAIN	0.108|0.836	0.115|0.845	0.109|0.836

Performance in terms of Area under receiver operator characteristic (AUROC) and Area under the precision-recall curve (AUPRC) for diseases occurring for the first time. Table adapted from Li et al[123].

4.2.4.3 Embedding analyses

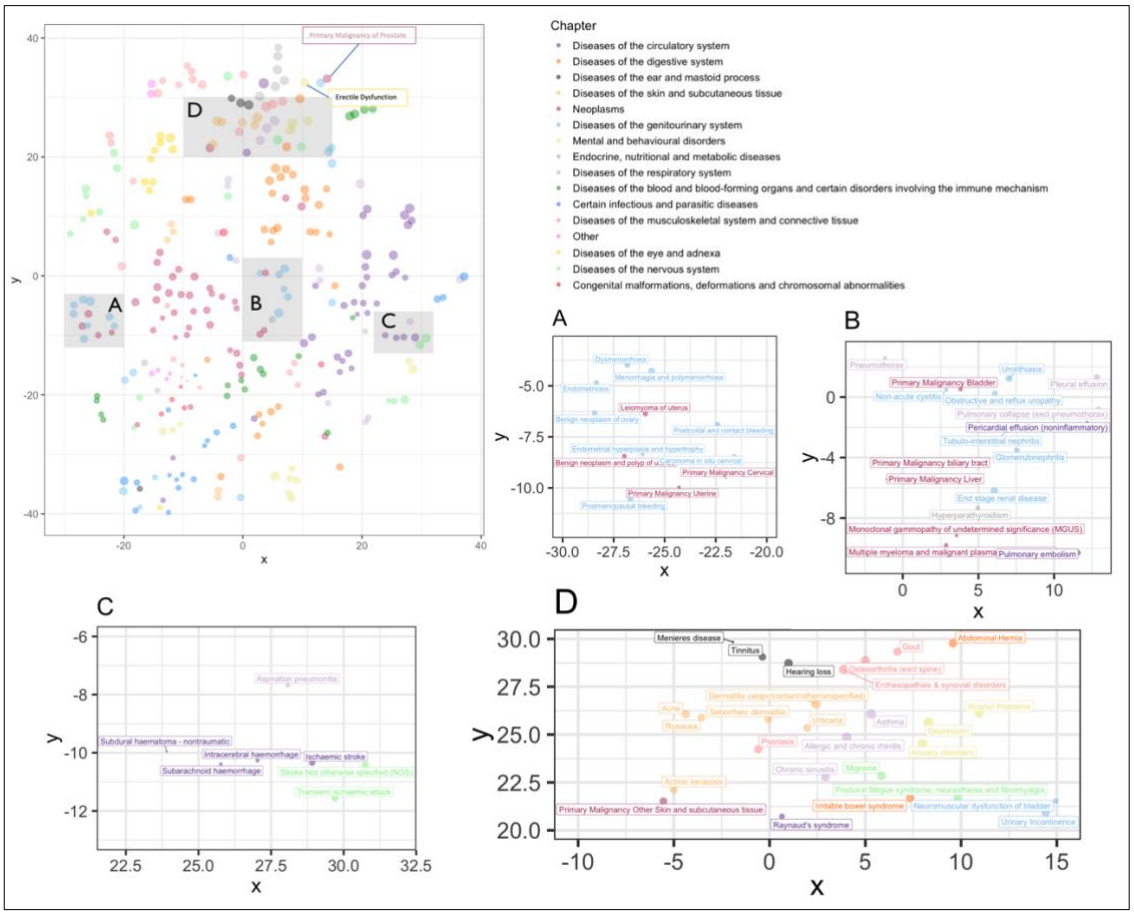
For the embedding analyses, we investigated the embeddings mapped into two-dimensional space (via projection by t-distributed stochastic neighbour embedding) as shown in Figure 4-5. Generally, the clusters naturally formed by the embedding visualisation were in line with clinical knowledge; with that being said however, there are some disease clusters, due to issues in dimensionality reduction are presented visually as more counter-intuitive.

One interesting and reassuring pattern captured by BEHRT is that the sex-specific diseases are stratified quite naturally. Female diseases (e.g., dysmenorrhea, endometriosis, and others) are quite distant from male ones (e.g., prostate cancer, erectile dysfunction, and other diseases). These patterns demonstrate that BEHRT embeddings might find these diseases contextually dissimilar; naturally while some diseases occur with one another, in one patient, male and female diseases cannot technically occur in the same medical history (barring patients who undergo sex-change procedures).

Additionally, the model is able to pick up clusters that are defined by Caliber (denoted by colour in Figure 4-5) in addition to other natural clusters not explicitly annotated by Caliber phenotyping. For example, eye diseases and musculoskeletal diseases are clustered around nervous system disorders even though all these conditions

do not belong to one homogenous Caliber group. Circulatory system diseases are also clustered around one another in addition to select nervous system diseases (e.g., stroke related).

Figure 4-5: Two-dimensional visualisations of condition embeddings



Caliber disease embedding projections in two-dimensional space. A, B, C, D are cut-out graphs of the top-left graph. This figure was adapted from Li et al[123].

Lastly, for more quantitative evaluation, for each disease that occurred in at least 1% of the population, the ten closest diseases were computed using cosine similarity metrics. Comparing these neighbourhoods against those provide by a clinical researcher in the Deep Medicine group, we found a 0.757 overlap. In other words, nearly 76% of the BEHRT-derived clusters were clinically valid and overlapping with clusters derived as a result of clinical knowledge. Furthermore, the clinical researcher notes that the

associations had clear overlap in symptomatology; however, some associations were poor disease associations. The researcher concluded that BEHRT presents a strong ability to understand latent characteristics of diseases without being offered explicit information of underlying pathophysiology and symptomatology.

4.2.5 Interpretation

In sum, in this work, we introduced a novel deep learning model for EHR, BEHRT – a personalised, risk assessment tool which can model health encounters with attributions of age, visit, and position information. With its powerful Transformer-based modular architecture, the model can be used in a variety of settings. Pre-trained on large datasets and fine-tuned on task-specific datasets has been demonstrated as a powerful learning strategy on linked CPRD data. In terms of predictive performance, this model outperforms benchmark convolutional and recurrent neural network modelling counterparts in the prediction of approximately 300 conditions. Additionally, the model can naturally capture attributes such as sex without explicit inclusion. Lastly, in both qualitative and quantitative evaluation, the contextualised embeddings space was found to be clinically meaningful implying that the model is not just powerful, but capable of grasping clinically valid associations latent in data.

In terms of architecture, the BEHRT model is a flexible deep learning feature extractor and prediction model that can assimilate four forms of sequential data found in comprehensive EHR: encounter, age, segment, and position. With these encounters and respective attributes, the model can learn about past diseases, the change in age as these diseases manifest, and the change in the frequency of visits as well. Since the embedding structure is flexible, the four embedding structures can be expanded upon; for example, year can be included and other types of encounters can be explored (E.g., medications).

In this work, the primary goal was to understand limitations in current approaches of modelling multimodal EHR and address them directly with a modelling solution. Furthermore, the goal was to understand the worth of the model in a variety of different ways: predictive performance on both occurrence and incidence of diseases, secondary prediction analyses, and embedding analyses to try and understand the concepts captured by the model. With predictive performance, we have shown that BEHRT has outperformed both benchmark models in both the occurrence and incidence prediction-based tasks. Furthermore, the model not only appropriately predicted sex-based disease, but also illuminated issues in mapping of certain sex-specific diseases (e.g., infertility-based disorders) unknown to us at the onset of the investigations. Lastly, the ablation study illuminated how certain embedding structures were more important than others such as the age and position structures.

In future works, some limitations of this study must be addressed. First, in terms of study design, we only included those patients with at least 5 visits with diagnoses data. This exposes the work to bias as the prediction tasks do not investigate predictive performance in those with fewer than 5 visits of data. However, this was conducted in order to specifically test the proposed model in high-risk settings; while statistical models have appropriately modelled health in lower-risk subgroups, the utility of BEHRT can truly be explored in higher risk subgroups as simpler models would be limited in appropriately capturing risk. Also, in terms of other cohort selection procedures, while all EHR was explored as a data source for this work from 1985 onwards, the HES data only become usable from 1998 onwards. Hence, the data before 1998 would be limited in capturing the patient medical journey and heterogenous data before and after 1998 would bias model training and feature capture. Additionally, the inclusion of data before 2004 might hamper robust evaluation of the models; with Quality and Outcomes Framework

(QOF) introduced in 2004 incentivising healthcare providers to report conditions accurately and in accordance with guidelines, usage of data for modelling after 2004 will be of higher quality than data before[129]. Hence, training one model on data from both before and after this period may be inappropriate. While both sources of bias discussed may be consequential, the intention was to evaluate and rank model performance. Indeed, all models were trained on the same data and tested on the same cut of test data; hence, internally this is a fairly conducted investigation and while the predictive metrics may be quantitatively different if data from 2004 onwards were considered, we expect ranking of model performance to be same as what we presented.

In terms of phenotyping, while Caliber was an incredibly rich phenotype for mapping and outcome ascertainment, other phenotyping methods should be explored for input encounters. In this work, models were trained Caliber phenotypes as opposed to more granular codes such as ICD-10 codes or the Read codes for the GP records. At the time of research, this was an initial exploration of Transformers for EHR data and the desire was to first investigate model performance on disease codes of a modest cardinality. While there are approximately 300 Caliber codes, there would have been many more codes if the model utilised ICD-10 codes leading to questions such as: (1) Can the model handle approximately all 70,000 codes in the ICD-10 coding system or will there be curse of dimensionality issues to handle? (2) If not, then at which level is appropriate (e.g., ICD-10 at 3-character level)? Hence, in this proof-of-concept exploration of Transformers for EHR, the language was limited to 300 Caliber phenotype codes. More granular disease codes (e.g., ICD-10) should be explored in future works.

While medications and procedures are indeed often beneficial for patient health and quality of life, often they can have adverse effect (e.g., iatrogenic risk of some pharmacological therapies). Hence, accounting for these modalities would be important;

however, the current data cut of CPRD offers limited procedures data. Future data supplements may be useful for getting access to procedures. Also, other methods of offering data to model may be useful. To understand cohort effects (i.e., changes across calendar years), perhaps, calendar year can be included as an embedding layer in future investigations. Furthermore, perhaps the source of data (i.e., GP or hospital) in the diagnoses (and possibly, medications) encounter input would be another orthogonal source of knowledge that would be useful for the model.

4.3 Deep learning modelling for electronic health records: incident heart failure prediction

4.3.1 Introduction

HF remains a major cause of global mortality and economic burden. Despite recent evidence suggesting improvements in the quality of care of HF patients and improving prognosis trends, HF incidence has remained relatively constant with little reduction[19]. Due to population ageing and growth in the past few decades, indeed the absolute incidence of HF has actually been increasing. Hence, there is a great need for better HF preventative strategies and deeper investigations into risk factors of the complex condition[19]. While several statistical approaches have been developed to predict incident HF, the models have been quite unsatisfactory in predictive performance.

4.3.2 Aims

In this work, the aim was to amend and implement a state-of-the-art sequential deep learning model, BEHRT, to predict incident HF using temporal multimodal EHR. The model was compared against state-of-the-art deep learning models, RETAIN-EX and

Deepr. Furthermore, a secondary ablation analysis was undertaken to better understand the predictive performance of BEHRT.

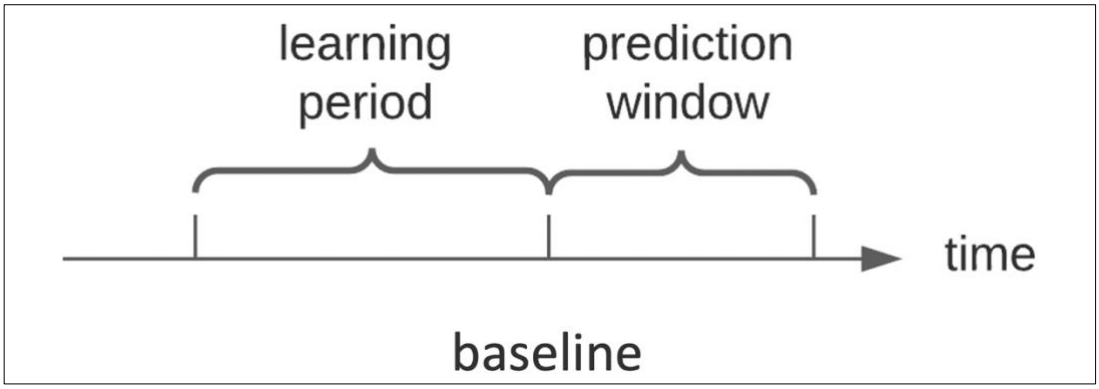
4.3.3 Methods

4.3.3.1 Data

In this study, we used the described cut of CPRD (see section 3.6)[5]. Initially with access to data from 8 million patients, in this study, we only considered patients from GP that are eligible for linkage with HES in order to only consider patients with documented comprehensive medical health history. Additionally, to only keep those patients that have enough records for prediction, we kept those patients with at least 5 visits in the medical history.

Diagnoses and medication encounters in medical history were extracted. Diagnoses codes from primary and secondary care were mapped to the Caliber phenotype codes[122]. The medication codes were encoded using the British National Formulary (BNF) hierarchical format[130]. The medication data in CPRD only indicate prescription as opposed to the retrieval or dispensation of the actual medications. While many forms of coding can be used for modelling, the codes at the BNF section level were used. In total, 299 diagnostic codes and 426 medications as well as patient age in months and calendar year were extracted for modelling. Specifically, only the data for medication in BNF code format was utilised; dosage and number of days of treatment data values were not utilised for modelling.

Figure 4-6: Incident heart failure prediction task



For a hypothetical patient, study entry is start of patient medical history. And baseline marks beginning of "follow-up". This figure was adapted from Rao et al[131].

In this study, the focus was the prediction of incident heart failure. Figure 4-6 describes the task: for each patient, all patient medical history before baseline was used for model training. The outcome was ascertained in a six-month window following baseline. The incidence of HF was defined as the first recorded HF diagnoses code (adopted from Caliber[122]) in EHR for each patient. The phenotype specifically for HF was "heart failure" as defined by Caliber. Only diagnoses codes were used to capture the incidence; historical diagnoses codes were not used. For those with at least one diagnosis of HF, the baseline was defined as a random timestamp within 6 months before the incidence of HF, and for those without HF in their medical history, the baseline was a randomly selected time stamp. Patient medical history considered for modelling started at the date of the GP registration for each patient. This cohort is henceforth referred to as the HF cohort.

The follow-up window of six months was chosen in order to simply test model prediction abilities on an important clinical outcome. As opposed to previous proof-of-concept prediction works (e.g., 12-month multi-condition prediction in section 4.2.3.4), this work was intended to test the predictive power of BEHRT on a clinically significant

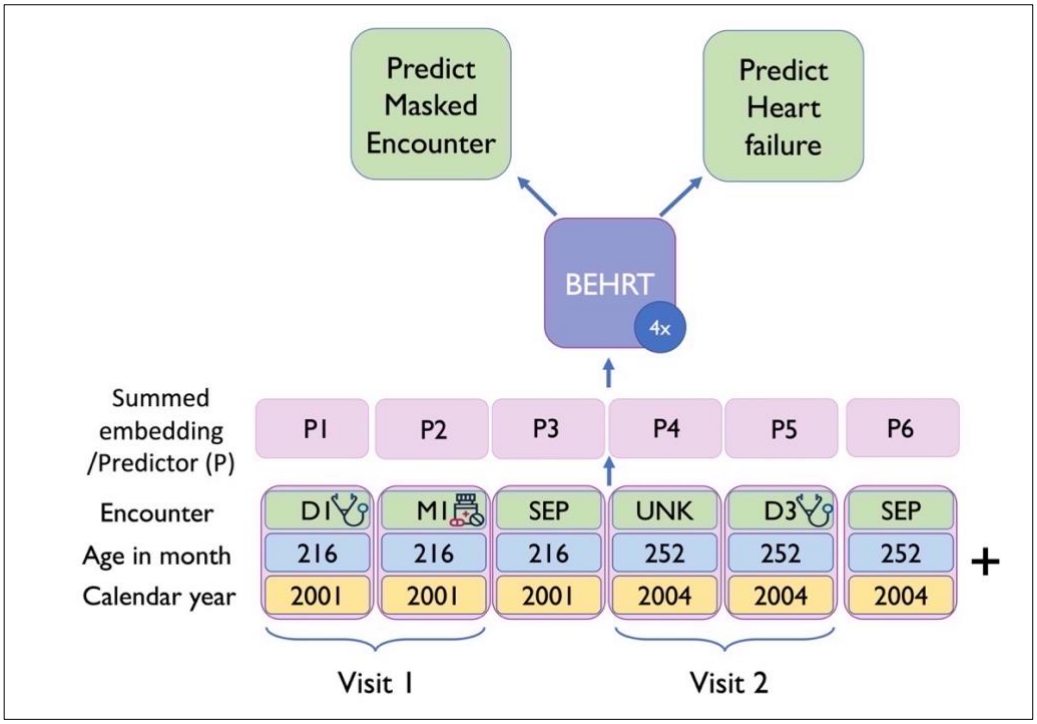
outcome: HF, a major cardiovascular outcome. Furthermore, the purpose was to train a model that can be ultimately probed in explainability studies (Chapter 5). Hence, the modest six-month window was chosen.

4.3.3.2 Model development

BEHRT as previously introduced was amendment to address the incident heart failure prediction task. The first amendment was inclusion of calendar year embedding structure in addition the embeddings of encounter, and age (Figure 4-7). The second amendment to the modelling was that both diagnoses and medications were used. Medication codes were incorporated into the BEHRT model similarly as the disease encounters; when a medication record was found for a particular visit, the respective BNF code was included for modelling as an encounter with appropriate age and year attributes. In terms of handling repeat records, all diagnoses and prescriptions were included for modelling and repeats were not excluded.

The model was pre-trained using the Masked Language Modelling derivation as previously introduced (see Pre-training). Instead of predicting multiple diseases at once as done in previous experiments, the model predicted a probability. Specifically, the latent space after pooling layer in the BEHRT architecture was mapped to (*1x1)* real space and transformed using sigmoid activation function to yield a probability as opposed to a real number.

Figure 4-7: BEHRT model for incident HF prediction task



The figure shows the embeddings used, and the two tasks investigated in this work: unsupervised training vias masked encounter modelling and incident heart failure prediction. This figure was adapted from Rao et al[131].

Furthermore, the proposed model was compared against two state-of-the-art models for EHR, Deepr, the previously introduced convolutional neural network, and the model, RETAIN-EX, an expansion of the RETAIN model.

For hyperparameters, we applied Bayesian optimisation for deriving optimal hyperparameters: number of attention heads, number of layers, intermediate size, and hidden size – all core components of the BEHRT architecture. 20 iterations of searching the hyperparameter space yielded optimal number of layers: 4, hidden size: 120, attention heads: 6, and intermediate size: 108. All models were coded, trained, and evaluated on Pytorch.

4.3.3.3 Evaluation

For implementation of the models and evaluation, we first cut the extracted HF cohort into random, non-overlapping partitions consisting of 60%, 20%, and 20% for training, tuning, and evaluation cohorts respectively. The training and the tuning cohorts were used for the hyper-parameter tuning, while the evaluation cohort was used for conducting downstream analyses (see 5.2). For predictive performance analyses, k-fold cross validation (k=5) was applied on the HF cohort, and AUROC and AUPRC were reported with confidence intervals derived over the folds.

Furthermore, an ablation study investigating the utility of the various modalities was conducted. Specifically, we investigated the utility of: diagnoses (D), medications (M), age (A), and year (Y) by alternatively including them in the modelling structure. The six experiments investigated the following various combinations of modalities: D, DA, DAY, DM, DMA, and DMAY (full model). Model performance for each of the 6 experiments was assessed with both AUROC and AUPRC and corresponding 95% confidence intervals (derived via the aforementioned 5-fold cross validation paradigm).

4.3.4 Results

4.3.4.1 Population characteristics

Of 100,071 patients for incident HF prediction (HF cohort), 13,050 patients got incident HF in the follow-up period. 58.3% were women, the median age in years at baseline was 70; 1st and 3rd quartile: (59, 79), 65.7% had history of hypertension, 9.3% a prior myocardial infarction, and 5.1% an ischaemic stroke (Table 4-4). Furthermore, there was low prevalence of those with rheumatoid arthritis, but a larger percentage of individuals had atrial fibrillation and diabetes.

Table 4-4: Population characteristics for the heart failure cohort

Number of incident cases of heart failure (%)	*13050 (13.1)*
Characteristics	
Women (%)	58331 (58.3)
Men (%)	41740 (41.7)
Median follow-up duration (year)	9
Median age (year); Interquartile Range	70; (59,79)
Diabetes Mellitus (%)	20606 (20.6)
Hypertension (%)	65760 (65.7)
Rheumatoid arthritis (%)	3288 (3.3)
Atrial fibrillation and flutter (%)	26257 (26.2)
Myocardial infarction (%)	9278 (9.3)
Chronic obstructive pulmonary disease (COPD) (%)	13897 (13.9)
Ischaemic stroke (%)	5124 (5.1)

Table presents the population characteristics for the heart failure (HF) cohort. This figure was adapted from Rao et al[131].

4.3.4.2 Model evaluation

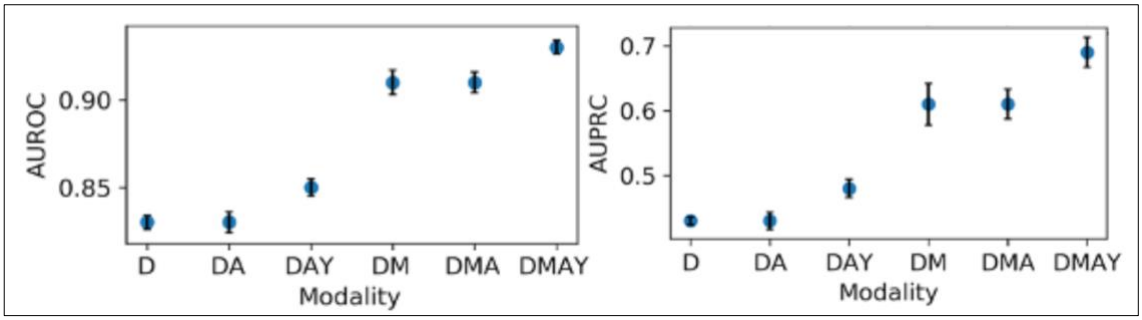
Table 4-5: Predictive performance of deep learning models for incident heart failure prediction

Model Name	*AUROC (95% CI)*	*AUPRC (95% CI)*
BEHRT	**0.93 (0.926, 0.934)**	**0.69 (0.667, 0.713)**
RETAIN-EX	0.90 (0.893, 0.901)	0.62 (0.596, 0.636)
Deepr	0.91 (0.901, 0.913)	0.61 (0.577, 0.633)

Performance in terms of Area under receiver operator characteristic (AUROC) and Area under the precision-recall curve (AUPRC) for incident heart failure prediction. 95% confidence interval (CI) is also given. This table was adapted from Rao et al[131].

The BEHRT model with four modalities: DMAY, achieved best predictive performance in terms of AUROC and AUPRC. Specifically, with respect to the benchmark models, the BEHRT model demonstrated absolute improvement of 2% and 7% in terms of AUROC and AUPRC respectively.

Figure 4-8: BEHRT model ablation study



The figure shows the ablation study conducted on the BEHRT model for incident heart failure prediction. Various modalities, D: Diagnosis, M: Medications, A: Age, Y: Year were alternatively given and removed from modelling and Area under receiver operator characteristic (AUROC) and Area under the precision-recall curve (AUPRC) were calculated. The figure was adapted from Rao et al[131].

Ablation study of the various modalities (DMAY) for BEHRT modelling illuminated two notable results. Medication data points were indeed important for predictive performance as seen in the greatest leap in predictive performance between D and DM in terms of both AUROC and AUPRC. Also, we saw that calendar year was found more useful by the model than inclusion of age denoting that this absolute capture of time was more important than the relative measure of time, age.

4.3.5 Interpretation

We expanded the BEHRT model for a more clinically relevant task: the prediction of incident heart failure. When compared to the known EHR deep learning models of Deepr and RETAIN-EX, the model demonstrated superior prediction performance across metrics. Furthermore, the ablation study showed that inclusion of medications was found to be important for prediction. Solely diagnoses were insufficient for high predictive performance. Furthermore, calendar year was found to be more useful for prediction than the relative concept of time, age.

We included more predictors of health than previously demonstrated in literature. The inclusion of age for risk prediction is well known; however, the utilisation of calendar year is not frequently done. Our ablation study found that calendar year is a powerful element of input for predictive performance. One potential explanation is that the temporal variability caused by changes in medical practice such as changes in disease pattern, policy, and availability/use of treatments can be tracked by calendar year. Hence, calendar year stands as an expressive proxy for more latent changes in medicine not exclusively captured by disease/medication/age modalities. BEHRT's flexible architecture allows for inclusion of these variables.

In terms of strengths, our study presented a comparison of three varieties of deep learning models: recurrent, convolutional, and Transformer neural networks. The proposed modelling approach achieved the highest AUPRC and AUROC as opposed to benchmark comparisons. Furthermore, the ablation study conducted offers some model transparency; specifically, the ablation of various modalities clarifies which input modalities are indeed important for prediction. Additionally, the Caliber phenotyping algorithm, utilising expert-driven phenotyping of approximately 300 conditions, allows nuanced identification of complex diseases. Our study also has some limitations. First, the cohort selection procedure only investigated patients with higher interactions with clinical services – i.e., those with sufficient number of records. This could potentially compromise model generalizability for prediction in low-risk groups who have fewer clinical encounters. Additionally, the BEHRT model for incident HF risk prediction needs to be evaluated on other datasets for a full examination of its external validity and generalisability.

5 EXPLAINABILITY

In this chapter, the research that address the second objective (see section 1.2) will be presented. After establishing that the proposed deep learning architecture learns useful representations for prediction, second, the following section attempts to answer the question: how can deep learning modelling be trusted?

To this end, I develop tools to better understand the decision-making processes of the "black-box" deep learning modelling. Continuing the study of risk prediction for incident heart failure, I first develop metrics to better understand the importance of temporal modalities of age and calendar year. Second, I develop tools to extract input variables and encounters that are important for the outcome prediction. With the developed tools, I proceed to discuss a pipeline of analyses to better trust deep learning modelling.

The contents of this chapter are published as a manuscript, *"An Explainable Transformer-Based Deep Learning Model for the Prediction of Incident Heart Failure"* in *Journal of Biomedical and Health Informatics (JBHI)* (doi.org/10.1109/JBHI.2022.3148820). The publication is a product of work by multiple authors. As a co-first author, my role consisted in designing the study, conducting literature review, processing data, conducting statistical/deep learning analyses, and writing the first draft of the manuscript. Material from the publication (including figures, tables, and text) have been amended for presentation in the following sections.

5.1 Introduction

Heart failure (HF) is a complex condition that remains a major cause of death around the world. The complexity is due to the multifactorial nature of the disease as described in the introduction section relevant to HF (see section 2.1.1). More importantly, given limited knowledge about risk factors of HF (e.g., diabetes, atrial fibrillation, and hypertension), further investigation into factors of risk and protection is needed to better inform preventative care.

The current understanding of the aetiology of HF comes from medical knowledge and evidence from randomised and observational investigations. Statistical models have been used to understand association of the various factors to the risk of HF. Through modelling, 27 clinical factors have been verified to be associated with incident HF across 15 studies[132]. Since the predictive performance is lacklustre, this implies of course, that many factors of risk are being omitted from modelling; alternatively, with more comprehensive capture of risk and protection at baseline, models can more accurately predict incident HF[60].

However, with data-driven modelling of the condition of HF, the potential for using models agnostically capturing those variables which are predictive of HF remains unexplored. On the other hand, unlike statistical models, which model a host of variables and allow interpretation of the contribution of these variables to the outcome prediction, the deep learning models are more black-box, and lack "explainability".

Given the BEHRT model's striking performance on a variety of predictive tasks as compared to benchmark RNN and CNN models (Chapter 4), the investigation of capturing data-driven associations to better understand incident HF should be optimally conducted on the BEHRT model. However, for Transformer-based models, that too for

EHR-based studies, there are a dearth of readily available methods for interpreting these complex multi-level attention-based networks.

5.2 Aims

In this work, on the task of incident HF prediction, I aim to explain the BEHRT predictive process in two ways:

1. I investigate the temporal modalities and their contribution to prediction of HF.
2. I develop a surrogate model – an auxiliary, simpler model – that uses parametrized noise to derive the contributions of each of the encounters in medical history (with contextualised age/year annotations) to the prediction of incident HF.

5.3 Methods

5.3.1 Data

In this work, the focus was explaining the predictions for the task of incident HF prediction. As introduced earlier, Figure 4-6 describes the task and the cohort, in which we conducted the analyses, was the HF cohort as described in the section, 4.3.3.1. Furthermore, non-overlapping partitions of 60%, 20%, and 20% of the HF cohort were extracted for training, tuning, and the explainability analyses.

5.3.2 Explainability investigations

Two forms of explainability were conducted in this study: embedding based analyses and perturbation theory driven analyses.

5.3.2.1 Temporal variability analyses

The first was the explanation of temporal modalities of specifically age and calendar year embedding structures. Considering passage of time in the relative sense

(age) and absolute sense (calendar year) should be theoretically useful for modelling ageing as well as changes in medical practice, disease patterns, and other properties of medicine that change over time, we expect that modelling both in the BEHRT model would be informative for prediction of incident HF.

To test the utility of both the embeddings, four candidate diseases were chosen. The diseases were prevalent in the HF dataset to ensure sufficient training of their respective embeddings. For each disease, the trained disease embedding was added to each and every trained age embedding (embeddings representing age 192 to 1200 in months). These summed embeddings represent the disease at various ages between 192 and 1200 months of age. In this work, cosine distance was applied; cosine distance is a distance metric that measures the similarity between representations of diseases at different ages. The greater the difference in the age-contextualised disease representation implies greater temporal variability. Similarly, summation and cosine distance were computed for year embeddings (embeddings representing years 1988 to 2014). The dissimilarity across the age spectrum was compared against the same across the year spectrum.

5.3.2.2 Contribution analyses

To further understand the BEHRT model's decision-making processes, a method was derived and implemented in the incident HF setting to derive the contribution of an encounter to final outcome prediction. We extended a perturbation-based technique created for Transformer-based language modelling[133]. Utilising the summed embedding as a predictor (encounter/age/year) to represent the encounter contextualised with age and calendar year data, the fundamental concept was to quantify change between the predicted probability given input of the predictor embedding perturbed by parametrised noise and the predicted probability given input of the untouched predictor embedding.

Algorithm 1 Perturbation algorithm.

***M*: Model, *N*: Number of encounters, *X*: input encounter embeddings $X \in \mathcal{R}^{N \times K}$, $\tilde{X}$: perturbed input embeddings, *Y*: indicator of HF patient, *L*: loss function, *S*: contribution score.**

PERTURBATION (X,Y)

Initialise $\epsilon \leftarrow [\epsilon_1, \epsilon_2, \dots, \epsilon_N]$ and $\epsilon_{1_i} \sim \mathcal{N}(0, \sigma_i^2 I)$ (5-1)

While (not converged)

$\tilde{X} \leftarrow X + \epsilon$

$O \leftarrow M(X), \tilde{O} \leftarrow M(\tilde{X})$

$Loss \leftarrow L(O, \tilde{O}, Y, \tilde{X})$

$update\ \epsilon$

$S \leftarrow transform(\epsilon)$

Return S

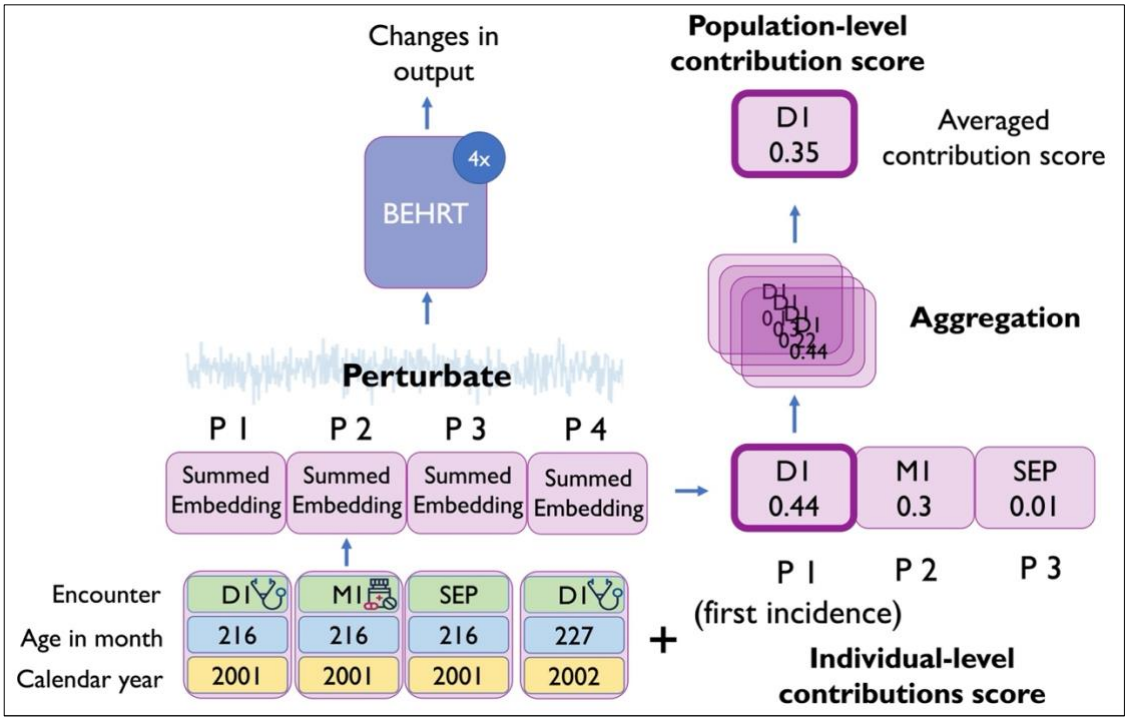
As a facilitating example, analysing a calcium tablet encounter (prescription/medication) at a given age/calendar year, if large perturbations of this calcium tablet predictor contribute to minimal change in output probability, that means that this predictor is unimportant for the output prediction. If even small perturbations of the same predictor amount to a great degree of change in predicted probability, then calcium tablets are indeed an important predictor of incident HF. In this work, an asymmetric loss function was proposed to prioritise encounters that maximally capture HF/non-HF predictions with accuracy. HF/non-HF represent those who have HF, and no HF respectively as a label. Equation(s) (5-1) presents the algorithm for the perturbation-based contribution extraction.

The perturbation modelling is a local surrogate model that can quantify contribution of the encounters at an individual, patient-by-patient basis. By aggregation of the individual-level encounters (i.e., diagnosis/medication) contribution over the entire population considered for analysis (Figure 5-1), we can understand the encounter

contribution at a global level (i.e., population level). This work only utilised the contribution of the first incidence of the disease/medication in medical history; contribution of the repeat encounters was not considered.

Further description of the loss function and the optimisation of the perturbation surrogate model is presented in supplementary section 9.2.

Figure 5-1: Contribution analyses utilising perturbation surrogate model



The pipeline for population-level contribution analyses. Trainable perturbation-based noise is used to understand contribution of a particular encounter to the outcome prediction. On the right, individual-level contribution scores (for example, D1) is aggregated and the final mean contribution score is presented (D1 – 0.35). The figure was adapted from Rao et al[131].

The analysis was developed to understand the association between a particular encounter in medical history and the outcome, HF. For this reason, the relative contribution (RC) metric was created (with associated 95% confidence interval (CI)). This metric is calculated by dividing the average contribution of the encounter in HF patients by the same in non-HF patients. In this way, RC>1.0 and <1.0 means that the

encounter is associating more towards the HF outcome and non-HF outcome respectively. In other words, there is a positive and negative association with HF respectively.

The analysis was conducted on the patient with confident prediction (predicted probability larger than 0.8 and less than 0.2) and focused on the encounters in medical history that had sufficient prevalence in order to ensure that associated embeddings were sufficiently trained for downstream analyses.

The pipeline to trust the model utilising perturbation analyses was developed and organised as follows:

1. For medically validated risk factors, investigate RC in medical history
2. Examine if the results from analysis are consistent with medical understanding
3. Identify novel factors of risk and protection captured by BEHRT in prediction of incident HF

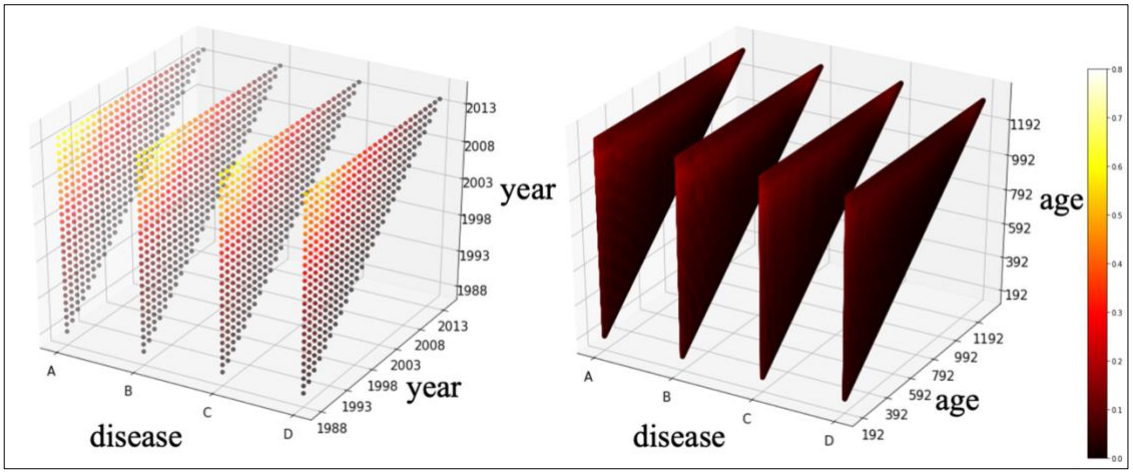
Specifically, for all three parts above, the differential contribution to HF by age and calendar year were investigated. Age stratified relative contribution analyses were conducted for age groups: (50-60], (60-65], (65-70], (70-75], (75-80], and calendar year stratified analyses were conducted for calendar year groups: [1990-1995], (1995-2000], (2000-2005], (2005-2010] when clinical events were first recorded. the "(" and "]/[" symbols represent exclusion and inclusion, respectively for age/year bands.

5.4 Results

5.4.1 Analysis of temporal variability

The cosine similarity matrices for both age and year embeddings summed with the four representative diseases are shown in Figure 5-2. The diseases are depression, peripheral arterial disease, anxiety disorders, and hypo or hyperthyroidism.

Figure 5-2: Temporal embeddings analysis

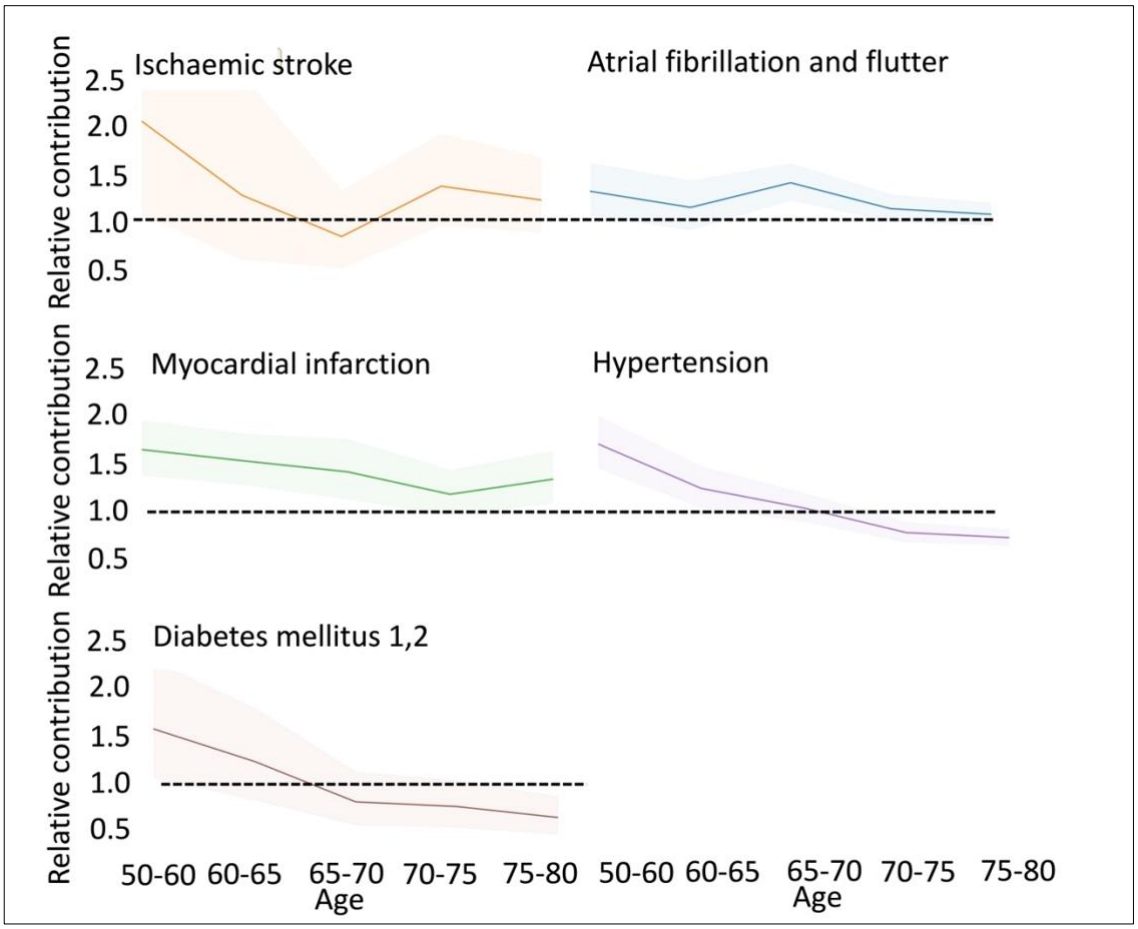


The age and year embedding analyses with cosine similarity measure. The cosine similarity is shown for year (left) and age (right) groups A: depression, B: peripheral arterial disease, C: anxiety disorders, D: hypo or hyperthyroidism. Age runs from 16-100 years in months and year, 1988-2014 in years. Lighter colours inherently mean greater dissimilarity and higher, lesser. The figure was adapted from Rao et al[131].

Calendar year (left) showed greater dissimilarity across the pairwise comparisons than age (as seen in the lighter colours prevalent in the year embedding figure as opposed to the age one). Disease representation of any of the four diseases were more sensitive to changes in year than changes in age. This suggests that calendar year as an embedding supplement to the encounter offers more signal to the raw encounter embedding than the age embedding; in other words, any of the four diseases at age 192 are immaterially different in terms of embedding values from the same at age 1192 or any other age between 192 and 1192.

5.4.2 Contribution analyses

Figure 5-3: Contribution analyses for validated risk factors



The age-stratified relative contribution (RC) analyses for established risk factors. X and y axes represent the age groups and relative contribution is presented as (mean; 95% confidence interval). The black dotted line denotes 1.0 RC. The figure was adapted from Rao et al[131].

As the pipeline presented, we first examine if BEHRT can capture established risk factors of HF with the proposed RC metric[134]. We derived the RC for known risk factors: hypertension, myocardial infarction, diabetes, ischaemic stroke, atrial fibrillation and flutter. For these risk factors the average RC was greater than 1.0 in the general and age-stratified analyses (Table 5-1 and Figure 5-3). In general, it is seen that the association is stronger in younger ages and diminishes closer to the line of parity across all known risk factors as age increases. These results were indeed consistent with evidence from past works[135,136].

Table 5-1: Relative contribution of validated risk factors of heart failure

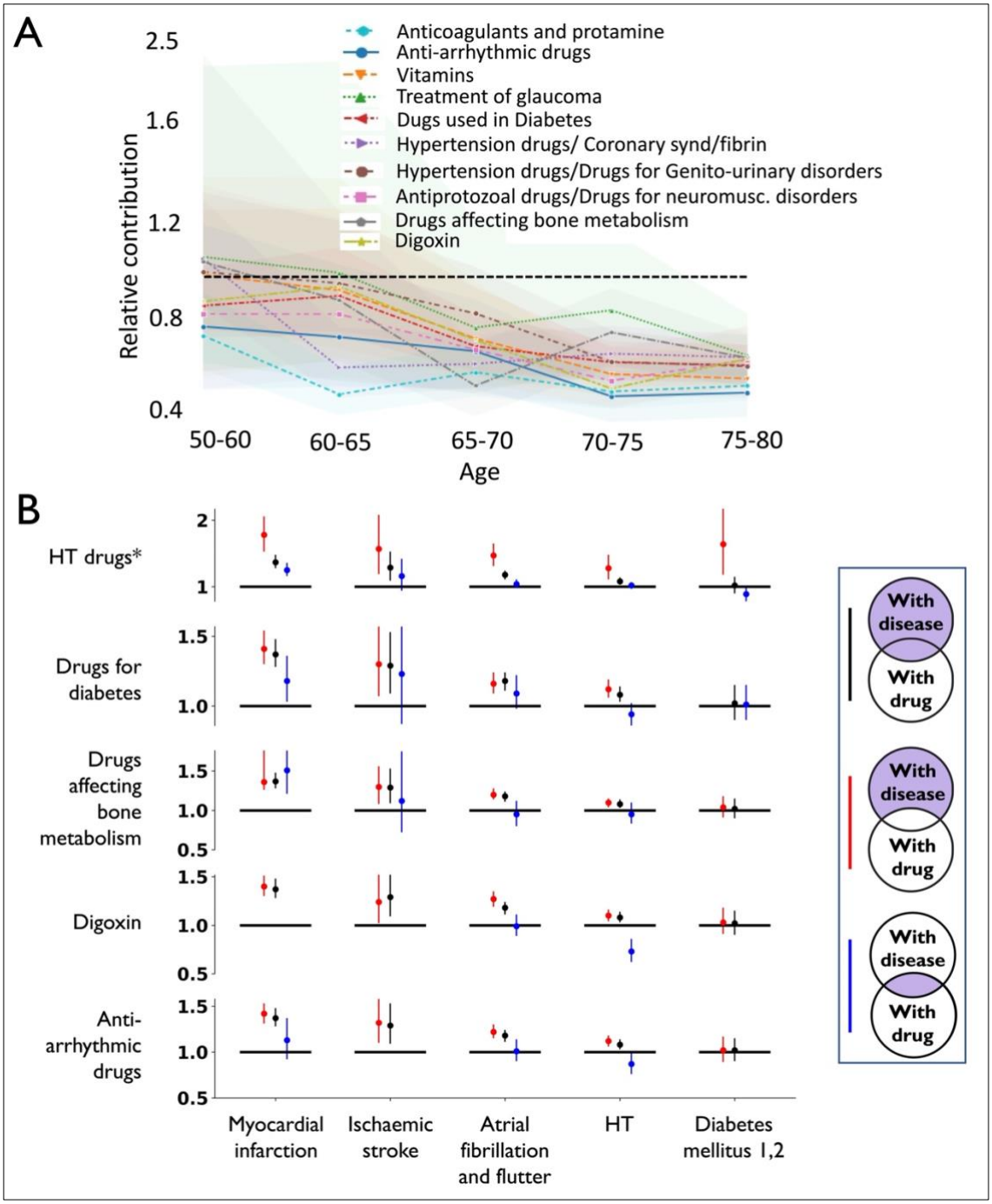
Disease	*Overall*	*50-60*	*60-65*	*65-70*	*70-75*	*75-80*
Myocardial infarction	1.37; (1.28, 1.48)	1.55; (1.33, 1.8)	1.45; (1.25, 1.68)	1.36; (1.13, 1.63)	1.16; (0.99, 1.37)	1.29; (1.09, 1.53)
Ischaemic stroke	1.29; (1.09, 1.53)	1.86; (1.05, 3.3)	1.21; (0.64, 2.27)	0.84; (0.57, 1.24)	1.29; (0.95, 1.75)	1.17; (0.88, 1.54)
Hypertension	1.18; (1.11, 1.24)	1.59; (1.39, 1.83)	1.21; (1.05, 1.4)	1.04; (0.93, 1.17)	0.83; (0.75, 0.92)	0.79; (0.73, 0.86)
Diabetes mellitus 1,2	1.08; (1.03, 1.14)	1.48; (1.05, 2.07)	1.19; (0.86, 1.65)	0.84; (0.64, 1.1)	0.8; (0.62, 1.03)	0.7; (0.56, 0.89)
Atrial fibrillation and flutter	1.02; (0.9, 1.15)	1.23; (1.02, 1.47)	1.08; (0.89, 1.32)	1.3; (1.15, 1.47)	1.07; (0.96, 1.2)	1.02; (0.93, 1.12)

Table presents the general relative contribution and associated 95% confidence interval for overall analyses and age-stratified analyses. This figure was adapted from Rao et al[131].

Interestingly, for hypertension and diabetes specifically, the RC trended slightly below the line of parity; we hypothesized that this occurrence might be due to contextualisation with treatments of the two diseases. Often patients with those diseases are indeed put on medications to directly address the issues – antihypertensives for hypertension and antidiabetic drugs for diabetes, and these drugs mitigate risk of incident HF. Hence, while medications indeed are negatively associated to incident HF, by contextualisation, these diseases, for which treatments are given, are also negatively associated due to consistent contextualisation across patients.

To properly test this hypothesis, we first investigated disease prevalence: 73% of patients older than 65 years of age with hypertension are treated with antihypertensives. And for diabetes, 70% of those with diabetes are treated with medication for diabetes. The contextualisation of the disease/treatment pair is indeed quite consistent across patients; i.e., often a patient with this disease will be treated, so to the model, the two might be interchangeable or indistinguishable.

Figure 5-4: Contribution analyses for medications and contextualisation analyses

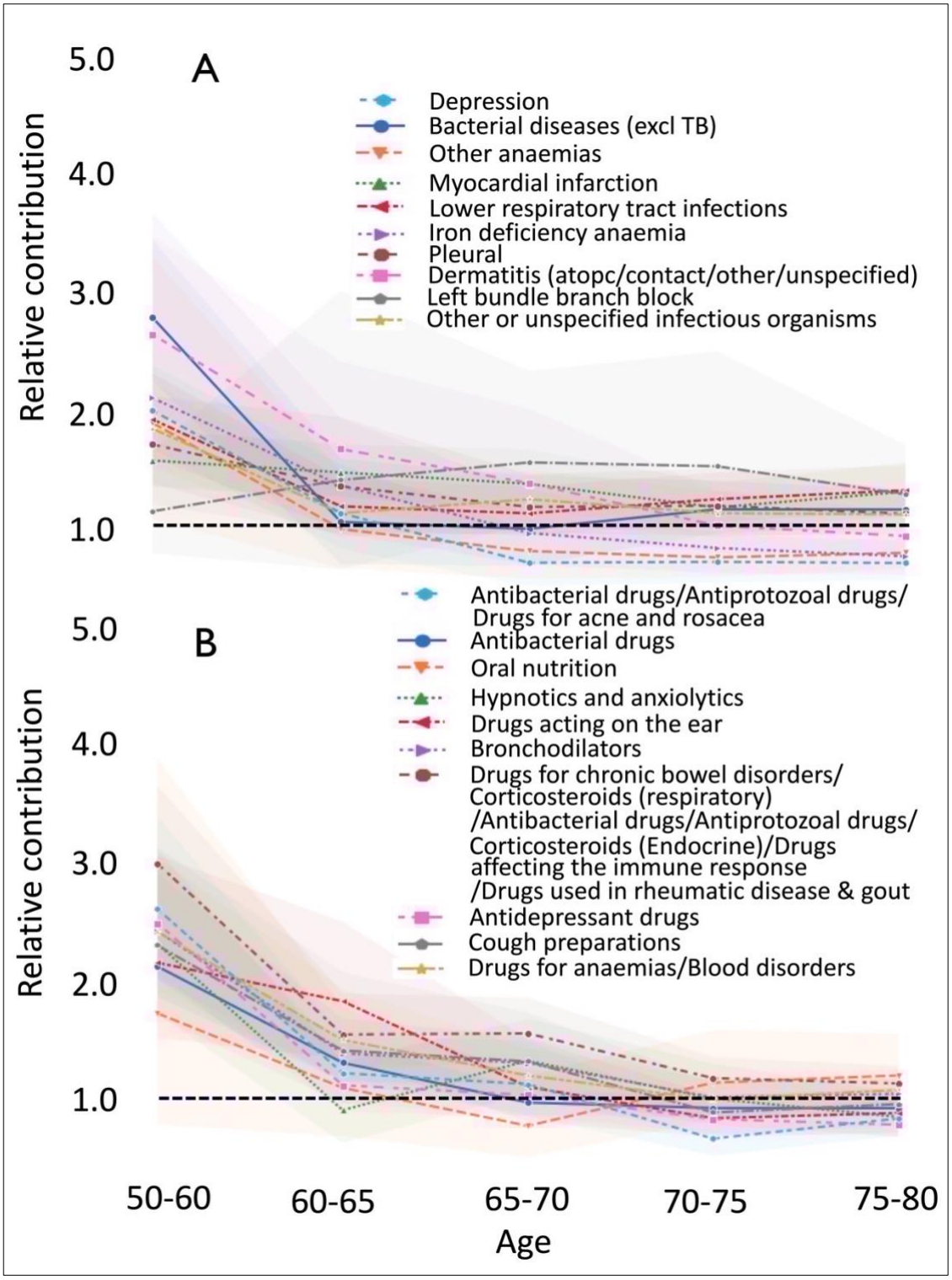


The lowest age-stratified relative contribution (RC) for medications (A) and the contextualisation of medication/risk factors (B). Specifically, in (A), x and y axis represent the age groups RC and the black dotted line is the line of parity for RC (1.0) (B), shows the RC of the risk factor in patients with variety of different stratifications. Black forest plot: general population of people with disease (column); red forest plot: those with disease (column) not treated with medication (row); blue forest plot: those with disease (column) and treated with medication (row). Some of the forest plots had insufficient sample size (e.g., drugs for diabetes and the risk factor, diabetes). The figure was adapted from Rao et al[131].

Given that these diseases frequently contextualise with paired treatments (especially in older age groups), we investigated if treatments associate with non-HF or RC<1. Indeed, we captured that antihypertensives, digoxin, and drugs for diabetes and other established treatments of validated risk factors were associated with non-HF (Figure 5-4). This confirms that the model was able to capture elements known to mitigate risk of HF by directly counteracting elements that exacerbate risk of HF (risk factors)[15,137].

In order to better understand the relationship between risk and treatment of risk, we conducted an analysis of disease encounter RC stratified by treatment status. As a motivating example, to understand the association between treated/untreated hypertension and incident HF, we computed the RC for hypertension in the subgroups of patients with hypertension and the desired antihypertensive treatment status (all patients with hypertension, those treated with antihypertensive, those not treated with antihypertensives). In Figure 5-4, with respect to all with hypertension generally (black lines), in treated patients (blue lines), there is an attenuation in the RC for most diseases (16 of 19 RC computations on disease encounters). While RCs of several risk factors were lower in the treated as opposed to the untreated stratifications, the RCs were most clearly attenuated in those groups treated with antihypertensives, digoxin, and medications for diabetes. In some of these cases, we note that the RC was not calculated because of insufficient sample size for the computations. With these analyses into various disease/medication pairs, it is made transparent that BEHRT can naturally capture the difference between untreated and treated risk. BEHRT can understand that untreated risk heightens association to HF while treatment mitigates it.

Figure 5-5: Contribution analyses for model derived risk factors



The age-stratified relative contribution (RC) for (A) top-10 diseases and (B) medications. X and y axis represent the age groups RC and the black dotted line is the line of parity for RC (1.0). 95% confidence intervals are shaded. The figure was adapted from Rao et al[131].

Continuing to the third step of the pipeline, after validating the risk and protective factors known in medical literature, we investigated other novel factors gleaned from the RC analyses conducted on disease/medication encounters. We found many diseases like myocardial infarction, pleural effusion, and lower respiratory tract infection and medications such as bronchodilators, corticosteroids, and acne drugs were all positively associated with HF (Figure 5-5) in age-stratified analyses of the ten highest RC diseases and medications. As age increased, these encounters most strongly associated with HF showed a trend similar to that of the age-stratified analysis of validated risk factors as well – i.e., limited discriminatory independent contribution to incident HF. Lastly, for some of the predictors such as left bundle branch block, the confidence intervals were too wide due to limited sample size and heterogeneity in contribution across the population in individual age-bands to allow for any conclusions about these predictors' differential RC contribution by age.

Similar to the analyses conducted on paired disease/treatment, the encounters shown in section B of Figure 5-5 were directly treatment of paired conditions in section A of the same. The model again is able to capture that contextually, some diseases and medications appear concurrently in medical history and might be causally associated. For example, dermatitis is treated with corticosteroids, which is linked to increased cardiovascular risk, as may be depression, and its paired medication, antidepressant[24,138]. Additionally, lower respiratory tract infections, asthma, and chronic obstructive lung disease are treated and hence paired contextually with cough preparations, bronchodilators, and antibacterial drugs[139,140]. These medications are associated with HF often because of misattribution of HF symptoms for respiratory conditions[140]. On the other hand, it may be because some of the aforementioned medications are in part, at

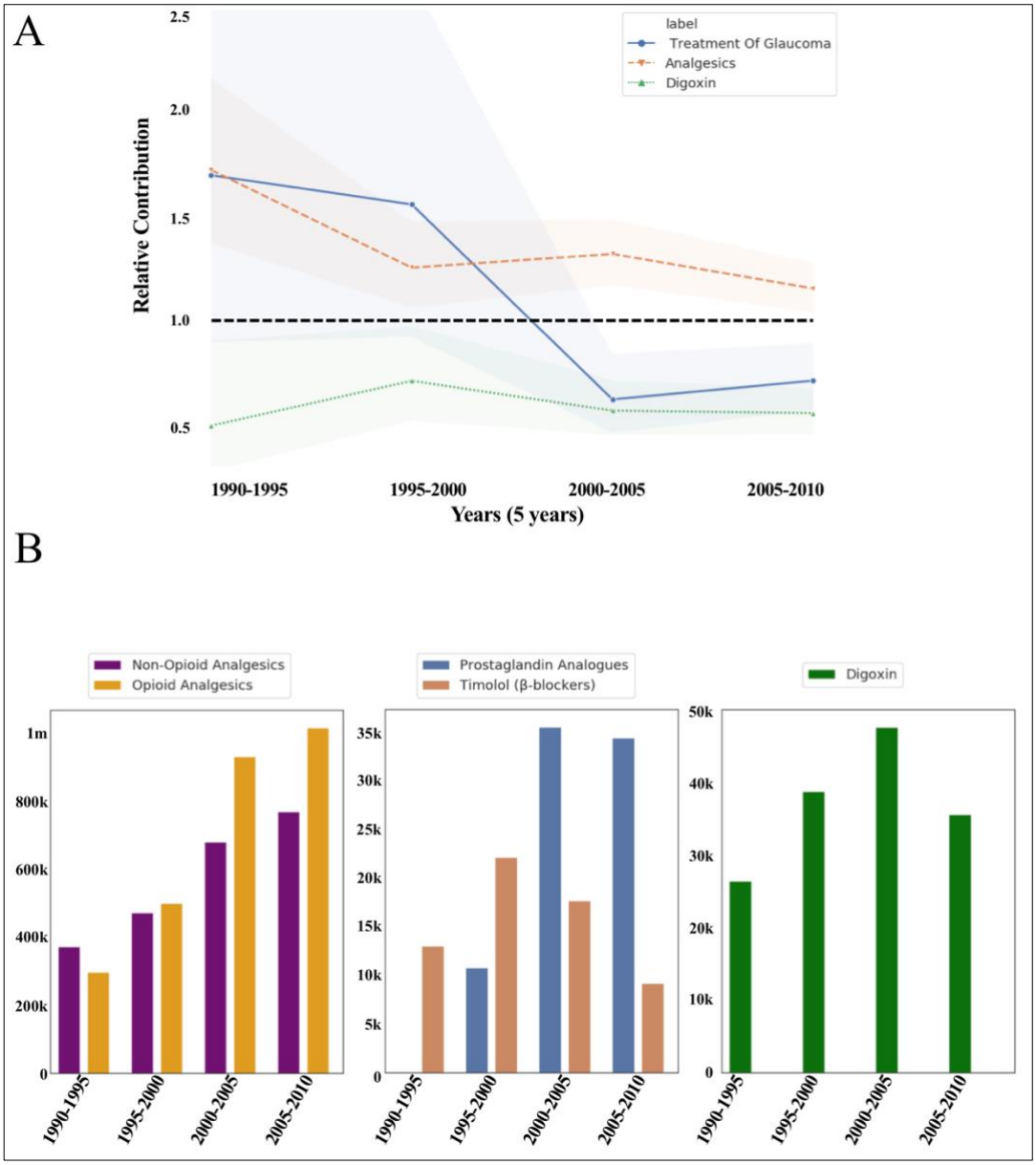
least, non-steroidal anti-inflammatory drugs – medications known to increase risk of HF[141].

In both the ablation study of incident HF prediction (see section 4.3.3.3) and the investigation of temporal variability analyses, calendar year is an important modality in the prediction task; hence, we wish to further investigate the differential RC by calendar year. For presentation, we have analysed three studies of <1 RC medications shown in Figure 5-6: digoxin, treatments for glaucoma, and analgesics.

Figure 5-6 part A shows the RC for the three medications stratified by calendar year. While digoxin was consistently negatively associated with incident HF across year groups, and analgesics, positively associated across the same, the results were more heterogenous for the medication, treatment for glaucoma. A hypothesis: while BNF coding was consistent over the decades, the underlying drug composition might have changed at approximately, the year 2000. Thus, we analyse the number of times different digoxin, analgesic, and glaucoma medications were first prescribed in patients between 1990 and 2010 in part B of the same figure. For this analysis, repeat medications were not counted.

We found that throughout the decade of the 1990s, timolol, a beta-blocker, was a common topical treatment for glaucoma[142]. With the introduction of novel medications in the 2000s, the use of the ophthalmic timolol began to decline[143].

Figure 5-6: Calendar year stratified relative contribution analyses



The year-stratified RC of medications. A, RC (mean; 95% CI) of three medications to HF prediction stratified by year. X and y represent year group and RC respectively; black line denotes 1.0 RC. B, frequency of drugs by components in different year groups in Dataset A. X/y represent the year group/counts of first-time drug (component) prescriptions to patients respectively. Individual drug components are represented with bars in different colour. Part of this figure is from Rao et al[131].

Our RC analyses in Figure 5-6 part A shows that BEHRT implicitly captured this change in the prescription of the treatment. Specifically, the BEHRT identifies that the treatment timolol, prescribed before 2000 highly associated with HF with RC>1[144].

Timolol indeed has had known cardiovascular side-effect such as bradycardia with the potential to exacerbate incident HF[144]. Following the year 2000, BEHRT can identify that the prevalent treatment, prostaglandin analogues, has <1 RC. As opposed to timolol, prostaglandins and related analogues – such as I2 and others – have known potential to reduce cardiovascular risk[145–147]. However, large-scale trials investigating potential preventative effect of are currently lacking[145,146].

In the case of analgesics, a declining trend, albeit firmly >1 RC trend emerges. Prescription of the analgesics were more non-opioid-based than opioid, prior to 1996. We see that BEHRT parallels this generational shift in drug component prescription through RC; non-opioid analgesics are primarily composed of nonsteroidal anti-inflammatory drugs and generally increase the risk of cardiovascular events[141,148]. Thus, RC prior to 1995 is shown to be quite high. Tracing the gradual change in majority preference to favour opioid based analgesics following 1995, RC captures this change in prescription behaviour and as a result, attenuates in the following decades.

On the other hand, prescription of digoxin wanes following 2005. However, the stable <1 RC across any and all years, further reinforces the hypotheses that the positive inotropic drug could help with preventative efforts for HF.

5.5 Interpretation

Our investigations into explainability had several outputs. The temporal variability analyses complemented and confirmed our results from previous investigations of calendar year. Furthermore, the contribution analyses appropriately worked through the proposed pipeline for trust in deep learning for clinical prediction, and captured validated risk factors of HF. Furthermore, the analyses provided insights

into a variety of potentially novel risk and protective factors in the forms of both diseases and medications.

In terms of novelties, this work, is by our understanding, the first work to introduce methods to better understand Transformer based models for incident HF prediction interpretability. While methods to query convolutional and recurrent models do exist, the methods have not been extended for the Transformer framework; thus, the perturbation method is a novel and useful contribution to model explainability, especially in the context of EHR and cardiovascular medicine.

With regards to the results from the contribution analyses, we showed that BEHRT can capture the dichotomy of risk and treatment with nuance. While BEHRT generally protective and risk factors appropriately, the age-stratified results had some interesting insights.

First, we show that in validation of established risk factors and the top disease/medication risk factors, as age increases, the individual contribution of risk factors attenuate. This is consistent with previous epidemiological evidence[135]. Additionally, this is in line with understanding of multimorbidity; as age increases, patients develop more illness (i.e., develop multimorbidity), and each individual illness might contribute less to incident HF. Second, we saw that risk factors sometimes associated with non-HF or equivalently, were negatively associated with HF. A cursory analysis might point towards an incorrect conclusion that BEHRT has incorrectly captured risk factors, such as hypertension, negatively associating with HF. This conclusion is biased by indication; the correct interpretation is that the medication for hypertension serves as a proxy for the disease itself, and ultimately has a negative relationship with HF, as noted in several studies[15,149]. Our analyses of risk and treatment of risk generally shows that while risk factors associate with HF, treated risk attenuates

while untreated risk exacerbates the association with HF – conclusions consistent with understanding of HF protection and risk.

In the age and year stratified analysis, the BEHRT model demonstrates that some medications might potentially provide preventative benefits. In the case of both digoxin and the prostaglandin analogues form of treatment for glaucoma, the consistent RC <1 in both age and year stratification signals that these drugs associate more so with non-HF, and hence, potentially preventative. However, as with conventional models, causal interpretations are not the most appropriate per se and must be presented with caution. On the other hand, this work does not justify causality but rather generates hypotheses; with triangulation of evidence from this source and others, further confirmatory studies can be appropriately crafted to appropriately identify the nature of the association.

6 CAUSAL INFERENCE AND ASSOCIATION ANALYSES

In this chapter, the research to meet the final objective (see section 1.2) will be presented. I have demonstrated, to a considerable degree, that deep learning models fit well and are useful for prediction tasks and can capture known factors of risk and protection well. With this, I proceed to developing models for conducting causal inference more efficiently than benchmark models. When confounding variables are not known or latent interactions in the data, the fundamental issue is that implemented models for understanding causal effect in a given observational study will be under-adjusted. Under-adjustment biases can lead to incorrect conclusions regarding both the strength and direction of the association. Hence, in this chapter, I develop a derivation of BEHRT for estimating causal effect that directly address these issues in addition to others and more accurately estimate effect size as compared to benchmark methods. Furthermore, I implement the model to study the association between antihypertensives and cancer, an association, which is well-studied in meta-analyses of randomised evidences.

The following sections are published in *IEEE Transactions on Neural Networks and Learning Systems* (doi.org/10.1109/TNNLS.2022.3183864). This publication is a product of work by multiple authors. As a first author, my role consisted of designing the study, conducting literature review, processing data, conducting statistical/deep learning

analyses, and writing the first draft of the manuscript. Material from the publication (including figures, tables, and text) have been amended for presentation in the following sections.

6.1 Introduction

Estimating causal effect of a hypothetical exposure (intervention) is a core problem for epidemiology. The following association is used as a facilitating example: the effect of antihypertensive drug classes on cancer. As this example centres around a study of medications, which are randomizable exposures, this association is optimally studied with a randomised control trial (RCT). In gist, as discussed in section 2.4.1, in RCTs, the intervention is randomly given to patients – some treated with Calcium channel blockers (CCBs) and others with Angiotensin receptor blockers (ARBs) – and incidence of cancer is compared in the two exposure groups over some finite follow-up period. RR, as discussed in section 2.4.1, is a common measure of comparing the risk of cancer in one group versus the other. Directly, the randomisation means that confounders are randomised and hence as a result, balanced between the two groups. Because of this balance in confounding factors, the confounding factors ultimately "washes" away. Hence, given sufficient sample size, the trials offer unconfounded estimates of causal effect.

In the case of the association of the various classes of the antihypertensives and cancer, numerous RCTs have indeed found the association to be null meaning that no class of antihypertensives cause cancer any more than any other class of the same[150].

In situations, in which RCTs cannot exist, are unfeasible, or fail to generalise, well conducted observational studies alternatively offer answers on the nature of the association [74,75,151,152]. However, important to carrying out observational studies, the

adjustment of variables that are confounding the association between exposure and outcome is necessary. Omission of confounding variables can render biased and hence, false conclusions concerning the strength and at times, even direction of the association.

As discussed in sections introducing causal inference (see section 2.4.3), traditionally, semi-parametric and parametric statistical modelling have been explored in observational causal inference in the field of epidemiology. Regression based models (e.g., log-binomial or logistic regression) incorporate the exposure variable into the modelling in addition to any adjustment variables and implement regression fitting for estimation of outcomes[153,154]. Another solution proposed is to adjust solely for the variables that are associated to exposure by propensity score modelling; however, naïve propensity score-based methods require correct specification of both the exposure and outcome prediction models, often not guaranteed[155,156]. Misspecification directly implies that the errors of the weights rapidly increase ultimately producing high-variance and volatile downstream causal estimates[157]. Furthermore, these statistical models all rely on conventional confounder selection (feature engineering); those confounders unknown to the experts conducting observational studies will be omitted from modelling. In observational settings involving high-risk, multimorbid, or simply, poorly understood patient cohorts, comprehensive adjustment of confounders is less guaranteed invariably hampering downstream association estimation.

"Doubly robust" modelling, a recent development in semi-parametric modelling, addresses issues of misspecification directly. As opposed to requiring consistency of both, these doubly robust estimators only require the consistency of either prediction of propensity score or prediction of outcome to produce unbiased causal effect estimates, and examples such as Targeted Maximum Likelihood Estimation (TMLE) and derivatives such as the Cross Validated TMLE (CV-TMLE) have been prolifically used to explore

causal inference problems of average treatment effect (ATE)[151,158–160]. TMLE-based methods have recently also been applied for epidemiological studies on routine EHR-based causal inference[161].

In addition to access to high quality and multimodal EHR allowing for more comprehensive adjustment if conducted appropriately, the developments in deep learning modelling have allowed for scalable modelling of a variety of data as discussed in section 2.5. Additionally, the development of representation learning methods such as unsupervised training strategies have given rise to deep learning frameworks that can conduct richer feature extraction offering better generalisability[162]. Research has shown that auxiliary unsupervised learning 1) adds an additional inductive bias (i.e., forces a relevant learning task) ultimately improving generalizability and 2) helps to learn representations shared or beneficial for the main task – in our case, the two tasks being propensity confounding adjustment and causal inference[162–164].

In the last decade, there have been advances in deep learning for causal inference. Models like Treatment Agnostic Representation Network (TARNET), Dragonnet, Causal Effect Variational Autoencoder (CEVAE), and others have been tested on synthetic and semi-synthetic derivations of static tabular data[152,165–167]. The TARNET model have also been applied in the epidemiological setting in COVID-19 related research[168]. Additionally, the Dragonnet model explores observational causal inference by exploiting the sufficiency of the propensity score to simultaneously model the propensity score and the outcomes[74,152]. However, these models have not been tested in the context of Transformer models and in the setting of routine EHR. And even though multimodal deep learning modelling is a staple approach for risk prediction and classification studies, few approaches firstly model both temporal and static variables for causal inference and secondly, develop appropriate environments to objectively test

estimation abilities of various models. Lastly, the considerable literature of deep learning for causal inference investigates conditional ATE/ Individualized Treatment Effect (ITE) almost exclusively; methods have rarely been evaluated for accuracy of RR estimation – a metric preferred by clinicians since RR captures relative utility of a hypothetical exposure variable (i.e., as compared to the risk in the control cohort).

6.2 Aims

Since RR is often the estimand of choice in clinical research, I aim to develop and evaluate methods that combine advances across deep learning and statistics in order to estimate RR more accurately than the conventional modelling benchmarks. This overarching objective is achieved through three independent contributions.

The first contribution is the design of a novel deep learning model, Targeted Bidirectional EHR Transformer, (Targeted-BEHRT) for more accurate RR estimation. The method synthesizes the following elements in a unique multi-task learning framework: 1) extenedd BEHRT architecture (Transformer-based feature extractor) for incorporation of both temporal and static variables, 2) auxiliary unsupervised learning framework for richer feature and hence, confounder extraction, and 3) doubly robust semi-parametric estimation for mitigating various selection biases including finite-sample estimation biases[123,156].

As the second contribution, a testing environment is developed to objectively evaluate accuracy of RR estimation of various models. Focusing on the aforementioned case study: the effect of various classes of antihypertensives on cancer, we form an observational dataset by including patients taking different classes of antihypertensives and investigate risk of cancer. Our reference exposure is Angiotensin Converting Enzyme Inhibitors (ACEIs), one of said classes of antihypertensives. Since the data generating

function for real-world patient data is unknown and counterfactual outcomes (i.e., outcome under a specific exposure status) are missing in the observational dataset, ground truth RR is inaccessible. Hence, objective comparison of model estimation is difficult. To directly address this issue, semi-synthetic derivations of the observational dataset are constructed with generated ground truth RR, and then the proposed model is applied against statistical and deep learning benchmarks in several experiments to identify the model with best RR estimation. Additionally, to test the model in situations of limited data, the utility of Targeted-BEHRT compared to other models is demonstrated in finite-sample estimation experiments.

As the third and final contribution: after validating the model on semi-synthetic derivations of routine clinical observational data, we demonstrate the model can be applied to the aforementioned observational study: the effect of ACEIs on cancer relative to other drug classes. Traditional observational studies have demonstrated conflicting results; however, these associations have been deemed null in numerous RCTs and meta-analyses of randomised evidences with narrow confidence intervals, across a wide range of patient groups, for multiple cancer subtypes[150,169].

6.3 Methods

6.3.1 Formal definition of task

The objective is to estimate RR in the setup of binary exposure and outcome. As described in section 2.4.2, we revisit the potential outcomes framework[67]. Consider the population of patients described by a tuple generated independently and identically: $(X_i, Y_i, T_i) \sim P$. Each patient *i* is described by medical records, X_i and is assigned exposure status, $T_i \in \{0,1\}$. The exposures, T_i in the presented work are two classes of antihypertensives with one of the classes acting as reference group. The variable Y_i

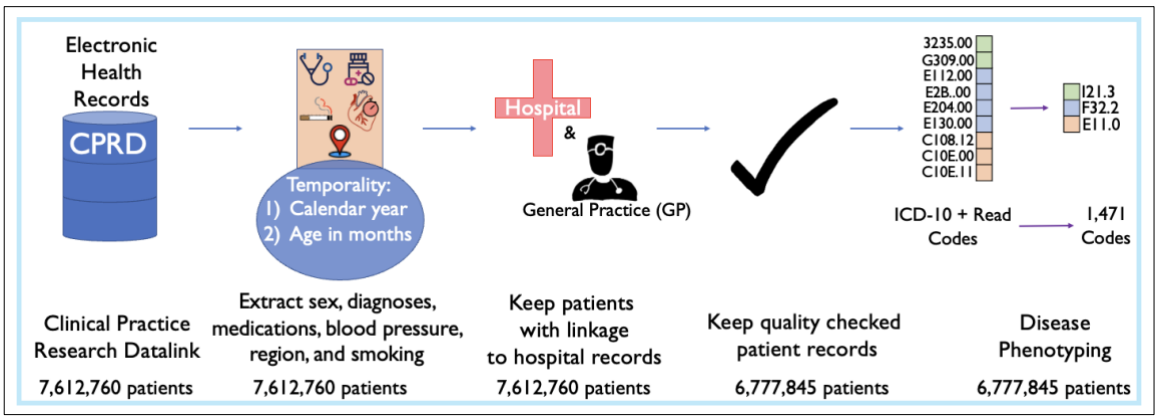
corresponds to the observed outcome – cancer – in the proposed investigations. In a fixed amount of "follow-up" time after hypothetical treatment, $T_i = 0$, outcome of cancer is notated as $Y_i(0)$, and similarly for treatment, $Y_i(1)$.

The RR cannot be directly computed since the counterfactual outcome is not available for inclusion. Hence, under the conditions of consistency, positivity, and unconfoundedness (see section 2.4.2.1), the exposure effect is identifiable and RR can be estimated as $\boldsymbol{RR} = \frac{\mathbb{E}[\boldsymbol{Y_u(1)}]}{\mathbb{E}[\boldsymbol{Y_u(0)}]}$.

6.3.2 Data

In addition to dataset pre-processing as discussed in section 3.6, the dataset for the investigations was restricted to patients who were (1) registered with the general practice for at least 12 months and (2) registered with a practice that provided consent for linking the data with national databases for hospitalizations and death registry.

Figure 6-1: Data selection for representation learning



We use Clinical Practice Research Datalink (CPRD) and extract diagnoses, medications, blood pressure, smoking, region, and sex records. We homogenize codes from ICD-10 and Read to one format. Unmapped Read codes were kept for completeness. The figure was adapted from Rao et al[170].

We extracted diagnoses, medications, blood pressure measurements, sex (male, female), region (10 regions in England), and smoking status (non, previous, or current

smoker). Diagnoses and medication codes were homogenized for machine readability. In sum, this processing led to a dataset of 6,777,845 patients, which was used for general representation learning (shown in Figure 6-1) for deep learning models.

For the causal inference investigation (i.e., investigating the effect of antihypertensive on incident cancer), a dataset containing five subpopulations had to be selected – one for each class of antihypertensives: ACEIs, diuretics, CCBs, Beta Blockers (BBs), and ARBs. Patients were selected in one of these groups based on first class of antihypertensive medications recorded before 2009 and if free of cancer diagnosis before this first prescription; the year 2009 was chosen to allow sufficient 'follow-up' time for the analysis of the occurrence of potential cancers. The date of this first prescription was defined as the "baseline" (a date between 1 January 1985 and 31 December 2008). Patients were then followed up from baseline until incident cancer report (including cancer diagnoses as cause of death) or end of the five-year follow-up period. The learning period included the entire patients' medical records up to a random point between 6 and 12 months before baseline. This feature of adjustment is to account for any potential inaccuracies in timing of prescription (or decision to prescribe) and to avoid possibility of antihypertensive prescription itself influencing the model training. "CPRD Product codes" are used for identifying classes of antihypertensives obtained from a dataset published by University of Bristol[171]. Codes for identifying cancer were found in published code sets validated for CPRD data[9].

6.3.3 Semi-synthetic data generation

Data generation of sequential, temporal variables is currently a difficult and arguably an unsolved task; many approaches have been suggested but have not undergone rigorous validation. Hence, instead of synthesizing all medical history variables including pre-exposure, exposure, and outcome variables, the existing medical history and exposure

variables in observational data was used to exclusively simulate binary factual and counterfactual outcomes.

Inspired by other semi-synthetic data simulations, intuitively, the association between a medical history variable Z_i (e.g., some diagnosis/medication) and exposure T_i with the empirical propensity in the dataset: $\lambda_i = P(T_i = 1|Z_i)$[167,172] were first modelled. If associated to an exposure ($\lambda_i \neq 0.5$), the potential outcomes, $Y_i(T = 1)$ and $Y_i(T = 0)$ as a function of λ_i and exposure $T_i = 1$ and $T_i = 0$ respectively were generated. In this way, semi-synthetic outcomes arose from an association between Z_i and exposure and Z_i and the outcome. Hence, the relationship between exposure and outcome is confounded via simulation by Z_i. While the empirical RR – the proportion of the outcome in one exposure group divided by the same in the other – would yield confounded causal conclusions, effectively adjusting for the confounder variable, Z_i , would yield identifiable (see section 6.3.1 and extended introduction in section 2.4.2) causal association between exposure and outcome.

In addition, to test model adjustment potential in situations of varying confounding intensity, the contribution of the confounding was weighted with a β factor: the greater the β implies the greater the confounding. More details of the semi-synthetic data generative process and functions modelled are in supplementary methods section 9.3.

In this work, investigations in semi-synthetic data utilizing two forms of confounders, persisting and transient confounding are presented. Persisting confounding is defined as confounders that are assigned at birth and persist through one's life course: ethnicity, sex, genes, and other variables assigned at birth that associate to variables later in age. Transient confounding is defined as confounders that manifest at a point or period of one's life effecting events downstream in time: disease diagnoses, age itself,

prescriptions, and other variables manifesting during one's life course (and specifically, after birth). These two distinctions of confounding are presented in this work because these are prevalent forms of confounding seen in population health research[173]. While these forms of confounding are not comprehensive, these two forms allow testing of various models.

From the observational dataset, two exposure groups – ACEIs and Diuretics are investigated; upon close examination of pre-exposure variables and associations with exposure status, it was found that female sex was associated to the Diuretics exposure status and hence, it is chosen to be a persistent confounder and conditional outcomes were generated. For another pair of exposures, i.e., ARBs and CCBs, the association of incidence of at least one of heart failure, hypertension, ischaemic heart disease, and diabetes mellitus to CCBs was found associated to exposure status. Thus, occurrence of at least one of these diseases as "cardiometabolic diseases" is set as the confounding variable and is utilized as a transient confounder for the second set of semi-synthetic data experiments. The various strengths: low, medium, and high confounding intensity are set for experiments with sex and cardiometabolic disease as confounders (β values: [1, 5, 10] and [25, 50, 75] respectively). In sum, with this confounding generation method, model confounding adjustment ability is tested with two forms of confounding at various degrees of intensity (β values) offering a total of 6 experiments (two forms of confounding at three levels of strength each).

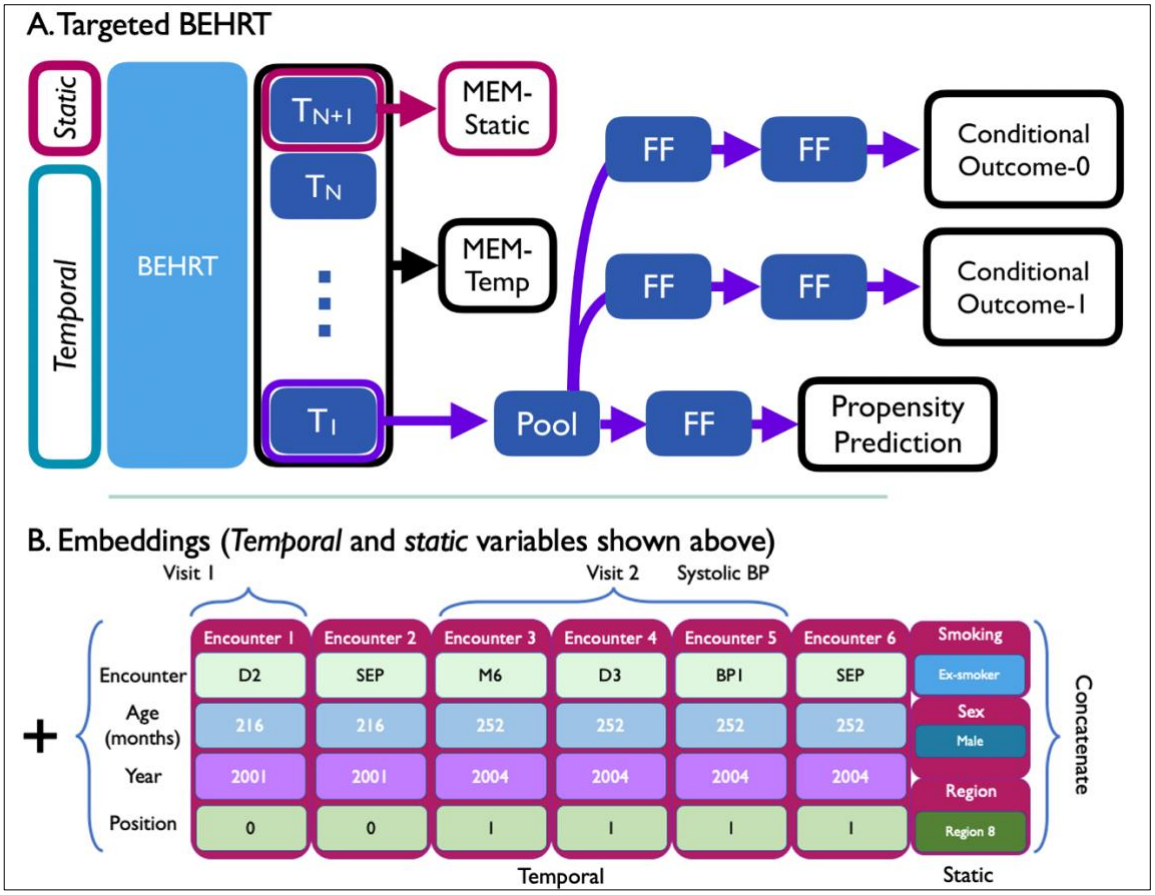
On the semi-synthetic dataset with highest intensity of cardiometabolic disease confounding, finite-sample causal estimation experiments are conducted. Since estimation in limited sample settings are known to be unstable in many cases (e.g., for inverse probability weighted estimators) despite asymptotic guarantees, the finite-sample estimation ability of models is important and worth testing[174]. And, the confounding

strength level is specifically set to the highest intensity level (β=75) because we wished to investigate how the model performs in estimation of RR in situations subject to the greatest/strongest confounding. The finite-sample estimation ability of the proposed model and other deep learning models are explored by applying the models on random sub-samples of this dataset: 2.5%, 5%, 10%, 25%, 50%, and finally, the entire dataset.

6.3.4 Model development

The model, Targeted-BEHRT, utilizes a modified feature BEHRT extractor to capture both static and temporal medical history variables and captures initial estimates of RR. After predicting propensity score and conditional outcomes with independent linear maps (and appropriate sigmoid activation functions for binary outcomes/exposures), we use CV-TMLE to correct for bias in initial RR estimate and compute corrected RR (see Figure 6-2).

Figure 6-2: Targeted-BEHRT model and embedding structure



A. Above, the model is shown. Generally, an input x (static and temporal variables) is fed to a feature extractor, which outputs a dense latent state (for EHR modelling, this feature extractor is BEHRT). The output of the final layer of the BEHRT feature extractor is fed to the Masked EHR modelling (MEM) prediction head to predict any masked encounters. T_{N+1} token state is fed to a Variational Autoencoder (VAE) neural network to predict masked static variables. The latent state of the first token (T1) is fed to a pooling layer to predict propensity and conditional outcomes with multiple prediction heads with feed forward (FF) neural network layers. The loss consists of the unsupervised loss from two MEM components – temporal (temp) and static (static) unsupervised data training - and the supervised loss of the propensity and factual outcomes. B. Below, the embedding structure for modelling rich EHR data is shown. Clinical encounters timestamped by age/year/position (visit number) are converted to vector representations and fed to model as temporal variables. Static data variable embeddings: patient sex, region in UK, and smoking status are concatenated to the temporal variable embeddings. The figure was adapted from Rao et al[170].

Intuitively, Targeted-BEHRT first extracts latent EHR features from static covariates and fixed sub-sequences of medical history with BEHRT. Second, the model predicts propensity of exposure and conditional outcome using these learned features.

Third, by additionally conducting auxiliary unsupervised learning, the model trains on reconstruction of both static and temporal data with two-part Masked EHR modelling (MEM).

The propensity prediction model is modelled as 1-hidden layer multilayer perceptron (MLP) and for each conditional outcome, we use a 2-hidden layer MLP with Exponential Linear Unit (ELU) activation.

With patient data tuple (X_i, Y_i, T_i) as described in section 6.3.1, parameters θ, propensity prediction head $g(X_i)$, and conditional outcome prediction heads, $H(X_i, T_i)$ for input X_i and exposure T_i for patient *i*, the loss is:

$$\widehat{\boldsymbol{O}}(\boldsymbol{X_i}; \boldsymbol{\theta}) = \boldsymbol{CrossEntropy}(\boldsymbol{H}(\boldsymbol{X_i}, \boldsymbol{T_i}; \boldsymbol{\theta}), \boldsymbol{Y_i}) + \boldsymbol{CrossEntropy}(\boldsymbol{g}(\boldsymbol{X_i}; \boldsymbol{\theta}), \boldsymbol{T_i}) \quad (6\text{-}1)$$

Next, we conduct MEM for two-part unsupervised learning: (1) temporal variable and (2) static variable modelling. The first part – unsupervised learning on temporal data – functions similarly to MLM in Natural Language Processing[119]. In MLM, the model receives a combination of masked, replaced, and unperturbed tokens (temporal or textual data) and the task is to predict the masked or replaced encounters. We do the same but additionally enforce another constraint: when replacing encounters, we do not replace encounters with those that define the exposure or outcome - antihypertensives and cancer in the current set of experiments. With encounter *j* for patient *i* represented as $E_{i,j} \subset X_i$ (i.e., encounters being a subset of the input X_i), masked/replaced encounters represented as $\tilde{E}_{i,j}$, BEHRT feature extractor *B*, temporal unsupervised prediction network M, neural network parameters $\phi_{MEM-Temp}$, we develop objective function:

$$\widehat{\mathcal{L}_{MEM-Temp}}\left(\boldsymbol{E}_{i,j};\boldsymbol{\phi}_{MEM-Temp}\right) = \sum_{j=1}^{|E_i|} \boldsymbol{CrossEntropy}\left(\boldsymbol{M}\left(\boldsymbol{B}\left(\widetilde{\boldsymbol{E}}_{i,j};\ \boldsymbol{\phi}_{MEM-Temp}\right)\right), \boldsymbol{E}_{i,j}\right) \tag{6-2}$$

For the second part of the MEM, static data modelling, variational autoencoders (VAEs) were chosen for the representation learning for the static data. We model static categorical variables: region, smoking status at baseline, and sex; the three variables are embedded in high dimensional embeddings (embedding dimensions for each variable are hyperparameters of the Targeted-BEHRT model) and mapped (via 1-layer MLP) to the size of the encounter (temporal) embeddings and finally, concatenated to the encounter embeddings. Hence, the BEHRT model functions as feature extractor for static/temporal variables and encoder for the VAE (see Figure 6-2). The temporal variables interact with the static variables through the multi-head self-attention mechanism of the BEHRT architecture[123]. For training the VAE, similarly to the temporal modelling, we mask some variables as input, and use a variable-specific decoder to decode the variable (if masked). Specifically, for static variable $X_{i,v}$ of a total of V static variables patient i, $q_{\phi_{Enc}}(Z_i|X_i)$ representing the encoder, and $p_{\phi_{Dec}}(X_{i,v}|Z_i)$ representing the multivariate Bernoulli decoder for variable v, the VAE loss is:

$$\widehat{\mathcal{L}_{MEM-Static}}\left(\boldsymbol{x}_i;\boldsymbol{\phi}_{Enc},\boldsymbol{\phi}_{Dec}\right) = \sum_{v=1}^{V}\sum_{i=1}^{n} \log\ \boldsymbol{p}_{\boldsymbol{\phi}_{Dec}}\left(\boldsymbol{X}_{i,v}|\boldsymbol{Z}_i\right) - \sum_{i=1}^{n} \boldsymbol{D}_{KL}\left(\boldsymbol{q}_{\boldsymbol{\phi}_{Enc}}(\boldsymbol{Z}_i|\boldsymbol{X}_i)\ ||\boldsymbol{p}_{\boldsymbol{\phi}_{Dec}}(\boldsymbol{Z}_i)\right) \tag{6-3}$$

The complete objective function to be minimized is the summation of Equations (6-1), (6-2), and (6-3) as shown in Equation (6-4):

$$\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\varepsilon}}, \widehat{\boldsymbol{\phi}}_{\boldsymbol{Enc}}, \widehat{\boldsymbol{\phi}}_{\boldsymbol{Dec}}, \widehat{\boldsymbol{\phi}}_{\boldsymbol{MEM-Temp}}$$

$$= \underset{\boldsymbol{\theta},\boldsymbol{\varepsilon},\boldsymbol{\phi}_{\boldsymbol{Enc}},\boldsymbol{\phi}_{\boldsymbol{Dec}},\boldsymbol{\phi}_{\boldsymbol{U}}}{\mathbf{argmin}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \widehat{\boldsymbol{O}}(\boldsymbol{X_i}; \boldsymbol{\theta}) \tag{6-4}$$

$$+ \boldsymbol{\delta}\left(\widehat{\boldsymbol{\mathcal{L}_{MEM-Temp}}}\left(\boldsymbol{E_{i,j}}; \boldsymbol{\phi_{MEM-Temp}}\right)\right.$$

$$\left. + \widehat{\boldsymbol{\mathcal{L}_{MEM-Static}}}(\boldsymbol{X_i}; \boldsymbol{\phi_{Enc}}, \boldsymbol{\phi_{Dec}})\right)$$

With hyperparameter δ for weighting the contribution of the unsupervised MEM loss terms.

6.3.5 Processing data for modelling

The modalities of CPRD considered for[130] deep learning modelling were sex, region, diagnoses from both primary and secondary care, medications, SBP measurements, and smoking status. We mapped Read codes from primary care and ICD-10 codes from secondary care to 1,471 unique ICD-10 diagnostic codes to harmonize disease codes in the dataset; unmapped codes were included for completion[120,175]. Furthermore, we mapped medication codes to 426 codes in the BNF coding format. Since SBP is a continuous variable and the feature extractor requires discretized elements (see section 6.3.4) SBP measurements (in mm Hg) were grouped into 16 categories based on pre-specified boundaries ([90-116], (116,121], (121,126], …, (181,186], >186). Furthermore, calendar year, age (months), and relative position (visit number) were utilised for the sequential/temporal modalities. Each patient p had n_p encounters, or instances of modalities: diagnoses, medications, and SBP measurements. Smoking status at baseline (non, previous, or current smoker), region (10 regions in England, and sex (male, female), were static variables included in modelling.

6.3.6 Benchmarks and causal estimation

Before pursuing the causal investigations with deep learning modelling, contextualized EHR embeddings and network weights are pre-trained via the MEM task on the pretraining dataset. This MEM task is used to generally train weights on all patients in CPRD before progressing to causal modelling (6,777,845 patients in Figure 6-1).

For semi-synthetic investigations, several statistical and deep learning models are implemented to serve as benchmarked comparison models for causal inference. The benchmarks include Bayesian Additive Regression Trees (BART), Logistic Regression (LR) and L1/L2 regularization variants, and LR with Targeted Maximum Likelihood Estimation (TMLE)[156,176]. The covariates for these models were chosen to be baseline age, smoking status, sex, region, incidence of 33 curated disease groups, and additionally prescription of four additional medication groups. While inclusion of baseline variables in epidemiological observational studies is standard practice, the disease/medication groups were included to enable a fairer comparison to deep learning modelling. Furthermore, diagnoses and medications are known to be confounders in observational studies, so adjustment of these variables is important for causal estimation. To ensure that the diagnoses and medication groups are medically valid clusters of diseases and medications respectively, groups compiled by past medical research are utilised in this research project[9,171]. A deeper explication is given in Supplementary section 9.3.2.

To serve as deep learning benchmarks, staple deep learning models for average causal effect are implemented: TARNET, TARNET + MEM (i.e., with unsupervised MEM component), and Dragonnet with BEHRT feature extractor and the embedding format presented in Figure 6-2A. These models are initialised with pretrained weights. After implementing and evaluating benchmarks, the proposed model, Targeted-BEHRT

with pre-trained network weights where applicable is implemented and modelling of semi-synthetic data investigations is pursued.

For the semi-synthetic data experiments, variables of cardiometabolic disease and sex respectively as input are not fed into models; it was necessary that the statistical and deep learning models infer confounding from remaining input variables. In routine clinical data, the observational studies would often not have access to all confounding variables. Hence it is important to test models' ability to adjust for confounding given limited input variables.

For all investigations, experiments with five-fold cross validation causal estimation were conducted. We calculated RR on the test dataset for each fold as advised by Chernozhukov et al and compute 95% Confidence Intervals (CI) over the five folds[158]. RR defined by naïve estimator on a finite sample: $\hat{\psi} = \mathbb{E}\left[\frac{\mathbb{E}[H(X,1)]}{\mathbb{E}[H(X,0)]}\right]$ for TARNET, TARNET-MEM, LR (and L1/L2 regularization variants), and BART are estimated. For Targeted-BEHRT, we use the CV-TMLE method for the estimation of RR. For Dragonnet, the model with the CV-TMLE estimator in order to directly compare the model with this benchmark model. In addition, we also implement the Dragonnet model with the naïve estimator (i.e., the original model without post-hoc estimator). For more information on the CV-TMLE method, advantages over TMLE, and implementation, please refer to supplementary material section 9.3.3. For models that utilized predicted propensity scores, we conducted propensity score trimming and exclude patients with predicted propensity score greater than 0.97 and less than 0.03 before pursuing RR calculation[177].

We identified the superior model by identifying the model with least Sum Absolute Error (SAE) over the three β values for each confounding experiment. We report

the Standard Error (SE) for the SAE; this metric was calculated using additive propagation of error[178]. For deep learning models, we also present the same results with an ablation study; the change in SAE is presented as modules are iteratively removed from the proposed Targeted-BEHRT model.

6.3.7 Implementation

We developed all statistical and deep learning models on python (deep learning models on Pytorch)[179]. Hyperparameters for the BEHRT feature extractor are reported in Supplementary Table S6. The Adam optimizer with exponential decay scheduler (decay rate=0.95) was used on all deep learning models to ensure training convergence[180]. For TARNET-MEM and Targeted-BEHRT, we pre-trained 5 epochs on exclusively the MEM task on the cohort investigated (in addition to pre-training on the ~6.7 million patients) before initiating joint MEM-causal task training.

After fitting deep learning and statistical models, in order to derive estimates for RR estimation, evaluation of the model on the test fold of the dataset using standard direct estimation methods was conducted[61]. For all patients in the test set, we first derived risk estimates (e.g., estimation of $P(Y|X,T=0)$) patients as if they were all assigned $T=0$, and similarly, derived estimates (e.g., estimation of $P(Y|X,T=1)$) as if they were all assigned $T=1$. In this way, the RR estimate, $\hat{\psi}$ can be derived as a function of these two quantities:

$$\widehat{\boldsymbol{\psi}} = \mathbb{E}\left[\frac{\mathbb{E}[\boldsymbol{H}(\boldsymbol{X},\boldsymbol{T}=\mathbf{1})]}{\mathbb{E}[\boldsymbol{H}(\boldsymbol{X},\boldsymbol{T}=\mathbf{0})]}\right] \tag{6-5}$$

LR (and regularization variants), BART, TMLE and CV-TMLE were implemented in python inspired by past works utilising TMLE[152]. To fit the nuisance parameter for the TMLE estimate update step, Nelder-mead optimisation was

utilised[181,182]. For deep learning models implemented with CV-TMLE, the naïve estimator (Equation (6-5)) was not used; rather, the CV-TMLE estimator was implemented utilising conditional outcome predictions, $H(X,T=1), H(X,T=0)$, and propensity score prediction, $g(X)$.

6.4 Results

6.4.1 Population statistics

Table 6-1: Characteristics for patients eligible for cohort study concerning antihypertensives and cancer

	Classes of antihypertensives				
	ACEIs	***BBs***	***CCBs***	***Diuretics***	***ARBs***
Number (%)	186709 (36)	150098 (29)	128597 (24)	28991 (5)	21970 (4)
Male (%)	101629 (54)	67794 (45)	60395 (46)	8134 (28)	10454 (47)
YOB (SD)	1938 (15)	1941 (15)	1936 (14)	1934 (16)	1940 (14)
Baseline Age (SD)	63 (14)	59 (14)	64 (13)	63 (15)	63 (13)
Number of visits (SD)	7 (4)	6 (4)	6 (4)	4 (4)	7 (4)
Baseline Year (SD)	2001 (4.2)	1999 (4.3)	2000 (4.9)	1996 (5.2)	2002 (2.8)

YOB: year of birth; baseline: the time of exposure assignment; SD: standard deviation; %: percentage. The table was adapted from Rao et al[170].

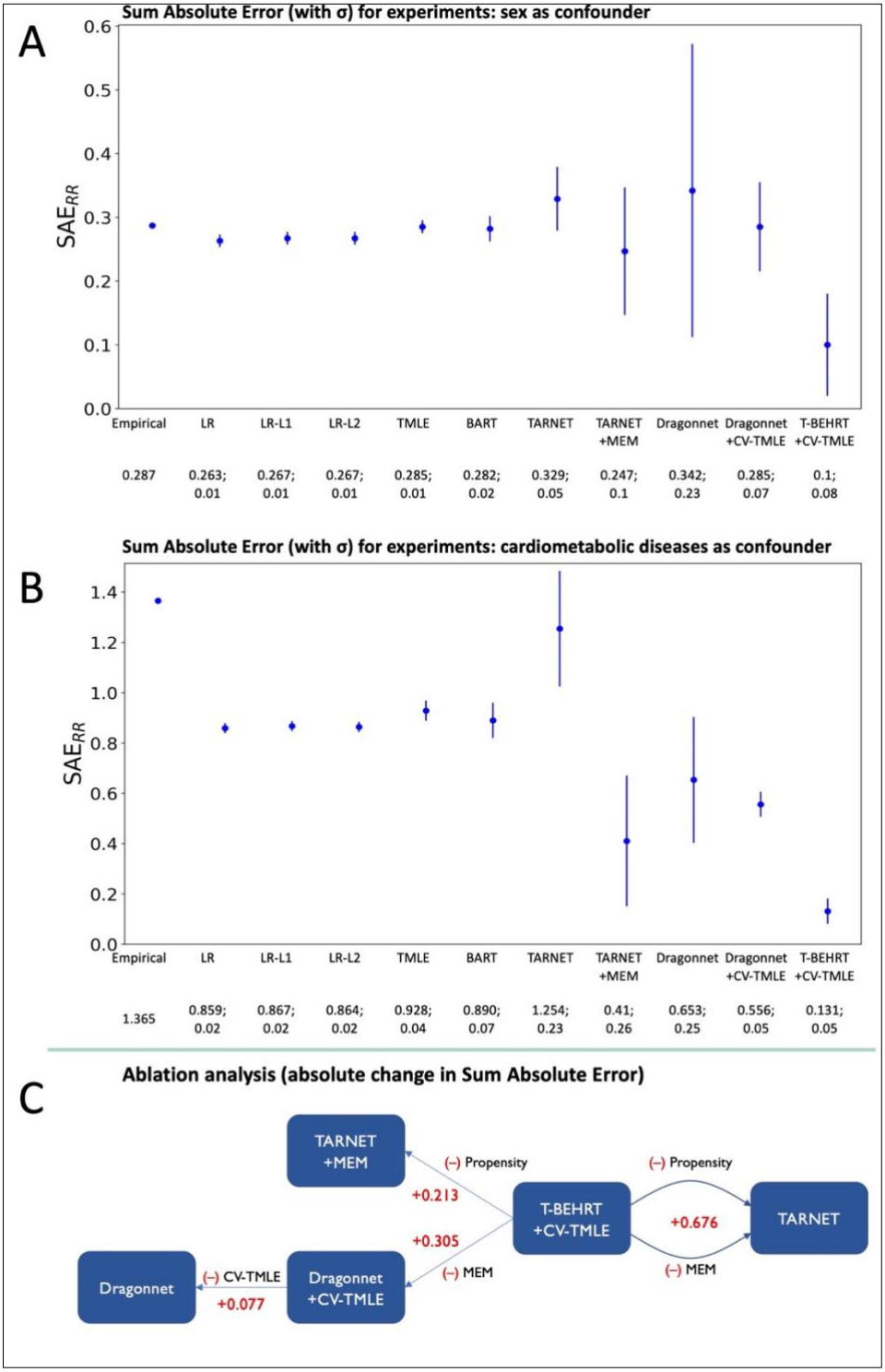
In the dataset for the investigation of antihypertensives on incident cancer, 186,709, 150,098, 128,597, 28,991, and 21,970 patients for ACEIs, BBs, CCBs, diuretics, ARBs were identified respectively totalling 516,365 patients. We demonstrate population statistics in Table 6-1. Cancer incidence counts/percentage of exposure group were 13,728 /7%, 9,819/7%, 10,232/8%, 1,784/6%, and 1,709/8% for ACEIs, BBs, CCBs, diuretics, ARBs respectively.

6.4.2 Semi-synthetic data experiments

In the semi-synthetic experiments on confounders cardiometabolic diseases and sex, we tested the Targeted-BEHRT models against several statistical and deep learning

benchmarks. In Figure 6-3, we show SAE with SE measures calculated over all β-specific semi-synthetic data experiments. We include more detailed experimental results in Supplementary Table S7 in section 9.3.5.

Figure 6-3: Semi-synthetic data experiments on various confounders and ablation analyses



Experiments on semi-synthetic data with sex (A) and cardiometabolic disease (B) as confounders; module inclusion analysis of causal modules (C). We show Sum Absolute Error (SAE) between ground truth risk ratio (RR) and estimated RR with standard error measures in both panels. The x axis is shown by the models implemented on these

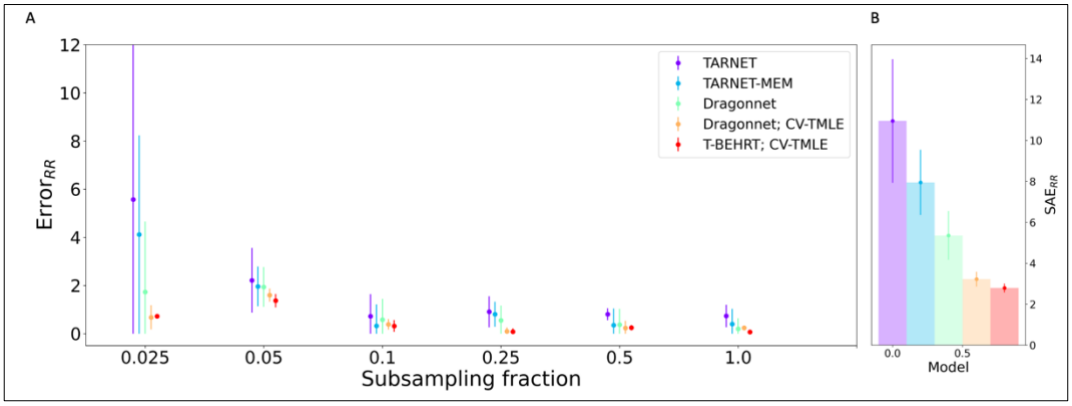
datasets, and the y axis is the SAE (lower is better). We present the numerical value and standard error measures underneath the model names. In C, we present the transformation from Targeted-BEHRT into other deep learning benchmarks. We show increase in average SAE (i.e., increase in error) across experiments of transient and persistent confounding in red as the model strips away components from its architecture indicated by (–). The figure was adapted from Rao et al[170].

Targeted-BEHRT was found to outperform all given deep learning and statistical model solutions in terms of SAE whilst maintaining narrow SE. Additionally, across both experiments, we found that deep learning models for EHR benefit from inclusion of CV-TMLE. This is seen by superior performance of both Dragonnet + CV-TMLE and Targeted-BEHRT in comparison with TARNET, which does not handle propensity score modelling. However, by investigating the exclusion of various modules from the chassis of Targeted-BEHRT shown in the ablation analysis (Figure 6-3C), we see that exclusion of MEM diminished RR estimation accuracy in a parallel way; the TARNET model with inclusion of MEM (SAE increase of 0.213) did approximately as well as Dragonnet + CV-TMLE (SAE increase of 0.305) averaged over experiments of persistent and transient confounding. Removal of CV-TMLE from Dragonnet + CV-TMLE further deteriorated performance of the Dragonnet model (SAE increase of 0.077). Ultimately, the improvement in combining both MEM and propensity/CV-TMLE modelling and forming Targeted-BEHRT demonstrated greatest SAE reduction of 0.676 - more so than the sum of its parts: 0.518 (0.231 + 0.305).

In the finite-sample estimation experiments shown in Figure 6-4, we showed that Targeted-BEHRT outperforms other models in RR estimation in individual and across data subsamples. While improvement of Targeted-BEHRT over Dragonnet + CV-TMLE is less pronounced than over other models, panel B shows that Targeted-BEHRT still demonstrates superior RR estimation performance with respect to the deep learning benchmarks. Furthermore, we found that inclusion of MEM aids more precise estimation of RR; TARNET + MEM and Targeted-BEHRT perform better than TARNET over all

finite samples as shown in Figure 6-4B. However, we note the application of CV-TMLE is more important than MEM in smaller datasets as seen by superior performance of Dragonnet + CV-TMLE as opposed to TARNET + MEM in Figure 6-4B. Furthermore, models equipped with propensity modelling (and CV-TMLE specifically) maintain relatively stable SAE across subsampling fractions while TARNET and derivatives suffer in RR estimation in smaller datasets. Lastly, across experiments in this work, while Figure 6-3 demonstrates that MEM is more important in observational settings with more samples (full dataset), Figure 6-4B shows that CV-TMLE provides greater utility in observational settings with limited samples. Implemented simultaneously (i.e., the Targeted-BEHRT model), both components ensure robust estimates across various sample sizes.

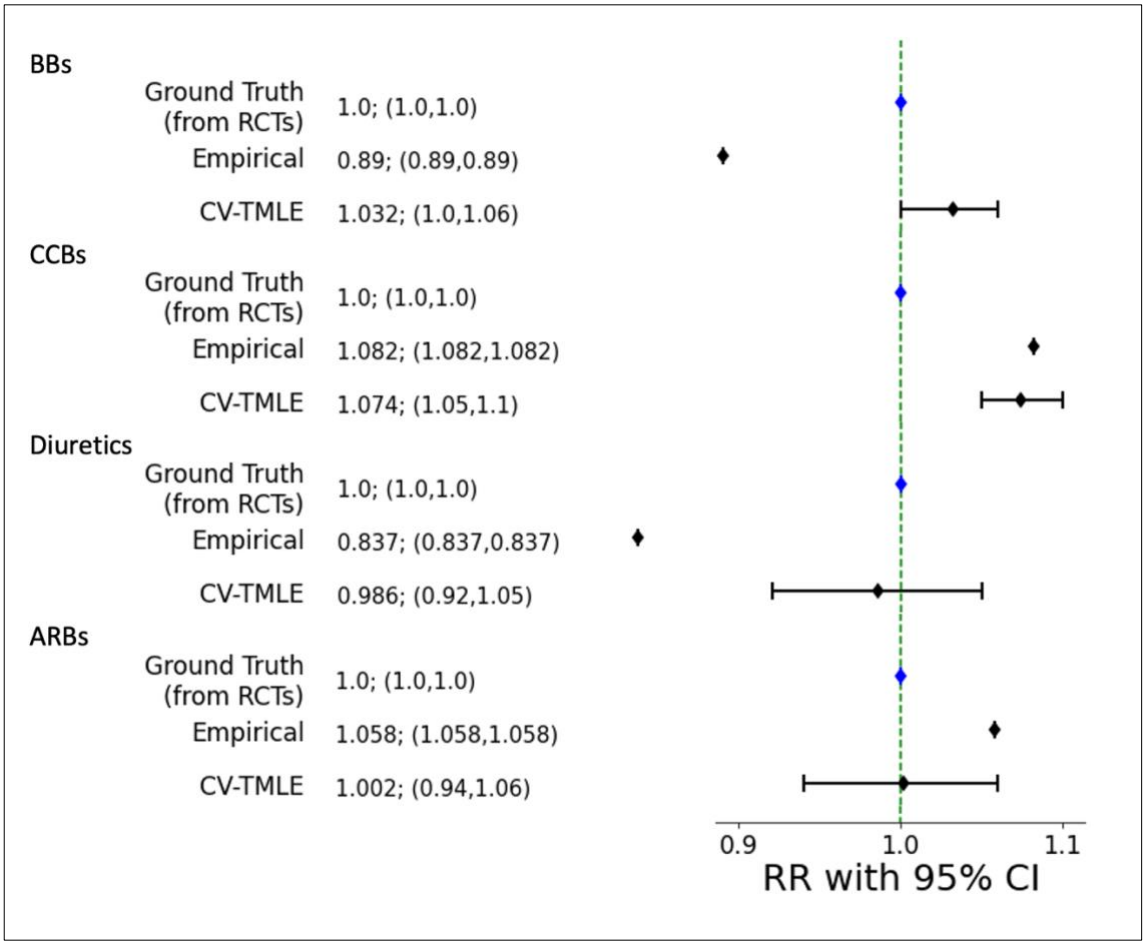
Figure 6-4: Finite-sample estimation experiments



A. We conduct experiments on finite subsamples of the semi-synthetic dataset for cardiometabolic confounding (β =75). The subsampling fraction of the dataset is shown on the x axis. The y axis shows error from ground truth risk ratio (RR). The models: TARNET (and with Masked EHR Modelling (MEM)), Dragonnet (and with CV-TMLE), and Targeted-BEHRT estimate RR on the fractional samples of the dataset. The point estimate is the mean value on five-fold cross validation and the error bars represent 95% confidence intervals for those point estimates of RR. B. Sum Absolute Error (SAE) across the seven subsamples of the dataset are shown for each model (Denoted by colour) is shown. The four models are represented by the four bars with interval defined by Standard Error (SE) and colour scheme is the same as part A. The figure was adapted from Rao et al[170].

As a trend, SAE across models began to converge as the dataset size increased as shown in Figure 6-4. Theoretically, as the number of samples increases, the finite-sample bias is mitigated, and hence, the performance of TARNET and derivations should be similar to those of models assisted by propensity modelling also noted by Shi et al[152].

Figure 6-5: Association of antihypertensives and incident cancer (fatal and non-fatal)



Application of Targeted-BEHRT on routine clinical data: Association of ACEI on incident cancer with respect to BBs, CCBs, Diuretics, and ARBs. We demonstrate Targeted-BEHRT with CV-TMLE risk ratio (RR) estimates with 95% Confidence Intervals (CI) on the Targeted-BEHRT model. In addition, we show empirical RR in the observational cohort selected for these experiments. The ground truth is assumed to be 1.0 (null) for all four associations validated by meta-analysis of RCTs. BBs: beta blockers; CCBs: calcium channel blockers; ACEIs: angiotensin-converting-enzyme inhibitors; ARBs: angiotensin receptor blockers; RR: risk ratio. The figure was adapted from Rao et al[170].

We applied the model on the routine clinical data study of association of ACEIs and incident cancer with respect to other antihypertensive drug classes and show the results in Figure 6-5. Across all four drug class comparisons, while the empirical RR often tended away from null implying a preventive or harmful effect, we showed that the model's 95% confidence interval for RR covered the null hypothesis (1.0 RR) across almost all drug class comparisons with exception of CCBs.

6.5 Interpretation

In this work, by utilizing large scale comprehensive EHR and deep learning methods, a model for observational causal inference was developed. The proposed model was validated against both statistical and deep learning models across six semi-synthetic experiments involving simulated confounding at various levels of strength (including finite-sample data experiments). Finally, we applied the model to a routine clinical data observational study to demonstrate ease of implementation in addition to the utility of the model in a case-study, in which all confounders are not explicitly given and perhaps not comprehensively measured (breakdown of strong ignorability).

Our work has contributions to the field of EHR-based deep learning research for causal inference and modelling generally. First, the Targeted-BEHRT model incorporates both static and temporal data embeddings into a unified embedding structure hence allowing adjustment over a spread of data types. Second, the model utilises a unique MEM unsupervised learning task combining MLM and VAE-based representation learning in tandem with the causal inference objective. The benefits of the unsupervised learning objective were clarified across multiple experiments as both TARNET and Targeted-BEHRT benefited with the MEM training. In our assessment, this is the first work conducting causal inference incorporating unsupervised learning on multiple EHR

data types. Third, CV-TMLE estimation correction was utilised for less biased RR estimation for deep learning causal models on EHR data. While MEM modelling for RR estimation was found to be especially useful in larger dataset sizes, we found that in the finite-sample estimation experiments, CV-TMLE is more critical for accurate RR estimation. Finally, we show that the model can be easily applied to test a clinical hypothesis regarding treatment effect in an observational setting.

Our work has some limitations and scope for future development. First and most fundamentally, while comprehensive EHR is very useful for observational studies, there is no guarantee of strong ignorability; in fact, realistically, confounders have been omitted in observational modelling. A variety of variables affecting outcome may be unadjusted (explicitly or through latent representation modelling) and further modality inclusion is necessary in future work to help mitigate residual confounding. While variational modelling strategies have been adopted in this work, latent confounding adjustment can be subsequently investigated in future works with more expressive latent variable modelling techniques to enrich EHR[183]. Also, in terms of encoding diseases in the model, the ICD-10 codes were used for encoding disease records. However, there are known issues of losing specificity of the code information when normalising to lower levels of ICD-10 (e.g., 3-character level). Future studies should investigate different resolutions of the ICD-10 encoding for inputting disease codes into Transformer models. In terms of data curation, we have allocated patients into an exposure group based on first prescription of class of antihypertensives. Subgroup investigations involving drug formulation, intensity, and duration of treatment should be additionally pursued in future studies. Other stratified analyses (e.g., sex, age, prior cardiometabolic disease) must also be actively pursued to perhaps identify heterogeneous treatment effect. In terms of applying the model to a case study, Targeted-BEHRT estimated null in most drug

comparisons in the routine clinical data study, but we note that the model finds the comparison to CCBs to deviate from the null (although quite proximal to the null hypothesis with <1.1 RR). While findings from the RCTs generally demonstrate that antihypertensives pose no effect on cancer, the evidence regarding CCBs is still conflicting and further research is required[169].

On the other hand, it must be noted that over-adjustment may also result in biased estimation. Although found to be an uncommon issue in observational modelling, M-structure bias variables (a special case of collider variables) might exacerbate estimation biases if included in adjustment; although in general, empirical studies have shown conditioning on all pre-treatment variables is still the optimal course of action[184,185]. The bias due to omission of confounders is found to be stronger than that due to over-adjustment[184,185]. However, further research must be conducted on the effect of complex variables such as M-structure variables specifically in the context of propensity-score modelling and the observational data setting.

7 ASSOCIATION ANALYSES IN AT-RISK PATIENTS

In this chapter, we apply the model, Targeted-BEHRT (see section 6.3.4) to better understand various associations in those with pre-existing conditions or multimorbidity. We have demonstrated in semi-synthetic experiments that Targeted-BEHRT is a powerful model that outperforms benchmark statistical and deep learning solutions for causal effect estimation on EHR data. Furthermore, the model's estimation of the association between antihypertensives and cancer was in line with meta-analyses of randomised evidence as well. Given the initial success of this model, I proceed to implement this model in observational cohorts, for which less is known about factors contributing to risk and protection. In this chapter, I implement the Targeted-BEHRT model in three analyses:

1. The association of systolic blood pressure and cardiovascular outcomes in patients with diabetes
2. The association of systolic blood pressure and cardiovascular outcomes in patients with COPD
3. The association of sodium-based paracetamol and cardiovascular outcomes, all-cause mortality, and systolic blood pressure as a continuous outcome in the elderly with respect to non-sodium-based formulations of the same.

The material in the first section has been organised in the manuscript titled *"Systolic blood pressure and cardiovascular risk in patients with diabetes: a prospective cohort study"* and has been published in *Hypertension* (doi.org/10.1161/hypertensionaha.122.20489). The material in the second section has been organised in the manuscript titled *"Systolic blood pressure, chronic obstructive pulmonary disease, and cardiovascular risk: a prospective cohort study"* and is under review at *Heart*. The material in the third and final section has been organised into a manuscript titled, "*Association of sodium-based paracetamol with changes in systolic blood pressure, cardiovascular events and all-cause mortality: a cohort study*" and has been submitted for publication. The manuscripts are a product of work by multiple authors. As first author on all three of the manuscripts, my role consisted of designing the study, conducting literature review, processing data, conducting statistical/deep learning analyses, and writing the first drafts of the manuscripts. Material from the manuscripts/publications (including figures, tables, and text) have been amended for presentation in the following sections.

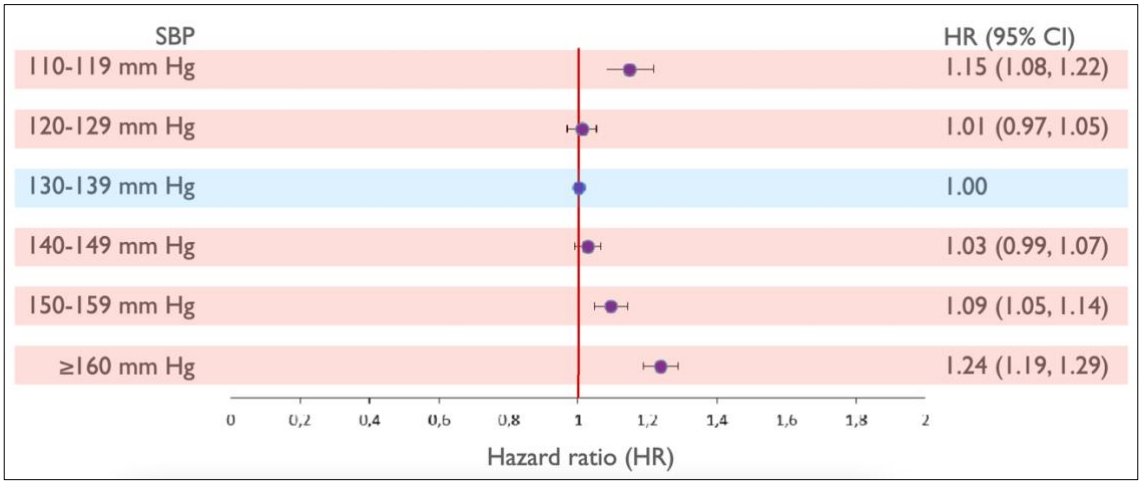
7.1 Systolic blood pressure, cardiovascular outcomes, and diabetes

7.1.1 Introduction

BP reduction is a well-known primary and secondary preventive strategy for cardiovascular events in addition to diabetes[15]. Observational studies conducted in the general cohort have suggested that the association between elevated BP and risk of major cardiovascular disease continuous or log-linear[186,187]. However, the association in patients with pre-existing cardiometabolic disorders is less well understood.

In people with diabetes, reports have been inconsistent in their conclusion about the nature of the association between SBP and risk of cardiovascular disease. In patients with known diabetes free of pre-existing cardiovascular conditions, prospective cohort studies have rendered conflicting conclusions. While some studies generally preserved the established log-linear relationship, others have concluded a "J-shaped" association in which, the lowest risk of cardiovascular events was at SBP between 135 and 139 mm Hg [16,188]. Moreover, in patients with diabetes with and without prior cardiovascular diseases, as seen in supplementary analyses conducted by Adamsson Eyrd et al., a clear J-shaped pattern is presented with a nadir of risk between 130 and 139 mm Hg for most cardiovascular outcomes (adapted below in Figure 7-1)[16]. This apparent discontinuous relationship has found some support from conventional meta-analyses of randomised controlled trials. For instance, in one study of BP-lowering in people with diabetes was found to increase the risk of cardiovascular death when the trial-average SBP was below 140 mm Hg [189]. However, previous observational studies using conventional statistical models are prone to reverse causation and uncontrolled confounding [15,16].

Figure 7-1: Association of systolic blood pressure and cardiovascular endpoints in patients with diabetes as conducted by previous studies



Association with composite of a composite of myocardial infarction, stroke, coronary heart disease, heart failure, and all-cause mortality estimated by Cox proportional hazards modelling and hazards ratio with 95% confidence interval (CI) given across the spectrum of systolic blood pressure (SBP). Nadir of risk illuminated with blue highlight. Figure redrawn from supplementary material of publication by Adamsson Eyrd et al[16]*.*

7.1.2 Aims

As a consequence, it has remained uncertain as to whether the effect of SBP on cardiovascular diseases in patients with pre-existing diabetes varies by baseline SBP. In this study, we applied Targeted-BEHRT to evaluate the relationship between SBP and cardiovascular events in a sample of 49,000 UK patients with diabetes using EHR data from the CPRD dataset.

7.1.3 Methods

We used retrospective anonymised EHR data from CPRD[5,131]. We used EHR from two data sources, primary care and secondary care (HES) within CPRD to identify a cohort of 49,000 individuals with diabetes: Those between 50 and 90 with at least one blood pressure measurement taken between the years 1990 and 2005 were included in this study with index date (baseline) being defined as the date of the first SBP measurement in this time period. We identified individuals as having diabetes at baseline using validated phenotyping methods[9,130,190]. Consistent with standard epidemiological studies, patients with heart failure before baseline were excluded from the study[15].

This cohort study followed the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guidelines.

7.1.3.1 Exposure

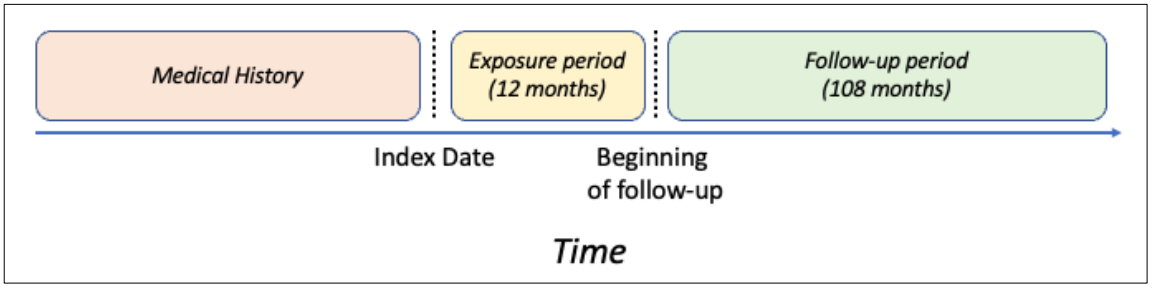
The exposure variable in this study was SBP and was derived from the CPRD measurements dataset. Blood pressure measurements are recorded by staff at the general practice (GP) during a visit/consultation[5]. In our study, we extracted SBP values and excluded measurements below 50 and above 300 mm Hg; this exclusion was conducted

based on recommendations by previously published methods to remove erroneity in measurements[191]. Following extraction and filtering, the exposure status for a patient was calculated as mean of the SBP measurements in the exposure period, defined as the first 12 months after index date. Patients were categorised into six exposure categories of this averaged measure of SBP over the course of the exposure period: ≤120 mm Hg (reference), 120–129 mm Hg, 130–139 mm Hg, 140–149 mm Hg, 150–159 mm Hg, and ≥ 160 mm Hg.

7.1.3.2 Outcomes

The primary outcome was fatal/non-fatal cardiovascular diseases defined as a composite of IHD, incident heart failure, stroke, and cardiovascular-related death. Secondary outcomes investigated in this study were individual components of the defined primary outcome (1) IHD, (2) incident heart failure, and (3) stroke. We identified cardiovascular events using three data sources in CPRD: (1) primary care, (2) secondary care (HES), and (3) the Office of National Statistics (cause-specific mortality) using previously published phenotyping algorithms [9]. Read codes were used to identify the conditions in the primary care setting while ICD-10 codes were used to identify cases in the secondary care and mortality setting. Follow-up period started one year from index date (i.e., following the exposure period). Events within 9 years of the follow-up period (i.e., between 12 and 120 months after index date) were captured for analysis; this feature of study design was incorporated to avoid conducting causal inference in the time period overlapping with the exposure period (i.e., the first 12 months following index date). Those who had events or left the study within the first 12 months following index date were removed from the analysis (Figure 7-2).

Figure 7-2: Study design: systolic blood pressure, cardiovascular endpoints, and diabetes



Study design of investigation of the association between systolic blood pressure (SBP) and cardiovascular outcomes in patients with diabetes. Index date (baseline) for a given patient is the date of the first SBP measurement recorded between 1990 and 2005 and ages 50 and 90.

7.1.3.3 Statistical and deep learning analyses

For analyses of the primary and secondary outcomes, the causal deep learning model, Targeted-BEHRT was implemented[170]. For inclusion of EHR variables in the Targeted-BEHRT model, we conducted data processing of raw CPRD data. First, diagnostic codes were harmonized by mapping Read codes from primary care and ICD-10 codes from hospital data to a total of 1,497 unique diagnostic codes[120]. Second, we mapped CPRD medication codes in the "Product code" format to 386 codes in the BNF coding format [130]. Third, we derived the variable, smoking status (current, former, never a smoker) identified by last known status in the 12 months before baseline. Fourth and finally, patient sex was extracted from CPRD for Targeted-BEHRT modelling. The sex/smoking status were included as static variables for modelling. Five-fold cross validation was used for the training/testing and evaluation of association. As necessary for the Targeted-BEHRT modelling approach, initial estimates were computed on patients from test set of each of the five iterations. Second, equipped with these initial estimates, we "updated" risk estimates utilising "doubly-robust" post-hoc estimator, CV-TMLE, to further mitigate biases[181]. RR and 95% CI were derived from this post-hoc

estimation procedure. More details on the Targeted-BEHRT model can be found in section 6.3.4.

To compare against the Targeted-BEHRT framework, logistic regression (LR) models were implemented to investigate the studied association. The exposure, SBP group was included as a categorical variable and adjusted for age, sex, smoking status at baseline, BMI at baseline, atrial fibrillation, chronic kidney disease, antihypertensive use at baseline, high density lipoprotein (HDL) at baseline, total cholesterol (TC) at baseline, and triglycerides (TG) at baseline. Baseline BMI, TC, HDL, and TG were calculated as the average of measurements in the 12 months before baseline. Antihypertensives were identified by BNF code[171]. Smoking status (current, former, never a smoker) was identified by last known status in the 12 months before baseline. Chronic kidney disease and atrial fibrillation were identified with established phenotyping algorithms for CPRD[9]. To ensure a fairer comparison to the deep learning model, we conducted imputation before inputting data to LR models. Multiple imputation using chained equations was used to impute missing variables BMI, HDL, TC, TG, and smoking status; 25 imputations were conducted. For the LR, an estimate for the RR was obtained utilising direct estimation[61]. We calculated RR as the average across the test sets of k-fold cross validation (k=5) and calculated 95% CI over the five runs[158]. In addition to the LR model, the crude risk ratio was calculated as the average risk of the outcome in a particular exposure group divided by the average risk of the same outcome in the reference exposure group. This crude measure is an unadjusted measure.

Additionally, seven sensitivity analyses were conducted using Targeted-BEHRT. First and second respectively, we conducted sex (male and female) and baseline age-stratified (≤75 and 75 years of age) analyses. Third, since antihypertensives have been shown to be preventative for various cardiovascular events, we restricted the analyses of

the primary outcome to patients who had not taken antihypertensives during the exposure or follow-up period since treatment can dilute association[192]. Fourth and fifth respectively, to assess the possible impact of reverse causality, we excluded individuals who had cardiovascular events in the first 12 and 24 months of follow-up period. Sixth, we conducted stratified analysis of the primary outcome by baseline antihypertensive use to better understand association in the context of antihypertensive treatment. Seventh, we analysed the primary outcome including patients who left the study or had an event during the exposure period (Figure 7-2).

7.1.4 Results

7.1.4.1 Population statistics

A total of 49,000 patients with diabetes were included in this study (Figure S1). Demographic and baseline characteristics of patients in the exposure groups are provided in Table 7-1. The median follow-up from baseline was 7.3 with lower baseline SBP having higher prevalence of baseline IHD. On the other hand, patients in higher SBP exposure groups had greater antihypertensive use at baseline. BMI at baseline generally indicated an overweight cohort across all exposure groups. 45% of the individuals in the cohort were women and 39% current or former smokers. Further analyses of event rates are given in Table S8.

Table 7-1: Characteristics for patients with diabetes

	Categories of baseline systolic blood pressure					
	<120 mm Hg	**120-129 mm Hg**	**130-139 mm Hg**	**140-149 mm Hg**	**150-159 mm Hg**	**≥160 mm Hg**
No. (%)	2706 (5.5)	5881 (12.0)	10793 (22.0)	12203 (24.9)	8704 (17.8)	8713 (17.8)
Follow-up, yrs (IQR)	6.9 (2.5-9.0)	7.8 (3.1-9.0)	7.5 (3.2-9.0)	7.5 (3.2-9.0)	7.3 (3.0-9.0)	6.5 (2.8-9.0)
Age, yrs (IQR)	60.0 (52.0-70.0)	61.0 (53.0-70.0)	63.0 (55.0-72.0)	65.0 (56.0-73.0)	66.0 (58.0-74.0)	69.0 (61.0-76.0)
Women (%)	1045 (38.6)	2309 (39.3)	4466 (41.4)	5350 (43.8)	4048 (46.5)	4638 (53.2)
YOB (IQR)	1941 (1931-1948)	1940 (1931-1948)	1938 (1929-1947)	1936 (1928-1945)	1934 (1926-1942)	1931 (1924-1939)
BMI, kg/m2 (IQR)†	29.3 (26.8-31.1)	29.6 (27.3-31.4)	29.5 (27.4-31.5)	29.5 (27.5-31.5)	29.4 (27.5-31.3)	29.1 (27.4-31.0)
HDL, mmol/L (IQR)†	1.3 (1.2-1.4)	1.2 (1.2-1.4)	1.3 (1.2-1.4)	1.3 (1.2-1.4)	1.3 (1.2-1.4)	1.3 (1.2-1.4)
TG, mmol/L (IQR)†	2.2 (1.6-2.6)	2.2 (1.7-2.6)	2.2 (1.7-2.6)	2.2 (1.8-2.6)	2.2 (1.8-2.6)	2.1 (1.8-2.5)
TC, mmol/L (IQR)†	5.1 (4.8-5.5)	5.1 (4.8-5.5)	5.1 (4.8-5.5)	5.2 (4.9-5.5)	5.2 (4.9-5.5)	5.3 (5.0-5.5)
Smoking status†:						
Current/former smoker (%)	1234 (45)	2604 (44)	4479 (41)	5005 (41)	3172 (36)	2928 (33)
Never smoker (%)	1472 (54)	3277 (55)	6314 (58)	7198 (58)	5532 (63)	5785 (66)
Disease at baseline:						
IHD (%)	409 (15.1)	763 (13.0)	1235 (11.4)	1219 (10.0)	870 (10.0)	823 (9.4)
Chronic kidney disease (%)	30 (1.1)	76 (1.3)	125 (1.2)	134 (1.1)	103 (1.2)	82 (0.9)
Stage 1 and 2 kidney disease (%)	37 (1.4)	67 (1.1)	123 (1.1)	143 (1.2)	79 (0.9)	103 (1.2)
Atrial fibrillation (%)	122 (4.5)	194 (3.3)	320 (3.0)	315 (2.6)	232 (2.7)	207 (2.4)
Medications at baseline:						
Antihypertensive use(%)	1036 (38.3)	2556 (43.5)	5459 (50.6)	6771 (55.5)	5045 (58.0)	5333 (61.2)

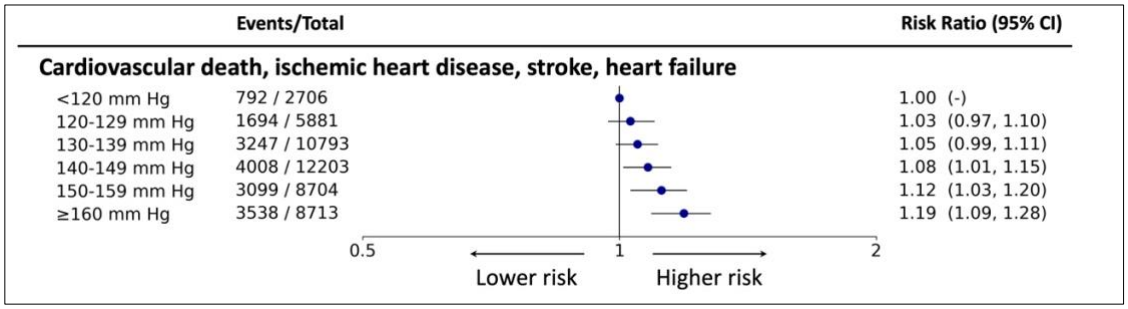
Values presented are median with interquartile range (IQR) or percentage (%) YOB: year of birth; BMI: body mass index; HDL: high density lipoprotein; TC: total cholesterol; TG: triglycerides; IHD: ischaemic heart disease; SD: standard deviation; %: percentage. †The percentage of missing variables – BMI (55.1%), smoking status (29.2%), HDL (75.3%), TC (49.9%), TG (68.7%).

7.1.4.2 Association analyses

RR estimates from adjusted Targeted-BEHRT model demonstrated a rise in the risk of cardiovascular events with a rise in SBP categories (Figure 7-3). The crude and LR estimates of RR both depicted a J-shaped pattern (Figure S2). Compared with

reference SBP group (<120 mm Hg), the adjusted LR model demonstrated a nadir of risk at SBP between 130 and 139 mm Hg.

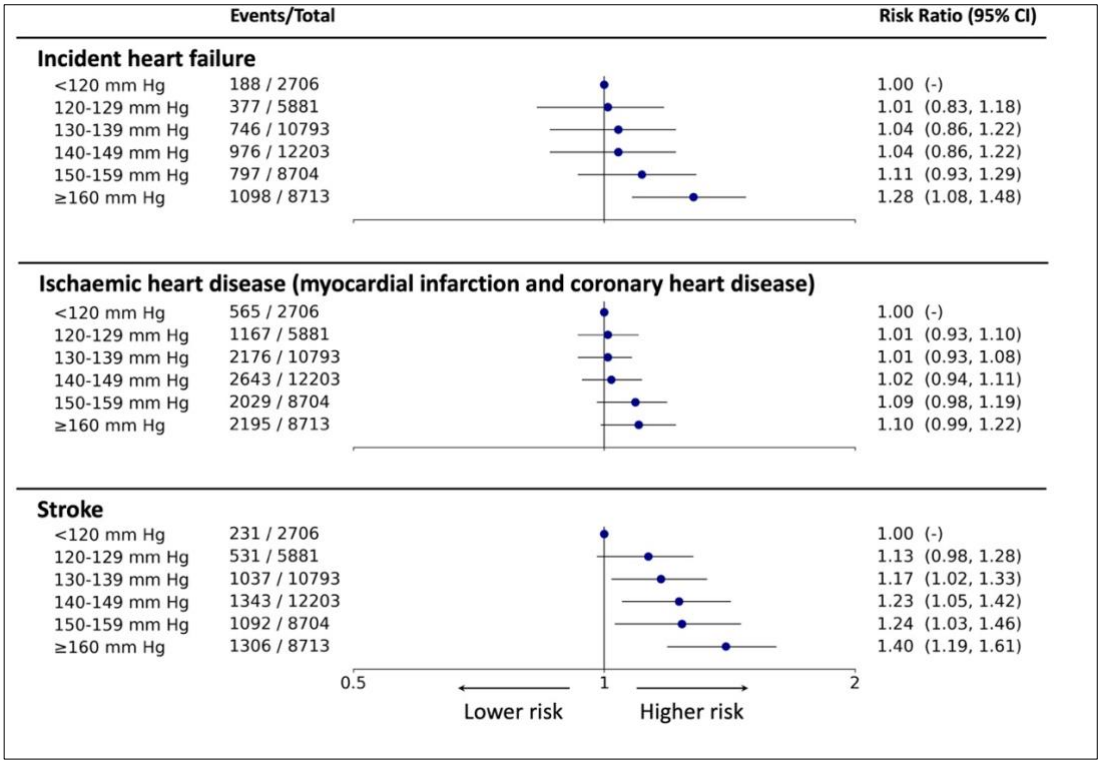
Figure 7-3: Association with primary composite outcome in patients with diabetes



Forest plot of risk ratio estimates of the Targeted-BEHRT model with 95% confidence intervals (CI) for association of systolic blood pressure and the primary composite outcome. From the left, the six exposure groups are shown in first column. Number of events and total number of patients in each exposure group is shown in second column. The forest plot and corresponding risk ratio estimates are shown in the right-most column relative to reference class, <120 mm Hg. The forest plot is plotted in logarithmic scale. For all estimates for reference class, there is no confidence interval.

Estimates from the Targeted-BEHRT model investigating the association between SBP and secondary outcomes showed that the lowest risk of all secondary cardiovascular outcomes was observed at <120 mm Hg (Figure 7-4). On the other hand, both crude and adjusted LR modelling estimated that SBP between 130 and 140 mm Hg exhibited lowest risk of the secondary outcomes: incident heart failure and IHD (Figure S3). However, specifically for the investigation of stroke, the crude and adjusted LR models estimated that <120 mm Hg SBP demonstrated lowest risk similar to the Targeted-BEHRT model. Lastly, all models showed that patients with ≥160 mm Hg exhibited highest risk of secondary outcomes.

Figure 7-4: Association with secondary outcomes in patients with diabetes

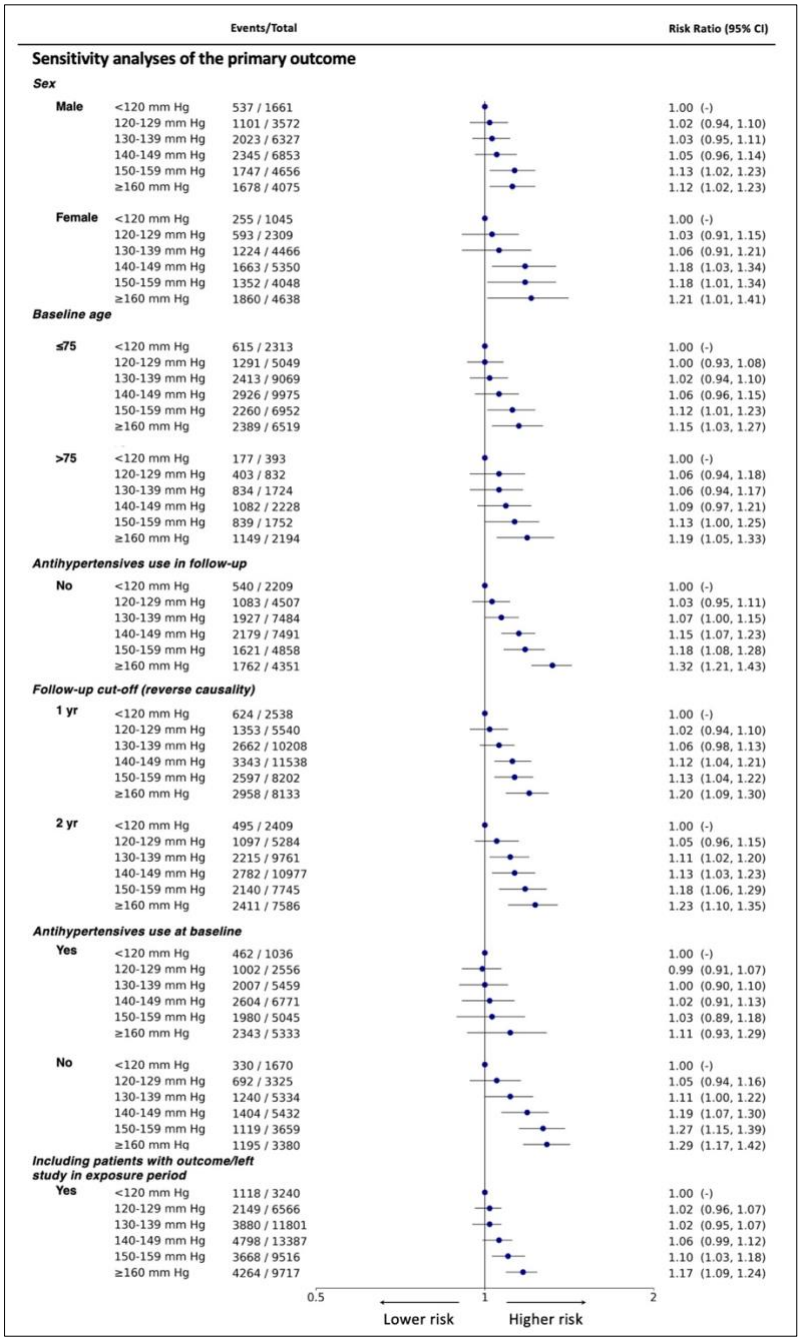


Forest plot of risk ratio estimates of the Targeted-BEHRT model with 95% confidence intervals (CI) for association of systolic blood pressure and secondary outcomes. From the left, the six exposure groups are shown in first column. Number of events and total number of patients in each exposure group is shown in second column. The forest plot and corresponding risk ratio estimates are shown in the right-most column relative to reference class, <120 mm Hg. The forest plot is plotted in logarithmic scale. For all estimates for reference class, there is no confidence interval.

Shown in Figure 7-5, the sensitivity analyses investigating the association between SBP and the primary composite outcome using the Targeted-BEHRT model preserved the trend found in the main analysis. In both sex and age-stratified analyses, the log-linear trend across the spectrum of SBP was generally preserved. In analysis of patients who have not taken antihypertensives during the exposure and follow-up periods, the RR estimates and corresponding 95% CI for each exposure group was slightly higher than their counterparts in the main analysis but overall mirrored the trend in the main analyses. Furthermore, excluding patients who had events in the first 12 and 24 months of follow-up also captured a similar trend as that of the main analysis. Stratifying by

antihypertensive usage at baseline, albeit the slight presence of a local minimum, the trend showed little material difference from the main result. Lastly, incorporating those who dropped out during the first 12 months following baseline, the trend presented was similar to that in the main analysis.

Figure 7-5: Sensitivity analyses of association with primary outcome in patients with diabetes



Forest plot of risk ratio estimates of the Targeted-BEHRT model with 95% confidence intervals (CI) for association of systolic blood pressure and primary outcomes in sensitivity analyses. The particular sensitivity analysis is italicised on the left with strata indented. From the left, the six exposure groups are shown in first column. Number of events and total number of patients in each exposure group is shown in second column. The forest plot and corresponding risk ratio estimates are shown in the right-most column relative to reference class, <120 mm Hg. The forest plot is plotted in logarithmic scale. For all estimates for reference class, there is no confidence interval.

7.1.5 Interpretation

In this study, using deep learning modelling on a comprehensive dataset of UK EHR in a cohort of 49,000 individuals with diabetes, we found SBP to be monotonically associated to cardiovascular risk. Patients with SBP below 120 mm Hg exhibited lowest risk of both the primary and secondary outcomes. Furthermore, the results from the sensitivity analyses were inconsequentially different from the main analyses.

Conventional statistical models are usually implemented in observational studies with curated, low-risk cohorts relatively free of multimorbidity at baseline. By ensuring the cohort is healthy, statistical models need to include a handful of established confounding variables and deliver adjusted and trustworthy estimates of causal effect. However, in cohorts like ours with high baseline BMI and a host of underlying conditions – including some cardiovascular in nature indicating prevalent multimorbidity – traditional models such as logistic regression might be insufficient for adjustment of confounding variables.

Our work implementing deep causal modelling addresses the issues and fills some of the gaps in this research of SBP and cardiovascular outcomes. With Targeted-BEHRT modelling, our research found no evidence of the J-shaped association between SBP and primary and secondary outcomes. While a recent individual-participant data meta-analysis of randomized evidence has indirectly dismissed the existence of a J-shaped in patients with or without prior cardiovascular disease[15], evidence of heterogeneity of

treatment effects in patients with diabetes has been controversial. In particular, a tabular meta-analysis of randomised trials in patients with diabetes has suggested that BP lowering might increase the risk of cardiovascular death when SBP is below 140 mm Hg. Our observational study provides complementary evidence against a potentially harmful effect of BP lowering on cardiovascular outcomes across a wide range of baseline SBP categories; moreover, in line with the findings of the work by Nazarzadeh et al, our work presents complementary, independent support for "the lower, the better" paradigm of hypertension management – especially in those with pre-existing conditions[14].

This study further shows that the Targeted-BEHRT framework requires high-dimensional longitudinal data. The Targeted-BEHRT estimator can function optimally with access to multiple EHR modalities (diagnoses, medications, etc) and associated temporal annotation (age, calendar year). Furthermore, the original methods paper demonstrated that when such rich data are provided the unsupervised process (masked EHR modelling) works well in reducing bias in estimation. When data are sparse or limited (finite sample), the doubly-robust estimation, with known benefits for finite-sample estimation, more accurately estimates RR than the variants without utilisation of doubly-robust estimation[181]. Furthermore, when positivity (overlap) between exposure groups is limited, the Targeted-BEHRT model fares better than other benchmark models[61]. Specifically, in our study of patients with cardiometabolic multimorbidity – where the traditional logistic regression model insufficiently adjusts for confounders, deep learning modelling in tandem with semi-parametric methods can be appropriately implemented for robust RR estimation. Further testing to assess model estimation accuracy is needed for settings in which the exposure is poorly defined (e.g., "generic painkiller") or when the outcome is categorical or continuous. Future confirmatory

investigations on additional data sources would be invaluable for studying the association from different perspectives.

We note in our cohort of high-risk patients, relatively modest associations were observed using the Targeted-BEHRT approach. This, however, was also the case in our conventional modelling and comparable with previous research[16]. The weak associations might be due to two main reasons. First, as shown by meta-analysis of randomised evidence, BP lowering's effect on cardiovascular outcomes was half as effective in people with diabetes as compared to those without diabetes [14]. This might be due to the fact that the prevention of diabetes itself may be mediating a part of the effect of BP lowering on CVD outcomes – a pathway that might not be relevant to people with pre-existing diabetes[44]. Second, our cohort included patients with several comorbidities and prescriptions in medical history. While all these attributes are typical in diabetes patients (and a suitable case for complex models such as Targeted-BEHRT), this manifestation of complex baseline health could lead to attenuation of the association with cardiovascular disease as compared to the association in lower risk cohorts [16,149]. Sensitivity analyses support this claim, where for example, exclusion of patients with use of antihypertensives at baseline or during follow-up led to stronger associations between SBP and risk of CVD.

In parallel, in comparison to past works, the general *strength* of associations from the supplementary analyses by Adamsson Eyrd et al. was more-or-less in line with our findings even though our work dismisses the trend demonstrated by the same work; including patients with cardiovascular diseases at baseline, the analyses presented effect sizes generally more diluted than those in patients free of cardiovascular diseases at baseline [16]. In sum, while effect sizes are modest, the log-linear pattern captured is in line with established understanding of the relationship between SBP and cardiovascular outcomes in the general population.

Previous observational studies investigating this association of SBP and cardiovascular events in diabetic patients have rendered conflicting conclusions using conventional statistical modelling. However, in totality, they have remarked upon a larger phenomenon: the shape of the association as a function of degree of "ignorability" (i.e., the "unconfoundedness" of estimation). Simple confounding adjustment of baseline age, sex, and demographic variables falls short of eliminating the J-shaped association in cohorts with diabetes free of cardiovascular diseases as shown in past studies such as the NDR-BP-II study with 54,000 patients and the ROSE study with 34,000 patients [188,193]. Extending the predictor set to include key cardiovascular risk factors better modelled the association as compared to the former two studies and thus rejected the J-shaped association in a cohort of 187,000 patients [16]. However, in the same study, in extended supplementary analyses including patients with pre-existing cardiovascular conditions, the conventional modelling approach failed to reject the J-shaped relationship for all outcomes, save those with outcome, stroke[16]. Similarly, in our direct implementation, we found that conventional crude and adjusted statistical modelling exhibited evidence of the J-shaped relationship in studies of the primary outcome and the secondary outcomes of incident heart failure and IHD.

These results have important implications for cardiovascular research and the clinical community. Many hypertension guidelines have changed their stance for recommended SBP in diabetic patients [194–197]. The current guidelines actively advocate a BP lowering treatment goal of <130 mm Hg as opposed to <120 mm Hg in those with concomitant diabetes. In elderly patients (>65 years), the recommendation is currently a SBP goal of <140 mm Hg [195,197].

Our analysis provides some clarity concerning the relationship between SBP and cardiovascular events in patients with diabetes. In parallel, a recent individualised patient

data meta-analyses of randomised studies investigating the effect of blood pressure lowering interventions on cardiovascular endpoints in diabetic patients concluded while effect sizes are diluted in the diabetic population with respect to those free of diabetes, there is indeed benefit of blood pressure lowering to diabetic patients across the spectrum of blood pressure[14]. In defence of "the lower, the better" paradigm for blood pressure in diabetic patients, the conclusions from the meta-analysis are consistent with ours. Hence, the results from both, our study and meta-analysis, independently illuminate that SBP lower than current guidelines might be justified for further reducing the risk of cardiovascular events.

In terms of strengths and limitations, our first strength is the large sample size that includes patients from a representative, administrative primary care database. The size and linkage capabilities of CPRD allow us to extract individuals eligible for our research. In addition to breadth, CPRD allows us a host of rich diagnosis, medication, and measurements variables. With repeat measurement data available, we were able to leverage a summary metric (average) of multiple measurements thereby mitigating issues of measurement error. Lastly, unlike previous studies that have excluded various stratifications of the population, we did not exercise strict exclusion on the cohort. Unlike previous studies of SBP and cardiovascular events that have excluded older aged individuals, patients between 50 and 90 years of age were included in this work. Exclusion based on baseline attributes was limited; only those with heart failure were excluded at baseline.

Second, deep learning feature extraction was used to automatically adjust for confounding variables and latent interactions in the input data. We also implemented conventional statistical approaches enabling direct comparison of Targeted-BEHRT to established methods in the study of SBP and cardiovascular outcomes. We showed that

by using superior adjustment methods, we can more effectively model the association in observational data, thereby refuting the J-shaped argument.

This study also has some limitations. EHR data might have some level of measurement error or misclassification. Despite evidence of the validity of diagnoses in the CPRD dataset[198], ascertainment of diabetes might have some possible misclassifications (e.g., metabolic syndromes and related disorders as opposed to diabetes). Furthermore, measurement errors are a natural issue with EHR data, especially that of BP, but attempts have been made to mitigate these issues in the case of SBP data by taking an average of multiple measurements over the course of 12 months following baseline. Survival modelling can alternatively be used to better address issues of capturing CVD risk in the time-to-event setting (i.e., more comprehensive accounting of censored data); currently, although, classical survival models have been extended to several deep learning variants, further methodological advancement is needed to appropriately alter the model for causal/association estimation. Lastly, as is the case for all observational research, residual confounding cannot be ruled out; the model, Targeted-BEHRT cannot fully capture all confounding variables with its adjustment processes. Further randomised investigations, given sufficient sample size, is the optimal way to validate findings captured in the observational setting.

7.2 Systolic blood pressure, cardiovascular outcomes, and COPD

7.2.1 Introduction

In subgroups with COPD, the association of SBP with cardiovascular outcomes is less well understood. Independently, SBP and COPD have both been associated with a higher risk of cardiovascular disease [15,37,199,200]. However, there is a dearth of evidence when it comes to conclusively understanding the effect of SBP on cardiovascular

endpoints in patients with COPD. A J-shaped association between SBP and cardiovascular events was found in a previous observational analysis using traditional statistical modelling in patients with COPD who had or had not previously developed cardiovascular disease [35]. However, observational studies utilising conventional statistical modelling might be limited in investigating this question. The adjusted variables need to be manually chosen, naturally exposing models to issues of residual confounding. Additionally, in subgroups of patients with multiple comorbidities at baseline and a large number of complicated factors of risk and prevention, confounding factors are lesser understood; as a result, conventional statistical models with insufficient adjustment can result in confounded or spurious J-shaped associations [15,34,201].

7.2.2 Aims

As a consequence, it has remained uncertain as to whether the effect of SBP on cardiovascular diseases in patients with pre-existing COPD varies by baseline SBP. In this study, we applied Targeted-BEHRT to evaluate the relationship between SBP and cardiovascular events in a sample of 39,602 UK patients with COPD using EHR data from the CPRD dataset.

7.2.3 Methods

We used prospectively collected EHR data from CPRD. We used EHR data from three resources in which we identified a cohort of 39,602 individuals with prevalent diabetes: primary care, secondary care (Hospital Episode Statistics), and the Office of National Statistics (cause-specific mortality). We included people between 55 and 90 years of age with at least one BP measurement taken between the years 1990 and 2009. The baseline was defined as the date of the first BP measurement. We identified individuals as having COPD at baseline using validated phenotyping methods[9].

The REporting of studies Conducted using Observational Routinely-collected Data (RECORD) reporting guidelines were followed for this cohort study.

7.2.3.1 Exposures

SBP was the exposure variable and was derived from CPRD measurement data. In CPRD, BP measurement is recorded by GP staff during an in-person visit or consultation[5]. The European Society of Cardiology (ESC) guidelines recommend three BP measurements measured 1–2 min apart with BP recording as the average of the last two BP readings[202]. In CPRD, the GPs follow the same approach but a single BP measurement is recorded from each visit[5]. We excluded the SBP values below 50 and above 300 mm Hg as suggested by previously published phenotyping methods to exclude outlier measurements[191]. All analyses were conducted with exposure status calculated as mean of SBP measurements in the first 12 months after baseline (i.e., exposure period). For example, for a hypothetical individual with four measurements in the first 12 months following baseline, the exposure would be the mean value of the four measurements. Patients were categorised into six exposure categories of SBP: less than 120 mm Hg, 120–129 mm Hg (reference), 130–139 mm Hg, 140–149 mm Hg, 150–159 mm Hg, and greater than or equal to 160 mm Hg.

7.2.3.2 Outcomes

The primary outcome was fatal or non-fatal cardiovascular disease, defined as a composite of IHD, heart failure, stroke, and cardiovascular death. Secondary outcomes were components of the primary outcome: (1) IHD, (2) heart failure, and (3) stroke. All outcomes were identified by Read codes (primary care) and ICD-10 codes (secondary care and mortality data) as reported previously[9]. Follow-up period started 12 months after baseline; this was done in order to avoid conducting inference within the exposure period (first 12 months after baseline). Thus, events that occurred between 12 and 72 months

after baseline (e.g., 60 months or 5 years of follow-up period) were captured for analysis and patients who had a cardiovascular outcome (i.e., event of HF, IHD, stroke, or cardiovascular death) or left the study within the first 12 months following baseline were removed from the analyses.

7.2.3.3 Statistical and deep learning analyses

For the deep learning approach, we used Targeted Bidirectional Electronic Health Records Transformer (Targeted-BEHRT) for risk ratio (RR) estimation of association between SBP and cardiovascular outcomes with SBP of 120-129 mm Hg considered as reference group[170]. For each of these comparisons, Targeted-BEHRT was first trained to jointly predict exposure category (propensity score) and risk of outcome with five-fold cross-validation implemented for training and testing. The Targeted-BEHRT model adjusted for the diagnoses/ medication in medical history prior to baseline in addition to baseline smoking status (current, former, never a smoker) – identified by last known status in the 12 months before baseline – and sex. RR and 95% CI were derived from the CV-TMLE post-hoc estimation procedure. More details and implementation of the Targeted-BEHRT approach can be found in section 6.3.4.

In order to compare the deep learning approach against established statistical modelling, logistic regression (LR) was implemented to estimate the studied association. The SBP exposure group was included as a categorical variable. Since we motivated the work with findings from the research conducted by Byrd et al, we adjusted for the same variables as those chosen in their research: sex, age, BMI, smoking status (current, former, never a smoker), BB use, long-acting beta agonist (LABA) use, and inhaled corticosteroid use [35]. In a second LR model with an expanded set of predictors including known cardiovascular risk factors, we additionally adjusted for TG, low density lipoprotein (LDL), TC, atrial fibrillation, rheumatoid arthritis, severe mental illness

(psychosis, schizophrenia, or bipolar disorder), chronic kidney disease, and diabetes. Diagnoses and medication use were identified using validated phenotyping algorithms [9,171,203]. For BMI, TC, TG, and LDL, average of the measurements recorded in the 36 months before index date were computed to minimise issues of random measurement error [15,204]. We conducted imputations on missing variables to ensure fairer comparison with the deep learning approach. Multiple imputations using chained equations were implemented (15 imputations) to impute the continuous and categorical missing variables: BMI, TC, TG, LDL, and smoking status. Estimation of RR was conducted using the direct standardisation method [61]. More details on LR modelling and tutorial of the direct estimation can be found in Supplementary section 9.4.2.1. Crude RR was also estimated to provide the unadjusted estimate of the association with primary and secondary outcomes.

Four sensitivity analyses were pursued in our studies using the Targeted-BEHRT model. First, we investigated the effect of SBP on cardiovascular risk in patients who had not taken antihypertensives during the exposure/follow-up period. Antihypertensives are established medications for lowering high blood pressure thereby potentially attenuating cardiovascular risk; thus, we conducted this sensitivity analysis in order to investigate the undiluted association between SBP and cardiovascular outcomes in COPD patients [36]. Second, to investigate the effects of time period, we limited the investigation to only include those with index date after January 1 2001. Third and fourth, to mitigate issues of reverse causality, we investigated the primary outcome excluding individuals who had cardiovascular events in the first 12 and 24 months of the follow-up period respectively.

7.2.4 Results

7.2.4.1 Population statistics

39,602 individuals with COPD at baseline were included in our analysis. The median follow-up time was 3.9 years (interquartile range [IQR]: 1.5-5.0) with 10,987 events, and the median age at baseline, 69 years (IQR: 60-76) shown in Table 7-2. Patients with lower SBP had a higher percentage of atrial fibrillation, chronic kidney disease, and IHD and were more likely to be current smokers at baseline. Also, patients with lower SBP had more clinical encounters (medications and diagnoses) recorded in GP/secondary care. However, individuals with a higher SBP had a higher percentage of antihypertensive usage.

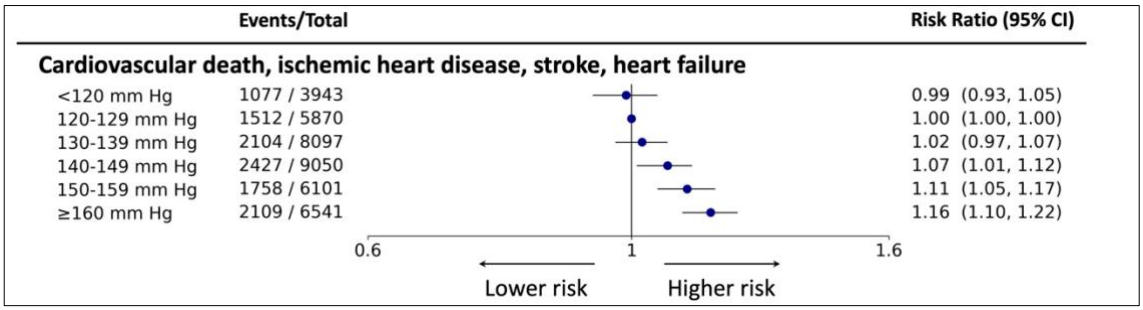
Table 7-2: Characteristics for patients with COPD

	Categories of systolic blood pressure					
SBP categories	**<120 mm Hg**	**120-129 mm Hg**	**130-139 mm Hg**	**140-149 mm Hg**	**150-159 mm Hg**	**≥160 mm Hg**
No. (%)	3943 (10.0)	5870 (14.8)	8097 (20.4)	9050 (22.9)	6101 (15.4)	6541 (16.5)
Follow-up, yrs (IQR)	3.4 (1.1-5.0)	3.0 (1.5-5.0)	4.0 (1.6-5.0)	4.0 (1.6-5.0)	4.0 (1.5-5.0)	3.7 (1.5-5.0)
Age, yrs (IQR)	66.0 (58.0-75.0)	67.0 (58.0-75.0)	68.0 (59.0-76.0)	69.0 (61.0-77.0)	70.0 (62.0-77.0)	72.0 (65.0-78.0)
Women (%)	1892 (48.0)	2695 (45.9)	3774 (46.6)	4179 (46.2)	2765 (45.3)	3142 (48.0)
YOB (IQR)	1937 (1927-1945)	1936 (1927-1945)	1935 (1926-1944)	1933 (1925-1942)	1931 (1924-1940)	1927 (1921-1935)
BMI† kg/m² (IQR)	25.7 (24.0-27.0)	26.0 (24.4-27.2)	26.0 (24.7-27.4)	25.9 (24.7-27.2)	25.9 (24.8-27.1)	25.6 (24.7-26.8)
LDL†, mmol/L (IQR)	3.1 (2.9-3.2)	3.1 (2.9-3.2)	3.1 (2.9-3.2)	3.1 (3.0-3.2)	3.1 (3.0-3.2)	3.1 (3.0-3.2)
TG†, mmol/L (IQR)	1.6 (1.4-1.8)	1.6 (1.4-1.8)	1.6 (1.4-1.8)	1.6 (1.4-1.8)	1.6 (1.4-1.8)	1.6 (1.4-1.7)
TC†, mmol/L (IQR)	5.3 (5.0-5.7)	5.3 (5.0-5.6)	5.3 (5.0-5.6)	5.3 (5.0-5.6)	5.3 (5.0-5.6)	5.3 (5.0-5.6)
Smoking status†:						
Current smoker %	1960 (49)	2831 (48)	3627 (44)	4074 (45)	2728 (44)	3049 (46)
Former smoker %	1453 (36)	2131 (36)	3148 (38)	3490 (38)	2336 (38)	2307 (35)
Never smoker (%)	530 (13)	908 (15)	1322 (16)	1486 (16)	1037 (16)	1185 (18)
Disease at baseline:						
IHD (%)	711 (18.0)	872 (14.9)	1131 (14.0)	1085 (12.0)	650 (10.7)	670 (10.2)
CKD (%)	41 (1.0)	36 (0.6)	38 (0.5)	51 (0.6)	30 (0.5)	34 (0.5)
Diabetes (%)	268 (6.8)	477 (8.1)	684 (8.4)	618 (6.8)	390 (6.4)	300 (4.6)
Severe Mental illness (%)	47 (1.2)	62 (1.1)	54 (0.7)	46 (0.5)	29 (0.5)	33 (0.5)
Atrial fibrillation (%)	290 (7.4)	319 (5.4)	397 (4.9)	386 (4.3)	220 (3.6)	225 (3.4)
Medications at baseline:						
Antihypertensive (%)	1283 (32.5)	1898 (32.3)	2851 (35.2)	3273 (36.2)	2337 (38.3)	2444 (37.4)
IC (%)	2221 (56.3)	3214 (54.8)	4557 (56.3)	5280 (58.3)	3617 (59.3)	3874 (59.2)
LABA (%)	637 (16.2)	873 (14.9)	1271 (15.7)	1263 (14.0)	756 (12.4)	602 (9.2)

Values presented are median with interquartile range (IQR) or percentage (%). YOB: year of birth; Yrs: years; BMI: body mass index; IHD: ischaemic heart disease; CKD: chronic kidney disease; LABA: long-acting beta agonists; IC: inhaled corticosteroids; TC: total cholesterol; TG: triglycerides; LDL: low density lipoprotein. †Percentage of missing variables – BMI (56.3%), smoking status (24.4%), TC (71.7%), TG (80.7%), LDL (85.6%).

7.2.4.2 Association analyses

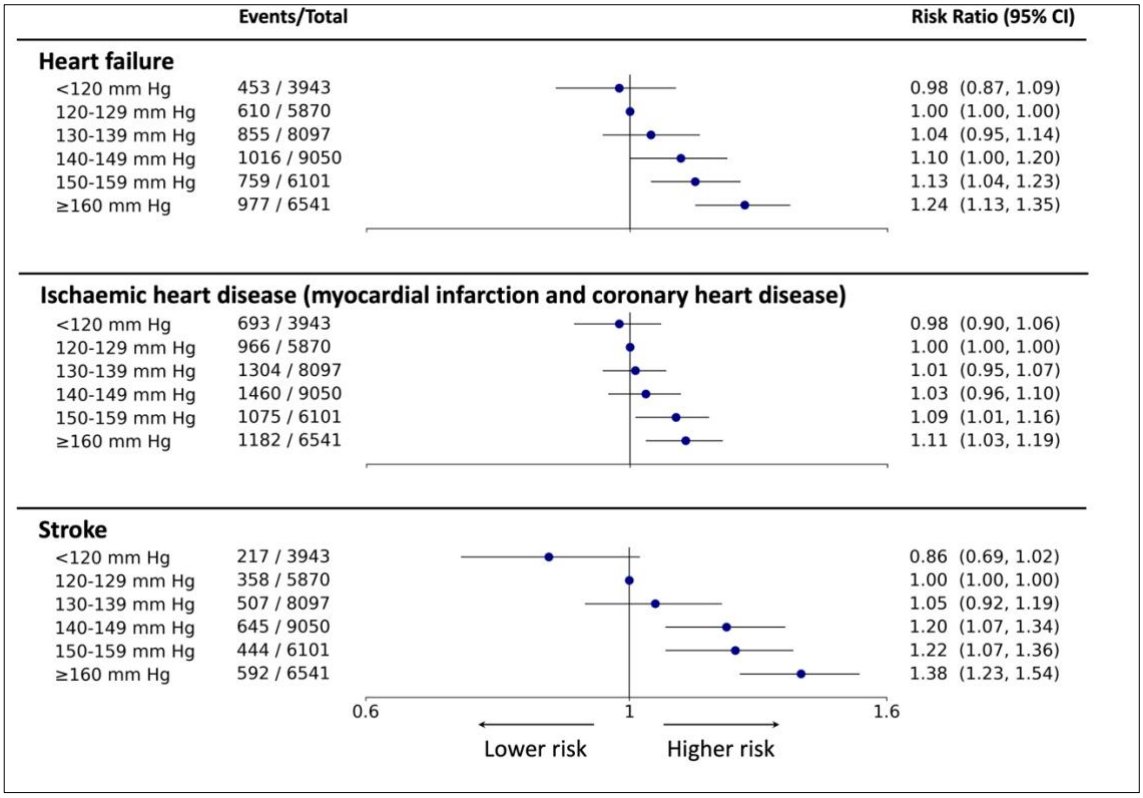
Figure 7-6: Association with primary outcome in patients with COPD



Forest plot of risk ratio estimates of the Targeted-BEHRT model with 95% confidence intervals (CI) for association of systolic blood pressure and primary outcome. From the left, the six exposure groups are shown in first column. Number of events and total number of patients in each exposure group is shown in second column. The forest plot and corresponding risk ratio estimates are shown in the right-most column relative to reference class, 120-129 mm Hg. The forest plot is plotted in logarithmic scale. For all estimates for reference class, there is no confidence interval.

The Targeted-BEHRT model estimated a continuous relationship between SBP and the primary outcomes in patients with COPD (Figure 7-6). By contrast, the crude and adjusted LR estimates of RR both demonstrate a nadir of risk at SBP between 130 and 139 mm Hg (Figure S4). The adjusted LR model with expanded set of predictors demonstrated similar trends as compared to the base adjusted LR model (i.e., predictors defined in Byrd et al.) for the analysis of the primary outcome (Figure S5)[35]. All models found that ≥160 mm Hg demonstrated greatest risk of cardiovascular events.

Figure 7-7: Association with secondary outcomes in patients with COPD

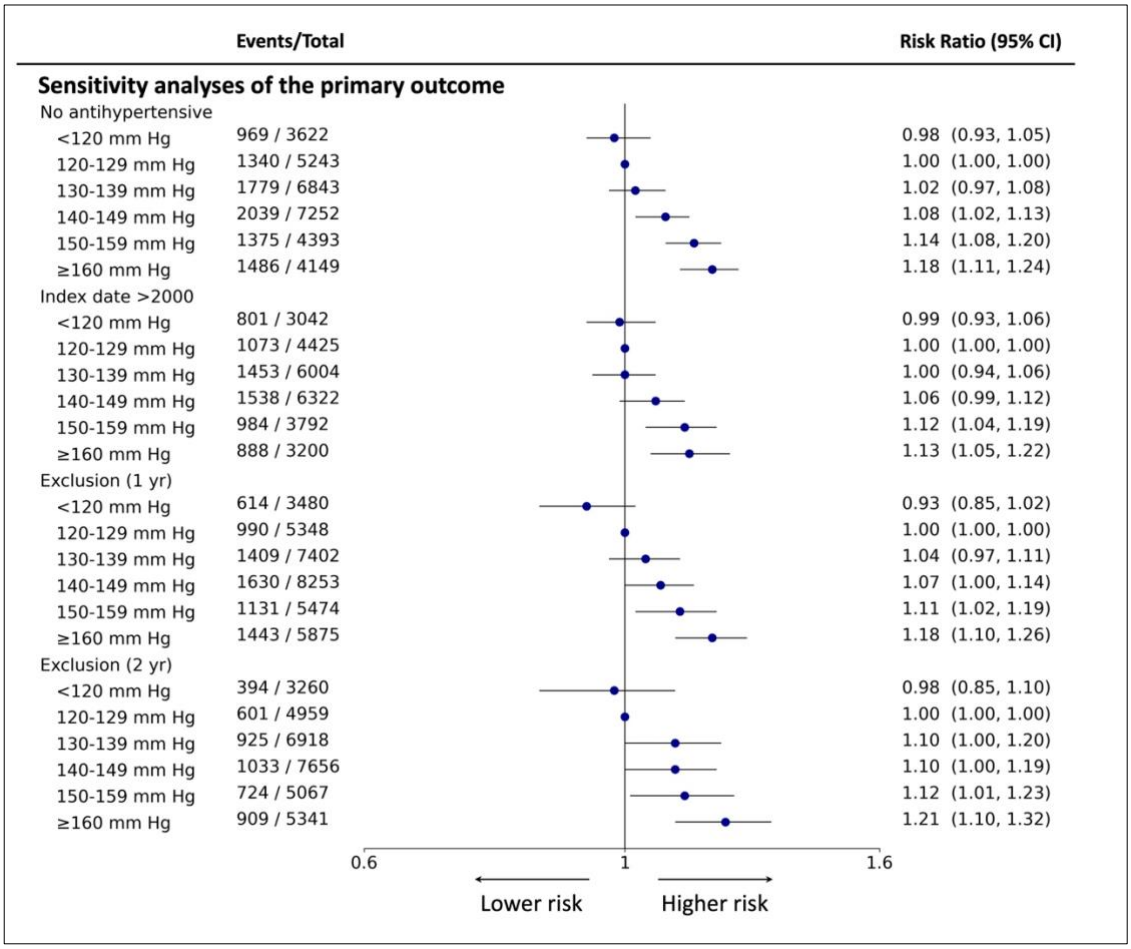


Forest plot of risk ratio estimates of the Targeted-BEHRT model with 95% confidence intervals (CI) for association of systolic blood pressure and secondary outcomes. From the left, the six exposure groups are shown in first column. Number of events and total number of patients in each exposure group is shown in second column. The forest plot and corresponding risk ratio estimates are shown in the right-most column relative to reference class, 120-129 mm Hg. The forest plot is plotted in logarithmic scale. For all estimates for reference class, there is no confidence interval.

In analyses of the components of the primary outcome, the Targeted-BEHRT model showed a continuous association between SBP and individual cardiovascular endpoints with lowest risk at <120 mm Hg in comparison to the reference category (Figure 7-7). Additionally, for endpoints of heart failure and IHD, the crude and adjusted LR estimates of RR found SBP between 130 and 150 mm Hg to contribute to the lowest risk of secondary outcomes (Figure S6) with little deviation in findings from the adjusted LR approach utilising the expanded predictor set (Figure S5). All four approaches found <120 mm Hg is associated with the lowest risk of stroke. Lastly, the trends discovered in

the four sensitivity analyses demonstrated little deviation from the patterns found in the main analysis (Figure 7-8).

Figure 7-8: Association with secondary outcomes in patients with COPD



Forest plot of risk ratio estimates of the Targeted-BEHRT model with 95% confidence intervals (CI) for association of systolic blood pressure and primary outcomes in sensitivity analyses. From the left, under a particular sensitivity analysis, the six exposure groups are shown in first column. Number of events and total number of patients in each exposure group is shown in second column. The forest plot and corresponding risk ratio estimates are shown in the right-most column relative to reference class, 120-129 mm Hg. The forest plot is plotted in logarithmic scale. For all estimates for reference class, there is no confidence interval.

7.2.5 Interpretation

Using a deep learning approach for assessing causality on longitudinal EHR, we found that SBP was continuously associated with cardiovascular risk in patients with COPD in a cohort of 39,602 patients. Individuals with SBP <120 mm Hg were found to

have the lowest risk of both the primary and secondary outcomes. Lastly, there was little material difference between the trends found in the sensitivity analyses and the main analysis.

SBP is established to be linearly associated with cardiovascular risk in the general population and in fact, naturally below average values in industrialized communities [37,205–207]. However, in subgroups with prior cardiovascular diseases and associated risk factors, the relationship remains insufficiently described. In this context of high-risk subgroups – such as those with diabetes, IHD, and other risk factors at study entry – many observational studies reject the monotonic relationship between SBP and cardiovascular risk concluding a "J -shaped" trend [13,16,35]. However, these observational studies are criticised for improperly dealing with manifestations of reverse causality and confounding. With cardiometabolic multimorbidity at baseline more prevalent in those with lower SBP than higher, additional variables capturing this poor baseline health and associated cardiovascular illnesses must be included for adjustment. Given an insufficient understanding of risk and protection in multimorbid patients currently, solely relying on expert-selection of known confounders (e.g., gender, age, BMI, known risk factors of CVD) exposes the modelling to issues of residual confounding [59]. As a result, unadjusted confounding due to multimorbidity in lower SBP groups can result in the J-shaped pattern: an optimum exists such that SBP below and above is associated with higher cardiovascular risk [35,193].

In our own implementation of LR adjusting for predictors defined in Byrd et al, the results captured this described trend and rejected the established linear relationship between SBP and cardiovascular outcomes [35,37]. In our cohort, we found that established risk factors of cardiovascular diseases were inversely related with the exposure, SBP. In patients with lower SBP, there was generally a higher incidence of IHD, chronic kidney

disease, diabetes, mental illness, and atrial fibrillation lending support to the hypothesis that issues of reverse causation may indeed be at play in our observational cohort. Even the fully adjusted LR model with the expanded set of predictors resulted in a non-monotonic association between SBP and cardiovascular risk in our cohort of COPD patients. This non-monotonicity or J-shaped pattern was preserved across analyses of both primary and secondary outcomes with the adjusted LR approaches.

Implementing the deep learning approach for assessing causality directly confronted these modelling issues. By utilising minimally processed diagnoses and medications data in routine clinical EHR, our deep learning approach accounts for a breadth of risk and protective factors potentially confounding the exposure-outcome relationship. In our cohort with COPD and cardiometabolic multimorbidity at baseline, in which traditional approaches failed to sufficiently captured confounding factors in observational data, our approach was appropriately implemented to model the association between SBP and cardiovascular events.

The continuous association concluded in this work raises important clinical questions for cardiovascular care. What is the optimal SBP in patients with COPD? Does this threshold differ from the recommendations for the general population (<120 mm Hg)? How should the decision calculus for blood pressure lowering treatment be formulated for those with COPD and hypertension? While guidelines for hypertension indeed endorse blood pressure lowering in patients with concomitant COPD and high blood pressure, the recommendations suggest a treatment target of <130 mm Hg (<140 mm Hg in the elderly) [48]. However, assuming causality, our results demonstrated an infimum of risk at SBP of <120 mm Hg – lower than currently recommended guidelines and consistent with the established log-linear understanding of the association between SBP and cardiovascular risk. Additionally, with median age of 69 years, our results

illuminated an SBP target of <120 mm Hg might be preferred over the recommended SBP target of <140 mm Hg in elderly COPD patients [48]. Given that randomised evidence of blood pressure lowering in COPD patients is unavailable and likely to remain unavailable in the near future, our study helps disentangle the relationship between SBP and cardiovascular events in COPD patients. However, further observational research independently evaluating the association between both, SBP and blood pressure lowering interventions, and cardiovascular outcomes in COPD patients utilising robust confounding adjustment methods like Targeted-BEHRT would be imperative for reassessing hypertension guidelines.

Looking at the strengths of the study, first, in terms of data, the comprehensive information provided by CPRD is a strength of our research. The linkage capabilities of CPRD allow the capture of rich health encounters (e.g., diagnoses, medications, measurements, static attributes) from various sources including primary care, secondary care, and mortality-based datasets. Strength of deep learning modelling is derived in part by richness of data; with access to rich EHR, the deep learning approach could better extract confounders, both known and latent in routine clinical data[170]. Second, with access to repeated SBP measurements specifically, we were able to derive a summary value (mean value of multiple SBP measurements) limiting issues of measurement error [204]. Third, we were able to capture many more patients than prior studies investigating this association, and also, unlike previous studies of SBP and cardiovascular risk, we included older aged patients and those with cardiovascular multimorbidity at baseline [35]. Exclusion from our study was limited; thus, this allowed us to understand the association of SBP and cardiovascular outcomes in high-risk subgroups with COPD. Fourth, rich longitudinal data in CPRD afforded us the opportunity to follow patients for a median of 3.9 years as opposed to the prior exploration of this association in patients with COPD,

which reported median follow-up of 1.9 years [35]. With a longer follow-up period, potential biases in RR estimation due to issues of reverse causation are mitigated. Fifth, we explored various sensitivity analyses in order to understand the role of unforeseen biases (e.g., reverse causality) and supplement the narrative of the main results. In terms of modelling, a strength of our work is the deep learning approach capable of extracting and adjusting for confounding factors in rich annotated EHR. Additionally, we implemented the conventional statistical approach with validated predictors set allowing direct comparison with the deep learning approach [35]. By utilising superior confounding adjustment methods, we demonstrated the utility in data-driven causal inference ultimately rejecting the evidence of a J-shaped relationship.

This study also had some limitations. The EHR data in CPRD has some degree of recording error, and the process of identifying COPD patients may have led to misclassifications. However, past studies have validated the use of primary care, secondary care, and mortality-based sources within the CPRD database for observational research; specifically, for COPD, there is ample evidence to suggest that the condition can be accurately identified in the CPRD dataset with a positive predictive value (PPV) of more than 80% as compared to clinician judgment [5,9,123,208]. Also, with continuous measurement like SBP measurements, random measurement error is a known issue; we have attempted to ameliorate this issue by taking an average of repeat measurements over the course of 12 months following index date – a recommended course of data processing in order to deal with issues of measurement error [204]. Lastly, as is the case with all observational studies, the proposed Targeted-BEHRT approach cannot fully capture all confounders and residual confounding may still bias estimation of the association; however, further validation with randomised trials would be prudent to fully disentangle

the nature of the relationship between blood pressure (lowering) and CVD in those with COPD.

7.3 Paracetamol, systolic blood pressure, incident cardiovascular diseases, and all-cause mortality

7.3.1 Introduction

Paracetamol, also known as acetaminophen, is the most commonly used analgesic worldwide and is recommended as a first-line treatment of pain in many acute or chronic conditions[209]. Although it is generally known to be safer than other frequently used analgesics, its conventional tablet formulation has certain disadvantages, such as low systemic bioavailability in oral administration and the potential for hepatotoxicity if taken in excess. To mitigate these issues, a soluble effervescent formulation of the drug was launched into the market. This is a compacted tablet that contains carbonates, acids, and sodium bicarbonate in addition to the active drug ingredients[210].

Access to effervescent pain treatments has become essential in some clinical circumstances, for example in people with swallowing disorders[211–213]. However, since effervescent formulations include sodium, concerns have been raised that it may increase the risk of elevated blood pressure, CVD, and mortality[214]. These concerns were recently supported by a study that found that initiating sodium-containing paracetamol is associated with an elevated risk of incident CVD and all-cause death [215,216]. This and another study that came to a similar conclusion, however, were based on conventional statistical approaches with their typical limitations[3]. The adjustment variables in such models are usually selected by experts, resulting in the omission of variables and interactions unknown to subject-area specialists. This can ultimately lead to residual

confounding and biased estimates, and can be particularly problematic in patients with multiple comorbidities and complicated risk factors, where complex interactions and hidden confounding are more likely[3].

With the development of deep learning approaches for causal inference, we have the opportunity to address pertinent issues of conventional confounding adjustments[123,170]. Specifically, concerning association analyses on EHR data, the Targeted-BEHRT model has demonstrated more accurate estimation of causal effect in a host of semi-synthetic data experiments and has even been applied to better understand risk and protection in multimorbid individuals[170].

7.3.2 Aims

In this study, we applied the Targeted-BEHRT deep learning approach to investigate the association of sodium-based paracetamol versus non-sodium formulations with SBP, incident CVD and all-cause mortality [214].

7.3.3 Methods

We used UK EHR from CPRD, validated for population-based epidemiological research[5,9,123]. Using EHR from primary care, linked with data from the ONS, we identified a cohort of 475,442 individuals. We included people between 60 and 90 years of age with at least one prescription of paracetamol between the years 2000 and 2014, aiming to replicate the approach in a previous study[216]. The index date was defined as the date of the first paracetamol prescription. Patients with any type of cancer, previous CVD (composite of heart failure, stroke, myocardial infarction), and prior use of compound paracetamol (e.g., paracetamol with codeine) were excluded similar to previous research[216]. Cancer and CVD were identified using previously validated disease

phenotyping methods while compound paracetamol was identified by CPRD "product code" (i.e., native coding system for medications designed by CPRD organisation)[217].

7.3.3.1 Exposures and outcomes

We compared forms of paracetamol containing sodium (i.e., formulations of "soluble" and "effervescent") as the exposure group with non-sodium-based formulations (i.e., formulations of "capsule", "tablet", and "oral suspension") as the comparison group. The information about the type of paracetamol was identified with the CPRD "product codes".

Three outcomes were investigated in a 1-year follow-up period: (1) systolic blood pressure (SBP) as a continuous outcome for patients with SBP measurements (as recorded in CPRD), (2) incident major CVD defined as a composite of MI, heart failure, and stroke, and (3) all-cause mortality, identified by death in the ONS registry. To mitigate measurement error, SBP was calculated as an average value of the measurements taken in a 6-months window around the 1-year mark (i.e., between 9 and 15 months following baseline)[204].

7.3.3.2 Deep learning and statistical analyses

We used Targeted-BEHRT, a causal deep learning model developed for causal inference on EHR data[170]. The Targeted-BEHRT modelling approach uses minimally processed EHR for accurate estimation of causal effect than benchmark statistical/machine learning approaches (see section 6.3.4). The model was adjusted for primary care diagnosis and medication records with attributions of age and calendar year of recording, as well as static attributes of sex and smoking status at baseline. By jointly predicting propensity score and outcome risk in one deep learning framework, the approach utilises CV-TMLE for downstream estimation of RR for binary outcomes and

mean difference (MD) for continuous outcomes[181]. Analysis was additionally conducted on patients with and without hypertension at index date.

For incident CVD, SBP, and all-cause mortality as outcomes, to directly compare our deep learning approach with conventional modelling, we implemented a two-stage propensity-based statistical modelling. In this conventional model, the propensity score was assessed with logistic regression; IPTW were derived and utilised for log-binomial modelling for binary outcomes, incident CVD and all-cause mortality, and linear regression for the continuous outcome, SBP; both models were regressed on the exposure variable[74,184]. We adjusted for a total of 52 variables based on past research, with the exception of the Townsend Deprivation Index, which was replaced with the Index of Multiple Deprivation (IMD)[216]. For BMI, the average of the measurements recorded in the 36 months before the index date were used to mitigate measurement error[204]. Multiple Imputation by Chained Equations (15 imputations) was conducted on missing continuous and categorical variables[218]. Estimation of RR for incident CVD and all-cause mortality investigations and MD for the SBP investigation, in addition to associated 95% confidence intervals were derived from model coefficients. Lastly, crude (unadjusted) effect size was calculated for all three outcomes as a naive approach. Analyses of those with and without hypertension was additionally conducted for all adjusted and unadjusted analyses.

We conducted several complementary and sensitivity analyses to check the robustness of the finding and investigate bias in estimation of all-cause mortality risk. First, in order to account for the possibility of reverse causality, we re-analysed the data, excluding those who died in the first month of follow-up repeated up to the sixth month of follow-up. Second, using conventional mediation analysis of direct effect estimation, we identified and analysed potentially mediating variables[219]. In order to conduct this

analysis, first, we identified factors reported in the time interval between exposure and outcome that were most associated with the exposure through unadjusted RR modelling. Second, we additionally conducted adjusted modelling of the association between exposure and individual mediators using log-binomial modelling controlled for sex, age, IMD, BMI, region, ethnicity, alcohol status, and smoking status. For the mediators that demonstrated a non-null association with the exposure in the analyses (i.e., implying mediation was present), we estimated direct effect utilising Targeted-BEHRT modelling. In essence, direct effect estimation aims to evaluate the strength of the association between exposure and outcome controlling for confounding and mediation[220]. Importantly, mediation is present if direct effect estimates are diluted with respect to the main analysis (i.e., association between paracetamol and all-cause mortality). Lastly, we pursued both aforementioned sensitivity analyses simultaneously to comprehensively mitigate the potential biases on the estimation for all-cause mortality risk. For these three sensitivity analyses, diagnosis records from secondary care were additionally included in Targeted-BEHRT modelling.

7.3.4 Results

A total of 475,442 eligible individuals were included in this study (Figure S7). The mean follow-up time was 11.0 months and 11.2 months for CVD and all-cause mortality as outcomes, respectively. Participants' characteristics by exposure categories are shown in Table 7-3 with extended data presented in Table S9. 460,980 and 14,462 patients were selected for the non-sodium and sodium-based paracetamol exposure groups respectively. Mean age at baseline was 74 (standard deviation: 8.6) years and 64% were women. While many characteristics at index date including BMI, year of birth, and several diseases and prescriptions were balanced between both exposure groups, smoking status, alcohol status, diabetes, hypertension, dementia, gout, and blood pressure lowering medications

were not, consistent with past studies investigating the same assocation[216]. For investigation of SBP as outcome, 235,699 patients were included; baseline characteristics for this subset of patients are comprehensively described in Supplementary materials (Table S10).

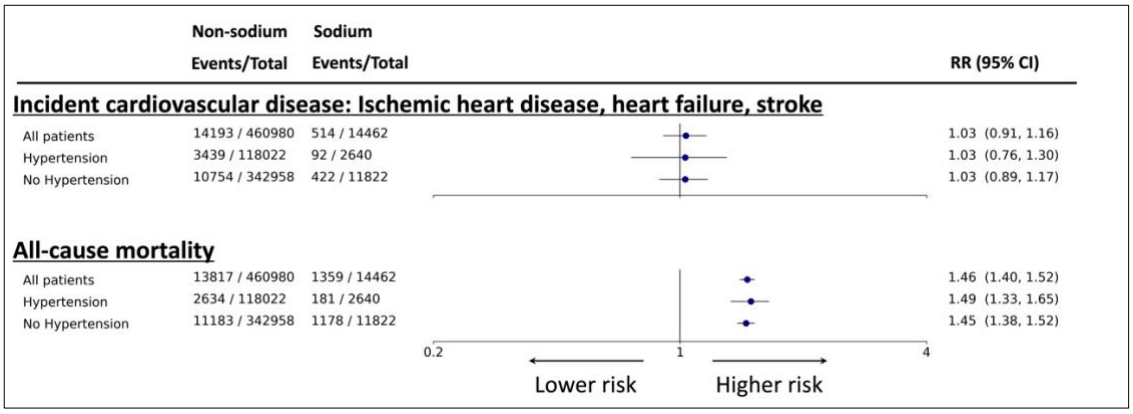
Table 7-3: Characteristics for the investigation of risk of sodium-based paracetamol on all-cause mortality and incident cardiovascular disease as outcomes

	Non-sodium formulation	***Sodium formulation***
No. (%)	460980 (97.0)	14462 (3.0)
Age, yrs (SD)	73.7 (8.6)	76.1 (9.1)
Women (%)	296190 (64.3)	10342 (71.5)
Ethnicity (White) (%)	126248 (27.4)	4060 (28.1)
IMD (SD [†]	1.9 (1.4)	1.9 (1.3)
SBP (SD)[†]	141.2 (13.7)	139.4 (14.5)
BMI (SD)[†]	27.7 (4.3)	26.0 (4.0)
Smoking status[†]		
Current or former smoker (%)	252522 (54)	5431 (37)
Never smoker (%)	208458 (45)	9031 (62)
Alcohol status[†]		
Current or former drinker (%)	343602 (74)	9221 (63)
Never drinker (%)	117356 (25)	5241 (36)
Comorbidity at baseline		
CKD (%)	3757 (0.8)	86 (0.6)
Diabetes (%)	41894 (9.1)	988 (6.8)
Hypertension (%)	118022 (25.6)	2640 (18.3)
Arthritis (%)	140161 (30.4)	2960 (20.5)
Gout (%)	16239 (3.5)	297 (2.1)
Rheumatoid arthritis (%)	7477 (1.6)	238 (1.6)
Hyperlipidaemia (%)	35861 (7.8)	737 (5.1)
Atrial fibrillation (%)	17563 (3.8)	443 (3.1)
Gastrointestinal bleeding (%)	5457 (1.2)	226 (1.6)
Reflux disease (%)	24349 (5.3)	663 (4.6)
Dementia (%)	9973 (2.2)	985 (6.8)
Medications use at baseline		
Statins (%)	112124 (24.3)	2304 (15.9)
Blood pressure lowering (%)	216574 (47.0)	5465 (37.8)
Anticoagulants (%)	21102 (4.6)	570 (3.9)
Antiplatelet (%)	120975 (26.2)	3612 (25.0)
Opioids (%)	139717 (30.3)	3031 (21.0)

SD: standard deviation; No: number; Yrs: years; BMI: body mass index; SBP: systolic blood pressure; CKD: chronic kidney disease; IMD: index of multiple deprivation Imputed variables Values presented are median with interquartile range (IQR) or percentage (%) YOB: year of birth; BMI: body mass index; HDL: high density lipoprotein; TC: total cholesterol; TG: triglycerides; IHD: ischaemic heart disease; SD: standard deviation; %: percentage. [†]The percentage of missing variables –alcohol status (54.6%), smoking status (36.9%), IMD (38.7), SBP (26.4%), BMI (40.9%)*

The RR of incident CVD in patients who had initiated sodium-based paracetamol compared to those who had initiated non-sodium-based paracetamol use is shown in Figure 7-9. In the overall analysis using Targeted-BEHRT model, we did not find any association between sodium-based paracetamol and incident CVD (RR 1.03; 95% CI: (0.91,1.16)). Likewise, in stratified analysis by hypertension status, the relative effects were similar between the two groups. By contrast, a positive association with the risk of CVD events was observed in analysis using two-stage statistical and crude modelling (Figure S8).

Figure 7-9: Association of sodium-based vs non-sodium-based paracetamol and incident cardiovascular disease and all-cause mortality

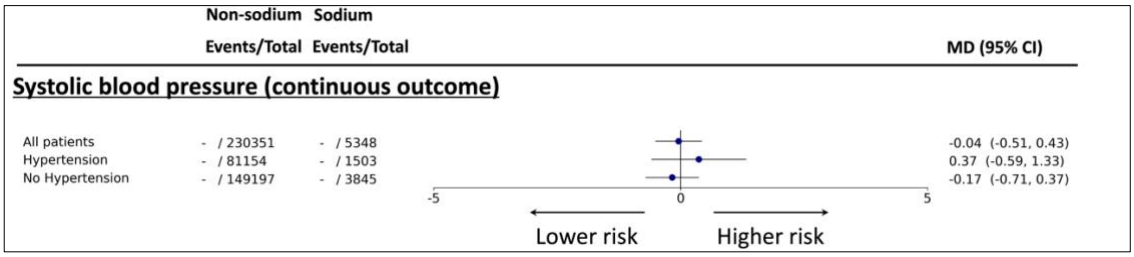


Forest plots of Targeted-BEHRT modelling for analyses of binary outcomes (all patients, stratified by hypertension status) is shown. Number of events and total number of patients in each exposure group is shown in second and third columns. The forest plot and corresponding risk ratio estimates are shown in the right-most column relative to the reference exposure, non-sodium paracetamol. The effect size is plotted on a logarithmic scale.

Additionally, in both the overall analysis and the subgroup analysis by the status of hypertension at baseline, we observed no association between the type of drug and SBP as a continuous outcome (Figure 7-10). While there was no association in the overall analysis of the crude model, there was some heterogeneity in the subgroup analysis by

hypertension status, with a rise in blood pressure in those with a history of hypertension at baseline and a decrease in those without (Figure S9).

Figure 7-10: Association of sodium-based vs non-sodium-based paracetamol and systolic blood pressure

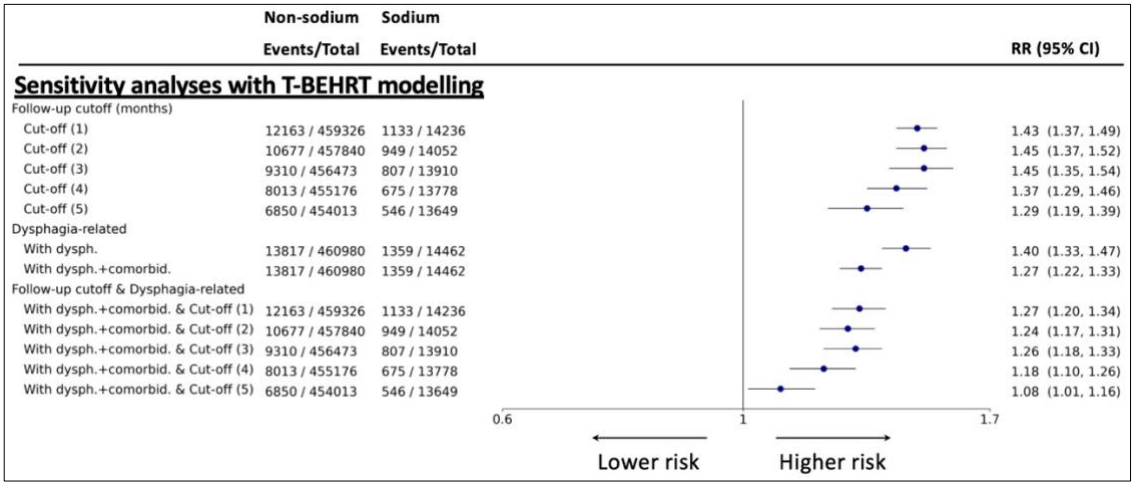


Forest plots of Targeted-BEHRT modelling for analyses of continuous outcome, systolic blood pressure (all patients, stratified by hypertension status) is shown. Number of events and total number of patients in each exposure group shown in second and third columns are left blank. The forest plot and corresponding mean difference (MD) estimates are shown in the right-most column relative to the reference exposure, non-sodium paracetamol. The effect size is plotted on a logarithmic scale.

Figure 7-9 shows the associations of drug types with all-cause mortality as the outcome. We found that sodium-based paracetamol is associated with an elevated risk of all-cause death, both in the overall analysis and in subgroups of hypertension status (Figure 7-9). Similar results were observed using two-stage statistical analysis and crude modelling, but with a greater magnitude of effect (Figure S8).

Because of a relatively short follow-up period of 12 months, we examined if the association with all-cause mortality was distorted by reverse causation. Excluding those who had died in the first month of follow-up, repeated up to six months, the Targeted-BEHRT effect size attenuated by roughly 50% as compared to the effect size on the entire cohort (Figure 7-11).

Figure 7-11: Association of sodium-based vs non-sodium-based paracetamol and all-cause mortality in sensitivity analyses



Forest plots of Targeted-BEHRT modelling for sensitivity analyses of all-cause mortality as outcome. From left, the type of sensitivity analyses is presented. Number of events and total number of patients in each exposure group is shown in second and third columns. The forest plot and corresponding risk ratio estimates are shown in the right-most column relative to the reference exposure, non-sodium paracetamol. The effect size is plotted on a logarithmic scale.

In order to investigate additional sources of bias, we analysed mediation on the pathway between paracetamol initiation and all-cause mortality. Table 7-4 presents the first 10 mediator variables ranked by the strength of the unadjusted exposure-mediator association. Furthermore, adjusted modelling demonstrated immaterial difference from the unadjusted findings (Table 7-4). Several neuro-cognitive and digestive tract disorders were identified as conditions associated with the exposure.

Table 7-4: Top ten conditions recorded post-index date with the largest difference in prevalence between the exposure groups (non-sodium-based and sodium-based paracetamol groups)

Disease	*Prevalence in non-sodium-based*	*Prevalence in sodium-based*	*Prevalence ratio association (unadjusted)*	*Prevalence ratio association (adjusted)*
Malignant neoplasm of oesophagus	0.001	0.009	5.826	5.34; (4.12, 6.933)
Pneumonitis due to solids and liquids	0.002	0.014	5.749	5.72; (5.002, 6.545)
Dementia in Alzheimer disease	0.002	0.008	4.274	3.78; (3.294, 4.334)
Dysphagia	0.004	0.017	4.232	3.39; (3.135, 3.669)
Alzheimer disease	0.007	0.031	4.117	2.76; (2.54, 3.009)
Multiple sclerosis	0.001	0.005	3.852	2.33; (1.898, 2.863)
Decubitus ulcer and pressure area	0.006	0.021	3.488	2.19; (1.874, 2.565)
Unspecified dementia	0.017	0.056	3.316	2.63; (2.455, 2.808)
Seizures	0.002	0.007	3.296	2.66; (2.353, 3.005)
Hemiplegia	0.002	0.007	2.916	2.67; (2.367, 3.015)

Association (unadjusted): relative prevalence of that disease in sodium-based paracetamol group versus the control exposure group (i.e., column 2 divided by column 3). Association adjusted: log-binomial modelling estimating association between exposure and condition adjusting for baseline covariates

Examination of these variables yielded insight that all of the variables were related to dysphagia, comorbid with dysphagia, or caused by dysphagia (Table 7-4) [212,221–224]. In light of this, for direct effect estimation utilising Targeted-BEHRT, we first assessed the association between jointly, exposure and the sole mediator, dysphagia, and the outcome of all-cause mortality. As compared to the main analyses (RR: 1.46; (1.4,1.52)), the direct effect estimated a slight reduction (RR: 1.39; (1.35, 1.45)) (Figure 7-11). Second, we assessed the association between jointly, the exposure and all of the ten mediators (Table 7-4), and the outcome of all-cause mortality using Targeted-BEHRT. The direct effect was estimated to be RR: 1.27; (1.22,1.30) attenuated with respect to the previous direct effect analysis and the main analysis of all-cause mortality (Figure 7-11).

Lastly, we investigated a combination of the presented strategies: in addition to excluding patients who had died in the first month of follow-up, repeated up to six months, we estimated the association between jointly, exposure and the ten mediators (Table 7-4), and the outcome. As the cut-off time was increased, the effect size diluted commensurately. Investigating the association in patients alive 6 months into the follow-up period ("Cut-off (5)" in Figure 7-11), Targeted-BEHRT estimated RR: 1.08; 1.01, 1.16.

7.3.5 Interpretation

Utilising a deep learning approach for assessing causality on a rich EHR data, we found that sodium-based paracetamol had no effect on incident CVD and SBP as outcomes with respect to non-sodium formulations. For all-cause mortality, our deep learning approach captured a positive association in the main analysis, however in sensitivity analysis we found attenuation of the effect towards the null after considering mediation and the presence of reverse causality. Similar to results from past works, conventional approaches captured higher risk for all three outcomes.

Previous research utilising conventional approaches have shown that sodium-based paracetamol increases the risk of both incident CVD and all-cause mortality[215,216]. Our conventional statistical modelling utilising IPTW-based propensity score methods regressing on predictors comparable to those from previous studies captured a similar positive association with all outcomes[215,216]. However, our analyses showed that more comprehensive adjustment of known and inferred health variables with deep learning modelling weakened the association.

While sodium-based paracetamol was found to increase risk of all-cause mortality by all models, sensitivity analyses using Targeted-BEHRT illuminated that the distorted

causal effect captured in the main analyses is multifactorial in nature. On one hand, noting the limited follow-up period of one year, we analysed our observational study for manifestation of reverse causation. We captured attenuation of effects as we excluded those who died early in the follow-up period, directly and clearly capturing distortion of effect size by reverse causation. On the other hand, excess risk captured in the main analysis due to dysphagia and associated comorbidities potentially yields two different interpretations.

One interpretation is that dysphagia and associated conditions as mediator variables are responsible for increasing all-cause mortality risk. In which case, utilising Targeted-BEHRT for direct effect estimation (association between jointly, exposure and mediators, and all-cause mortality) addressed the root of the excess mortality risk and demonstrated that the association between sodium-based paracetamol and all-cause mortality converges towards the null hypothesis when mediation is appropriately handled. A second, perhaps orthogonal interpretation is that as opposed to indirect effects by mediation, dysphagia and associated comorbidities are rather confounding variables with biases of recording delay/missingness and are responsible for increasing all-cause mortality risk. As previously discussed, dysphagia is a major reason that patients are prescribed effervescent formulations of paracetamol as opposed to solid formulations (e.g., tablets or capsules). Associated with exposure status and, in many cases, all-cause mortality in the elderly, dysphagia-driven confounding is challenging to assess using administrative EHR. Albeit noted to be quite common in the elderly, dysphagia is underdiagnosed for numerous reasons (e.g., patient embarrassment), and even if diagnosed, recording is delayed and often recorded as a secondary diagnosis to often more serious associated comorbidities including head and neck diseases, complications due to paralysis, cancer, epilepsy, and neurodegenerative diseases[212,221–224]. Additionally,

diagnostic delays of the associated neuro-cognitive conditions of dementia and Alzheimer's disease in addition to cancer is established[225,226]. In the presence of these conditions, effervescent formulations of medications, if available, are indeed recommended by prescription guidelines in the UK for those complaining of difficulties with swallowing – even if formal recording of dysphagia or associated conditions is absent in EHR[227,228].

In our retrospective data, at least partly due to complaints of swallowing difficulties, clinicians prescribed patients effervescent formulations for paracetamol, forming the exposure group. Thus, in addition, considering the short follow-up period of 1 year, while these conditions are sometimes recorded only after index date likely due to the more sub-clinical and perhaps, asymptomatic nature of many of these conditions at index date, the dysphagia and associated comorbidities are likely present at the time of exposure initiation. Moreover, the ten diseases (Table 7-4) are associated with exposure status and mortality. Therefore, since Targeted-BEHRT only accounts for predictors up to index date, patently an issue of residual confounding, omission of these conditions from confounding adjustment would demonstrate a positive association between sodium-based formulations and all-cause mortality (Figure S10)[212].

Given (1) the short follow-up period of one year, (2) the evidence on recording biases concerning dysphagia and associated comorbidities, and (3) known issues in association modelling in failing to capture all confounders, there is compelling evidence in support of this interpretation for excess all-cause mortality risk. With this interpretation, our direct effect analyses alternatively function as simply, confounding adjusted analyses using Targeted-BEHRT – additionally adjusting for confounding by dysphagia and related conditions (allowing for delayed recordings after index date).

While each interpretation has merits and substantial supporting evidence, by either mediation or confounding, ultimately, the association between sodium-based paracetamol and all-cause mortality is diluted when appropriately handling variables independently increasing mortality risk. Interestingly, solely accounting for dysphagia in sensitivity analyses resulted in limited risk reduction as compared to the main analyses. Perhaps, this is because while dysphagia itself might be associated with mortality, the diseases associated with dysphagia exert greater risk on mortality in comparison (e.g., oesophageal cancer and dementia in (Table 7-4)[223,229]. Estimating the association between exposure and all-cause mortality accounting for dysphagia and additionally, the associated conditions as mediators (or conversely, confounders under the second interpretation), indeed further diminished the effect size. Lastly, in parallel, accounting for both reverse causation and dysphagia and associated conditions, the association tended towards the line of parity hinting that the two pathways might be independently contributing to excess risk of all-cause mortality. Depression of the effect size in these analyses reflects the heavy presence of bias in the main analysis of all-cause mortality as outcome.

Given the compelling evidence in support of the second interpretation, a question regarding confounding emerges: how can we ensure modelling adjusts for these incipient confounders in the observational study? While impossible to truly guarantee the causal assumption that all confounders are adjusted, models that can take into account both known and latent confounding such as Targeted-BEHRT can more accurately estimate causal effect than conventional approaches[170,230]. As an example, while we note dementia is more prevalent in the exposure group (

Table 7-3), modelling by past works and our independent statistical implementation omit this variable from the predictor set – perhaps simply overlooked for a useful adjustment variable[215,216]. However, this variable is indeed imbalanced in exposure groups; independently, dementia is established to be associated with mortality[229]. Therefore, omission of this variable is the omission of a confounder rendering downstream estimates to be exaggerated. While we have uncovered that dementia, a single variable, is confounding the association, there may be several other confounder factors not explicitly measured. On the other hand, Targeted-BEHRT extracts confounding elements from minimally processed EHR more comprehensively than statistical model estimation, rendering attenuated effect sizes even in the main analyses of all of the three outcomes investigated.

Especially in multimorbid elderly patients, pain is a common symptom of many conditions. Inappropriate clinical recommendations derived from biased estimations of risk will limit access to the only pain-management option available[212,231]. The immaterial effects on CVD and SBP as outcomes should help mitigate concerns regarding the effect of sodium-based paracetamol. In the investigation of all-cause mortality, the strong positive effect size was found to likely manifest from various biases as opposed to the exposure. Further confirmatory randomized evidence would be beneficial to validate findings from our study; a trial would (1) yield greater sample size for the exposure group and (2) would provide conclusive, unconfounded evidence of the association since confounders, known and unknown, would be theoretically randomised at index date.

In terms of strengths of our study, first, in terms of data, the comprehensive EHR provided by CPRD is a strength of our work. Whilst replicating Zeng et al's work, we restricted our work to GP-based adjustment; for sensitivity analyses, the access to secondary care data enabled more comprehensive adjustment. Second, the deep learning

approach extracted both known and latent confounders in comprehensive EHR, an asset in our approach not available in more conventional approaches in modelling. Third, with access to repeat measurements, we were able to mitigate issues of measurement error with our definition of SBP as an outcome. Fourth, our sensitivity analyses into all-cause mortality as outcome enabled better understanding of biases prevalent in this observational study. Specifically, reverse causality analyses demonstrated the issues with a short follow-up period while inspection of mediator variables illuminated how unadjusted confounding was present and distorting effect sizes. Lastly, we also implemented conventional models to compare directly against the deep learning approach whilst mirroring Zeng et al's work[216].

In terms of limitations, CPRD, as an administrative dataset, has known issues of recording biases and measurement errors. In fact, uncontrolled confounding at baseline due to recording biases played a prevalent role in the estimation; incipient diseases at baseline due to latent nature of conditions hampered estimation of treatment effect in main analyses. Furthermore, more accurate modelling of the outcome and patient censoring is necessary[232]. However, because survival deep learning modelling is still in its nascent stages and has not been appropriately developed for assessing causality, we were unable to pursue a more nuanced evaluation of risk with time-to-event deep learning modelling. In the same vein, there have been numerous advancements in mediation analysis, but few have been tailored for deep learning approaches. Thus, further work combining deep learning with, for example, doubly-robust estimation for mediation analysis needs to be explored in future works. Lastly, a frequent criticism levied at direct effect estimation is that the presence of mediator-outcome confounding can explain away the direct effect association[220,233]. Assuming there exists unadjusted variables perhaps that associate with both the mediator and the outcome, it is possible that the direct effect

measured is overestimated. With the RR estimate of 1.27, as an example, there could hypothetically exist a mediator-outcome confounder (e.g., another neurological condition associated with both dysphagia/comorbidities and all-cause mortality) with RR of 4.5 with prevalence of 15% in the exposure group and 5% in the non-exposure group, that could explain away the direct effect[234,235]. Albeit unlikely, the existence of this mediator-outcome confounder would only further weaken the captured association and ultimately nullify the relationship between sodium-based paracetamol and all-cause mortality.

8 DISCUSSION

The presented doctoral research explored deep learning modelling for causal inference. The following sections will discuss the summary of main findings from the doctoral research, strengths and limitations, the methodological implications of the findings, the clinical implications of the findings, future directions for research, and salient concluding remarks concerning the doctoral research.

8.1 Summary of main findings

In chapter 2, an in-depth review of literature relevant to the thesis was presented. First, this review introduced the cardiovascular diseases, risk factors, and related comorbidities studied in this thesis. The national (UK) and global burden of these diseases were discussed in these initial sections. Second, an introduction of electronic health records was presented. Third, risk prediction was discussed; the motivations, theory, and evaluation of risk prediction models are discussed and conventional models and their strengths and limitations are discussed in the context of epidemiological studies. Fourth, following this material, with risk prediction introduced as the foundation of conventional statistical modelling in epidemiology, the foundations of causal inference were discussed in the section. In this section, motivations for causal investigations were introduced, followed by discussion of the theoretical framework of causal inference including elucidation of assumptions for identifiability of causal effect, and methods for causal inference and association analyses in the observational setting. Fifth and finally, deep

learning is explored and material concerning some notable neural network architectures are discussed in this section.

In chapter 3, the CPRD database, the dataset used for the thesis, was discussed. Information concerning the data collection, processing, regulation, and validation was presented in this section. Particular focus was devoted to ascertaining the legitimacy, validity, and reliability of the CPRD database. Establishing validity of the CPRD database was necessary since my doctoral research utilises this dataset for deep learning and statistical analyses of various exposures and cardiovascular-related outcomes in the observational capacity. Lastly, the data cut used in this research was described in full in this section.

In chapter 4, model development and risk prediction were conducted and two studies were presented: (1) the development of the Transformer-based BEHRT model and prediction of occurrence of subsequent diseases in different settings, and (2) the risk prediction of incident heart failure utilising BEHRT. In these works, the model, BEHRT was developed to utilise minimally processed EHR and consolidate them in a unified Transformer-based feature extractor. Specifically, the model incorporated raw diagnosis and medications records from the primary and secondary setting along with attributions of time (patient age). With this rich modelling of longitudinal health variables, the model achieved state-of-the-art predictive performance on subsequent disease prediction tasks. Furthermore, the sex and embedding analyses demonstrated that the model was able to capture clinically valid concepts in high dimensional latent space. In the heart failure risk prediction investigation, I showed that the BEHRT architecture can be flexibly modified for further annotation of medical history with inclusion of the calendar year embedding layer. With this inclusion, I demonstrated that BEHRT can achieve superior predictive performance in terms of AUROC and AUPRC metrics on incident heart failure prediction

task. Furthermore, ablation analyses of modalities utilised for modelling illustrated that certain modalities are more important than others for predictive performance. In particular, I found that the medication and year modality was important for predictive gain. While medications complemented the diagnostic medical history data, the year modality on the other hand likely provided more latent elements that allows for capture of the "birth cohort effect" in the cohort extracted for the investigation.

In chapter 5, while BEHRT demonstrated superior risk prediction performance on a host of tasks, I investigated if the model could be trusted by breaking open the "black box" Transformer architecture and deriving what the model captured as elements of risk and protection in medical history with respect to prediction of incident heart failure. The temporal embedding analyses complemented analyses presented in Chapter 4; year is indeed an informative embedding structure that provides utility for the prediction task. Furthermore, the contribution analyses demonstrated that (1) BEHRT captured validated risk and protective factors of heart failure, and (2) BEHRT captures potentially novel risk and preventative factors that can be formally tested in a hypothesis testing framework.

In chapter 6, given that deep learning has demonstrated superior predictive performance and can be trusted, a derivation of the BEHRT model was created for causal inference and association analyses. The model, Targeted-BEHRT, utilised a two-step procedure consisting of (1) deep learning-based confounding adjustment and (2) targeted learning that enabled more accurate RR estimation with mitigated selection/finite sample estimation bias. Furthermore, the model successfully estimated RR more accurately as compared to several benchmark statistical and deep learning models in semi-synthetic data experiments. Ablation analyses demonstrated that the unsupervised modelling was instrumental for mitigating bias in estimation when the model was allowed access to large-scale, rich EHR, but doubly-robust estimation was instrumental in mitigating bias

in data sparse settings (i.e., finite sample estimation). Lastly, the Targeted-BEHRT model was implemented to investigate the effect of antihypertensives on incident cancer in a population free of cancer at baseline. The Targeted-BEHRT analysis captured a null association with exception of the association between CCBs and cancer with respect to control, ACEIs; these results were consistent with established meta-analyses of randomised evidences.

In chapter 7, I implemented Targeted-BEHRT in three more association studies investigating at-risk patients either due to existing comorbidity or old age. First and second, in the investigation of the association of SBP and cardiovascular endpoints in those with diabetes and COPD respectively, while conventional modelling generally estimated a J-shaped pattern (with exception of stroke as outcome), the Targeted-BEHRT model utilising both longitudinal and static medical history captured a continuous, log-linear trend. Several sensitivity analyses pursued using the Targeted-BEHRT modelling demonstrated little deviation from the results of the main analyses. Third, in the investigation of the association between sodium-based paracetamol on SBP, incident CVD, and all-cause mortality with respect to non-sodium formulations in elderly patients, the Targeted-BEHRT model captured a null association for outcomes, SBP and incident CVD. However, the model estimated significant increased risk of all-cause mortality. In sensitivity analyses of all-cause mortality as outcome, I found that both reverse causality and dysphagia-related confounding were likely biasing the association. Appropriately accounting for these issues in observational data, the association tended towards the null. While Targeted-BEHRT captured a null association between sodium-based paracetamol and SBP, incident CVD, and all-cause mortality (with appropriate handling of the elements distorting the effect size), the conventional statistical modelling demonstrated increased risk for all three outcomes. In gist, the Targeted-BEHRT model, with better

confounding adjustment and bias mitigation abilities than conventional approaches, introduced robust observational evidence for hypotheses, which otherwise are not likely to be tested in a randomised trial setting.

A summary of doctoral research conducted for this thesis is presented in Table 8-1.

Table 8-1: Summary of doctoral research

Chapter	*Objective*	*Cohort*	*Tasks*	*Results/Conclusions*
4	Develop risk prediction model for subsequent disease prediction using minimally processed multimodal, longitudinal EHR with attributions of time	1.6 million registered patients with linkage to HES and more than 5 visits	Predict subsequent diseases: 1) Next visit 2) Next 6 months 3) Next 12 months	1) BEHRT outperforms recurrent and convolutional neural network models on subsequent disease occurrence task 2) BEHRT captures sex and diseases in latent space in line with clinical knowledge
4	Develop risk prediction model for incident heart failure prediction	100,071 registered patients with linkage to HES, free of heart failure, and more than 5 visits	Predict incident heart failure	1) BEHRT outperforms recurrent and convolutional neural network models on incident heart failure prediction task 2) In ablation analyses, BEHRT finds medication and calendar year to be important sources of data
5	Develop tools to explain deep learning model	100,071 registered patients with linkage to HES, free of heart failure, and more than 5 visits	Develop tools for understanding contribution of various disease/medications to incident heart failure prediction	1) BEHRT independently captures risk factors in medical literature 2) The explainability tool generates novel factors of prevention (e.g., prostaglandin analogues)
6	Develop deep learning model for accurate causal effect estimation	516,365 registered patients with linkage to HES free of cancer at baseline	1) Develop deep learning model that can more accurately estimate causal effect than benchmark models 2) Develop testing environment using semi-synthetic data to objectively test estimation capabilities of various models	1) Targeted-BEHRT more accurately estimates RR utilising unsupervised modelling in tandem with propensity score/outcome modelling and mitigates selection biases utilising doubly robust estimation
6	Implement Targeted-BEHRT for estimation of effect of antihypertensives on incident cancer	516,365 registered patients with linkage to HES free of cancer at baseline	Estimate effect of BBs, CCBs, diuretics, and ARBs on cancer with respect to ACEIs.	1) Null effect for all association with exception of the one with CCBs 2) Effect estimates were consistent with meta-analyses of randomised evidence
7	Estimate association of SBP and cardiovascular endpoints in patients with diabetes	49,000 registered patients with diabetes and free of heart failure at baseline	In patients with diabetes, estimate association between SBP and cardiovascular endpoints with Targeted-BEHRT and conventional modelling	1) While conventional modelling estimated a J-shaped association, the Targeted-BEHRT model estimated a log-linear relationship in both main and sensitivity analyses
7	Estimate association of SBP and cardiovascular endpoints in patients with COPD	39,602 registered patients with COPD at baseline	In patients with COPD, estimate association between SBP and cardiovascular endpoints with Targeted-BEHRT and conventional modelling	1) While conventional modelling estimated a J-shaped association, the Targeted-BEHRT model estimated a log-linear relationship in both main and sensitivity analyses
7	Estimate association of paracetamol and SBP, all-cause mortality, incident CVD in patients with COPD	475,442 registered patients free of cancer, CVD, prior compound paracetamol use at baseline	Estimate association between sodium-based paracetamol and SBP, incident CVD, and all-cause mortality with respect to non-sodium formulations.	1) While conventional modelling estimated increased risk for all outcomes, Targeted-BEHRT model estimated null effect estimate for outcomes of SBP and incident CVD with increased risk of all-cause mortality. 2) In sensitivity analyses, accounting for reverse causality and dysphagia related confounding, the estimates tended towards the null.

Left most column is chapter number. CVD: cardiovascular disease, COPD: chronic obstructive pulmonary disorder, HES: hospital episode statistics; SBP: systolic blood pressure; EHR: electronic health records; RR: risk ratio;

8.2 Strengths and limitations

In the following sections, the strengths and limitations of the doctoral research will be presented.

8.2.1 Strengths: Data

One of the greatest assets in this research is the dataset, CPRD. The CPRD dataset is an incredibly rich dataset providing access to multiple data modalities for patients registered in the primary care setting. Furthermore, the dataset offers linkages to other secondary care and mortality-based datasets. In sum, this is a powerful dataset for nationally representative epidemiological research. In fact, the breadth and depth of CPRD is in many ways, unparalleled. Few other datasets offer as many attributes of health for as many patients. Furthermore, the work presented in this research is conducted on UK patients; for this reason, the results and conclusions derived from these research works are likely representative of other high-income countries. Ultimately, this implies that the analyses conducted on CPRD data is clinically important in the global context in addition to the national context.

8.2.1.1 Validity

While the reliability and validity of CPRD is generally presented in Chapter 3, there are also numerous other research works that demonstrate the reliability of the dataset. As examples, previous works have demonstrated that the diagnoses in the primary healthcare setting for a host of disease groups are generally concordant with national statistics. Generally, the PPV of the diagnosis records found in CPRD is 89% (92% completeness) when compared with statistics from national registries[107]. Additionally, previous research has also found that utilising linked data sources of rich EHR improves the completeness and validity of diagnostic medical history of registered patients[5,104,107].

Also, the works in this thesis also employs appropriate and validated data extraction and processing methods preserve the consistency of the raw CPRD data. Specifically, the data processing and cohort selection was carefully conducted to mitigate injections of bias into the analysis. Furthermore, several reliability checks were conducted to ensure the data points were "up to standard" and of high quality for downstream research[5]. For several works, erroneous data entries were removed with data filtering steps and carefully designed criteria. In addition, diagnosis and medication codes were mapped to harmonised codes in order for ease of machine-readability using validated phenotyping algorithms[120,122]. Caliber phenotyping was crucial for conducting disease code harmonisation for research presented in Chapters 4 and 5[122]. Overall, with these methods of cleaning the data, the validity of the original CPRD is maximally retained in the processed data.

8.2.1.2 Sample size

The research conducted in this thesis benefits from the large sample size offered by the CPRD data. In risk prediction and explainability investigations, 1.6 million and 100 thousand patients with higher number of interactions with healthcare providers (i.e., at least 5 visits) were selected for investigation of subsequent diseases and incident heart failure prediction respectively (Chapters 4 and 5). In investigations of association analyses in the observational setting, 516 thousand patients were selected for the investigation of the association between various antihypertensive drug classes and cancer. Furthermore, 49 thousand, 40 thousand, and 475 thousand eligible patients were selected for the three studies involving high-risk and elderly patients respectively.

The large sample size offered by CPRD is beneficial for epidemiological research for a host of reasons. In association analyses, the 7 million patients registered between 1985 and 2015 offers the possibility of high-powered studies with slimmer

confidence intervals than past works. Additionally, larger sample sizes can allow for identification of smaller effect measures as well; for example, in the investigation of antihypertensives and cancer, while most drug classes presented a null association with ACEIs, a slight statistically significant positive association was captured with CCBs[170]. These results were consistent with meta-analyses of randomised evidences, and the large sample size allowed for high-powered investigations of drug class comparisons. Furthermore, in the investigations studying SBP and cardiovascular endpoints in at-risk patients, the selection criteria restricted analyses to those with prior illness and at specified age ranges. However, each exposure group and outcomes (both primary and secondary) had sufficient numbers for analyses. In fact, in the investigation in COPD patients, the analysis was conducted on a cohort with far greater numbers than those in previous research[35]. For the investigation of paracetamol and various outcomes, large sample sizes were available for both the analyses of cardiovascular and all-cause mortality as outcomes as well as SBP as a continuous outcome. For all association analyses, the large sample size also allowed for numerous sensitivity analyses ultimately allowing the opportunity to trust the robustness of the results of the main analyses in various subgroups.

Additionally, given that Transformer models require large amounts of data for training, the CPRD dataset is an optimal dataset for Transformer-driven risk prediction and causal inference[119]. Specifically, for BEHRT and related models, the large sample size allows for pre-training of embeddings and model weights and downstream fine-tuning. For the risk prediction works, the extracted cohort of 1.6 million patients allowed for pre-training of model weights prior to task-specific fine tuning. This strategy helped secure the superior predictive performance on the subsequent disease and incident heart failure prediction tasks (Chapter 4). In the works for causal inference, the Targeted-

BEHRT model was initially pre-trained on 6.8 million patients prior to further tandem unsupervised (MEM) and supervised modelling on the specified cohort for association analyses (Chapters 6 and 7). In the case of high-risk cohorts, since the cohort of eligible patients for analysis is naturally smaller than other healthier cohorts, the large amounts of data offered by CPRD for pre-training is even more important for downstream Targeted-BEHRT modelling of the association of interest.

8.2.1.3 Richness

Additionally, unlike many other datasets for EHR, the CPRD dataset is especially rich in two ways: longitudinal nature of data available and number of variables.

In terms of the longitudinal nature of variables provided, the benefits are manyfold. Previously, many epidemiological research works were exclusively conducted on cross-sectional data. However, cross-sectional datasets directly imply that there will be certain restrictions on the research. For example, data will be collected at certain time points hence the collected variables might have less information concerning time (e.g., age, calendar date, etc). Ultimately, for both statistical and deep learning modelling, adjustment might be impacted; with longitudinal data, the models like BEHRT and Targeted-BEHRT can more accurately capture the medical trajectory of the patient thereby allowing for more nuanced adjustment of variables at baseline for both risk assessment and association analyses (Chapters 4, 6, and 7). Furthermore, for ascertainment of variables such as pre-exposure and exposure variables; longitudinal data variables are necessary to ensure that modelling is not adjusting for variables "on the causal pathway" – i.e., happening between the time of exposure and outcome as done in Chapters 6 and 7. Similarly, outcome ascertainment is easier and more robust with access to longitudinal data variables. Also, when dealing with issues of reverse causation as shown in investigations in Chapter 7, longitudinal data allows for careful cohort selection

procedures that alleviate the presence of reverse causation and associated issues of uncontrolled confounding.

In terms of number of variables, the CPRD dataset offers so many variables for more comprehensive modelling. The data used in this doctoral research includes: diagnoses in the primary and secondary care settings, prescription data, measurements, death (cause/date). In addition to this, the phenotyping for specifically the CPRD dataset is validated allowing for seamless processing of diagnosis and prescription data.

8.2.1.4 Other benefits

Use of routine clinical EHR offers other benefits in terms of mitigations of biases. Unlike self-reported data, recall bias is limited with the use of routine clinical EHR[236]. Sensitive diseases like developmental disorders can be studied with lesser concerns of recall bias distorting conclusions[236]. Also, the CPRD dataset offers accurate time information on records from the primary and secondary care settings that are also not susceptible to issues of recall bias.

8.2.2 Strengths: Modelling

8.2.2.1 Transformer-driven modelling: Maximal preservation of EHR

The proposed BEHRT architecture and derivations take great care in preserving the natural complexity and longitudinal structure of the CPRD database. Disease and medication data serve as input into the model; repeat diagnoses/medications are included as opposed to discarded to provide further latent information concerning duration and, in some cases, intensity of the condition[123,170]. Especially, this repetition implicitly informs the model if a disease is chronic versus acute; diseases, which are chronic or at the least, episodic, are more likely to repeat as opposed to those which are acute or isolated (e.g., "one-off" conditions)[123]. The idea of a visit to the healthcare services is also inputted for the model in the form of the "SEP" token; summarisation of the medical history into one

single vector is also easily possible with the pooling layer acting on the "CLS" token's hidden state[123]. Furthermore, both relative and absolute forms are used to provide time-based information for the diseases[123,170]. Position is encoded as a pre-determined function in order to alleviate issues of biased and/or weak learning of position of visits due to imbalanced distribution of history length in EHR[86]. This positional encoding in a Transformer model is the substitute for representing sequential data in the recurrent and convolutional neural network framework. Lastly, while longitudinal variables are indeed important for modelling, some variables are static and unchanging (e.g., ethnicity). In the Targeted-BEHRT model, the inclusion of static variables in the architecture allowed more comprehensive modelling of patient characteristics; not only were longitudinal variables included to model the health trajectory of the patient but also static variables were included that could form latent interactions with these temporal variables in deeper layers of the network[170].

In addition, for a given visit in medical history, the age, position, year, and segment embeddings are identical; this ultimately makes the BEHRT and other derived models order-invariant (i.e., within a visit, the order of diagnoses do not matter). In stark contrast to recurrent and convolutional networks, for which ordering of encounters matter, the BEHRT model allows for more generalised learning of sequential data. Specifically, the attention mechanism in the BEHRT model investigates intra-visit relationships amongst diseases[123]. Furthermore, due to the feed-forward structure of BEHRT's model and therefore, its ability to handle longer medical history sequences as compared to the recurrent neural network structures such as RETAIN/RETAIN-EX (suffering from issues of vanishing/exploding gradient), more comprehensive medical history modelling is possible with Transformer-driven modelling[124]. In this way, the

model can maximally preserve the complexity and the high-dimensional nature of patient medical history.

8.2.2.2 Explainability

While the BEHRT model is not fully explainable, some aspects are quite transparent ultimately paving the way for stronger confidence in deep learning models in the healthcare setting. When investigating embeddings and time-related embeddings (e.g., age and calendar year), certain examinations can directly illuminate how the model assimilates and processes input and what the model finds as important input for learning. In Chapter 4, the embedding analyses demonstrate that the model captures clinical encounters in concordance with clinical understanding; in terms of sex-related, cardiovascular, renal, and other disease groups, the model naturally clustered these diseases without manual feature engineering or prompting. Specifically, the model's trained clinical encounter embeddings were in line with symptomatology and progression of the diseases.

In addition, the ablation analyses in Chapter 4 itself can be a simple way of explaining model predictions. While this is an indirect way of quantifying utility of certain modalities in the model, the ablation analyses directly illuminated that certain predictors were key: medications and calendar year embedding structures provided information perhaps, orthogonal to diagnoses and age, respectively allowing for more nuanced modelling. Also, the importance of calendar year was in line with understanding of the "birth cohort effect" in epidemiological modelling[131]. Furthermore, the temporal embedding analyses confirmed the utility of the year embedding separately from the ablation analyses[131].

These forms of direct evaluation of the embedding space and relevant ablation studies performed in this thesis is one strong step in the path to trusting deep learning models.

For those elements of BEHRT not readily explainable, a major strength of this research is the development of the perturbation-based tool to explain incident heart failure predictions (Chapter 5). This tool captured several associations that enable more trust in deep learning solutions: (1) several medically validated risk factors were indeed captured by the BEHRT model in both main and age-stratified analyses, (2) while not causal, understanding of risk, treated risk, and untreated risk was also in line with clinical understanding of heart failure and associated care, (3) the year-stratified analyses confirmed the underlying change in the prescription composition of several medications.

Lastly, in addition to instilling more confidence in deep learning approaches, the explainability tool (Chapter 5) also helped to generate hypotheses concerning incident heart failure prevention. Prostaglandin analogues and digoxin were captured as medications that could potentially prevent incident heart failure; hypothesized to be beneficial in previous works, the presented work for this doctoral research provides another source of evidence. Further downstream hypothesis testing in a formal observational or if possible, randomised, setting would be crucial to further clarify the strength and direction of the association.

8.2.2.3 Association analyses

While many models exist for association analyses in the observational setting for healthier patients, when cohorts are lesser understood, conventional modelling strategies do not work. The confounding adjustment strategies may be weak for a variety of reasons: (1) the understanding of risk/protection in the context of a particular exposure-outcome relationship may be poorly studied, (2) the population may be very unhealthy

(e.g., high number of comorbidities, high BMI, etc) implying conventional confounders selection may be insufficient for comprehensive identification of confounding variables, (3) confounding variables may be missing/recorded late. While data collection may have to be improved to fix reason (3), given enough data and modalities of EHR data, data-driven and automatic confounder selection as opposed to conventional expert-driven confounder selection may help address issues, (1) and (2). Given the proposed deep learning model is fitting adequately (Chapter 4) and the model can be trusted (Chapter 5), the model can be amended for data-driven causal/association modelling.

The developed model, Targeted-BEHRT is a major focal point and strength of this thesis (Chapter 6)[170]. In theory, the model, utilising both deep learning adjustment processes and doubly-robust estimation for selection bias mitigation, is a robust approach for causal inference and association analyses. Furthermore, in simulation experiments, the model estimated ground truth more accurately than benchmark statistical models[170]. Furthermore, the ablation analyses and finite-sample estimation experiments further illuminated that the unsupervised modelling provided greater utility in large/rich datasets, and the CV-TMLE procedure provided greater utility in finite sample data situations[170]. Also, with utilisation of both longitudinal and static variables, the model is an appropriate choice for association analyses when confounding adjustment is difficult due to reasons (1) and (2) described above. In sum, a major strength of works presented in Chapters 6 and 7 is the deep learning approach capable of extracting and adjusting for confounding factors in rich annotated EHR. With more comprehensive adjustment of confounders at baseline, the studies converge towards operating in the "strong ignorability" setting.

In addition, in this doctoral thesis, the statistical models were also implemented to directly compare with the proposed approach. The results from benchmark statistical models directly demonstrated the utility of conventional confounders (and in section 7.3,

pre-exposure variable) selection. Specifically, in both investigations of SBP, the Targeted-BEHRT modelling reject the J-curve hypothesis while the conventional modelling with expert selected confounders demonstrated a statistically significant J-shaped pattern (Sections 7.1 and 7.2). Also, in the investigations of paracetamol and cardiovascular/all-cause mortality outcomes, even adjustment of over 50 variables in the conventional modelling paradigm was insufficient to dismiss elevated risk. The association was quantitatively larger than the Targeted-BEHRT approach likely implying residual confounding in play in the case of the conventional modelling. However, in the investigations of paracetamol and SBP, the estimation from conventional modelling was qualitatively indifferent from that of the deep learning approach.

Furthermore, in the case of point (3) when confounders are missing/sparsely recorded, a mixed strategy of automatic and expert-driven/conventional confounder selection must be conducted. Targeted-BEHRT allows for this hybrid strategy; in the case of paracetamol of all-cause mortality, manually including variables recorded post-exposure but likely recorded late and pre-exposure as static variables explicitly in the Targeted-BEHRT model was instrumental to accurately modelling the clinically validated DAG (section 7.3). This example further demonstrates that data-driven modelling left without supervision of subject area specialists is a faulty and even more importantly, a dangerous approach. Supervised data-driven modelling allowing for careful expert-driven selection of confounding variables is necessary for high-quality investigations.

In sum, the Targeted-BEHRT approach is a major strength of this thesis; the model has the ability to deliver trustworthy evidence specifically in association studies for which the scope of confounding is not fully understood. In collaboration with subject area specialists, reliable interpretation of model estimates can be conducted.

8.2.3 Limitations: Data

8.2.3.1 Noise and missingness

Missing and erroneous data are indeed a problem that needs to be carefully handled. While for deep learning modelling, imputation was not conducted, several variables in investigations in Chapter 7 underwent imputation for inclusion in statistical modelling. Also, since patient data variables are recorded in the primary and secondary care settings, certain diagnoses in medical history might bias number and frequency of future visits. As an example, given a patient is diagnosed with diabetes, they might be requested to come to the GP more frequently for management of diabetes and health checks. Furthermore, the same patient may be measured for SBP and BMI more frequently than others in the same age group due to their chronic illness. Another issue with CPRD research is that absence of a code in records is considered as the absence of the condition itself; while this is accepted practice for CPRD data, this may not be true[5]. Sometimes, the patient may not go to GP appointments, the GP/hospital may not record the event, or there may be other issues with regards to how the CPRD organisation organised the GP data. With all this being said, the validity and reliability of the CPRD dataset is well established and the dataset has been used for ground-breaking research into cardiovascular and associated diseases[5,104,107].

8.2.3.2 Standardization of data

In routine clinical data, not all of the data for all patients will be collected in the same way. Across patients, the recording of diagnoses, measurements, prescriptions might be subject to recording practices differing across GPs and hospitals. For this reason, the algorithms for case ascertainment are even more important whilst using datasets such as CPRD. Hence, the phenotyping algorithms are crucial for comprehensively assessing a particular exposure or outcome in a particular population. With adoption of the

established Caliber phenotyping algorithms as well as phenotyping algorithms for CPRD developed by Tran et al, standardisation issues were directly addressed in the doctoral research works.

8.2.3.3 Unrecorded conditions

Because of how the CPRD dataset collects data, some conditions are more difficult to identify in patients and hence harder to study. Specifically, since the variables are routine clinical care data from primary and secondary data sources, only those conditions, for which the patient seeks medical attention are more regularly and reliably recorded in the clinical care setting. Milder conditions (e.g., slight headache, light muscular pain) or diseases that have stigma associated with them often go unreported. In our research dysphagia is one such condition (Chapter 7); as both a mild condition in many cases and one that causes embarrassment, this condition is poorly recorded, recorded late, or wholly unreported[212,222–224]. In the case of conditions like dysphagia and others as confounding variables, this remains problematic for variable adjustment. However, an interdisciplinary research environment with data scientists, machine learning scientists, epidemiologists, and clinical experts can identify potential issues such as the ones illustrated in Section 7.3 and can work together to address them.

8.2.3.4 Other factors relevant to the data

CPRD only has patient records on patients who are using the primary care services. This ultimately means that the patients who are contributing to records in the CPRD dataset are those who understand that they must seek primary care services for the betterment of health. As a result, those who are not in the CPRD dataset might be unhealthier as they are rejecting preventative care, screening services, and the professional advice of health care professionals in the clinical care setting[4] While population selection bias endemic to the CPRD dataset may be an issue, the dataset is still

a far better source of health data for epidemiological studies than others; in other studies, participants have to explicitly wish to be included (i.e., opt in) for study ultimately exposing the study to other more problematic forms of selection bias[4].

Also, the CPRD dataset is not a good source of data to understand the views and perspectives of the patient. There is quite a discrepancy between patient's perception of illness and clinical diagnoses of the condition as reported in past prospective cohort studies[237]. For studies that rely on patient-reported symptoms, the self-reported data are often a better source of data-driven analyses than datasets like CPRD[237].

8.2.4 Limitations: Modelling

8.2.4.1 Risk prediction

In risk prediction studies including subsequent diseases and incident heart failure prediction presented in Chapter 4, there are certain methodological limitations in modelling. First, the phenotyping in the Caliber encoding simplifies the diagnoses into approximately 300 codes meaning that certain inaccuracies in disease identification manifest[122]. However, the causal inference and association studies conducted in Chapters 6 and 7 adopt more granular methods of representing the patient health (i.e., ICD-10 encoding). Additionally, only the diagnosis, measurement, and medication modalities were used in this thesis. While data for procedures were sparse in the cut of CPRD used in this thesis, further inclusion of procedures and even laboratory tests in addition to static variables (e.g., ethnicity) would be prudent for downstream research.

Second, in both investigation in Chapter 4, filtering was implemented to conduct risk prediction studies on an enriched set of patients – i.e., those with greater interactions with clinical care providers. While this is useful to assess model performance in a cohort of patients with more clinical interactions or perhaps even those with higher baseline risk

(by proxy of higher interactions with healthcare providers), this is not useful for assessing risk prediction in the general population[131].

Third, for more robust assessment of the BEHRT model's predictive ability and generalisability, validation was not conducted on another external dataset. The models were internally trained and tested in CPRD; while CPRD is a reliable and validated nationally representative dataset as discussed in Chapter 3, a more comprehensive test of model predictive ability would involve external validation. Hence, external validation is left as an important goal to be addressed in future works (see section 8.5).

Indeed, the models trained and tested in these settings are not suited for deployment in the clinical setting as further validation is needed. However, the risk prediction studies pursued in these studies were not intended for deployment in the clinical setting. In fact, the prediction of subsequent diseases – especially including conditions that may repeat – presented in the first half of Chapter 4 has little clinical significance and impact. Rather, BEHRT was built and compared to other convolutional and recurrent architectures in the two risk prediction studies to better understand, which models demonstrated promise for causal inference and association analyses. By analysing BEHRT's prediction performance measures across tasks, it was understood that BEHRT's feature extractor captured signal that aided risk prediction as compared to the convolutional and recurrent architectures.

8.2.4.2 Causal inference and association analyses

First, all of the four association analyses fail to fully satisfy the assumptions of conducting causal inference and hence not causal. In all four investigations, the works are conducted in the observational setting implying that residual confounding cannot be fully ruled out. Hence, the assumption of "strong ignorability" is not met. With this being stated, however, ignorability in the observational setting is a spectrum, and with access

to more informative variables concerning patient health and more robust estimation/adjustment methods, association studies converge towards meeting this assumption. Also, in the investigations of SBP and cardiovascular endpoints in at-risk patients, SBP fundamentally cannot be randomised directly implying that these works are not causal by nature.

Second, while the Targeted-BEHRT model was tested in a host of semi-synthetic data experiments, further experimentation is needed to validate the model. Similar to research presented in Chapter 6, further assessment of statistical and deep learning models needs to be conducted on two fronts: assessment of estimation with respect to 1) generated ground truth and 2) clinically validated/established ground truth. Consistent reliable estimation in both settings can better instil trust in more contemporary modelling solutions such as Targeted-BEHRT.

Third, the proposed Targeted-BEHRT model may be suffering from collider adjustment and other over-adjustment biases. Regarding colliders, study design is an important consideration; study designs were chosen that has a clear baseline with Targeted-BEHRT modelling health variables up to baseline. Hence, potential collider bias is already mitigated in this way with strict study design methods. Furthermore, for certain special colliders (e.g., the M-structure), these special colliders are actually quite uncommon in real world data settings. Nevertheless, the M-structure variables might indeed be a source of bias if adjusted[184]. However, empirical research has demonstrated that fully conditioning on all pre-exposure variables is still far more optimal avoiding adjustment of variables hypothesized to be M-structure variables[185]. Fundamentally, this is because effects of confounding malign the estimation far more heavily than hypothetical M-structure bias. Hence, the collider bias should not be a substantial source of bias in Targeted-BEHRT estimation.

Fourth, much like trusting deep learning models in the context of risk prediction, a fundamental question arises: can we trust Targeted-BEHRT and generally, deep learning modelling for causal inference and association analyses? The Targeted-BEHRT deep learning model builds on the fundamental BEHRT feature extractor[123]. In past works, the opaque, "black-box" BEHRT model has been shown to capture clinically meaningful and validated signal[131,170]. Specifically, the feature extractor can (1) capture known risk and preventative factors of cardiovascular conditions, (2) capture established progression and advances in pharmacotherapy across time (e.g., changes in treatment for glaucoma). Since Targeted-BEHRT's feature extraction builds on the BEHRT model (allowing for static variables as well), cracking open the "black-box" BEHRT architecture and probing the learning process also mitigates concerns about the Targeted-BEHRT's learning process[170]. Nevertheless, directly addressing the issue and conducting analyses that make the Targeted-BEHRT model more transparent would immensely benefit causal inference research and perhaps facilitate better acceptance of deep learning based causal modelling in epidemiological research as well.

Fifth, bias and confounding are endemic to observational studies. Hence, for the three observational studies presented in Chapter 7, a variety of sensitivity analyses were conducted to ensure the robustness of results. While for the SBP studies, the sensitivity analyses affirmed the results presented in the main analyses, the sensitivity analyses in the paracetamol investigation were particularly useful to illuminate the bias present in the analysis of the outcome of all-cause mortality. In this way, the natural limitations of modelling in the observational setting have been partially addressed. Lastly, residual confounding can never fully be captured in observational studies; orthogonal methods of addressing the research question (including randomised sources of evidence) can help in clarifying the nature of the association

8.2.4.3 Modelling complexity and environmental impact

For modelling with BEHRT and derivations, one of the foremost issues of the model are the number of parameters and training time. Because of the numerous parameters associated with multi-head self-attention architecture, the model is slow to train and requires large amounts of data[238]. The BEHRT model has several million parameters to be tuned implying that GPUs are necessary for efficient training. However, further research is needed to understand the necessity of this high parameter count and if more efficient versions of the model can be made that can retain predictive/estimation performance. In terms of data, CPRD indeed is a good source of data for such investigations; however, the training time required for such large amounts of data is in the unit of days and not hours. Hence, research must be better conducted on making model training more efficient in terms of time and computational complexity. While some solutions exist for accelerating model training (e.g., PyTorch lightning), these solutions must be thoroughly investigated for deep learning for EHR data[239].

Additionally, there are environmental costs when conducting deep learning modelling[238]. As an example, training a large Transformer model just once leaves a carbon footprint roughly 1/10th of that of a passenger travelling from New York to San Francisco[238]. While training a model just once may be acceptable and even negligible, during the course of this doctoral research, hundreds if not thousands of experimental models are trained and fine-tuned for prediction or association estimation. Hence, the environmental impact is indeed not nominal. However, in this doctoral research, the models are much smaller than those used in NLP research. In terms of vocabulary, the EHR vocabulary is only a few thousand disease/medication concepts while the original BERT model is trained on more than 30,000 English words; downsizing vocabulary ultimately directly means that there are fewer parameters to be trained (i.e., fewer

encounter embeddings) and shorter training times. Furthermore, while initial BEHRT model development in Chapter 4 were conducted with 12 layers and 12 attention heads for the BEHRT model much like the heavier BERT model (i.e., the $BERT_{BASE}$ model) implementations, in the majority of the following works presented in this thesis, only 4 layers were used with 6 attention heads for both BEHRT and Targeted-BEHRT models (Chapters 5, 6, and 7) [119,123]. Hence, while the environmental costs of deep learning are generally high, there have been efforts to mitigate the computational complexity and hence the environmental impact in this thesis for EHR related studies.

Furthermore, as stated in section 8.2.2.3, for causal modelling specifically, if the association studied is in relatively healthy populations, the recommendation offered is to use conventional models. These models are simpler to code, train, and evaluate, and additionally, leave a much lower carbon footprint. Hence, while the modelling proposed in this doctoral research is valuable in certain observational settings, the conventional modelling and associated benefits must be capitalised upon where appropriate.

8.3 Implications for methodological research

This doctoral research has contributed methods to the fields of deep machine learning, explainability, and causal inference.

In deep learning, while the Transformer and BERT model was developed for natural language processing tasks in mind, Transformers for EHR were not yet developed at the onset of the doctoral study[86,119]. Given the natural differences between EHR and text data, the modelling of the record data are also different. BEHRT directly handled the unique nature of EHR with its embedding structure[123]. Furthermore, the embedding structure was extended to not only include calendar year as a feature of absolute time in

the heart failure prediction studies, but also include static features in Targeted-BEHRT studies[131,170].

In terms of explainability, at the time, there were few methods for "explaining" Transformer models. While work by Guan et al helped pave the way for my explainability research, this perturbation-based method was expanded with a novel weighted loss function[131,133]. Given our understanding of risk and protection, this customisation was necessary for modelling the explainability studies appropriately. Furthermore, the relative contribution, to my understanding, is a novel metric that quantifies contribution of a variable to final predictive probability. Rather than providing an absolute view of predictor contribution, the relative contribution metric provides how much greater is the contribution in those predicted to be at risk of the outcome versus those who are not. While absolute measures of contribution for some predictors may be negligible, the relative contribution allows for more nuanced analyses of all predictors (Chapter 5).

For causal inference and association analyses, in terms of both modelling and data, there have been notable contributions to methodological research. First, the model Targeted-BEHRT is a contribution to causal inference methodological research and brings together many advancements in deep learning and statistics in one unified structure for more accurate estimation of causal effect. The unsupervised learning element of the model aids confounding adjustment as shown in the ablation analyses (Figure 6-3). Furthermore, the tandem propensity score estimation and outcome prediction allows for an end-to-end framework for generating initial estimates and propensity scores for doubly-robust estimation via CV-TMLE ultimately mitigating selection biases as shown in finite-sample estimation experiments (Figure 6-4). Also in another way, in the survey of past literature, BEHRT and other known EHR deep learning models (e.g., RETAIN, Deepr) have not been extended for EHR-driven causal inference or association analyses.

In my view, this is the first expansion of multimodal EHR deep learning models for conducting causal inference and association analyses. In downstream application works, I also demonstrate that the Targeted-BEHRT approach can be flexibly and easily applied to cardiovascular epidemiological research questions. Lastly, the development of the semi-synthetic data environment is another contribution to the field. While past works have indeed developed simulation datasets for testing models, the presented research is the first in our knowledge that derives semi-synthetic data from large-scale, multimodal EHR. Furthermore, this synthesizes propensity score-based data simulation and multimodal EHR processing in a novel way.

8.4 Implications for medical sciences

The research presented has been developed from its infancy in theory to application in epidemiological research. While the Targeted-BEHRT model has initially been explored in a sandbox environment with strict adherence to framework of causal inference, the model has been implemented in several association studies in the observational capacity.

Specifically, the model has been implemented to study four associations. While the evidence on the association between antihypertensives and cancer is well established due to well conducted meta-analyses on randomised studies, observational studies have presented more conflicting results. The work presented in this thesis is the first work that presents well-adjusted observational evidence on the matter that is in line with meta-analyses of randomised investigations.

Furthermore, the model was used to better disentangle the relationship between SBP and cardiovascular outcomes in at-risk patients with diabetes and COPD. The Targeted-BEHRT model provided comprehensively adjusted evidence on the association

rejecting the J-curve hypothesis in both cases across both primary and secondary outcome investigations and numerous sensitivity analyses. This advances our understanding of SBP and the relationship it has with cardiovascular endpoints in those at heightened risk and pre-existing conditions. With lowest risk at <120 mm Hg, the research confirms that the relationship remains log-linear in these sicker cohorts in addition to those at lower risk (i.e., general population).

Furthermore, due to issues of limited power, associations between blood pressure lowering medications and cardiovascular outcomes in those with pre-existing disease are likely to remain untested in the randomised nor in the individual level meta-analysis setting. For this reason, the deep learning-driven association analyses provide well-adjusted observational evidences for re-examination of guideline recommendations for blood pressure lowering treatments in at-risk patients. With lowest risk at <120 mm Hg, the presented research affirms "the lower, the better" paradigm for hypertension management. On the whole, further independent examination of this association in other datasets would also be valuable for ascertaining the generalisability of the results presented. Lastly, while solely observational evidence is insufficient for revising the guidelines, observational evidences in the past have indeed been seminal for the re-examination and discussion of guidelines if not the revision itself[240].

In the analyses of paracetamol and various outcomes, the results demonstrate that, for the most part, any excess risk of any outcome is a result of confounding and bias as opposed to actual signal. Furthermore, this observational work has clinical implications concerning treatment of the elderly; given that elderly patients often have little recourse but to take effervescent forms of medication, this research allays fears of excess risk associated with taking this medication. Furthermore, much like the case with high-risk individuals, it is unlikely that this association can and will be tested in a randomised

setting. It is all the more important that Targeted-BEHRT be implemented to study these associations. Also, given the noisiness of dysphagia recording, this research raises an important point for observational research on EHR: datasets like CPRD are more appropriate to study conditions that are not considered "mild" or go under-reported (raised in 8.2.3.3). It is important to only derive significance for clinical care when the study is fully appreciative of the strengths and weaknesses of the underlying EHR.

8.5 Future directions

There are several directions for future research. For risk prediction investigations, the binary prediction setting has several drawbacks for precision medicine research. With a binary label indicating presence or absence of the desired outcome, the nuances of the time of the outcome in the follow-up time window are ignored in analyses. Furthermore, patient censoring is not appropriately accounted in modelling. Hence, the more optimal method of modelling risk is with survival models in the time-to-event setting. There have been several initial explorations of deep learning models for prediction in the survival modelling setting; however, more research needs to be conducted to amend models for EHR data-based risk prediction tasks[241]. Future cardiovascular research should investigate the utility of the modifying the BEHRT model for the time-to-event prediction setting.

Furthermore, the external validation of risk prediction models is a necessity. The While CPRD was the dataset of choice for this doctoral research and was crucial for model development and testing, future research should focus on validating the models presented on different EHR data. Also, this ultimately implies the vocabulary for representing encounters must be harmonised across data settings. While the Caliber disease phenotyping has been validated for CPRD, it has not been extensively tested for

other datasets[122]. On the other hand, for encoding disease data, the ICD-10 or SNOMED disease encoding is a far more universal dictionary for identifying conditions across electronic health recording systems around the world[242]. For the medications, while BNF is UK-specific, more universal medication encodings such as the "RxNorm" encoding exists and mapping from BNF to RxNorm is offered through validated data dictionaries[243]. With these advancements in phenotyping, validating BEHRT and similar models in different data settings may be more feasible.

Furthermore, for causal inference and association analyses as well, developing causal models in the time-to-event setting is crucial for more sensitive modelling of risk. While this implies that first, survival DL models must be built for representation learning and risk prediction as previously discussed, the models should be amended for association analyses in the time-to-event setting. While the proportional hazards framework may be useful for deep learning, several other frameworks exist (e.g., logistic hazards) that do not rely on the assumptions and conditions for proportional hazards modelling exist and must be fully explored[244].

Also, further investigation of heterogenous treatment effect estimation must be pursued. While doubly-robust estimators discussed in this doctoral research and as an extension, Targeted-BEHRT, are not intended for heterogenous treatment effect estimation, there have been some recent explorations of doubly-robust estimators for stratified effects estimation[245]. Furthermore, other methods exist for stratified analyses and must be appropriately amended for EHR and rigorously tested in the EHR data setting[166].

Also, while external validation is not per se recommended by clinical guidelines, further research into SBP and paracetamol must be conducted and would be beneficial to advancing knowledge concerning generalisability of findings. Specifically, while in this

doctoral research, the relationship between SBP and cardiovascular endpoints were explored, future investigations should directly investigate the effect of blood pressure lowering medications on cardiovascular endpoints in high-risk patients perhaps using the trial emulation framework for designing the study[246].

8.6 Conclusions

This doctoral research advances knowledge in cardiovascular disease research through methodological advances in risk prediction, explainability, and causal inference. First, the research develops the BEHRT model and presents its utility in the risk prediction setting. Second, while the proposed deep learning model indeed demonstrated better prediction performance than benchmark models, the explainability study of the model instilled more confidence in the deep learning approach. Third and finally, Targeted-BEHRT demonstrated that Transformer-driven confounding adjustment in tandem with targeted learning for causal analyses enables better confounding adjustment and more accurate estimation of associations in the observational setting.

In terms of epidemiological and clinical impact, the research has advanced risk prediction for cardiovascular studies – specifically, the risk prediction of incident HF. Furthermore, the explainability investigations of the "black-box" BEHRT model for incident HF prediction has independently captured known risk factors and generated other factors of potential protective utility that could be formally assessed in future hypothesis testing investigations. Also, the Targeted-BEHRT model has advanced understanding of SBP, antihypertensives, paracetamol, and CVD; especially, in the context of elderly or at-risk patients, while conventional approaches fall short of appropriately adjusting for complex confounding, the deep learning Targeted-BEHRT approach provided well-adjusted evidences in the observational setting.

9 APPENDICES

In the following sections, the supplementary material relevant to the body of the thesis (i.e., Chapters 4, 5, 6, and 7) is presented ordered by chapter. A full accounting of tables and figures for these supplementary sections are provided below. Lastly, all code from the research projects presented can be found on both the Deep Medicine group GitHub site (https://github.com/deepmedicine) and my GitHub site (https://github.com/srn284).

LIST OF SUPPLEMENTARY TABLES

LIST OF SUPPLEMENTARY FIGURES

9.1 Supplement for Chapter 4: Model Development and Risk Prediction

9.1.1 Deep learning modelling for electronic health records: model development and subsequent disease prediction

Supplementary tables from section 4.2 is found below.

Table S1: Hyperparameters for BEHRT model

Iteration	*Hidden size*	*Layers*	*Attention heads*	*Intermediate Size*	*Precision*
1	216	3	6	256	0.6191
2	288	9	12	512	0.6399
3	216	3	12	512	0.6175
4	432	3	18	512	0.6397
5	288	6	6	784	0.638
6	216	6	18	512	0.6262
7	288	3	18	512	0.6292
8	432	3	6	784	0.6426
9	288	6	12	512	0.6356
10	288	3	12	256	0.6283
11	432	9	18	512	0.6466
12	576	9	6	1024	0.6538
13	432	3	18	1024	0.6411
14	432	9	6	1024	0.6508
15	576	6	6	256	0.6503
16	576	6	12	256	0.651
17	360	9	18	512	0.6404
18	576	9	6	512	0.6513
19	288	6	6	512	0.6363
20	288	3	6	512	0.6297
21	288	6	12	512	0.6597
22	576	3	12	512	0.6487
23	360	6	6	784	0.6412
24	432	9	6	512	0.6497
25	360	6	12	512	0.6423

The table was adapted from Li et al[123]*.*

Table S2: Hyperparameters for Deepr model

Iteration	*Filters*	*Kernel Size*	*FC 1*	*FC 2*	*FC 3*	*Dropout 1*	*Dropout 2*	*Dropout 3*	*Learning rate*	*Average Precision*
1	37	7	10	46	28	0.4139	0.4997	0.3718	0.0004	0.2599
2	24	5	28	19	13	0.4936	0.1722	0.4643	0.0062	0.2319
3	17	7	16	48	40	0.225	0.3945	0.3264	0.031	0.1815
4	33	7	6	10	25	0.1602	0.4476	0.3544	0.0026	0.264
5	32	7	4	30	32	0.3714	0.3382	0.3573	0.0353	0.2005
6	13	4	50	17	47	0.3424	0.183	0.3056	0.0008	0.3274
7	12	3	3	6	43	0.4201	0.2387	0.3884	0.0123	0.2112
8	30	7	16	26	41	0.2055	0.3343	0.1541	0.0011	0.3256
9	35	5	50	24	25	0.1567	0.4056	0.1329	0.0004	0.3433
10	41	7	9	35	19	0.4885	0.3548	0.308	0.0004	0.2343
11	4	4	33	12	39	0.1811	0.1251	0.1066	0.001	0.3051
12	48	4	36	45	37	0.2143	0.3486	0.1222	0.0015	0.3504
13	36	3	39	10	48	0.1992	0.4164	0.1183	0.005	0.3291
14	45	4	44	40	26	0.2329	0.29	0.128	0.0903	0.0042
15	40	4	35	48	11	0.16	0.4704	0.1211	0.0019	0.3125
16	49	3	47	41	40	0.1002	0.2541	0.1284	0.0019	0.3588
17	47	3	50	47	12	0.2519	0.2125	0.1668	0.0005	0.2988
18	47	3	37	35	50	0.1059	0.4225	0.112	0.0034	0.3487
19	47	4	48	34	39	0.2177	0.3607	0.1301	0.0016	0.3567
20	46	3	46	45	43	0.2007	0.1609	0.1196	0.0044	0.3356

FC: Fully connected layer. The table was adapted from Li et al[123].

Table S3: Hyperparameters for RETAIN model

Iteration	*Embedding size*	*Recurrent Size*	*Dropout embedding*	*Dropout context*	*L2*	*Average precision*
1	142	90	0.3846	0.0224	0.0891	0.1822
2	124	43	0.3922	0.0382	0.0003	0.1815
3	173	90	0.3929	0.2238	0.0014	0.3479
4	145	91	0.4117	0.0404	0.0102	0.2049
5	153	92	0.4569	0.0642	0.0116	0.2049
6	120	37	0.4335	0.4017	0.0728	0.1815
7	180	102	0.3567	0.3307	0.0039	0.2469
8	195	38	0.3805	0.2711	0.001	0.374
9	165	119	0.4969	0.1964	0.0047	0.2292
10	174	92	0.3862	0.2078	0.0019	0.3329
11	145	110	0.3928	0.1926	0.0891	0.1813
12	195	83	0.4418	0.4787	0.0011	0.3543
13	187	110	0.3456	0.2083	0.0123	0.2122
14	144	80	0.3717	0.0932	0.0032	0.3038
15	193	68	0.4528	0.395	0.0022	0.328
16	198	45	0.4344	0.3324	0.0442	0.1828
17	145	64	0.4213	0.195	0.0626	0.1813
18	171	116	0.4166	0.495	0.0028	0.3197
19	186	38	0.4062	0.4916	0.0011	0.3309
20	136	54	0.3503	0.1678	0.0067	0.2162

L2 is regularization weight. The table was adapted from Li et al[123].

Table S4: BEHRT prediction performance metrics for individual diseases

Description	*APS*	*AUROC*	*Ratio*	*Chapter*
Gastritis and duodenitis	0.066765	0.723828	0.011198	Diseases of the digestive system
Diaphragmatic hernia	0.108185	0.797633	0.01149	Diseases of the digestive system
Hearing loss	0.118093	0.742646	0.021964	Diseases of the ear and mastoid process
Spondylosis	0.132567	0.773249	0.013459	Diseases of the musculoskeletal system and connective tissue
Pleural effusion	0.142594	0.844596	0.010229	Diseases of the respiratory system
Other anaemias	0.162186	0.798654	0.023303	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
Bacterial Diseases (excluding TB)	0.163355	0.822707	0.023979	Certain infectious and parasitic diseases
Iron deficiency anaemia	0.167457	0.804545	0.02078	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
Urinary Tract Infections	0.169674	0.83661	0.022534	Diseases of the genitourinary system
Diverticular disease of intestine (acute and chronic)	0.170852	0.811759	0.015966	Diseases of the digestive system
Allergic and chronic rhinitis	0.17717	0.810068	0.023241	Diseases of the respiratory system
Osteoporosis	0.181182	0.847938	0.013982	Diseases of the musculoskeletal system and connective tissue
Dyslipidaemia	0.184	0.790655	0.02601	Endocrine, nutritional and metabolic diseases
Oesophagitis and oesophageal ulcer	0.184629	0.799592	0.022426	Diseases of the digestive system
Primary Malignancy Other Skin and subcutaneous tissue	0.203218	0.855593	0.012013	Neoplasms
Gastro-oesophageal reflux disease	0.206455	0.776367	0.026271	Diseases of the digestive system
Type 1 Diabetes Mellitus, Type 2 Diabetes Mellitus, and Diabetes Mellitus – other or not specified	0.208524	0.814481	0.021226	Endocrine, nutritional and metabolic diseases
Actinic keratosis	0.219637	0.869366	0.01249	Diseases of the skin and subcutaneous tissue
Irritable bowel syndrome	0.220017	0.874319	0.011182	Diseases of the digestive system
Urinary Incontinence	0.223863	0.824114	0.020057	Diseases of the genitourinary system
Osteoarthritis (excluding spine)	0.234444	0.785766	0.043714	Diseases of the musculoskeletal system and connective tissue
Other or unspecified infectious organisms	0.245906	0.834062	0.030963	Diseases of the respiratory system
Glaucoma	0.249208	0.894444	0.011367	Diseases of the eye and adnexa
Hyperplasia of prostate	0.250573	0.885547	0.020842	Diseases of the genitourinary system
Peripheral arterial disease	0.264687	0.879325	0.010951	Diseases of the circulatory system
Enthesopathies	0.265863	0.762939	0.047036	Diseases of the musculoskeletal system and connective tissue

Erectile dysfunction	0.267187	0.905812	0.017873	Mental and behavioural disorders
Lower Respiratory Tract Infections	0.268504	0.867094	0.023518	Certain infectious and parasitic diseases
Dermatitis (atopic/contact/other/unspecified)	0.271027	0.753816	0.049051	Diseases of the skin and subcutaneous tissue
Macular degeneration	0.292598	0.893802	0.010752	Diseases of the eye and adnexa
Stable Angina	0.296798	0.889236	0.032039	Diseases of the circulatory system
Coronary heart disease not otherwise specified	0.301041	0.900088	0.035177	Diseases of the circulatory system
Stroke Not otherwise specified (NOS)	0.307238	0.911618	0.023395	Diseases of the nervous system
Cataract	0.319433	0.863447	0.042099	Diseases of the eye and adnexa
Abdominal Hernia	0.319972	0.845171	0.01918	Diseases of the digestive system
Carpal tunnel syndrome	0.325143	0.84348	0.012013	Diseases of the nervous system
Heart failure	0.334902	0.912117	0.024918	Diseases of the circulatory system
Obesity	0.335131	0.87967	0.017442	Endocrine, nutritional and metabolic diseases
Asthma	0.348523	0.885055	0.026133	Diseases of the respiratory system
Gout	0.349361	0.882694	0.018058	Diseases of the musculoskeletal system and connective tissue
Diabetic ophthalmic complications	0.350132	0.942604	0.018919	Endocrine, nutritional and metabolic diseases
Migraine	0.368465	0.911541	0.012028	Diseases of the nervous system
Psoriasis	0.398842	0.904686	0.011751	Diseases of the musculoskeletal system and connective tissue
Anxiety disorders	0.410899	0.858498	0.041914	Mental and behavioural disorders
Menorrhagia and polymenorrhoea	0.433645	0.969406	0.015504	Diseases of the genitourinary system
Hypo or hyperthyroidism	0.489456	0.905032	0.047897	Endocrine, nutritional and metabolic diseases
Vitamin B12 deficiency anaemia	0.491672	0.855823	0.014489	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
Chronic obstructive pulmonary disease (COPD)	0.501496	0.923082	0.036869	Diseases of the respiratory system
Atrial Fibrillation and flutter	0.514881	0.901268	0.077629	Diseases of the circulatory system
Hypertension	0.531597	0.819527	0.200618	Diseases of the circulatory system
Dementia	0.542223	0.950442	0.024656	Mental and behavioural disorders
Depression	0.553561	0.877904	0.076876	Mental and behavioural disorders
Female genital prolapse	0.57388	0.934049	0.015781	Diseases of the genitourinary system
Primary Malignancy Prostate	0.575574	0.964776	0.011844	Neoplasms
Alcohol Problems	0.583305	0.952656	0.014535	Mental and behavioural disorders
Polymyalgia Rheumatica	0.647243	0.955062	0.013213	Diseases of the musculoskeletal system and connective tissue
Epilepsy	0.648763	0.977907	0.016104	Diseases of the nervous system

APS: Average precision score; AUROC: Area under the receiver operator characteristic. The table was adapted from Li et al[123].

Table S5: Sex-based analysis of BEHRT predictions

Caliber phenotype	*Sex*	*Male prediction (count)*	*Female prediction (count)*
Hyperplasia of prostate	M	384	0
Hydrocoele (including infected)	M	36	0
Male Infertility	M	1	24
Primary Malignancy Prostate	M	557	0
Erectile Dysfunction	M	425	1
Menorrhagia and polymenorrhoea	F	0	697
Endometriosis	F	0	47
Female Genital Prolapse	F	0	865
Female Infertility	F	2	36
Benign neoplasm of ovary	F	0	69
Postmenopausal bleeding	F	0	140
Primary Malignancy Breast	F	0	11
Primary Malignancy Ovarian	F	1	193

M: male, F: female. The table was adapted from Li et al[123].

9.2 Supplement for Chapter 5: Explainability

9.2.1 Explanation of the perturbation-based method for explainability

In sections below, the perturbation method for the explainability analyses is described.

In addition to the methods proposed by Guan et al, we developed an asymmetric loss function for focused learning of a specific objective in addition to the original information entropy-based loss term[133]. The objective function is shown below in the equation (9-1).

$$\boldsymbol{alpha}(\boldsymbol{y},\boldsymbol{x},\boldsymbol{s}) = \begin{cases} \boldsymbol{\beta_1}, & \boldsymbol{if\ y = 1, \mathbf{M}(\tilde{x}) - s \geq 0} \\ \boldsymbol{\beta_2}, & \boldsymbol{if\ y = 0, \mathbf{M}(\tilde{x}) - s \leq 0} \\ \boldsymbol{\beta_3}, & \boldsymbol{otherwise} \end{cases} \tag{9-1}$$

$$\boldsymbol{L(\sigma) = alpha(y,x,s) \times \mathbf{E}_{\epsilon}\|(\mathbf{M}(x) - s)\|^2\ \lambda \sum_{i=1}^{n} H(\tilde{x}\,|s)|_{\epsilon_i \sim \mathcal{N}(0,\sigma_i^2 I)}}$$

In this equation, $\tilde{x}$ represents the input encounter embedding that has been perturbed (the original is *x*), the value, *n,* represents the number of encounters in patient medical history, and the value, *s*, represents the latent output state of the original input (*x* i.e., the input without perturbations), $\mathrm{M}(\tilde{x})$ is the output latent state of perturbed input, the values of $\beta_1 \; and \; \beta_2$ are weight hyperparameters with the weights designed such that ($\beta_1 < \beta_2$). If the values equal one another ($\beta_1 = \beta_2$), then the loss function becomes a symmetric loss formulation. *y* is the outcome (heart failure incidence). The value, $\mathbb{E}_\epsilon \|\mathrm{M}(\tilde{x}) - \mathrm{s}\|^2$ represents mean squared error described in Guan et al[133].

The equation in words can be explained as the following: when the intention is to understand contribution of the prediction to this with heart failure in follow-up (label=1), the perturbations to be prioritized are those that increase the outcome probability that those that decreases the probability – i.e., how much more confident (closer to 1.0 prediction) can prediction be if input is perturbed. However, if the patient does not get heart failure in follow-up (label=0), then we want to see how much more confident can the prediction of non-heart failure be (i.e., closer to 0.0 prediction) as a function of perturbing the input space. Hence, we penalize the loss function with the above $alpha(y, x, s)$ term; asymmetric losses are often used in modelling situations, for which the error in one direction is "worth" more than the in the other direction.

This method outputs certain elements that is discussed here. The learned $\epsilon = [\epsilon_1, \epsilon_2, \ldots, \epsilon_n]$ with ϵ_1 per predictor is the delivered quantity in this training process. This ϵ_1 in this training process is the allowable variance that indicated predictor contribution; the maximal allowable variance is defined by a user-specified hyperparameter value (in this work, the hyperparameter is set to 0.5). To actually derive the individual-level, predictor embedding contribution, we transform the value, ϵ_1 to the value of 0.5-ϵ_1 in order to reflect the following understanding of contribution: the lower the value of ϵ_1, the higher the

contribution of the predictor to the heart failure prediction. The vice-versa statement holds true here as well. Specifically, this is elucidated in the algorithm showed as Equation(s) (5-1). Furthermore, as seen in Figure 5-1, we first derive the individual-level contribution and then derive the population-based understanding of the relative contribution presented in the following figures in the chapter.

9.3 Supplement for Chapter 6: Causal Inference and Association Analyses

9.3.1 Semi-synthetic data generation

Data generation of the sequential, temporal variables is a difficult task; it is well noted that there are few (if any) medically validated methods of generating realistic EHR medical history. Hence, in gist, I utilise routine clinical data components in medical history data: (1) medical history and (2) known exposure status – observed exposure status in medical history. With these two constructs, the potential outcomes can be born, and with the potential outcomes – both counterfactual and naturally, factual, outcomes, a ground truth risk ratio (RR) can be synthesized. Comparison of model estimation accuracy can be conducted with respect to this ground truth RR.

In order to create this semi-synthetic dataset, we first form the dataset for the investigation: effect of antihypertensives on incident cancer allowing us access to components (1) and (2). Since confounding often manifests partly due to imbalanced variables between exposure groups, we find an imbalanced variable, Z_i in medical history. We then force this imbalanced variable, Z_i to be a confounder and generate conditional outcome from a sampling function:

$$\boldsymbol{Y^{T_i}{}_i = Bernoulli\left(\sigma\left(aT_i + m\beta(\lambda_i + c)\right)\right)} \quad (9\text{-}2)$$

In Equation (9-2), λ_i represents $P(T_i|Z_i)$, $T_i \in \{0,1\}$ is the value that represents the intervenable exposure status for patient *i*, $Y^{T_i}{}_i$ is the outcome for patient i given exposure T_i, σ is the sigmoid function, and β, the intensity of confounding. Variables *a*, *m*, and *c* are coefficients to terms weighting their importance in the function.

Intuitively, we first model the association between a variable Z_i and exposure (($P(T_i|Z_i)$) with λ_i. Next, we generate $Y^{T_i}{}_i$ with two variables: the variable Z_i and T_i. In this way, we form an association between Z_i and exposure and Z_i and the outcome; with association to both exposure and outcome, Z_i becomes a confounder in this data generating process. This process synthesizes controlled confounded observational data; by generating the outcome with this function, we control confounding with a confounder Z_i. Thus, in this way we can generate factual/counterfactual outcomes and consequently ground truth RR. Lastly, we can modify β value to vary the degree of confounding in the data generation process.

9.3.2 Statistical model development and adjustment

RR was the estimand of interest for statistical modelling as well. The covariates adjusted were: baseline age (continuous variable: [0,1]), sex (male/female), region, smoking status (smoker/non-smoker), chronic kidney disease (presence/absence), human immunodeficiency virus/acquired immune deficiency syndrome (presence/absence), ischaemic heart disease (presence/absence), cardiac arrhythmia (presence/absence), stroke (presence/absence), heart failure (presence/absence), anemia (presence/absence), diabetes mellitus (presence/absence), hypertension (presence/absence), osteoporosis (presence/absence), arthritis (presence/absence), connective tissue disorder (presence/absence), gout (presence/absence), rheumatoid arthritis (presence/absence),

peptic ulcer disease (presence/absence), liver disease (presence/absence), asthma (presence/absence), peripheral arterial disease (presence/absence), chronic obstructive pulmonary disorder (presence/absence), hemiplegia (presence/absence), epilepsy (presence/absence), dementia (presence/absence), learning disorder (presence/absence), eating disorder (presence/absence), adjustment (presence/absence), anxiety (presence/absence), affective disorder (presence/absence), depression (presence/absence), bipolar disorder (presence/absence), psychoses (presence/absence), schizophrenia (presence/absence), hyperlipidemia (presence/absence), obesity (presence/absence), substance abuse (presence/absence), anticholinergics (presence/absence), drugs that cause gastrointestinal bleedings (presence/absence), statins (presence/absence), drugs for diabetes (presence/absence). The exposure variable was antihypertensive medications (class 1 as control and class 2 as the exposure). The outcome was defined as the synthetic outcome/cancer (presence/absence). The models were fit and tested using five-fold validation. The naïve risk ratio estimates were calculated on the testing dataset in each fold and mean risk ratio (RR) estimate and 95% confidence intervals (CI) for estimates were derived.

The TMLE was a two-stage propensity score-based approach that utilised two logistic regression models. One of the models was for outcome prediction and the other for exposure prediction. The outcome prediction model adjusted for covariates and exposure variable listed above, and the exposure prediction model used just the covariates. The TMLE algorithm was fit and tested using five-fold validation. The TMLE RR estimates were calculated on the testing dataset in each fold and mean RR estimate and 95% (CI) for estimates were derived from the TMLE estimation procedure.

9.3.3 CV-TMLE

After using the Targeted-BEHRT model to compute initial estimates, the CV-TMLE estimation procedure was used to "correct" these initial estimates by removing selection bias. Source material for the TMLE and the cross validated form, CV-TMLE, can be found in publications by van der Laan[156,181]. In brief, the original formulation of the CV-TMLE algorithm requires targeting steps for each of the k folds pre-defined in the iterative version of TMLE. However, Levy forms a simpler construction of the CV-TMLE which is less computationally cumbersome; the advised method is to pool all the initial estimates across folds and compute the estimation update vis-à-vis a standard TMLE update step[156,181]. Albeit procedurally different, the original formulation and Levy's more recent formulation of CV-TMLE are mathematically identical. Upon scoping relevant literature, in our understanding, this is the first work utilizing CV-TMLE paired with deep learning methods.

CV-TMLE provides a host of benefits to observational causal inference. CV-TMLE is a form of TMLE which is robust to issues of fold-wise overfitting whilst conducting k-fold cross validation[181]. Furthermore, previous works show that the CV-TMLE estimator provides more robustness than other cross-validated estimators (e.g. CV-augmented inverse probability treatment weighting) in the case of violations of the assumption of overlap[247]. Lastly, operating with fewer assumptions than the original TMLE estimator, the CV-TMLE process does not require the necessity of assuming the Donsker class condition and hence allows for initial estimators of the nuisance parameters to be overfitted given a remainder term is asymptotically negligible[181].

9.3.4 Targeted-BEHRT modelling details

Table S6: Hyperparameters for the Targeted-BEHRT modelling

Hyperparameter	*Attribute*
Hidden size (BEHRT)	150
Intermediate size (BEHRT)	108
Region embedding size	7
Sex embedding size	1
Smoking status embedding size	2
Hidden dropout probability	0.3
Attention dropout probability	0.4
Number of hidden layers (BEHRT model)	4
Hidden activation functions	GeLU
Initialiser range of parameters	0.02
N – number of encounters	200
D – weight of the coefficient for the unsupervised learning component	0.1

The table was adapted from Rao et al[170].

9.3.5 Semi-synthetic data experimentation results

Table S7: Estimation of risk ratio on semi-synthetic data experiments

Confounder and type of model	***Risk ratio ground truth and estimates from semi-synthetic data experiments***				
		Risk ratio			**Error**
Cardio-metabolic disease	**Beta**	**25**	**50**	**75**	
	Modelling				
	Ground Truth	2.207	2.727	3.178	1.555
	Empirical	2.532	3.251	3.883	1.365
Statistical	LR	2.398; (2.37, 2.43)	3.003; (2.97, 3.03)	3.569; (3.5, 3.64)	0.859; 0.02
	LR-L1	2.399; (2.37, 2.43)	3.005; (2.97, 3.04)	3.576; (3.51, 3.64)	0.867; 0.02
	LR-L2	2.399; (2.37, 2.43)	3.004; (2.97, 3.03)	3.574; (3.5, 3.64)	0.864; 0.02
	TMLE	2.411; (2.28, 2.54)	3.005; (2.87, 3.14)	3.622; (3.4, 3.84)	0.928; 0.04
	BART	2.398; (2.37, 2.43)	3.011; (2.98, 3.04)	3.592; (3.53, 3.65)	0.890; 0.07
Deep learning	TARNET	2.283; (2.21, 2.35)	3.183; (2.76, 3.6)	3.899; (3.43, 4.36)	1.254; 0.23
	TARNET + MEM	2.226; (2.14, 2.31)	2.719; (2.35, 3.09)	3.561; (2.94, 4.19)	0.41; 0.26
	Dragonnet	2.308; (2.12,2.50)	3.098; (2.57,3.62)	3.12; (2.55,3.69)	0.529; 0.25
	Dragonnet+CV-TMLE	2.281; (2.26, 2.31)	2.954; (2.91, 3.0)	2.922; (2.85, 2.99)	0.556; 0.05

	Targeted-BEHRT+CV-TMLE	2.263; (2.24, 2.29)	2.753; (2.71, 2.8)	3.227; (3.14, 3.31)	0.131; 0.05
			Risk Ratio		**Error**
Sex	**Beta**	**1**	**5**	**10**	
	Modelling				
	Ground Truth	1.465	1.926	2.154	
	Empirical	1.456	1.823	1.979	0.287
Statistical	LR	1.455; (1.45, 1.46)	1.83; (1.81, 1.85)	1.996; (1.95, 2.04)	0.263; 0.01
	LR-L1	1.455; (1.45, 1.46)	1.83; (1.81, 1.85)	1.992; (1.95, 2.04)	0.267; 0.01
	LR-L2	1.455; (1.45, 1.46)	1.829; (1.81, 1.85)	1.993; (1.95, 2.04)	0.267; 0.01
	TMLE	1.453; (1.44, 1.47)	1.824; (1.79, 1.86)	1.982; (1.93, 2.03)	0.285; 0.01
Deep learning	BART	1.455; (1.45, 1.46)	1.826; (1.8, 1.85)	1.981; (1.94, 2.02)	0.282;0.02
	TARNET	1.465; (1.44, 1.49)	1.803; (1.71, 1.89)	1.948; (1.83, 2.02)	0.329;0.05
	TARNET+MEM	1.457; (1.44, 1.47)	1.863; (1.77, 1.96)	1.977; (1.72, 2.24)	0.247;0.1
	Dragonnet	1.479; (1.45,1.49)	1.969; (1.84,2.10)	2.439; (1.82,3.06)	0.342; 0.23
	Dragonnet+CV-TMLE	1.469; (1.45, 1.49)	1.827; (1.78, 1.87)	1.973; (1.85, 2.09)	0.285; 0.07
	Targeted-BEHRT+CV-TMLE	1.47; (1.45, 1.49)	1.854; (1.81, 1.9)	2.132; (1.98, 2.29)	0.1; 0.08

This table shows the risk ratio and standard deviation (five-fold) for statistical and deep learning models over the two semi synthetic experiments with cardiometabolic diseases and sex as confounders (top and bottom respectively). Over various values of Beta, confounding experiments are conducted. Ground truth risk ratio is calculated and displayed for both experiments. Risk ratio and 95% confidence interval for each model is presented in the table. The sum absolute error from ground truth risk ratios for models over all the confounding experiments and standard error is shown in the far-right column. LR: Logistic Regression; LR-L1; Logistic Regression with L1 penalty; LR-L2; Logistic Regression with L2 penalty; TMLE: Targeted Maximum Likelihood Estimation; BART: Bayesian Additive Regression Trees. The table was adapted from Rao et al[170].

9.4 Supplement for Chapter 7: Association analyses in at-risk patients

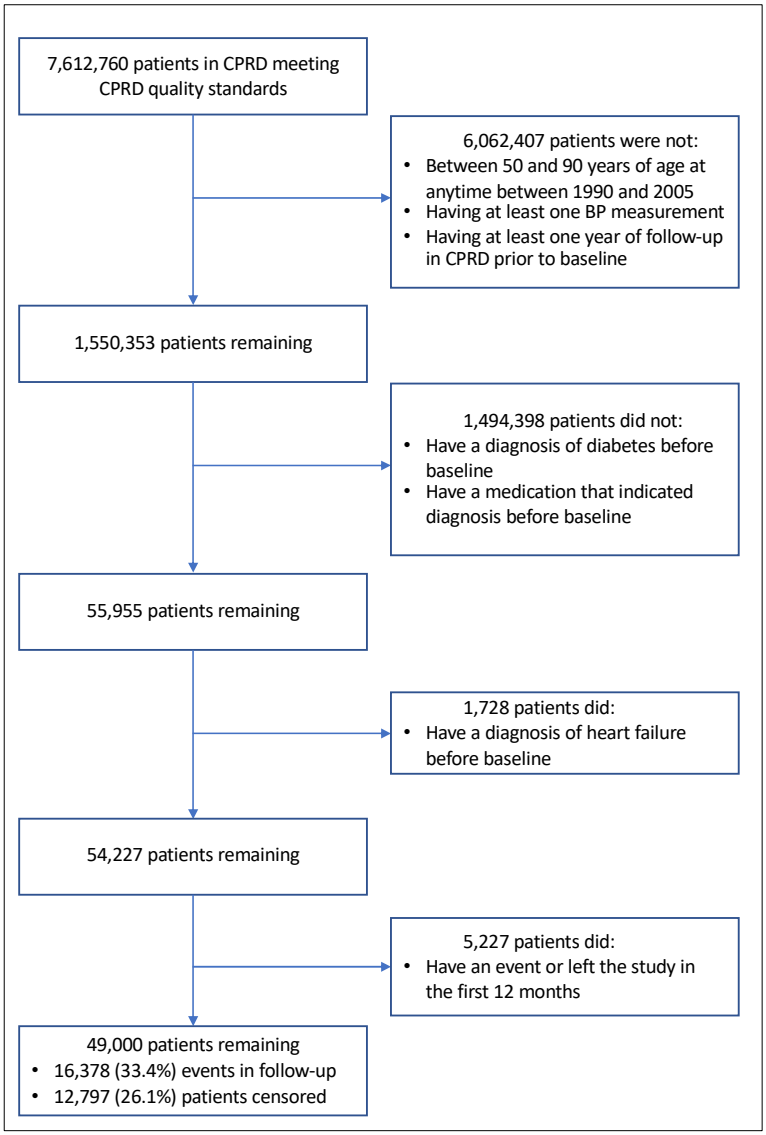
9.4.1 Systolic blood pressure, cardiovascular outcomes, and diabetes

Table S8: Statistics for primary and secondary outcome event rates stratified by exposure status

	Primary outcome	*IHD*	*HF*	*Stroke*
<120 mm Hg	32.52	23.2	7.72	9.49
120-129 mm Hg	32.01	22.05	7.12	10.03
130-139 mm Hg	33.43	22.4	7.68	10.68
140-149 mm Hg	36.49	24.07	8.89	12.23
150-159 mm Hg	39.56	25.9	10.17	13.94
≥160 mm Hg	45.12	27.99	14	16.65

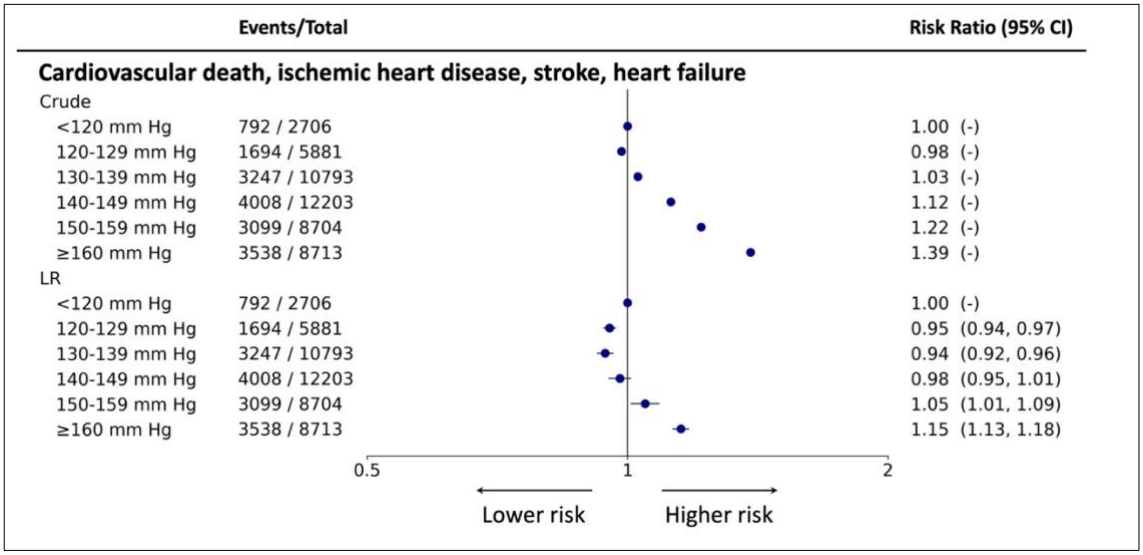
HF: heart failure; IHD: ischaemic heart disease.

Figure S1: Flowchart for cohort selection for patients with diabetes



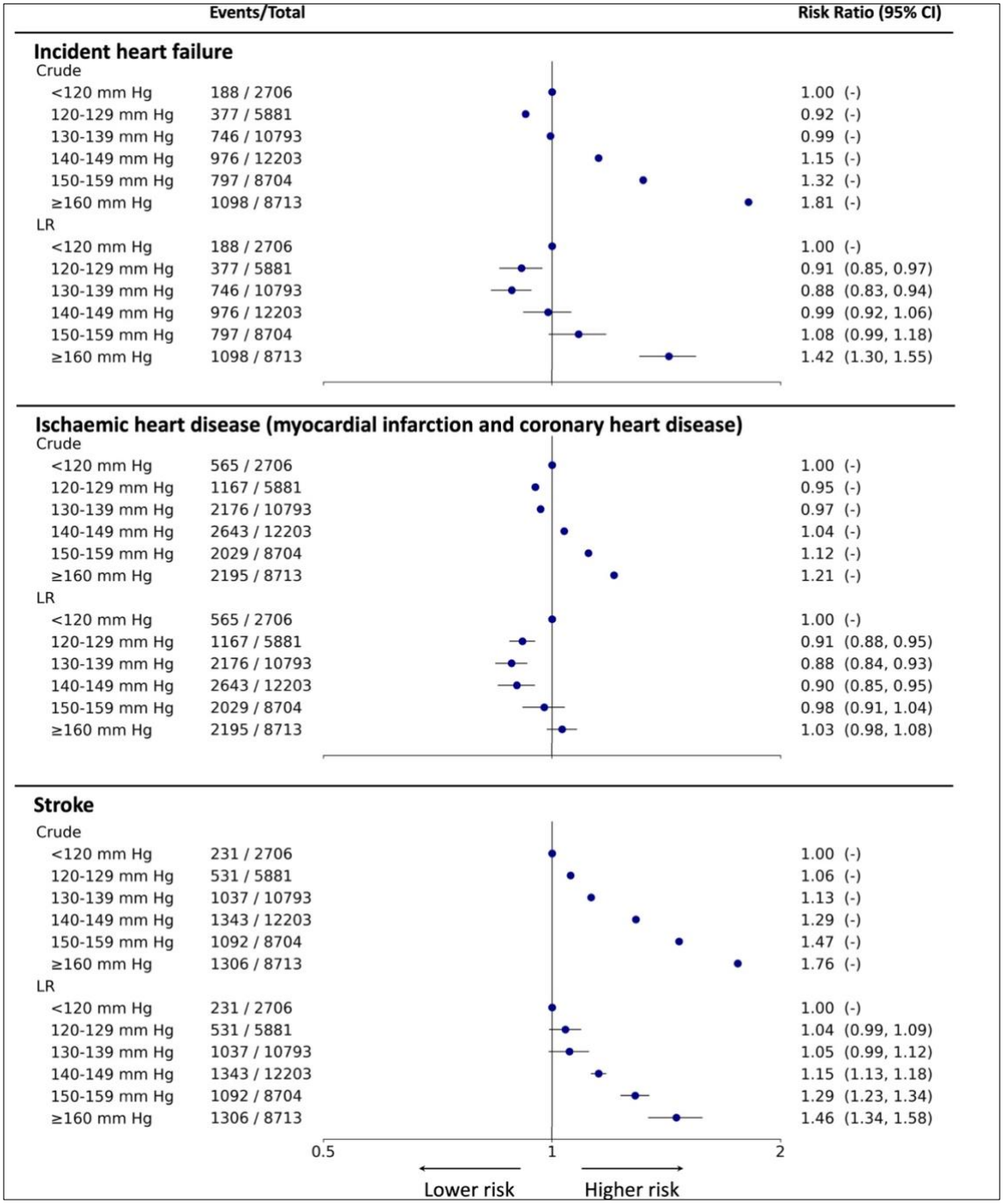
CPRD: Clinical Practice Research Datalink; BP: blood pressure;

Figure S2: Association with primary outcomes in patients with diabetes (statistical benchmark modelling)



Forest plot of risk ratio estimates with 95% confidence intervals (CI) for association of systolic blood pressure and primary outcome. From the left, the six exposure groups are shown in first column. Number of events and total number of patients in each exposure group is shown in second column. The forest plot and corresponding risk ratio estimates are shown in the right-most column relative to reference class, <120 mm Hg. The forest plot is plotted in logarithmic scale. For all crude estimates and estimates for reference class, there is no confidence interval. LR: logistic regression; CI: confidence interval.

Figure S3: Association with secondary outcomes in patients with diabetes (statistical benchmark modelling)



Forest plot of risk ratio estimates with 95% confidence intervals (CI) for association of systolic blood pressure and secondary outcomes. From the left, the six exposure groups are shown in first column. Number of events and total number of patients in each exposure group is shown in second column. The forest plot and corresponding risk ratio estimates are shown in the right-most column relative to reference class, <120 mm Hg. The forest plot is plotted in logarithmic scale. For all crude estimates and estimates for reference class, there is no confidence interval. LR: logistic regression; CI: confidence interval.

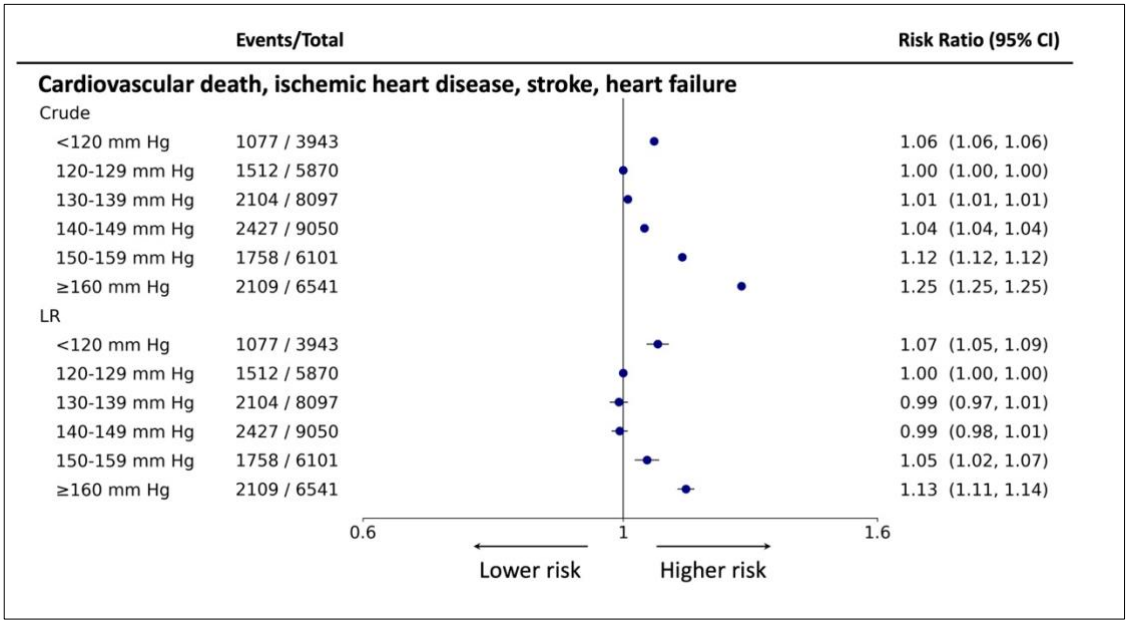
9.4.2 Systolic blood pressure, cardiovascular outcomes, and COPD

9.4.2.1 Details on the logistic regression modelling

Logistic regression modelling (LR) was used for the conventional approach in this work. The modelling utilised direct estimation method for estimation of the RR[61]. As an example, to estimate the effect of 150-159 mm Hg on cardiovascular outcomes with respect to the reference exposure, the trained LR model predicted risk with exposure for all patients set to the categorical variable representing 150-159 mm Hg and predicted risk with exposure similarly set to the reference group. The RR was derived as the ratio of the average of these two sets of predictions. For theoretical guarantees, we implemented k-fold cross-validation (k=10) for causal estimation. RR was calculated as the average of RR estimations on the 10 individual test sets, and the 95% CI was calculated via bootstrapping[58]. Lastly, the crude RR was calculated as the ratio between the average empirical risk of outcome in a particular exposure group divided by the same in the reference exposure group.

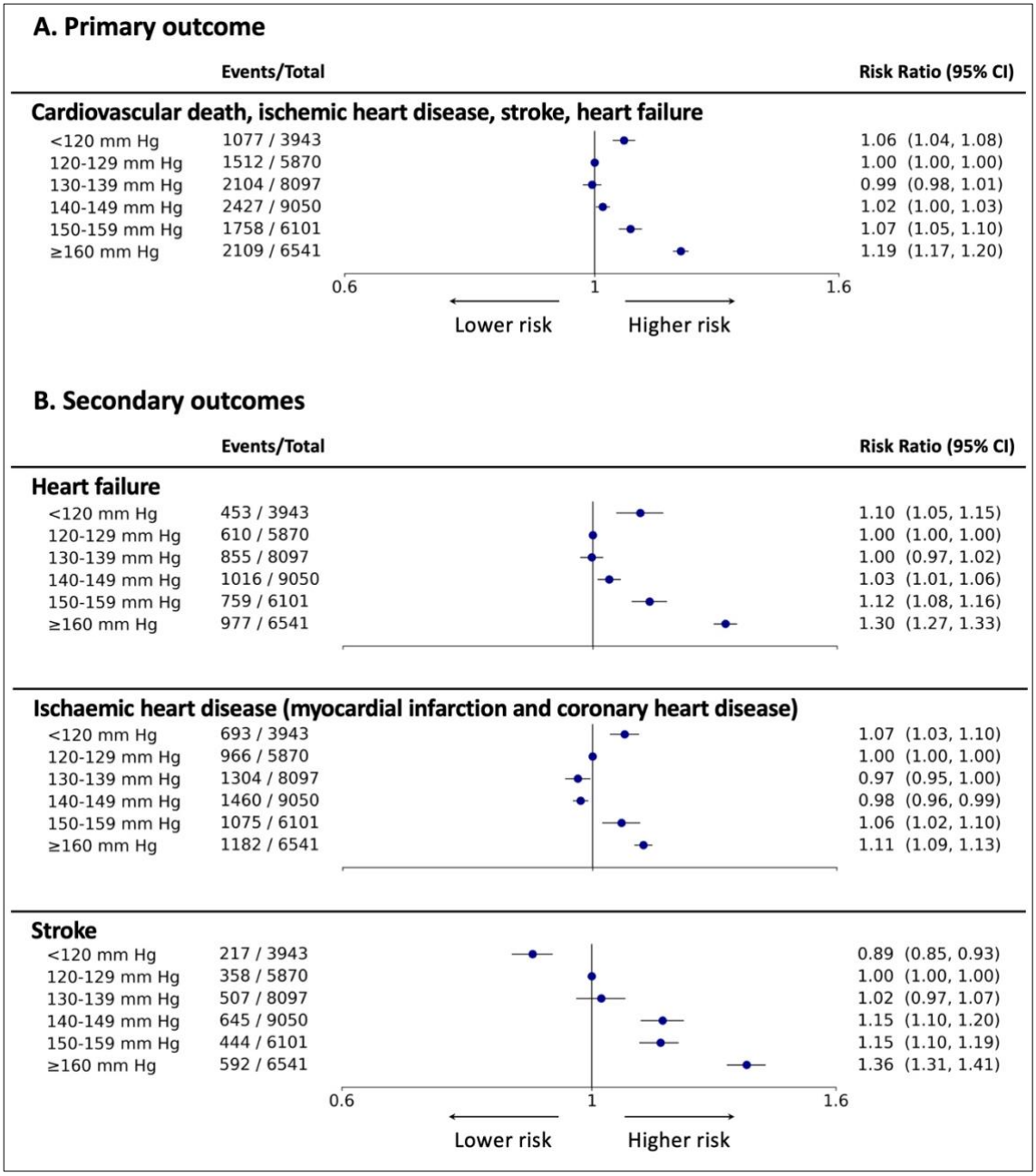
9.4.2.2 Statistical benchmark modelling

Figure S4: Association with primary outcomes in patients with COPD (statistical benchmark modelling)



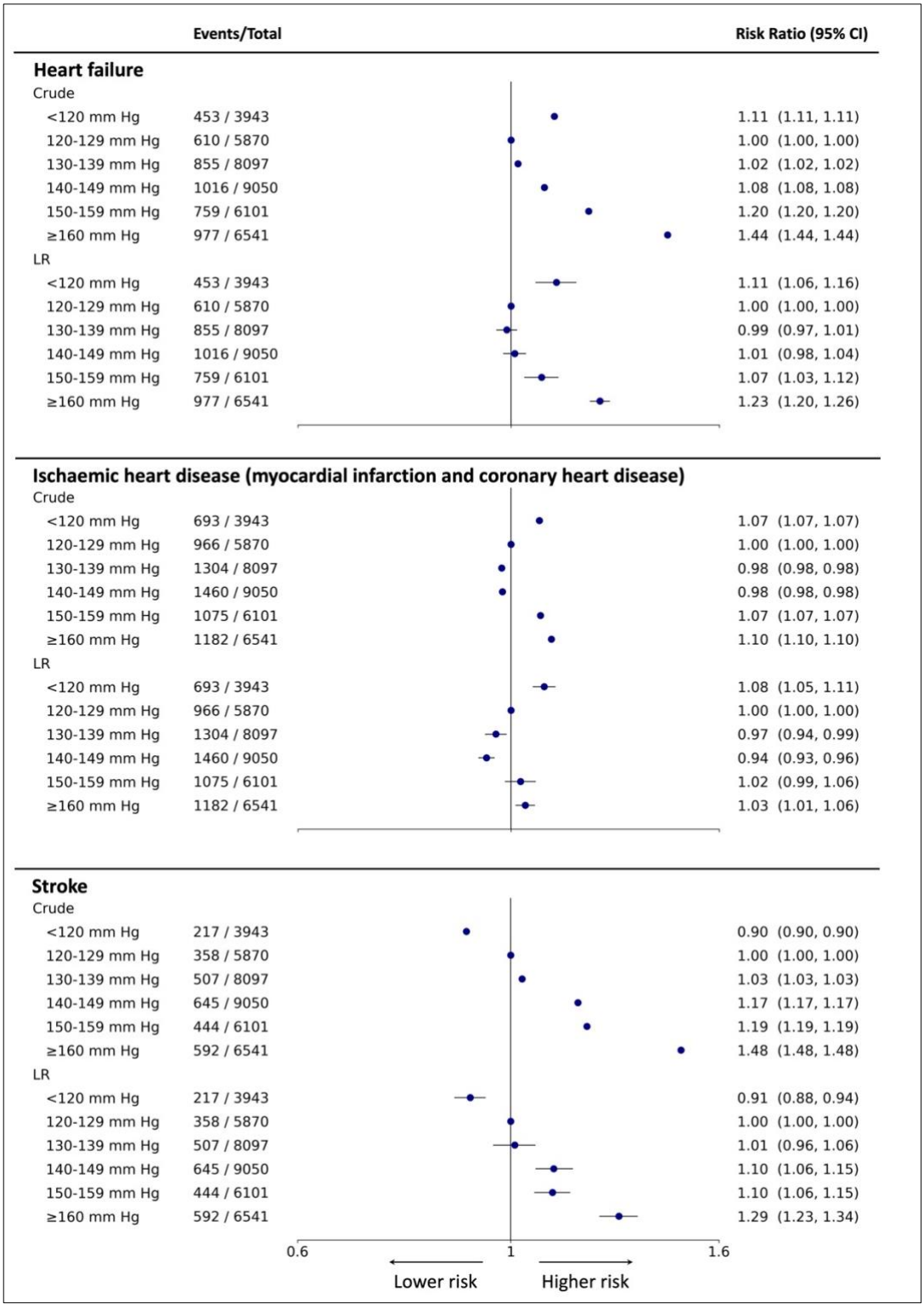
Forest plot of risk ratio estimates with 95% confidence intervals (CI) for association of systolic blood pressure and primary outcome. From the left, the six exposure groups are shown in first column. Number of events and total number of patients in each exposure group is shown in second column. The forest plot and corresponding risk ratio estimates are shown in the right-most column relative to reference class, <120 mm Hg. The forest plot is plotted in logarithmic scale. For all crude estimates and estimates for reference class, there is no confidence interval. LR: logistic regression; CI: confidence interval.

Figure S5: Association with all outcomes in patients with COPD (logistic regression modelling with expanded predictors)



Forest plot of risk ratio estimates with 95% confidence intervals (CI) for association of systolic blood pressure and all outcomes with logistic regression modelling with expanded predictor set. From the left, the six exposure groups are shown in first column. Number of events and total number of patients in each exposure group is shown in second column. The forest plot and corresponding risk ratio estimates are shown in the right-most column relative to reference class, <120 mm Hg. The forest plot is plotted in logarithmic scale. For all crude estimates and estimates for reference class, there is no confidence interval. LR: logistic regression; CI: confidence interval.

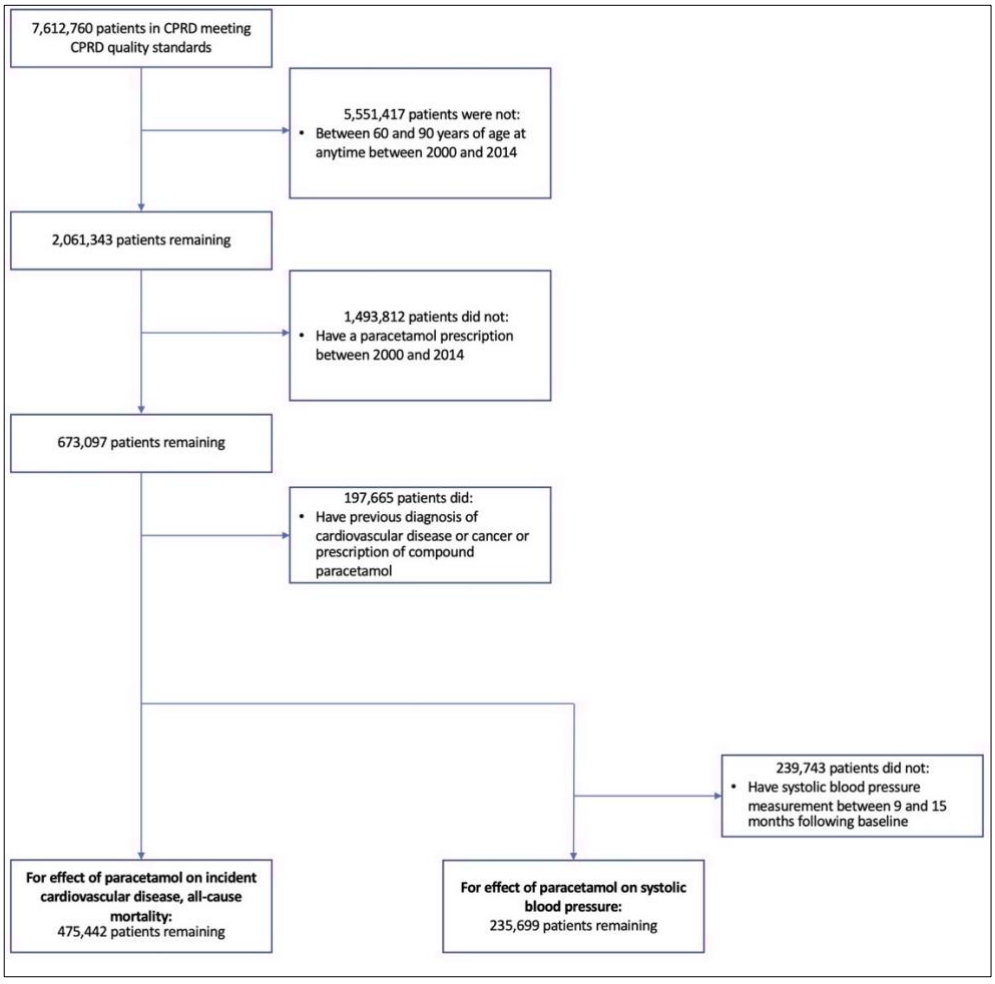
Figure S6: Association with secondary outcomes in patients with COPD (statistical benchmark modelling)



Forest plot of risk ratio estimates with 95% confidence intervals (CI) for association of systolic blood pressure and secondary outcomes. From the left, the six exposure groups are shown in first column. Number of events and total number of patients in each exposure group is shown in second column. The forest plot and corresponding risk ratio estimates are shown in the right-most column relative to reference class, <120 mm Hg. The forest plot is plotted in logarithmic scale. For all crude estimates and estimates for reference class, there is no confidence interval. LR: logistic regression; CI: confidence interval.

9.4.3 Paracetamol, systolic blood pressure, incident cardiovascular diseases, and all-cause mortality

Figure S7: Flow chart for patient selection for paracetamol study cohort



This is the flow chart to select eligible patients for the study investigating paracetamol and various outcomes. CPRD: Clinical Practice Research Datalink.

Table S9: Extended baseline characteristics among patients initiating non-sodium-based or sodium-based paracetamol

Exposure	*Non-sodium*	*Sodium*
No. (%)	460980 (97.0)	14462 (3.0)
Age, yrs (STD)	73.7 (8.6)	76.1 (9.1)
Women (%)	296190 (64.3)	10342 (71.5)
Ethnicity (White) (%)	126248 (27.4)	4060 (28.1)
No. of GP visits (STD)	3.3 (3.5)	3.3 (3.7)
No. of secondary care visits (STD)	1.3 (8.7)	1.7 (6.7)
IMD (STD) †	1.9 (1.4)	1.9 (1.3)
YOB (STD)	1933.4 (9.7)	1930.2 (10.3)
SBP (STD) †	141.2 (13.7)	139.4 (14.5)
BMI (STD) †	27.7 (4.3)	26.0 (4.0)
Smoking status†*:*		
Current or former smoker (%)	252522 (54)	5431 (37)
Never smoker (%)	208458 (45)	9031 (62)
Alcohol status†*:*		
Current or former drinker (%)	343602 (74)	9221 (63)
Never drinker (%)	117356 (25)	5241 (36)
Disease at baseline:		
CKD (%)	3757 (0.8)	86 (0.6)
Diabetes (%)	41894 (9.1)	988 (6.8)
Hypertension (%)	118022 (25.6)	2640 (18.3)
Arthritis (%)	140161 (30.4)	2960 (20.5)
Gout (%)	16239 (3.5)	297 (2.1)
Rheumatoid arthritis (%)	7477 (1.6)	238 (1.6)
Liver disease (%)	1300 (0.3)	41 (0.3)
PUD (%)	6224 (1.4)	176 (1.2)
Asthma (%)	31659 (6.9)	939 (6.5)
COPD (%)	25775 (5.6)	794 (5.5)
PAD (%)	12949 (2.8)	338 (2.3)
Epilepsy (%)	3815 (0.8)	213 (1.5)
Dementia (%)	9973 (2.2)	985 (6.8)
Depression (%)	45231 (9.8)	1324 (9.2)
Substance abuse (%)	3260 (0.7)	97 (0.7)
Hyperlipidaemia (%)	35861 (7.8)	737 (5.1)
Venous thromboembolism (%)	18915 (4.1)	471 (3.3)
Atrial fibrillation (%)	17563 (3.8)	443 (3.1)
Fracture (%)	50478 (11.0)	1596 (11.0)
Pneumonia (%)	6471 (1.4)	329 (2.3)

Fall (%)	1227 (0.3)	42 (0.3)
Gastrointestinal bleeding (%)	5457 (1.2)	226 (1.6)
Reflux disease (%)	24349 (5.3)	663 (4.6)
Gastritis (%)	15600 (3.4)	474 (3.3)
Medications at baseline:		
Anticholinergics (%)	243209 (52.8)	7676 (53.1)
Statins (%)	112124 (24.3)	2304 (15.9)
Bisphosphonates (%)	38965 (8.5)	1194 (8.3)
Calcium (%)	52365 (11.4)	1839 (12.7)
Benzodiazepines (%)	61521 (13.3)	2254 (15.6)
Dementia (%)	4241 (0.9)	370 (2.6)
Antihypertensives (%)	216574 (47.0)	5465 (37.8)
Anticoagulants (%)	21102 (4.6)	570 (3.9)
Antiplatelet (%)	120975 (26.2)	3612 (25.0)
Anxiolytics and hypnotics (%)	112258 (24.4)	4033 (27.9)
Opioids (%)	139717 (30.3)	3031 (21.0)
Antipsychotic (%)	85393 (18.5)	3324 (23.0)
Steroids (%)	83460 (18.1)	2222 (15.4)
Nitrates (%)	29985 (6.5)	868 (6.0)
Loop diuretics (%)	60549 (13.1)	2094 (14.5)
Thiazide diuretics (%)	129818 (28.2)	3133 (21.7)
Potassium sparing diuretics (%)	27226 (5.9)	1013 (7.0)
Anti-diabetic (%)	6212 (1.3)	122 (0.8)
Calcium channel blockers (%)	108754 (23.6)	2528 (17.5)
ACE inhibitors (%)	106362 (23.1)	2361 (16.3)
Angiotensin receptor blockers (%)	38717 (8.4)	794 (5.5)
Beta blockers (%)	103282 (22.4)	2483 (17.2)
Oestrogen (%)	65340 (14.2)	1656 (11.5)
Insulin (%)	8357 (1.8)	209 (1.4)
H2 blockers (%)	66324 (14.4)	2018 (14.0)
Proton pump inhibitors (%)	159288 (34.6)	4379 (30.3)
DMARDs (%)	7386 (1.6)	198 (1.4)
Glucocorticoid (%)	92014 (20.0)	2391 (16.5)

%: percent; STD: standard deviation; No: number; Yrs: years; GP: general practice; YOB: year of birth; BMI: body mass index; SBP: systolic blood pressure; CKD: chronic kidney disease; PUD: peptic ulcer disease; COPD: chronic obstructive pulmonary disease; PAD: peripheral artery disease; ACE: Angiotensin-converting enzyme; DMARDs: Disease-modifying antirheumatic drugs; †The percentage of missing variables –alcohol status (54.6%), smoking status (36.9%), IMD (38.7), SBP (26.4%), BMI (40.9%).

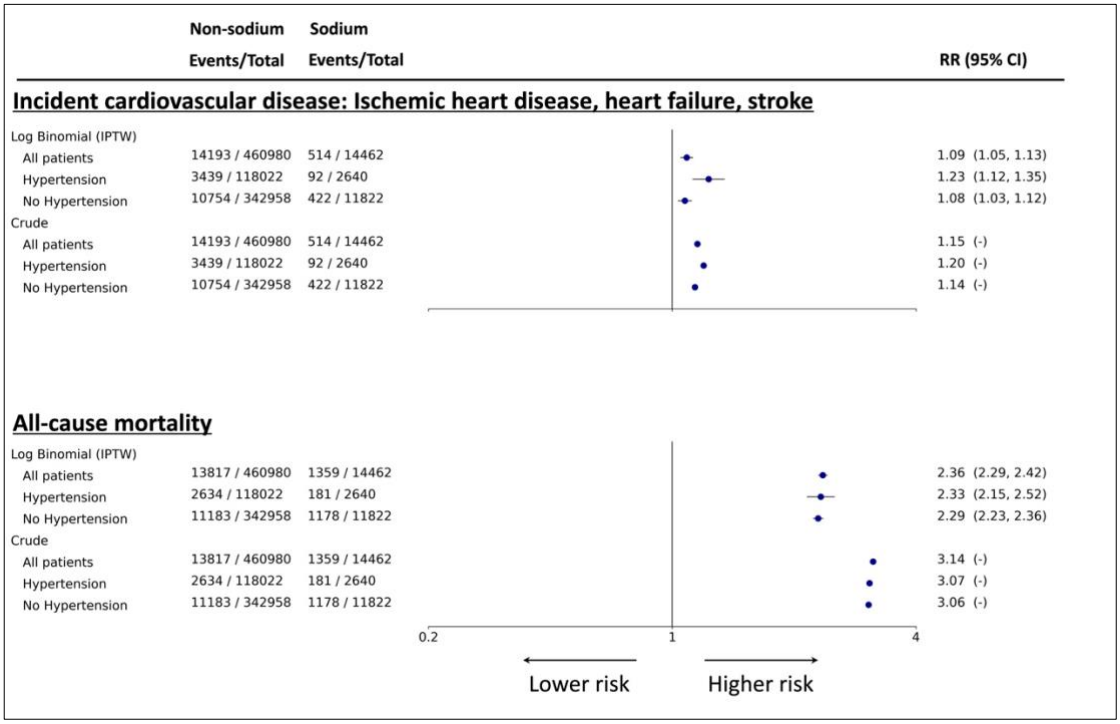
Table S10: Baseline characteristics among patients with systolic blood pressure measurements initiating non-sodium-based or sodium-based paracetamol

Exposure	*Non-sodium*	*Sodium*
No. (%)	230351 (97.7)	5348 (2.3)
Follow-up, yrs (STD)	11.7 (1.5)	11.7 (1.5)
Age, yrs (STD)	73.7 (8.3)	75.4 (8.9)
Women (%)	149345 (64.8)	3929 (73.5)
Ethnicity (White) (%)	64540 (28.0)	1618 (30.3)
No. of GP visits (STD)	3.6 (3.6)	3.6 (3.7)
No. of secondary care visits (STD)	1.2 (8.3)	1.5 (5.6)
*IMD (STD)**	1.9 (1.4)	1.9 (1.4)
YOB (STD)	1933.5 (9.3)	1931.2 (10.1)
SBP (STD)	142.4 (14.3)	141.8 (15.6)
*BMI (STD)**	28.1 (4.6)	26.6 (4.5)
LDL (STD)	3.1 (1.4)	3.1 (0.7)
TG (STD)	1.6 (1.5)	1.6 (0.7)
TC (STD)	5.2 (1.7)	5.2 (0.9)
Smoking status:*		
Current or former smoker (%)	122388 (53)	2018 (37)
Never smoker (%)	107963 (46)	3330 (62)
Alcohol status:*		
Current or former drinker (%)	171599 (74)	3496 (65)
Never drinker (%)	58742 (25)	1852 (34)
Disease at baseline:		
CKD (%)	2433 (1.1)	41 (0.8)
Diabetes (%)	29231 (12.7)	555 (10.4)
Hypertension (%)	81154 (35.2)	1503 (28.1)
Arthritis (%)	74113 (32.2)	1301 (24.3)
Gout (%)	9592 (4.2)	133 (2.5)
Rheumatoid arthritis (%)	3799 (1.6)	99 (1.9)
Liver disease (%)	605 (0.3)	13 (0.2)
PUD (%)	3041 (1.3)	66 (1.2)
Asthma (%)	16927 (7.3)	439 (8.2)
COPD (%)	12721 (5.5)	304 (5.7)
PAD (%)	7238 (3.1)	149 (2.8)
Epilepsy (%)	1689 (0.7)	73 (1.4)
Dementia (%)	3459 (1.5)	207 (3.9)
Depression (%)	22608 (9.8)	518 (9.7)
Substance abuse (%)	1360 (0.6)	25 (0.5)
Hyperlipidaemia (%)	22207 (9.6)	411 (7.7)

Venous thromboembolism (%)	9884 (4.3)	182 (3.4)
Atrial fibrillation (%)	10162 (4.4)	198 (3.7)
Fracture (%)	24039 (10.4)	589 (11.0)
Pneumonia (%)	3095 (1.3)	107 (2.0)
Fall (%)	611 (0.3)	11 (0.2)
Gastrointestinal bleeding (%)	2756 (1.2)	81 (1.5)
Reflux disease (%)	13139 (5.7)	297 (5.6)
Gastritis (%)	8211 (3.6)	206 (3.9)
Medications at baseline:		
Anticholinergics (%)	123050 (53.4)	2852 (53.3)
Statins (%)	71388 (31.0)	1220 (22.8)
Bisphosphonates (%)	19423 (8.4)	460 (8.6)
Calcium (%)	25976 (11.3)	706 (13.2)
Benzodiazepines (%)	30821 (13.4)	810 (15.1)
Dementia (%)	1513 (0.7)	81 (1.5)
Antihypertensives (%)	139362 (60.5)	2807 (52.5)
Anticoagulants (%)	11607 (5.0)	236 (4.4)
Antiplatelet (%)	71042 (30.8)	1476 (27.6)
Anxiolytics and hypnotics (%)	55343 (24.0)	1412 (26.4)
Opioids (%)	70787 (30.7)	1196 (22.4)
Antipsychotic (%)	42913 (18.6)	1142 (21.4)
Steroids (%)	43083 (18.7)	938 (17.5)
Nitrates (%)	17495 (7.6)	355 (6.6)
Loop diuretics (%)	32194 (14.0)	753 (14.1)
Thiazide diuretics (%)	86469 (37.5)	1749 (32.7)
Potassium sparing diuretics (%)	14320 (6.2)	366 (6.8)
Anti-diabetic (%)	4436 (1.9)	71 (1.3)
Calcium channel blockers (%)	72588 (31.5)	1369 (25.6)
ACE inhibitors (%)	72549 (31.5)	1301 (24.3)
Angiotensin receptor blockers (%)	27605 (12.0)	498 (9.3)
Beta blockers (%)	65951 (28.6)	1262 (23.6)
Oestrogen (%)	34010 (14.8)	731 (13.7)
Insulin (%)	5742 (2.5)	119 (2.2)
H2 blockers (%)	34181 (14.8)	842 (15.7)
Proton pump inhibitors (%)	82365 (35.8)	1763 (33.0)
DMARDs (%)	3987 (1.7)	93 (1.7)
Glucocorticoid (%)	47660 (20.7)	1015 (19.0)

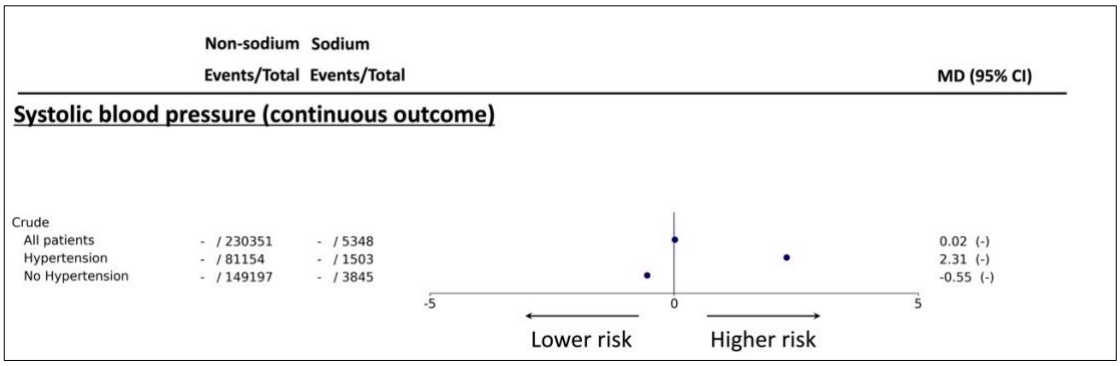
*%: percent; STD: standard deviation; No: number; Yrs: years; GP: general practice; YOB: year of birth; BMI: body mass index; SBP: systolic blood pressure; LDL: low-density lipoprotein; TC: total cholesterol; TG: triglycerides; CKD: chronic kidney disease; PUD: peptic ulcer disease; COPD: chronic obstructive pulmonary disease; PAD: peripheral artery disease; ACE: Angiotensin-converting enzyme; DMARDs: Disease-modifying antirheumatic drugs; *: imputed variables.*

Figure S8: Association of sodium-based vs non-sodium-based paracetamol and incident cardiovascular disease and all-cause mortality (conventional modelling)



Forest plots of log binomial with IPTW and crude modelling for analyses of binary outcomes (all patients, stratified by hypertension status) is shown. Number of events and total number of patients in each exposure group is shown in second and third columns. The forest plot and corresponding risk ratio estimates are shown in the right-most column relative to the reference exposure, non-sodium paracetamol. The effect size is plotted on a logarithmic scale.

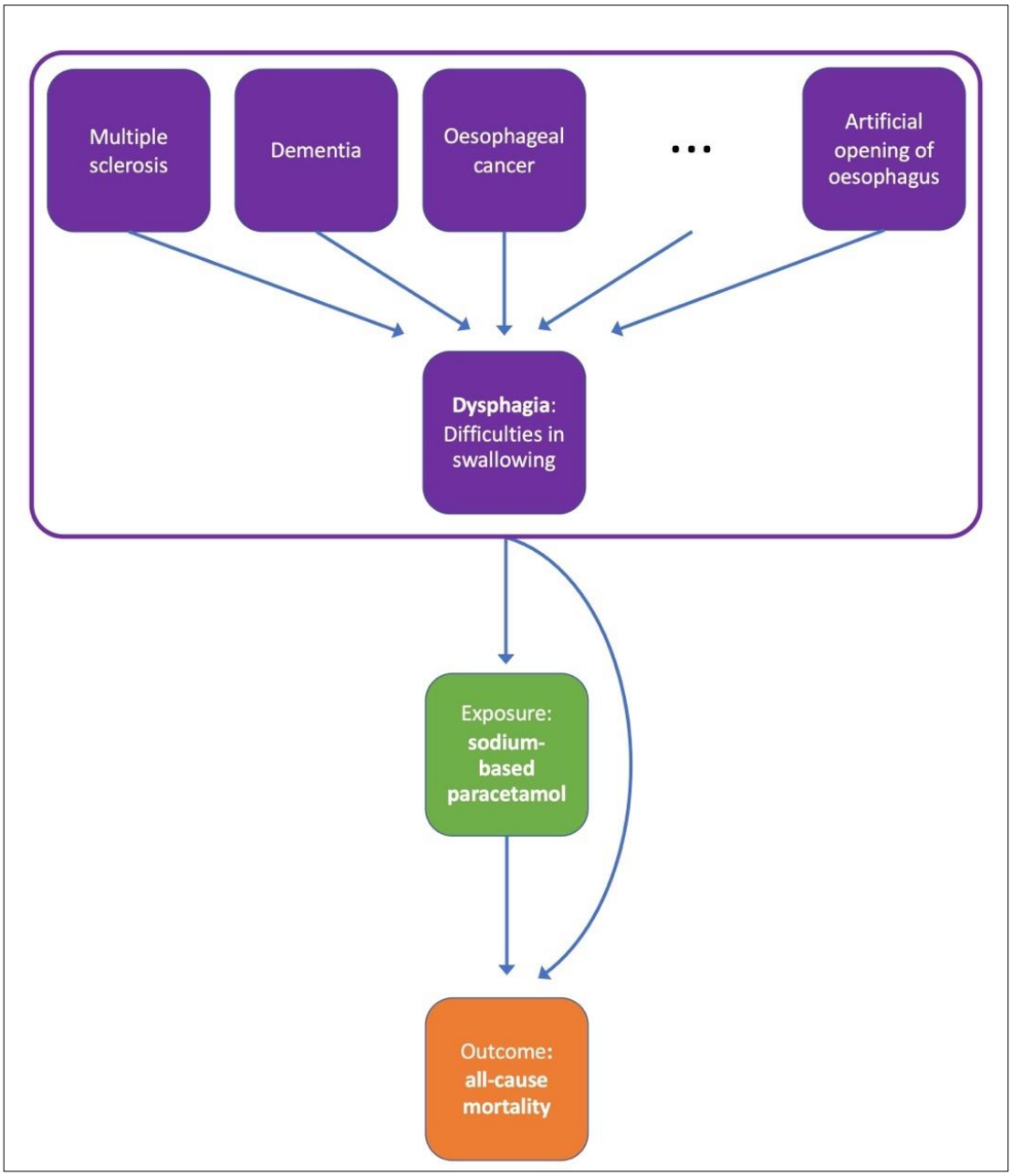
Figure S9: Association of sodium-based vs non-sodium-based paracetamol and systolic blood pressure (conventional modelling)



Forest plots of log binomial with IPTW and crude modelling for analyses of continuous systolic blood pressure as outcome (all patients, stratified by hypertension status) is

shown. Number of events and total number of patients in each exposure group is shown in second and third columns. The forest plot and corresponding risk ratio estimates are shown in the right-most column relative to the reference exposure, non-sodium paracetamol. The effect size is plotted on a logarithmic scale.

Figure S10: Diagram of confounding due to dysphagia and related comorbidities



The purple panel contains dysphagia and associated disorders connected to both exposure and outcome. The green box represents the exposure of sodium-based paracetamol, connected to outcome. The orange box is outcome of all-cause mortality.

10 REFERENCES

1. Pearl J. Causal inference in statistics: An overview. *Stat Surv*; 3. Epub ahead of print 2009. DOI: 10.1214/09-SS057.

2. Kingston A, Robinson L, Booth H, Knapp M, Jagger C, Adelaja B, Avendano M, Bamford SM, Banerjee S, Berwald S, Bowling A, Burgon C, Bustard E, Comas-Herrera A, Dangoor M, Dixon J, Farina N, Greengross S, Grundy E, et al. Projections of multi-morbidity in the older population in England to 2035: Estimates from the Population Ageing and Care Simulation (PACSim) model. *Age Ageing*. Epub ahead of print 2018. DOI: 10.1093/ageing/afx201.

3. Schneeweiss S, Maclure M. Use of comorbidity scores for control of confounding in studies using administrative databases. *Int J Epidemiol*; 29. Epub ahead of print 2000. DOI: 10.1093/ije/29.5.891.

4. Casey JA, Schwartz BS, Stewart WF, Adler NE. Using Electronic Health Records for Population Health Research: A Review of Methods and Applications. *Annual Review of Public Health*; 37. Epub ahead of print 2016. DOI: 10.1146/annurev-publhealth-032315-021353.

5. Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, Staa T van, Smeeth L. Data Resource Profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol* 2015; 44: 827–836.

6. Collaborators GB of D 2016 C of D. Global, regional, and national age-sex specific mortality for 264 causes of death, 1980–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet* 2017; 390: 1151–1210.

7. Cardiovascular disease prevention: applying All Our Health. *GOV*, https://www.gov.uk/government/publications/cardiovascular-disease-prevention-applying-all-our-health/cardiovascular-disease-prevention-applying-all-our-health (2022).

8. Bromfield S, Muntner P. High blood pressure: The leading global burden of disease risk factor and the need for worldwide prevention programs. *Curr Hypertens Rep*; 15. Epub ahead of print 2013. DOI: 10.1007/s11906-013-0340-9.

9. Tran J, Norton R, Conrad N, Rahimian F, Canoy D, Nazarzadeh M, Rahimi K. Patterns and temporal trends of comorbidity among adult patients with incident cardiovascular disease in the UK between 2000 and 2014: A population-based cohort study. *PLoS Med* 2018; 15: e1002513.

10. Romero-Corral A, Montori VM, Somers VK, Korinek J, Thomas RJ, Allison TG, Mookadam F, Lopez-Jimenez F. Association of bodyweight with total mortality and with cardiovascular events in coronary artery disease: a systematic review of cohort studies. *Lancet*; 368. Epub ahead of print 2006. DOI: 10.1016/S0140-6736(06)69251-9.

11. Yusuf S, Hawken S, Ôunpuu S, Bautista L, Franzosi MG, Commerford P, Lang CC, Rumboldt Z, Onen CL, Lisheng L, Tanomsup S, Wangai P, Razak F, Sharma AM, Anand SS. Obesity and the risk of myocardial infarction in 27 000 participants from 52 countries: A case-control study. *Lancet*; 366. Epub ahead of print 2005. DOI: 10.1016/S0140-6736(05)67663-5.

12. The National Institute for Health and Care Excellence. *Cardiovascular disease: risk assessment and reduction, including lipid modification*, www.nice.org.uk/guidance/cg181 (27 September 2016).

13. Vidal-Petiot E, Ford I, Greenlaw N, al. et. Cardiovascular event rates and mortality according to achieved systolic and diastolic blood pressure in patients with stable coronary artery disease: an international cohort study. *Lancet* 2016; 388: 2142–2152.

14. Nazarzadeh M, Bidel Z, Canoy D, Copland E, Bennett DA, Dehghan A, Davey Smith G, Holman RR, Woodward M, Gupta A, Adler AI, Wamil M, Sattar N, Cushman WC, McManus RJ, Teo K, Davis BR, Chalmers J, Pepine CJ, et al. Blood pressure-lowering treatment for prevention of major cardiovascular diseases in people with and without type 2 diabetes: an individual participant-level data meta-analysis. *Lancet Diabetes Endocrinol* 2022; 10: 645–654.

15. The Blood Pressure Lowering Treatment Trialists' Collaboration. Pharmacological blood pressure lowering for primary and secondary prevention of cardiovascular disease across different levels of blood pressure: an individual participant-level data meta-analysis. *The Lancet* 2021; 397: 1625–1636.

16. Adamsson Eryd S, Gudbjörnsdottir S, Manhem K, Rosengren A, Svensson AM, Miftaraj M, Franzén S, Björck S. Blood pressure and complications in individuals with type 2 diabetes and no previous cardiovascular disease: national population based cohort study. *BMJ* 2016; 354: 4070.

17. Heart failure - NHS, https://www.nhs.uk/conditions/heart-failure/ (accessed 12 July 2022).

18. Tanai E, Frantz S. Pathophysiology of heart failure. *Compr Physiol* 2016; 6: 187–214.

19. Conrad N, Judge A, Tran J, Mohseni H, Hedgecott D, Crespillo AP, Allison M, Hemingway H, Cleland JG, McMurray JJV, Rahimi K. Temporal trends and patterns in heart failure incidence: a population-based study of 4 million individuals. *The Lancet*. Epub ahead of print 2018. DOI: 10.1016/S0140-6736(17)32520-5.

20. Groenewegen A, Rutten FH, Mosterd A, Hoes AW. Epidemiology of heart failure. *Eur J Heart Fail* 2020; 22: 1342–1356.

21. NATIONAL HEART FAILURE AUDIT (NHFA) NATIONAL CARDIAC AUDIT PROGRAMME ii The National Institute for Cardiovascular Outcomes Research (NICOR), www.hqip.org.uk/ (2021, accessed 12 July 2022).

22. McDonagh TA, Metra M, Adamo M, Gardner RS, Baumbach A, Böhm M, Burri H, Butler J, Celutkiene J, Chioncel O, Cleland JGF, Coats AJS, Crespo-Leiro MG, Farmakis D, Gilard M, Heymans S. 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. *Eur Heart J* 2021; 42: 3599–3726.

23. Epstein FH, Levin ER, Gardner DG, Samson WK. Natriuretic peptides. *N Engl J Med* 1998; 339: 321–328.

24. Sholter DE, Armstrong PW. Adverse effects of corticosteroids on the cardiovascular system. *Canadian Journal of Cardiology* 2000; 16: 505–511.

25. Coronary heart disease - NHS, https://www.nhs.uk/conditions/coronary-heart-disease/ (accessed 13 July 2022).

26. Gofman JW, Young W, Tandy R. Ischemic Heart Disease, Atherosclerosis, and Longevity. *Circulation* 1966; 34: 679–697.

27. World HO (WHO). World Health Report 2002: Reducing Risks, Promoting Healthy Life World Health Organization. *Agricultural Economics*.

28. Warlow CP. Epidemiology of stroke. *Lancet*; 352. Epub ahead of print 1998. DOI: 10.1016/s0140-6736(98)90086-1.

29. Cardiovascular disease - NHS, https://www.nhs.uk/conditions/cardiovascular-disease/ (accessed 20 July 2022).

30. Roth GA, Johnson C, Abajobir A, Abd-Allah F, Abera SF, Abyu G, Ahmed M, Aksut B, Alam T, Alam K, Alla F, Alvis-Guzman N, Amrock S, Ansari H, Ärnlöv J, Asayesh H, Atey TM, Avila-Burgos L, Awasthi A, et al. Global, Regional, and National Burden of Cardiovascular Diseases for 10 Causes, 1990 to 2015. *J Am Coll Cardiol*; 70. Epub ahead of print 2017. DOI: 10.1016/j.jacc.2017.04.052.

31. Johansson BB. Hypertension mechanisms causing stroke. In: *Clinical and Experimental Pharmacology and Physiology*. 1999. Epub ahead of print 1999. DOI: 10.1046/j.1440-1681.1999.03081.x.

32. Organization WH. A global brief on hypertension: Silent killer, global public health crisis. 2013. *WHO_DCO_WHD_2013*.

33. Pickering G. Hypertension. Definitions, natural histories and consequences. *Am J Med* 1972; 52: 570–583.

34. Kang Y-Y, Wang J-G. The J-Curve Phenomenon in Hypertension. *Pulse*; 4. Epub ahead of print 2016. DOI: 10.1159/000446922.

35. Byrd JB, Newby DE, Anderson JA, Calverley PM, Celli BR, Cowans NJ, Crim C, Martinez FJ, Vestbo J, Yates J, Brook RD. Blood pressure, heart rate, andmortality in chronic obstructive pulmonary disease: The SUMMIT trial. *Eur Heart J*; 39. Epub ahead of print 2018. DOI: 10.1093/eurheartj/ehy451.

36. Blood Pressure Lowering Treatment Trialists' Collaboration, Sundström J, Arima H, Woodward M, Jackson R, Karmali K, Lloyd-Jones D, Baigent C, Emberson J, Rahimi K, MacMahon S, Patel A, Perkovic V, Turnbull F, Neal B. Blood pressure-lowering treatment based on cardiovascular risk: a meta-analysis of individual patient data. *Lancet* 2014; 384: 591–8.

37. Whelton SP, McEvoy JW, Shaw L, Psaty BM, Lima JAC, Budoff M, Nasir K, Szklo M, Blumenthal RS, Blaha MJ. Association of Normal Systolic Blood Pressure Level with Cardiovascular Disease in the Absence of Risk Factors. *JAMA Cardiol*; 5. Epub ahead of print 2020. DOI: 10.1001/jamacardio.2020.1731.

38. 2020 International Society of Hypertension Global Hypertension Practice Guidelines | Hypertension, https://www.ahajournals.org/doi/10.1161/HYPERTENSIONAHA.120.15026 (accessed 28 February 2022).

39. Diabetes, https://www.who.int/news-room/fact-sheets/detail/diabetes (accessed 12 July 2022).

40. Number of people with diabetes reaches 4.7 million | Diabetes UK, https://www.diabetes.org.uk/about_us/news/new-stats-people-living-with-diabetes (accessed 12 July 2022).

41. World Health Organization. Global Report on Diabetes. *Isbn*; 978. Epub ahead of print 2016. DOI: ISBN 978 92 4 156525 7.

42. Standards of medical care in diabetes--2015: summary of revisions. *Diabetes Care* 2015; 38: S4.

43. Emdin C, Anderson S, Woodward M, Rahimi K. Usual blood pressure and risk of new-onset diabetes: evidence from 4·1 million adults and a meta-analysis of prospective studies. *J Am Coll Cardiol* 2015; 66: 1552–1562.

44. Nazarzadeh M, Bidel Z, Canoy D, Copland E, Wamil M, Majert J, Smith Byrne K, Sundström J, Teo K, Davis BR, Chalmers J, Pepine CJ, Dehghan A, Bennett DA, Smith GD, Rahimi K. Blood pressure lowering and risk of new-onset type 2 diabetes: an individual participant data meta-analysis. *The Lancet* 2021; 398: 1803–1810.

45. Chronic obstructive pulmonary disease (COPD) statistics | British Lung Foundation, https://statistics.blf.org.uk/copd (accessed 12 July 2022).

46. Morgan AD, Zakeri R, Quint JK. Defining the relationship between COPD and CVD: what are the implications for clinical practice? *Ther Adv Respir Dis*; 12. Epub ahead of print 19 January 2018. DOI: 10.1177/1753465817750524.

47. Vogelmeier CF, Criner GJ, Martinez FJ, Anzueto A, Barnes PJ, Bourbeau J, Celli BR, Chen R, Decramer M, Fabbri LM, Frith P, Halpin DMG, Varela MVL, Nishimura M, Roche N, Rodriguez-Roisin R, Sin DD, Singh D, Stockley R, et al. Global strategy for the diagnosis, management, and prevention of chronic obstructive lung disease 2017 report. *American Journal of Respiratory and Critical Care Medicine*; 195. Epub ahead of print 2017. DOI: 10.1164/rccm.201701-0218PP.

48. Unger T, Borghi C, Charchar F, Khan NA, Poulter NR, Prabhakaran D, Ramirez A, Schlaich M, Stergiou GS, Tomaszewski M, Wainford RD, Williams B, Schutte AE. 2020 International Society of Hypertension Global Hypertension Practice Guidelines. *Hypertension* 2020; 75: 1334–1357.

49. National Trends in Hospital and Physician Adoption of Electronic Health Records | HealthIT.gov, https://www.healthit.gov/data/quickstats/national-trends-hospital-and-physician-adoption-electronic-health-records (accessed 20 July 2022).

50. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: Towards better research applications and clinical care. *Nature Reviews Genetics*; 13. Epub ahead of print 2012. DOI: 10.1038/nrg3208.

51. Thygesen JH, Tomlinson C, Hollings S, Mizani MA, Handy A, Akbari A, Banerjee A, Cooper J, Lai AG, Li K, Mateen BA, Sattar N, Sofat R, Torralbo A, Wu H, Wood A, Sterne JAC, Pagel C, Whiteley WN, et al. COVID-19 trajectories among 57 million adults in England: a cohort study using electronic health records. *Lancet Digit Health* 2022; 4: e542–e557.

52. Kivimäki M, Batty GD, Singh-Manoux A, Britton A, Brunner EJ, Shipley MJ. Validity of Cardiovascular Disease Event Ascertainment Using Linkage to UK Hospital Records. *Epidemiology* 2017; 28: 735.

53. S Buuren KG-O. mice: multivariate imputation by chained equations in R. *J Statistical Software* 2011; 45: 67.

54. Clinical Practice Research Datalink - CPRD.

55. Lloyd-Jones DM. Cardiovascular Risk Prediction. *Circulation* 2010; 121: 1768–1777.

56. Pencina MJ, d'Agostino RB, d'Agostino RB, Vasan RS. Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Stat Med* 2008; 27: 157–172.

57. Risk Prediction | Columbia Public Health, https://www.publichealth.columbia.edu/research/population-health-methods/risk-prediction (accessed 21 July 2022).

58. Burnham KP, Efron B. The Jackknife, the Bootstrap and Other Resampling Plans. *Biometrics*; 39. Epub ahead of print 1983. DOI: 10.2307/2531123.

59. Imison C. Multiple long-term conditions (multimorbidity): making sense of the evidence. *National Institute for Health Research*.

60. Sahle BW, Owen AJ, Chin KL, Reid CM. Risk Prediction Models for Incident Heart Failure: A Systematic Review of Methodology and Model Performance. *J Card Fail* 2017; 23: 680–687.

61. Hernan M, Robins J. Causal inference: what if.

62. Deaton A, Cartwright N. Understanding and misunderstanding randomized controlled trials. *Soc Sci Med* 2018; 210: 2.

63. Speich B, von Niederhäusern B, Schur N, Hemkens LG, Fürst T, Bhatnagar N, Alturki R, Agarwal A, Kasenda B, Pauli-Magnus C, Schwenkglenks M, Briel M. Systematic review on costs and resource use of randomized clinical trials shows a lack of transparent and comprehensive data. *J Clin Epidemiol* 2018; 96: 1–11.

64. Haidich AB. Meta-analysis in medical research. *Hippokratia* 2010; 14: 29.

65. Beta-blockers | Prescribing information | Hypertension | CKS | NICE, https://cks.nice.org.uk/topics/hypertension/prescribing-information/beta-blockers/ (accessed 22 July 2022).

66. Rubin DB. Assignment to Treatment Group on the Basis of a Covariate. *Journal of Educational Statistics*; 2. Epub ahead of print 1977. DOI: 10.3102/10769986002001001.

67. Rubin DB. Causal Inference Using Potential Outcomes. *J Am Stat Assoc*. Epub ahead of print 2005. DOI: 10.1198/016214504000001880.

68. Neyman J. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*; 5.

69. Johansson FD, Shalit U, Kallus N, Sontag D. Generalization Bounds and Representation Learning for Estimation of Potential Outcomes and Causal Effects. 2020; 1–42.

70. Cochran WG. The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies. *Biometrics*; 24. Epub ahead of print 1968. DOI: 10.2307/2528036.

71. Rothman KJ, Greenland S, Associate TLL. Modern Epidemiology, 3rd Edition. *Hastings Cent Rep*; 44 Suppl 2.

72. McNamee R. Regression modelling and other methods to control confounding. *Occupational and Environmental Medicine*; 62. Epub ahead of print 2005. DOI: 10.1136/oem.2002.001115.

73. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res*; 46. Epub ahead of print 2011. DOI: 10.1080/00273171.2011.568786.

74. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. Epub ahead of print 1983. DOI: 10.1093/biomet/70.1.41.

75. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *J Am Stat Assoc*; 79. Epub ahead of print 1984. DOI: 10.1080/01621459.1984.10478078.

76. Rosenbaum PR. Model-based direct adjustment. *J Am Stat Assoc*; 82. Epub ahead of print 1987. DOI: 10.1080/01621459.1987.10478441.

77. Hammer GP, Prel J-B du, Blettner M. Avoiding Bias in Observational Studies. *Dtsch Arztebl Int*. Epub ahead of print 2009. DOI: 10.3238/arztebl.2009.0664.

78. Nawata K, Nagase N. Estimation of sample selection bias models. *Econom Rev*; 15. Epub ahead of print 1996. DOI: 10.1080/07474939608800363.

79. Gauss CF. THEORIA MOTUS CORPORUM COELESTIUM IN SECTIONIBUS CONICIS SOLEM AMBIENTIUM. In: *Werke*. 2012. Epub ahead of print 2012. DOI: 10.1017/cbo9781139058285.001.

80. Bishop CM. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. 2006.

81. Rumelhart DE, Hinton GE, Williams RJ. Learning Internal Representations by Error Propagation. In: *Readings in Cognitive Science: A Perspective from Psychology and Artificial Intelligence*. 2013. Epub ahead of print 2013. DOI: 10.1016/B978-1-4832-1446-7.50035-2.

82. Hochreiter S, Schmidhuber J. Long Short-Term Memory. *Neural Comput*; 9. Epub ahead of print 1997. DOI: 10.1162/neco.1997.9.8.1735.

83. Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate. 2014; 1–15.

84. Cho K, van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*. 2014. Epub ahead of print 2014. DOI: 10.3115/v1/d14-1179.

85. Cho K, van Merriënboer B, Bahdanau D, Bengio Y. On the properties of neural machine translation: Encoder–decoder approaches. In: *Proceedings of SSST 2014 - 8th Workshop on Syntax, Semantics and Structure in Statistical Translation*. 2014. Epub ahead of print 2014. DOI: 10.3115/v1/w14-4012.

86. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention Is All You Need. *Adv Neural Inf Process Syst*, https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (2017).

87. LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Comput*; 1. Epub ahead of print 1989. DOI: 10.1162/neco.1989.1.4.541.

88. Biran O, Cotton C. Explanation and Justification in Machine Learning: A Survey. *IJCAI Workshop on Explainable AI (XAI)*.

89. Roscher R, Bohn B, Duarte MF, Garcke J. Explainable Machine Learning for Scientific Insights and Discoveries. *IEEE Access*; 8. Epub ahead of print 2020. DOI: 10.1109/ACCESS.2020.2976199.

90. Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. In: *34th International Conference on Machine Learning, ICML 2017*. 2017.

91. Nguyen A, Yosinski J, Clune J. Understanding Neural Networks via Feature Visualization: A Survey. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 2019. Epub ahead of print 2019. DOI: 10.1007/978-3-030-28954-6_4.

92. Erhan D, Bengio Y, Courville A, Vincent P. Visualizing higher-layer features of a deep network. *Bernoulli*.

93. Serrano S, Smith NA. Is attention interpretable? In: *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*. 2020. Epub ahead of print 2020. DOI: 10.18653/v1/p19-1282.

94. Data access | CPRD, https://cprd.com/data-access (accessed 29 July 2022).

95. Delgado-Rodríguez M, Llorca J. Bias. *J Epidemiol Community Health (1978)* 2004; 58: 635–641.

96. Ayala Solares JR, Canoy D, Raimondi FED, Zhu Y, Hassaine A, Salimi-Khorshidi G, Tran J, Copland E, Zottoli M, Pinho-Gomes A-C, Nazarzadeh M, Rahimi K. Long-Term Exposure to Elevated Systolic Blood Pressure in Predicting Incident Cardiovascular Disease: Evidence From Large-Scale Routine Electronic Health Records. *J Am Heart Assoc* 2019; 8: e012129.

97. Pearson-Stuttard J, Bennett J, Cheng YJ, Vamos EP, Cross AJ, Ezzati M, Gregg EW. Trends in predominant causes of death in individuals with and without diabetes in England from 2001 to 2018: an epidemiological analysis of linked primary care records. *Lancet Diabetes Endocrinol*; 9. Epub ahead of print 2021. DOI: 10.1016/S2213-8587(20)30431-9.

98. Canoy D, Tran J, Zottoli M, Ramakrishnan R, Hassaine A, Rao S, Li Y, Salimi-Khorshidi G, Norton R, Rahimi K. Association between cardiometabolic disease multimorbidity and all-cause mortality in 2 million women and men registered in UK general practices. *BMC Med*; 19. Epub ahead of print 2021. DOI: 10.1186/s12916-021-02126-x.

99. Rahimi K, Mohseni H, Kiran A, Tran J, Nazarzadeh M, Rahimian F, Woodward M, Dwyer T, MacMahon S, Otto CM. Elevated blood pressure and risk of aortic valve disease: a cohort analysis of 5.4 million UK adults. *Eur Heart J* 2018; 44: 3596–3603.

100. Nazarzadeh M, Pinho-Gomes A-C, Smith Byrne K, Canoy D, Raimondi F, Ayala Solares JR, Otto CM, Rahimi K. Systolic Blood Pressure and Risk of Valvular Heart Disease. *JAMA Cardiol* 2019; 4: 788.

101. Pathak N, Zhang CX, Boukari Y, Burns R, Mathur R, Gonzalez-Izquierdo A, Denaxas S, Sonnenberg P, Hayward A, Aldridge RW. Development and Validation of a Primary Care Electronic Health Record Phenotype to Study Migration and Health in the UK. *International Journal of Environmental Research and Public Health 2021, Vol 18, Page 13304* 2021; 18: 13304.

102. Mathur R, Bhaskaran K, Chaturvedi N, Leon DA, van Staa T, Grundy E, Smeeth L. Completeness and usability of ethnicity data in UK-based primary care and hospital databases. *Journal of Public Health (United Kingdom)*; 36. Epub ahead of print 2014. DOI: 10.1093/pubmed/fdt116.

103. Bhaskaran K, Forbes HJ, Douglas I, Leon DA, Smeeth L. Representativeness and optimal use of body mass index (BMI) in the UK Clinical Practice Research

Datalink (CPRD). *BMJ Open*; 3. Epub ahead of print 2013. DOI: 10.1136/BMJOPEN-2013-003389.

104. Herrett E, Gadd S, Jackson R, al. et. Eligibility and subsequent burden of cardiovascular disease of four strategies for blood pressure-lowering treatment: a retrospective cohort study. *Lancet* 2019; 394: 663–671.

105. Herbert A, Wijlaars L, Zylbersztejn A, Cromwell D, Hardelid P. Data Resource Profile: Hospital Episode Statistics Admitted Patient Care (HES APC). *Int J Epidemiol*; 46. Epub ahead of print 2017. DOI: 10.1093/ije/dyx015.

106. Deaths - Office for National Statistics, https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths (accessed 29 July 2022).

107. Herrett E, Thomas SL, Schoonen WM, Smeeth L, Hall AJ. Validation and validity of diagnoses in the General Practice Research Database: A systematic review. *Br J Clin Pharmacol*. Epub ahead of print 2010. DOI: 10.1111/j.1365-2125.2009.03537.x.

108. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*; 25. Epub ahead of print 2019. DOI: 10.1038/s41591-018-0300-7.

109. Attia ZI, Noseworthy PA, Lopez-Jimenez F, Asirvatham SJ, Deshmukh AJ, Gersh BJ, Carter RE, Yao X, Rabinstein AA, Erickson BJ, Kapa S, Friedman PA. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *The Lancet* 2019; 394: 861–867.

110. Poplin R, Varadarajan A v., Blumer K, Liu Y, McConnell M v., Corrado GS, Peng L, Webster DR. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering 2018 2:3* 2018; 2: 158–164.

111. Liu D, Shin WY, Sprecher E, Conroy K, Santiago O, Wachtel G, Santillana M. Machine learning approaches to predicting no-shows in pediatric medical appointment. *npj Digital Medicine 2022 5:1* 2022; 5: 1–11.

112. Liang Z, Zhang G, Huang JX, Hu QV. Deep learning for healthcare decision making with EMRs. *Proceedings - 2014 IEEE International Conference on Bioinformatics and Biomedicine, IEEE BIBM 2014* 2014; 556–559.

113. Tran T, Nguyen TD, Phung D, Venkatesh S. Learning vector representation of medical objects via EMR-driven nonnegative restricted Boltzmann machines (eNRBM). *J Biomed Inform*. Epub ahead of print 2015. DOI: 10.1016/j.jbi.2015.01.012.

114. Nguyen P, Tran T, Wickramasinghe N, Venkatesh S. Deepr: A Convolutional Net for Medical Records. *IEEE J Biomed Health Inform*. Epub ahead of print 2017. DOI: 10.1109/JBHI.2016.2633963.

115. Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. *JMLR Workshop Conf Proc* 2016; 56: 301–318.

116. Choi E, Bahadori MT, Kulas JA, Schuetz A, Stewart WF, Sun J. RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism, http://arxiv.org/abs/1608.05745 (2016).

117. Kwon BC, Choi M, Kim JT, Choi E, Kim Y Bin, Kwon S, Sun J, Choo J. RetainVis : Visual Analytics with Interpretable and Interactive Recurrent Neural Networks on Electronic Medical Records. Epub ahead of print 2019. DOI: 10.1109/TVCG.2018.2865027.

118. Kim L, Kim J-A, Kim S. A guide for the utilization of Health Insurance Review and Assessment Service National Patient Samples. *Epidemiol Health*; 36. Epub ahead of print 2014. DOI: 10.4178/epih/e2014008.

119. Devlin J, Chang M-W, Lee K, Google KT, Language AI. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North* 2019; 4171–4186.

120. NHS Digital. Read Code Map, https://isd.digital.nhs.uk/trud3/user/guest/group/0/pack/9 (2020).

121. Organization WH. ICD-10 : international statistical classification of diseases and related health problems : tenth revision. 2004; Spanish version, 1st edition published by PAHO as.

122. Kuan V, Denaxas S, Gonzalez-Izquierdo A, Direk K, Bhatti O, Husain S, Sutaria S, Hingorani M, Nitsch D, Parisinos CA, Lumbers RT, Mathur R, Sofat R, Casas JP, Wong ICK, Hemingway H, Hingorani AD. A chronological map of 308 physical and mental health conditions from 4 million individuals in the English National Health Service. *Lancet Digit Health* 2019; 1: e63–e77.

123. Li Y, Rao S, Solares JRA, Hassaine A, Ramakrishnan R, Canoy D, Zhu Y, Rahimi K, Salimi-Khorshidi G. BEHRT: Transformer for Electronic Health Records. *Sci Rep* 2020; 10: 7155.

124. Pascanu R, Mikolov T, Bengio Y. On the difficulty of training Recurrent Neural Networks, http://arxiv.org/abs/1211.5063 (2012).

125. Hammer B. Neural Smithing – Supervised Learning in Feedforward Artificial Neural Networks. *Pattern Analysis & Applications*; 4. Epub ahead of print 2001. DOI: 10.1007/s100440170029.

126. Goodfellow I, Bengio Y, Courville A. *Deep learning (Vol.1, No.2)*. 2016.

127. Choi E. RETAIN issue #3, https://github.com/mp2893/retain/issues/3 (2017, accessed 3 August 2022).

128. Snoek, Jasper; Larochelle, Hugo; Adams R. Practical Bayesian Optimization of Machine Learning Algorithms. *NIPS* 2017; 2: e540.

129. The Health and Social Care Information Centre. The Quality and Outcomes Framework (QOF) 2004/05. *The Health and Social Care Information Centre.*

130. Trowell WJ. "British National Formulary". *British Medical Journal (Clinical research ed.)*. Epub ahead of print 1981. DOI: 10.1136/bmj.282.6269.1078.

131. Rao S, Li Y, Ramakrishnan R, Hassaine A, Canoy D, Cleland JG, Lukasiewicz T, Salimi-Khorshidi G, Rahimi K. An explainable Transformer-based deep learning model for the prediction of incident heart failure. *IEEE J Biomed Health Inform* 2022; 1–1.

132. Yang H, Negishi K, Otahal P, Marwick TH. Clinical prediction of incident heart failure risk : a systematic review and meta-analysis. 2015; 1–8.

133. Guan C, Wang X, Zhang Q, Chen R, He D, Xie X. Towards a Deep and Unified Understanding of Deep Neural Models in {NLP}. *Proceedings of the 36th International Conference on Machine Learning*.

134. Hippisley-Cox J, Coupland C. Development and validation of risk prediction equations to estimate future risk of heart failure in patients with diabetes: a prospective cohort study. *BMJ Open* 2015; 5: e008503.

135. Tromp J, Pangiagua S, Lau E, Allen N, Blaha M. Age dependent associations of risk factors with heart failure: pooled population based cohort study. *Br Med J*; 372.

136. Jacobsen SJ, Freedman DS, Hoffmann RG, Gruchow HW, Anderson AJ, Barboriak JJ. Cholesterol and coronary artery disease: Age as an effect modifier. *J Clin Epidemiol*; 45. Epub ahead of print 1992. DOI: 10.1016/0895-4356(92)90145-D.

137. Campbell TJ, MacDonald PS. Digoxin in heart failure and cardiac arrhythmias. *Medical journal of Australia* 2003; 179: 98–102.

138. Goodwin RD, Davidson KW, Keyes K. Mental disorders and cardiovascular disease among adults in the United States. *J Psychiatr Res* 2009; 43: 239–246.

139. Butler CC, Hood K, Kelly MJ, Goossens H, Verheij T, Little P, Melbye H, Torres A, Mölstad S, Godycki-Cwirko M, Almirall J, Blasi F, Schaberg T, Edwards P, Rautakorpi UM, Hupkova H, Wood J, Nuttall J, Coenen S. Treatment of acute cough/lower respiratory tract infection by antibiotic class and associated outcomes: A 13 European country observational study in primary care. *Journal of Antimicrobial Chemotherapy* 2010; 65: 2472–2478.

140. Macie C, Wooldrage K, Manfreda J, Anthonisen N. Cardiovascular morbidity and the use of inhaled bronchodilators. *International Journal of COPD* 2008; 3: 163–169.

141. Bleumink GS, Feenstra J, Sturkenboom MCJM, Stricker BHC. Nonsteroidal anti-inflammatory drugs and heart failure. *Drugs* 2003; 63: 525–534.

142. Mäenpää J, Pelkonen O. Cardiac safety of ophthalmic timolol. *Expert Opin Drug Saf*. Epub ahead of print 2016. DOI: 10.1080/14740338.2016.1225718.

143. Harasymowycz P, Birt C, Gooi P, Heckler L, Hutnik C, Jinapriya D, Shuba L, Yan D, Day R. Medical Management of Glaucoma in the 21st Century from a Canadian Perspective. *Journal of Ophthalmology*. Epub ahead of print 2016. DOI: 10.1155/2016/6509809.

144. Abraham WT. β-blockers: The new standard of therapy for mild heart failure. *Archives of Internal Medicine*. Epub ahead of print 2000. DOI: 10.1001/archinte.160.9.1237.

145. Lièvre M, Morand S, Besse B, Fiessinger JN, Boissel JP. Oral beraprost sodium, a prostaglandin I2 analogue, for intermittent claudication: A double-blind, randomized, multicenter controlled trial. *Circulation*. Epub ahead of print 2000. DOI: 10.1161/01.CIR.102.4.426.

146. Mohler ER, Hiatt WR, Olin JW, Wade M, Jeffs R, Hirsch AT. Treatment of intermittent claudication with beraprost sodium, an orally active prostaglandin I2 analogue: Double-blinded, randomized, controlled trial. *J Am Coll Cardiol*. Epub ahead of print 2003. DOI: 10.1016/S0735-1097(03)00299-7.

147. Pass HI, Pogrebniak HW. Potential uses of prostaglandin E1 analog for cardiovascular disease [5]. *Journal of Thoracic and Cardiovascular Surgery*. Epub ahead of print 1994. DOI: 10.1016/s0022-5223(94)70312-4.

148. Varga Z, Sabzwari S rafay ali, Vargova V. Cardiovascular Risk of Nonsteroidal Anti-Inflammatory Drugs: An Under-Recognized Public Health Issue. *Cureus*; 9. Epub ahead of print 8 April 2017. DOI: 10.7759/cureus.1144.

149. Rahimi K, Bidel Z, Nazarzadeh M, Copland E, Canoy D, Wamil M, Majert J, McManus R, Adler A, Agodoa L, Algra A, Asselbergs FW, Beckett NS, Berge E, Black H, Boersma E, Brouwers FPJ, Brown M, Brugts JJ, et al. Age-stratified and blood-pressure-stratified effects of blood-pressure-lowering pharmacotherapy for the prevention of cardiovascular disease and death: an individual participant-level data meta-analysis. *The Lancet*; 398. Epub ahead of print 2021. DOI: 10.1016/S0140-6736(21)01921-8.

150. Copland E, Canoy D, Nazarzadeh M, Bidel Z, Ramakrishnan R, Woodward M, Chalmers J, Teo KK, Pepine CJ, Davis BR, Kjeldsen S, Sundström J, Rahimi K, Adler A, Agodoa L, Algra A, Asselbergs FW, Beckett N, Berge E, et al. Antihypertensive treatment and risk of cancer: an individual participant data meta-analysis. *Lancet Oncol* 2021; 22: 558–570.

151. Rose S, van der Laan MJ. Targeted Learning: Causal Inference for Observational and Experimental Data. *Targeted Learning: Causal Inference for Observational and Experimental Data*.

152. Shi C, Blei DM, Veitch V. Adapting Neural Networks for the Estimation of Treatment Effects. 2019; 1–14.

153. Reichenheim ME, Coutinho ES. Measures and models for causal inference in cross-sectional studies: Arguments for the appropriateness of the prevalence odds ratio and related logistic regression. *BMC Medical Research Methodology*; 10. Epub ahead of print 2010. DOI: 10.1186/1471-2288-10-66.

154. Robins JM, Hernán MÁ, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*; 11. Epub ahead of print 2000. DOI: 10.1097/00001648-200009000-00011.

155. Glynn AN, Quinn KM. An introduction to the augmented inverse propensity weighted estimator. *Political Analysis*. Epub ahead of print 2009. DOI: 10.1093/pan/mpp036.

156. van der Laan MJ, Rubin D. Targeted maximum likelihood learning. *International Journal of Biostatistics*. Epub ahead of print 2006. DOI: 10.2202/1557-4679.1043.

157. Kang JDY, Schafer JL. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*; 22. Epub ahead of print 2007. DOI: 10.1214/07-STS227.

158. Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, Robins J. Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal* 2018; 21: C1–C68.

159. Decruyenaere A, Steen J, Colpaert K, Benoit DD, Decruyenaere J, Vansteelandt S. The obesity paradox in critically ill patients: a causal learning approach to a casual finding. *Crit Care* 2020; 24: 1–11.

160. Zhang Y, Lin LA, Starkopf L, Chen J, Wang WWB. Estimation of causal effect in integrating randomized clinical trial and observational data – An example application to cardiovascular outcome trial. *Contemp Clin Trials*; 107. Epub ahead of print 2021. DOI: 10.1016/j.cct.2021.106492.

161. Sofrygin O, Zhu Z, Schmittdiel JA, Adams AS, Grant RW, van der Laan MJ, Neugebauer R. Targeted learning with daily EHR data. *Stat Med*. Epub ahead of print 2019. DOI: 10.1002/sim.8164.

162. Le L, Patterson A, White M. Supervised autoencoders: Improving generalization performance with unsupervised regularizers. In: *Advances in Neural Information Processing Systems*. 2018.

163. Melamud O, Bornea M, Barker K. Combining unsupervised pre-training and annotator rationales to improve low-shot text classification. In: *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*. 2020. Epub ahead of print 2020. DOI: 10.18653/v1/d19-1401.

164. Liebel L, Körner M. Auxiliary tasks in multi-task learning. *arXiv*.

165. Louizos C, Shalit U, Mooij J, Sontag D, Zemel R, Welling M. Causal Effect Inference with Deep Latent-Variable Models, http://arxiv.org/abs/1705.08821 (2017).

166. Shalit U, Johansson FD, Sontag D. Estimating individual treatment effect: generalization bounds and algorithms, http://arxiv.org/abs/1606.03976 (2016).

167. Veitch V, Sridhar D, Blei DM. Using Text Embeddings for Causal Inference. 2019; 1–11.

168. Izdebski A, Thoral PJ, Lalisang RCA, McHugh DM, Gommers D, Cremer OL, Bosman RJ, Rigter S, Wils E-J, Frenzel T, Dongelmans DA, de Jong R, Peters MAA, Kamps MJA, Ramnarain D, Nowitzky R, Nooteboom FGCA, de Ruijter W, Urlings-Strop LC, et al. A pragmatic approach to estimating average treatment effects from EHR data: the effect of prone positioning on mechanically ventilated COVID-19 patients, http://arxiv.org/abs/2109.06707 (2021).

169. Bangalore S, Kumar S, Kjeldsen SE, Makani H, Grossman E, Wetterslev J, Gupta AK, Sever PS, Gluud C, Messerli FH. Antihypertensive drugs and risk of cancer: Network meta-analyses and trial sequential analyses of 324 168 participants from randomised trials. *Lancet Oncol*. Epub ahead of print 2011. DOI: 10.1016/S1470-2045(10)70260-6.

170. Rao S, Mamouei M, Salimi-Khorshidi G, Li Y, Ramakrishnan R, Hassaine A, Canoy D, Rahimi K. Targeted-BEHRT: Deep Learning for Observational Causal Inference on Longitudinal Electronic Health Records. *IEEE Trans Neural Netw Learn Syst* 2022; 1–12.

171. Payne R, Denholm R. CPRD product code lists used to define long-term preventative, high-risk, and palliative medication. *University of Bristol*. Epub ahead of print 2018. DOI: 10.5523/BRIS.K38GHXKCUB622603I5WQ6BWAG.

172. Yao L, Huai M, Li S, Gao J, Li Y, Zhang A. Representation learning for treatment effect estimation from observational data. *Adv Neural Inf Process Syst* 2018; 2018-Decem: 2633–2643.

173. Nørgaard M, Ehrenstein V, Vandenbroucke JP. Confounding in observational studies based on large health care databases: problems and potential solutions – a primer for the clinician. *Clin Epidemiol*; Volume 9. Epub ahead of print March 2017. DOI: 10.2147/CLEP.S129879.

174. Funk MJ, Westreich D, Wiesen C, Stürmer T, Brookhart MA, Davidian M. Doubly robust estimation of causal effects. *Am J Epidemiol* 2011; 173: 761–767.

175. Read Codes. *NHS*, https://digital.nhs.uk/services/ terminology-and-classifications/read-codes (accessed 19 October 2018).

176. Chipman HA, George EI, McCulloch RE. BART: Bayesian additive regression trees. *Annals of Applied Statistics*. Epub ahead of print 2012. DOI: 10.1214/09-AOAS285.

177. Stürmer T, Rothman KJ, Avorn J, Glynn RJ. Treatment effects in the presence of unmeasured confounding: Dealing with observations in the tails of the propensity score distribution-A simulation study. *Am J Epidemiol*. Epub ahead of print 2010. DOI: 10.1093/aje/kwq198.

178. Weisstein EW. Normal Sum Distribution. *MathWorld*.

179. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Köpf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library, http://arxiv.org/abs/1912.01703 (2019).

180. Kingma DP, Ba JL. Adam: A method for stochastic optimization. In: *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. 2015.

181. Levy J. An Easy Implementation of CV-TMLE. *arXiv*.

182. Nelder JA, Mead R. A Simplex Method for Function Minimization. *Comput J*; 7. Epub ahead of print 1965. DOI: 10.1093/comjnl/7.4.308.

183. Zhang L, Wang Y, Ostropolets A, Mulgrave JJ, Blei DM, Hripcsak G. The Medical Deconfounder: Assessing Treatment Effects with Electronic Health Records. 2019; 1–22.

184. Pearl J. Remarks on the method of propensity score. *Statistics in Medicine* 2009; 28: 1415–1416.

185. Ding P, Miratrix LW. To Adjust or Not to Adjust? Sensitivity Analysis of M-Bias and Butterfly-Bias. *J Causal Inference*; 3. Epub ahead of print 2014. DOI: 10.1515/jci-2013-0021.

186. Collaboration PS. Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies. *Lancet* 2002; 360: 1903–1913.

187. Age-specific association between blood pressure and vascular and non-vascular chronic diseases in 0·5 million adults in China: a prospective cohort study - The Lancet Global Health, https://www.thelancet.com/journals/langlo/article/PIIS2214-109X(18)30217-1/fulltext (accessed 15 December 2021).

188. Sundström J, Sheikhi R, Östgren CJ, Svennblad B, Bodegard J, Nilsson PM, Johansson G. Blood pressure levels and risk of cardiovascular events and mortality in type-2 diabetes: Cohort study of 34009 primary care patients. *J Hypertens* 2013; 31: 1603–1610.

189. Brunström M, Carlberg B. Effect of antihypertensive treatment at different blood pressure levels in patients with diabetes mellitus: systematic review and meta-analyses. *BMJ*; 352.

190. Emdin CA, Anderson SG, Woodward M, Rahimi K. Usual Blood Pressure and Risk of New-Onset Diabetes Evidence from 4.1 Million Adults and a Meta-Analysis of Prospective Studies. *J Am Coll Cardiol*. Epub ahead of print 2015. DOI: 10.1016/j.jacc.2015.07.059.

191. Denaxas S, Gonzalez-Izquierdo A, Direk K, Fitzpatrick N, Banerjee A, Dobson R, Fatemifar G, Kuan V, Lumbers T, Pasea L, Patel R, Hingorani A, Sudlow C,

Hemingway H. UK phenomics platform for developing and validating EHR phenotypes: CALIBER. *bioRxiv* 2019; 539403.

192. Sundström J, Arima H, Jackson R, al. et. Effects of blood pressure reduction in mild hypertension: a systematic review and meta-analysis. *Ann Intern Med* 2015; 162: 184–191.

193. Cederholm J, Gudbjörnsdottir S, Eliasson B, Zethelius B, Eeg-Olofsson K, Nilsson PM. Blood pressure and risk of cardiovascular diseases in type 2 diabetes: Further findings from the Swedish National Diabetes Register (NDR-BP II). *J Hypertens* 2012; 30: 2020–2030.

194. Association AD. 9. Cardiovascular Disease and Risk Management. *Diabetes Care* 2017; 40: S75–S87.

195. Cosentino F, Grant PJ, Aboyans V, Bailey CJ, Ceriello A, Delgado V, Federici M, Filippatos G, Grobbee DE, Hansen TB, Huikuri H v., Johansson I, Juni P, Lettino M, Marx N, Mellbin LG, Ostgren CJ, Rocca B, Roffi M, et al. 2019 ESC Guidelines on diabetes, pre-diabetes, and cardiovascular diseases developed in collaboration with the EASD. *Eur Heart J* 2020; 41: 255–323.

196. Williams B, Masera G. *2018 ESC / ESH Guidelines for the management of arterial hypertension*. 2018. Epub ahead of print 2018. DOI: 10.1093/eurheartj/ehy339.

197. Whelton PK, Carey RM, Aronow WS, Casey DE, Collins KJ, Himmelfarb CD, DePalma SM, Gidding S, Jamerson KA, Jones DW, MacLaughlin EJ, Muntner P, Ovbiagele B, Smith SC, Spencer CC, Stafford RS, Taler SJ, Thomas RJ, Williams KA, et al. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: Executive summary: A report of the

American college of cardiology/American Heart Association task force on clinical practice guidelines. *Hypertension* 2018; 71: 1269–1324.

198. Arana A, Margulis A v., Varas-Lorenzo C, Bui CL, Gilsenan A, McQuay LJ, Reynolds M, Rebordosa C, Franks B, de Vogel S, Appenteng K, Perez-Gutthann S. Validation of cardiovascular outcomes and risk factors in the Clinical Practice Research Datalink in the United Kingdom. *Pharmacoepidemiol Drug Saf* 2021; 30: 237.

199. Chen W, Thomas J, Sadatsafavi M, FitzGerald JM. Risk of cardiovascular comorbidity in patients with chronic obstructive pulmonary disease: A systematic review and meta-analysis. *Lancet Respir Med*; 3. Epub ahead of print 2015. DOI: 10.1016/S2213-2600(15)00241-6.

200. Huang Y, Cai X, Mai W, Li M, Hu Y. Association between prediabetes and risk of cardiovascular disease and all cause mortality: Systematic review and meta-analysis. *BMJ (Online)*; 355. Epub ahead of print 2016. DOI: 10.1136/BMJ.I5953.

201. Rahimi K, Emdin C, MacMahon S. The epidemiology of blood pressure and its worldwide management. *Circ Res* 2015; 116: 925–936.

202. Williams B, Mancia G, Spiering W, al. et. 2018 ESC/ESH guidelines for the management of arterial hypertension. *Eur Heart J* 2018; 39: 3021–3104.

203. Chalmers JD, Poole C, Webster S, Tebboth A, Dickinson S, Gayle A. Assessing the healthcare resource use associated with inappropriate prescribing of inhaled corticosteroids for people with chronic obstructive pulmonary disease (COPD) in GOLD groups A or B: An observational study using the Clinical Practice Research Datalink (CPRD). *Respir Res*; 19. Epub ahead of print 2018. DOI: 10.1186/s12931-018-0767-2.

204. Hutcheon JA, Chiolero A, Hanley JA. Random measurement error and regression dilution bias. *BMJ (Online)*; 340. Epub ahead of print 2010. DOI: 10.1136/bmj.c2289.

205. MacMahon S, Peto R, Collins R, Godwin J, MacMahon S, Cutler J, Sorlie P, Abbott R, Collins R, Neaton J, Abbott R, Dyer A, Stamler J. Blood pressure, stroke, and coronary heart disease. Part 1, prolonged differences in blood pressure: prospective observational studies corrected for the regression dilution bias. *The Lancet*; 335. Epub ahead of print 1990. DOI: 10.1016/0140-6736(90)90878-9.

206. Lewington S, Clarke R, Qizilbash N, Peto R, Collins R. Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies. *Lancet* 2002; 360: 1903–1913.

207. Mueller NT, Noya-Alarcon O, Contreras M, Appel LJ, Dominguez-Bello MG. Association of Age with Blood Pressure Across the Lifespan in Isolated Yanomami and Yekwana Villages. *JAMA Cardiology*; 3. Epub ahead of print 2018. DOI: 10.1001/jamacardio.2018.3676.

208. Quint JK, Müllerova H, DiSantostefano RL, Forbes H, Eaton S, Hurst JR, Davis K, Smeeth L. Validation of chronic obstructive pulmonary disease recording in the Clinical Practice Research Datalink (CPRD-GOLD). *BMJ Open*. Epub ahead of print 2014. DOI: 10.1136/bmjopen-2014-005540.

209. Ennis ZN, Dideriksen D, Vægter HB, Handberg G, Pottegård A. Acetaminophen for Chronic Pain: A Systematic Review on Efficacy. *Basic and Clinical Pharmacology and Toxicology*; 118. Epub ahead of print 2016. DOI: 10.1111/bcpt.12527.

210. USP NF 2003, USP NF 21. The United States pharmacopeia – The national formulary. *United States Pharmacopeial Convention, Rockville, MD*.

211. Schofield P. The assessment of pain in older people: UK national guidelines. *Age Ageing*; 47. Epub ahead of print 2018. DOI: 10.1093/ageing/afx192.

212. Sura L, Madhavan A, Carnaby G, Crary MA. Dysphagia in the elderly: Management and nutritional considerations. *Clinical Interventions in Aging*; 7. Epub ahead of print 2012. DOI: 10.2147/CIA.S23404.

213. Hyllested M, Jones S, Pedersen JL, Kehlet H. Comparative effect of paracetamol, NSAIDs or their combination in postoperative pain management: A qualitative review. *Br J Anaesth*; 88. Epub ahead of print 2002. DOI: 10.1093/bja/88.2.199.

214. Mente A, O'Donnell M, Rangarajan S, McQueen M, Dagenais G, Wielgosz A, Lear S, Ah STL, Wei L, Diaz R, Avezum A, Lopez-Jaramillo P, Lanas F, Mony P, Szuba A, Iqbal R, Yusuf R, Mohammadifard N, Khatib R, et al. Urinary sodium excretion, blood pressure, cardiovascular disease, and mortality: a community-level prospective epidemiological cohort study. *The Lancet*; 392. Epub ahead of print 2018. DOI: 10.1016/S0140-6736(18)31376-X.

215. George J, Majeed W, Mackenzie IS, MacDonald TM, Wei L. Association between cardiovascular events and sodium-containing effervescent, dispersible, and soluble drugs: Nested case-control study. *BMJ (Online)*; 347. Epub ahead of print 2013. DOI: 10.1136/bmj.f6954.

216. Zeng C, Rosenberg L, Li X, Djousse L, Wei J, Lei G, Zhang Y. Sodium-containing acetaminophen and cardiovascular outcomes in individuals with and without hypertension. *Eur Heart J*. Epub ahead of print 2022. DOI: 10.1093/eurheartj/ehac059.

217. Tran J, Norton R, Conrad N, Rahimian F, Canoy D, Nazarzadeh M, Rahimi K. Patterns and temporal trends of comorbidity among adult patients with incident cardiovascular disease in the UK between 2000 and 2014: A population-based cohort study. *PLoS Med* 2018; 15: e1002513.

218. van Buuren S, Groothuis-Oudshoorn K. *Journal of Statistical Software mice: Multivariate Imputation by Chained Equations in R*, http://www.jstatsoft.org/ (2011).

219. Baron RM, Kenny DA. The Moderator-Mediator Variable Distinction in Social Psychological Research. Conceptual, Strategic, and Statistical Considerations. *J Pers Soc Psychol*; 51. Epub ahead of print 1986. DOI: 10.1037/0022-3514.51.6.1173.

220. Richiardi L, Bellocco R, Zugna D. Mediation analysis in epidemiology: Methods, interpretation and bias. *Int J Epidemiol*; 42. Epub ahead of print 2013. DOI: 10.1093/ije/dyt127.

221. Almirall J, Rofes L, Serra-Prat M, Icart R, Palomera E, Arreola V, Clavé P. Oropharyngeal dysphagia is a risk factor for community-acquired pneumonia in the elderly. *European Respiratory Journal*; 41. Epub ahead of print 2013. DOI: 10.1183/09031936.00019012.

222. Leslie P, Smithard DG. Is Dysphagia Under Diagnosed or is Normal Swallowing More Variable than We Think? Reported Swallowing Problems in People Aged 18–65 Years. *Dysphagia*; 36. Epub ahead of print 2021. DOI: 10.1007/s00455-020-10213-z.

223. Kawashima K, Motohashi Y, Fujishima I. Prevalence of dysphagia among community-dwelling elderly individuals as estimated using a questionnaire for

dysphagia screening. *Dysphagia*; 19. Epub ahead of print 2004. DOI: 10.1007/s00455-004-0013-6.

224. Chadwick DD, Jolliffe J. A descriptive investigation of dysphagia in adults with intellectual disabilities. *Journal of Intellectual Disability Research*; 53. Epub ahead of print 2009. DOI: 10.1111/j.1365-2788.2008.01115.x.

225. Bradford A, Kunik ME, Schulz P, Williams SP, Singh H. Missed and delayed diagnosis of dementia in primary care: Prevalence and contributing factors. *Alzheimer Disease and Associated Disorders*; 23. Epub ahead of print 2009. DOI: 10.1097/WAD.0b013e3181a6bebc.

226. Round T, Steed L, Shankleman J, Bourke L, Risi L. Primary care delays in diagnosing cancer: What is causing them and what can we do about them? *J R Soc Med*; 106. Epub ahead of print 2013. DOI: 10.1177/0141076813504744.

227. International Dysphagia Diet Standardisation Initiative Foundation. International Dysphagia Diet Standardisation Initiative.

228. *Safe administration of medications for adults with Swallowing Difficulties (Dysphagia)*.

229. Agüero-Torres H, Fratiglioni L, Guo Z, Viitanen M, Winblad B. Mortality from dementia in advanced age: A 5-year follow-up study of incident dementia cases. *J Clin Epidemiol*; 52. Epub ahead of print 1999. DOI: 10.1016/S0895-4356(99)00067-0.

230. Pearl J. *Causality: Models, reasoning, and inference, second edition*. 2011. Epub ahead of print 2011. DOI: 10.1017/CBO9780511803161.

231. Noroozian M, Raeesi S, Hashemi R, Khedmat L, Vahabi Z. Pain: The neglect issue in old people's life. *Open Access Maced J Med Sci*; 6. Epub ahead of print 2018. DOI: 10.3889/oamjms.2018.335.

232. Li Y, Sperrin M, Ashcroft DM, van Staa TP. Consistency of variety of machine learning and statistical models in predicting clinical risks of individual patients: Longitudinal cohort study using cardiovascular disease as exemplar. *The BMJ*; 371. Epub ahead of print 2020. DOI: 10.1136/bmj.m3919.

233. Zheng W, van der Laan MJ. Targeted maximum likelihood estimation of natural direct effects. *International Journal of Biostatistics*; 8. Epub ahead of print 2012. DOI: 10.2202/1557-4679.1361.

234. Vanderweele TJ. Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology*; 21. Epub ahead of print 2010. DOI: 10.1097/EDE.0b013e3181df191c.

235. Cole SR, Hernán MA. Fallibility in estimating direct effects. *Int J Epidemiol*; 31. Epub ahead of print 2002. DOI: 10.1093/ije/31.1.163.

236. Last JM. A dictionary of epidemiology. *International Journal of Epidemiology*; 15. Epub ahead of print 1986. DOI: 10.1093/ije/15.2.277.

237. Hansen H, Schäfer I, Schön G, Riedel-Heller S, Gensichen J, Weyerer S, Petersen JJ, König HH, Bickel H, Fuchs A, Höfels S, Wiese B, Wegscheider K, van den Bussche H, Scherer M. Agreement between self-reported and general practitioner-reported chronic conditions among multimorbid patients in primary care - Results of the MultiCare Cohort Study. *BMC Fam Pract*; 15. Epub ahead of print 2014. DOI: 10.1186/1471-2296-15-39.

238. Strubell E, Ganesh A, McCallum A. Energy and policy considerations for deep learning in NLP. In: *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*. 2020.

239. Falcon W, Borovec J, Wälchli A, Eggert N, Schock J, Jordan J, Skafte N, Bereznyuk V, Harris E, Murrell T. PyTorchLightning/pytorch-lightning: 0.7. 6 release. *Zenodo: Geneva, Switzerland*.

240. Bakris G, Ali W, Parati G. ACC/AHA Versus ESC/ESH on Hypertension Guidelines: JACC Guideline Comparison. *Journal of the American College of Cardiology*; 73. Epub ahead of print 2019. DOI: 10.1016/j.jacc.2019.03.507.

241. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med Res Methodol*; 18. Epub ahead of print 2018. DOI: 10.1186/s12874-018-0482-1.

242. SNOMED CT.

243. Nelson SJ, Zeng K, Kilbourne J, Powell T, Moore R. Normalized names for clinical drugs: RxNorm at 6 years. *Journal of the American Medical Informatics Association*; 18. Epub ahead of print 2011. DOI: 10.1136/amiajnl-2011-000116.

244. Gensheimer MF, Narasimhan B. A scalable discrete-time survival model for neural networks. *PeerJ*; 2019. Epub ahead of print 2019. DOI: 10.7717/peerj.6257.

245. Kennedy EH. Towards Optimal Doubly Robust Estimation of Heterogeneous Causal Effects.

246. Danaei G, Rodríguez L. Observational data for comparative effectiveness research: an emulation of randomised trials to estimate the effect of statins on primary prevention of coronary heart disease. *… in Medical Research* 2013; 22: 1–26.

247. Díaz I, van der Laan MJ. Targeted Data Adaptive Estimation of the Causal Dose–Response Curve. *J Causal Inference* 2013; 1: 171–192.