
# Deep Learning for Electronic Health Records:

# Risk Prediction, Explainability, and Uncertainty

**Yikuan Li**

St Cross College, University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas Term 2022

Supervised by:

Professor Kazem Rahimi

Professor Thomas Lukasiewicz

Doctor Mohammadhossein Mamouei

To my wife and my parents for their enormous love and support.

# Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text. This dissertation contains approximately 46,000 words including appendices, footnotes, tables, and equations.

Yikuan Li

September 2022

# Abstract

**Background**: Risk models are essential for care planning and disease prevention. The unsatisfactory performance of the established clinical models has raised broad awareness and concerns. An accurate, explainable, and reliable risk model is highly beneficial but remains a challenge.

**Objective**: This thesis aims to develop deep learning models that can make more accurate risk predictions with the provision of uncertainty estimation and the ability to provide medical explanations using a large and representative electronic health records (EHR) dataset.

**Methods**: We investigated three directions in this thesis: risk prediction, explainability, and uncertainty estimation. For risk prediction, we investigated deep learning tools that can incorporate the minimal processed EHR for modelling and comprehensively compared them with the established machine learning and clinical models. Additionally, the post-hoc explanations were applied to deep learning models for medical information retrieval, and we specifically looked into explanations in risk association and counterfactual reasoning. Uncertainty estimation was qualitatively investigated using probabilistic modelling techniques. Our analyses relied on Clinical Practice Research Datalink, which contains anonymised EHR collected from primary care, secondary care, and death registration and is representative of the UK population.

**Results**: We introduced a deep learning model, named BEHRT, that can incorporate minimal processed EHR for risk prediction. Without expert engagement, it learned meaningful representations that can automatically cluster highly correlated diseases. Compared to the established machine learning and clinical models that relied on expert-selected predictors, our proposed deep learning model showed superior performance on a wide range of risk prediction tasks and highlighted the necessity of recalibration when applying a risk model to a population with severe prior distribution shifts, and the importance of regular model updating to preserve the model's discrimination performance under temporal data shifts.

Additionally, we showed that the deep learning model explanation is an excellent tool for discovering risk factors. By explaining the deep learning model, we not only identified factors that were highly consistent with the established evidence but also those that have not been considered in expert-driven studies. Furthermore, the deep learning model also captured the interplay between risk and treated risk and the differential association of medications across different years, which would be difficult if the temporal context was not included in the modelling. Besides the explanations in terms of association, we introduced a framework that can achieve accurate risk prediction, while enabling counterfactual reasoning under hypothetical interventions. This offers counterfactual explanations that could inform clinicians for selection of those who will benefit the most. We demonstrated the benefit of the proposed framework using two exemplary case studies.

Furthermore, transforming a deterministic deep learning model to probabilistic can make predictions with an uncertainty range. We showed that such information has many potential implications in practice, such as quantifying the confidence of a decision, indicating data insufficiency, distinguishing the correct and incorrect predictions, and indicating risk associations.

**Conclusions:** Deep learning models led to substantially improved performance for risk prediction. The ability of uncertainty estimation can quantify the confidence of risk prediction to further inform clinical decision-making. Deep learning model explanation can generate hypotheses to guide medical research and provide counterfactual analysis to assist clinical decision-making. This encouraging evidence supports the great potential of incorporating deep learning methods into electronic health records to inform a wide range of health applications such as care planning, disease prevention, and medical study design.

# Acknowledgement

I am most grateful to my supervisors, Prof. Kazem Rahimi, Prof. Thomas Lukasiewicz, and Dr. Mohammadhossein Mamouei, for their ongoing guidance and support. Their expertise, perspective, and advice have been invaluable in forming me as a researcher.

Besides my supervisors, I would like to thank Dr. Gholamreza Salimi-Khorshidi for his mentorship and guidance throughout my DPhil. His insightful comments have led to considerable improvements to my research and have helped me develop my research skills.

I am also very thankful to all my colleagues, and in particular to Shishir Rao, Dexter Canoy, Milad Nazarzadeh, Abdelaali Hassaine, Emma Copland, and Naseem Akhtar, for the many inspiring research conversations and their enormous support outside of research.

The British Heart Foundation has supported this work, and I would like to express my gratitude for the generous funding provided.

Lastly, I would like to thank my family and friends who have been there for me throughout the journey, particularly my mom and wife, who always believed in me, encouraged me to persevere, and supported me unconditionally.

# Publications

**Publications contributing to DPhil thesis**

**Li, Y.**[1], Rao, S.[1], Solares, J.R.A., Hassaine, A., Ramakrishnan, R., Canoy, D., Zhu, Y., Rahimi, K., Salimi-Khorshidi, G., BEHRT: Transformer for Electronic Health Records. Scientific Reports (2020). doi:10.1038/s41598-020-62922-y.

**Li, Y.**, Rao, S., Hassaine, A., Ramakrishnan, R., Canoy, D., Salimi-Khorshidi, G., Mamouei, M., Lukasiewicz, T., Rahimi, K., Deep Bayesian Gaussian processes for uncertainty estimation in electronic health records. Scientific Reports (2021). doi: 10.1038/s41598-021-00144-6.

Rao, S. [1], **Li, Y.** [1], Ramakrishnan, R., Hassaine, A., Canoy, D., Cleland, J., Lukasiewicz, T., Salimi-Khorshidi, G., Rahimi, K., An Explainable Transformer-Based Deep Learning Model for the Prediction of Incident Heart Failure. IEEE Journal of Biomedical and Health Informatics (2022), doi:10.1109/JBHI.2022.3148820.

**Li, Y.**, Salimi-Khorshidi, G., Rao, S., Canoy, D., Hassaine, A., Lukasiewicz, T., Rahimi, K., Mamouei, M., Validation of risk prediction models applied to longitudinal electronic health record data for the prediction of major cardiovascular events in the presence of data shifts. European Hear Journal Digital Health (2022), doi:10.1093/ehjdh/ztac061.

[1] Equal first author

**Li, Y.**, Mamouei, M., Salimi-Khorshidi, G., Rao, S., Hassaine, A., Canoy, D., Lukasiewicz, T., Rahimi, K., 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), doi: 10.1109/JBHI.2022.3224727.

**Li, Y.**, Mamouei, M., Rao, S., Hassaine, A., Canoy, D., Lukasiewicz, T., Rahimi, K., Salimi-Khorshidi, G., Clinical Outcome Prediction under Hypothetical Interventions - A Representation Learning Framework for Counterfactual Reasoning. (Under review)

**Other relevant publications published during DPhil**

Hassaine, A., Canoy, D., Solares, J.R.A., Zhu, Y., Rao, S., **Li, Y.**, Learning multimorbidity patterns from electronic health records using Non-negative Matrix Factorisation. Journal of Biomedical Informatics (2020), doi:10.1016/j.jbi.2020.103606.

Canoy, D., Nazarzadeh, M., Copland, E., Bidel, Z., Rao, S., **Li, Y.**, Rahimi, K., How Much Lowering of Blood Pressure Is Required to Prevent Cardiovascular Disease in Patients With and Without Previous Cardiovascular Disease? Current Cardiology Reports (2022). doi:10.1007/s11886-022-01706-4.

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 Medicine (2021). doi:10.1186/s12916-021-02126-x.

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 Transactions on Neural Networks and Learning Systems (2022). doi: 10.1109/TNNLS.2022.3183864.

Azhir, A., Talebi, S., Merino, L.H., **Li, Y.**, Lukasiewicz, T., Argulian, E., Narula, J., Mihaylova, B., BEHRTDAY: Dynamic Mortality Risk Prediction using Time-Variant COVID-19 Patient Specific Trajectories. AMIA Annual Symposium Proceedings (2022). PMID: 35854750; PMCID: PMC9285184.

# Contents

# List of abbreviations

| | |
|---|---|
| **ACEI** | Angiotensin-converting enzyme inhibitors |
| **AF** | Atrial fibrillation |
| **APC** | Admitted Patient Care |
| **APS** | Average precision score |
| **ARB** | Angiotensin II receptor blockers |
| **AUROC** | Area under the receiver operating characteristics curve |
| **BB** | β blockers |
| **BDL** | Bayesian deep learning |
| **BMI** | Body mass index |
| **BNF** | British national formulary |
| **BP** | Blood pressure |
| **CCB** | Calcium channel blockers |
| **CHD** | Coronary heart disease |
| **CKD** | Chronic kidney disease |
| **CNN** | Convolutional neural network |
| **CPH** | Cox proportional hazards |
| **CPRD** | Clinical Practice Research Datalink |
| **CVD** | Cardiovascular disease |
| **DKL** | Deep kernel learning |
| **DNN** | Deep neural network |
| **EHR** | Electronic health records |
| **ELBO** | Evidence lower bound |
| **FNN** | Feed-forward neural network |
| **GELU** | Gaussian error linear units |
| **GP** | Gaussian process |
| **GRU** | Gated recurrent unit |
| **HDL** | High-density lipoprotein cholesterol |
| **HES** | Hospital Episode Statistics |
| **HF** | Heart failure |
| **ICD-10** | International Classification of Diseases and Health-Related Problems, Tenth Edition |
| **IMD** | Index of multiple deprivation |
| **KISS-GP** | Kernel interpolation for scalable structured Gaussian processes |
| **KL** | Kullback-Leibler |
| **LR** | Logistic regression |
| **LSOA** | Lower Super Output Area |
| **LSTM** | Long short-term memory |
| **MC** | Monte Carlo |
| **MLM** | Masked language model |
| **MLP** | Multilayer perception |
| **NHS** | National Health Service |
| **NICE** | National Institute for Health and Care Excellence |
| **NLP** | Natural language processing |
| **ONS** | Office for National Statistics |

| | |
|---|---|
| **OPCS** | Office of Population Censuses and Surveys Classification of Interventions and Procedures |
| **PCB** | Partial concept bottleneck |
| **PSI** | Population stability index |
| **RC** | Relative contribution |
| **RCT** | Randomised clinical trail |
| **RD** | Risk difference |
| **RELU** | Rectified linear |
| **RF** | Random forest |
| **RNN** | Recurrent neural network |
| **RR** | Risk ratio |
| **SD** | Standard deviation |
| **SGP** | Sparse Gaussian process |
| **STRL** | Straight-through representation learning |
| **Tanh** | Hyperbolic tangent function |
| **t-SNE** | T-distributed stochastic neighbour embedding |
| **UK** | United Kingdom |
| **UTS** | Up to standard |

# 1. Introduction

## 1.1 Motivation

Prognostication is one of the core tasks in medical practitioners' daily work. Current evidence has proved the importance of the risk models in prognostication and their benefit in improving medical prescribing. Still, the unsatisfactory performance of the established clinical models has also raised broad awareness.[1,2] This is in part due to the current risk models being, in general, overly simplistic and insufficient to represent the relationship between the predictors and the outcome.[3] Also, the selection of predictors highly relies on expert knowledge; this can inevitably limit the full potential of using the complete medical information to enable prognostication.[4] Given the central role of prediction in prognostication, together with the growing access to large-scale datasets such as electronic health records (EHR) from millions of individuals, it seems likely that machine learning, especially deep learning, will have transformative effects on medical care. Despite the importance of having accurate risk prediction, the model explainability and the ability for uncertainty quantification are also important.[5,6] These properties are essential to obtain trust and understanding for a model, which is crucial in clinical decision-making. However, the deep learning models are perceived as 'black-box' models with little opportunity for explaining medical phenomena, which hinders their wide application in healthcare.[7] Therefore, this thesis research investigates these difficulties (i.e., risk prediction, explainability, and uncertainty) in prognostication.

## 1.2 Aims of the thesis

This doctoral thesis investigates the applications and clinical benefits of deep learning for prognostic risk prediction, focusing on cardiovascular-related outcomes using large-

scale longitudinally linked EHR that is representative of the United Kingdom (UK) population. It aims to address the following three objectives:

(1) How to accurately predict risks using deep learning and longitudinal multi-modal EHR?

(2) How to retrieve medical knowledge from deep learning risk models to inform risk association analysis and clinical decision-making?

(3) How to deliver reliable risk prediction with the provision of measurement of uncertainty?

Although cardiovascular outcomes are the main interest of this research, the knowledge, methodological approach, and statistical analysis techniques in this thesis can be applied to other medical conditions and disease groups.

### 1.3 Structure of the thesis

To achieve these goals, we will start the thesis with an introduction that covers the critical components of the prognostic risk prediction task (Chapter 2). It will clarify the fundamental concepts and provide the necessary tools for the readers to better engage with the thesis. This part should be easily accessible for both experts and non-experts in the field. Afterwards, the thesis is mainly divided into four parts, and each part will cover one of the following topics: data source (Chapter 3), risk prediction (Chapter 4), explainability (Chapter 5), and uncertainty (Chapter 6).

The data source chapter (Chapter 3) aims to thoroughly introduce a dataset called Clinical Practice Research Datalink (CPRD). It is an anonymised EHR dataset from the UK population that contains information from primary care, secondary care, and death certificates.[8] It is the primary data source for all analyses conducted in the thesis. The risk prediction chapter (Chapter 4) will discuss a deep-learning risk model named Bidirectional Transformer for EHR (referred to as BEHRT), including its application in risk prediction (§4.1), its performance in the presence of data shifts (§4.2) and with the inclusion of more modalities (§4.3), and the benefit of applying BEHRT for survival analysis compared to the

established clinical models (§4.4). This chapter intends to establish evidence and illustrate the great potential of deep learning for EHR in risk prediction and healthcare. With an accurate deep learning risk model, we assume the model has extracted some interesting patterns. However, these patterns are not easily explainable due to the large size of deep learning models, which also raises concerns about trusting the model and hinders its broader application in practice. Therefore, it is a natural move for the next step to investigate methods that can explain the model's underlying mechanism for risk prediction and retrieve medical knowledge from the model to assist clinical research and decision-making. This will be discussed in Chapter 5. So far, we mainly focus on investigating the deterministic model, which provides point estimates for the risk of the outcomes without any uncertainty quantification. However, uncertainty can be critical in healthcare decision-making. Considering the severe impact of the consequence, Chapter 6 will discuss methods that can transform BEHRT for probabilistic modelling and the applications of uncertainty in aiding decision-making. In the end, A synthesis of results, overall implications for clinical practice, and potential future works will be discussed in Chapter 7.

# 2. Background

## 2.1 Prognostic risk models in healthcare

Prognostication is, in essence, the prediction of the course of a disease or future outcomes based on the information available at present, and the prognostic risk model is the tool for conducting the prediction. It is essential for care planning and disease prevention.[2] For example, the cardiovascular disease (CVD) risk score, more specifically, QRISK2[2], is used as the risk assessment tool for CVD prevention for the general population, followed by the National Institute for Health and Care Excellence (NICE) guideline[9]. It recommends that patients with a 10% or greater 10-year CVD risk be treated with atorvastatin which has been proved to be an effective approach for prevention of CVD outcomes.

### 2.1.1 Definition of risk prediction

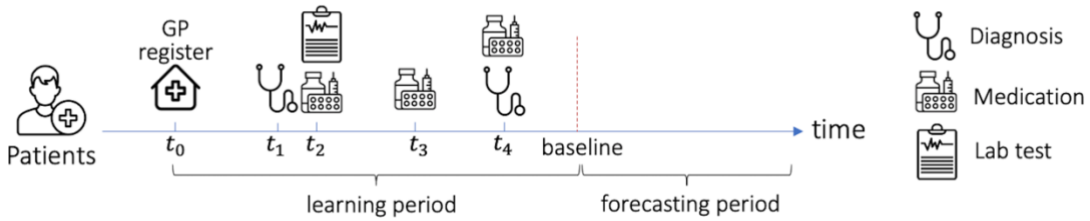



*Figure 2.1: Visual illustration for risk prediction. The horizontal axis represents time, and GP means general practices. A patient starts to generate medical information once they are registered with GP. A patient can visit GP or hospital irregularly whenever needed, represented by the irregular interval between* $t_i$ $(i = 1, 2, 3, 4)$. Additionally, a patient can have multiple events within the same visit, and we use diagnoses, medications, and lab tests as examples for visualisation.

Let us first define the risk prediction task to formalise our discussion on prognostic risk prediction. For the rest of the thesis, we will interchangeably use prognostic risk prediction and risk prediction. As shown in Figure 2.1, we use a hypothetical patient to illustrate several essential components in risk prediction: baseline, learning period, and forecasting period. The baseline is the time point for risk prediction. It uses all available information about a patient before that time point (i.e., learning period) to predict the future

outcomes after it (i.e., forecasting period). Therefore, the risk prediction task can be broadly grouped into classification and survival analysis.

### 2.1.2 Classification

For a given cohort with *N* patients, the input-output pair in the classification task can be represented as $\mathcal{D} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$, where $x_i \in \mathbb{R}^D$ represents the covariates collected from the learning period and $y_i \in \mathbb{R}^Q$ describes the outcomes from the forecasting period, which is usually binary or categorical. $D$ and $Q$ represent the input and output dimensions, respectively. For example, CVD risk prediction would have a binary outcome: CVD positive or CVD negative. The objective of binary classification is to find a transformation $f^{\boldsymbol{\omega}}$ that can map $x_i$ to $y_i$: $p(y_i) = f^{\boldsymbol{\omega}}(x_i)$, different parameters $\boldsymbol{\omega}$ represent different transformations, and the transformation minimises the cross-entropy error: $\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log f^{\boldsymbol{\omega}}(x_i) + (1 - y_i) \cdot \log(1 - f^{\boldsymbol{\omega}}(x_i))$.

### 2.1.3 Survival analysis

To introduce the survival analysis, let us start with the components quite different from the classification task. Survival analysis is designed to analyse the expected duration of time (i.e., survival time) until the event of interest occurs. Therefore, the prediction is time-dependent. As shown in Figure 2.2, instead of having binary or categorical outcomes for the classification task, the survival analysis's outcome includes survival time and event. The event is binary and similar to the classification task, representing if an event of interest is observed for a patient within the forecasting period. The survival time means the time to event, either the duration to the event of interest or until a patient is censored. Censoring may arise in the following situations: (1) a patient has not (yet) experienced the event of interest by the end of the forecasting period, or (2) a patient is lost to follow-up or experiences a different event that makes further follow-up impossible (e.g., death) during the forecasting period.[10] Conceptually, the survival prediction is formatted by $\mathcal{D} =$

$\{(x_1, t_1, y_1), \ldots, (x_N, t_N, y_N)\}$, where $x_i \in \mathbb{R}^D$ represents the covariates collected from the learning period, $t_i \in \mathbb{R}^1$ and $y_i \in \mathbb{R}^1$ represent the survival time and event, respectively. The objective is to find a hazard function $f^{\boldsymbol{\omega}}$ determined by the covariates: $p(y_i|t) = f^{\boldsymbol{\omega}}(x_i, t)$, where $\boldsymbol{\omega}$ represents the function parameters. The hazard function is the probability of developing an event at a different survival time for an individual.

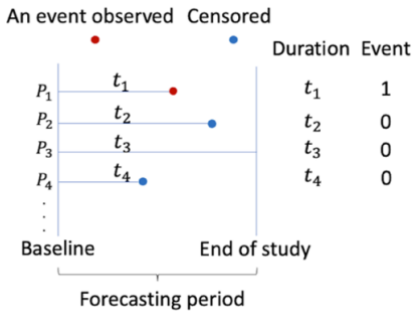



*Figure 2.2: Illustration for survival analysis. P represents patients, t represents the survival time, the red dot indicates that an event of interest is observed, and the blue dots represent a patient lost to follow-up (i.e., censored) before the end of the study. For example, an event of interest is observed at the time $t_1$ for the patient $P_1$. Therefore, the duration and event are $t_1$ and positive (i.e., 1), respectively. Patient $P_2$ drops out of the study at $t_2$ and the event of interest has not been observed until that time; thus, patient $P_2$ is still negative (i.e., 0) until time $t_2$ (i.e., duration).*

#### 2.1.3.1 Cox proportional hazards (CPH) regression model

CPH is a frequently used survival model in healthcare. It investigates the relationship between the time to event outcome $y$ and the input covariates $x$. CPH model assumes that the hazard rate of a person with one set of covariates is proportional to the hazard rate of a person with a different set of covariates.[11] Therefore, the difference between the two hazards can be explained by a multiplicative constant. CPH can be defined as:

$$p(y_i|t) = f^{\boldsymbol{\omega}}(x_i, t) = h(t, x_i) = h_0(t) \cdot e^{\lambda(x_i)}, \qquad (2.1.1)$$

where $h_0(t)$ is the baseline hazard function and $\lambda(x_i)$is the log-risk function (i.e., a function of a set of covariates). For the classic CPH model, the effect of covariates on the hazard (i.e., log-risk function) is modelled as a linear function. Note that, only the log-risk function is parametrised, and the baseline hazard function is derived from the observations (without

covariates). Thus, CPH is often called the "semi-parametrised" model. The objective of CPH is to find a function $\lambda$ that maximises the partial likelihood[12]:

$$\prod_{i=1}^{N}\left(\frac{e^{\lambda(x_i)}}{\sum_{t_j \le t_i} e^{\lambda(x_j)}}\right)^{y_i}.$$

Additionally, the baseline hazard function $h_0(t)$ is estimated using the Breslow estimator[12]. Given cumulative baseline hazard function $\Lambda_0(t) = \int_0^t h_0(u)du$, Breslow suggested to estimate $\lambda$ and $\Lambda_0(t)$ in the same maximum likelihood framework and treat $h_0$ as a piecewise constant between uncensored event time. This led to an estimated cumulative baseline hazard function:

$$\Lambda_0(t) = \sum_{i=1}^{N} \frac{I(t_i,t)y_i}{\sum_{t_j \le t_i} e^{\lambda(x_j)}},$$

where $I(t_a, t_b)$ is an indicator function to represent if $t_a \le t_b$ (with 0 and 1 indicating false and true, respectively).

## 2.2 Electronic health records

EHR refer to a systematised collection of medical information about individual and populations. It is typically captured during a health care encounter in a digital format.[13] EHR may contain various types of data, both structured and unstructured, such as demographics, medical history, medication, lab test results, and free-text clinical notes.[8] The records are also associated with detailed information such as dates, dosages, and measured clinical values to store data accurately and capture a patient's state across time.[14]

For the last decades, many large-scale EHR have been proved to be a valuable resource for clinical research as they provide longitudinal health data that are representative of the population in terms of age, gender, ethnicity, and socioeconomic status on a regional or even national-wide scale at low-cost.[8,15,16] Some of the main applications of EHR have included the re-evaluation of prior findings, testing associations in subpopulations, and studying rare outcomes using a large sample size; enhancing social epidemiology using patients

distributed across a wide range of physical and social environments; research on stigmatised conditions; and predictive modelling.[15]

As risk prediction is the main interest of this thesis, we will narrow the scope and discuss the EHR for predictive modelling. Historically, data from epidemiologic cohorts or clinical trials have been the primary resource for risk modelling.[17] Although these data sources are often high quality, the cons sometimes can outweigh the pros. First, the increasing challenge in recruitment and the burdensome data collection process has raised wide concerns; this constantly adds to the growing cost of clinical research and compromises the investment in the field.[17,18] Also, the population included in the cohort or trial is usually narrowly selected and not representative of the general population. Thus, the generalisability of the risk scores is overall unsatisfactory.[17,19] Additionally, data collected in the trials and cohort studies may not match the data collected in the clinical settings, which can further increase the discrepancy between modelling and clinical application.

The rise of comprehensive large-scale EHR offers unprecedented opportunities to tackle the abovementioned difficulties. Major advantages of using EHR data include a large sample size, long-term follow-up, and rich medical information.[8] As EHR have already been collected in routine clinical encounters, they are more cost, time, and resource efficient and involve larger and more representative populations than the trials and cohort studies.[17,20] Additionally, EHR that are available for research purposes usually have already registered and followed patients for many years across various health services (e.g., primary and secondary care).[14] They contain fine-grained longitudinal patient health profiles and are representative of the data collected from the clinical practices.[17,20,21] Although the data quality can vary across different data sources, we believe the practitioner-recorded data overall are more accurate than the self-reported data.

Despite the many benefits, EHR also have limitations. First, EHR are collected from clinical encounters; therefore, the records do not necessarily reflect the actual time an event happens but highly depend on a patient's health status and when and how a patient seeks care. Additionally, data consistency and quality can vary across different times within and between practitioners as it is the physician and patient, not the researcher, who stipulate the amount of time under observation and generate the data.[22] Due to the same reason, another common challenge in EHR is data missingness. Because the physician usually only records medical information related to patients' current health status, not all information is available for various research purposes.

## 2.3 Deep learning

To introduce deep learning, we shall start from the most straightforward statistic tool: linear regression[23–25]. For a given dataset $\mathcal{D} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$, where $x_i \in \mathbb{R}^D$ and $y_i \in \mathbb{R}^1$ represent the input and output, respectively. The linear regression is to find a linear transformation $f^{\boldsymbol{\omega}}$ that maps $\boldsymbol{X}$ to $\boldsymbol{Y}$: $\boldsymbol{Y} = f^{\boldsymbol{\omega}}(\boldsymbol{X}) = \boldsymbol{XW} + b$, with $\boldsymbol{W} \in \mathbb{R}^{D\times 1}$ and $b \in \mathbb{R}^1$. Different $\boldsymbol{W}$ and $b$ represent different transformations. The objective is to find a transformation that minimises the mean squared error over the observed data: $\frac{1}{N}\sum_{i=1}^{N}(y_i - f^{\boldsymbol{\omega}}(x_i))^2$.

One of the fundamental assumptions of linear regression is that $\boldsymbol{Y}$ and $\boldsymbol{X}$ have a linear relationship.[25] This is a stringent requirement as, in most cases in practice, the relationship between $\boldsymbol{Y}$ and $\boldsymbol{X}$ can be non-linear. We resort this to the linear basis function models[26], generalisations of linear regression that regress the output on functions of the predictors. The relationship can be represented as $\boldsymbol{Y} = f^{\boldsymbol{\omega}}(\boldsymbol{X}) = \boldsymbol{\Phi}(\boldsymbol{X})\boldsymbol{W} + \boldsymbol{b}$, where $\boldsymbol{\Phi}(\boldsymbol{X}) = [\phi_1(x^1), \ldots, \phi_D(x^D)]$, and the basis functions $\boldsymbol{\phi}$ are chosen to suitably model the non-linear relationship between the inputs and the output. In deep learning, the most fundamental models are described as a hierarchy of the basis functions, and the hierarchy is referred to

as the neural network, and each basis function is referred to as a neuron. This process has been described as replicating the brain function because it is believed that the brain contains many levels of processing, and each level learns features and representations at an increasing level of abstraction.[24] For instance, some evidence suggests that the visual cortex of the brain extracts edges first, then patches, then surfaces, the objects, and the neural network can replicate such information extraction.[24] The features in the intermediate layer of the neural network are called latent features, and they are usually learned by backpropagation using parametrised basis functions.[27]

To further abstract the neural network, we can consider the basis functions in each layer of the neural network as a "building block", and the investigation of the building blocks in a neural network has been a rapidly growing area in the last decades.[28–31] The modularity and the composition of building blocks ensure the versatility and flexibility of deep learning models. This has led to many successful applications in various fields. Here, we will review some of the most popular building blocks in deep learning.

### 2.3.1 Feed-forward neural networks (FNNs)

FNNs are also called multilayer perceptions (MLPs). We will focus on an FNN with a single hidden layer to simplify the description. This can be further extended to multiple hidden layers to represent general and more expressive FNNs. A feed-forward layer in an FNN includes a weight matrix $\boldsymbol{W}$, a bias term $\boldsymbol{b}$ and an activation function $\sigma$. For the abovementioned data $\mathcal{D}$ with $\boldsymbol{X} \in \mathbb{R}^{N\times D}$ and $\boldsymbol{Y} \in \mathbb{R}^{N\times 1}$, we use $\boldsymbol{W_1} \in \mathbb{R}^{D\times D_1}$ and $\boldsymbol{b_1} \in \mathbb{R}^{D_1}$ to represent the weight matrix and bias for the hidden layer, the $\sigma$ is usually an element-wise non-linear transformation function (e.g., rectified linear (RELU), sigmoid, and hyperbolic tangent function (Tanh)). It transforms the input $\boldsymbol{X}$ to obtain hidden vector $\boldsymbol{\Phi}(\boldsymbol{X}) = \sigma(\boldsymbol{X}\boldsymbol{W_1} + \boldsymbol{b_1})$, which is further followed by an output layer to map the hidden

vector to the output with weight matrix $\boldsymbol{W_o} \in \mathbb{R}^{D_1 \times 1}$ and bias $b_o \in \mathbb{R}^1$. The relationship between input and output can be represented as $\widehat{\boldsymbol{Y}} = \boldsymbol{\Phi}(\boldsymbol{X})\boldsymbol{W_o} + b_o$.

### 2.3.2 Convolutional neural networks (CNNs)

CNN is inspired by the organisation of the animal visual cortex and is designed to process data with a grid pattern, such as images.[32] It is typically composed of three functions: convolution, pooling, and feed-forward layers, and they are recursively applied to learn spatial hierarchies of features, from low-level to high-level patterns.[33] More specifically, the convolution layer computes the dot product between the filters (also referred to as kernel) and the image patches (depicted in Figure 2.3), where each filter is a set of learnable parameters. The pooling layer is applied to reduce the size of the volume to reduce memory and prevent overfitting. For example, max pooling uses a filter to take the maximum of the patches.[34] Because the filters are shared across the image patches, CNNs can substantially reduce the number of weights that must be learned, and they are invariant to the spatial changes of objects in the image.[33]

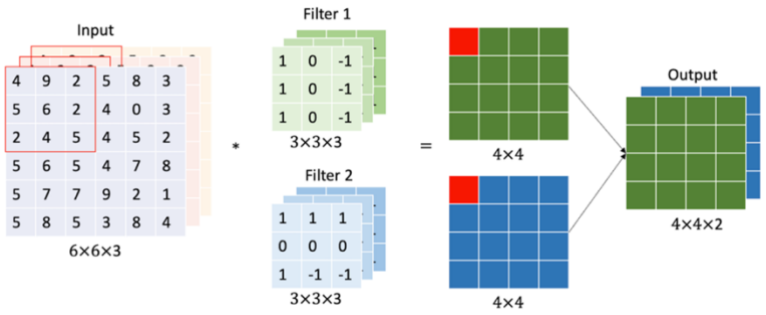



*Figure 2.3: Convolution operation in a CNN. The given input has a height of 6, width of 6, and 3 channels. There are two filters (with kernel size 3 × 3), and each is convoluted with the image patches (the red square in the input represents a single image patch). The red square after the filter is the convolution output for the marked patch in the input.*

### 2.3.3 Recurrent neural networks (RNNs)

RNNs are widely adopted for handling sequential data, such as text, audio, and video.[27] In general, the inputs of RNNs are a sequence of symbols; each symbol represents a time step, and we describe data with N observations as $\mathcal{D} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$, where

$x_i = \{x_{i,1}, \ldots, x_{i,T_i}\}$, $x_{i,t} \in \mathbb{R}^D$, and $T_i$ represents the maximum sequence length $L$ for observation $i$. For a given observation $i$, RNNs are applied to a single symbol $x_{i,t}$ at each time step $t$ and the output of an RNN at $t$ is represented as a hidden state $\boldsymbol{h_{i,t}}$ (as shown in Figure 2.4). It is a combination of the input $x_{i,t}$ and the information encoded from the previous time steps $\boldsymbol{h_{i,t-1}}$, and can be described as $\boldsymbol{h_{i,t}} = f^{\boldsymbol{\omega}}(x_{i,t}, \boldsymbol{h_{i,t-1}})$. For a simple RNN, the hidden state at time step $t$ is generated as:

$$\boldsymbol{h_{i,t}} = \sigma(\boldsymbol{x_{i,t}W_h} + \boldsymbol{h_{t-1}U_h} + \boldsymbol{b_h}),$$

where both $\boldsymbol{W_h} \in \mathbb{R}^{D \times D_1}$, and $\boldsymbol{U_h} \in \mathbb{R}^{D_1 \times D_1}$ are weight matrices, $\boldsymbol{b_h}$ is the bias term, and $\sigma$ represents the non-linear activation function. Afterwards, a linear transformation can be applied to map the hidden state of the last time step to the output as:

$$\widehat{\boldsymbol{y_i}} = \boldsymbol{h_{i,T}W_o} + b_o,$$

where $\boldsymbol{W_o} \in \mathbb{R}^{D_1 \times 1}$ and $b_o \in \mathbb{R}^1$.

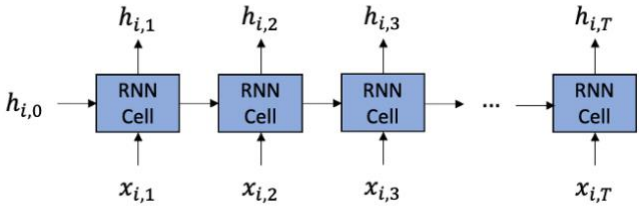



*Figure 2.4: Visual illustration for RNN. x represents the input at different time steps, and h represents the hidden state at each time step.*

Because simple RNNs face gradient vanishing and exploding issues, especially when the sequence gets longer, more sophisticated RNN models such as long short-term memory (LSTM) network and gated recurrent unit (GRU) are more preferred in practical applications.[35] For example, instead of consistently generating the hidden state using information from both the current time step and the previous time steps, LSTM uses the "input" gate, "forget" gate, and "output" gate to control the information flow.[31] If the input from the current time step $x_{i,t}$ is not informative, LSTM can potentially turn off the "input" gate and only generates the hidden state $\boldsymbol{h_{i,t}}$ using information from $\boldsymbol{h_{i,t-1}}$. Similarly, if the

information from the previous time steps $\boldsymbol{h_{i,t-1}}$ is not informative, LSTM can also turn off the "forget" gate to generate the hidden state $\boldsymbol{h_{i,t}}$ using only the input $x_{i,t}$. Otherwise, both $x_{i,t}$ and $\boldsymbol{h_{i,t-1}}$ can contribute to the generation of $\boldsymbol{h_{i,t}}$. This gating mechanism in LSTM has been proved to be an effective method for learning valuable features and alleviating the gradient vanishing and exploding. The detailed equations for LSTM are defined as follows:

$$f_t = \sigma_g\left(\boldsymbol{x_{i,t}W_f} + \boldsymbol{h_{t-1}U_f} + b_f\right), \qquad \widetilde{\boldsymbol{c_t}} = \sigma_c\left(\boldsymbol{x_{i,t}W_c} + \boldsymbol{h_{t-1}U_c} + b_c\right),$$
$$in_t = \sigma_g\left(\boldsymbol{x_{i,t}W_{in}} + \boldsymbol{h_{t-1}U_{in}} + b_{in}\right), \qquad \boldsymbol{c_t} = f_t \circ \boldsymbol{c_{t-1}} + in_t \circ \widetilde{\boldsymbol{c_t}},$$
$$out_t = \sigma_g\left(\boldsymbol{x_{i,t}W_{out}} + \boldsymbol{h_{t-1}U_{out}} + b_{out}\right), \qquad \boldsymbol{h_{i,t}} = out_t \circ \sigma_h(\boldsymbol{c_t}),$$

where $\boldsymbol{\omega}_W = \{\boldsymbol{W_f}, \boldsymbol{W_{in}}, \boldsymbol{W_{out}}, \boldsymbol{W_c}\} \in \mathbb{R}^{D \times D_1}$ , $\boldsymbol{\omega}_U = \{\boldsymbol{U_f}, \boldsymbol{U_{in}}, \boldsymbol{U_{out}}, \boldsymbol{U_c}\} \in \mathbb{R}^{D_1 \times D_1}$ are weight matrices for the forget gate, input gate, output gate, and cell state, respectively; $\boldsymbol{\omega}_b = \{\boldsymbol{b_f}, \boldsymbol{b_{in}}, \boldsymbol{b_{out}}, \boldsymbol{b_c}\} \in \mathbb{R}^{D_1}$ are bias terms for the transformations; $\sigma_g$, $\sigma_c$, and $\sigma_h$ are sigmoid, tanh, and tanh functions, respectively. GRU follows the same philosophy as the LSTM (i.e., gate mechanism), but with fewer parameters.[36]

### 2.3.4 Transformer

RNNs are firmly established approaches for sequential data modelling. However, one of the major disadvantages is that the RNNs must generate a sequence of hidden states $\boldsymbol{h_t}$ as a function of the previous state $\boldsymbol{h_{t-1}}$ for time step $\boldsymbol{t}$. This inherence nature of RNNs prevents the parallelisation within observation (across different timesteps) and becomes a critical bottleneck when the sequence length gets longer.[37] To resolve this limitation, Vaswani et al.[37] proposed a landmark model, Transformer, for sequence modelling. It only relies on stacked self-attention and pointwise FFNs to draw global dependencies across the sequence. This mechanism allows for in-sequence parallelisation and achieves state-of-the-art performance on multiple tasks in natural language processing (NLP).[38]

To introduce Transformer, I shall start with the attention mechanism. In psychology, attention is a cognitive mechanism in that our mind selectively concentrates on one or a few

things and ignores the others.[39] Since neural networks are inspired by how the brain operates, researchers attempt to implement similar properties in the neural network to enable them to concentrate on the relevant information and ignore the rest. For example, given a photo with a group of kids in the garden, if anyone asks, "how many kids are there?", we might immediately solve this question by counting the heads. Thus, we naturally concentrate on the heads and ignore the background and other information. The attention mechanism was first proposed to model the long-term dependencies for the RNN-based encoder-decoder model for machine translation.[40] Because of its outstanding performance, this mechanism raised wide popularity and has been applied to various tasks and models.[41–43]

The attention function generally includes four parts, query, key and value pair, and output, each of which is a vector.[37] For the convenience of description, we represent them as $\boldsymbol{Q}$, $\boldsymbol{K}$, $\boldsymbol{V}$, and $\boldsymbol{O}$, respectively, in this section. Mathematically, a model's attention calculates the similarity between the representation of the $\boldsymbol{Q}$ and the contexts $\boldsymbol{K}$. For example, the $\boldsymbol{Q}$ can be a vector representing the question in the head counting example abovementioned, and the $\boldsymbol{K}$ can be the vectors representing different patches in the photo. We expect that the $\boldsymbol{Q}$ will have a higher similarity (i.e., attention score) with the patches that include the heads. Then, the information from all patches $\boldsymbol{V}$ can be summarised based on the attention score to deliver the output $\boldsymbol{O}$ (the output mainly consists of the information from the patches with heads in this example). Similarly, for sequential data, $\boldsymbol{Q}$ could be the representation of a time step, and $\boldsymbol{K}$ could be the representation of the contexts. For a given sequence data with context length T, $\boldsymbol{Q} \in \mathbb{R}^{1\times D_q}$, $\boldsymbol{K} \in \mathbb{R}^{T\times D_q}$, $\boldsymbol{V} \in \mathbb{R}^{T\times D_v}$, the attention score and the output can be calculated as below:

$$\boldsymbol{S} = softmax(\boldsymbol{Q}\boldsymbol{K}^T),$$

$$\boldsymbol{O} = \boldsymbol{S}\boldsymbol{V},$$

where $\boldsymbol{S} \in \mathbb{R}^{1\times T}$ represents the attention scores.[40] When both query and key-value pair come from the same source (e.g., same sequence or same image), we call it intra-attention or self-attention. Otherwise, it is called inter-attention[44].

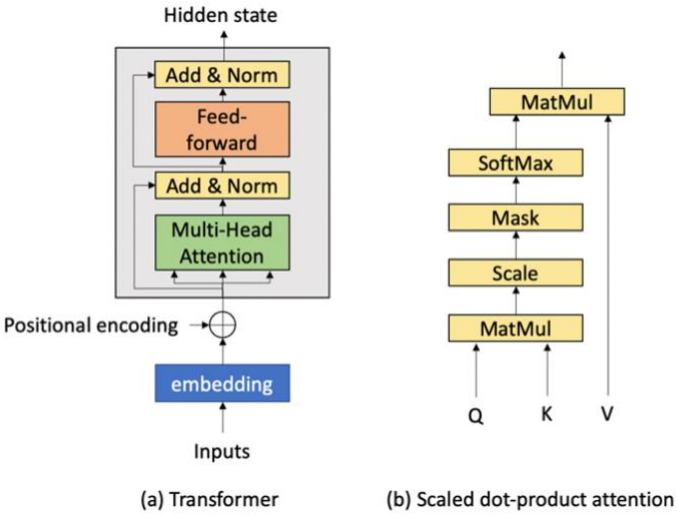



*Figure 2.5: Transformer model architecture and scaled dot-product attention. (a) Transformer architecture mainly includes a multi-head attention layer, a feed-forward layer, and two residual connections. (b) Scaled dot-product attention computes the dot products of the query with all keys and applies a softmax function to obtain the weights on the values.*

The Transformer is a model that relies on the self-attention mechanism to capture global dependencies. As shown in Figure 2.5, Transformer includes a multi-head attention layer, a feed-forward layer, and two residual connections. The multi-head attention is a breakthrough and is the most crucial component of Transformer[37]. It concatenates the representations from multiple scaled dot-product attention heads. This mechanism has been proved to benefit modelling by allowing the model to attend to information at different positions jointly. For a single scaled dot-product attention head, the representations are calculated as:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{D_Q}})V,$$

where $Q \in \mathbb{R}^{T\times D_Q}$ is the query, $K \in \mathbb{R}^{T\times D_Q}$ is the key, and $V \in \mathbb{R}^{T\times D_V}$ is the value; $D_Q$ and $D_V$ are the key and value dimensions, respectively, and $T$ represents the maximum sequence length. Because the queries, keys, and values are all derived from the same source in

Transformer with different linear transformations. This mechanism is also called multi-head self-attention, and it is the primary strategy for learning the long-range dependencies with a sequence in Transformer.[37]

Additionally, unlike RNNs, which preserve the sequence order with recurrence, Transformers do not contain any order information.[45] To make use of the order of a sequence, Transformers inject the order information into the inputs with the position embeddings, which can be learned while model training or manually coded. The manual positional encoding is described as below:

$$PE_{(pos,2i)} = \sin\left(pos/10000^{2i/D}\right),$$

$$PE_{(pos,2i+1)} = \cos\left(pos/10000^{2i/D}\right),$$

where $pos$ is the position, $i$ is the dimension, and $D$ is the dimension of the embeddings.

#### 2.3.4.1 Bidirectional Encoder Representations from Transformers (BERT)

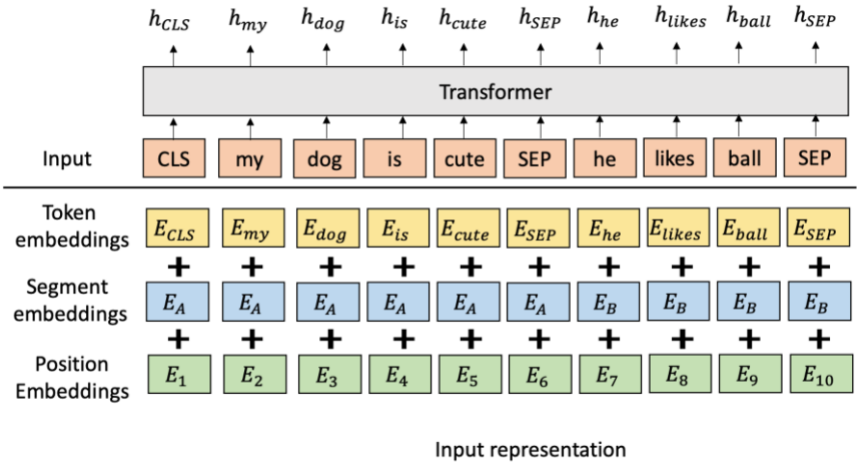



*Figure 2.6: Input representation for Transformer. In the NLP task, the inputs of the transformations are the summation of three representations: token embeddings, segment embeddings, and position embeddings. The representations of all time steps are fed into the Transformer in parallel to produce the hidden outputs. Segment embeddings are two symbols (A and B) that alternatively change between consecutive sentences.*

BERT[38] is a Transformer-based model with several critical improvements for sequential modelling. First, BERT more comprehensively defines the model inputs. As shown in Figure 2.6, for a given token, the input representation of BERT is constructed as

the summation of the token, segment, and position embeddings; the first token of a sequence is always a special token "CLS", and the final hidden state of this special token (from an additional single FFN "pooling" layer) is considered as the aggregate sequence representation for classification tasks. Additionally, BERT packs multiple sentences as a single sequence for modelling, and it differentiates the sentences in two ways: (1) separating sentences with a special token "SEP" and (2) adding a learned segment embedding to each token indicating whether a token belongs to sentence A or sentence B.

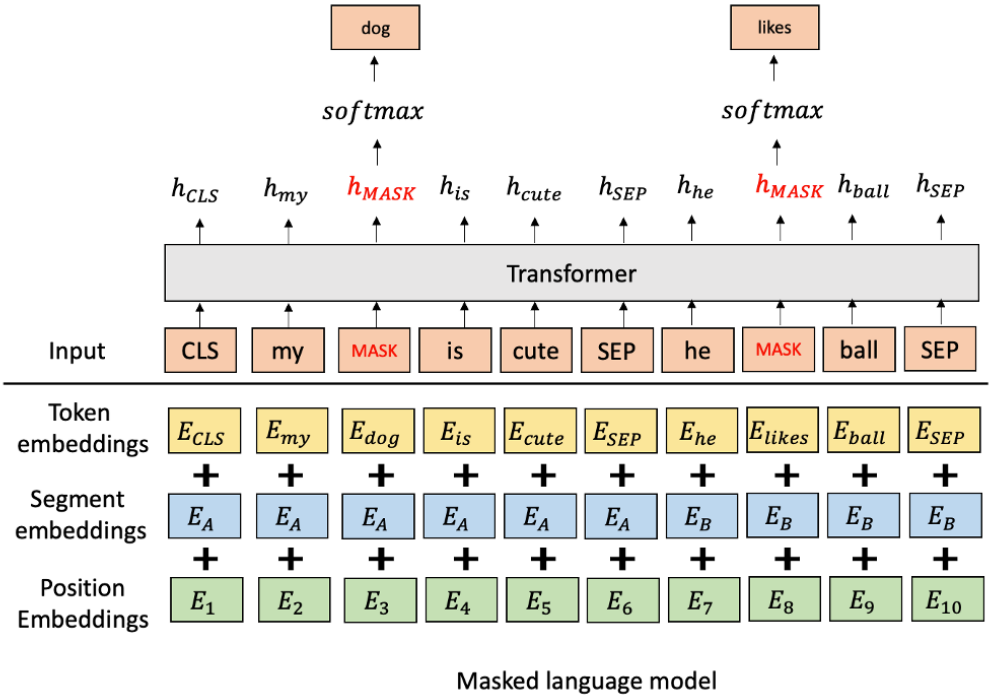



*Figure 2.7: Architecture for MLM. Some inputs are randomly masked by the "MASK" token, and the objective is to learn contextual representations that can recover the masked tokens.*

Another important improvement is proposing a bidirectional representation pretraining method, named masked language model (MLM). Intuitively, it is reasonable to believe the contexts on both sides (left and right) of a token are informative and indirectly allow the word to "see itself". The presentation of a token that is jointly conditioned on both left and right context is called bidirectional representation. To this end, MLM proposes to construct a reconstruction problem, which corrupts the inputs by randomly masking some percentage of the tokens using "MASK" in the sequence and then uses the uncorrupted

context to learn meaningful representations that can recover the corrupted information in the original sequence (as shown in Figure 2.7). Additionally, because the token "MASK" only exists in MLM but not the downstream task (e.g., classification), instead of just masking a token, BERT proposes to conduct the token corruption by either (1) masking the token, or (2) replacing with a random token, or (3) keeping the token unchanged.

## 2.4 Model explainability

A key property of neural networks that helped their success is their depth (i.e., many layers and parameters).[46] This means that neural networks are composed of a large amount of non-linear functions, which ensures their versatility and flexibility in learning complex relations between the inputs and the outputs.[47] However, this also comes at the cost of explainability. Each layer of the neural network can increase the abstraction of features, hence, failing to provide simple human-understandable explanations for a model's underlying mechanism. This raises concerns about trusting them in healthcare applications and fails to provide meaningful medical knowledge to inform medical research and clinical decision-making.[48,49]

Recently, an increasing number of works has aimed to shed light on neural networks with different methods for providing explanations.[50–54] However, in general, we can classify such explanations under the three-level "causal ladder" proposed by Pearl and Mackenzie: association, intervention, and counterfactual insights.[55] The association is at the lowest level of the causal ladder; it represents a conditional probability of observing output variable $y$ given input variable $x$: $P(y|x)$. Some of the representative methods for conducting explanations at the association level are LIME and SHAP.[6,50,56] Both are instance-wise explainers, and the associations are measured by post-hoc perturbation. For example, LIME explains a prediction for a fixed model by learning an interpretable model (e.g., linear regression) using pseudo-observations generated by perturbating the selected instance; then,

the instance-wise association is aggregated across all samples in the dataset to measure the data-level association.[50]

The second level of the causal ladder is intervention, which introduces an additional level of complexity. Instead of just seeing, it can answer the "what if" question by investigating the change in the outcome resulting from an exogenous intervention. This is referred to as the $do$ action: $P(y|do(x))$. However, to provide the ladder two explanation with the ladder one model, the intervention must be conducted when the data of the key surrogates are available and the key surrogates are carefully controlled.[55,57] For instance, a change in the price of an item can lead to a change in customer behaviour, which nullifies the historical sales data for modelling future customer behaviour. Recently, some concept-based methods have aimed to provide a second-level explanation.[58–60] They work on user-defined high-level concepts and allow users to interact with the model (i.e., $do$ action) under carefully controlled conditions. Note that, without intervention, the relation between the output and input variable is equal to the relationship described at the association level. Therefore, the intervention level subsumes the association level in the ladder.

While the second level of the ladder allows us to investigate the outcome by changing the environment (i.e., $do$ action), the third level allows us to go back in time and change the history to ask "what would have happened if I had done things differently?" [55] That is, instead of the "factual" observations, we are interested in an alternative reality or counterfactual scenario, which was not actually observed. This would involve three components: the observed history (referred to as variable $z$), the exogenous intervention $x$ given history $z$, and the output variable $y$.[57] The relation can be represented as $P(y|do(x), z)$. This level highlights that the hypothetical intervention $x$ cannot conflict with the observation $z$ because the alternative reality is just a hypothetical. We require a model with given observation $z$ to connect reality to the hypothetical world.[61,62]

## 2.5 Uncertainty in decision making

Assume that we want to confirm a conclusion from an experiment. One natural move we would take is to repeat the experiment multiple times to ensure the first experiment's observation is not by chance. However, we rarely get the same results when an experiment is reproduced. This can be caused by various reasons, such as the variability of the experimental process or the precision limits of the measurement.[63] Because of the difference in results from each experiment, the more experiments we conduct, the better estimation we have for the uncertainty of the conclusion. We call the uncertainty of the conclusion predictive uncertainty, a combination of epistemic and aleatoric uncertainty.[5] Epistemic uncertainty is caused by a lack of knowledge. In our case, it represents the experimental design, and a better experimental design can reduce epistemic uncertainty. Aleatoric uncertainty represents the noise inherent in the data and the experiment, such as the imprecision of measurement due to the limits of the equipment.[64]

Similarly, the modelling process, including neural networks, follows the same philosophy to conclude a prediction. It predicts the outputs $y$ for an input $x$ using a model $f^{\boldsymbol{\omega}}$. The epistemic uncertainty is caused by the uncertainty of the model structure and the uncertainty of the model parameters. More observations can potentially improve the model design and benefit the model parameter training, therefore, reducing epistemic uncertainty. However, the aleatoric uncertainty comes from the noises in the observed data. Henceforth, it is irreducible.[5]

Because uncertainty's essential role in quantifying the credibility of decisions, it has raised great attention in healthcare.[5,65,66] Recent studies in probabilistic modelling, especially Bayesian modelling and Gaussian processes (GPs), provide a powerful framework to model the uncertainty in neural networks rigorously.[67,68] To formalise our

discussion, let us review the main ideas about Bayesian modelling, approximate interference, and GPs.

### 2.5.1 Bayesian modelling

Let us assume that we have a dataset $\mathcal{D} = \{(x_1, y_1), \dots, (x_N, y_N)\}$ with $\boldsymbol{X} = \{x_1, \dots, x_N\}$ and $\boldsymbol{Y} = \{y_1, \dots, y_N\}$ represent the observed inputs and outputs, respectively. The objective is to find a model with parameters $\boldsymbol{\omega}$ that is likely to generate the outputs $\boldsymbol{Y}$ for inputs $\boldsymbol{X}$: $y = f^{\boldsymbol{\omega}}(x) = f(x, \boldsymbol{\omega})$. Following the Bayesian theorem[24], we would need a prior belief on the distribution of the model parameters. It indicates the prior belief on the parameters $\boldsymbol{\omega}$ that are likely to generate $\boldsymbol{Y}$ for $\boldsymbol{X}$ without any observations and can be represented as $p(\boldsymbol{\omega})$. Once some of the data is observed, we can define a likelihood distribution $p(y|x, \boldsymbol{\omega})$ to assess how likely the input $\boldsymbol{X}$ can generate $\boldsymbol{Y}$ for a set of given parameters $\boldsymbol{\omega}$. For the binary classification task in the prognostic risk prediction, we may use the cross-entropy[69] as a measurement for the likelihood. Afterwards, the posterior distribution can be calculated as

$$p(\boldsymbol{\omega}|\mathbf{X}, \mathbf{Y}) = \frac{p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\omega})p(\boldsymbol{\omega})}{p(\boldsymbol{Y}|\boldsymbol{X})}, \tag{2.5.1}$$

which represents the updated belief on the model parameters considering both the prior belief and the observed data. With such belief, we can conduct any inference for an unobserved input $x^*$ by integrating over the model parameters:

$$p(y^*|x^*, \boldsymbol{X}, \boldsymbol{Y}) = \int p(y^*|x^*, \boldsymbol{\omega})\, p(\boldsymbol{\omega}|\mathbf{X}, \mathbf{Y}) d\boldsymbol{\omega}, \tag{2.5.2}$$

where $p(y^*|x^*, \boldsymbol{X}, \boldsymbol{Y})$ is defined as predictive distribution.

As shown in eq. (2.5.1), we have discussed the prior distribution $p(\boldsymbol{\omega})$ and the likelihood distribution $p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\omega})$ in the evaluation of the posterior distribution, another important component in the equation is $p(\boldsymbol{Y}|\boldsymbol{X})$, which is called model evidence. It represents the integration of likelihood over the $\boldsymbol{\omega}$ (i.e., marginal likelihood):

$$p(\boldsymbol{Y}|\boldsymbol{X}) = \int p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\omega}\,)\, p(\boldsymbol{\omega}) d\boldsymbol{\omega}.$$

For simpler models such as the Bayesian linear regression, we can specify convenient parametric forms for the prior distributions and the likelihood to retain the prior's parametric form through to the posterior (i.e., prior conjugate).[24,70] Therefore, the integration can be solved with known analytic tools in calculus. However, the integration is intractable for more complicated models such as neural networks and cannot be solved with analytical tools. Thus, other strategies such as variational inference are required to approximate the posterior distribution.

### 2.5.2 Variational inference

Because the posterior distribution $p(\boldsymbol{\omega}|\mathbf{X}, \mathbf{Y})$ is usually intractable in the neural networks. One potential solution is to define a variational distribution $q_\theta(\boldsymbol{\omega})$, parametrised by $\theta$, whose structure is easy to evaluate to approximate the posterior distribution. In this case, the objective is to minimise the Kullback-Leibler (KL) divergence (a distance metric)[67,71] between the true posterior distribution and the variational distribution:

$$KL\big(q_\theta(\boldsymbol{\omega}) \,||\, p(\boldsymbol{\omega}|\mathbf{X}, \mathbf{Y})\big) = \int q_\theta(\boldsymbol{\omega}) log \frac{q_\theta(\boldsymbol{\omega})}{p(\boldsymbol{\omega}|\mathbf{X}, \mathbf{Y})} d\,\boldsymbol{\omega}. \tag{2.5.3}$$

Afterwards, the inference can be conducted by replacing the posterior distribution $p(\boldsymbol{\omega}|\mathbf{X}, \mathbf{Y})$ in eq. (2.5.2) with the variational distribution $q_\theta(\boldsymbol{\omega})$:

$$p(y^*|x^*, \boldsymbol{X}, \boldsymbol{Y}) = \int p(y^*|x^*, \boldsymbol{\omega}\,)\, q_\theta(\boldsymbol{\omega}) d\boldsymbol{\omega}.$$

As shown in eq. (2.5.3), it is infeasible to directly minimise the KL divergence between the posterior and the variational distribution. However, instead of minimising the KL divergence, we perhaps can maximise a function if the summation of the function and the KL divergence equals a constant. To this end, let us reorganise the eq. (2.5.1) as below:

$$\log\big(p(\boldsymbol{Y}|\boldsymbol{X})\big) \quad = \log\Big(\frac{p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\omega})p(\boldsymbol{\omega})}{q_\theta(\boldsymbol{\omega})}\Big) - \log\Big(\frac{p(\boldsymbol{\omega}|\mathbf{X}, \mathbf{Y})}{q_\theta(\boldsymbol{\omega})}\Big),$$

$$= \int q_\theta(\boldsymbol{\omega}) \log\left(\frac{p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\omega})p(\boldsymbol{\omega})}{q_\theta(\boldsymbol{\omega})}\right) d\boldsymbol{\omega} - \int q_\theta(\boldsymbol{\omega}) \log\left(\frac{p(\boldsymbol{\omega}|\mathbf{X},\mathbf{Y})}{q_\theta(\boldsymbol{\omega})}\right) d\boldsymbol{\omega},$$

$$= \int q_\theta(\boldsymbol{\omega}) \log\left(\frac{p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\omega})p(\boldsymbol{\omega})}{q_\theta(\boldsymbol{\omega})}\right) d\boldsymbol{\omega} + \mathrm{KL}(q_\theta(\boldsymbol{\omega}) \,||\, p(\boldsymbol{\omega}|\mathbf{X},\mathbf{Y})),$$

$$\geq \int q_\theta(\boldsymbol{\omega}) \log\big(p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\omega})\big)\, d\boldsymbol{\omega} - KL(q_\theta(\boldsymbol{\omega}) \,||\, p(\boldsymbol{\omega})),$$

where $\log\big(p(\boldsymbol{Y}|\boldsymbol{X})\big)$ is a constant, and it equals the summation of the KL divergence between the variational and the posterior distribution and the evidence lower bound (ELBO) (i.e., $\int q_\theta(\boldsymbol{\omega}) \log\left(\frac{p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\omega})p(\boldsymbol{\omega})}{q_\theta(\boldsymbol{\omega})}\right) d\boldsymbol{\omega}$). Therefore, the KL divergence minimisation problem can be appropriately redefined as maximising the ELBO. The ELBO also includes two terms. The first term $\int q_\theta(\boldsymbol{\omega}) \log\big(p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\omega})\big)\, d\boldsymbol{\omega}$ is the expected log-likelihood, and the optimisation encourages $q_\theta(\boldsymbol{\omega})$ to better describe the observed data. The second term $KL(q_\theta(\boldsymbol{\omega}) \,||\, p(\boldsymbol{\omega}))$ acts as a regulariser. The optimisation of ELBO can penalise the complexity of the variational distribution by encouraging $q_\theta(\boldsymbol{\omega})$ to be as close as possible to the prior. This procedure is known as variational inference, a standard Bayesian inference technique.

### 2.5.3 Gaussian processes (GPs)

GP is a popular and elegant Bayesian non-linear non-parametric model with natural properties for uncertainty estimation.[72] For the convenience of discussion, we only consider regression at this stage, but they can be easily used for binary classification by wrapping a logistic regression.[24] For a given training data $\mathcal{D} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$, where $x_i$ and $y_i$ represent input and output, respectively. GPs usually place a GP prior on the regression function, denoted by $f(\boldsymbol{x}) \sim GP(v(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}'))$, where $v$ is the mean function. The kernel function $k$ controls the smoothness of GPs, and it is required to be a positive definite function. For any finite set of observed points, we define a join Gaussian: $p(f|\boldsymbol{X}) =$

$\mathcal{N}(f|\mu, \boldsymbol{K})$, where $\boldsymbol{K}_{i,j} = k(x_i, x_j)$ and $\mu = (v(x_1), \ldots, v(x_N))$, and $v(x)$ is often taken as zero since the GPs are flexible enough to model the mean arbitrarily well.

To predict new data $\mathcal{D}^* = \{({x_1}^*, {y_1}^*), \ldots, ({x_{N^*}}^*, {y_{N^*}}^*)\}$, where $y^*$ and $x^*$ are new output and input, respectively, GPs have the posterior as following:

$$p(f^*|\boldsymbol{X}^*, \boldsymbol{X}, f) = \mathcal{N}(f^*|\mu^*, \Sigma^*),$$

$$\mu^* = {\boldsymbol{K}^*}^T \boldsymbol{K}^{-1} \boldsymbol{y},$$

$$\Sigma^* = \boldsymbol{K}^{**} - {\boldsymbol{K}^*}^T \boldsymbol{K}^{-1} \boldsymbol{K}^*,$$

where $\boldsymbol{K} = k(\boldsymbol{X}, \boldsymbol{X})$ is $N \times N$, $\boldsymbol{K}^* = k(\boldsymbol{X}, \boldsymbol{X}^*)$ is $N \times N^*$, and $\boldsymbol{K}^{**} = k(\boldsymbol{X}^*, \boldsymbol{X}^*)$ is $N^* \times N^*$.

Additionally, if we assume that the observations are obtained by adding Gaussian noise to the underlying latent function, we can define $y_i = f(x_i) + \varepsilon_i$, where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. This will lead to a GP posterior specified as following:

$$p(f^*|\boldsymbol{X}^*, \boldsymbol{X}, f) = \mathcal{N}(f^*|\mu^*, \Sigma^*),$$

$$\mu^* = {\boldsymbol{K}^*}^T (\boldsymbol{\sigma^2 I} + \boldsymbol{K})^{-1} \boldsymbol{y},$$

$$\Sigma^* = \boldsymbol{K}^{**} - {\boldsymbol{K}^*}^T (\boldsymbol{\sigma^2 I} + \boldsymbol{K})^{-1} \boldsymbol{K}^*,$$

The posterior GP depends on the marginal likelihood for selection hyperparameters, which is shown in eq. (2.5.4)

$$\log p(y) = -\frac{1}{2} \boldsymbol{y}^T {\boldsymbol{K_n}}^{-1} \boldsymbol{y} - \frac{1}{2} \log|\boldsymbol{K_n}| - \frac{N}{2} \log(2\pi), \tag{2.5.4}$$

where $\boldsymbol{K_n} = \boldsymbol{K} + \boldsymbol{\sigma^2 I}$.

### 2.5.4 Sparse Gaussian processes (SGPs)

GPs need to calculate the determinant and inversion of the covariance matrix $\boldsymbol{K}$, the complexity of which increases dramatically as the number of observations increases. Many approaches have been proposed to approximate $\boldsymbol{K}$ with a lower rank matrix. One popular approach is a posterior approximation using variational free energy proposed by Titsias[73]. The method suggests using a pseudo dataset (with $M$ pseudo points $U = \{u_1, \ldots, u_M\}$, where

$M \ll N$) as an approximation of the observed dataset. Therefore, instead of calculating the covariance matrix for the observations, it conducts inference using the pseudo points, which can substantially simplify the calculation. In terms of optimisation, it uses ELBO as the objective (as shown in eq. (2.5.5)) and the pseudo points as a part of the variational parameters to rigorously minimise the distance (i.e., KL divergence) between the posterior and variational posterior.

$$L_{SGP} = -\frac{1}{2}\boldsymbol{y}^T \boldsymbol{Q_n}^{-1}\boldsymbol{y} - \frac{1}{2}\log|\boldsymbol{Q_n}| - \frac{N}{2}\log(2\pi) - \frac{t}{2\boldsymbol{\sigma^2}}, \tag{2.5.5}$$

where $\boldsymbol{Q_n} = \boldsymbol{Q} + \boldsymbol{\sigma^2 I}$, $\boldsymbol{Q} = \boldsymbol{K_{un}}^{\boldsymbol{T}}\boldsymbol{K_{uu}}\boldsymbol{K_{un}}$, $\boldsymbol{K_{un}} = \boldsymbol{K}(\boldsymbol{U},\boldsymbol{X})$ is $M \times N$, $\boldsymbol{K_{uu}} = \boldsymbol{K}(\boldsymbol{U},\boldsymbol{U})$ is $M \times M$, $t = Tr(\boldsymbol{K} - \boldsymbol{Q})$.

### 2.5.5 Kernel interpolation for scalable structured Gaussian processes (KISS-GP)

In addition to posterior approximation, Wilson and Nickisch[74] proposed a structured kernel interpolation framework to produce a more scalable kernel approximation, named KISS-GP. This method combines structure-exploiting approaches, inducing points, and sparse interpolation to further reduce the inference time cost and storage costs from $\mathcal{O}(M^2N + M^3)$ and $\mathcal{O}(MN + M^2)$ in SGP to $\mathcal{O}(N)$. The main idea of KISS-GP is to impose the grid constraint on the inducing points. Therefore, the kernel matrix $\boldsymbol{K_{uu}}$ admits the Kronecker structure or a Toeplitz covariance matrix for a much easier calculation. For the cross-kernel matrix $\boldsymbol{K_{un}}$, it can be approximated, for example, by a local linear interpolation with adjacent grid inducing points as shown below:

$$\mathrm{k}(x_i, u_j) = w_i k(u_a, u_j) + (1 - w_i)k(u_b, u_j),$$

where $u_a$ and $u_b$ are two inducing points on the grid that are closest to $x_i$, and $w_i$ is an interpolation weight that represents the distance to the inducing point. Eventually, KISS-GP can be trained the same as SGP, but with matrix $\boldsymbol{Q}$ approximated as $\boldsymbol{Q} \approx \boldsymbol{W_{un}}^{\boldsymbol{T}}\boldsymbol{K_{uu}}\boldsymbol{K_{un}}$.

# 3. Data source: CPRD

CPRD is the primary data source for studies conducted in this thesis. Therefore, we shall give readers an introduction in this chapter to get familiar with CPRD. This serves as an overview of the CPRD dataset, and the subpopulation selected for each study will be further discussed in Chapters 4, 5, and 6, respectively.

## 3.1 Background

The CPRD is an ongoing primary care database that provides EHR to support public health and clinical research. This service is jointly sponsored by the Medicines and Healthcare products Regulatory Agency and the National Institute for Health and Care Research, as part of the Department of Health and Social Care.[75] CPRD collects de-identified patient data from a network of primary care (general) practices across the UK, and the data can be further linked to a range of other health-related services.[75] Therefore, it is one of the most comprehensive longitudinal datasets in the world and representative of the UK population.[76] For the last 30 years, it has been widely used for studies in drug safety, use of medicines, the effectiveness of health policy, health care delivery, and disease risk factors.[75]

As of May 2022, CPRD collected EHR from 63 million de-duplicated patients, with 16 million currently registered using the Egton Medical Information Systems (EMIS) and Vision software system.[75] Due to the difference in structures and the clinical coding in these two software, CPRD provided data from two different systems as two separate databases: CPRD GOLD and CPRD Aurum. CPRD GOLD database contained patient EHR collected from 985 general practices using the Vison software, covering 4.9% of the UK general practices and 4.61% of the UK population.[75] Additionally, CPRD Aurum was contributed

from 1,491 general practices using EMIS with 19.83% UK population coverage.[75] The CPRD contains rich health data from various sources, including demographics, symptoms, tests, diagnoses, therapies, procedures, and health-related behaviours.[8,76,77] Additionally, as data are prospectively recorded as part of routine care, they offer great generalisability of the findings to the routine clinical settings than the traditional observational studies.[78]

## 3.2 Structure of the CPRD

CPRD contains de-identified patients' health information recorded by the primary care providers during routine clinical encounters. It can be broadly structured into clinical, referral, immunisation, test, and therapy data.[8] Each is referred to as a "topic" and can be linked using unique identifiers. For those primary care practices that participate in the CPRD linkage scheme, the patient data can be individually linked to other health and area-based datasets, such as Hospital Episode Statistics (HES) Admitted Patient Care data, death registration data from the Office for National Statistics (ONS), or socioeconomics data.[75] Additionally, all patient data entries are considered "consultations", and one consultation can have multiple "events" with each associated with a date.[8] A visual illustration can be found in Figure 3.1.

Data in CPRD are largely recorded using version 2 Read codes[79] by general practices staff. It includes over 96,000 codes to describe a wide range of conditions, such as diagnoses, past medical history, symptoms, type of test performed (e.g., blood pressure measurement), and lifestyle measurement (e.g., drinking and smoking status). Each code is also associated with the detailed numerical clinical measures during consultations (e.g., height, weight, and blood pressure). The prescriptions issued by general practices are recorded using both British National Formulary (BNF) code[80] and product code[81], with the corresponding dosage instructions and quantity. Additionally, the uncoded free text is also recorded by the general

practices. However, such information is identifiable. Hence, it is not available as a standard part of CPRD to researchers.[8]

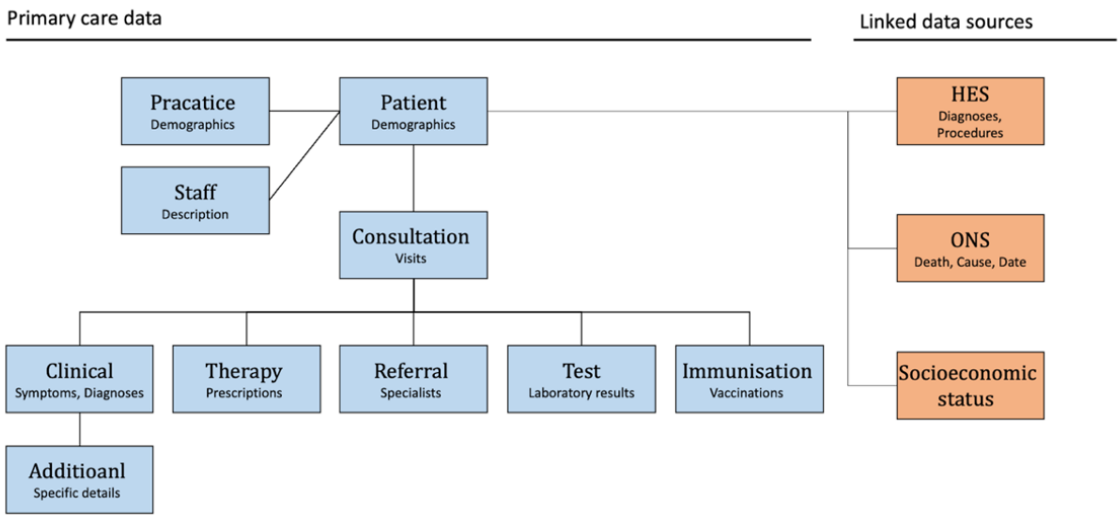



*Figure 3.1: Overview of the data structure of CPRD and linked data sources. Patients consult with practice staff, where clinical therapy, referral, test and immunisation information is coded in the medical record.*[8]

## 3.3 Linked databases

CPRD can be linked to a wide range of health-related databases via a trusted third party (the Health and Social Care Information Centre), including HES (hospitalisation), ONS (death registration), socioeconomic status (index of multiple deprivations (IMD)), and disease registries (the National Cancer Intelligence Network and the Myocardial Ischaemia National Audit Project). However, the linkage is only available for a subset of English practices that have consented to participate in the CPRD linkage scheme, covering approximately 75% of the English practice and 50% of the UK practices in CPRD.[8] Because our study only involves HES, ONS, and socioeconomic status, we will mainly introduce these three linkages.

### 3.3.1 Hospital episode statistics (HES)

HES includes all admissions to National Health Service (NHS) in England and the admissions to independent sector providers (private or charitable hospitals) paid for by the NHS. The NHS funds an estimated 98%-99% of the hospital activity in England.[82] This

thesis mainly uses the HES dataset related to "Admitted Patient Care (APC)", which is hospital admissions about secondary care-based activity requiring a hospital bed.[82] It includes emergency and planned admissions, day cases, births and associated deliveries. Accident and emergency (A&E, emergency department), attendances or outpatient bookings are not covered in APC. HES APC contains hospital episode information at the patient level, including admission and discharge dates, diagnoses, specialists seen, and procedures undertaken. Each admission can also have multiple diagnoses. One of them is the primary diagnosis, and the other diagnoses are referred to as comorbidities. All diagnoses are recorded using the International Classification of Diseases and Health-Related Problems, Tenth Edition[83] (ICD-10) coding system. The procedures follow the same protocol for recording and are coded using the Office of Population Censuses and Surveys Classification of Interventions and Procedures[84] (OPCS) version 4.7.

### 3.3.2 Death registration data from Office for National Statistics (ONS)

ONS mortality data is a rich source containing all deaths registered in England and Wales with information on the place of death, the date of death, and the underlying causes.[85] The causes of death are recorded using ICD-10 and comprise one primary cause of death and up to 15 secondary causes.

### 3.3.3 Index of Multiple Deprivation (IMD)

The IMD measures multiple deprivations in 32,844 small areas or neighbourhoods, called Lower Super Output Area (LSOA), in England, where an LSOA contains between 400 and 1,200 households.[82,86] It is calculated based on 37 separate indicators across seven distinct domains, including income deprivation, employment deprivation, health deprivation and disability, education deprivation, crime deprivation, barriers to housing and services deprivation, and living environment deprivation.[87] The LSOAs are further ranked by the index from the most deprived area (1) to the least deprived area (32,844) and divided into

five equal groups based on the quintiles they fall into. CPRD provides information about the deprivation quintiles instead of the ranks by matching a patient's postcode to the neighbourhood it refers to. The IMD dataset used in this thesis is the 2015 version.

### 3.3.4 Representativeness

CPRD patients broadly represent the UK population in age, sex, ethnicity, and body mass index (BMI) distribution.[8] However, the CPRD may not be representative of all practices in the UK based on geography and size.[8]

### 3.3.5 Data quality

Although CPRD is not specifically designed for research, it has established a consistent approach to ensure that the data quality adheres to the standards required for research purposes, including acceptability for patients and up-to-standard (UTS) date for practices. Acceptability is an indicator that only patients that are "acceptable" are recommended for use in research. It is evaluated based on registration status, recording of events in the patient record, and valid age and gender.[8] The UTS date is a practice-based quality metric based on the continuity of recording and the number of recorded deaths.[75] It is deemed the date at which data in the practice is considered to have continuous high quality for use in research. Therefore, only records that contribute to the dataset after the UTS dates and only patients marked as "acceptable" by CPRD are included in the analyses in this thesis.

# 4. Deep learning for risk prediction

As introduced in Chapter 1, the first objective of this doctoral thesis is to investigate how we can deliver accurate risk models for prognostication using deep learning and temporal and multimodal EHR. To this end, this chapter will start by introducing a proposed Transformer-based risk model, BEHRT, for accurate risk prediction (Chapter 4.1) and the model performance in the presence of data shifts compared to the established machine learning and clinical models (Chapter 4.2). Afterwards, we will discuss the limitations of the BERT model and the potential approach to empower risk prediction with a patient's entire medical trajectory (i.e., comprehensive multimodal EHR) (Chapter 4.3). In the end, topics about adapting the deep learning model for survival analysis (Chapter 4.4) and the potential future works (Chapter 4.5) will also be discussed. We will be presenting the chapters according to the TRIPOD guideline[88]. For the remainder of the paper, CPRD is the main data source for the DPhil project and it recorded admission records as time-stamped codes and measures rather than clinical notes and free text (§3.2).

## 4.1 Transformer for electronic health records

This has been joint work with Shishir Rao and published in Scientific Reports at: https://www.nature.com/articles/s41598-020-62922-y. The manuscript benefitted from the input of several co-authors. I am the first author and the work presented in this chapter is my own work. The sensitivity analyses are jointly produced with Shishir Rao. My role included designing the model, conceiving the experiments and analyses, extracting and cleaning data, and writing the manuscript.

### 4.1.1 Motivations

In traditional research on risk prediction, individuals are usually represented as a vector of attributes, called "features" or "predictors".[89] This approach primarily relies on the understanding of the outcome and the experts' ability to define the appropriate features. Therefore, the essential question here is "what are the key features for this prediction?" or "what features should have interactions with one another?", and it is becoming the fundamental bottleneck for risk prediction even with the accessibility to all medical data would not solve it better. However, recent developments in deep learning provided us with models that can learn meaningful representations from raw or minimally processed data with little expert guidance.[90] This is achieved through a sequence of layers in the model, each employing a large number of linear or nonlinear transformations to learn representations at an increasing level of abstraction; this process across layers leads to a final representation in which the patterns of data points with different outcomes are distinguishable.

Besides the developments in deep learning algorithms, large-scale data is another critical factor contributing to its success. The deep learning approach is known to be data-driven and data-hungry. It must be fuelled with massive data to gain the ability for representation learning. The growing availability of EHR systems in recent years provides an unprecedented opportunity to create an enormous influx of large multimodal EHR data. Hospitals that have adopted EHR systems now exceed 84% and 94% in the US and UK, respectively.[91,92] Therefore, EHR systems of a national medical organisation nowadays are likely to capture comprehensive medical data from millions of individuals over many years. This creates a great potential for us to investigate the impact of deep learning on risk prediction using large-scale EHR.

### 4.1.2 Related work

Deep learning with large-scale medical data has led to great progress in many fields in healthcare, such as cardiovascular medicine, radiology, neurology, dermatology, ophthalmology, and pathology,[93–96] and demonstrated its utility in learning meaningful representations to inform accurate risk prediction. One of the earliest works from Liang et al.[97] showed strong evidence that deep neural networks (DNNs) can outperform classic machine learning models that rely on manual feature engineering over several prediction tasks on several different datasets. Tran et al.[98] applied restricted Boltzmann machines for EHR representation learning, and it outperformed the manual feature extraction for predicting suicide risk. Similarly, Miotto et al.[99] employed a stack of denoising autoencoders for representation learning using EHR to predict the onset of a number of diseases, and it outperformed many popular feature extraction and transformation approaches, such as principal component analysis, independent component analysis, and Gaussian mixture models.[100]

Despite these earlier works showing great potential to use deep learning to improve care provision, they ignored some of the most important properties in EHR data, for instance, the irregularity of the inter-visit intervals and the temporal order of events or encounters. To address this, Nguyen et al.[101] proposed to treat a patient's medical history as a sequence of events (e.g., diagnoses and medications) and inserted a special event between each pair of consecutive visits to denote the time difference between them for sequential modelling. In another similar attempt, Choi et al.[102] employed a shallow RNN to firstly encode all events within the same visit as a single high-dimensional visit representation. Thus, a patient's medical history was represented as a sequence of high-dimensional visit representations for further sequential modelling.

In addition to sequential modelling, another major improvement in recent years for applying deep learning to EHR was modelling long-term dependencies among events. For instance, key diagnoses like diabetes or hypertension can stay a risk factor over a patient's life even decades after their first incidence; certain surgeries may prohibit specific future interventions. Therefore, the ability to model the long-term dependencies can be critical and informative in certain health applications. Pham et al.[103] showed that the LSTM with attention mechanism outperformed the standard LSTM and RNN for predicting the onset of diabetes. Choi et al.[104] proposed a RETAIN model, which employed LSTM with a reverse-time attention mechanism to consolidate past influential visits for heart failure (HF) risk prediction. It outperformed most of the models at the time of its publication and provided a decent baseline for the EHR learning research.

### 4.1.3 Aims

In this work, we aim to build on some of the latest developments in deep learning (Transformer) while considering various EHR-specific challenges and providing improved accuracy for predicting future diagnoses using large-scale EHR. We named our model BEHRT (i.e., BERT for EHR) due to its architectural similarities with BERT[38].

### 4.1.4 Methods

#### 4.1.4.1 Data source

This study was conducted using EHR from CPRD linked to HES and ONS. Primary data was available from 1 January 1985 to 31 December 2015, and the linkage was available from 1 January 1998 to 30 September 2015.

#### 4.1.4.2 Study design

This study investigated subsequent visit risk prediction tasks: prediction of disease in the next visit, next six months, and next 12 months, respectively. Each task was essentially a multi-label risk prediction task that predicted all potential diseases in subsequent visits

using available information before the baseline date. The prediction of each disease is considered a binary classification task. Note that, the prediction tasks in this work were designed as "force to predict", and only patients with at least one event in the future visit/six months/12 months are included. The baseline date for risk prediction was randomly selected from one's EHR history for each individual.

#### 4.1.4.3 Eligible participants

In this study, we selected a cohort with patients who contributed to data between 1 January 1985 and 31 December 2015, eligible to be linked to HES, marked as "acceptable" by CPRD quality control. This led to 4 million patients. Afterwards, we applied a two-sage patient selection. For the first stage, patients eligible for inclusion in the study were men and women, aged at least 18, had enough useful history records (at least five visits). This led to a cohort with 1,609,024 patients. For the second stage, patients with minimal sufficient history records (i.e., at least five visits) before the baseline date for subsequent visit risk prediction were kept. A total of 699, 391, and 324 thousand patients were included for prediction of disease in the next visit, next six months, and next 12 months, respectively.

*Table 4.1: Population characteristics of selected cohorts for subsequent visit risk prediction*

| Characteristics | | Next visit | Next six months | Next 12 months |
|---|---|---|---|---|
| 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% |
| | Black Caribbean | 0.20% | 0.30% | 0.20% |
| | Other Asian | 0.10% | 0.10% | 0.10% |
| | Black African | 0.10% | 0.10% | 0.10% |
| | Mixed | 0.10% | 0.10% | 0.10% |
| | Bangladeshi | 0.08% | 0.07% | 0.07% |
| | Black Other | 0.07% | 0.06% | 0.06% |
| | Chinese | 0.06% | 0.06% | 0.05% |
| Age start | Median | 58 | 60 | 53 |
| Age end | Median | 70 | 71 | 71 |
| Unique codes | Median | 9 | 10 | 11 |
| Number of visits | Median | 15 | 20 | 20 |

#### 4.1.4.4 Outcomes

Because diagnoses are recorded using different coding systems for primary care and HES (Read code for primary care and ICD-10 for HES). We mapped diagnoses from both systems to CALIBER codes[105], an expert-checked phenotyping dictionary provided by University College London. Therefore, we can consistently define the same disease recorded by different systems. Eventually, this led to a total of 301 unique CALIBER codes for diagnoses in our dataset. The outcomes of the risk prediction task were the 301 conditions defined by CALIBER codes.

#### 4.1.4.5 Data preparation

Diagnosis records collected from primary care and HES were included for modelling, and all diagnoses were mapped to 301 CALIBER codes. Additionally, each record was associated with an event date and the chronological age (calculated based on event date and date of birth). Further, to use minimal processed EHR for deep learning modelling, we assumed that no recording of an event corresponds to lack of event occurrence (or the awareness of the physician of it). Thus, data imputation was not relevant for events.

#### 4.1.4.6 BEHRT: A Transformer-based model for EHR

We proposed BEHRT for EHR modelling in this study. It considered four fundamental challenges in EHR modelling: (1) complex and nonlinear interactions among past, present, and future events; (2) long-term dependencies among events (events occurring in a patient's early life affecting events far later in the future); (3) difficulties of representing heterogeneous events in various size and forms to the model; (4) the irregular interval between consecutive visits.[103]

The proposed BEHRT model was inspired by the recent success of Transformers[37], more specifically, BERT[38], because of the similarity between NLP and EHR in sequential

modelling. As shown in Figure 4.1, by depicting events as words, each visit as a sentence, and a patient's entire medical history as a document, BEHRT follows the same philosophy to model EHR as BERT for NLP. Additionally, compared to BERT, several important changes have been made to adopt the Transformer for EHR modelling. First, a combination of four embeddings: event, "position", age, and "visit segment" (or "visit" for short), were employed to represent the evolution of one's EHR. Such a combination provides the flexibility to represent one's medical trajectory as detailed as possible. More specifically, the event embedding informs the model about all the events recorded in the past in one's EHR. That is, knowing the past disease trajectories and morbidity patterns can inform the prediction for future diagnoses. Position embedding follows the same philosophy as the Transformers to indicate the relative position of events in the EHR sequence. Therefore, it enables the model to be aware of the order of the events and capture the positional interaction among them.

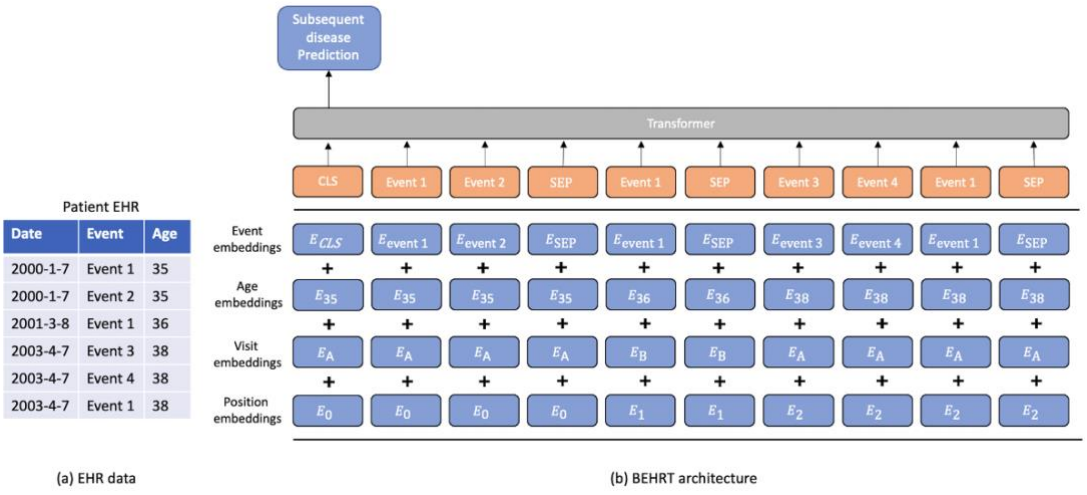



*Figure 4.1: BEHRT architecture and input representation for a hypothetical patient. (a) and (b) show the process of transforming the raw EHR to the input sequence for BEHRT. (a) EHR records for a hypothetical patient, where the date, event, and age represent the date that a corresponding event is recorded, the recorded event, and the age of the patient at the time of recording (i.e., the difference in years between date and the date of birth), respectively. (b) BEHRT architecture includes four embedding layers (event, age, visit, and position), and the final embedding of an event is their summation. "CLS" and "SEP" are two special symbols injected into the EHR sequence. "CLS" is always at the beginning of the sequence, and "SEP" s are injected between two consecutive visits.*

Age and visit segment embeddings are explicitly designed for BEHRT to empower the model further to deal with the abovementioned challenges. Age is known to be a key risk factor for most diseases (i.e., the increase in age can increase the risk of many diseases). By contextualising the events with the age embedding, it provides a sense of time on the irregular interval between visits and serves as a universal epidemiological notion of when things happen (comparable across patients). The visit segment uses two trainable embeddings (referred to as "A" and "B") that alternatively change between visits to add an additional level of information to indicate the separation between two consecutive visits (i.e., two adjacent visits of a patient will always have different visit segment embeddings).

Note that, the embedding design of BEHRT ensures that there is no prescribed order for multiple events within the same visit. That is, the age, visit segment, and position embeddings are identical for all events within the same visit. This is consistent with a human's intuition and makes BEHRT order-invariant for intra-visit events. Therefore, the inter-visit and intra-visit relationship can be maximally preserved, and their interaction only depends on the attention mechanism.

Through a combination of the abovementioned four embeddings, we present the model with events along with their medical context, a precise sense of time, and the details about the delivery of care. In other words, we present the EHR data to the model as how a doctor reviews a patient's medical trajectory. Therefore, it provides the model with full potential to learn from the medical history that, traditionally, we might have sought to capture through many features manually extracted from EHR.

#### 4.1.4.7 Model derivation

In this study, we developed the BEHRT model on two tasks: MLM (§2.3.4.1) and subsequent visit risk prediction. For the convenience of model assessment, we randomly separated the selected cohort from stage one into a training and validation cohort (containing

80% and 20% patients, respectively). For each of the subsequent visit risk prediction cohort, patients within the training and validation cohort were used for training and validation, respectively. BEHRT for MLM was developed on the training set.

Additionally, BEHRT incorporated all available diagnosis records for MLM training. However, for subsequent visit risk prediction, BEHRT only used all available diagnosis records before the baseline date for learning to predict the disease in the next visit, next six months, and next 12 months, respectively, after the baseline date.

Regarding model training, BEHRT was jointly trained on 301 binary risk prediction tasks using averaged binary cross-entropy loss for subsequent visit risk prediction. Bayesian optimisation[106] was applied for hyper-parameter tuning on the MLM task for 25 searches (Table A.1), and the model that achieved the best precision for MLM prediction was selected as the optimal architecture. This led to a BEHRT model with six layers, 12 attention heads, 512 for the hidden size of the intermediate layer, 288 for hidden size, 0.2 for dropout rate, and 100 for maximum sequence length. For reproducibility purposes, the model was trained for 100 epochs using Adam optimiser[107] with a learning rate of $3e^{-5}$ and weight decay $0.01$ on the MLM task, and we used age in months instead of years with maximal age of 110 years old for training. The BEHRT models for risk prediction were finetuned on the BEHRT model trained on the MLM task. The finetuning was conducted on all trainable parameters. Adam optimiser with learning rate $3e^{-6}$ and weight decay $0.01$ was also used for the subsequent visit prediction tasks.

Additionally, two state-of-the-art deep learning risk models, (CNN-based) Deepr [101] and (LSTM-based) RETAIN[104] were developed as the comparators in our study. Bayesian optimisation was also applied for hyper-parameter tuning for both Deepr (Table A.2) and RETAIN (Table A.3), and they were trained the same way as BEHRT for subsequent visit risk prediction.

#### 4.1.4.8 Model validation and analysis

Precision was used to continuously monitor the progress of MLM training (a measurement of how many masked tokens can be correctly recovered). To further validate the MLM training, we devised a visual investigation and an association analysis on the embeddings compared with established medical knowledge. For the visual validation, we mapped high-dimensional event embeddings to two-dimensional vectors that can be directly plotted for visualisation using T-distributed stochastic neighbour embedding[108] (t-SNE) and analysed the association between different events and their neighbours. For the association analysis, we found the top 10 disease with the highest association (i.e., closest cosine distance[109]) for each disease with prevalence in at least 1% of the population. Then, a clinical researcher manually validated the consistency between the identified association and the expert's knowledge.

For subsequent visit risk prediction, we evaluated the model performance on the validation set using the area under the receiver operating characteristics curve (AUROC)[110] and average precision score (APS)[111] averaged across 301 outcomes and compared the performance of BEHRT with two state-of-the-art deep learning risk models (Deepr[101] and RETAIN[104]). We also conducted a series of subgroup analyses to further investigate BEHRT's predictive performance. We investigated (1) if BEHRT can implicitly learn the difference in gender from the context without the incorporation of gender information in training; (2) how selectively deactivating the age, visit, or position embeddings can affect the risk prediction (referred to as ablation study); and (3) performance of BEHRT for predicting new incidences of diseases in the subsequent visit (i.e., only predicting events in the label but had not appeared in the history at the time of prediction).

### 4.1.5 Results

#### 4.1.5.1 Disease embedding visual analysis

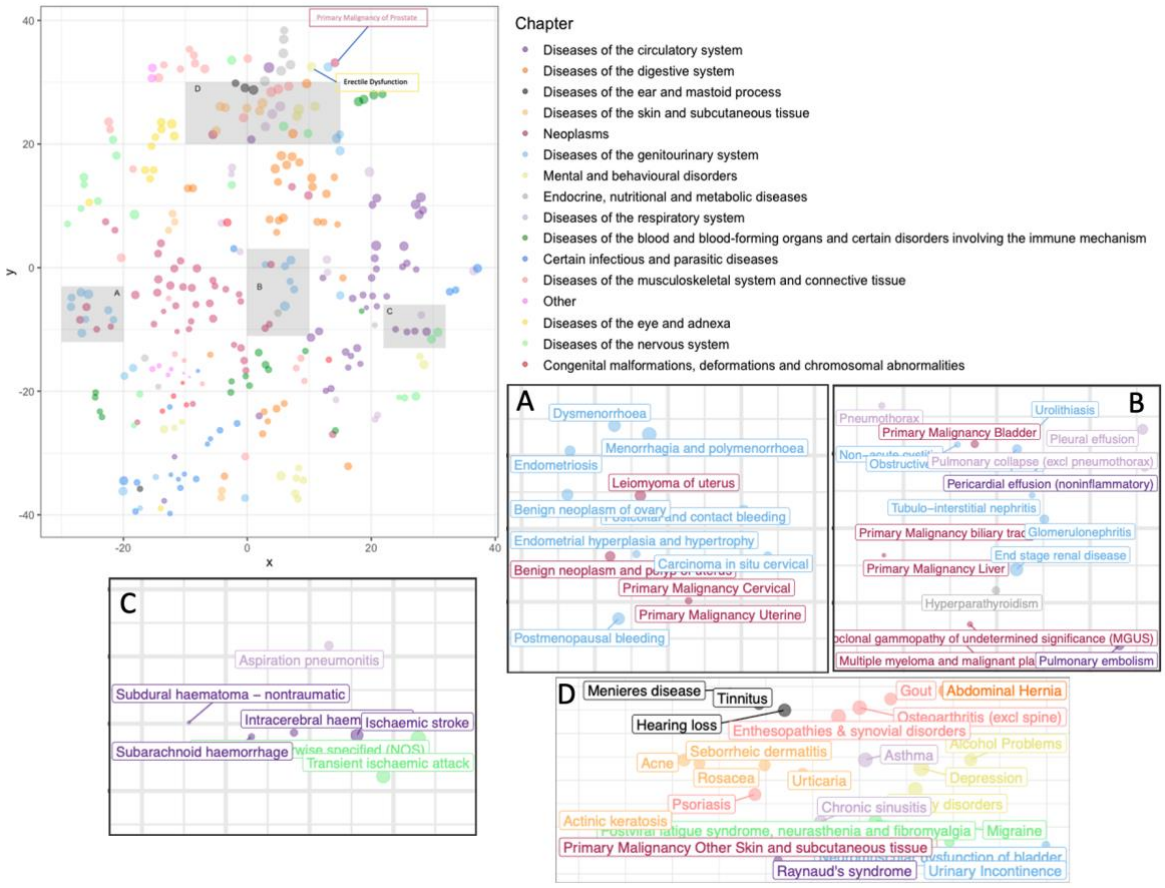



*Figure 4.2: Visual illustration of event embeddings. The high-dimensional event embeddings are projected into two-dimensional vectors where the distance between events indicates their closeness (i.e., closer events have a higher association). The colours represent the CALIBER chapters, and events within the same chapter have the same colour. Some areas are zoomed in for better visualisation.*

Figure 4.2 is a visual illustration that indicates the event associations projected from high-dimensional space using t-SNE. It shows that, in general, many diseases known to co-occur and/or belong to the same clinical groups are clustered together. For instance, we can see that many diseases that belong to the same CALIBER chapter[105] (i.e., same colour) are located in the same area. CALIBER is an expert-defined phenotyping dictionary that groups diseases within the same system under the same chapter. Therefore, the observation indicated that the event representations learned by BEHRT were highly consistent with the established expert knowledge.

Additionally, beyond the expert-based phenotyping, the learned representations clustered correlated diseases that were not identical to a specific chapter. For example, many eye and adnexa diseases were among many nervous system diseases, and many nervous system diseases were also among many musculoskeletal diseases. Note that, although most of the clusters and the neighbourhoods are broadly consistent with medical knowledge, there are clusters that might seem counter-intuitive, perhaps because the extreme dimensionality reduction for visualisation can cause potential distortion, hence cannot perfectly preserve their distance (i.e., association) structure in the high-dimensional space. Another interesting pattern that can be seen in Figure 4.2 is the natural stratification of gender-specific diseases. For instance, the male-specific diseases (e.g., erectile dysfunction and primary malignancy of prostate) are quite distant from the female-specific diseases (e.g., endometriosis, menorrhagia, and dysmenorrhea). Such observation indicates the medical contexts of gender-specific diseases are considerably different; thus, even the information about gender is not explicitly available to the model, it can infer such factors based on the understanding of the context.

#### 4.1.5.2 Disease embedding association analysis

In terms of analysing diseases and their corresponding top ten associations. Nearly 76% (i.e., 75.7%) of the identified associations were aligned with the expert's expectations. The clinical researcher also notes that while many of the closest diseases had clear overlap in symptomatology, some were graded to be poor disease associations. Therefore, it is believed that BEHRT can learn meaningful characteristics of the diseases from the medical EHR without them being explicitly given to it. An **online supplementary**, accessible under https://www.researchgate.net/profile/Yikuan-Li, was set up to make the association table available for other researchers to download and assess. Space constraints prohibit the inclusion of the association table in the supplementary appendix of this thesis.

### 4.1.5.3 Subsequent visit risk prediction

*Table 4.2: Model performance in subsequent visit prediction tasks. APS and AUROC indicate the average APS and AUROC of all (i.e., 301) diseases for the multi-label risk prediction task.*

| | APS\|AUROC | | |
|---|---|---|---|
| Model name | Next visit | Next six months | Next 12 months |
| **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 |

The embedding and attention analyses showed solid evidence that BEHRT can learn meaningful representations from the raw EHR data. We further validated the BEHRT model for subsequent visit risk prediction tasks and compared it with the state-of-the-art benchmark models (Deepr and RETAIN) in the literature. Table 4.2 shows that BEHRT substantially outperforms both Deepr and RETAIN for all three prediction tasks, with an 8% to 13.2% improvement in APS. We also include a prevalence-based analysis in Figure A.1.

### 4.1.5.4 Subgroup analysis for gender-related events

*Table 4.3: Predictions of gender-specific disease (next six months). Male and female predictions in the table represent the number of patients with a predictive probability above 0.5 for that gender, and disease gender indicates whether a disease is a male-specific or female-specific disease (M and F represent male and female, respectively).*

| Disease Name | Disease gender | Male predictions | Female predictions |
|---|---|---|---|
| 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 |

To investigate the model performance further, we conducted several subgroup analyses. The first analysis assessed the model's ability to distinguish the gender-related events (without being given gender as an input variable). As shown in Table 4.3, most of the gender-specific diseases are correctly categorised by BEHRT, except "male infertility". This might seem like an error in the results; however, after careful investigation, we found that

the “male infertility” condition actually included 354 male and 1,734 female patients in the dataset. The same mistake also existed in “female infertility” (had 192 male and 2,306 female patients). This was because the Read code in CPRD was gender agnostic, and some codes were mapped to both “male infertility” and “female infertility” by the CALIBER system.

#### 4.1.5.5 Subgroup analysis for ablation study

Since BEHRT uses a combination of event, age, visit, and position embeddings as inputs, it is reasonable to investigate how different embedding layers affect the risk prediction. Therefore, the second analysis we conducted was called the ablation study, which selectively deactivated certain embedding layers for model performance assessment. Table 4.4 shows that the change of embedding layers has a more significant impact on the APS than the AUROC. Overall, the age and position embeddings are more important than the visit embedding for risk prediction as their inclusion can significantly improve the model performance. However, the inclusion of visit embedding results in minimal changes in model performance.

*Table 4.4: Ablation study for embedding layer importance comparison (next six months)*

| Embedding layers | | | | Metrics | |
|---|---|---|---|---|---|
| Event | Age | Visit | Position | APS | AUROC |
| **1** | **1** | **1** | **1** | **0.525** | **0.958** |
| 1 | 0 | 1 | 1 | 0.515 | 0.954 |
| 1 | 1 | 0 | 1 | 0.511 | 0.954 |
| 1 | 1 | 1 | 0 | 0.498 | 0.954 |
| 1 | 0 | 1 | 0 | 0.451 | 0.952 |
| 1 | 1 | 0 | 0 | 0.501 | 0.955 |
| 1 | 0 | 0 | 0 | 0.446 | 0.951 |
| 1: Embedding layer activated; 0: Embedding layer deactivated | | | | | |

#### 4.1.5.6 Subgroup analysis for the prediction of new incidence

Lastly, to shed light on the model performance for first incidence risk prediction, we calculated AUROC and APS for a subset of diseases in the label occurring for the first time.

Table 4.5 shows a consistent pattern that BEHRT substantially outperforms the other two models.

*Table 4.5: Model performance in risk prediction tasks – first incidence of disease*

| | APS\|AUROC | | |
|---|---|---|---|
| Model name | Next visit | Next six months | Next 12 months |
| **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 |

### 4.1.6 Discussion

In this work, we introduced a novel deep learning model for EHR called BEHRT. It can scale across various diseases and incorporate a wide range of events in EHR for modelling. BEHRT achieved superior performance and outperformed the best deep EHR models in the literature by more than 8% (absolute improvement in APS) for predicting the most likely diseases in one's future. The results were validated on CPRD, one of the largest longitudinally linked EHR databases.

In addition to predictive performance, BEHRT also illustrated its strength in learning meaningful representations without experts' engagement. For instance, the disease (i.e., event) embedding from the BEHRT can provide great insights into the disease associations that are largely aligned with experts' expectations. It goes beyond the co-occurrence and uses the concept of closeness to indicate the similarity of disease trajectory in a large population. Furthermore, without the inclusion of gender, BEHRT can make a robust and gender-specific prediction, implying that the representations learned by BEHRT are not strictly temporal but rather contextual.

These advantages are partly because of the architectural strength of the Transformer and the combination of four key concepts from EHR for modelling: event, age, visit, and position. This combination presents the model with a patient's comprehensive medical trajectory, which allows the model to not only learn about the inter-association between

events and the outcome, but also gain insights about the long-term intra-associations in the medical history and the underlying generating process of EHR.

Although this work considers diagnosis as the event for modelling, future work can also include other modalities such as medications, tests, and procedures to comprehensively represent a patient's medical trajectory. This would require minimal architectural changes by only expanding the events along with corresponding age, visit, and position embeddings.

## 4.2 Risk prediction in the presence of data shifts

This work has been published in European Heart Journal Digital Health at: https://academic.oup.com/ehjdh/advance-article/doi/10.1093/ehjdh/ztac061/6765054. The manuscript benefitted from the input of several co-authors. I am the first author and the work presented in this chapter is my own work. My role consisted in designing the study, conceiving the experiments and analyses, conducting the literature searches, data extraction and cleaning, and writing the manuscript.

### 4.2.1 Motivation

Risk prediction models are important tools to guide clinical decision-making in routine health care. They can help clinicians to identify at-risk patients and initiate preventive measures. Most of the current risk models in use rely on (simple) conventional statistical models, for example, QRISK3 [112], Framingham [113], and Assign [114]. These are the most commonly used models for the prediction of cardiovascular diseases (CVD). However, the limitations of these models are obvious. First, they largely rely on the exert-selected predictors. The usefulness and comprehensiveness of the predictors not only depend on the understanding of a disease, but they are also a key factor that determines the performance of a risk model. Furthermore, these models assume a linear relationship between the predictors and the outcome. This overly simplistic assumption can limit the predictive power and accuracy of the risk models in some applications.[104,115,116]

Over the past decade, deep learning models have gained growing popularity due to their ability to learn high-level features that capture the complex interactions of input variables. As demonstrated by BEHRT and many other works,[104,117] deep learning models can achieve state-of-the-art predictive performance across different fields without explicitly incorporating the experts' knowledge. Indeed, these models have a great potential to improve prediction performance. However, their usability in clinical practice would require

a comprehensive assessment of the advantages of deep learning models compared to the established clinical models.

### 4.2.2 Related work

Several previous studies demonstrated that even simple machine learning models could outperform statistical models and suggested their advantages in improving healthcare.[118,119] But some have argued that beyond internal validations that are prone to be optimistic and are not representative of the practical use cases of risk models, evidence of achieving higher performance using deep learning and machine learning models is lacking. Other studies have shown that simpler statistical models can perform better or as well as machine learning models.[120,121] These contradictory findings are partly because, in the absence of benchmarking datasets, the data used in different studies – in terms of size and quality – may not always merit complex modelling approaches. More importantly, model evaluation methods and metrics vary across different studies. Lastly, deep learning models encompass a wide range of architectures and depending on the data, the benefits of different architectures can vary significantly.[90,117]

### 4.2.3 Aims

This research aimed to comprehensively evaluate the proposed BEHRT model with widely used clinical models (QRISK3[112], Framingham[122,123], and ASSIGN [112,114]) for the prediction of major cardiovascular events in the general population. In addition to internal validation, we investigated model performance under data shifts on external cohorts (1) from different geographies and (2) from different time periods.

### 4.2.4 Methods

#### 4.2.4.1 Data source

This study was conducted using EHR from CPRD linked to HES and ONS. Primary data was available from 1 January 1985 to 31 December 2015, and the linkage was available from 1 January 1998 to 30 September 2015.

#### 4.2.4.2 Study design

This work mainly investigated the five-year risk prediction of the incidence of three major cardiovascular events (HF, stroke, and coronary heart disease (CHD)). The baseline date of each patient for risk prediction was determined by randomly selecting a date during the eligible record period. We created three sub-datasets based on the selected baseline date for these three tasks, respectively. All datasets followed the same protocol for data preparation. For the convenience of description, we use HF risk prediction as an example. For the HF dataset, we filtered out all patients with prevalent HF before the baseline. A patient was identified as HF (+) if diagnosed with HF from either general practices and HES records or death registration within a five-year window after the baseline date. Moreover, a patient was identified as HF (-) if one had records for at least five years after the baseline, and no HF was identified during the five-year window. The rest of the patients were considered lost to follow-up before developing HF by year five with an uncertain outcome (i.e., censored patients).

#### 4.2.4.3 Eligible participants

In this study, we included four million patients who contributed to data between 1985-01-01 and 2015-12-31, were linked to HES, and marked as “acceptable” by CPRD quality control. Furthermore, we selected a cohort with both men and women, aged at least 35 years old, registered with general practices for at least two years, and with the valid IMD recorded. This led to a cohort of 3,052,290 patients.

#### 4.2.4.4 Outcomes

The diagnosis codes for identifying HF, stroke, and CHD were adapted from the CALIBER code repository[105]. In brief, HF was defined as a composite condition of congestive HF, left ventricular failure, cor pulmonale, cardiomyopathy, hypertensive heart disease with (congestive) HF, cardiac failure, and HF; stroke was defined as a composite condition of ischaemic stroke, transient ischaemic attack, and stroke not otherwise specified; and CHD was defined as a composite condition of atherosclerotic heart disease, coronary or ischaemic heart disease, aborted myocardial infarction, or coronary artery disease.

#### 4.2.4.5 Study population

Overall, 3,052,290 patients contributed to CPRD between 1985-01-01 and 2015-12-31, can be linked to HES, had IMD recorded, aged at least 35, and registered with the general practices for at least two years. With the exclusion of censored patients and those with prevalent HF, stroke, and CHD before the baseline for the prediction of incident HF, stroke, and CHD, respectively, we identified 954,983, 983,086, and 1,003,554 patients for HF, stroke, and CHD datasets, respectively, which were substantially smaller than the overall cohort.

#### 4.2.4.6 Predictors

Six sets of predictors were included for risk prediction. Two of them were adopted from QRISK and ASSIGN,[112,114] respectively, and they were used for all three risk prediction tasks. As suggested by the previous research, the Framingham model used three different sets of predictors for HF[122,123], stroke [124], and CHD [113] risk prediction, respectively. Another set of predictors was used for training the BEHRT models. Table 4.6 presents the complete list of predictors selected by the clinical models.

*Table 4.6: Summary of predictors for risk prediction. QRISK and ASSIGN use the presented predictors for all prediction task, while the Framingham model use three modified sets of predictors for three prediction tasks, respectively. BEHRT follows a different philosophy that incorporates all available information in the medical history for modelling, therefore, not presented in the table.*

| Predictor | QRISK | Framingham | | | ASSIGN |
|---|---|---|---|---|---|
| | | HF | Stroke | CHD | |
| Age | * | * | * | * | * |
| Ethnicity | * | | | | |
| Indices of multiple deprivations | * | | | | * |
| Systolic blood pressure | * | * | * | * | * |
| SD of systolic blood pressure | * | | | | |
| Body mass index | * | | | | |
| Total cholesterol/HDL ratio | * | | | | * |
| Smoking status | * | | * | * | * |
| Family history of CHD | * | | | | * |
| Diabetes | * | * | * | | * |
| Treated hypertension | * | | * | * | |
| Rheumatoid arthritis | * | | | | * |
| Atrial fibrillation | * | | * | | |
| CKD (stage 3,4,5) | * | | | | |
| Migraine | * | | | | |
| Corticosteroid use | * | | | | |
| Systemic lupus erythematosus | * | | | | |
| Atypical antipsychotic drugs | * | | | | |
| Severe mental illness | * | | | | |
| HIV or AIDS | * | | | | |
| Erectile dysfunction | * | | | | |
| Sex | | * | | * | * |
| CHD | | * | * | | |
| Left ventricular hypertrophy | | * | | | |
| Valve disease | | * | | | |
| Heart rate | | * | | | |
| Total cholesterol | | | | * | |
| HDL | | | | * | |
| HDL: High-density lipoprotein cholesterol; CKD: chronic kidney disease; SD: standard deviation | | | | | |

The demographic factors, clinical diagnoses, medications, and clinical values were extracted from the general practice and HES records. We identified all clinical diagnoses, except left ventricular hypertrophy, using a list of codes adapted from the CALIBER code repository.[105] The predictors were defined the same way as described in the QRISK3[112]. The left ventricular hypertrophy was identified using left ventricular hypertrophy verified by electrocardiogram test or diagnostic codes for left ventricular hypertrophy (Table A.4). We identified clinical values using records within a one-year window before the baseline and calculated the mean for two or more recorded values. Regarding data missingness, clinical diagnoses and medications were assumed to be present only if recorded. Smoking status and

ethnicity were marked as unknown if not recorded. Missing data for body mass index (76%), systolic blood pressure (56%), total cholesterol to high-density lipoprotein cholesterol ratio (85%), and heart rate (96%) were imputed five times using multivariable imputation with chained equations.[125]

To prepare data for BEHRT, we incorporated records from diagnoses, medications, lab tests, and procedures before each patient's baseline. Diagnosis codes from primary care were mapped to ICD-10 using a phenotyping dictionary proposed by Hassaine et al.[126]. Compared to the CALIBER phenotyping, which mapped diagnosis codes to 301 disease categories, this approach can better preserve information about diseases in the original granularity and avoid potential bias in the phenotyping due to the expert's preference. More specifically, all Read codes were mapped to ICD-10 using two vocabularies provided by NHS digital[127] and SNOMED-CT[128]. The NHS digital vocabulary had higher priority than the SNOMED-CT wherever a conflict occurred. Afterwards, we kept diagnosis codes at ICD-10 level four. For medications, we kept codes at the BNF section level. The originally recorded OPCS codes were kept for the procedures. The Read codes were preserved for the lab tests came from the CPRD. This procedure led to 3858, 390, 1439, and 679 unique medical codes for diagnoses, medications, lab tests, and procedures, respectively. Additionally, age in year (with maximal age of 110 years old) was used for age embedding.

#### 4.2.4.7 Model derivation

This study considered three types of models for comparison – statistical models, machine learning models, and deep learning models. Statistical models were CPH models implemented by "Lifelines"[129] that relied on various abovementioned clinical validated predictors. The same predictors were used for the machine learning models, which are random forest (RF) models in this study. Because current software packages of survival RF cannot handle large data very well to our best knowledge, we use the classic RF models

from “scikit-learn”[130] instead. We named each model based on the predictor and the model type to distinguish different model and predictor pairs. For instance, we named the CPH model based on the QRISK predictor QRISK and the RF model based on the QRISK predictor RF (QRISK). For the deep learning models, we followed the identical procedure described in section §4.1 to implement BEHRT. However, instead of predicting multiple outcomes, we used a feed-forward layer with a sigmoid activation function as the classifier to map the aggregated patient representation (final hidden state of “CLS”) after the pooling layer to predict a single outcome. We performed hyper-parameter tuning for machine learning and deep learning models by randomly searching a parameter space for 30 iterations. We chose the optimal hyperparameters that achieve the best accuracy on a small subset of randomly selected samples held-out in the training dataset for validation purpose. For brevity, we only reported the model with the best performance (details can be found in Appendix Table A.5 and Table A.6).

Regarding training, BEHRT was implemented with PyTorch[131] and trained using Adam optimiser[107] with a three-stage learning rate scheduler over 100 epochs on 2 GPUs. The scheduler included 10%, 40%, and 50% epochs for warm-up, hold, and cosine decay, respectively. We used batch size 32, hold learning rate $5e^{-5}$ with 4-step gradient accumulation. The early-stopping strategy was also included for training, and we stopped training once the training loss did not decrease for five epochs.

#### 4.2.4.8 Model validation methods

We employed a range of model evaluation procedures to compare the performance of models beyond the narrowly defined internal validation procedures (Figure 4.3). In addition to a standard five-fold cross-validation (or “internal validation”), we investigated three “external validation” approaches: out-of-region validation, temporal (i.e., time-shift) validation, and time series cross-validation.

**Out-of-region validation**: We followed the recommendations[89] and chose different regions (i.e., strategic health authorities) for forming the training and validation datasets (i.e., regions in the validation dataset were not part of the training data). For validation, we selected three regions with different socioeconomic statuses[89] from north England that account for 23% of the total population. The remaining regions were used for training. We argue that geographically different patients with different statistical characteristics are similar to using an external dataset with a certain level of data shift.[89]

**Temporal validation:** The changes in demographics, care policy, and practice over the years can introduce data shifts.[132] Therefore, we trained and validated models on cohorts from different years. More specifically, since a baseline has been chosen for each individual, all those with baseline years of 2000 and earlier were used for training models, and those with baseline years between 2000 to 2010 were used for validation (Figure 4.3). Note that each patient contributes only once to either training or validation.

**Time series cross-validation:** In routine clinical settings, a risk prediction model can only be trained with the data in the past to predict an event in the future. Assume we wanted to use a risk model in 2005; this means we need a model that has been trained and validated with data before 2005. To assess the models' generalisation power in settings similar to their typical use case, and over a period of time, our third external validation performed "time-series cross-validation"[133]. In this approach, we split the data into three time windows, denoted as $w_{train}$, $w_{label}$, and $w_{val}$. For example, we trained a model using patients with baseline years before 2000 ($w_{train}$), left a five-year gap (2000-2005) for labelling ($w_{label}$), and validated ($w_{val}$) on patients with baseline one year after the labelling window (2005-2006). The time-series cross-validation repeated the abovementioned process by moving $w_{train}$ and $w_{val}$ from 2000 to 2004, and 2006 to 2010, respectively, with a one-

year interval. This procedure also mimics using newly collected patients to update the risk model annually.

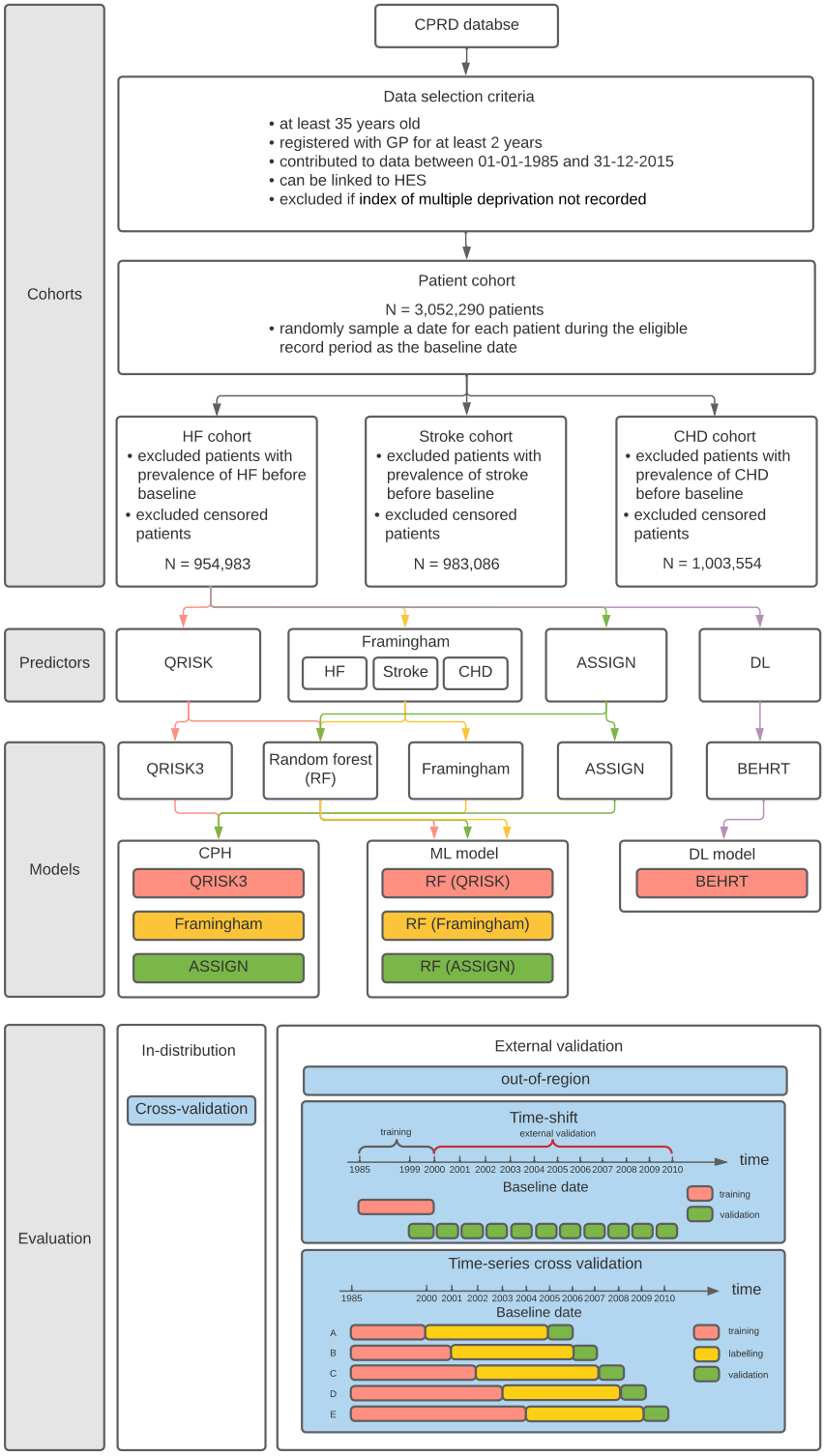



*Figure 4.3: Flow diagram of the cohort selection, modelling, and evaluation. The colour in predictor and modelling section represents a path that flows from the cohort to a specific set of predictors and how the predictors are used by the models.*

#### 4.2.4.9 Validation metrics and statistical analysis

While our outcome was binary, instead of predicting positive or negative, we predicted the probability of a patient that belonged to positive or negative. This allows us to measure the discrimination of models using AUROC and APS. The model calibration was assessed with the calibration curves[134]. The deep learning and machine learning model predict the estimated risk naturally within the binary classification framework. However, for the CPH survival models, we combined the regression coefficients as the weight with the baseline survival function. It allowed us to estimate the risk for each year of follow-up, and we only focused on the 5-year risk estimation.[112] This allows us to calculate the AUROC, APS, and calibration curves that are directly comparable with the binary classification models. Additionally, the precision-related metrics (e.g., APS) were not considered for evaluation in the presence of data shifts (i.e., external validations) as they are sensitive to the changes in event rate. Naturally, the event rate can change across time for all outcomes; therefore, these metrics are not comparable in this case.

To provide additional information about how the time shift influence the risk prediction, we analysed the incidence rate of HF, stroke, and CHD across different years on the selected cohort. Furthermore, we used population stability index[135] (PSI) to identify the covariates distribution shift between the training and validation dataset.

### 4.2.5 Results

#### 4.2.5.1 Population statistics

Due to the considerable overlap of cohorts across three risk prediction tasks, we summarised the baseline characteristics of the selected cohorts after data imputation to simplify the presentation. Table 4.7 shows that the overall study population included 1,096,275 patients, and HF has a lower percentage of positive cases than stroke and CHD. For the out-of-region cohort, expectedly, the training and validation cohorts were different

in some respects, such as ethnicity, age, smoking status, and family history of CHD. The proportions of men and women were comparable, but the IMD in the validation cohort was slightly higher than in the training cohort.

*Table 4.7: Baseline characteristics of risk prediction cohorts*

| | General cohort | Out-of-region | |
|---|---|---|---|
| Predictor | | Training | Validation |
| No. of patients | 1,096,275 | 838,961 | 257,314 |
| No. of patients (% of positive cases) for HF cohort | 954,983 (3.3) | 729,613 (3.3) | 225,370 (3.6) |
| No. of patients (% of positive cases) for Stroke cohort | 983,086 (6.7) | 752,321 (6.8) | 230,765 (6.8) |
| No. of patients (% of positive cases) for CHD cohort | 1,003,554 (10.3) | 769,304 (10.2) | 234,250 (10.9) |
| Sex, n (%) | | | |
| Male | 534,657 (48.7) | 409,522 (48.8) | 125,105 (48.6) |
| Female | 561,618 (51.2) | 429,409 (51.2) | 132,209 (51.4) |
| Age, mean (SD) | 58.0 (15.3) | 58.2 (15.4) | 57.3 (15.0) |
| IMD score, mean (SD) | 2.7 (1.4) | 2.6 (1.3) | 3.1 (1.4) |
| Family history of CHD, n (%) | 341,042 (31.1) | 246,416 (29.4) | 94,626 (36.7) |
| BMI, mean (SD) | 27.6 (3.2) | 27.6 (3.2) | 27.7 (3.2) |
| Strategic health authority (region), n (%) | | | |
| North East | 27,195 (2.5) | - | 27,195 (2.5) |
| North West | 175,416 (16.0) | - | 175,416 (16.0) |
| Yorkshire and the Humber | 54,655 (5.0) | - | 54,655 (5.0) |
| East Midlands | 38,242 (3.5) | 38,242 (3.5) | - |
| West Midlands | 137,824 (12.6) | 137,824 (12.6) | - |
| East of England | 134,314 (12.3) | 134,314 (12.3) | - |
| South West | 130,170 (11.9) | 130,170 (11.9) | - |
| South Central | 136,189 (12.4) | 136,189 (12.4) | - |
| London | 126,175 (11.5) | 126,175 (11.5) | - |
| South East Coast | 136,047 (12.4) | 136,047 (12.4) | - |
| Ethnicity, n (%) | | | |
| Unknown | 683,961 (62.4) | 536,157 (63.9) | 147,804 (57.4) |
| White | 399,533 (36.4) | 291,908 (34.8) | 107,625 (41.8) |
| Other Asian | 1,016 (0.1) | 905 (0.1) | 111 (0.0) |
| Pakistani | 1,237 (0.1) | 802 (0.1) | 435 (0.2) |
| Indian | 3,070 (0.3) | 2,794 (0.3) | 276 (0.1) |
| Other | 3,162 (0.3) | 2,603 (0.3) | 559 (0.2) |
| Caribbean | 1,659 (0.2) | 1,562 (0.2) | 97 (0.0) |
| Mixed | 790 (0.1) | 697 (0.1) | 93 (0.0) |
| Bangladeshi | 338 (0.0) | 283 (0.0) | 55 (0.0) |
| Chinese | 656 (0.1) | 495 (0.1) | 161 (0.1) |
| Black African | 853 (0.1) | 755 (0.1) | 98 (0.0) |
| Smoking status, n (%) | | | |
| Not recorded | 776,927 (70.9) | 596,842 (71.1) | 180,085 (70.0) |
| Non-smoker | 154,991 (14.1) | 119,100 (14.2) | 35,891 (13.9) |
| Ex-smoker | 106,512 (9.7) | 81,872 (9.8) | 24,640 (9.6) |
| Light smoker (<10 cigarettes/day) | 15,576 (1.4) | 11,436 (1.4) | 4,140 (1.6) |
| Moderate smoker (10-20 cigarettes/ day) | 23,269 (2.1) | 16,619 (2.0) | 6,650 (2.6) |
| Heavy smoker (>20 cigarettes/day) | 19,000 (1.7) | 13,092 (1.6) | 5,908 (2.3) |
| Clinical values | | | |
| Systolic blood pressure, mean (SD) | 135.3 (13.0) | 135.4 (13.1) | 135.3 (12.9) |
| Cholesterol/HDL | 3.99 (0.67) | 3.98 (0.69) | 4.00 (0.63) |
| Comorbidity, n (%) | | | |
| Diabetes | 38,972 (3.6) | 29,478 (3.5) | 9,494 (3.7) |
| Rheumatoid arthritis | 7,507 (0.7) | 5,584 (0.7) | 1,923 (0.7) |

| | | | | |
|---|---|---|---|---|
| | Atrial fibrillation | 34,137 (3.1) | 26,804 (3.2) | 7,333 (2.8) |
| | Chronic kidney disease | 6,727 (0.6) | 5,222 (0.6) | 1,505 (0.6) |
| | Migraine | 34,789 (3.2) | 25,872 (3.1) | 8,917 (3.5) |
| | Severe mental illness | 6,113 (0.5) | 4,460 (0.5) | 1,653 (0.6) |
| | Systemic lupus erythematosus | 947 (0.1) | 667 (0.1) | 280 (0.1) |
| | HIV or AIDS | 1,755 (0.2) | 1,316 (0.2) | 439 (0.2) |
| | Erectile dysfunction | 27,885 (2.5) | 21,421 (2.6) | 6,464 (2.5) |
| Prescribed medication, n (%) | | | | |
| | Treated hypertension | 233,056 (21.3) | 177,009 (21.1) | 56,047 (21.8) |
| | Antipsychotic | 5,778 (0.5) | 4,381 (0.5) | 1,397 (0.5) |
| | Corticosteroid | 44,388 (4.0) | 32,983 (3.9) | 11,405 (4.4) |
| HDL, high-density lipoprotein; IMD, Index of Multiple Deprivation; SD, standard deviation | | | | |

### 4.2.5.2 Internal validation

*Table 4.8: Mean and 95% confidence interval of AUROC and APS from 5-fold internal cross-validation.*

| | AUROC (95% confidence interval) | | |
|---|---|---|---|
| Models | HF | Stroke | CHD |
| **BEHRT** | **0.954 (±0.004)** | **0.957 (±0.002)** | **0.951 (±0.002)** |
| QRISK3 | 0.895 (±0.002) | 0.874 (±0.005) | 0.838 (±0.005) |
| RF (QRISK) | 0.897 (±0.005) | 0.877 (±0.005) | 0.850 (±0.003) |
| ASSIGN | 0.885 (±0.002) | 0.858 (±0.002) | 0.829 (±0.001) |
| RF (ASSIGN) | 0.884 (±0.005) | 0.859 (±0.003) | 0.833 (±0.002) |
| Framingham | 0.883 (±0.004) | 0.869 (±0.005) | 0.831 (±0.005) |
| RF (Framingham) | 0.884 (±0.002) | 0.868 (±0.004) | 0.836 (±0.003) |
| | APS (95% confidence interval) | | |
| **BEHRT** | **0.801 (±0.020)** | **0.876 (±0.011)** | **0.897 (±0.013)** |
| QRISK3 | 0.270 (±0.002) | 0.435 (±0.007) | 0.463 (±0.002) |
| RF (QRISK) | 0.313 (±0.031) | 0.471 (±0.008) | 0.487 (±0.006) |
| ASSIGN | 0.230 (±0.006) | 0.360 (±0.004) | 0.386 (±0.005) |
| RF (ASSIGN) | 0.278 (±0.007) | 0.368 (±0.008) | 0.398 (±0.008) |
| Framingham | 0.270 (±0.002) | 0.451 (±0.007) | 0.431 (±0.007) |
| RF (Framingham) | 0.302 (±0.016) | 0.452 (±0.007) | 0.442 (±0.006) |

Table 4.8 shows the model performance of the statistical, machine learning, and deep learning models for HF, stroke, and CHD risk prediction. The BEHRT model substantially outperformed the other statistical and machine learning models for all risk prediction tasks, achieving AUROC 0.954, 0.957, and 0.951; APS 0.801, 0.876, and 0.897 for HF, stroke, and CHD risk prediction, respectively. Compared to the best performing clinical models, BEHRT improved AUROC 5.8%, 8.2%, and 11.3%, APS 53.1%, 42.5%, and 43.4%, for HF, stroke, and CHD, respectively. However, machine learning models showed similar or slightly better performance than the statistical models with the same predictors. Furthermore, Figure 4.4 shows calibration curves, indicating that BEHRT and QRISK models,

particularly BEHRT, generally have better calibration performance than the other models, especially for the high-risk predictions, in all tasks.

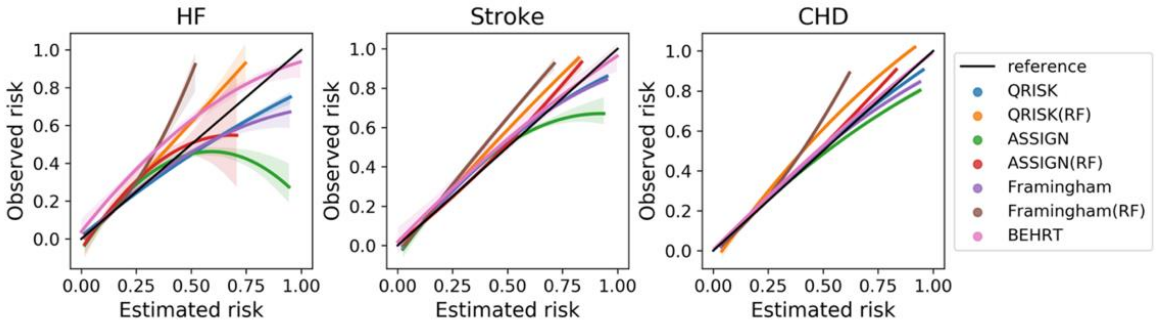



*Figure 4.4: Mean and 95% confidence interval of calibration curve from 5-fold internal cross-validation*

#### 4.2.5.3 Out-of-region validation

Table 4.9 further shows that all models show a slight decline in AUROC compared to the internal validation for the out-of-region validation; BEHRT showed more degradations than the other machine learning and statistical models. On average, the decrease in AUROC was less than 1.5% for both machine learning and statistical models and around 3% for the BEHRT models. Yet, the BEHRT models still achieved the best performance for all risk prediction tasks.

*Table 4.9: External validated model performance on a cohort with patients from the non-random selected regions.*

| | AUROC (absolute decline compared to the internal performance) | | |
|---|---|---|---|
| Models | HF | Stroke | CHD |
| **BEHRT** | **0.909 (-0.044)** | **0.932 (-0.025)** | **0.929 (-0.022)** |
| QRISK3 | 0.883 (-0.012) | 0.865 (-0.009) | 0.830 (-0.008) |
| RF (QRISK) | 0.883 (-0.014) | 0.866 (-0.011) | 0.840 (-0.010) |
| ASSIGN | 0.873 (-0.012) | 0.852 (-0.003) | 0.823 (-0.006) |
| RF (ASSIGN) | 0.874 (-0.010) | 0.853 (-0.006) | 0.827 (-0.006) |
| Framingham | 0.871 (-0.012) | 0.862 (-0.007) | 0.821 (-0.010) |
| RF (Framingham) | 0.873 (-0.011) | 0.855 (-0.013) | 0.826 (-0.010) |

#### 4.2.5.4 Temporal validation

Figure 4.5 shows the AUROC of all models when predicting the outcome for patients with external baseline. In general, all models suffered a performance decline under temporal data shifts, and the AUROC values from all models decreased as the gap between the reference (i.e., 1999-2000) and the baseline of the validated cohort increased. Noticeably,

HF and stroke had a worse decline compared to CHD. Furthermore, there was no substantial difference between statistical models (i.e., Framingham, QRISK, and ASSIGN) and the machine learning models. The QRISK model, in general, has the best performance among the statistical models. Additionally, we see the BEHRT and machine learning models have a more considerable decline between internal (i.e., 1999-2000) and external (i.e., 2000-2010) performance.

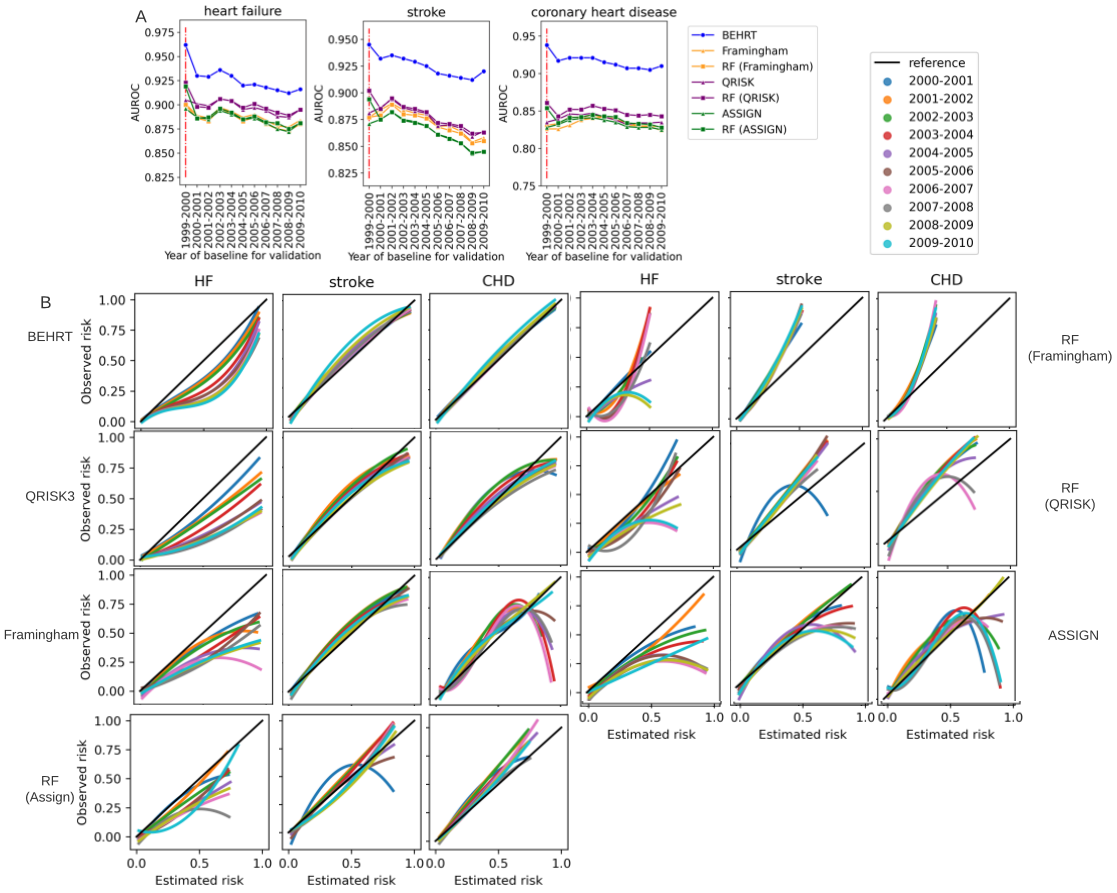



*Figure 4.5: Model performance validated under temporal data shifts. A: Validation of discrimination. B: Validation of calibration. All models were trained on patients who had baseline before 2000 and validated on patients who had baseline between 2000 to 2010 with one-year interval. The internally validated AUROC (red line) is provided for comparison in A; and the black line in B is the reference calibration curve. Colours in A represent the models for validation and the colours in B represent the year of baseline for validation.*

Additionally, Figure 4.5B shows the calibration curve of models under temporal data shifts. The result indicates that the temporal data shifts can negatively affect the calibration of models. These negative effects are more pronounced in the calibration of the HF risk prediction models relative to stroke and CHD. However, BEHRT models still achieved the

best calibration among all models for all three risk prediction tasks. This was followed by QRISK. The two models substantially outperformed the others. An important difference between the BEHRT and QRISK models pertains their performance in the high-risk region of the calibration curve. The QRISK models have relatively poor calibration in this region, especially under temporal data shift.

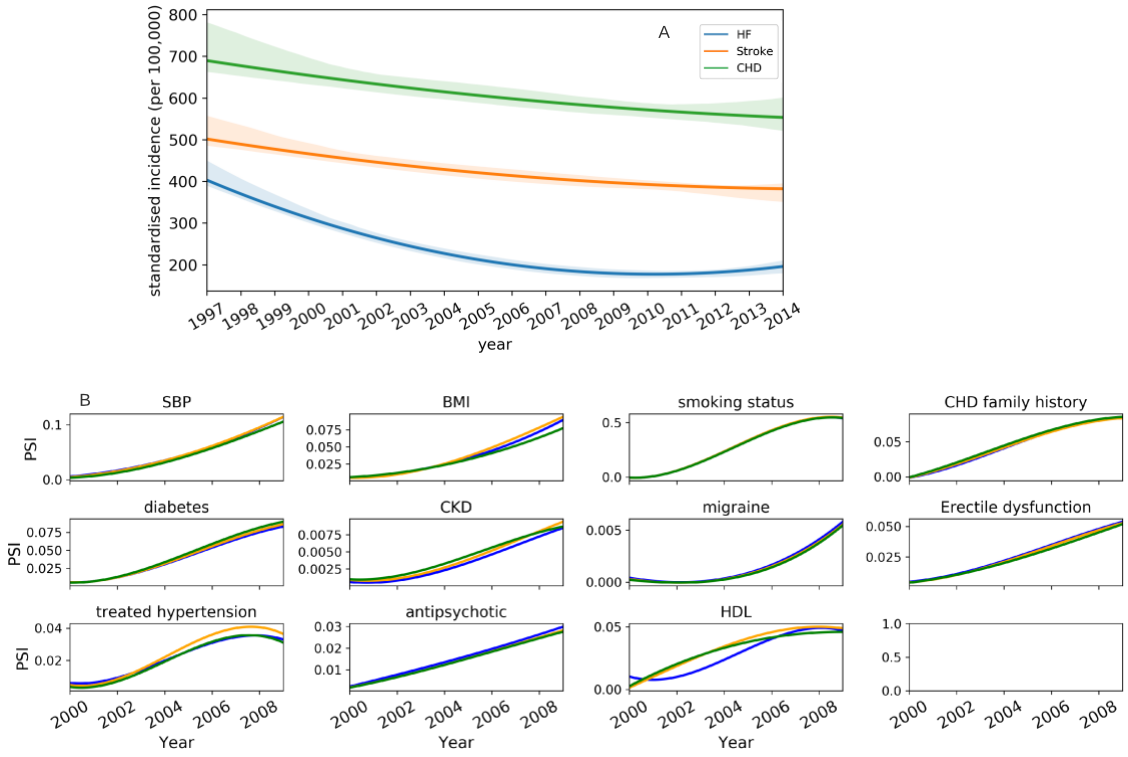



*Figure 4.6: Analysis of data shift in the covariates and the outcome. A. Temporal trends of incidence rate presented with fitted local polynomial regression lines with 95% confidence interval. The presented period covers the 95% patients' baseline in the selected cohort. B: PSI of predictors between training and temporal validation set (predictors with SPI>0·001 are presented). The colours in both A and B represent the diseases.*

Data shifts were commonly classified into three categories: covariate shift (i.e., distribution shift in the covariates), prior probability shift (i.e., distribution shift in the outcome), and concept shift (i.e., the change of relation between covariates and the outcome).[136] We specifically investigated the prior probability shift and covariate shift. As shown in Figure 4.6A, the incidence rate of HF substantially changes across different year (prior probability shifts), but it is not the case for CHD and stroke. Therefore, the prior probability shifts can be a potential reason that causes the calibration shift in HF in Figure 4.5. Figure 4.6B shows the covariate shifts in general are not substantial (i.e., PSI <0·1) for

all three diseases, except systolic blood pressure and smoking status. Considering the performance decline as shown in Figure 4.5A, the concept shift can potentially play an important role in temporal data shift, and it suggested that the relation between covariates and the outcome can change over time.

#### 4.2.5.5 Time-series cross-validation

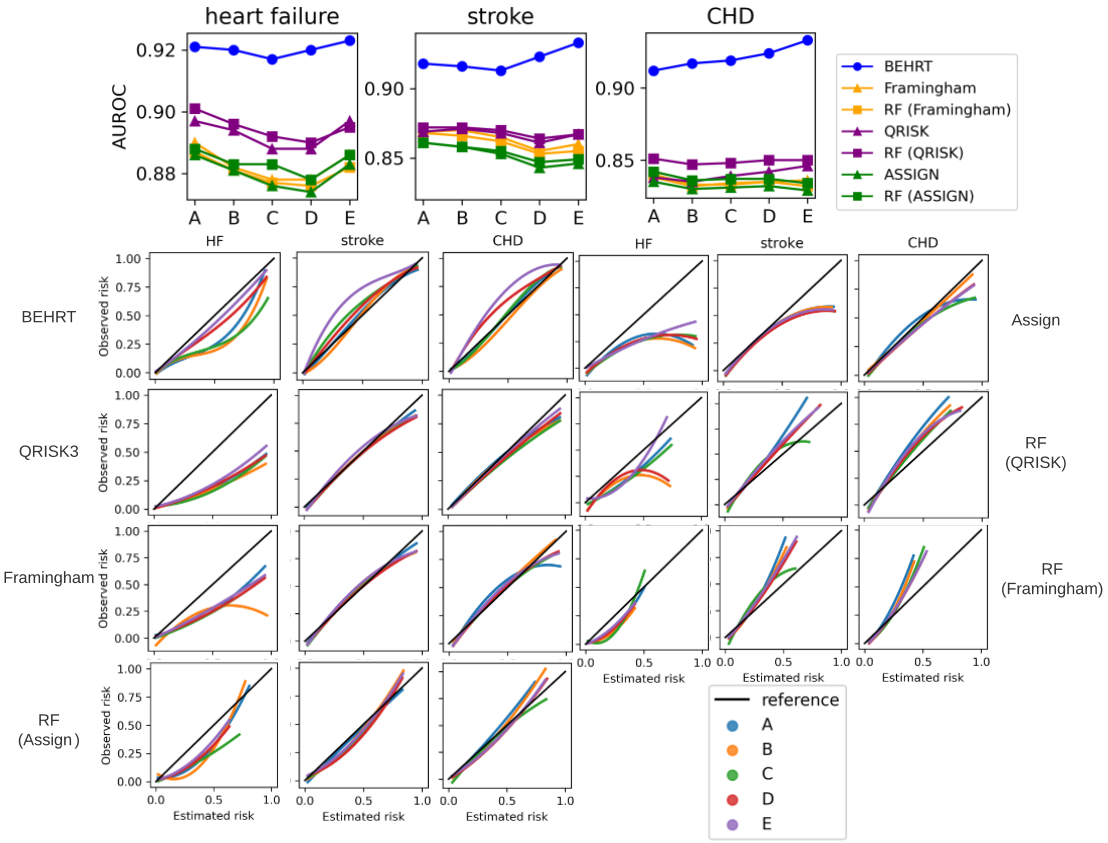



*Figure 4.7: AUROC and calibration curve using time-series cross-validation. A, B, C, D, and E represent experiments with $w_{train}$ from 2000 to 2004 and $w_{eval}$ from 2006 to 2010, respectively (Figure 4.3)*

Figure 4.7 shows the AUROC is relatively stable with less than 2% changes for all risk prediction tasks across different time, therefore, suggesting that all models benefit from using updated information to prevent performance decline in AUROC. However, this has negligible impact on the calibration curve. This is partly because using the updated information can mitigate concept shift (i.e., the relation between the covariates and the outcome), but not the prior probability shift (i.e., the distribution shift in the outcome).

Therefore, the result suggests that updating model only improves the calibration when the prior probability shift is not severe

### 4.2.6 Discussion

In this study, we showed that in large cohorts with comprehensive medical information, deep learning models outperformed machine learning and statistical models in predicting HF, CHD, and stroke. This was to be expected. A sequential deep learning model mapped the entire history of individuals to risk profiles, whereas machine learning and statistical models relied on limited, expert-selected predictors and were oblivious to the temporal dimension of medical events. Our findings also corroborated similar studies that reported no or minor differences between machine learning models and statistical models.[121,137] Combined these findings underlined that additional functional complexity, as seen in RF models compared to statistical models, had a minor effect on CVD risk predictions. Some recent studies claimed that the minimal performance gain in the machine learning models (e.g., RF) is due to the limited access to the EHR, and those models can benefit from having more comprehensive features.[138] However, even with the presence of the entire EHR, machine learning models still have to rely on the expert-knowledge for predictor selection, which is still the main obstacle to further improve the performance. On the contrary, incorporating the temporal dimension of medical events and capturing the entire trajectory of patients, as seen in deep learning compared to machine learning and statistical models, leads to major improvements.

Additionally, we evaluated the models under data shifts and showed that the performance of all models declined compared to validation with a random train-test split. Although deep learning maintained the best performance, it had the largest performance decline compared to the statistical models when the test set was from a distinct region. This is because deep learning models can learn the in-distribution associations very well, which,

however, inevitably may compromise its generalisability when the target population is mismatched (e.g., out-of-distribution datasets and datasets in the presence of data shifts).[139] People may interpret such observation as deep learning models are more prone to overfitting. However, an overfitted model would not work well even for the in-distribution population, which is not the case in our study. Therefore, we conclude deep learning risk model would require more careful finetune when the target population substantially changes. However, it does not necessarily represent overfitting.

Similarly, the accuracy of all models declined as the time gap widened between the training and validation cohorts, especially for HF and stroke risk prediction. We showed that this could be remedied by regularly updating the model. Almost all models retained performance across different periods when updated every year. Our results also suggested that the miscalibration under temporal shifts can potentially link to the prior probability shift: the bigger the prior probability shift, the larger the calibration shift (e.g., HF). Model updating can only mitigate the concept shift but not the prior probability shift. Therefore, its usefulness in alleviating the calibration shift is limited.

The present study also has several limitations. We focused on the risk of CHD, HF, and stroke. As such, the conclusions may not generalise to other diseases. Another limitation is that censored patients were excluded from the analysis. Despite the importance of the survival framework in risk prediction, most deep learning models remained binary classification models, thus, unable to handle the censored patients. To this end, some commonly used approaches are (I) discarding those patients, and (II) considering them as event free.[121,140] The latter can substantially underestimate the risk in classification models but discarding censored patients will lead both survival and binary classification models to a similar risk prediction task and produce a more comparable result.[121] Although this can yield a less representative cohort, we believe the results are still valid in terms of model

comparison with the presence of data shifts. Our future work will explore the impact of survival framework versus binary classification framework on risk prediction and the benefit of transforming deep learning binary classification models to deep learning survival models.

## 4.3 Risk prediction using comprehensive electronic health records

This work has been published in IEEE Journal of Biomedical and Health Informatics at: https://ieeexplore.ieee.org/document/9964038. The manuscript benefitted from the input of several co-authors. I am the first author and the work presented in this chapter is my own work. My role consisted in designing the study, conceiving the experiments and analyses, conducting the literature searches, data extraction and cleaning, and writing the manuscript.

### 4.3.1 Motivation

EHR provide up-to-date and comprehensive information about patients. It allows clinicians to assess the entire patient journey and presents what is available in clinical practice.[141] However, most of the existing works applied to EHR have mainly relied on only a small fraction of information available in the datasets (typically diagnoses and medications from hospitals).[142] This would typically involve maximally a few hundred records from a patient.[104,143,144] A realistic demand of using more comprehensive information to represent a patient's medical history and capturing the whole history of medical encounters is expected to lead to more accurate predictions. However, this can substantially increase the number of available records and expand the EHR length for modelling. Therefore, adapting a risk prediction model to handle patients with thousands of records or even longer EHR sequences and avoid missing important historical information is highly desired but remains under-investigated.

### 4.3.2 Related work

Previous studies demonstrated that the complexity of the Transformer-based model grows quadratically as the sequence length grows.[37] This prevented it from applying to sequential data with a sequence length of more than 512.[38] Therefore, it also prohibited the utility of BEHRT for processing comprehensive large-scale EHR.

Recent research has proposed two potential solutions to assist the Transformer-based model in processing long sequences. One is to use sparse attention to replace the classic self-attention mechanism[145–147] and the other is to use a hierarchical model structure to firstly extract local temporal features that are substantially shorter than the raw sequential data, then feed the local features into the higher-level Transformer for further global feature extraction and prediction.[148] Both methods can dramatically reduce the model complexity and adapt them to better handle data with longer sequences.

### 4.3.3 Aims

In this work, considering that medical events naturally have stronger local correlations (i.e., the closer two events are in time, the more likely they are also semantically related), we aimed to employ the hierarchical model architecture[148] to enhance the proposed BEHRT model to investigate the benefit of incorporating rich and complete patient's medical trajectory for risk prediction. We refer to this model as hierarchical BEHRT or Hi-BEHRT.

### 4.3.4 Methods

#### 4.3.4.1 Data source

This study was conducted using EHR from CPRD linked to HES and ONS. Primary data was available from 1 January 1985 to 31 December 2015, and the linkage was available from 1 January 1998 to 30 September 2015.

#### 4.3.4.2 Study design

In this work, we conducted risk models to predict incident HF, diabetes, CKD, and stroke. The incidence of a condition was defined as the first record in primary care or hospital admission records from any diagnostic position. To do so, two important dates were identified for each individual – an incident date of the event of interest and a baseline date. The risk prediction task uses all available records before the baseline date as a learning

period to predict the 5-year risk of an event of interest after it. Due to the known inaccuracies in recording the timing of events in EHR, we also ignored the events that occurred within a 1-year window after the baseline date. We achieved this by firstly identifying the incident date for all positive cases and then randomly selecting a baseline date within a 1-year to a 5-year window before the incident date. For patients who did not have recorded events of interest, we considered them as negative patients and randomly selected a baseline date for each of them with a guarantee of having at least five years of records after the baseline date. This would ensure that all negative patients were absolutely free from the event of interest during the predicting period.

#### 4.3.4.3 Eligible participants

In this study, we selected a cohort with patients who contributed to data between 1985-01-01 and 2015-12-31, were linked to HES, and marked as "acceptable" by CPRD quality control. This led to 4 million patients. Afterwards, we selected patients with both men and women, aged at least 16 years old and excluded patients with less than three years of records for the learning period (i.e., before the baseline date). The procedures for cohort selection and study design led to 1,975,630, 1,885,924, 1,889,925, and 1,730,828 patients for HF, diabetes, CKD, and stroke risk prediction, respectively. Among the identified patients, we split them into 60%, 10%, 30% for training, tuning, and validation, respectively.

#### 4.3.4.4 Outcomes

In this work, we defined the prediction objectives as HF, diabetes, CKD, and stroke. To be explicit, HF was defined as a composite condition of rheumatic HF, hypertensive heart disease with (congestive) heart and renal failure, ischemic cardiomyopathy, chronic cor pulmonale, congestive HF, cardiomyopathy, left ventricular failure, and cardiac, heart, or myocardial failure;[149] diabetes was defined as a composite condition of type 1 and type 2 diabetes mellitus, malnutrition-related diabetes mellitus, other specified diabetes mellitus,

and pre-existing malnutrition-related diabetes mellitus;[150] CKD included CKD from stage 1 to 5, kidney transplant failure and rejection, obstructive and reflux uropathy, acute renal failure, nephrotic syndrome, hypertensive renal failure, type 1 and 2 diabetes mellitus with kidney complications, and chronic tubule-interstitial nephritis;[151–153] and stroke is identified as cerebrovascular diseases, subarachnoid haemorrhage, intracerebral haemorrhage, sequelae of cerebrovascular disease, cerebral infarction, and occlusion and stenosis of cerebral arteries.[151] The list of ICD-10 codes used for case identification can be found in Table A.7 to Table A.10.

#### 4.3.4.5 Data preparation

We included all records from diagnoses, medications, procedures, lab tests, blood pressure (BP) measurements (both systolic and diastolic BP), drinking status, smoking status, and BMI collected from primary care and HES. For diagnoses, medications, procedures, and lab tests, we followed the identical protocol for data preparation as stated in §4.2. In brief, diagnoses were mapped to ICD-10 level four, medications were kept as BNF codes at the section level, and lab tests and procedures were preserved as Read codes and OPCS codes. Additionally, for drinking and smoking status, both were recorded as categorical values in CPRD, including current drinker/smoker, ex, and non. For continuous values, we included measurements with systolic BP, diastolic BP, and BMI within the range 80 to 200 mmHg, 50 to 140 mmHg, and 16 to 50 $kg/m^2$, respectively. Afterwards, we categorised systolic and diastolic BP into bins with a 5-mmHg step size (e.g., 80-85 mmHg). BMI was processed similarly with a step size of 1 $kg/m^2$. In the end, all patient records were formatted as a sequence and ordered by the event date, with age in year annotated to each record.

#### 4.3.4.6 Hi-BEHRT

Figure 4.8 uses a hypothetical patient to illustrate the architectural difference between BEHRT and Hi-BEHRT. Instead of processing the entire EHR sequence at the same

time as BEHRT. We adopted the idea of using hierarchical architecture to process EHR as the work proposed by Pappagari et al.[148] to process NLP data. We call this transformed BEHRT model Hi-BEHRT. The Hi-BEHRT model used a sliding window to segment the entire medical history into smaller segments and applied a Transformer as a local feature extractor to focus on extracting temporal interactions within each segment. Because medical records naturally have a stronger correlation when they are closer in time, we would expect the local feature extractor learns the most representative latent representation for each segment. Afterwards, we applied another Transformer as a feature aggregator to globally summarise the local features extracted in all segments. Because the size of the sliding window and the length (or number) of segments were substantially shorter (or smaller) than the raw EHR sequence, Hi-BEHRT would greatly reduce the model complexity compared to BEHRT, thus can handle a longer EHR sequence. More specifically, the space and time complexity of the self-attention are $O(L^2 + Ld)$ and $O(L^2)$, respectively, where $L$ is the sequence length and $d$ is the hidden dimension size. Assume both BEHRT and Hi-BEHRT have the same hidden dimension size $d$, because the window size ($L_w$) and length of segments ($L_{seg}$) are much small than the EHR sequence length ($L_w \ll N$ and $L_{seg} \ll N$), the Hi-BEHRT substantially reduces the space and time complexity, and the longer the EHR sequence, the bigger the difference.

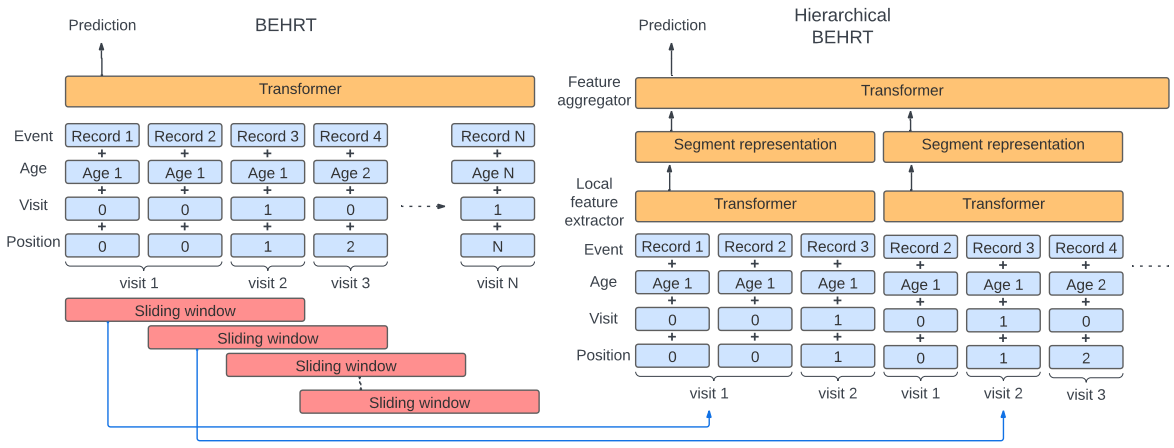



*Figure 4.8: Model architecture illustration for BEHRT and Hi-BEHRT. Despite the segment representation appearing non-over-lapping (right), they are built on overlapping events in the sliding window (left).*

#### 4.3.4.7 Model derivation

In this work, we developed both BEHRT and Hi-BEHRT for comparison. The BEHRT model mostly adopted hyperparameters reported in §4.2, and similar parameters were also used for the Hi-BEHRT. Specifically, the BEHRT model used hidden size 150, 6 attention heads, intermediate size 108, 8 layers, maximum sequence length 256, dropout rate 0.2 and attention dropout rate 0.3. For the Hi-BEHRT model, we extended the maximum sequence length to 1220, which covered the full EHR for approximately 97% of patients across all four risk prediction tasks. We used 50 and 30 as the window size and stride side (for sliding window), respectively. Similarly, we used hidden size 150, 5 attention heads, intermediate size 108, dropout rate 0.2, and attention dropout rate 0.3. The number of layers for the local feature extractor and feature aggregator is 4 and 4, respectively. Therefore, these two models. (A sensitivity analysis regarding model performance with different window size and stride size can be found in *Methods A.1*)

In terms of training, we used Adam optimiser[107] with a three-stage learning rate scheduler[154], over 100 epochs, to train both BEHRT and Hi-BEHRT models. The three-stage scheduler included 10%, 40%, and 50% epochs for warm-up, hold, and cosine decay, respectively. Batch size 128 was used, and we swept the hold learning rate among $5e^{-5}$,

$1e^{-4}$, and $5e^{-4}$, and reported the one with the best performance for each risk prediction task. In terms of early stopping strategy, we stopped training once the validation loss did not decrease for six epochs. We followed the same strategy to train Hi-BEHRT for risk prediction tasks.

#### 4.3.4.8 Model validation

In this study, we evaluated the model performance using AUROC and APS on the validation set for each risk prediction task. Additionally, to better understand how Hi-BEHRT can benefit risk prediction for patients with longer EHR sequences, we conducted a subgroup analysis to compare model performance for patients with EHR lengths 0 – 256 and longer than 256 in the learning period. Note that, 256 is the maximum capacity BERHT can handle in our study, and it will only use the latest 256 records for modelling if patients have records longer than 256. Furthermore, because patients with long EHR (i.e., >256) have a higher percentage of positive cases, which can negate the APS comparison between the short and the long EHR subgroups (as APS is sensitive to the changes in positive), we further included one more experiment that preserved all negative samples and bootstrapped a subset of positive samples in the long EHR group to manually create an evaluation dataset with a similar proportion of positive cases as the subgroup of patients with short EHR, for a fair comparison. We repeated bootstrap five times and reported the averaged results.

### 4.3.5 Results

#### 4.3.5.1 Population statistics

Table 4.10 shows that HF and diabetes have lower prevalence compared to CKD and stroke, thus, they are more imbalanced. Additionally, all cohorts have similar characteristics, but HF and CKD cohorts have sequence length slightly longer than diabetes and stroke cohorts.

*Table 4.10: Descriptive analysis of selected cohorts.*

| | HF | Diabetes | CKD | Stroke |
|---|---|---|---|---|
| **General characteristics** | | | | |
| No. of patients | 1,975,630 | 1,885,924 | 1,889,925 | 1,730,828 |
| No. (%) of positive cases | 94,495 (4.7) | 91,659 (4.9) | 175,440 (9.3) | 224,442 (12.9) |
| **Baseline characteristics** | | | | |
| Mean No. of visits per patient | 64.11 | 60.65 | 65.33 | 59.22 |
| Mean No. of codes per visit | 4.44 | 4.32 | 4.44 | 4.45 |
| Mean (SD) learning period (year) | 8.38 (3.99) | 8.31 (3.96) | 8.40 (4.02) | 8,18 (3.85) |
| Mean (SD) age (year) | 53.04 (18.18) | 52.76 (18.33) | 53.89 (18.16) | 52.58 (18.11) |
| % of patients with at least one diagnosis code | 100 | 100 | 100 | 100 |
| % of patients with at least one medication code | 95.9 | 95.9 | 96.2 | 95.1 |
| % of patients with at least one procedure code | 60.9 | 60.5 | 59.5 | 61.6 |
| % of patients with at least one test code | 88.8 | 88.3 | 88.9 | 88.3 |
| % of patients with at least one BP measurement | 88.1 | 87.7 | 88.2 | 87.4 |
| % of patients with at least one BMI measurement | 70.2 | 68.7 | 70.2 | 69.4 |
| % of patients with at least one drinking status | 62.9 | 61.9 | 63.0 | 62.0 |
| % of patients with at least one smoking status | 83.5 | 82.8 | 83.6 | 82.9 |
| Mean No. of diagnosis codes per patient | 18.84 | 18.35 | 18.57 | 17.27 |
| Mean No. of medication codes per patient | 94.70 | 85.83 | 98.79 | 85.73 |
| Mean No. of procedure codes per patient | 4.40 | 4.28 | 4.04 | 4.39 |
| Mean No. of test codes per patient | 80.13 | 72.03 | 80.76 | 75.80 |
| Mean No. of BP measurement per patient | 15.43 | 14.28 | 15.58 | 14.51 |
| Mean No. of BMI measurement per patient | 2.69 | 2.28 | 2.71 | 2.56 |
| Mean No. of drinking status per patient | 1.38 | 1.28 | 1.40 | 1.32 |
| Mean No. of smoking status per patient | 18.84 | 18.35 | 18.57 | 17.27 |
| Median (IQR) EHR length for learning period | 156 (267) | 148 (252) | 158 (285) | 146 (255) |
| No. (%) of patients with EHR length for learning period > 256 | 666,450 (33.7) | 596,072 (31.6) | 649,464 (34.4) | 546,437 (31.5) |

SD: standard deviation, IQR: interquartile range

#### 4.3.5.2 Risk prediction performance evaluation

Table 4.11 shows the model performance comparison between BEHRT and Hi-BEHRT for the prediction of HF, diabetes, CKD, and stroke on the validation. With a smaller model size and less model complexity, the Hi-BEHRT model outperformed the BEHRT model on all risk prediction tasks with 1% to 5% and 3% to 8% improvement for

AUROC and APS, respectively. While the results prove the superior performance of Hi-BEHRT in risk prediction in the general population, its strength is to enhance the prediction for patients with longer EHR.

*Table 4.11: Average performance over 5 random seeds. The same population is used to evaluate both models with the identical inputs.*

| Task | BEHRT | | Hi-BEHRT | |
|---|---|---|---|---|
| | AUROC | APS | AUROC | APS |
| HF | 0.93 | 0.71 | 0.96 | 0.77 |
| Diabetes | 0.88 | 0.66 | 0.93 | 0.74 |
| CKD | 0.91 | 0.76 | 0.93 | 0.79 |
| Stroke | 0.89 | 0.76 | 0.90 | 0.79 |

#### 4.3.5.3 Subgroup analysis in respect of EHR length

*Table 4.12: Model performance evaluation on subgroups with different EHR length.*

| | | | BEHRT | | Hi-BEHRT | |
|---|---|---|---|---|---|---|
| Sample size | % of positive cases | EHR length | AUROC | APS | AUROC | APS |
| HF | | | | | | |
| 392,161 | 2.3 | 0 - 256 | 0.91 | 0.57 | 0.95 | 0.69 |
| 189,908 | 9.0 | > 256 | 0.91 | 0.75 | 0.94 | 0.81 |
| 176,802 | 2.3 | > 256 | 0.91 | 0.64 | 0.95 | 0.72 |
| Diabetes | | | | | | |
| 386,373 | 3.2 | 0 - 256 | 0.88 | 0.61 | 0.92 | 0.68 |
| 171,777 | 8.2 | > 256 | 0.87 | 0.7 | 0.93 | 0.80 |
| 169,928 | 3.2 | > 256 | 0.86 | 0.63 | 0.93 | 0.74 |
| CKD | | | | | | |
| 371,577 | 5.3 | 0 - 256 | 0.89 | 0.68 | 0.90 | 0.69 |
| 184,597 | 15.7 | > 256 | 0.9 | 0.81 | 0.93 | 0.84 |
| 164,373 | 5.3 | > 256 | 0.90 | 0.72 | 0.93 | 0.75 |
| Stroke | | | | | | |
| 354,698 | 10.3 | 0 - 256 | 0.89 | 0.74 | 0.89 | 0.75 |
| 157,284 | 18.3 | > 256 | 0.88 | 0.79 | 0.92 | 0.84 |
| 143,298 | 10.3 | > 256 | 0.88 | 0.74 | 0.92 | 0.80 |

Table 4.12 shows that Hi-BEHRT outperforms BEHRT model in almost all subgroup analysis, both long and short EHR groups, with approximately 3% to 6% improvement on AUROC and 3% to 10% improvement on APS. Additionally, for long and short EHR groups with similar percentage of positive cases, Hi-BEHRT in general showed a more substantial performance improvements than the BEHRT model with the inclusion of more records. This observation supported the advantage of Hi-BEHRT for processing long EHR sequence. Moreover, we also noticed that for HF risk prediction, the performance of BEHRT was considerably worse than the Hi-BEHRT model. Even within the BEHRT model

itself, its performance on short EHR group is far worse than its performance on long EHR group.

### 4.3.6 Discussion

In this work, we applied a hierarchical framework to enhance the BEHRT model and named the proposed model Hi-BEHRT. It can incorporate long EHR sequences from various modalities, address the shortcomings of vanilla Transformers in processing long sequential data, and avoid missing important historical information for accurate risk prediction. With the capability of using the full available medical records, Hi-BEHRT outperformed the BEHRT model in terms of AUROC (1% - 5%) and APS (3% - 8%) across all four investigated risk prediction tasks (HF, diabetes, CKD, and stroke).

Additionally, we surprisingly found that the Hi-BEHRT model not only substantially outperformed the BEHRT model on risk prediction tasks for patients with long EHR but also greatly improved model performance for patients within the short EHR group. Especially for HF risk prediction, we noticed that the performance of BEHRT was considerably worse than the Hi-BEHRT model. Even within the BEHRT model, its performance in the short EHR group was far worse than in the long EHR group. One potential reason is that HF is a rare condition, and most of the positive cases in our dataset were included in the long EHR group. If we assume patients with long and short EHR have different medical patterns and can be represented as samples from two different distributions. The BERHT model with global attention can be overpowered by samples with long EHR. Therefore, this explains the observation that BEHRT performs considerably better HF risk prediction in the long EHR group than in the short EHR group.

On the other hand, Hi-BEHRT can benefit from the local feature extractor, which has a constrained receptive field (compared to the global attention mechanism in BEHRT) because of the relatively small sliding window size. This can focus on training the model

weights to extract local features first, which are far less sensitive to the long-term differences in medical trajectory. Afterwards, Hi-BEHRT has the feature aggregator to focus on distinguishing the differences between long-term and short-term patterns. Therefore, it achieved a more comprehensive risk estimation under both circumstances.

One of the major contributions of this work is the provision of a framework for risk prediction with the potential to include a long and comprehensive EHR. With the growing accessibility and usability of EHR systems, risk prediction using long EHR is inevitable and has important implications for medical practice. To our best knowledge, long sequence modelling and its application in the context of healthcare and EHR remain under-investigated. Our work proposed a potential solution to tackle this problem and investigated its benefit compared to a model that makes a prediction using only a fraction of the EHR.

Our study also has limitations. First, we focused on the risk of HF, diabetes, CKD, and stroke. As such, the conclusion may not generalise to other diseases. Additionally, our work relied on internal validation and the model performance under data shifts or in the external cohorts requires further investigation.

## 4.4 Deep learning for survival analysis

This has been joint work with Shishir Rao. My role consisted in conducting the literature searches, the data extraction and cleaning, conceiving the experiments and analyses, and writing the draft of the manuscript.

### 4.4.1 Motivation

Conventional statistical risk models that largely relied on expert-selected predictors and assumed an overly simplistic linear relationship between the covariates and the outcome have led to unsatisfactory risk prediction performance.[2,3] This has raised great awareness in the fields. To establish more accurate prediction, recent studies have applied machine learning and deep learning models and reported comparisons with the conventional statistical models.[118,121,155] However, these findings largely relied on a comparison of models from inconsistent framework[121] (statistical survival models and binary classification machine learning and deep learning models), or on a less representative cohort[118,156] (patients who were lost to follow-up were excluded from the study), or on a biased cohort[121] (patients who were lost to follow-up were considered as event free). This is because, unlike the statistical models, most of the current machine learning and deep learning models were binary classification models, which were unable to handle time to event and censoring. Therefore, those studies were designed in a beneficial way that the models can be compared directly. However, this of course compromised the generalisability of the findings, and the utility of the machine learning and deep learning models has raised wide criticism and remains elusive. So far, this is also a major limitation for our previously proposed studies.

### 4.4.2 Aims

Using CVD risk prediction as an example, in this study, we aimed to propose a framework that can adapt the BEHRT model for survival analysis and better handle censoring. Additionally, we would comprehensively assess the merits of survival models

versus the binary classification models, and the data-driven deep learning models versus the machine learning and statistical models that relied on expert-selected predictors using the CPRD dataset.

### 4.4.3 Methods

#### 4.4.3.1 Data source

This study was conducted using EHR from CPRD linked to HES and ONS. Primary data was available from 1 January 1985 to 31 December 2015, and the linkage was available from 1 January 1998 to 31 September 2015.

#### 4.4.3.2 Eligible participants

In this work, we followed a similar protocol from a landmark study for CVD risk prediction, QIRSK3[112], for population selection. The cohorts included patients who contributed to the CPRD between 1 January 1998 and 31 December 2015, both men and women, aged at least 25 years old, registered with general practices for at least one year, can be linked to HES, and excluded those without IMD data. The index (baseline) date for measuring risk factors was randomly selected from one's EHR history for each individual. As suggested by Li et al.[121], this approach can alleviate the time-related variability of information with more representative spread of calendar time and age. All patients were censored at the earliest of the last date in practice, last collection from practice, death, incident CVD, 10 years after the index date, or the study end date (i.e., 31 December 2015). Additionally, we excluded patients who had a prior history of CVD, prescription of statin, or no available records for risk factors before the index date. This led to a cohort that was representative of the general population, and we referred to this cohort as the general cohort. Another selected cohort only included a subset of patients with diabetes before the baseline (referred to as the diabetes cohort).

#### 4.4.3.3 Study design

In this study, we investigated 10-year CVD risk prediction tasks. For both general cohort and diabetes cohort, we randomly allocated three quarters of patients to a derivation dataset and the remainder to a validation dataset.

#### 4.4.3.4 Outcomes

CVD was defined as a composite of coronary heart disease, ischaemic stroke, or transient ischaemic attack.[121] The CALIBER[105] repository was used for disease phenotyping.

#### 4.4.3.5 Predictors

The data preparation for data-driven deep learning models followed the identical protocol described in §4.2. It considered the medical history as a sequence and incorporated all recorded information from the following modalities in primary care records: diagnoses, medications, laboratory tests, and procedures available for model training.

For the statistical models, QRISK3 is a widely accepted and clinical validated CVD risk model. Therefore, we selected and defined the expert-selected risk factors as the ones used in QRISK3[112]. More specifically, risk factors included age, sex, ethnicity, IMD, systolic BP, standard deviation (SD) of systolic BP, BMI, total cholesterol to high-density lipoprotein cholesterol ratio, smoking status, family history of coronary heart disease in a first degree relative aged under 60 years, diabetes, rheumatoid arthritis, atrial fibrillation, chronic kidney disease (stage 3, 4, 5), migraine, lupus erythematosus, severe mental illness, HIV or AIDS, erectile dysfunction, treated hypertension, history of prescribing of atypical antipsychotic drugs, and corticosteroid use. For CVD risk prediction in the diabetes cohort, we incorporated the classic CVD risk factors and excluded those diabetes-specific risk factors as done by previous research.[157] All measurements were extracted using the mean value within a 2-year window before the baseline, and the prevalence of diagnoses and

medications were identified using the phenotyping provided by the CALIBER repository[105] and Payne et al.[81], respectively.

#### 4.4.3.6 Model derivation

In this work, we compared binary classification models versus the survival models, and linear models versus the non-linear models. The binary classification models make predictions by mapping the output to a probability between zero and one using a logistic function and are trained using binary cross-entropy[27]. The survival models were implemented within the CPH framework: $h(t|x) = h_0(t)e^{\lambda(x)}$, where $h_0$ and $\lambda$ represent baseline hazard function and log-risk function, respectively, and the survival framework is trained using the partial log-likelihood objective function[158]. Note that, the objective function (i.e., partial log-likelihood) has to access the entire dataset to calculate the loss function. Thus, it does not scale well on large models and large-scale datasets.[159,160] To address this limitation, we applied the stochastic approximation proposed by Kvamme et al.[160] to train large-scale survival models. It randomly samples a subset of the population as an approximation of the overall population to calculate the partial log-likelihood loss and update the model parameters. (A pilot analysis in terms of the model convergence using stochastic approximation is included in *Methods A.2*.)

We developed five models for comparison. Our main deep learning model was an adaptation of the BEHRT model. BEHRT has demonstrated state-of-the-art performance for a variety of risk prediction tasks. In this study, we extended and modified BEHRT from a binary classification model into a survival model and refer it to as BEHRTSurv. This was achieved by replacing the linear log-risk function in the CPH model with an a more general deep learning model and trained the model within the survival framework. Then, BEHRTSruv was compared with four different models. The first two models were the logistic regression and the classic CPH models. The logistic regression model is a binary

classification model, and the CPH model is a survival model. Both models assume a linear relationship between the predictors and the outcome. The second pair of models was MLP and DeepSurv[158] (with an MLP for the log-risk function). Both models assume a non-linear relationship between the selected predictors and the outcome. Each pair of the models had an identical underlying model architecture, and all these models relied on expert-selected predictors. These five models were applied to both the general and diabetes cohorts for comparison.

Additionally, because diabetes cohort is relatively small and previous research has demonstrated that transfer learning[161] can benefit modelling on small dataset, instead of initialising random model weights for training, we conducted an additional investigation by transferring the weights of the BEHRTSurv trained on the general population to the diabetes cohort (details can be found in *Methods A.3*). We expect to show that transfer learning can be a potential strategy to facilitate modelling and risk prediction on small population.

In terms of the model implementation, BEHRTSurv adopted the BEHRT's hyperparameter reported in §4.2. Additionally, several Python packages were used to assist our study. The logistic regression models were trained using the "scikit-learn"[129,130] package; the classic CPH models were trained using the "lifelines"[129]; the MLP, DeepSurv, and BEHRTSurv were implemented using PyTorch [131]. This study used a Random search process on hyperparameter tuning to acquire high-performing DeepSurv and MLP models. Details of these models are summarised in *Methods A.3*.

#### 4.4.3.7 Model validation

We evaluated model performance in CVD risk prediction using five-fold cross validation[162] and compared their performance directly using five complementary approaches. (1) We compared discrimination performance in terms of AUROC, APS, and C-statistics [163]. C-statistics were exclusively used for survival models, and both AUROC and APS were

calculated by treating censored patients as event free for the convenience of comparing binary classification models and survival models. Note that, survival models can derive individual risk estimation over each year of the follow-up.[112] Here, we used the 10-year risk estimation for the calculation of AUROC and APS. (2) The calibration performance was evaluated using the calibration curve. It estimates the consistency between the predicted risk and the observed risk. Similarly, we used individual risk estimated at the 10$^{th}$ year as the predicted risk for the survival models, and the binary classification models can directly produce the predicted risk. With the inclusion of time-to-event data, a smoothed calibration curve was interpolated using restricted cubic splines with three knots.[164] (3) We further compared the consistency of predicted risk of individuals across different models using scatter plot analysis in a random subset of individual. (4) we conducted net benefit analysis[165] to visually compare the consequences of different decision thresholds for each model. Due to the uncertainty of the outcome for patients who are lost to follow-up before end of study, we excluded them for the net benefit analysis. Lastly, (5) we investigated a strategy for finding an optimal treatment recommendation threshold using the F1 score – an omnibus statistic combining precision and recall in one metric. Specifically, we compared the F1 score of models across different thresholds and derived the optimal treatment recommendation threshold as the threshold with the highest F1 score.

#### 4.4.3.8 Statistical analysis

For models that relied on expert-selected predictors, we conducted data imputation five times using the multiple imputation by chained equations[166]. It applied to systolic BP (38% missing in the general cohort; 3% missing in the diabetes cohort), the SD of systolic BP (62%; 10%), BMI (58%; 10%), and the ratio of total cholesterol to high-density lipoprotein cholesterol (80%; 22%). Because BEHRTSurv models followed a different

philosophy, which uses minimal processed EHR rather than selected predictors, for modelling, data imputation was not conducted and did not apply to them.

### 4.4.4 Results

#### 4.4.4.1 Study population

The general and diabetes cohorts included 3,099,020 million (with 5.3% positive cases) and 63,760 (13.2%) patients, respectively, with a mean follow-up of 3.8 years for the general cohort and 3.2 years for the diabetes cohort. Compared to the general cohort, patients in the diabetes cohort were relatively older, had lower socioeconomic status, higher systolic BP, higher BMI, and higher prevalence of CVD risk factors. Table 4.13 shows the baseline characteristics for the general and the diabetes cohorts.

*Table 4.13: Baseline characteristics for general and diabetes cohorts.*

| Characteristics | General Cohort | Diabetes Cohort |
|---|---|---|
| No. of patients | 3,099,020 | 63,760 |
| CVD cases (%) | 164,275 (5.3) | 8,439 (13.2) |
| Women (%) | 1,695,892 (54.7) | 29,746 (46.7) |
| Age in years (SD) | 49.1 (17.6) | 61.7 (16.0) |
| Ethnicity | | |
| Unknown (%) | 2,051,545 (66.2) | 38,776 (60.8) |
| White (%) | 978,312 (31.6) | 22,507 (35.3) |
| Other Asian (%) | 7,234 (0.2) | 264 (0.4) |
| Pakistani (%) | 7,059 (0.2) | 360 (0.6) |
| Indian (%) | 12,359 (0.4) | 523 (0.8) |
| Other (%) | 13,211 (0.4) | 323 (0.5) |
| Caribbean (%) | 6,443 (0.2) | 351 (0.6) |
| Mixed (%) | 5,051 (0.2) | 104 (0.2) |
| Bangladeshi (%) | 2,164 (0.1) | 132 (0.2) |
| Chinese (%) | 3,188 (0.1) | 71 (0.1) |
| Other Black (%) | 3,755 (0.1) | 120 (0.2) |
| Black African (%) | 8,699 (0.3) | 229 (0.4) |
| Index of multiple deprivation (IMD) | | |
| IMD 1 (%) | 725,418 (23.4) | 11,697 (18.3) |
| IMD 2 (%) | 686,786 (22.2) | 13,255 (20.8) |
| IMD 3 (%) | 649,736 (21.0) | 13,442 (21.1) |
| IMD 4 (%) | 566,205 (18.3) | 13,070 (20.5) |
| IMD 5 (%) | 470,875 (15.2) | 12,296 (19.3) |

| | | |
|---|---|---|
| Systolic blood pressure (SD) | 129.6 (14.4) | 136.9 (14.3) |
| BMI (SD) | 27.0 (3.9) | 30.4 (5.9) |
| HDL (SD) | 1.5 (0.3) | 1.3 (0.4) |
| Smoking status (number of cigarettes per day) | | |
| Unknown (%) | 1,478,574 (47.7) | 8,337 (13.1) |
| Non-smoker (%) | 846,466 (27.3) | 28,481 (44.7) |
| Ex-smoker (%) | 425,383 (13.7) | 18,565 (29.1) |
| Light (<10 cigarettes/day) (%) | 104,013 (3.4) | 2,319 (3.6) |
| Moderate (10-20 cigarettes/day) (%) | 146,528 (4.7) | 3,098 (4.9) |
| Heavy (>20 cigarettes/day) (%) | 98,056 (3.2) | 2,960 (4.6) |
| Disease at baseline | | |
| Diabetes (%) | 63,760 (2.1) | 63,760 (100) |
| Rheumatoid arthritis (%) | 15,455 (0.5) | 3,166 (5.0) |
| Atrial fibrillation (%) | 51,855 (1.7) | 2,795 (4.4) |
| CKD (%) | 12,902 (0.4) | 1,426 (2.2) |
| Migraine (%) | 105,581 (3.4) | 59 (0.1) |
| Lupus erythematosus (%) | 2,352 (0.1) | 948 (1.5) |
| Mental illness (%) | 24,313 (0.8) | 206 (0.3) |
| HIV/AIDS (%) | 5,863 (0.2) | 5,416 (8.5) |
| Erectile dysfunction (%) | 60,803 (2.0) | 5,414 (8.5) |
| Medication use at baseline | | |
| Antihypertensives (%) | 146,992 (4.7) | 5,414 (8.5) |
| Antipsychotics (%) | 6,139 (0.2) | 176 (0.3) |
| Corticosteroids (%) | 41,220 (1.3) | 834 (1.3) |

#### 4.4.4.2 Discrimination performance

*Table 4.14: Model discrimination performance with 95% confidence intervals on general and diabetes cohorts.*

| | General Cohort | | | Diabetes Cohort | | |
|---|---|---|---|---|---|---|
| Model | AUROC | APS | C-statistics | AUROC | APS | C-statistics |
| BEHRTSurv | **0.92±0.00** | **0.32±0.01** | **0.92±0.00** | **0.81±0.03** | **0.34±0.03** | **0.81±0.02** |
| DeepSurv | 0.85±0.00 | 0.19±0.00 | 0.86±0.00 | 0.69±0.02 | 0.23±0.02 | 0.71±0.01 |
| MLP | 0.84±0.00 | 0.19±0.01 | — | 0.70±0.00 | 0.24±0.02 | — |
| CPH | 0.84±0.00 | 0.19±0.00 | 0.85±0.00 | 0.69±0.02 | 0.23±0.02 | 0.71±0.01 |
| LR | 0.84±0.00 | 0.20±0.00 | — | 0.70±0.00 | 0.24±0.01 | — |

LR: logistic regression

Table 4.14 shows the model discrimination performance in the general and the diabetes cohorts. We see that the BEHRTSurv model, which incorporated minimal processed EHR, substantially outperformed the linear and non-linear models (logistic regression, CPH, MLP, and DeepSurv) that relied on expert-selected predictors.

Additionally, model performance in the diabetes cohort was generally worse than in the general cohort. We did not observe a substantial difference in predictive performance among the four benchmark models.

#### 4.4.4.3 Calibration performance

As shown in Figure 4.9**,** the survival models (CPH, DeepSurv, and BEHRTSurv) demonstrate superior calibration performance compared to the binary classification models (LR and MLP). The LR and MLP models, which ignored censoring (i.e., assume censored patients to be event free), critically underestimated risk of incident cardiovascular events compared to the survival framework modelling.

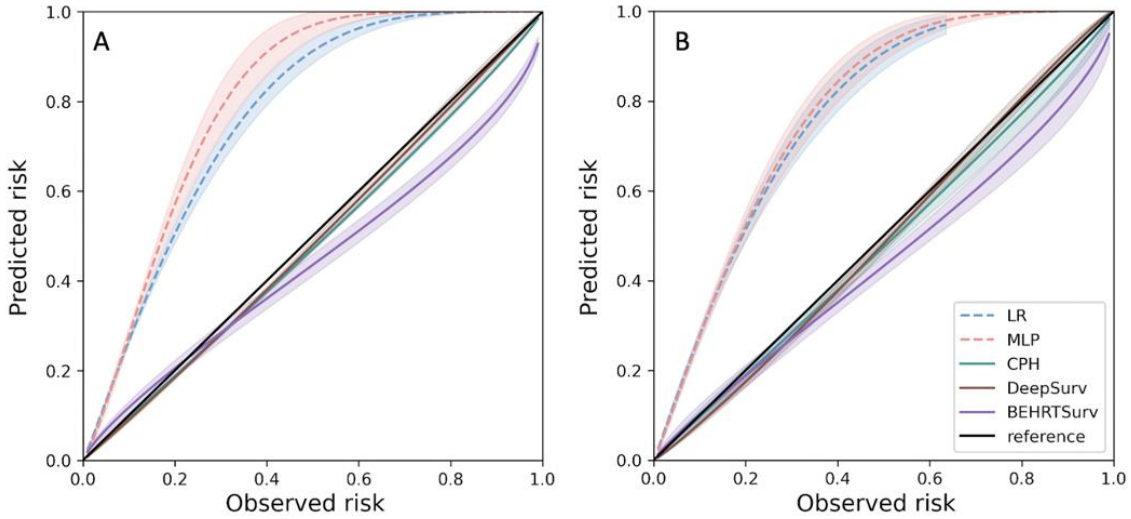



*Figure 4.9: Calibration curves for general cohorts (A) and diabetes cohort (B). Reference line represents that a model is well calibrated. LR represents logistic regression model.*

#### 4.4.4.4 Consistency of model predictions

Using the validated CPH model as the reference, we compared the consistency of predictions for a random subset of individuals across different models (Figure 4.10). LR and DeepSurv demonstrated good correlation of individual-level prediction to CPH modelling. However, the LR substantially underestimated the risk of CVD across the entire risk distribution compared with CPH (consistent with the calibration analysis). Additionally, both models ranked a fraction of individuals who suffered and event during follow-up as

low-risk (as seen by the distribution of red dots in the lower left section). The predictions of BEHRTSurv and CPH varied substantially. BEHRTSurv captured positives better than CPH modelling (high density of red dots above the diagonal line). That is, a large, fraction of positive cases were predicted as low-rosk patients by the CPH model but were correctly identified as high-risk patients by the BEHRTSurv model. This poor correlation appears to be largely due to the CPH's limited ability to assign individuals, for example, to a risk > 0.7 in the general cohort.

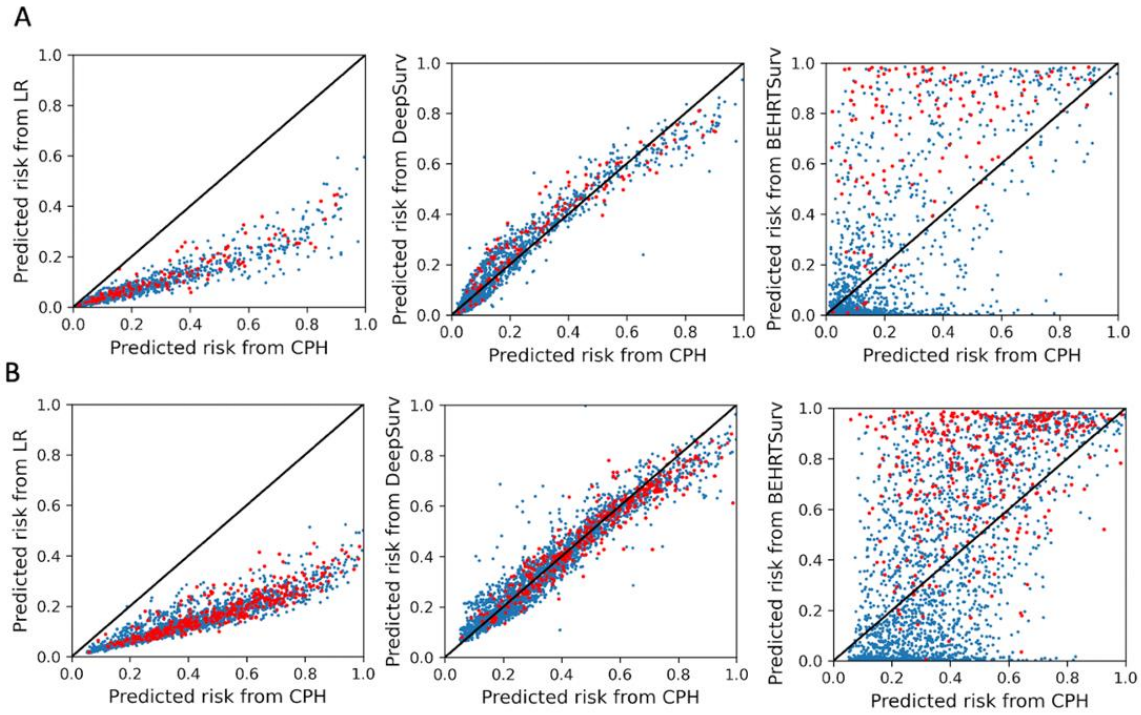



*Figure 4.10: Model consistency on individual predictions. CPH is used as the benchmark for model comparison on the general cohort (A) and the diabetes cohort (B). For the convenience of visualisation, 2,500 patients randomly selected patients were used for comparison. The red and blue dots represent patients with and without an event (or censored) in follow-up, respectively.*

Further analyses of predicted risk distributions for all models in both cohorts are provided in Figure 4. 11. While the survival benchmarks generally demonstrated a unimodal distribution (with one maximum at the low-risk range), the BEHRTSurv model better discriminated the high and low risk patients with two peaks close to probability 0.0 and 1.0.

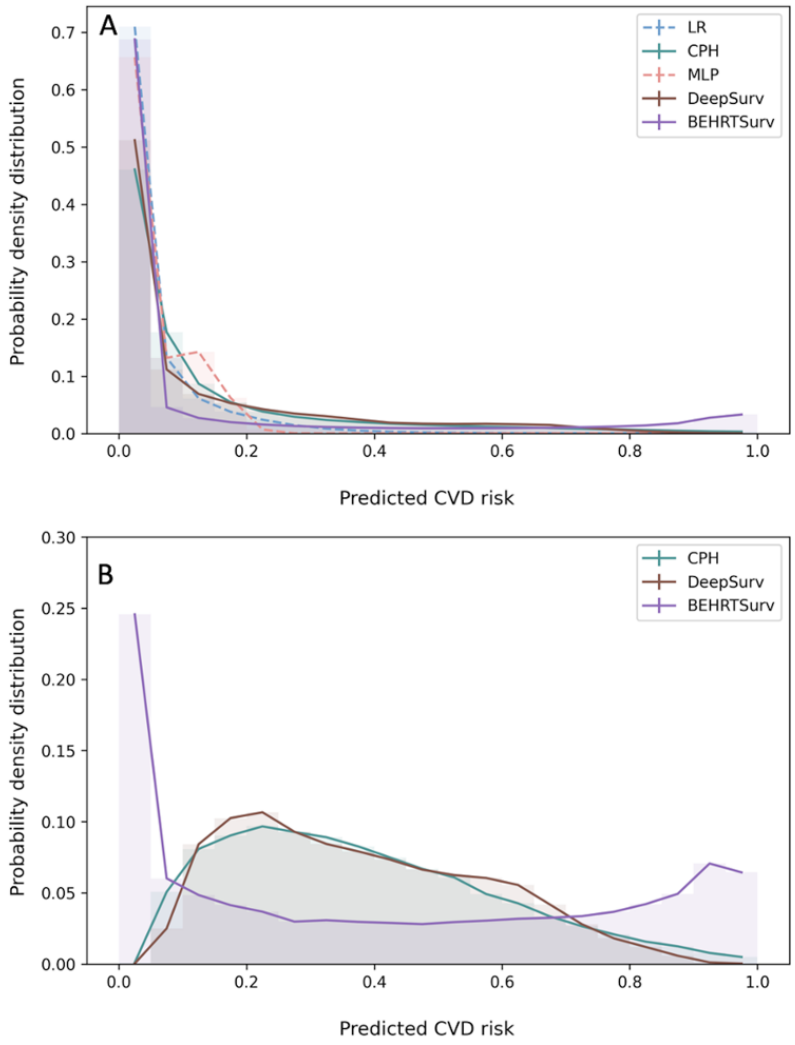



*Figure 4. 11: Risk distribution of models in general (A) and diabetes (B) cohort.*

#### 4.4.4.5 Net benefit analysis

The previous analyses have demonstrated that the survival models were preferable for risk prediction. Therefore, we further conducted a net benefit analysis for the survival models to better understand the consequence of decision-making at different risk thresholds. Figure 4.12 shows that the CPH and DeepSurv models have similar net benefit patterns across the full spectrum of decision thresholds, in both the general and the diabetes cohorts. However, the BEHRTSurv model outperformed the DeepSurv and CPH model across all decision thresholds. In both the general and diabetes cohorts, the area under the net benefit curve (AUC-NB) was larger for BEHRTSurv than the other models.

As recommended by the NICE [9] and other guidelines such as European Society of Cardiology[167], patients with a predicted 10-year risk of CVD of 10% or more are to be recommended for primary prevention treatment. In patients who are not lost to follow-up, applying this threshold to the general cohort would classify 172,310 (37%) and 247,876 (53%) patients as eligible for treatment when decisions are based on BEHRTSurv and QRISK, respectively. Among those, BEHRTSurv identified 159,413 (93%) true positive patients and QRISK identified 143,645 (58%) true positive patients. This means that keeping the currently recommended risk threshold for patient identification, the BEHRTSurv model would reduce the treatment eligible population by about 30% and even then, it would only overtreat 8% of those who would not suffer an event, compared with 42% when QRISK is applied.

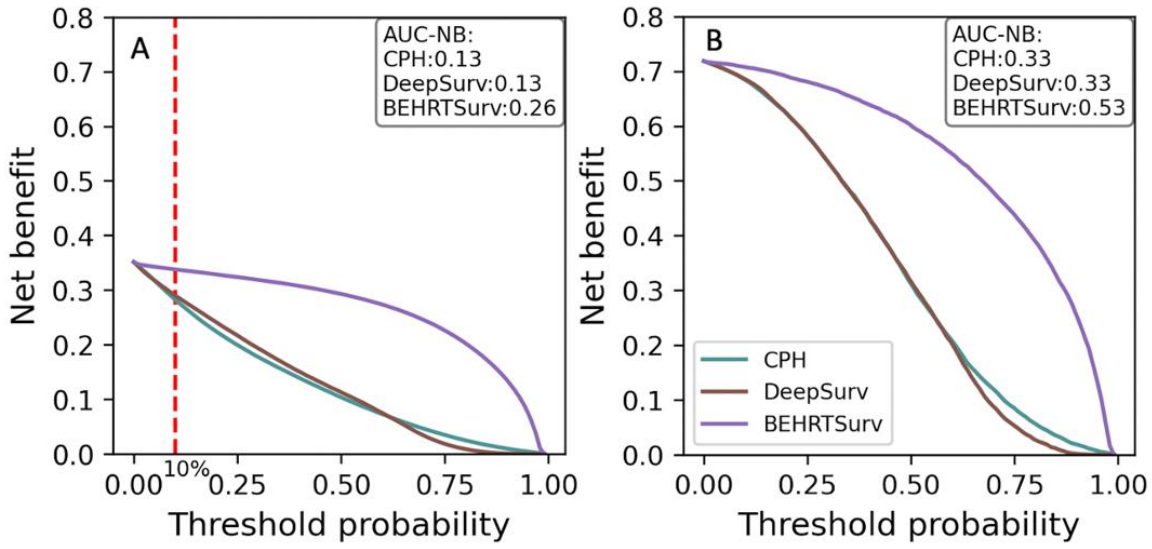



*Figure 4.12: Net benefit of survival models on the general cohort (A) and the diabetes cohort (B). AUC-NB represents the area under the net benefit.*

In the diabetes cohort, the BEHRTSurv approach similarly provided greater net benefit than benchmark survival framework comparisons (Figure 4.12B). At any threshold across the spectrum of decision thresholds, the BEHRTSurv model captured more true positives than the other two survival models. At the 10% threshold as an example, 11,493 and 8,593 patients were classified as high risk by QRISK and BEHRTSurv models,

respectively. Of these predictions, 8,380 (73%), and 8,338 (97%) were true positives, respectively.

#### 4.4.4.6 Treatment recommendation analysis

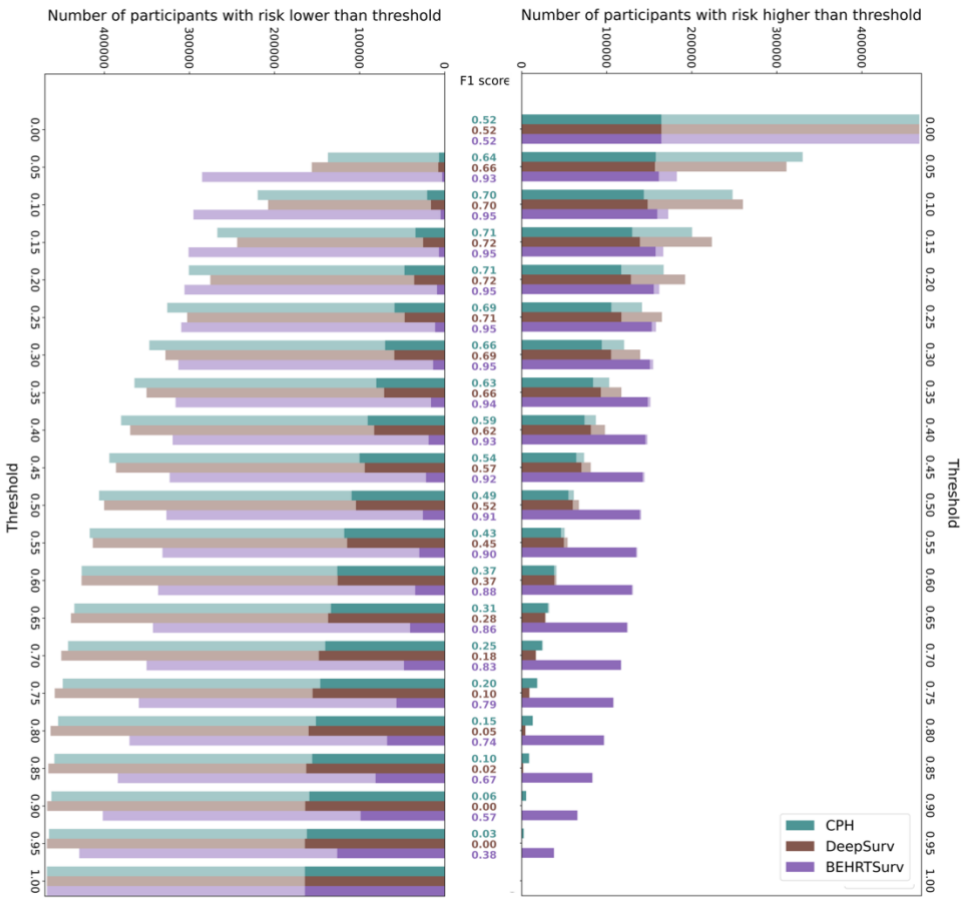
In part due to the reduced costs of statin-based therapy and its safety profile, the guideline committees have indeed embraced a policy of recommending treatment for those at 10% predicted risk or higher for patients free of pre-existing risk factors. However, as Figure 4.12A shows, BEHRTSurv showed a relatively small attrition in net benefit even when the decision threshold was higher than 10% risk threshold. For instance, selecting patients at 30% risk of CVD would lead to treatment being offered to 154,531 (33%) patients with 150,661 (97% of those offered treatment) being true positive patients. While a 30% decision boundary using BEHRTSurv may be a better alternative to the recommended 10% threshold for selecting those at risk for downstream treatment, a natural question emerges: is there an optimal decision boundary for the BEHRTSurv model?

To assess whether there is an alternative and more efficient decision boundary based on predicted risk of CVD, the number of false positives and false negatives need to be factored into the performance metric. For all survival models across ventiles (20 thresholds), Figure 4.13A demonstrates the proportion of individuals for whom treatment would be recommended (right panel), and those who would be above the risk threshold and hence, not eligible for treatment (left panel). Among each of these groups, a fraction will suffer an event in the future and a fraction will not. The optimal threshold would have few people without an event below the decision threshold and few people with an event above the decision threshold. The 20% decision threshold yields the highest F1 score making this threshold the optimal decision boundary for assessing risk in the general cohort using the proposed model. At this optimal decision boundary of 20%, the BEHRTSurv model demonstrates a 26% absolute and 37% relative improvement in terms of F1 score. As compared to the benchmark

CPH and DeepSurv risk prediction models, there is a substantial reduction in false positive/negative predictions for the BEHRTSurv model while simultaneously, true positive/negative predictions are generally preserved.

In patients with diabetes (Figure 4.13B), to assess whether there is a more optimal treatment strategy than the current "treat all" strategy [9], a similar analysis with F1 score was conducted. For the BEHRTSurv approach, with respect to the baseline decision boundary (threshold =0), there is a 14% absolute and 17% relative improvement at threshold of 10% (optimal decision threshold with maximal F1 score for treatment recommendation by BEHRTSurv). Compared to the benchmark CPH and Deepsurv models, BEHRTSurv achieves substantially higher performance regarding F1 score. However, increasing the decision threshold above 0.3 diminishes all models' ability to capture the positives correctly (decrease in more true than false positives). Thus, while guidelines recommend treatment for all patients with diabetes regardless of estimated risk, we can see in Figure 4.13B that BEHRTSurv would provide greater at a relaxed threshold (around 10%).

A


Number of participants with risk lower than threshold
Number of participants with risk higher than threshold
F1 score
Threshold
CPH
DeepSurv
BEHRTSurv

B

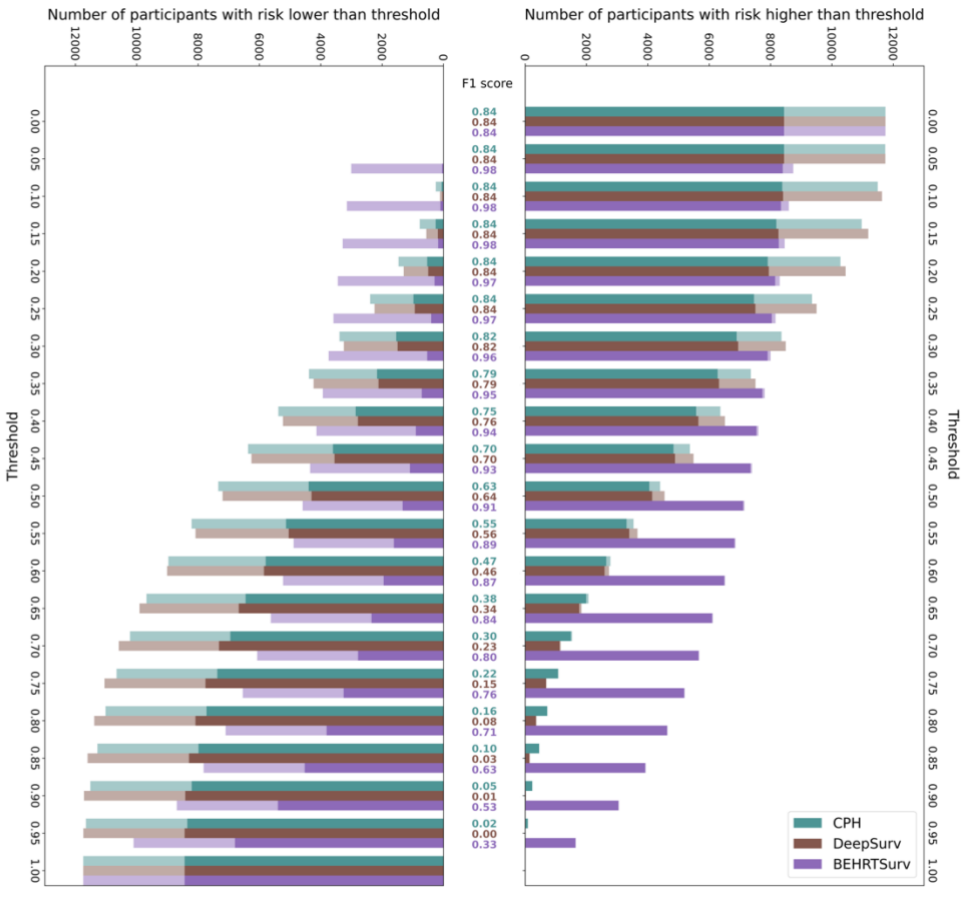

Number of participants with risk lower than threshold
Number of participants with risk higher than threshold
F1 score
Threshold
CPH
DeepSurv
BEHRTSurv

*Figure 4.13: F1 score and error analysis of various models across different thresholds in the general (A) and diabetes cohort (B). The lighter coloured (transparent) stack of the bar graph represents the number of predictions of no event during follow-up period while the darker coloured stack (solid) represents the number of predictions of an event during follow-up period. The y-axis represents the thresholds and a patient with predictive probability less than a given threshold is considered a negative prediction. Otherwise, the patient is considered as a positive prediction. The F1 score for all models is given in between the two graphs in appropriate coloured text.*

### 4.4.5 Transfer learning for BEHRTSurv on diabetes cohort

*Table 4.15: Discrimination performance of BEHRTSurv with and without transfer learning in the diabetes cohort. BEHRTSurv and BEHRTSurv (transfer) represent models trained without and with transfer learning, respectively.*

| | Diabetes Cohort | | |
|---|---|---|---|
| Model | AUROC | APS | C-statistics |
| BEHRTSurv | 0.70 (±0.02) | 0.24 (±0.01) | 0.75 (±0.02) |
| BEHRTSurv (transfer) | 0.81 (±0.03) | 0.34 (±0.03) | 0.81 (±0.02) |

Table 4.15 and Figure 4.14 summarise the comparison between BEHRTSurv trained with and without transfer learning in the diabetes cohort. The results show that transferring the model learned in the general cohort can greatly improve the model discrimination performance in a subset cohort with diabetes compared to a model trained from randomly initialised weights. This further led to better net benefit across the full spectrum of decision thresholds.

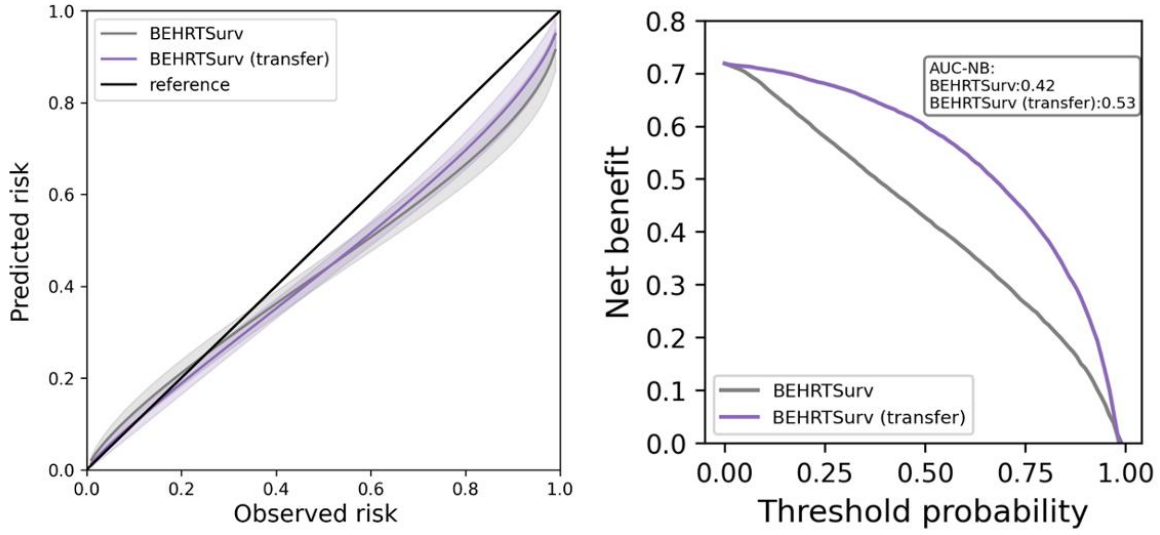



*Figure 4.14: Calibration performance (left) and net benefit (right) of BEHRTSurv with and without transfer learning in the diabetes cohort. BEHRTSurv and BEHRTSurv (transfer) represent models trained without and with transfer learning, respectively.*

### 4.4.6 Discussion

In the presence of a large cohort with comprehensive medical records and a wide range of analyses, we found that survival models were more preferable for risk prediction than the binary classification models, and the data-driven deep learning models that relied on minimal processed EHR substantially outperformed the linear, and non-linear models relied on expert-selected predictors. We also illustrated that transfer learning can be beneficial for applying deep learning models to a relatively small diabetes cohort.

Previous studies have investigated the benefit of various machine learning benchmark models compared to the statistical models for risk prediction.[121] However, there are many shortcomings in these studies that limit their generalisability, including but not limited to issues in study design (e.g., biased cohort selection [121], exclusion of censored patients [118,155,168]) and model comparison (e.g., inconsistent comparison between binary classification models and survival models[121]). Furthermore, while the previous works have explored the comparisons between the traditional machine learning models (e.g., random forest and gradient boosting models [121]) and the clinically validated models, the investigation of the newly developed deep learning models using large-scale EHR is lacking.

The research demonstrated novelty in several aspects. It designed a study that consistently and comprehensively probed model performance in various ways (i.e., discrimination, calibration, consistency, net benefit, and performance gain) that ultimately shed light on the strengths and weaknesses of various models. While the discrimination metrics showed that all benchmark models performed similarly in AUROC and APS, they failed to illustrate the severe miscalibration of the binary classification models. A more holistic image of model capabilities emerged by complementing evaluation with other analyses.

Additionally, this study proposed a novel deep learning survival model, BEHRTSurv, that can incorporate data-driven feature extraction, rich EHR, and a better-calibrated survival framework to achieve outstanding CVD risk prediction. Compared to the benchmark models, its superiority was not limited to the general cohort but also to the diabetes cohort, which has a higher baseline cardiovascular risk. The utility of BEHRTSurv was especially highlighted in investigating the higher risk (diabetes) cohort. While the investigation for CVD prediction in the general cohort has a long history and a list of staple risk factors has been established, modelling for the at-risk (e.g., diabetes) cohort is still poorly explored. With an unhealthier baseline in the diabetes cohort, such as higher BMI, greater socioeconomic deprivation, and higher prevalence of known risk factors, a risk assessment must be conducted with nuance and with the inclusion of risk factors specific to that cohort. However, even with the common awareness that the risk factors have to be specified for different cohorts, the major obstacle remains: the variables chosen are only those known to be risk factors, and those unknown risk factors not included because of lack of knowledge can severely impact the predictive performance. Therefore, in those cases, BEHRTSurv can provide a great alternative that automatically captures latent features from the raw EHR without the need for specifying risk factors.

Also, compared to a more heterogeneous general cohort with large sample size, the diabetes cohort is relatively small, which can limit the potential of a data-driven model. Thus, we conducted an initial investigation on the utility of transfer learning. By finetuning the model weights (train on the general cohort) that capture rich features (in the general population) in the diabetes cohort, BERTSurv achieved better risk estimation than training from randomly initialised weights. Thus, with this success in knowledge transfer from the general population to sub-population in cardiovascular risk prediction, the utility of transfer

learning in other sub-populations and different clinical focus areas should be explored in the future study.

In terms of clinical benefit, the proposed BEHRTSurv model shows promising performance for CVD risk prediction. In the general cohort, considering patients with more than 10% CVD risk should be recommended for treatment as stated in the NICE guideline, the BEHRTSurv model can capture seven more true positive predictions for every 100 patients than the current clinical model (QIRSK). Such improvement can considerably enhance care planning and CVD prevention. In the diabetes cohort, the NICE guideline recommended treatment for all with diabetes. Such a conservative strategy was beneficial because current risk models had unsatisfactory predictive performance and could not correctly identify the true positive predictions as shown in the consistency analysis. However, perhaps more discriminative risk tools, such as BEHRTSurv, can avoid the "treatment for all" paradigm and more selectively assign treatment to the subset of high-risk patients with diabetes, while the others might be strongly urged to follow non-pharmacological forms of therapy (healthy diet, exercise, etc.).

This study also has limitations. One limitation is that the data in EHR are not missing at random. Therefore, for models that rely on expert-selected predictors, there is always a risk of bias when imputation is employed. However, this limitation is largely relevant to epidemiological studies that seek to identify risk factors rather than risk prediction. The non-linear and deep learning model can also benefit from more comprehensive hyper-parameter tuning. However, we believe the reported models have achieved sufficient performance and led to reasonable conclusions.

## 4.5 Conclusion and potential future works

This chapter introduced a deep learning framework, BEHRT, that can incorporate minimal processed EHR for accurate risk prediction. Compared to the conventional statistical and machine learning models, the proposed approach can achieve superior performance for predicting cardiovascular-related outcomes, both internally and externally (in the presence of data shifts). This is primarily because the proposed data-driven approach can automatically learn meaningful representations from the raw EHR data, in which the comprehensive medical histories are maximally preserved. However, on the other hand, the conventional models mainly rely on manual feature engineering (e.g., risk factor selection and data imputation). It requires an adequate understanding of the medical conditions. The risk factors also have to be carefully reconstructed for specific cohorts (e.g., risk factors in the diabetes cohort may differ from the general cohort). Otherwise, a model may fail to meet the expectations. Such a rigorous requirement in feature engineering becomes a major obstacle in these models, and even the accessibility to the most comprehensive medical dataset would not resolve it. This limitation in the conventional models reinforces the great potential of the proposed deep learning framework in those scenarios. Building on this, we further enhanced our proposed framework with the flexibility to handle patients with long EHR sequences (that important information would not be missed while making a prediction) and censoring (i.e., survival analysis). This empowers our proposed framework with broad applications for various risk prediction tasks in healthcare.

There are also several limitations. First, we mainly focused on BEHRT, the Transformer-based model, in this thesis, which can potentially miss a significant part of the other deep learning models in this fields and might not be the most cutting-edge models in the literature. This is mainly because the Transformer model has consistently demonstrated exceptional performance across a wide range of tasks and proved its superiority to some

established LSTM and CNN-based deep learning models in our work (§4.1). Instead of always exploring the latest models and methods, we aim to streamline our focus and conduct comprehensive evaluations on the Transformer model, then leverage the Transformer model to address numerous existing challenges in EHR. Our work will serve as a foundation to the usefulness of deep learning models in risk prediction and EHR, and encourage researchers for further investigations, such as improving model generalisability and better handle data shifts using contrastive learning, adversarial learning, or federated learning[169]. These methods should be directly applicable to the proposed BEHRT model.

Additionally, it is also important to acknowledge that our focus has been primarily on the CPRD dataset in this DPhil project. The generalisability of the proposed model can benefit from further evaluations using other large-scale EHR datasets. However, it should be noted that these datasets are often not publicly available and are typically limited to specific authorised purposes. Due to this reason, we were unable to conduct further evaluation.

# 5. Explaining neural networks for medical knowledge retrieval

This chapter will focus on the second objective of this doctoral thesis to provide medical insights by explaining deep learning models. More specifically, we will investigate techniques that can interrogate a model to uncover disease associations (Chapter 5.1). This can help find risk and preventative factors, especially for rare or poorly studied conditions. Additionally, we will investigate models that can provide counterfactual explanations (Chapter 5.2). Such explanations can inform actionable suggestions that directly impact clinical decision-making, which would be more desirable in practice.

## 5.1 Post-hoc interpretation for risk association analysis

This has been joint work with Shishir Rao and parts of this work have been published in IEEE Journal of Biomedical and Health Informatics at: https://ieeexplore.ieee.org/abstract/document/9706318. The manuscript benefitted from the input of several co-authors. I am the co-first author and the work presented in this chapter is my own work. The perturbation-based analyses are jointly produced with Shishir Rao. My role consisted in designing the study, conceiving the experiments and analyses, conducting the literature searches, the data extraction and cleaning, and writing the manuscript.

### 5.1.1 Motivation

Previous evidence demonstrated that deep learning models can learn meaningful representations from large-scale EHR that lead to accurate risk prediction.[101,102,104,170,171] This data-driven mechanism fueled a new hope to discover potential risk associations that are less dependent on expert knowledge. Nevertheless, due to their deep, complex

architecture, it is difficult to retrieve medical knowledge from these models for further clinical usage, hindering their contributions to risk factor discoveries and wider applications in healthcare. Therefore, there is a need for further work on the explainability of these models.

### 5.1.2 Related work

Recent research has shown great progress in model explanations,[53,172,173] in which post-hoc explainers are currently the most widespread form of explanations. They aim to explain models that are already trained and with a fixed target. For instance, given a trained model, LIME[50] perturbates a sample to estimate the neighbourhood of its prediction, then learns an interpretable model (e.g., linear regression) to explain the relationship between the inputs and the outputs; SHAP[56] learns explanations by permutating the inputs of a sample and observing how the additive features change the outputs of the model. Other state-of-the-art approaches such as DeepLIFT[174] and saliency maps[175] also follow a similar philosophy that assigns weights to different attributes to indicate their importance and associations to the prediction. Many fields, such as natural language processing and computer vision, have applied these methods to assist the model explanation. However, explainable deep learning with rich EHR is still in its nascency. Hence, tailoring known methods to improve model explainability in the medical context is crucial.

### 5.1.3 Aims

In this study, we used heart failure (HF) as an example, and applied the proposed BEHRT model for predicting incident HF using large-scale EHR dataset. In addition to validating BEHRT's performance for HF risk prediction, we aimed to investigate model expainability, especially risk and preventative factors identified by the deep learning model, by applying an established post-hoc perturbation-based explainer.

### 5.1.4 Methods

#### 5.1.4.1 Data source

This study was conducted using EHR from CPRD linked to HES and ONS. Primary data was available from 1 January 1985 to 31 December 2015, and the linkage was available from 1 January 1998 to 30 September 2015.

#### 5.1.4.2 Study design

We defined a risk prediction task that used all available information before the baseline to predict the incident HF risk within a six-month window after the baseline. The incident HF was defined as the first recorded HF diagnosis code in one's EHR (both primary care and hospital care). In terms of the baseline, we randomly selected a date up to six months before the incident HF for patients with at least one HF diagnosis; our process maximally preserved HF patients and ensured that no HF diagnosis was in the learning period. For patients without HF, a baseline date was randomly selected between the registration date and the date of the last available record.

#### 5.1.4.3 Eligible participants

In this study, we selected patients who contributed to data between 1985-01-01 and 2015-12-31, were linked to HES, and marked as "acceptable" by CPRD quality control. This led to a cohort of four million patients. Furthermore, we selected patients with both men and women, aged at least 16 years old, and with richer information for the prediction of incident HF. More specifically, we included patients with at least ten (primary care or hospital) visits, at least three years of records, and at least ten unique medical codes recorded before the baseline. This led to a cohort of 100,071 patients,among those, 13,050 (13%) patients were positive cases.

#### 5.1.4.4 Outcomes

We adopted the HF phenotyping from the CALIBER[105] code repository. In brief, HF was defined as a composite condition of congestive HF, left ventricular failure, cor pulmonale, cardiomyopathy, hypertensive heart disease with (congestive) HF, cardiac failure, and HF.

#### 5.1.4.5 Data preparation

In this study, we incorporated records from diagnoses and medications for modelling. Primary and hospital care diagnoses were mapped to 299 clinically meaningful disease categories using the CALIBER code repository. Medications were recorded using the BNF coding system, indicating medication prescriptions instead of medication retrieval or dispensation. We used the BNF codes at the section level prevalent in the selected population. This led to 426 unique medication codes. Additionally, all records were associated with corresponding chronological age in month and calendar year.

#### 5.1.4.6 Model derivation

We developed and validated HF risk prediction using BEHRT with an additional embedding layer – calendar year – to inform the model about the potential change in medical trajectories across a different calendar year. Therefore, each event is represented as the summation of event, age, visit, position, and calendar year embeddings in BEHRT (as shown in Figure 5.1).

The model was implemented using PyTorch[131], and Bayesian optimisation[106] was applied for hyperparameter tuning on the number of layers, hidden size and intermediate size. After 20 iterations, we chose the optimal hyperparameters that achieve the best accuracy on a small subset of samples held-out in the training dataset for validation purpose, with number of layers: 4, hidden size: 120, number of attention heads: 6, and intermediate

size: 108. In this work, we consider the risk prediction a binary classification task. Thus, the BEHRT was trained using the binary cross-entropy loss[27].

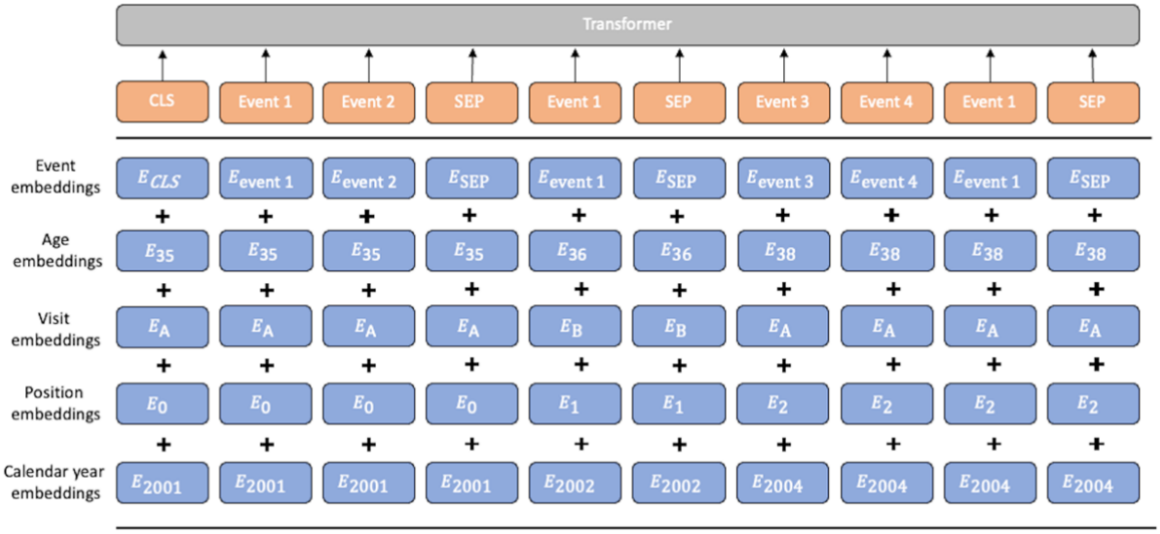



*Figure 5.1: Embedding representation of the BEHRT architecture for incident HF risk prediction.*

#### 5.1.4.7 Model validation

We evaluated model performance using five-fold cross-validation and reported model performance in terms of AUROC and APS with 95% confidence intervals. Furthermore, by selectively removing available information (diagnoses (D), medications (M), age (A), and calendar year (Y)) in modelling, we assessed how the inclusion of different data modalities affected the model predictive performance (referred to as ablation study). Lastly, because BEHRT is expected to model medical trajectory, and the temporal variability is intrinsic to EHR as many factors such as population and disease patterns can change over time, we conducted embedding similarity analysis (using cosine distance[109]) for the learned representation of age and calendar year as a way of assessing the capability of BEHRT for capturing the temporal variability. The large the difference (i.e., higher dissimilarity) in the learned representation, the more substantial the change (or temporal variability).

#### 5.1.4.8 Post-hoc perturbation-based explainer

In this work, we aimed to apply a post-hoc explainer to the BEHRT model to better understand the event contributions to the prediction of incident HF, as a way of making the

deep learning model explainable. To this end, we extended an established perturbation-based explainer in language modelling[173] to the EHR, applied the perturbation to the summed embeddings, and treated an event in the context of age, visit, position, and calendar year as a whole. The summed embedding of an event is referred to as a predictor for the convenience of description. The fundamental concept of the perturbation explainer is to measure the change in the output after perturbing the predictor to indicate the contribution of predictors. If large perturbations of predictors lead to minimal changes in the output, then predictors are unimportant for prediction. On the contrary, if minimal perturbation greatly changes the output, the respective predictors are highly important. In this work, we proposed an asymmetric loss function to prioritise the predictors that enhance the HF/non-HF predictions. HF and non-HF represent HF positive and negative, respectively. The explainer is summarised by Algorithm 1, and more implementation details can be found in Methods A.4.

Algorithm 1 Perturbation analysis

$M$: model, $N \in \mathbb{N}$: number of predictors, $X \in \mathbb{R}^{N \times D}$: embeddings of input predictor, $\tilde{X} \in \mathbb{R}^{N \times D}$: perturbed input embeddings, $Y \in \mathbb{R}^{1}$: indicator of HF patient, $L$: loss function, $S \in \mathbb{R}^{N \times 1}$: contribution score, $O \in \mathbb{R}^{1}$ and $\tilde{O} \in \mathbb{R}^{1}$: model output.

PERTURBATION ()

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

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 $S$

The perturbation-based explainer is a local surrogate method, which means it can only provide explanations (contribution of predictors) at the individual level. Therefore, to carry out an explanation at the population level, we proposed to aggregate the contribution of predictors across the population and use the average contribution to represent a predictor's

contribution. In this study, we focused on the contribution of the first incidence (if repetition was applicable) of a disease/medication in the context of age and calendar year for each patient (Figure 5.2). However, simply aggregating the contributions across patients can potentially cancel out the contribution of a predictor across the population (a discussion can be found in Methods A.5), and it also makes little sense that a predictor with identical contribution scores to the HF positive and HF negative predictions are considered the same in terms of HF association. Therefore, in this study, we proposed to consider true positive prediction (HF group) and true negative predictions (non-HF group) as two separate groups and then calculate the relative contribution (RC) between these two groups to indicate the event associations.

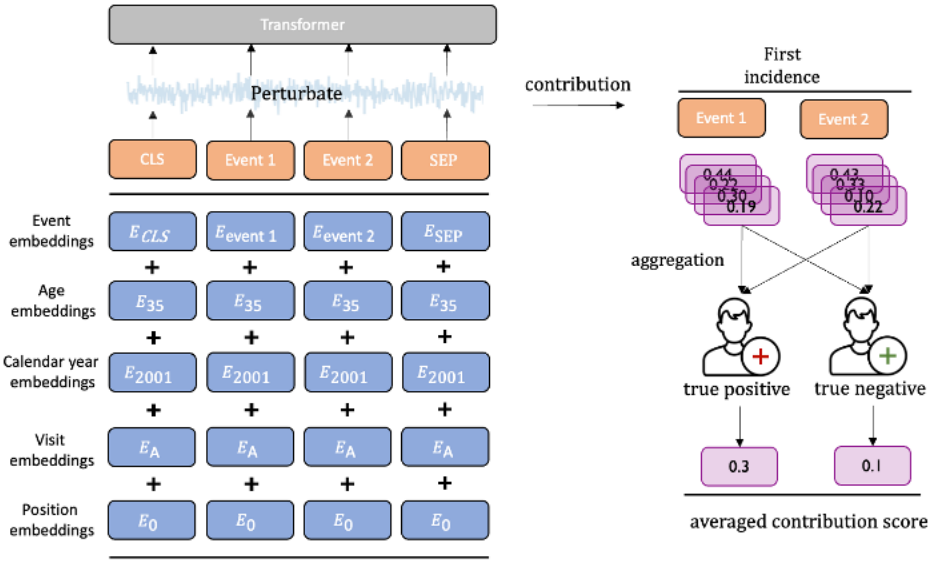



*Figure 5.2: Illustration for conducting population-level relative contribution. We applied trainable perturbation noise to indicate the contribution of each predictor to the prediction. Then the contributions of the first incidence of events are aggregated across true positive and true negative groups to deliver the averaged population-level contribution score.*

#### 5.1.4.9 Association analysis

We randomly allocated 60% of patients to a training dataset and the remainder to a testing dataset (for association analysis). The association analysis used RC to indicate the association between a disease/medication and HF with 95% confidence intervals[176]. It is calculated by dividing the average contribution of the disease/medication in the true positive

(HF) group by the true negative (non-HF) group. RC greater than one and less than one implies a disease/medication is positively and negatively associated with HF, respectively. Additionally, our analyses had an emphasis on patients with a relative confident prediction (predictive probability large than 0.8 and smaller than 0.2) and diseases/medications that were sufficiently trained (with at least 5% prevalence in the population).

In this study, we investigated (1) if BEHRT could capture risk factors that are consistent with the established evidence. To this end, we explicitly validated several established risk factors: hypertension, atrial fibrillation and flutter, myocardial infarction, diabetes, and ischaemic stroke using RC metric; (2) the capability of BEHRT to identify other novel risk and preventative factors. Also, age and calendar year stratified analyses were included to better understand the differential contribution to HF by age and calendar year. More specifically, age was stratified by groups: (50-60], (60-65], (65-70], (70-75], (75,80], and calendar year was stratified by groups: [1990-1995], (1995-2000], (2000-2005], (2005-2010]. For clarification, "()" and "[]" represent exclusion and inclusion, respectively.

### 5.1.5 Results

#### 5.1.5.1 Population characteristics

*Table 5.1: Baseline characteristics for patients in the overall, training, and testing datasets*

| | Overall | Training | Testing |
|---|---|---|---|
| Total number of patients | 100,071 | 60,043 | 40,028 |
| Number of incident cases of heart failure (%) | 13050 (13.0) | 7,853 (13.1) | 5,197 (13.0) |
| Women (%) | 58,331 (58.3) | 35,094 (58.4) | 23,237 (58.1) |
| Men (%) | 41,740 (41.7) | 24,949 (41.6) | 16,791 (41.9) |
| Median follow-up duration (year) | 9 | 9 | 9 |
| Median age (year); Interquartile Range | 70; (59,79) | 70; (59,79) | 70; (59,79) |
| Diabetes Mellitus (%) | 20,598 (20.5) | 12,348 (20.5) | 8,250 (20.6) |
| Hypertension (%) | 65,760 (65.7) | 39,427 (65.6) | 26,333 (65.7) |
| Rheumatoid arthritis (%) | 3,288 (3.3) | 1,936 (3.2) | 1,352 (3.4) |
| Atrial fibrillation and flutter (%) | 26,257 (26.2) | 15,826 (26.3) | 10,431 (26.1) |
| Myocardial infarction (%) | 9,278 (9.3) | 5,588 (9.3) | 3,690 (9.2) |
| Chronic obstructive pulmonary disease (COPD) (%) | 13,897 (13.9) | 8,364 (13.9) | 5,533 (13.8) |
| Ischemic stroke (%) | 5,124 (5.1) | 3,022 (5.0) | 2,102 (5.2) |

The overall, training, and testing datasets included 100,071, 60,043, and 40,028 patients, respectively, for HF risk prediction. Because training and testing were randomly

allocated, Table 5.1 shows that patients within these three datasets have similar baseline characteristics.

#### 5.1.5.2 Predictive performance evaluation

The BEHRT model achieved a great performance for HF risk prediction with 0.930 (95% confidence interval: 0.926, 0.934) and 0.690 (0.667, 0.713) for AUROC and APS, respectively. Without this prerequisite, the model explanation seems less meaningful.

#### 5.1.5.3 Ablation study

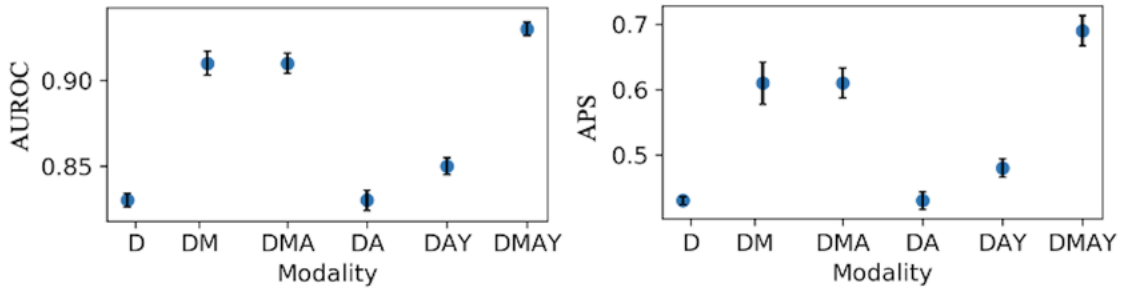



*Figure 5.3: Model performance evaluation in ablation study. The x-axis indicates the inclusion of modalities for BEHRT modelling. The modalities include diagnoses (D), medications (M), age (A), and calendar year (Y).*

The ablation study was conducted to better understand the contribution of different modalities for BEHRT modelling. Figure 5.3 shows that the BEHRT using all available information (i.e., diagnoses, medications, age, and calendar year) for modelling has the best performance, and the inclusion of medications outperforms the model with just diagnoses in both AUROC and APS. Furthermore, the inclusion of calendar year for modelling achieves substantial improvements as compared to using age only. Therefore, the results suggested that medications were important for learning meaningful representations, and calendar year might be more informative for contextualising predictors than chronological age.

#### 5.1.5.4 Embedding analysis

Additionally, by calculating the dissimilarity among age and calendar embeddings, respectively, Figure 5.4 shows that calendar year embeddings have substantial dissimilarity across years (from 0 to 0.8). In contrast, the dissimilarity was less pronounced across ages

with a range between 0 and 0.3. The results indicated that diseases (and medications) in the context of the calendar year show more substantial temporal variability than them in the context of age. Hence, the calendar year was considered more informative for the incident HF prediction than chronological age.

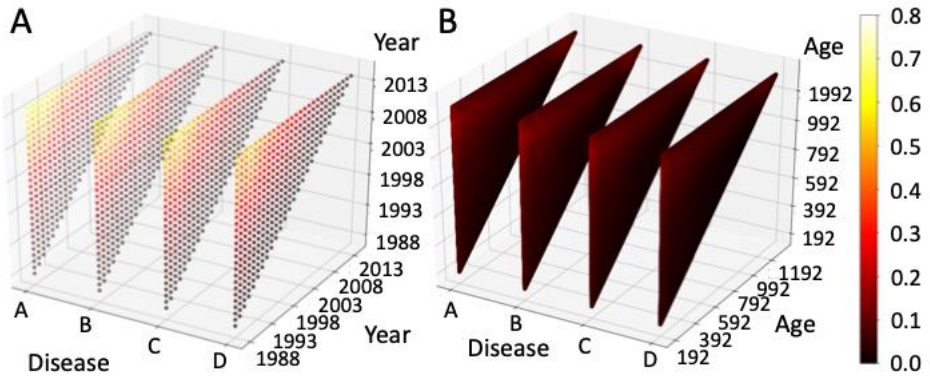



*Figure 5.4: Embedding similarity analysis for age and calendar year embeddings. Figures A and B show cosine similarity measurements for the summed embedding of diseases with different years and ages, respectively. A, B, C, and D represents four randomly selected disease: depression, peripheral arterial disease, anxiety disorders, and hypo or hyperthyroidism, respectively. The age axis represents age in months from 16 to 100 years, and the year axis represents the calendar year from 1988 to 2014. In terms of similarity, smaller values (darker colours) indicate higher similarity and larger values (lighter colours) indicate higher dissimilarity.*

#### 5.1.5.5 Association for established risk factors

*Table 5.2: Relative contribution with 95% confidence interval for medically validated risk factors.*

| Disease | RC | 95% confidence interval | |
|---|---|---|---|
| | | Lower bound | Upper bound |
| Myocardial infarction | 1.37 | 1.28 | 1.48 |
| Ischaemic stroke | 1.29 | 1.09 | 1.53 |
| Atrial fibrillation and flutter | 1.18 | 1.11 | 1.24 |
| Hypertension | 1.08 | 1.03 | 1.14 |
| Diabetes mellitus 1,2 | 1.02 | 0.9 | 1.15 |

Table 5.2 shows that the medically validated risk factors are largely consistent with the expectation, with average RCs greater than one. The same patterns were also seen in the age-stratified analysis (Figure 5.5). These analyses indicated that the BEHRT model associated these diseases with HF. Furthermore, the age-stratified analysis showed the RCs were generally higher for those aged 50-60 years and lower in older ages, implying that the risk factors individually had little contribution to the HF in older ages. This finding aligned with the evidence from past epidemiological studies[177,178].

However, Figure 5.5 also shows that hypertension and diabetes have RCs lower than one in the older age groups. This was against our expectations because both diseases were well-known risk factors for HF. Thus, we hypothesised that the associations of these diseases might be biased because they are contextually associated with the treatment in older age patients. For instance, hypertension is commonly treated with antihypertensive drugs, and these drugs are associated with non-HF. Therefore, the explainer can potentially conclude an explanation that both antihypertensive drugs and, by contextualisation, hypertension are associated with non-HF.

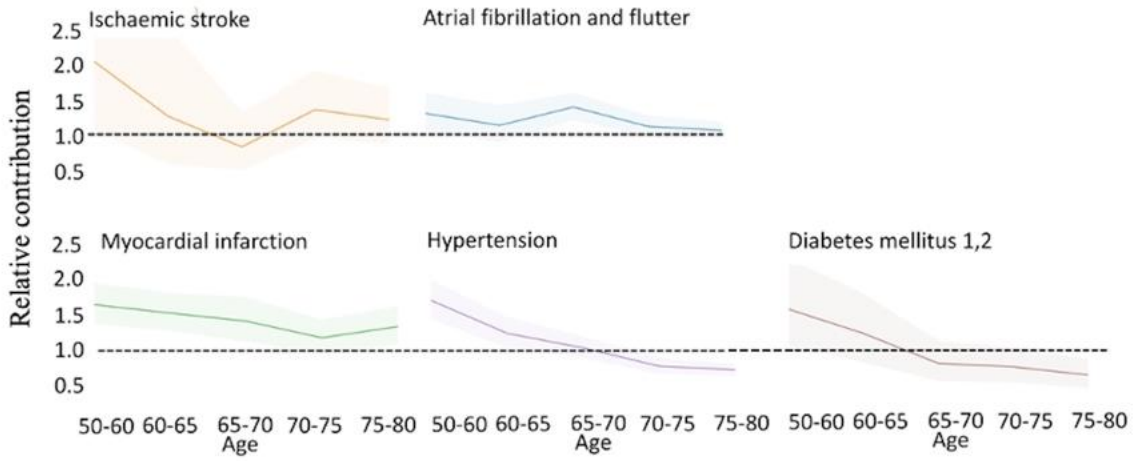



*Figure 5.5: Age-stratified relative contribution with 95% confidence interval for medically validated risk factors. The black line denotes that relative contribution equals to one.*

To further test our hypothesis, we found that 73% of patients with hypertension and older than 65 years and 70% of patients with diabetes were treated with antihypertensive drugs and medications for diabetes, respectively. Furthermore, these treatments that were frequently contextualised with their respective diseases were largely associated with non-HF. As shown in Figure 5.6, a lot of treatments of validated risk factors such as antihypertensive drugs, digoxin, and drugs for diabetes were largely associated with non-HF.

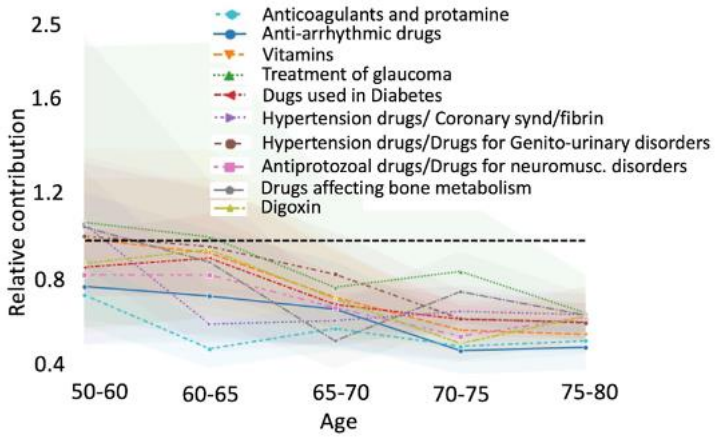



*Figure 5.6: Age-stratified analysis for 10 medications with the lowest relative contributions. The x and y axis represent age group and relative contribution with mean and 95% confidence interval, respectively. Black line represents relative contrition equals to one.*

While the previous analysis showed that treatments of known risk factors were associated with non-HF, we were also interested in understanding how those treatments affected the risk associations of risk factors. To this end, we conducted a stratified analysis on patients within the treated and untreated groups. For example, we derived RC of hypertension in the subgroups of patients that were treated and untreated with antihypertensive drugs as opposed to RC of hypertension in the general population. Figure 5.7 shows that diseases within the treated group generally have lower RC than the general population (16 of 19 cases), indicating less association to HF. While RCs of risk factors were generally lower in treated groups compared to untreated groups, RCs showed a substantial decrease in groups that were treated with antihypertensive drugs, digoxin, and medications for diabetes. Therefore, these experiments demonstrated that BEHRT can naturally capture untreated risk factors associated strongly with HF while treated risk factors are mitigated in risk due to treatment and thus, appropriately associate less with HF.

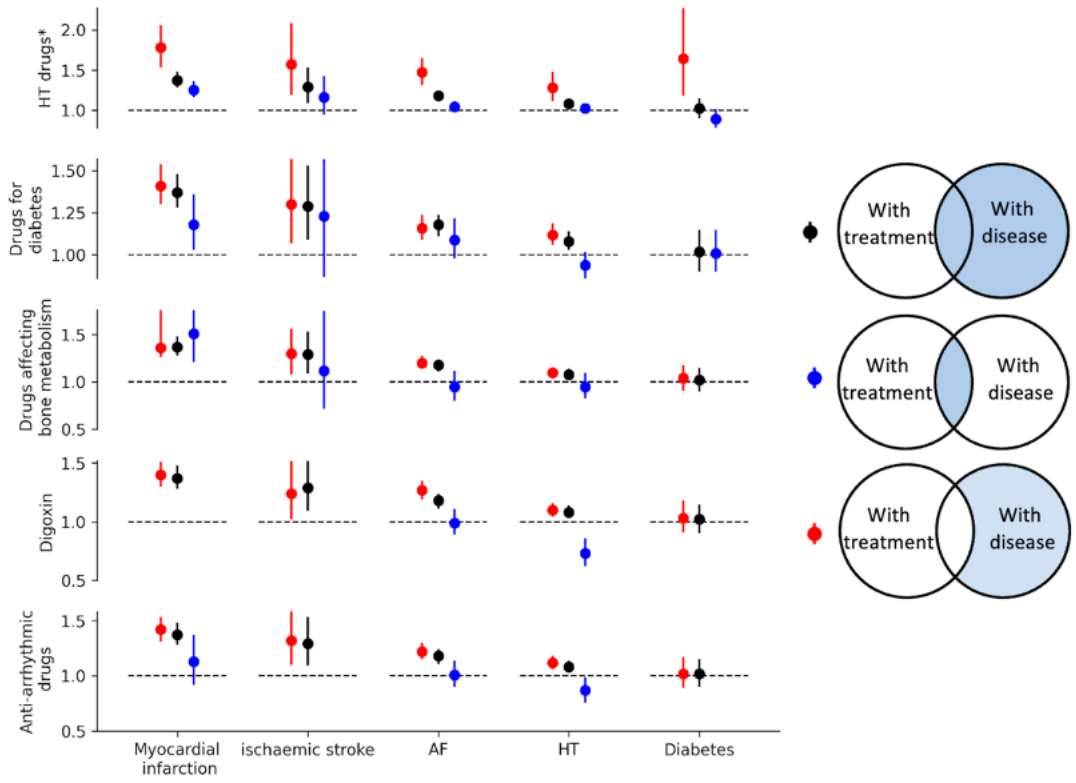



*Figure 5.7: Relative contribution analysis for contextuality of medications and risk factors. Each subplot represents a medication (row) and disease (column) pair. It represents the relative contribution of a disease in general population (black forest plot), untreated population (red forest plot), and treated population (blue forest plot). Some graphs fail to have treated/untreated forest plot due to insufficient sample size.*

### 5.1.5.6 Risk factor identification

After validating our model's capability of identifying known risk factors and capturing the interplay between risk and treatments, we further applied the explainer to discover potential risk factors (with RC) that were independent of experts' knowledge and directly derived from the model. Some examples of the identified risk factors (both diseases and medications) were bacterial diseases, lower respiratory tract infections, myocardial infarction and pleural effusion, corticosteroid and antibacterial drugs, bronchodilators, and acne and rosacea drugs (more examples can be found in Table A.14 and Table A.15). These identified risk factors were also largely consistent with the established evidence[179–182]. Moreover, Figure 5.8 shows an age-stratified analysis for ten diseases and medications with the highest RC. It demonstrated a consistent pattern that risk factors individually showed limited discriminatory contribution in older ages. However, some diseases and medications,

such as left bundle branch block, had wide confidence intervals across different age groups. Thus, it was hard to allow firm conclusions about any differential RC by age.

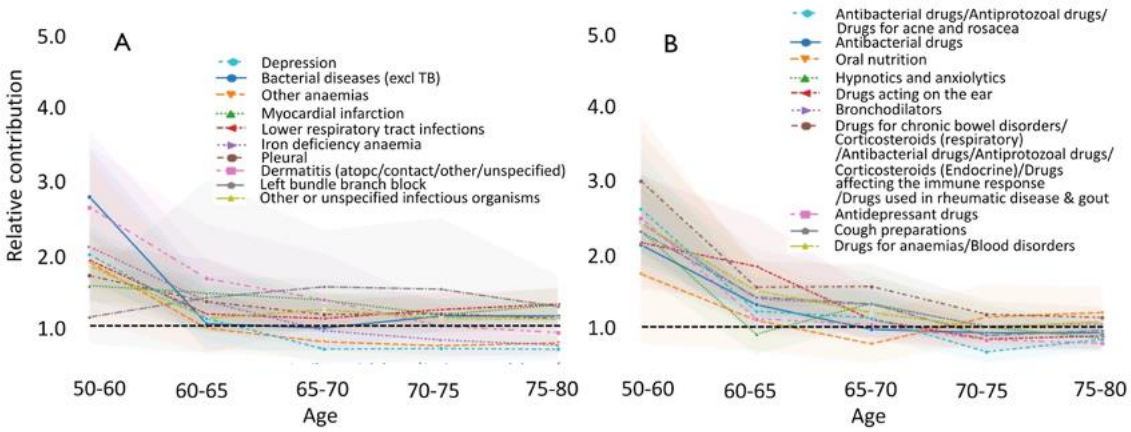



*Figure 5.8: Age-stratified analysis for top 10 diseases (A) and medications (B) with the highest relative contribution. x and y represent age groups and relative contribution with 95% confidence interval, respectively. Black lines denote relative contribution equals to one.*

### 5.1.5.7 Calendar year stratified analysis

Our previous analysis showed that the calendar year also played an important role in risk prediction. Therefore, we further investigated the RC stratified by calendar year. For the convenience of description and presentation, we have analysed two medications with RC less than one depicted in Figure 5.6: digoxin and treatment for glaucoma. Figure 5.9A shows that digoxin is constantly associated with non-HF while the association of treatment of glaucoma is more heterogeneous with RC greater than one before 2000 and less than one after 2000. We hypothesised that the model considered medications at the BNF section level, which was constant across different years. However, the underlying drug composition might have changed around the year 2000, and the model was able to capture such changes when the medication was contextualised with the calendar year.

To this end, we further analysed the prevalence of medications by components in patients between 1990 and 2010 (not counting repeat prescriptions). As shown in Figure 5.9B, for treatment for glaucoma, timolol (beta blocker) was commonly prescribed throughout the 1990's, while in the 2000's, its prevalence started to decline because of the introduction of new medications. The previous evidence also supported the findings.[183,184]

Our RC analysis in Figure 5.9A explicitly demonstrates that BEHRT can capture this change when the treatment of glaucoma is contextualised with the calendar year. Specifically, BEHRT identified that treatment of glaucoma before 2000 was positively associated with HF (RC greater than one), which was consistent with the established evidence that timolol has known cardiovascular side effects such as bradycardia with the potential to exacerbate HF.[185] After 2000, the BEHRT identified that treatment of glaucoma was positively associated with non-HF. This was explained by the change of treatment of glaucoma from timolol to prostaglandin analogues after 2000, which have known vasodilating properties with the potential of reducing cardiovascular risk (e.g., prostaglandin $I_2$). However, large-scale randomised trials to investigate preventative effects are currently lacking.[186–188] In contrast, digoxin was described by only one drug composition, and it had stable RC across different years. Therefore, it supported the hypothesis that digoxin could play a role in the prevention of HF.

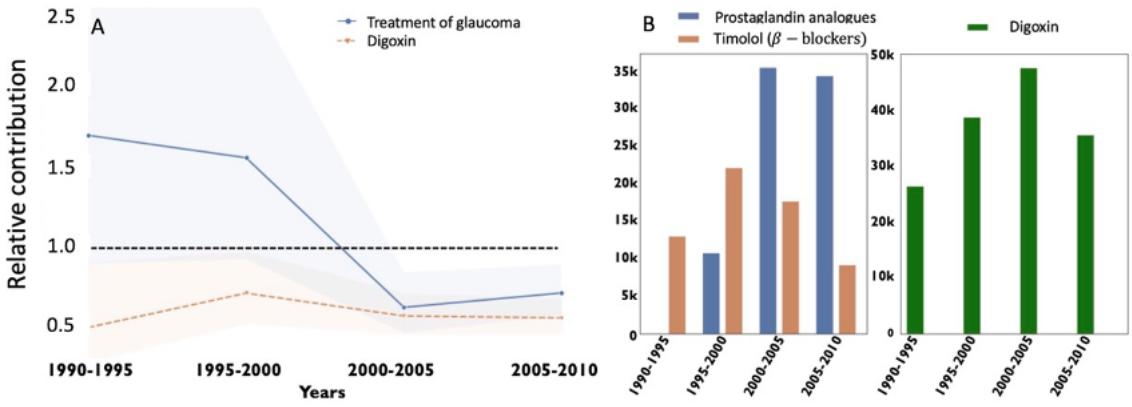



*Figure 5.9: Calendar year-stratified relative contribution analysis for two selected cases. (A) relative contribution of two cases stratified by calendar year. Black line denotes relative contribution equals to one (B) prevalence of drugs by components in different year. x and y represent the year group and counts of first-time drug component prescriptions to patients, respectively.*

### 5.1.6 Discussion

Using large-scale EHR data, BEHRT achieved great performance for incident HF risk prediction. Additionally, it indicated that diagnoses and medications together were strong predictors for risk prediction, and the calendar year was better at capturing temporal variabilities than the chronological age.

While risk prediction is important, the strength of the work is to further provide a post-hoc explanation regarding risk associations. The post-hoc explainer applied to BEHRT not only proved that the BEHRT model was able to identify risk factors that were largely aligned with the established evidence, but also discovered that BEHRT can capture the interplay between risk and treated risk. More specifically, we saw some known risk factors associated with non-HF in some age groups. A cursory analysis might lead to a false conclusion that, for instance, hypertension decreases the HF risk in older age groups. However, the conclusion was biased by indication; the correct interpretation is that the medication serves as a proxy for hypertension, which has a strong effect on HF incidence. Our analysis of risk and treated risk generally demonstrates while risk factors positively associate with HF, treated risk attenuates that association, which is consistent with the medical understanding of HF risk.

Additionally, the explainers also discovered medications with potential preventative effects on HF but have not been considered in expert-driven studies. Through both age and year stratified analysis, we found that digoxin and prostaglandin analogues had stable RC less than one, implying their potential preventative effects on HF. However, as with standard statistical models, a causal interpretation should be made with great caution. Rather, our method was able to generate hypotheses, which depending on the totality of evidence from this work and other sources, should provide the impetus for further confirmatory studies.

Furthermore, the BEHRT with the post-hoc explainer also allowed us to quantitively measured the differential association of medications over different years to HF prediction, which would be missed if the temporal context was not included in modelling. It is well known to clinicians that changes in medications over time or more subtle changes in disease patterns, for instance, due to advances in technologies or policy, can affect risk and influence prognostication. BEHRT was able to consider those changes for modelling.

Our study also has limitations. First, the diagnosis phenotyping method (i.e., CALIBER) restricted our study to 299 disease categories and potentially caused bias in the definition of disease, by extension, the association analysis. Additionally, we kept patients with sufficient information for training and risk prediction during cohort selection. This can lead to a cohort that is less representative of the general population, hence, compromising the generalisability of the model and the association analysis.

## 5.2 Clinical outcome prediction under hypothetical intervention

### 5.2.1 Motivation

Deep learning is commonly referred to as "representation learning" because of its ability to learn useful representation (e.g., of patients, images, or sentences) from raw or minimally processed input data. For most of the deep learning frameworks, we can refer them to as "straight-through representation learning (STRL)": They map an input $x$ to an output $y$, via representation $l$, that is not constrained (e.g., to correspond to certain a priori known concepts). That is, while entries in $l$ might happen to correspond to variables familiar to researcher/users (e.g., presence of a dog in an image, or cardio-metabolic health issues in a patient), there is no guarantee for this to be the case. While the presence of such correspondences may not be important in many prediction-only tasks, in many high-stake decision-making fields (such as those in healthcare), one's inability to understand and interact with such complex representations (e.g., through comparing them against the external knowledge and observation) is essential. Therefore, any solution that can help provide simple insights about such deep learning models, can help domain experts understand the rationale behind the predictions and act with better insights.

Suppose that a doctor is using a deep learning model to predict the risk of an event such as HF. It is common for such a doctor to expect some form of explanation about the model's predictions: What factors have the biggest influence on the predictions? Are there any interventions that can reduce certain risks for some patients? Could a different set of events (such as diagnosis, medication, or intervention) in a patient's record change his/her current predictions?

Based on Pearl and Mackenzie's three-level "causal ladders"[57], one can classify such explanations under the association, intervention, and counterfactual categories. Most current methods (e.g., LIME[50], SHAP[54], including our work as discussed in §5.1) for providing

insights about deep learning models operate at the lowest level of the causal ladder (i.e., association). Despite their utility in explaining important attributes, it is more desirable in practice to better understand how the change in the outcome can result from an exogenous intervention (i.e., explanation at the second and the third level of the causal ladders).

In epidemiology, one popular approach for counterfactual analysis is restriction[55]. It selects patients with similar characteristics, then assuming individuals with the same covariates are comparable across intervention and control groups. However, the selection of covariates highly relies on expert's experience. Additionally, the more covariates we select, the less patients can be matched. This problem is often called curse of dimensionality. Considering the exceptional performance of deep learning in learning abstracted high-level latent representation, we are interested in using deep learning model to learn a small number of highly informative latent covariates. We consider patients who have the same latent covariate share similar characteristics and can be applied for counterfactual analysis.

### 5.2.2 Related work

There are recent developments in machine learning that investigate the explanations at the second and the third levels of the ladder – answering the "do" and "what-if" questions by conducting post-hoc intervention as a separate tool on high-level (latent or human understandable) concepts[189–191]. The nature of counterfactual questions, however, makes a model-free approach implausible (§2.4). Therefore, in a recent study, Koh et al.[58] proposed a framework called the concept bottleneck model. In this approach, instead of the common STRL approaches, the model first learns to predict high-level human specified "concepts" (i.e., intermediate labels provided by humans) from raw inputs such images (i.e., referred to as bottleneck); next, these learned concepts are mapped to the outputs/predictions. This enables a correspondence between latent representations and concepts, which then allows one to conduct interventions on concepts/representations (e.g., what if we change a certain

concept's value from zero to one?) to investigate their effects on the prediction/outcome. One of the drawbacks of this framework is that a complete list of pre-defined concepts is needed (i.e., feature engineering) to ensure sufficient predictive power (i.e., accuracy); the more comprehensive the pre-defined concepts, the better the prediction, also the more difficult to conduct counterfactual analyses because of the curse of dimensionality (referred to as counterfactual plausibility in the following sections).

In counterfactual analyses, the degree to which one can change some aspects of the input should be consistent with the empirical evidence (i.e., "plausible counterfactuals").[55] For instance, we will be less likely to simply add HF to a young patient's counterfactual record, if the empirical evidence suggests that HF almost always happens at older ages; similarly, if the empirical evidence suggests that diabetes is very likely to happen in a particular group of patients with certain characteristics, it is more plausible to add diabetes to such patients' counterfactual records. The assessment of these plausibility criteria can have a varying degree of difficulty, when going from one domain to another. In epidemiology and causal inference, the commonly used methods to ensure the plausibility of counterfactuals are randomisation, adjustment, and propensity score matching.[192] These methods must be applied for cohort selection before causal reasoning and restrict our attention to a well-specified causal question (e.g., how does a concept affect an outcome?). This makes them less suitable for conducting counterfactual explanations for a risk prediction model that works for the general population with multiple interventions of interest. Therefore, a risk model that can provide accurate risk prediction for the general population and with the ability to provide counterfactual explanations for individuals seems to be the most ideal solution for guiding clinical decision-making.

### 5.2.3 Aims

In this study, using the incident five-year HF risk prediction and five-year cancer risk prediction tasks as two exemplary study cases, we aimed to build on recent developments in concept bottleneck models and propose a new framework, partial concept bottleneck (PCB), which can provide accurate risk prediction (i.e., comparable to the black-box models), while enabling counterfactual reasoning under hypothetical interventions. More specifically, we showcased a PCB design for incident HF risk prediction with an emphasis on investigating how having atrial fibrillation (AF) as a comorbidity in patients with hypertension and/or diabetes can influence the risk of HF; and showcased a PCB design for incident cancer risk prediction with an emphasis on investigating the effects of antihypertensive medication use on cancer risk.

### 5.2.4 Methods

#### 5.2.4.1 Data source

This study was conducted using EHR from CPRD linked to HES and ONS. Primary data was available from 1 January 1985 to 31 December 2015, and the linkage was available from 1 January 1998 to 30 September 2015.

#### 5.2.4.2 Study design

In this work, we selected two cohorts for predicting five-year incident HF risk (referred to as HF cohort) and incident cancer risk (referred to as cancer cohort), respectively. We defined the risk prediction task as using all the information available before the baseline date to predict the five-year risk of an outcome after the baseline date. The baseline date is a randomly selected date within a patient's medical history. A patient was identified as positive if diagnosed with the outcome of interest from either general practices and HES records or death registration within a five-year window after the baseline date; A patient was identified as negative if one had records for at least five years after the baseline date and no

outcome of interest was identified during the five-year window. The rest of the patients were considered lost to follow-up before developing outcome of interest by year five with an uncertain outcome (i.e., censored patients).

#### 5.2.4.3 Eligible participants

In this study, we selected patients who contributed to data between 1985-01-01 and 2015-12-31, were linked to HES, and marked as "acceptable" by CPRD quality control. This led to a cohort with four million patients. Additionally, we selected patients with both men and women, aged at least 16 years old, and we excluded censored patients, patients with an outcome of interest identified before the baseline date, and patients who were registered with general practices for less than three years before the baseline date. This led to 1,975,630 and 1,221,589 patients for the HF cohort and cancer cohort, respectively.

#### 5.2.4.4 Outcomes

The incidence of HF or cancer was defined as the first record of HF or cancer recorded in the primary care, hospital admission, and death registration records, respectively. We considered HF as a composite condition of rheumatic HF, hypertensive heart disease with (congestive) HF and renal failure, ischemic cardiomyopathy, chronic cor pulmonale, congestive HF, cardiomyopathy, left ventricular failure, and cardiac, heart, or myocardial failure;[193] and considered cancer as a composite condition of Leukaemia, urinary cancer, bladder cancer, renal cancer, male reproductive cancer, prostate cancer, female reproductive cancer, ovarian cancer, cervical cancer, breast cancer, skin cancer, gastrointestinal cancer, respiratory cancer, other unspecified cancer, lung cancer, head and neck cancer, pancreatic cancer, liver cancer, stomach cancer, oesophageal cancer, rectal cancer, colon cancer, lymphatic cancer, and metastatic cancer. The code list for identifying HF can be found in Table A.7, and the code list for identifying cancer was adopted from the CALIBER repository.

#### 5.2.4.5 Study population

We randomly allocated 70% and 30% of patients to the derivation and validation datasets. This selection procedure led to 1,382,941 and 592,689 patients for the derivation and validation datasets, respectively, in the HF cohort, and 855,192 and 366,397 patients for the derivation and validation datasets, respectively, in the cancer cohort.

#### 5.2.4.6 Data preparation

In this study, we followed the identical procedure for data preparation as mentioned in §4.3. In brief, we incorporated records from diagnoses, medications, lab tests, procedures, BP measurements, BMI, smoking status, and drinking status for modelling. Diagnoses, medications, lab tests, and procedures were used in the format of ICD-10 level four, BNF codes at the section level, Read code, and OPCS codes, respectively. BP and BMI values were categorised by 5 $mmHg$ and 1 $kg/m^2$, respectively. Both drinking and smoking status were recorded as categorical value, including current drinker/smoker, ex, and non. We also calculated the corresponding age in year for each record using event date and date of birth.

#### 5.2.4.7 Partial concept bottleneck (PCB) model

Before introducing PCB, we need to firstly discuss two critical concepts and components in PCB: concept bottleneck model and discrete representation learning.

##### *5.2.4.7.1 Concept bottleneck model*

Unlike the STRL models that predict the output $y$ directly from the input $x$ (i.e., $x \rightarrow y$), the concept bottleneck model[58] constructs an intermediate layer between the $x$ and $y$ with only high-level (human understandable) concepts $c$ (i.e., $x \rightarrow c \rightarrow y$). The concept bottleneck model learns to predict the concepts from the input, and then use the concepts to predict the outputs. Since the intermediate layer is expected to capture all the useful information in $x$ in the form of high-level concepts (i.e., through forming and information bottleneck), it is also called a bottleneck layer. Mapping inputs to the concepts, requires

human-generated annotations for the concepts (somewhat similar to classic feature engineering). The mapping from concepts to the outputs can be done using a classifier (or regressor). Therefore, in theory, one can manipulate the value of concepts and observe their effect on the prediction and hence provide model interpretations and counterfactual explanations.

#### *5.2.4.7.2 Discrete representation learning*

Representation learning, especially for learning continuous representations, enables deep learning models to extract latent features that can be useful for various tasks. One of the challenges of dealing with such representations is the lack of a priori-known correspondence between their values and real-world concepts. That is, their continuous range, can hamper one's ability to find simple meaning and interpretation in their values. Learning discrete representations, on the other hand, can provide a simpler solution for reasoning and scenario analysis in the representation space (e.g., for model interpretation).

Additionally, if assuming patients with the same representations belong to the same patient cluster, in such discrete representation spaces, the number of patient clusters will be substantially smaller (compared to the continuous space), and more interpretable. This allows one to reason about hypothetical interventions in each cluster to further investigate various what-if questions. The main concern regarding the use of discrete (instead of continuous) representations is the loss of information in the mapping and hence less accurate risk estimation. Therefore, it is important to investigate the trade-off between the gain in interpretability and reasoning, against the potential loss in model accuracy.

An effective approach for learning discrete representation is vector quantisation[194,195]. It uses differentiable Gumbel-Softmax distribution:

$$y_i = softmax\left(\frac{\log(\pi_i)+g_i}{\tau}\right) \text{ for } i=1, \ldots, k,$$

to approximate sampling from Bernoulli or categorical distribution with class probability $\pi_i$, a scaling factor $\tau$, and independent and identical noise samples $g_i$ from standard Gumbel distribution. It then uses a codebook to map the discrete representation to the embedding space (as shown in Figure 5.10). A more practical application of discrete representation is to apply quantisation with multiple variable groups[194]. Instead of having one codebook, it chooses quantised representations from multiple codebooks and concatenates them. For a codebook with only two entries (i.e., binary variable), $n$ codebooks can lead to maximally $2^n$ combinations or groups. In this work, $n$ can be tuned as a hyperparameter (based on model performance). Ideally, we would like to have $n$ as small as possible, without substantially sacrificing the model accuracy (i.e., interpretability and accuracy trade-off and the bias-variance trade-off)

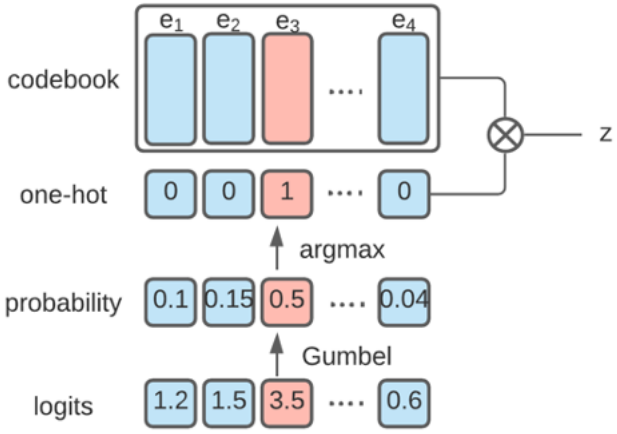



*Figure 5.10: Vector quantisation for a categorical variable. The logits are transformed into a one-hot vector using Gumbel-Softmax quantisation and represented with corresponding embedding representation in the codebook.*

#### *5.2.4.7.3 PCB model*

Concept bottleneck model allows users to interact with the model to investigate the differential effects of certain intervention (on the human understandable concepts). However, the definition and selection of a comprehensive list of concepts (i.e., feature engineering) for accurate prediction are difficult. It not only requires a significant amount of expert labour, but also highly relies on how the underlying causes of a condition are understood today.

Even with a complete list of concepts, due to the curse of dimensionality, it is difficult to guarantee the plausibility of counterfactual for a given concept when holding the other concepts constant. These restrictions hinder the usability of concept bottleneck models in EHR applications for counterfactual analysis.

Building on concept bottleneck, we propose PCB (as shown in Figure 5.11), which substantially relaxes the restrictions that one faces when using concept bottleneck models for counterfactual reasoning. It modifies concept bottleneck models of representation learning, by introducing a combination of discrete latent representation $l$, and concept representation $c$, which corresponds to human-defined concepts of interest (for hypothetical interventions). Compared to the simple concept bottleneck models, which require a large amount of work on concept annotation and a robust list of concepts to preserve predictive performance, $l$ takes care of the rest of the latent feature extraction with relatively small number of latent dimensions (alleviate the curse of dimensionality). This can be applied as a strategy for latent feature matching and further lead to the capability of conducting counterfactual inference (by intervening with $c$).

The PCB can be trained in a similar manner to the concept bottleneck models. It includes three components, a function $g(x)$ that maps $x$ to $c$, a function $h(x)$ that maps $x$ to $l$, and a classifier $f(c, l)$ that predicts $y$ based on $c$ and $l$. This results in the following optimisation objective:

$$\hat{f}, \hat{g}, \hat{h} = argmin_{f,g,h} \left(L_Y(f(c, h(x)), y) + L_C(g(x), c)\right),$$

Where $L_Y$ and $L_C$ represent loss function for the main objective and concepts, respectively. While $\hat{f}$ is trained with the ground truth concepts, it still takes the $\hat{g}(x)$ as input at test time. Additionally, the objective can be modified as below if having a shared model $m(\cdot)$ for representation learning before mapping to $c$ and $l$:

$$\hat{f}, \hat{g}, \hat{m}, \hat{h} = argmin_{f,g,m,h} \left(L_Y(f(c, h(m(x))), y) + L_C(g(m(x)), c)\right).$$

Ideally, $m(\cdot)$, $h(\cdot)$, $g(\cdot)$, and $f(\cdot)$ can be any model that serves the purpose.

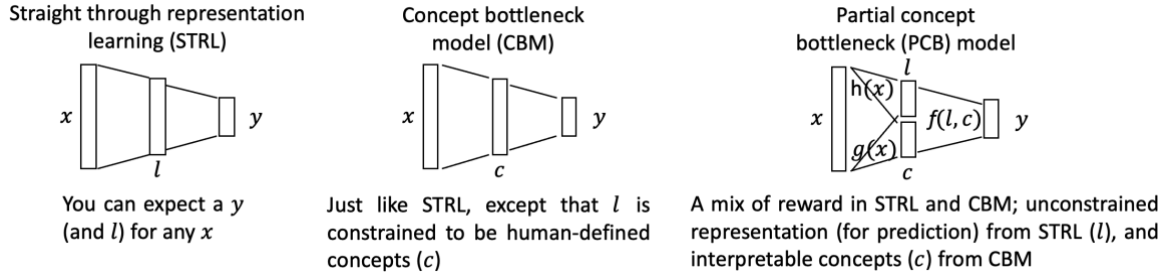



*Figure 5.11: Architecture illustration for STRL, concept bottleneck model, and PCB model. STRL maps $x$ to $y$ through latent representation learning $l$, hence, this entire process is hardly explainable. By contrast, concept bottleneck model maps $x$ to human-defined concepts $c$, then uses $c$ to predict $y$. Therefore, the bottleneck layer $c$ is human understandable and user can interact with concepts to investigate how the change of a concept can affect the outcome. PBC combines representation learning $l$ and human-defined concepts $c$. Similarly, user can interact with $c$ to investigate the differential outcomes caused by the change of $c$, and $l$ takes care of latent representation to maximally preserve prediction performance. Note that, there is no guarantee that all interactions are plausible, and the plausibility of interaction separates the counterfactual analysis from the association analysis.*

#### *5.2.4.7.4 Counterfactual analysis using PCB*

Counterfactual prediction refers to the prediction of an outcome for a patient under two (or more) hypothetical intervention. Since a patient cannot have two outcomes at once, counterfactual prediction is typically achieved by learning a distribution of outcomes from two or more groups of patients who share similar characteristics (or covariates) but undertook different interventions. These two conditions are summarised as positivity (individual has a positive probability of receiving all interventions) and conditional exchangeability (individuals with the same covariates are compared across different interventions).

The PCB model establishes a relationship from $x$ to $y$ through $c$ and $l$ (i.e., $x \rightarrow c, l \rightarrow y$), where the path for concepts $c$ is not contaminated with other variables. Therefore, we assume patients with the same latent feature $l$ share similar characteristics and outcome $y$ only depends on $c$. Therefore, any modification or intervention (referred to as $do(\cdot)$) on $c$ is a counterfactual from the model's perspective ($p(y|do(c), l)$).

Additionally, the counterfactual questions should satisfy the counterfactual plausibility (i.e., positivity)[55]. Thus, the hypothetical intervention cannot contradict the facts

(i.e., observations), and in the PCB model, the facts are restricted by $l$. For a specific cluster of patients with latent representation $l_i$ ($i = 1, 2, 3, \dots, 2^n$), only plausible interventions can be conducted, which can be represented as $p(y|do(c), l = l_i)$.

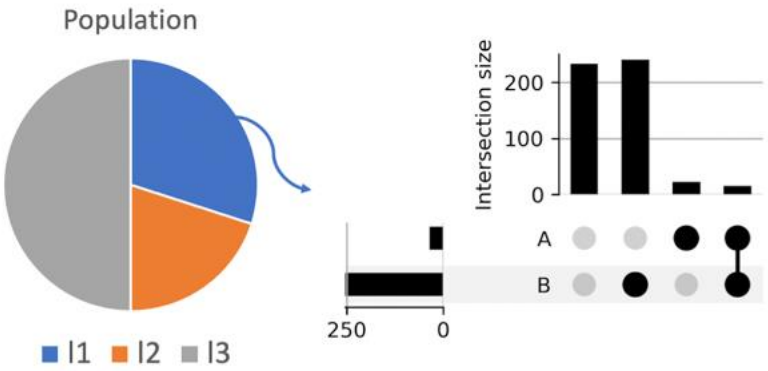



*Figure 5.12: Illustration of the upset plot for analysing counterfactual plausibility. In this toy example, patients are clustered into three groups (i.e., $l_1$, $l_2$, and $l_3$) by a PCB model with two concepts of interest (A and B) for a hypothetical population (left). The black and grey dots in UpSet plot (right) represent exposure and non-exposure to a concept, respectively. For patients in the group $l_1$, the UpSet plot shows (right) that 250 and approximately ten patients have B (+) and A (+), respectively, on the x-axis, while the numbers for A (-) and B (-) are unknown. Most of the patients within this group are A (-) (approximate 220 patients), and A (-) and B (+) (about 225 patients), as shown on the y-axis (intersection size). The summation of the intersection size across sets is the total number of patients within the group.*

In this work, we employ UpSet plot[196] as the systematic solution for analysing the plausibility of counterfactuals using empirical evidence. As shown in Figure 5.12, the UpSet plot carries out a comprehensive descriptive analysis, illustrating the possibility and probability of various counterfactuals that can happen for a specific cluster of patients. For example, for a patient who is A (-) and within the cluster $l_1$ in Figure 5.12, it is plausible to ask the counterfactual question: what would be the outcome if the patient is B (+) instead of B (-)? In contrast, although it is possible to ask what would be the outcome if the patient is A (+) instead of A (-) for a patient who is B (-) and within cluster $l_1$? Due to the rare prevalence in the alternative scenarios, we might consider the counterfactual as less reliable. Therefore, when asking a counterfactual question, we can consider a counterfactual is plausible if its prevalence in the group is within a certain range (i.e., threshold). The range can vary depending on the purpose of the task and what is the question at hand. 100% and 0% prevalence of a concept or concept combination in a cluster means that the counterfactual

is impossible for the specific cluster, hence, implausible. Instead of using a threshold, we can also rely on the confidence interval of the risk estimation because it reflects the uncertainty of the interpretation, the plausibility of a counterfactual, and the size for conducting the estimation (wider confidence interval for smaller prevalence).

#### 5.2.4.8 Model derivation

In this work, we proposed to use the PCB model as a solution to provide accurate risk prediction (i.e., comparable to the black-box baseline model), while enabling counterfactual reasoning. We trained all models on two P100 GPUs with a batch size of 128 per GPU, giving a total batch size of 256. All model was trained identically for 100 epochs with 10%, 40%, and 50% for warm-up, hold, and cosine decay, respectively. The learning rate for the hold stage was $1e^{-4}$, and the early stop was applied when loss did not decrease for ten epochs.

**Standard black-box model**: To this end, we adopted Hi-BEHRT and its parameters as the baseline end-to-end black-box model for comparison. As introduced in §4.3, Hi-BEHRT is a hierarchical BEHRT model that makes accurate risk prediction by incorporating a patient's complete medical trajectory. More specifically, it had four layers of feature extractor and four layers of feature aggregator, hidden dimension 150, number of attention heads five, intermediate size 108, dropout rate 0.2, maximum sequence length 1220, and 50 and 30 for window size, and stride size, respectively.

**PCB (AF-HF)**: In this study, we showcased a PCB design for incident HF risk prediction with an emphasis on investigating how having AF as a comorbidity in patients with hypertension and/or diabetes can influence the risk of HF. To this end, the concepts of the PCB model were defined as AF, hypertension, and diabetes. We defined a concept as

positive (or 1) for a patient if the corresponding condition (a list of codes defined by the CALIBER repository) is recorded in the primary care or hospital admission before the baseline date. Otherwise, a concept was negative (or 0). In this experiment, we selected diseases that were well-studied in medical research. Therefore, we can validate whether the interpretations from the proposed framework align with the established evidence.

To make the PCB model comparable to the baseline black-box model (i.e., Hi-BEHRT), we implemented a similar architecture. More specifically, we constructed the initial representation learning architecture $m(\cdot)$ by adopting the Hi-BEHRT model architecture and its parameters. Therefore, our PCB model took the identical input data as the Hi-BEHRT model. Next, a two-layer MLP $g(\cdot)$ and a vector quantisation component $h(\cdot)$ were trained to map $m(x)$ to $c$ and $l$, respectively. The MLP had 64 units, and the number of units was equivalent to the number of concepts for the first and the second layer, respectively. The Vector quantisation was employed with $n$ codebooks, and each codebook had two entries in this study. Therefore, we considered each codebook as a binary variable (with 0 and 1 representing two entries, respectively), and each entry corresponded to a codebook vector. Therefore, for the convenience of visualisation, we present the latent representation of the vector quantisation as an $n$-dimensional binary vector in the remaining work. Additionally, to have the output dimension of vector quantisation at a similar level as the concepts of interest, we concatenated the selected codebook vectors and applied a linear transformation to obtain an n-dimensional latent representation (the same as the number of codebooks) as a part of the hyperparameter tuning process. The classifier $f(\cdot)$ that predicts $y$ based on $c$ and $l$ was implemented as a three-layer MLP with 16, 8, and 1 unit for each layer, respectively.

We searched the number of codebooks $n$ over [4, 6, 8], and 6 was selected for PCB (AF-HF) as it achieved the best model performance with minimal $n$. Also, although the high-

level concepts of interest in this experiment can be extracted using rule-based methods, we believe it is still beneficial to show that the rule can be learned by the neural networks, which can be informative as a framework for other works in which the concepts cannot be extracted by rule-based methods.

**PCB (HTN-cancer)**: In this experiment, we showcased a PCB design for incident cancer risk prediction with an emphasis on investigating the effects of antihypertensive medication use on cancer risk. To this end, the concepts of the PCB model were defined as angiotensin-converting enzyme inhibitors (ACEI), angiotensin II receptor blockers (ARB), β blockers (BB), calcium channel blockers (CCB), and diuretics, which are commonly used antihypertensive treatments. We defined a concept as positive if a patient received the corresponding treatment before the baseline date, otherwise, a concept was negative. PCB (HTN-cancer) had a similar architecture as the PCB (AF-HF) and was trained the same way as PCB (AF-HF). However, instead of using a neural network to learn the rules for concept extraction, we used an expert-defined (rule-based) phenotyping method[81] provided by the Univerisyt of Bristol. The rest of the model architecture remained the same as PCB (AF-HF). Additionally, we searched the number of codebooks $n$ over [8, 10], and 8 was selected for PCB (HTN-cancer).

#### 5.2.4.9 Model validation

We evaluated the baseline Hi-BEHRT model and PCB model on the validation set using AUROC and APS. The purpose was to demonstrate that the PCB can achieve comparable risk prediction performance as the standard baseline deep learning model. Without this prerequisite, the utility of PCB seems less meaningful even it can provide counterfactual explanations. We also evaluated the accuracy for concepts prediction using the F1 score, which validated whether the network can learn the rule-based phenotyping for

concept extraction. For counterfactual analysis, we applied the UpSet plot to analyse the plausibility of different counterfactuals and reported the risk difference (RD) and risk ratio (RR) for each counterfactual. RD and RR are commonly used epidemiological metrics for measuring the association between and exposure and an outcome by comparing the risk of an outcome in the exposed group to the risk of an outcome in the unexposed group. RR greater than one and less than one (RD greater than zero and less than zero) indicates the risk of an outcome is increased and decreased by an exposure, respectively. The sanity check for counterfactual analysis was conducted by comparing the consistency between the estimated risk ratio and the observed (unadjusted) risk ratio for a counterfactual scenario in each group, as well as validating the counterfactual interpretation with the experts' knowledge and previous established evidence.

### 5.2.5 Results

#### 5.2.5.1 Population characteristics

Expectedly, the patient characteristics are very similar in the derivation and validation datasets (as shown in Table 5.3) since patients in each dataset were randomly allocated.

*Table 5.3: Descriptive analysis of HF and cancer cohort.*

| | HF cohort | | Cancer cohort | |
|---|---|---|---|---|
| | Training | Validation | Training | Validation |
| Total number of patients | 1,382,941 | 592,689 | 855,192 | 366,397 |
| Number of incident cases of heart failure (%) | 65,949 (4.7) | 28,546 (4.8) | 124,481 (14.5) | 53,465 (14.6) |
| Men (%) | 567,832 (41.0) | 243,032 (41.0) | 411,240 (48.1) | 175,673 (47.9) |
| Baseline characteristics | | | | |
| Median (IQR) No. of visits per patient | 38 (64) | 38 (64) | 16 (33) | 16 (33) |
| Median (IQR) No. of codes per visit | 4.23 (1.45) | 4.23 (1.45) | 3.21 (1.17) | 3.2 (1.17) |
| Mean (SD) learning period before baseline (year) | 8.38 (3.99) | 8.37 (3.99) | 8.38 (5.29) | 8.38 (5.30) |
| Median (IQR) baseline age (year) | 52 (29) | 52 (29) | 50 (26) | 50 (26) |
| Hypertension, n (%) | 303,504 (21.9) | 129,815 (21.9) | - | - |
| Diabetes, n (%) | 91,599 (6.6) | 39,329 (6.6) | - | - |
| Atrial fibrillation, n (%) | 44,683 (3.2) | 19,216 (3.2) | - | - |
| ACEI, n (%) | - | - | 47,179 (5.5) | 19,964 (5.4) |

| | | | | |
|---|---|---|---|---|
| BB, n (%) | - | - | 44,048 (5.2) | 18,737 (5.1) |
| CCB, n (%) | - | - | 43,773 (5.1) | 18,607 (5.1) |
| Diuretics, n (%) | - | - | 10,737 (1.3) | 4,573 (1.2) |
| ARB, n (%) | - | - | 14,116 (1.7) | 5,887 (1.6) |

SD: standard deviation, IQR: interquartile range

### 5.2.5.2 Model performance evaluation

To the best of our knowledge, there is no similar work to the proposed approach. Hence, we mainly compared the PCB model to the standard black-box model. Table 5.4 shows that the PCB model achieves comparable performance to the standard end-to-end black box models on both five-year incident HF risk prediction and cancer risk prediction. Additionally, PCB (AF-HF) can accurately predict each concept with F1 scores 0.991, 0.996, and 0.994 for AF, diabetes, and hypertension, respectively. While the result proves the viability of PCB in risk prediction, their strength is to further provide counterfactual explanations.

*Table 5.4: Model performance for incident HF risk prediction and cancer risk prediction. 95% confidence intervals are calculated over models trained on 5 random seeds. Overall, PCB achieves comparable performance as to standard end-to-end models.*

| | HF risk prediction | | Cancer risk prediction | |
|---|---|---|---|---|
| | AUROC | APS | AUROC | APS |
| Hi-BEHRT | 0.96 (± 0.01) | 0.77 (± 0.01) | 0.72 (±0.01) | 0.33 (±0.01) |
| PCB | 0.95 (± 0.00) | 0.75 (± 0.01) | 0.72 (±0.01) | 0.32 (±0.01) |

### 5.2.5.3 Counterfactual analysis for PCB (AF-HF)

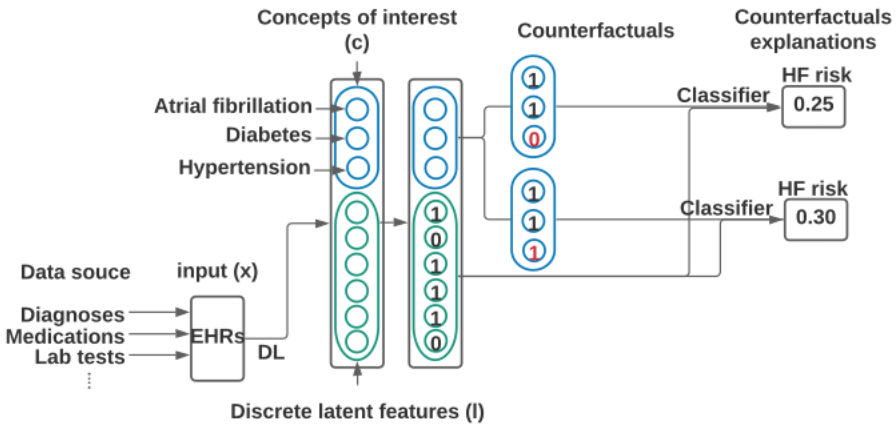
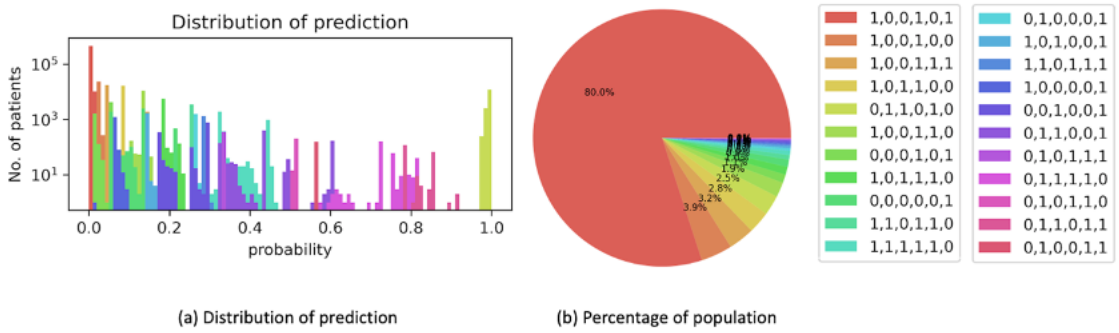



*Figure 5.13: Example of counterfactual explanations, where we intervene on the concepts of interest to investigate the "what if" question. For a group of patients with the latent feature "1,0,1,1,1,0", the model*

*thinks having hypertension as a comorbidity of atrial fibrillation and diabetes would increase the risk of HF by 20%. DL here represents deep learning model.*

In this experiment, we applied PCB model to HF risk prediction on the population with AF, diabetes, and hypertension as the concepts of interest. Figure 5.13 provides an example of conducting counterfactual analyses by intervening with the concepts of interest. Note that, counterfactual analysis was conducted at the (patient) cluster level rather than the individual level. Using the empirical data, Figure 5.14 further shows PCB model phenotypes the entire population into 22 clusters with a six-dimensional latent representation (defined by vector quantisation), and the majority of patients are mapped to the low-risk group, which is expected as HF only has 4.8% positive rate in the validation dataset.




(a) Distribution of prediction (b) Percentage of population

*Figure 5.14: Descriptive analysis for patients within different latent groups. Patients with the same latent feature are considered as the same latent group. (a) Risk distribution of predicted HF risk on the validation set. (b) proportion of latent groups within the validation cohort. Colours represent different latent groups.*

Using latent group "0,0,0,0,0,1" as an example (more examples can be found in Figure A.4), Figure 5.15 shows that all patients phenotyped to this group have diabetes. Thus, it is not plausible to ask counterfactual questions "what if a patient does not have diabetes?". Additionally, our PCB model indicated that patients within this group are at high risk of developing hypertension and this can substantially increase the HF risk; there was a relatively low chance of developing AF, however, the risk of HF in patient with AF could increase by about 300% compared with someone without AF. Therefore, the counterfactual analysis highlighted the potential role of preventing hypertension and AF on the downstream risk of HF in this specific patient group with diabetes.

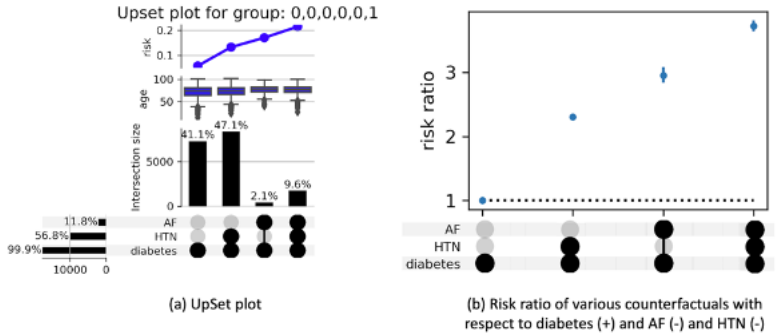



*Figure 5.15: Counterfactual analysis for patients within latent group "0,0,0,0,0,1". (a) UpSet plot analysis with the provision of age distribution and predicted HF risk for counterfactuals. (b) HF Risk ratio of counterfactuals with respect to AF (-) and HTN (-) and diabetes (+). HTN represents hypertension.*

In addition to providing group specific analysis, we can aggregate analyses across groups with plausible counterfactuals to carry out population level analyses. For example, to investigate the effect of AF on HF risk for patients with diabetes and/or hypertension, we identified six, four, and three groups with plausible counterfactuals for patients with diabetes and hypertension, hypertension only, and diabetes only, respectively, as shown in Figure 5.16.

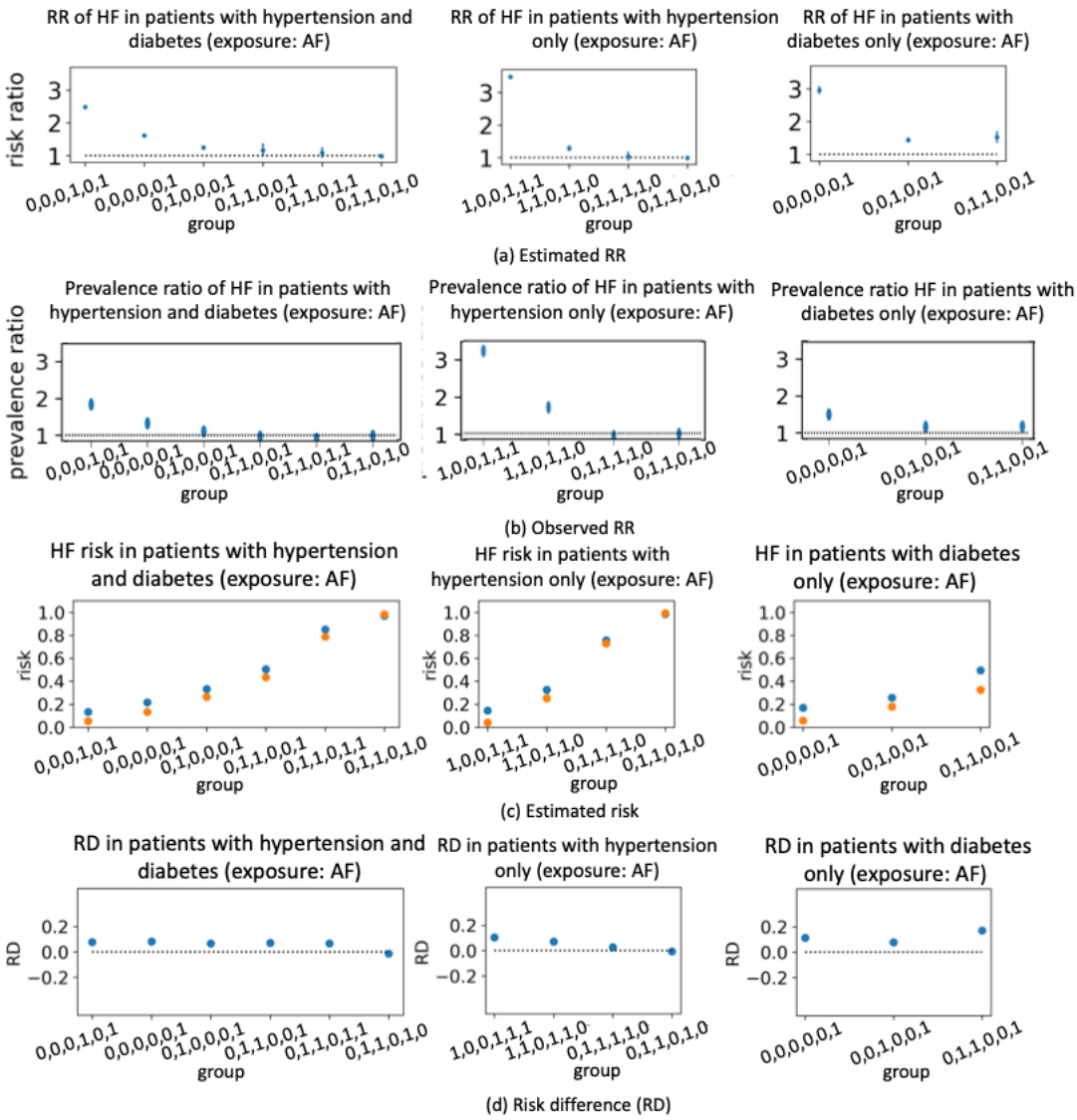



*Figure 5.16: RR, prevalence ratio (unadjusted RR), estimated risk, and risk difference (RD) of HF across various latent groups with AF as the exposure (sorted by ascending order in terms of the HF risk in each group). (a) RR is calculated by estimated risk. (b) prevalence (observed risk) ratio of HF in the exposure group compared to the unexposed group. (c) estimated HF risk with blue and orange dots representing exposure and non-exposure group, respectively. (d) RD calculates the risk difference between the exposure and non-exposure group.*

There are several interesting patterns. First, Figure 5.16(a) shows that AF can substantially increase the risk of HF (RR>1) across almost all groups that have AF as a plausible counterfactual, and the confidence intervals are relatively high for groups with high baseline HF risk due to relatively small sample size for the estimation. This is largely aligned with the previous medical research that AF is associated with an approximately 3-fold to 5-fold increase in the risk of HF.[197,198] Also, as an additional plausibility check, we provided the observed (unadjusted) RRs for reference (as shown in Figure 5.16(b)). They

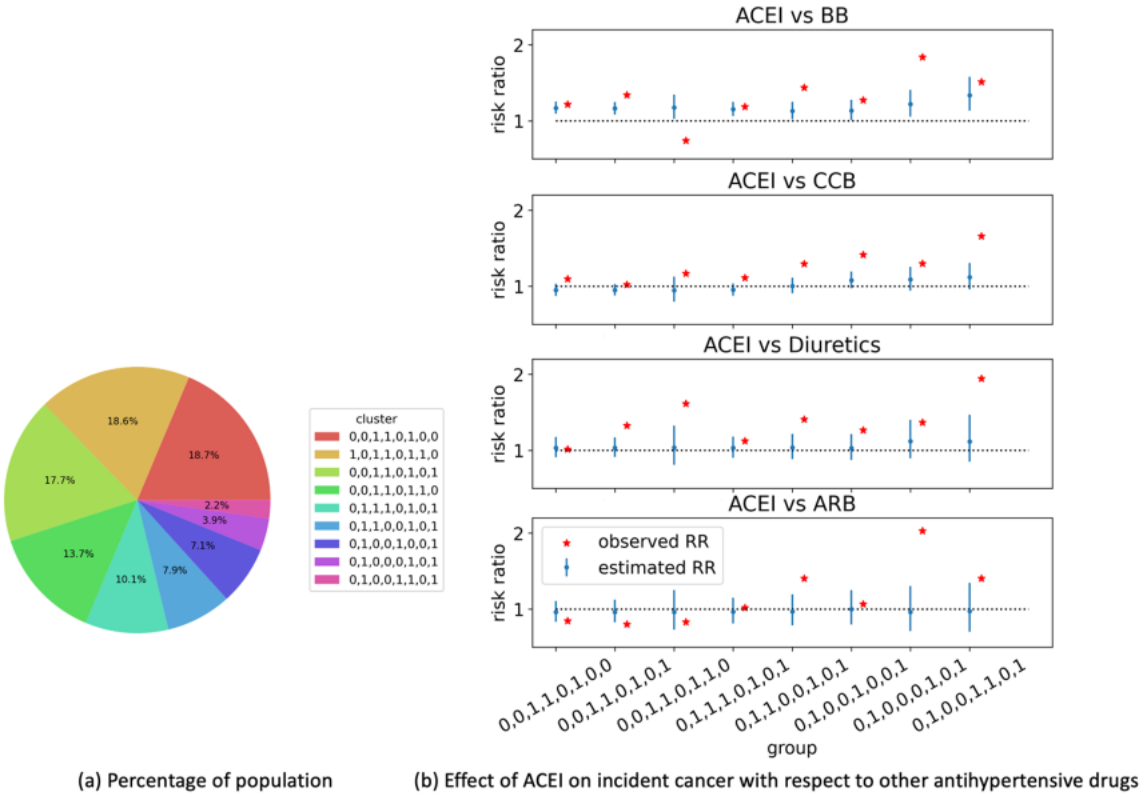
showed a consistent trend as the estimated RRs, and both were aligned with the established evidence.

Furthermore, we found that the RR decreased as the baseline HF risk increased (Figure 5.16(a)). This was consistent with evidence from past epidemiological studies that risk factors conferred greater relative risk in younger (or disease-free) compared with older ages (or multimorbid).[177,178] As a variety of factors, including age, can exacerbate the risk of HF, our analysis implies that the baseline risk might be the underlying reason for those observations. Expectedly, the baseline risk played an important role in RR calculation. A similar magnitude of the increased risk (in the exposure group) can lead to less substantial RR with the increase of the baseline risk.

Therefore, we further included the RD analysis. As shown in Figure 5.16(d), the impact of AF on HF (both RR and RD) clearly decreases as the baseline HF risk increases for patients with hypertension only. However, for patients with diabetes (including hypertension and diabetes, or diabetes only), although the RRs decrease with the increase of baseline HF risk, the RDs are relatively stable across different HF risk groups. This implies that AF as a comorbidity of diabetes can constantly and consistently increase the HF risk for patients within almost all (HF) risk groups and highlights the importance of preventative efforts for AF across the trajectory of HF development in patients with diabetes.

Additionally, Figure 5.16 shows that patients with only diabetes or hypertension as a comorbidity of AF are phenotyped to completely different groups without intersection. This potentially suggests that these two groups of patients can have different trajectories or underlying mechanisms that cause HF, hence, requiring different prevention strategies. Of course, this will need further confirmation in future medical studies.

### 5.2.5.4 Counterfactual analysis for PCB (HTN-cancer)




(a) Percentage of population (b) Effect of ACEI on incident cancer with respect to other antihypertensive drugs

*Figure 5.17: (a) Proportion of patients within each latent group. (b) Effects of ACEI on incident cancer with respect to BB, CCB, diuretics, and ARB (sorted by ascending order in terms of the cancer risk in each group). We demonstrated RR estimates from PCB with 95% confidence intervals on latent groups with plausible counterfactuals. The observed RRs are represented as red star marks for reference. The ground truth is assumed to be RR=1.0 (null) for all four associations validated by a meta-analysis from randomised control trials*[199].

Using the empirical data, Figure 5.17(a) shows that the PCB model phenotypes the population into nine latent groups with an eight-dimensional binary discrete latent representation (defined by vector quantisation). Among those, eight groups had plausible counterfactuals to investigate the effects of ACEI on incident cancer with respect to the other antihypertensive drugs. While an established meta-analysis[199] found no evidence of differential effects of antihypertensive medication use on cancer risk, some research regarding BB exhibited potential anticancer effects[200,201]. Such observation was highly consistent with our analysis (as shown in Figure 5.17(b)) that three out of four comparisons covered the null hypothesis (RR=1) with exception of BB, which requires further confirmation. Compared to the estimated RR from PCB, the observed RR often tended away

from the ground truth, implying a preventative or harmful effect, which contradicted the established evidence.

### 5.2.6 Discussion

In this study, we proposed a PCB framework, which uses a combination of discrete latent representations (as auxiliary predictive and phenotyping features), and high-level human-understandable concepts (as intervention candidates). It achieved a similar predictive accuracy to its black-box counterpart, while providing counterfactual explanations. Furthermore, it avoided intensive expert-driven feature engineering, by focusing on untangling the relations between the outcome and the concepts of interest. By automatically phenotyping patients into different clusters using the latent representations (i.e., latent feature matching), practitioners can investigate various counterfactuals for patients within a specific group to assist decision-making (e.g., provide potential actional suggestions), or get a better understanding of the model's inner workings.

A risk model that experts can interact with to investigate what-if scenarios can be a natural fit in precision medicine and might be the most practical tool to assist decision-making. However, most of the previous studies usually consider risk prediction and model interpretability as two individual components. They focused on delivering accurate risk prediction first, then conducting post-hoc interpretation using for example, perturbation-based methods[173], saliency mas[172], and local or global surrogate methods[50,202,203]. Without expert knowledge as guidance during model training, it is difficult to drive the post-hoc interpretation towards a specific question. The best interpretation these methods can provide is the association between a feature and the outcome based on the observations. Despite their usefulness in generating causal hypotheses to guide further clinical investigations, the interpretations fail to assist clinical decision-making or provide actionable suggestions. These can be more important and desired properties in risk prediction.

On the contrary, epidemiologists have a particular interest in understanding the what-if question for a well-specified causal claim: How an exposure A causes the outcome B to change. This is achieved by conducting a randomised clinical trial (RCT) or emulating RCTs using observational data. Although common causal frameworks secure the interpretation of the results, they are not suitable for delivering a precise risk estimation on the general population. One of the major contributions of our work is that we embedded the causal framework within the risk prediction model and proposed a potential solution to fill the gap between prediction and causality. The proposed model can provide personalised risk prediction and counterfactual explanations (based on the group a patient belongs to). Additionally, instead of considering the concept of causality as emulating the RCTs, we proposed to think of causation in a pluralistic perspective[204] and used the empirical data to inform the possibility and plausibility of counterfactuals. This framework can not only answer our question regarding one exposure, but also can be as flexible as necessary to solve more complex and multi-dimensional questions of interest.

Nevertheless, this study also has several limitations. More hyperparameter searching for the PCB models could have been considered, even though the reported models have achieved reasonably high performance and can lead to reasonable counterfactual analysis. Furthermore, the concepts of interest and latent representation are not independent. Thus, the descriptive analysis (i.e., UpSet plot) on the plausibility of counterfactuals is mandatory, and not all counterfactuals are plausible to deliver explanations. Another limitation is that the (latent) group phenotyping is achieved by latent representation learning, which makes the mechanisms for phenotyping relatively difficult to explain. To alleviate this, for instance, a descriptive analysis of patients' baseline characteristics within each cluster can be conducted to gain more insights and profile the patient groups.

A model that can achieve accurate HF risk prediction and with the ability to provide what-if explanations is highly desired for disease prevention and patient management. By deploying such a model, physicians can assess a patient's current HF risk and hypothetically investigate how different actions can influence the risk of HF and take the action that can benefit the outcome the most. Although this study mainly validated the proposed framework using established medical questions, it will be applied to investigate more clinical meaningful cases in the future.

## 5.3 Conclusion and potential future works

In this chapter, we introduced two methods to retrieve medical explanations (or knowledge) from deep learning models at a different level of the causal ladders. Explanations at different levels do not necessarily mean one explanation is better than the other. They are designed for different purposes, thus having different implications in healthcare.

The post-hoc association analysis showed that a deep learning model can provide a data-driven perspective to untangle the associations between medication events (e.g., diseases and medications). Using HF as an example, we demonstrated that the deep learning model was not only able to identify risk factors that were largely aligned with the established evidence, but also new associations which have not been considered in expert-driven research. This method proves to have a great potential to generate hypotheses for risk factor related medication research, especially for conditions that are still poorly understood.

In addition to association, another important medical question in clinical care is "what-if": How can the change in a certain scenario affect the outcome differently? This would provide counterfactual explanations that inform clinicians to make decisions that benefit the patient the most. Counterfactual reasoning is equally important as risk prediction in decision-making; however, they were considered as two independent topics in the past. The PCB framework described in this chapter paves the way for combining both together to better inform clinical decision-making.

Additionally, the findings in this chapter mainly rely on CPRD using limited study cases, and the generalisability of the proposed methods to other study cases and datasets is still unclear. It would be highly desired to conduct more comprehensive validation in future works.

# 6. Uncertainty in risk prediction

This chapter will focus on the third objective of this doctoral thesis to investigate the role of uncertainty in risk prediction. Parts of this work have been published in Scientific reports at: https://www.nature.com/articles/s41598-021-00144-6. The manuscript benefitted from the input of several co-authors. I am the first author and the work presented in this chapter is my own work. My role consisted in designing the study, conceiving the experiments and analyses, conducting the literature searches, data extraction and cleaning, and writing the manuscript. We will present this chapter according to the TRIPOD guideline[88].

## 6.1 Motivation

A lot of research in the field of deep learning has been focusing on estimating and improving "point predictions" in form of personalised risk scores for a given medical event in one's future, as for instance reported in innovative deep learning models such as RETAIN[104], BEHRT, and Doctor AI[102]. However, point predictions in absence of uncertainty estimates is impossible to derive credibility quantification (e.g., p-value). Considering the significant consequences of decision-making in clinical practice that is guided by model predictions, quantifying the uncertainty of predictions is proving to be a key step in putting these models to practice in medicine.

## 6.2 Related work

In the last several years, a new subfield of deep learning, called probabilistic deep learning, has drawn wide interest in providing prediction and uncertainty estimations at the

same time. The most promising methods are Bayesian deep learning[71] (BDL) and sparse Gaussian processes[205] (SGP). In BDL, by placing a distribution over each of the model weights instead of treating them as point values, the uncertainty of the weights can be passed layer by layer to eventually estimate the uncertainty in the predictions. However, this approach usually requires a compromise between model complexity and expressiveness of variational distributions.

On the contrary, the Gaussian process (GP), as a non-parametric model, is largely more flexible and expressive than the assumption of variational distributions in BDL. This advantage comes, however, at the expense of the need to store and process the data points for the covariance matrix. This usually takes cubic time $\mathcal{O}(n^3)$ to calculate the inversion and determinant of the covariance matrix for inference and becomes a challenge when working with large-scale datasets.[206]

The state-of-the-art solution to this challenge is to use a small number of pseudo-points (i.e., inducing points) to approximate the data points. This enables the covariance matrix to be approximated by a lower-rank representation.[207] Since the entire dataset is summarised by a small number of inducing points, this method is called sparse GP, and its performance highly depends on the robustness of the inducing points and kernel parameters. Wilson et al.[208] upgraded this framework to be more flexible and scalable by implementing a deep architecture beneath the kernel function as a feature extractor, which is known as deep kernel learning (DKL).

## 6.3 Aims

In this study, using incident heart failure (HF), diabetes, and depression risk prediction as examples, we aimed to adopt the established probabilistic modelling methods to transform BEHRT from deterministic to probabilistic and comprehensively evaluate the role of uncertainty in risk prediction.

## 6.4 Methods

### 6.4.1 Data source

This study was conducted using EHR from CPRD linked to HES and ONS. Primary data was available from 1 January 1985 to 31 December 2015, and the linkage was available from 1 January 1998 to 30 September 2015.

### 6.4.2 Study design

We defined three risk prediction tasks that used all available information before the baseline to predict the incident HF, or diabetes, or depression risk within a six-month window after the baseline. The incident HF, diabetes, and depression were defined as the first recorded HF, diabetes, and depression diagnosis code in one's EHR (both primary care and hospital care, respectively. The baseline dates were created independently for each risk prediction task. More specifically, we randomly selected a baseline date within a six-month window before the first incidence of the condition and marked those patients as positive cases. This procedure can maximally preserve all positive patients. For patients who did not have the pre-defined condition, we randomly selected a baseline date within one's EHR history and marked them as negative cases. Note that, we made sure that negative cases had more than six months medical records after the baseline date to guarantee each of them was an absolute negative sample. Therefore, avoiding any assumption for the unseen future.

### 6.4.3 Eligible participants

In this study, we selected a cohort with both men and women, contributed to data between 1985-01-01 and 2015-09-30, were linked to HES, marked as "acceptable" by CPRD quality control, and registered with general practices for more than three years before the baseline date.

### 6.4.4 Study population

The population selection procedure led to 788,880 (8.3% positive samples), 913,799 (11.3%), and 1,453,012 (16.1%) patients for HF, diabetes, and depression cohorts, respectively. Afterwards, we randomly allocated 50%, 20%, and 30% of patients to a training, tuning, and validation dataset, respectively, for each cohort.

### 6.4.5 Outcomes

We adopted the phenotypings for HF, diabetes, and depression from the CALIBER code repository[105]. We considered HF as a composite condition of rheumatic HF, hypertensive heart disease with (congestive) HF and renal failure, ischemic cardiomyopathy, chronic cor pulmonale, congestive HF, cardiomyopathy, left ventricular failure, and cardiac, heart, or myocardial failure;[193] diabetes as both type I, type II, and other unspecified diabetes; depression as depressive episode, depressive disorder, and dysthymia. The code lists for the identification of HF, diabetes, and depression can be found in Table A.7, Table A.8, and Table A.16, respectively.

### 6.4.6 Data preparation

In this work, we incorporated the diagnoses and medications for risk prediction, and they were prepared the same way as introduced in §4.2. In brief, diagnoses were mapped to ICD-10 level four, medications were kept as the BNF code at the section level. Each record was annotated with age in year.

### 6.4.7 Model derivation

In this work, we applied the popular probabilistic modelling approaches from GPs and BDL to transform (deterministic) BEHRT into a probabilistic model for risk prediction with the provision of uncertainty estimation (we refer reader to §2.5.1 and §2.5.2 to refresh the fundamental concepts about probabilistic modelling). More specifically, we employed Bayesian deep learning with variational inference and DKL. DKL is a GP hybrid neural

network (sparse GPs on the top of deep neural networks (DNNs) and trained end-to-end) that considers a trainable neural network as a part of the GP kernel function. Additionally, we noticed that GP hybrid Bayesian deep neural networks, which considered a probabilistic Bayesian neural network as a part of the GP kernel function, have not been explored in the past. Thus, we included it as an additional comparator and named it GPBDNN.

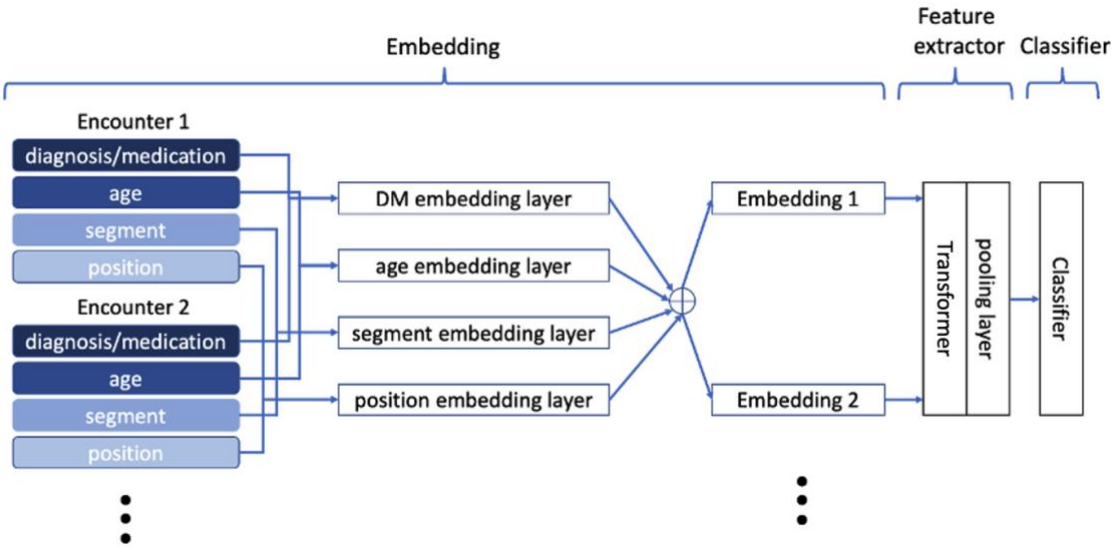



*Figure 6.1: BEHRT model architecture. The model includes an embedding layer to encode the input features, a feature extractor for representation learning, and a classifier for prediction. DM represents diagnosis and medication; the pooling layer extracts the latent representation of the first timestep for classification. An encounter is represented as the summation of the DM, age, segment, and position embeddings.*

Essentially, BEHRT has three components: embedding layer, feature extractor, and classifier (as shown in Figure 6.1). The probabilistic transformations were applied to these three components and all of the probabilistic BEHRT models share the same fundamental BEHRT architecture. The modifications that transfer BEHRT to a probabilistic model, as well as the definitions of different architectures are listed below:

**Bayesian embedding (BE)**: Transform the embedding layer in BEHRT from deterministic parameters to stochastic parameters while keeping all other architectures unchanged.

**Bayesian classifier (BC)**: Transform the classifier layer in BEHRT from deterministic parameters to stochastic parameters while keeping all other architectures unchanged.

**Bayesian embedding + Bayesian classifier (BE+BC)**: Transform the embedding layer and the classifier layer in BEHRT from deterministic parameters to stochastic parameters while keeping all other architectures unchanged.

**SGP**: Transform the classifier in BEHRT from deterministic parameters to Sparse GP (§2.5.4) while keeping all other architectures unchanged.

**KISS-GP**: Transform the classifier in BEHRT from deterministic parameters to KISS-GP (§2.5.5) while keeping all other architectures unchanged.

**Bayesian embedding + KISS-GP (GPBDNN)**: Transform the classifier in BEHRT from deterministic parameters to KISS-GP and the embedding layer from deterministic parameters to stochastic parameters while keeping all other architectures unchanged.

Dusenberry et al.[66] suggested that for BDL, models with Bayesian embedding and Bayesian classifier usually work better than fully Bayesian models. Therefore, we followed the same strategy and mainly investigated the BDL models with stochastic embedding and classifier.

This study implemented models in PyTorch and GPyTorch [209], and the optimisation objective for model training was evidence lower bound (ELBO) (§2.5.2). In terms of the model architecture, the fundamental BEHRT model used maximum sequence length 256, hidden size 150, 4 layers of Transformer, 6 attention heads, 108 intermediate hidden size, and 0.29 dropout rate. For BE, BC, BE+BC, 150 was used for pooling layer, while 24 was used for SGP, KISS-GP, and GPBDN. For stochastic weights, we used mean field distribution[5] as the variational posterior, where each weight is represented by a normal distribution with learnable mean and diagonal variance parameters, and the normal distributions with 0 mean and 0.374 standard deviation were used as priors. For GP classifiers, 40 inducing points and radial basis function kernel were used, and they were

implemented with a multivariate distribution with 0 mean and identity covariance matrix for the prior. These parameters were selected with Bayesian hyperparameter optimisation tuned over 30 iterations. Additionally, 64 batch size and Adam optimiser with learning rate $3e^{-5}$ and weight decay 0.01 were used for training, and Bayes by Backprop[210] with Blundell KL weight penalty[211] were used for training BDL models, and we also conducted oversampling for the positive cases with ratio 1:4 (positive: negative) within each training batch.

### 6.4.8 Model validation

We evaluated model performance on the validation dataset. For each probabilistic model, we applied Monte Carlo sampling[5] to sample a patient's prediction for 30 times as an estimation of the predictive distribution. The predictive distribution was assumed as a normal distribution because of the central limit theorem[212]. Additionally, we conducted model evaluation from three perspectives: generalisability, ability of rejecting overconfident predictions, and uncertainty estimation.

For generalisability, we evaluated model performance for AUROC and APS; both were calculated using the mean predictive probability, which is a probability averaged over the samples sampled from the predictive distribution. As for the ability of rejecting overconfident predictions, we evaluated the accuracy of predictions as a function of the mean predictive probability. In medicine, it is highly desirable to avoid overconfident and incorrect predictions. Therefore, it is more beneficial to evaluate the model performance for predictions above a user-specified threshold. One tends to trust the predictions more when the confidence is high, and resorts to a different solution when the prediction is not confident. Thus, the better rejection ability can be directly reflected by having a higher performance for high-confident predictions.

In terms of uncertainty estimation, quantifying the quality of uncertainty estimation is still an open question. One potential way to properly assess the correctness of the posterior

is to compare with the ground truth established by, for example, Hamiltonian Monte Carlo[213]. However, this method scales poorly on high dimensional space, thus, it is only feasible to a relatively simple model. Another popular approach for quality quantification is log-likelihood [214,215]. However, the previous evidence[213] also illustrated that this metric evaluated the posterior mean rather than the uncertainty (variance). Therefore, in this work, we proposed two qualitative methods for uncertainty evaluation, and they represented the most fundamental demands of uncertainty in health applications.

First, we propose to compare the uncertainty (variance) between the correct and incorrect predictions (i.e., true positive versus false positive, and true negative versus false negative). We would intuitively expect that the variance of correct predictions is distinguishably lower than the one of incorrect predictions. If the variances for both correct and incorrect predictions are similar, we would say the model cannot provide a meaningful uncertainty estimation. Secondly, for an imbalance risk prediction task, we would also intuitively expect the model to capture such information and indicate a higher uncertainty for the predictions of minority class.

### 6.4.9 Applications

In addition to model evaluation, we explored two potential uncertainty applications: uncertainty for hypothesis testing, and uncertainty to inform risk associations.

#### 6.4.9.1 Uncertainty for hypothesis testing

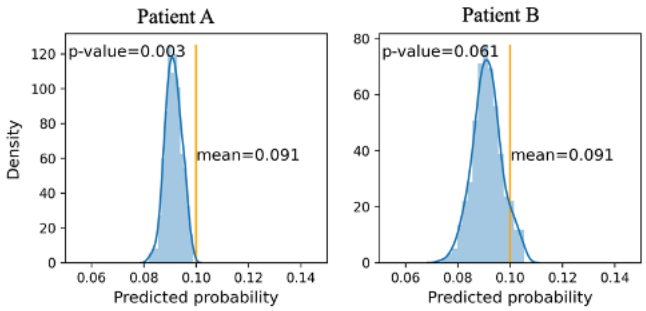



*Figure 6.2: Predictive distribution of two hypothetical patients. Both patients have predicted risk 9.1%. Patient A can confidently reject the 10% threshold at 5% significance level, while patient B cannot.*

Risk models and risk thresholds are effective tools to inform clinical decision-making. For instance, the NICE guideline[9] recommends treating patients with lipid-lowering drugs for patients with more than 10% CVD risk (estimated by QRISK2). However, a patient with a 9.99% CVD risk also seems reasonable to be treated as these two values are approximately the same. Without uncertainty quantification, it is hard to know when a patient should be rejected for treatment. For instance, Figure 6.2 shows two hypothetical patients with a mean predicted risk of 9.1%. Even though they have the same mean predicted risk, it is obvious that patient A can significantly reject the 10% threshold while patient B fails to reject the 10% threshold at 5% significance level. Thus, it is reasonable to include patient B for treatment even if one's predicted risk is less than 10%. In this study, we followed a similar strategy to investigate how the inclusion of patients who fail to reject the decision thresholds for treatment can affect the miss rate.

#### 6.4.9.2 Uncertainty analysis in embeddings

As shown in §4.1, embeddings that represent events (diagnoses or medications) usually contain important information regarding event associations. In this work, we wish to further explore that linkage between embedding uncertainty and risk associations. The uncertainty of an embedding was represented by summarising the entropy across all embedding dimensions. Entropy[216] is a commonly used metric to quantify the uncertainty of a probabilistic distribution. For diagnoses or medications with high uncertainty in the embeddings, we assume they would contribute more uncertainty to the latent representations as well as the predictions. Therefore, high uncertainty in the embedding indicates an unclear association between a disease or a medication and the target disease. Otherwise, the association would be more certain even though the magnitude of the association is strong or weak.

## 6.5 Results

### 6.5.1 Data descriptive analysis

Table 6.1 shows the baseline statistics for HF, diabetes, and depression cohorts. In general, HF dataset is more imbalanced compared to the diabetes and depression datasets; and diabetes dataset has relatively longer records than the HF and depression datasets.

*Table 6.1: Statistics of EHR datasets for HF, diabetes, and depression cohorts. Encounter represents medical records, and each visit may have multiple encounters.*

| | HF cohort | Diabetes cohort | Depression cohort |
|---|---|---|---|
| Total number of patients | 788,880 | 913,799 | 1,453,012 |
| Number of incident cases (%) | 65,477 (8.3) | 103,585 (11.3) | 233,726 (16.1) |
| Mean (SD) age (year) | 58.4 (18) | 60.3 (18.0) | 58.0 (18.3) |
| Median number of visits per patient | 55 | 71 | 45 |
| Mean number of encounters per visit | 2.37 | 2.33 | 2.29 |
| Median number of encounters per patient | 87 | 134 | 79 |

### 6.5.2 Evaluation for generalisability

Table 6.2 shows that all implemented probabilistic BEHRT models have a comparable performance in terms of AUROC and APS. HF has better performance than diabetes, and depression has the worst performance among these three prediction tasks. Expectedly, all models share the fundamental BEHRT architecture, and we would expect them to have a similar predictive performance. The result further reassured that different probabilistic modelling methods do not substantially affect the overall model generalisability.

*Table 6.2: Metrics for marginalised predictions on HF, diabetes, and depression. 95% confidence intervals are computed via a validation set bootstrapping with 50 bootstrap sets.*

| | HF | | Diabetes | | Depression | |
|---|---|---|---|---|---|---|
| Model | AUROC | APS | AUROC | APS | AUROC | APS |
| GPDBNN | $0.941 \pm 1e^{-3}$ | $0.625 \pm 5e^{-3}$ | $0.834 \pm 2e^{-3}$ | $0.533 \pm 4e^{-3}$ | $0.776 \pm 2e^{-3}$ | $0.416 \pm 3e^{-3}$ |
| SGP | $0.945 \pm 1e^{-3}$ | $0.645 \pm 1e^{-3}$ | $0.834 \pm 1e^{-3}$ | $0.538 \pm 3e^{-3}$ | $0.782 \pm 1e^{-3}$ | $0.433 \pm 2e^{-3}$ |
| KISS-GP | $\mathbf{0.945 \pm 1e^{-3}}$ | $\mathbf{0.649 \pm 5e^{-3}}$ | $\mathbf{0.837 \pm 2e^{-3}}$ | $\mathbf{0.538 \pm 4e^{-3}}$ | $\mathbf{0.782 \pm 1e^{-3}}$ | $\mathbf{0.433 \pm 2e^{-3}}$ |
| BE | $0.942 \pm 1e^{-3}$ | $0.631 \pm 7e^{-3}$ | $0.830 \pm 1e^{-3}$ | $0.529 \pm 5e^{-3}$ | $0.774 \pm 1e^{-3}$ | $0.409 \pm 2e^{-3}$ |
| BC | $0.933 \pm 1e^{-3}$ | $0.645 \pm 6e^{-3}$ | $0.825 \pm 1e^{-3}$ | $0.525 \pm 5e^{-3}$ | $0.765 \pm 1e^{-3}$ | $0.425 \pm 2e^{-3}$ |
| BE+BC | $0.941 \pm 1e^{-3}$ | $0.628 \pm 5e^{-3}$ | $0.835 \pm 2e^{-3}$ | $0.538 \pm 2e^{-3}$ | $0.778 \pm 1e^{-3}$ | $0.419 \pm 2e^{-3}$ |

### 6.5.3 Accuracy as a function of confidence

In this section, we re-used the results from the previous section to evaluate the ability of rejecting overconfident and incorrect predictions based on the mean predictive probability. Figure 6.3 shows the model accuracy for predictions with confidence above different thresholds. Considering that people are usually interested in predictions with confidence higher than a given threshold, the accuracy curve illustrated the model performance in a more practical way. In general, the results showed that GP-based methods (SGP, KISS-GP, and GPBDNN) could produce more robust high confident risk predictions as they have accuracy consistently higher than the BDL-based methods (BE, BC, BE+BC) across different thresholds. Although GPBDNN had relatively low accuracy for HF risk prediction compared to other methods for low confidence thresholds. It still outperformed others when the confidence threshold increased. This indicated that most of the misclassifications for GPBDNN were in the low confidence area, and its high confident predictions were very reliable, which is a desired feature in practice.

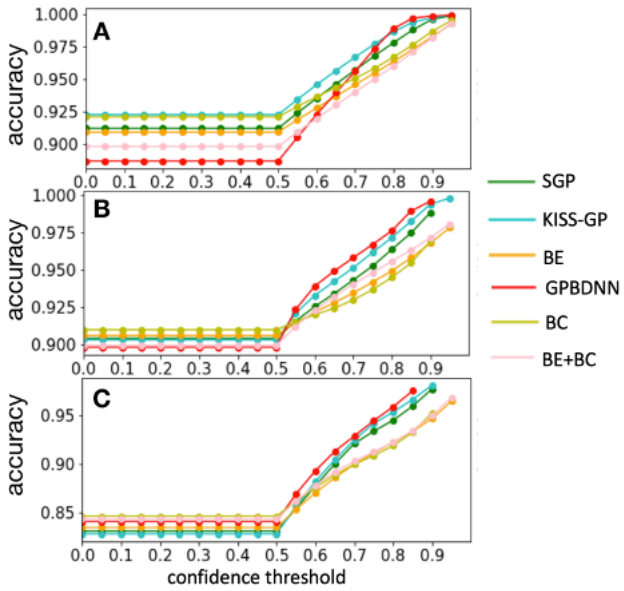



*Figure 6.3: Accuracy as a function of confidence. A, B, and C represent results for HF, diabetes, and depression risk prediction tasks, respectively. Figure shows the accuracy for predictions with confidence higher than a given threshold. For instance, confidence ≥ 0.6 means mean predictive probability ≥ 0.6 for positive predictions and mean predictive probability ≤ 0.4 for negative predictions.*

Additionally, instead of giving overconfident predictions, GP-based methods (especially GPBDNN) also showed a better capability of penalising (avoid making) highly confident predictions than the BDL-based methods, and such effects became clearer when the model prediction performance (AUROC and APS) dropped. For instance, depression risk prediction models had the worst performance. Thus, GPBDNN rarely produced risk prediction with confidence higher than 0.8; SGP and KISS GP rarely produced risk prediction higher than 0.9 (Figure 6.3C). A potential reason is that GP is a non-parametric model that uses a kernel to measure the similarity between the observations and the new inputs. The unsatisfactory model performance means either the similarity is badly measured, thus leading to false prediction, or the new inputs are not similar to the observations, and the predictions are quite uncertain, thus leading to incorrect predictions. The analysis suggests the latter, proving that GP-based methods can better reject overconfident predictions.

### 6.5.4 Uncertainty evaluation

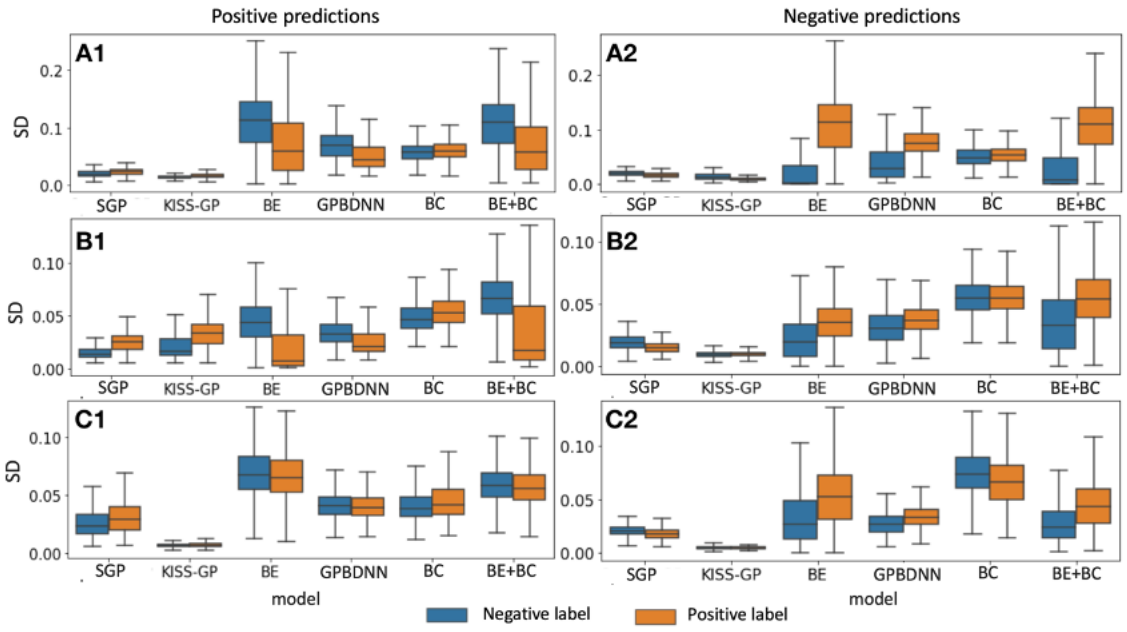



*Figure 6.4: Distribution of standard deviation (SD) for positive and negative predictions. A, B, and C represent the HF, diabetes, and depression prediction tasks, respectively. Figures on the left (1) and right (2) represent samples with positive prediction, and negative prediction, respectively. BDL-based methods and GPBDNN show the false positive and negative predictions have a higher uncertainty than true positive and negative predictions.*

To evaluate the difference in uncertainty between correct (true positive and true negative) and incorrect (false positive and false negative) predictions, we represented each patient with a mean predictive probability and a corresponding SD calculated from samples sampled from predictive distribution. Furthermore, we considered patients with a mean predictive probability higher and lower than 0.5 to be predicted as positives and negatives, respectively. Figure 6.4 shows that Bayesian embedding-based methods (BE, BE+BC, and GPBDNN) can better capture uncertainty that can distinguish the correct and incorrect predictions. For instance, Figure 6.4 (left) shows these methods produce significantly lower SD for the true positive predictions than the false positive predictions. On the contrary, GP-based methods (SGP and KISS-GP), in general, either provided an indistinguishable uncertainty estimation for correct and incorrect predictions, or provided uncertainty estimations that seem wrong (i.e., lower uncertainty for incorrect predictions than the correct ones).

Additionally, we employed the calibration curve with the provision of uncertainty to analyse the models' capability of indicating data insufficiency. By sampling model weights from the posterior distribution, we could consider each set of weights as an independent model. Therefore, the uncertainty of a calibration curve reflects the uncertainty of a model toward the prediction based on the data it is trained on. The uncertainty is caused by the uncertainty of the model parameters, thus, indicating the lack of knowledge (e.g., insufficiency of samples).

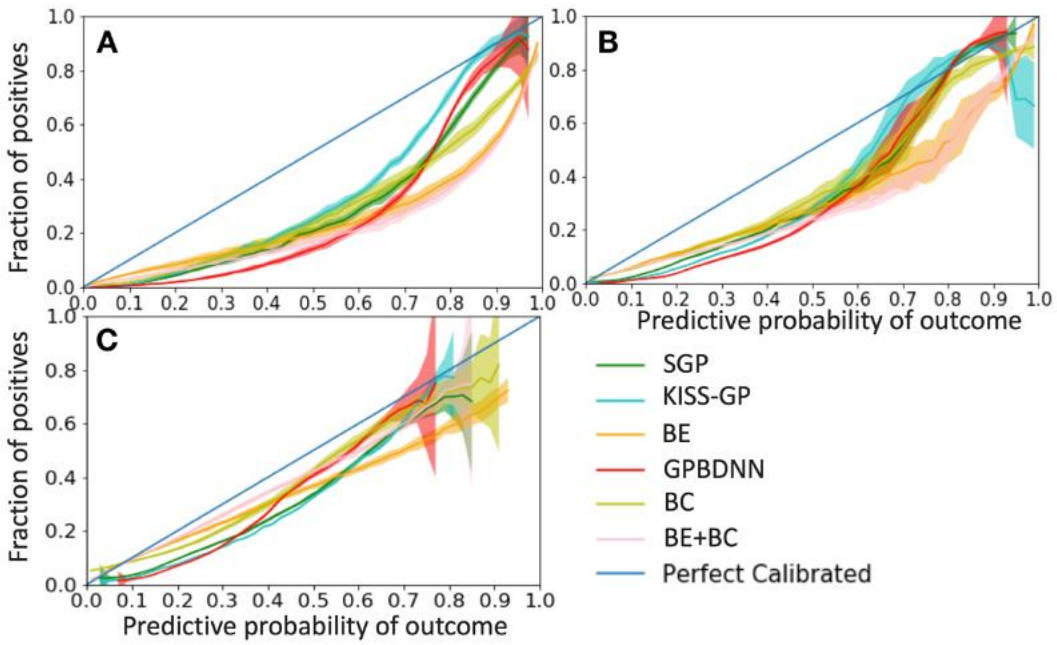



*Figure 6.5: Calibration curves with 95% confidence intervals. x axis and y axis represent the predictive probability and the fraction of positive cases. Respectively. A, B, and C represent HF, diabetes, and depression task, respectively. The oversampling primarily causes the miscalibration for HF and diabetes risk prediction and can be alleviated with post-hoc recalibration.*

Figure 6.5 shows that GP-based methods, in general, are less confident in making positive predictions compared to the negative predictions, indicating the data insufficiency for the positive cases. In contrast, even though the prediction performance in terms of APS (around 0.6 for HF prediction) is relatively poor, the BDL-based methods are still very certain (low confidence intervals) for all positive predictions; and the uncertainty only starts to show when the model performance is even worse (as shown in Figure 6.5C). The potential reasons can be that the GP classifiers make inference based on the observations (i.e., inducing points in our case) and the imbalance of the data is naturally preserved by the non-parametric properties. Therefore, they are more sensitive to the data imbalance and are better at representing the incomplete coverage of the domain knowledge from the training samples than the stochastic BDL modules with a mean-field variational distribution. Additionally, the miscalibration for HF and diabetes risk prediction was potentially caused by the oversampling strategy. It changed the prior distribution in the training dataset, thus, leading to the mismatch between the prior distribution in the training and the validation dataset.

### 6.5.5 Uncertainty for hypothesis testing

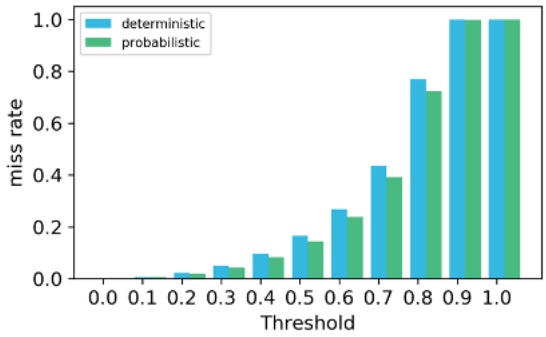



*Figure 6.6: Miss rate (false negative rate) comparison using GPBDNN. Deterministic analysis represents a benchmark that the miss rate at different threshold is calculated using mean predictive probability; mean predictive probability lower and higher than a threshold is considered as negative and positive prediction, respectively. The calculation for probabilistic analysis treats a prediction as negative only if its mean predictive probability is lower than a threshold and can reject the threshold at 5% significance level. Otherwise, we consider it as positive prediction.*

Figure 6.6 uses GPBDNN for HF risk prediction as an example to analyse the benefit of including uncertainty in decision-making. The results show that considering uncertainty and hypothesis testing as a part of the decision-making process can greatly reduce the miss rate compared to using mean predictive probability only.

### 6.5.6 Uncertainty in embeddings

Table 6.3 shows that 28 out of 30 of the low entropy (uncertainty) diagnoses and medications are closely associated to the outcome. The results were validated by two clinicians and evidence from previous research can also support the linkage. For instance, Coughlin's work[217] and Fernandez et al.'s[218] work indicated a significant association between influenza virus, chronic low back pain, and anxiety and depression, respectively. However, there is no clear evidence for the association between torticollis, cellulitis, and HF. Our analysis showed that most of the diagnoses and medications align with the prior knowledge and suggested a link between uncertainty and risk factors. Therefore, it has a potential to assist hypothesis generation.

*Table 6.3: Top 10 events with the lowest entropy for HF, diabetes, and depression prediction.*

| **Entropy** | **HF** |
|---|---|
| − 236.14 | Diuretics |
| − 235.67 | Acute myocardial infarction |
| − 235.43 | Unspecified acute lower respiratory infection |
| − 235.17 | Other ill-defined heart diseases |
| − 235.12 | Antiplatelet Drugs |
| − 235.08 | Chronic obstructive pulmonary disease with (acute) exacerbation |
| − 234.97 | Chronic obstructive pulmonary disease, unspecified |
| − 234.95 | Torticollis |
| − 234.90 | Cellulitis and acute lymphangitis of other parts of limb |
| − 234.68 | Type 2 diabetes mellitus with circulatory complications |
| **Entropy** | **Diabetes** |
| − 230.52 | Drugs used in diabetes |
| − 230.28 | Detection strips, urine for glycosuria |
| − 229.18 | Unspecified diabetes mellitus |
| − 229.16 | Hypoglycemia, unspecified |
| − 229.00 | Obesity, unspecified |
| − 228.42 | Essential (primary) hypertension |
| − 227.24 | Positive inotropic drugs |
| − 226.98 | Lipid-regulating drugs |
| − 226.96 | Chronic tubulo-interstitial nephritis, unspecified |
| − 226.77 | Sex hormones |
| **Entropy** | **Depression** |
| − 229.68 | Antidepressant drugs |
| − 221.35 | Anxiety disorder, unspecified |
| − 220.69 | Influenza due to unidentified influenza virus with other respiratory manifestations |
| − 219.65 | Drugs used in psychoses and related disorders |
| − 219.58 | Analgesics |
| − 219.36 | Hypnotics and anxiolytics |
| − 219.24 | Laxatives |
| − 219.15 | low back pain |
| − 219.13 | General anaesthesia/hypnotics and anxiolytics |
| − 219.09 | Dyspepsia and gastro-oesophageal reflux disease |

## 6.6 Discussion

In this study, we evaluated a range of methods that transformed BEHRT from deterministic to probabilistic for risk prediction with the provision of uncertainty estimation. As an additional comparator, we proposed a hybrid architecture, named GPBDNN, which

has not been considered in the past. It combines the strengths of both GPs and BDL, showing a comparable generalisation performance with other GP and BDL-based methods in terms of AUROC, APS, and accuracy, but with substantially better uncertainty estimation.

In our experiments, with APS only 0.62 to 0.65 for HF, and 0.52 to 0.54 for diabetes, only GP-based methods showed a clear pattern of having high model uncertainty for positive predictions (minority class) in highly imbalanced datasets; BDL-based methods were not sensitive to this information until the APS became very poor (0.40 to 0.43 for depression prediction). On the contrary, BDL-based methods were better at capturing the general trend for predictive uncertainty estimation. For instance, those methods can correctly show the correct (true positive and true negative) predictions were more confident than the incorrect (false positive and false negative) predictions. But the patterns from GP-based methods did not match human intuition and expectation, having more confident predictions for the wrong prediction than the correct predictions.

By combining both methods, GPBDNN showed correct patterns in both ways, providing a better capability for uncertainty estimation. This can be used to indicate data insufficiency and guide further model improvement. It also can identify potential false positive predictions and resort those patients to seek medical check-ups to carry out more careful conclusions in practice rather than being ignored because of low-risk predictions from the risk model. Furthermore, with uncertainty estimation, we can also conduct hypothesis testing to quantify the confidence of treating or not treating a patient. This can be critical statistical information to guide clinicians in decision-making. Also, we investigated the associations between uncertainty and risk factors, and the results showed strong evidence for the relation. Therefore, interpretability and causality analysis are interesting topics for future work.

Our work also has a limitation. We evaluated various probabilistic modelling approaches using qualitative methods rather than conducting a comprehensive quantitative comparison. This is mainly because quantitative metrics for uncertainty evaluation are still an open question, and CPRD lacks ground truth to numerically evaluate how well the posterior is estimated. Therefore, validating probabilistic modelling approaches on other clinical datasets with different populations and an evaluation of the posterior estimation of GPBDNN on simpler questions with ground truth can be beneficial in future works.

## 6.7 Conclusions and potential future works

In this chapter, we drew attention to the fact that uncertainty can provide critical information to assist decision-making and improve the quality of care. For example, the ability to tell right from wrong, quantify the confidence of whether treating a patient or not, and avoid making overconfident predictions. These are all desired properties for a risk model, yet still lacking for the current systems. To this end, we introduced a hybrid architecture GPBDNN. We showed that combining GPs and BDL can produce better uncertainty estimation than each individually and equip a risk model with the valuable properties abovementioned. Important directions for future work include: (1) conducting a more comprehensive and quantitative evaluation on cohorts from other populations; (2) investigating the utility of uncertainty in informing temporal distribution shifts, which is an inherent challenge in risk prediction; and (3) the effectiveness of uncertainty estimation in survival framework.

# 7. General conclusion and perspectives

In this thesis, we investigated a crucial clinical application – prognostication - with an emphasis on three topics: risk prediction, explainability, and uncertainty.

To determine an appropriate dataset for the research questions, Chapter 3 first outlined foundational works that provided a comprehensive assessment of the strengths and limitations of EHRs compared with traditional epidemiological studies. Specifically, the CPRD was assessed in detail regarding representativeness, validity, and comprehensiveness. This led to the conclusion that the CPRD dataset provides large-scale longitudinal EHR with comprehensive records collected from multiple sources and is representative of the UK population. Therefore, it is appropriate for the purpose of this study.

Furthermore, Chapter 4 narrowed the scope and focused on the investigation of risk prediction. It introduced a deep learning risk model, BEHRT, and its variations that can handle long EHR sequences (Hi-BEHRT) and censoring (BEHRTSurv), showing its capability of learning meaningful representations from minimally processed EHR for accurate risk prediction. Compared to established clinical and machine learning models that relied on expert-selected predictors, we demonstrated the superior performance of data-driven deep learning models on a range of risk prediction tasks both internally and on populations with the presence of data shifts. Our findings also highlighted the necessity of recalibration when applying a risk model to a population with severe prior distribution shifts and the importance of regular model updating to preserve the model's discrimination performance under temporal data shifts.

In terms of clinical benefit, the direct impact of deploying an accurate risk model in practice is the ability to better discriminate between positive and negative patients. This is presented as reducing false positive (or overtreatment for low-risk patients) and false

negative predictions (or undertreatment for high-risk patients). By extension, an accurate deep learning risk model can also offer a great opportunity to re-evaluate the conservative "treatment for all" paradigm due to the unsatisfactory risk prediction performance of established models for populations with unknown risk factors or risk factors that were poorly understood. Such improvement can considerably enhance care planning and disease prevention.

As for future work, the generalisability of the proposed BEHRT model can be future improved. One interesting direction could be federated learning, which allows a model to be trained anonymously across different datasets with data privacy secured. With access to more heterogeneous populations, we are more likely to carry out a highly generalisable risk model. Another way to improve model generalisability could be developing models that encode the data more effectively or causally because the population and the data recording system can change across regions, countries, or even at different times. It is highly desired to develop methods that can reduce the variability of information encoding and find the anchor (or key) information for modelling. This could involve but is not limited to embedding knowledge graphs for modelling.

In addition to outstanding risk prediction performance, Chapter 5 investigated another important property of the risk model: explainability. We demonstrated that BEHRT cannot only identify risk and preventative factors aligned with the established evidence without any expert's engagement, but also those that have not been considered in previous expert-driven studies. Moreover, BEHRT can capture the interplay between risk and treated risk, and the differential association of medications across different years, which would be difficult if the temporal context was not included in modelling. Although heart failure (HF) was investigated as an exemplary case in our study, the proposed framework paved the way for

further development of generating hypotheses for other conditions using deep learning models to assist medical research.

Besides the explanations in terms of association, we also investigated a framework that can achieve accurate risk prediction, while enabling counterfactual reasoning under hypothetical intervention. It allowed a user to interact with the risk model to explore the differential effects of an outcome resulting from an exogenous intervention, thus, informing clinicians to make decisions that benefit the patients the most. The framework introduced in the thesis, while providing encouraging results, can still benefit from further validation using new datasets and more established clinical questions. Additionally, this framework can not only answer our question regarding one exposure, but also can be as flexible as necessary to solve more complex and multi-dimensional questions of interest. Therefore, an interesting direction for future work would be applying the proposed framework to investigate meaningful clinical questions that were difficult to be conducted in clinical trials, such as disease phenotyping, treatment responsiveness, and multimorbidity patterns.

Lastly, Chapter 6 investigated uncertainty estimation methods, specifically Bayesian deep learning and Gaussian processes, that can transform a risk model from deterministic to probabilistic. We showed that a hybrid architecture that mixed Gaussian processes and Bayesian deep learning could achieve better uncertainty estimation than them alone. Its benefits include but are not limited to avoiding making overconfident predictions, indicating data insufficiency, and distinguishing the correct predictions from the wrong predictions. Quantitative evaluation of the reliability of uncertainty estimation remains an open question. However, future work to further evaluate the proposed framework using questions with the ground truth and other domain questions such as natural language and image processing can be highly beneficial.

# Appendix

## Supplementary Methods

### Methods A.1: Hi-BEHRT with different sliding window sizes and stride sizes

For the Hi-BEHRT model, we used a sliding window to segment the raw EHR into segments. As shown in Table A.11 when the window size is relatively small (i.e., 50), the stride size does not significantly impact predictive performance, and the bigger stride size can potentially decrease the number of segments and reduce model complexity. However, for the larger window size (i.e., 100), the stride size becomes more important, and some overlap between segments is necessary. Without any overlap for window size 100, the APS decreases by 4% compared to the model with stride size 50. Additionally, the analysis shows that larger window size is not always the better choice. For instance, APS of window size 100 without overlap decreases by 2% compared to APS of window size 50 without overlap. Without overlap, a larger window can lead to a shorter length at the segment level, and a balance between window size and segment length might be more preferred in the hierarchical structure.

### Methods A.2: Training model with stochastic approximation

Partial log-likelihood requires to access the entire dataset to identify the order of the positive events for loss calculation. Therefore, it may fail when dataset is too large or computational resources are limited. Recent studies[159,160] proposed an alternative approach to use a randomly selected subset of population as an approximation of the overall population for loss calculation rather than using the entire dataset. The authors mathematically proved that both methods can converge similarly. The implementation was achieved by randomly shuffling the dataset every epoch and training the model using

stochastic gradient descent. In this work, we used DeepSurv to investigate how different sample (batch) size for population approximation can affect the model performance. The experiment was conducted on the general population and the performance in terms of calibration and discrimination are shown in Figure A.2 and Table A.12, respectively. The results show that models trained with small sample size can achieve similar calibration and discrimination performance as models trained with large sample size or even the entire population.

**Methods A.3: Details for model training in survival analysis**

We randomly searched for 30 sets of hyperparameters for DeepSurv and reported the one with the highest discrimination performance (c-statistic). The same parameters were applied to MLP (as shown in Table A.13). MLP, DeepSurv, and BEHRT were trained using the Adam optimiser[107]. MLP and DeepSurv used a fixed learning rate for training, while the BEHRTSurv models were trained with a Cosine annealing learning rate scheduler [219]. We conducted a learning rate sweeping on a range of values [0.01, 0.001, 0.0001, 0.00001] for MLP and DeepSurv and reported the one with the best performance. For BEHRTSurv, the searching range was [0.00005, 0.0001, 0.0005]. All models were trained for 100 epochs and enabled early stopping if the validation loss did not decrease for 10 epochs. Additionally, we applied transfer learning as a comparator for risk prediction on the diabetes cohort. More specifically, we initialised the BEHRTSurv with weights trained on the general cohort. Afterwards, we only finetuned the last linear layer of log-risk function of BEHRTSurv and froze the weights of the rest layers. Because the prior distribution of the diabetes cohort was dramatically different from the general cohort, Platt scaling[220] was applied for post-hoc model recalibration.

***Methods A.4: Details of perturbation explainer***

The perturbation-based explainer was derived from Guan et al.'s work[173]. It aims to approximate the amount of information discarded by the hidden states in the intermediate layers of a deep learning model using perturbation. The equations below show the full description of the loss function conducted over one patient's EHR history.

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

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

where $\tilde{x}$, $x$, and $y$ represent the perturbed predictor, the original predictor, and the HF label, respectively. $S$ represents the original output state, which equals to $M(x)$, $M(\tilde{x})$ is the output state for perturbed input, and $M$ represents the model for the explanation. $\beta_1$ and $\beta_2$ are weights, and $\beta_1 < \beta_2$; if $\beta_1 = \beta_2$, the loss function is symmetric. $E_{\epsilon}\|(M(\tilde{x}) - s)\|^2$ is described as the mean square loss between the perturbed output and the original output, and $\sum_i^n H(\tilde{x}|s)|\epsilon_i \sim \mathcal{N}(0, {\sigma_i}^2 \boldsymbol{I})$ is the entropy loss introduced by Guan et al.[173].

For each patient, the perturbation method carries out $\epsilon = [\epsilon_1, \dots, \epsilon_n]$, where $n$ is the number of predictors in a patient's EHR. It is a list of variances with a range between 0 and 0.5 that each element indicates the contribution of the corresponding predictor. The variances reflect an inverse relationship: the higher the variance, the lower the contribution to the (positive or negative) prediction. Because a human would expect a higher value to represent higher contribution, and a lower value to represent lower contribution, we further conducted a transformation $0.5 - \epsilon_i$ to reverse the relationship.

***Methods A.5: Pilot analysis for the explainer consistency***

In this section, we conducted a pilot analysis to investigate the consistency of various model explainers and the interpretable model (i.e., linear regression), and how the

aggregation strategy for the local surrogate models can impact the importance of attributes at the population level.

More specifically, we applied several established explainers, LIME[50], SHAP[54], and DeepLIFT[174], to provide contribution analysis for CPH and DeepSurv models from Chapter 4.4 (survival analysis). Both models relied on expert-selected predictors, and we considered the coefficients from the CPH (linear) model as the benchmark. Additionally, for all explainers, we simply aggregated the importance of attributes at the individual level across the population and used the average importance to represent the importance of an attribute.

Figure A.3 shows that all explainers have good consistency in terms of feature importance. It means all explainers agree on the features that are important or unimportant. However, the explainers are not always consistent with the linear coefficients (from the CPH model). For instance, diabetes is considered an important feature for CVD risk prediction, and the CPH coefficient that represents diabetes also indicates the high importance of diabetes. However, almost all explainers consider that diabetes has little contribution to CVD risk prediction. In contrast, if we only aggregate the feature importance across the diabetes population, all explainers have a correct interpretation that diabetes is an important feature for risk prediction. Therefore, the analysis highlighted that simply aggregating the feature importance from local surrogate models across the population can fail to reveal the true feature importance, as the feature importance from different patients can cancel out each other during aggregation.

## Supplementary Figures

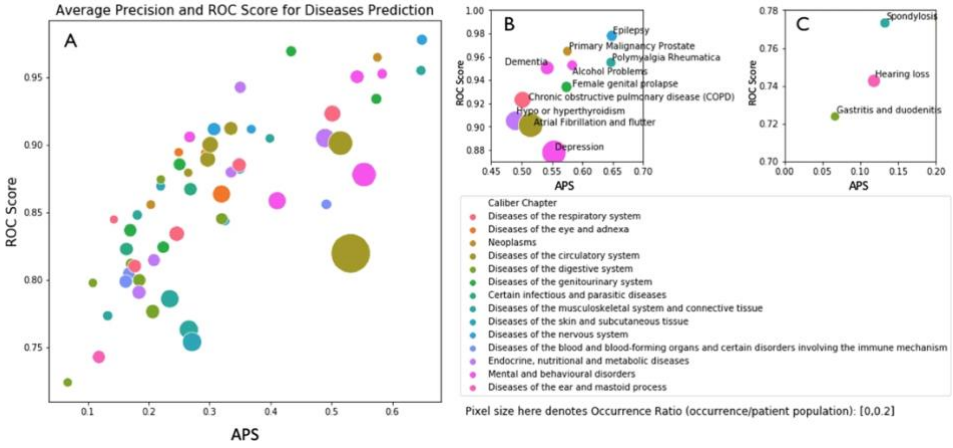



*Figure A.1: Disease-wise precision analysis. Each circle in these graphs represents a disease, and its colour and size denote the Caliber chapter and prevalence, respectively. Also, in these plots, we show APS and AUROC on the x- and y-axis, respectively. Therefore, the further right and higher a disease, the better BEHRT's job at predicting its occurrence in the next 6 months. Subplot (A) illustrated the full results, and subplots (B and C) illustrate the best and worst sections of the plot, in terms of BEHRT's performance.*

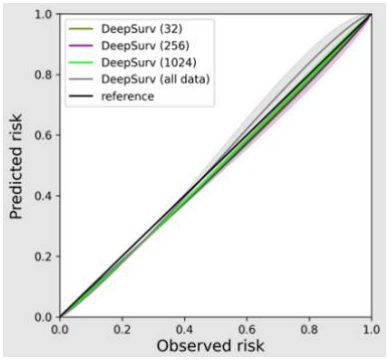



*Figure A.2: Calibration curve of DeepSurv trained with different batch size. All models are converged similarly in terms of calibration.*

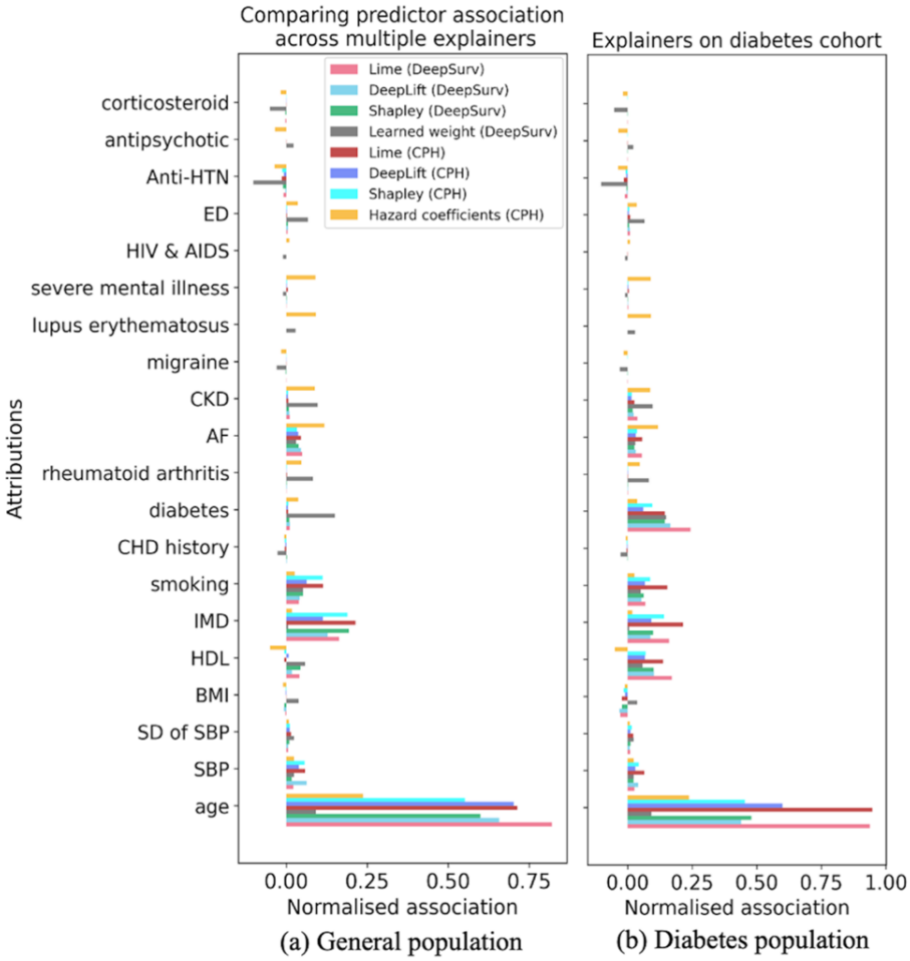



*Figure A.3: Population level feature importance identified by CPH coefficients and various model explainers. CPH coefficients is the same for both figure a and figure b. Figure (a) shows feature importance of explainers aggregated across the entire population and figure (b) shows feature importance of explainers aggregated across patients with diabetes. Figure (b) indicates explainers have good consistency in terms of feature importance. However, the explainers have different interpretation compared to the linear coefficients for features such as diabetes. Figure (b) further shows that if we only aggregate feature importance from explainers across the diabetes patients, both linear coefficients and explainers agree that diabetes is an important feature for CVD risk prediction.*

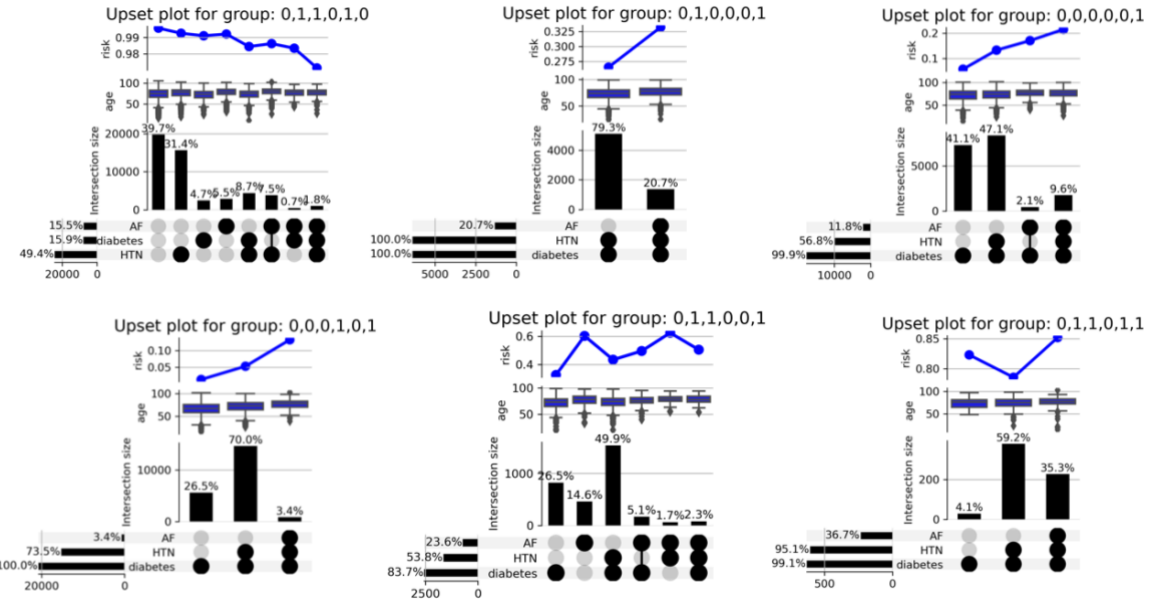



*Figure A.4: More examples of UpSet plot for patients within different latent groups.*

## Supplementary Tables

*Table A.1: BEHRT hyper-parameter searching on the MLM task.*

| 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 | 0.6380 |
| 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 | 6 | 6 | 1024 | 0.6508 |
| 15 | 576 | 6 | 6 | 256 | 0.6503 |
| 16 | 576 | 6 | 12 | 256 | 0.6510 |
| 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 |

*Table A.2: Hyper-parameter searching for Deepr for subsequent visit prediction.*

| Iteration | Filter | Kernel Size | FF I | FF II | FF III | Dropout I | Dropout II | Dropout III | Learning rate | APS |
|---|---|---|---|---|---|---|---|---|---|---|
| 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 |
| FF represents the feed-forward layer | | | | | | | | | | |

*Table A.3: Hyper-parameter searching for RETAIN for subsequent visit prediction.*

| Iteration | Embedding Size | Recurrent Size | Dropout Embedding | Dropout Context | L2 | APS |
|---|---|---|---|---|---|---|
| 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 |

*Table A.4: Codes for the identification of Left ventricular hypertrophy*

| Code | Type | Source | Description |
|---|---|---|---|
| 13857 | Medcode | CPRD (Test) | ECG: shows LVH |
| 23142 | Medcode | CPRD (Test) | ECG: LVH NOS |
| 6319 | Medcoe | CPRD (Test) | ECG: left ventricular hypertrophy |
| 526 | Medcode | CPRD (Clinical) | Left ventricular hypertrophy |

*Table A.5: Parameters of Framingham, QIRSK, ASSIGN-based random forest models*

| Parameters | Framingham | QRISK | ASSIGN |
|---|---|---|---|
| n_estimator | 800 | 800 | 800 |
| criterion | Gini | Gini | Gini |
| max_depth | 5 | 10 | 8 |
| max_features | auto | auto | auto |
| Random search space:<br>n_estimator: [200, 400, 600, 800, 1000]<br>max_depth: [3 – maximum number of predictors] | | | |

*Table A.6: Parameters for BEHRT*

| Max sequence length | 512 | Hidden size | 256 |
|---|---|---|---|
| No. of layers | 6 | Hidden activation | GELU |
| Intermediate size | 512 | No. of attention heads | 8 |
| Dropout probability | 0.2 | | |

*Table A.7: ICD-10 codes for the identification of HF*

| ICD Code | Description |
|---|---|
| I09.9 | Rheumatic heart failure |

| | |
|---|---|
| I11.0 | Hypertensive heart disease with (congestive) heart failure |
| I13.0 | Hypertensive heart and renal disease with (congestive) heart failure |
| I13.2 | Hypertensive heart and renal disease with both (congestive) heart failure and renal failure |
| I25.5 | Ischemic cardiomyopathy |
| I27.9 | Chronic cor pulmonale |
| I38 | Congestive heart failure due to valvular disease |
| I42.0 | Congestive cardiomyopathy |
| I42.1 | Obstructive hypertrophic cardiomyopathy |
| I42.2 | Nonobstructive hypertrophic cardiomyopathy |
| I42.6 | Alcoholic cardiomyopathy |
| I42.8 | Other cardiomyopathies |
| I42.9 | Cardiomyopathy NOS |
| I50.0 | Congestive heart failure |
| I50.1 | Left ventricular failure |
| I50.2 | Systolic (congestive) heart failure |
| I50.3 | Diastolic (congestive) heart failure |
| I50.8 | Other heart failure |
| I50.9 | Cardiac, heart or myocardial failure NOS |

*Table A.8: ICD-10 codes for the identification of diabetes*

| ICD Code | Description |
|---|---|
| E10 | Type 1 diabetes mellitus |
| E11 | Type 2 diabetes mellitus |
| E12 | Malnutrition-related diabetes mellitus |
| E13 | Other specified diabetes mellitus |
| E14 | Unspecified diabetes mellitus |
| O24.2 | Pre-existing malnutrition-related diabetes mellitus |

*Table A.9: ICD-10 codes for the identification of chronic kidney diseases*

| ICD Code | Description |
|---|---|
| N18.1 | Chronic kidney disease, stage 1 |
| N18.2 | Chronic kidney disease, stage 2 |
| N18.3 | Chronic kidney disease, stage 3 |
| N18.4 | Chronic kidney disease, stage 4 |
| N18.5 | Chronic kidney disease, stage 5 |
| N18.9 | Chronic kidney disease, unspecified |
| T86.1 | Kidney transplant failure and rejection |
| I12.0 | Hypertensive renal failure |
| N00 | Acute nephritic syndrome |
| N03 | Chronic nephritic syndrome |
| N04 | Nephrotic syndrome |
| N05 | Unspecified nephritic syndrome |
| N11 | Chronic tubulo-interstitial nephritis |
| N13 | Obstructive and reflux uropathy |
| N17 | Acute renal failure |
| N19 | Unspecified kidney failure |
| E10.2 | Type 1 diabetes mellitus with kidney complications |
| E11.2 | Type 2 diabetes mellitus with kidney complications |

*Table A.10: ICD-10 codes for the identification of stroke*

| ICD Code | Description |
|---|---|
| I60 | Subarachnoid haemorrhage |
| I61 | Intracerebral haemorrhage |
| I62 | Other nontraumatic intracranial haemorrhage |

| | |
|---|---|
| I63 | Cerebral infarction |
| I64 | Stroke, not specified as haemorrhage or infarction |
| I65 | Occlusion and stenosis of precerebral arteries, not resulting in cerebral infarction |
| I66 | Occlusion and stenosis of cerebral arteries, not resulting in cerebral infarction |
| I67 | Other cerebrovascular diseases |
| I68 | Cerebrovascular disorders in diseases classified elsewhere |
| I69 | Sequelae of cerebrovascular disease |
| G45.9 | Transient cerebral ischaemic attack, unspecified |
| G46 | Vascular syndromes of brain in cerebrovascular diseases |

*Table A.11: Hi-BEHRT performance evaluation with different window and stride size*

| Window size | stride size | AUROC | AUPRC |
|---|---|---|---|
| 50 | 30 | 0.96 | 0.77 |
| 50 | 50 | 0.95 | 0.76 |
| 100 | 50 | 0.96 | 0.78 |
| 100 | 100 | 0.95 | 0.74 |
| 150 | 150 | 0.95 | 0.74 |

*Table A.12: Discrimination performance of DeepSurv trained with different batch size. All models are converged similarly in terms of discrimination performance.*

| Batch size | AUROC | AUPRC | C-statistics |
|---|---|---|---|
| 32 | 0.848 (±0.002) | 0.193 (±0.003) | 0.856 (±0.002) |
| 256 | 0.849 (±0.002) | 0.193 (±0.001) | 0.856 (±0.002) |
| 1024 | 0.849 (±0.002) | 0.192 (±0.003) | 0.856 (±0.002) |
| All | 0.845 (±0.004) | 0.193 (±0.003) | 0.853 (±0.004) |

*Table A.13: Model parameters for DeepSurv and MLP*

| Number of layers | Neurons | Batch norm | Dropout rate | Learning rate |
|---|---|---|---|---|
| 3 | 14, 14, 14 | True | 0.2 | 0.0001 |

*Table A.14: Relative contribution for diseases that occurred in at least 5% of the population*

| Disease | RC | 95% Confidence intervalS | |
|---|---|---|---|
| | | Lower bound | Upper bound |
| Dermatitis (atopic/contact/other/unspecified) | 1.76 | 1.56 | 1.98 |
| Bacterial diseases (excluding tuberculosis) | 1.56 | 1.42 | 1.71 |
| Other or unspecified infectious organisms | 1.47 | 1.37 | 1.58 |
| Lower respiratory tract infections | 1.47 | 1.37 | 1.57 |
| Depression | 1.45 | 1.29 | 1.63 |
| Iron deficiency anaemia | 1.4 | 1.22 | 1.61 |
| Pleural effusion | 1.38 | 1.24 | 1.53 |
| Myocardial infarction | 1.37 | 1.28 | 1.48 |
| Left bundle branch block | 1.34 | 1.13 | 1.60 |
| Other anaemias | 1.33 | 1.17 | 1.50 |
| Hypo or hyperthyroidism | 1.26 | 1.13 | 1.41 |
| Atrial fibrillation and flutter | 1.18 | 1.11 | 1.24 |
| Acute kidney injury | 1.15 | 1.01 | 1.31 |
| Hypertension | 1.08 | 1.03 | 1.14 |
| Hearing loss | 1.03 | 0.86 | 1.22 |
| Urinary tract infections | 1.00 | 0.88 | 1.12 |
| Enthesopathies & synovial disorders | 0.9 | 0.79 | 1.02 |
| Stroke not otherwise specified (NOS) | 0.89 | 0.79 | 0.99 |
| Chronic obstructive pulmonary disease (COPD) | 0.87 | 0.80 | 0.95 |

| | | | |
|---|---|---|---|
| Cataract | 0.84 | 0.75 | 0.94 |
| Hyperplasia of prostate | 0.82 | 0.72 | 0.95 |
| Stable angina | 0.81 | 0.75 | 0.88 |
| Osteoarthritis (excluding spine) | 0.75 | 0.68 | 0.84 |
| Diabetic ophthalmic complications | 0.7 | 0.62 | 0.79 |

RC: relative contribution

*Table A.15: Relative contribution for medications that occurred in at least 5% of the population*

| Medication | RC | 95% confidence intervals | |
|---|---|---|---|
| | | Lower bound | Upper bound |
| Chronic bowel disorders/ corticosteroids (respiratory)/ etc | 1.71 | 1.59 | 1.84 |
| Anaemias and other blood disorders | 1.53 | 1.42 | 1.66 |
| Bronchodilators | 1.52 | 1.43 | 1.63 |
| Cough preparations | 1.45 | 1.28 | 1.64 |
| Oral nutrition | 1.41 | 1.68 | 1.68 |
| Antibacterial drugs/ antiprotozoal drugs/ acne and rosacea | 1.41 | 1.25 | 1.58 |
| Hypnotics and anxiolytics | 1.39 | 1.22 | 1.57 |
| Antidepressant drugs | 1.38 | 1.28 | 1.50 |
| Drugs acting on the ear | 1.36 | 1.2 | 1.54 |
| Antibacterial drugs | 1.32 | 1.26 | 1.39 |
| Antibacterial drugs/ acne and rosacea | 1.31 | 1.2 | 1.44 |
| Drugs used in nausea and vertigo | 1.27 | 1.16 | 1.40 |
| Soft-tissue disorders and topical pain relief | 1.24 | 1.13 | 1.36 |
| Drugs acting on the oropharynx | 1.22 | 1.07 | 1.41 |
| Hypnotics and anxiolytics/ general anaesthesia | 1.22 | 1.06 | 1.40 |
| Analgesics/ analgesics | 1.21 | 1.13 | 1.29 |
| Corticosteroids (respiratory) | 1.20 | 1.09 | 1.32 |
| Drugs used in rheumatic diseases and gout | 1.19 | 1.11 | 1.28 |
| Anti-infective eye preparations | 1.17 | 1.05 | 1.05 |
| Laxatives | 1.16 | 1.08 | 1.25 |
| Anti-infective skin preparations | 1.14 | 1.01 | 1.29 |
| Vaccines and antisera | 1.13 | 1.06 | 1.20 |
| Antihistamine, hyposensitivity and allergic emergent | 1.11 | 1.01 | 1.22 |
| Dyspepsia and gastro-oesophageal reflux disease | 1.09 | 0.98 | 1.20 |
| Antiplatelet drugs | 1.08 | 1.02 | 1.14 |
| Emollient and barrier preparations | 1.08 | 1.00 | 1.16 |
| Sex hormones | 1.07 | 0.92 | 1.25 |
| Nitrates, calcium-channel blocker and other antianginal drugs | 1.06 | 1.01 | 1.12 |
| Hypnotics and anxiolytics/ antiepileptic drugs/ etc | 1.06 | 0.94 | 1.20 |
| Miscellaneous ophthalmic preparations | 1.05 | 0.92 | 1.19 |
| Topical corticosteroids | 1.05 | 0.98 | 1.11 |
| Local preparations for anal and rectal disorders | 1.03 | 0.9 | 1.17 |
| Corticosteroids and other anti-inflammatory preparations | 1.02 | 0.88 | 1.17 |
| Drugs acting on the nose | 1.01 | 0.93 | 1.11 |
| Acute diarrhoea | 1.01 | 0.88 | 1.15 |
| Topical corticosteroids/ anti-infective skin preparations | 0.99 | 0.86 | 1.13 |
| Analgesics | 0.98 | 0.93 | 1.03 |
| Hypertension and heart failure | 0.95 | 0.91 | 1.00 |
| Drugs used in psychoses and Related disorders/ drugs used in nausea and vertigo | 0.95 | 0.82 | 1.11 |
| Diuretics | 0.93 | 0.88 | 0.98 |
| Acute diarrhoea/ cough preparations/ analgesics | 0.92 | 0.82 | 1.03 |
| Bronchodilators/ corticosteroids (respiratory) | 0.92 | 0.83 | 1.01 |
| Lipid-regulating drugs | 0.91 | 0.87 | 0.97 |
| Analgesics/ drugs used in rheumatic diseases and gout | 0.91 | 0.79 | 1.05 |
| Antidepressant drugs/ analgesics/ analgesics | 0.91 | 0.82 | 1.01 |

| | | | |
|---|---|---|---|
| Beta-adrenoceptor blocking drugs | 0.87 | 0.79 | 0.95 |
| Antisecretory drugs and mucosal protectants | 0.85 | 0.81 | 0.90 |
| Diuretics/ minerals | 0.85 | 0.77 | 0.93 |
| Drugs for Genito-urinary disorders | 0.81 | 0.73 | 0.91 |
| Thyroid and antithyroid drugs | 0.81 | 0.73 | 0.9 |
| Treatment of glaucoma | 0.77 | 0.64 | 0.92 |
| Hypertension and heart failure/ drugs for genito-urinary disorders | 0.71 | 0.65 | 0.78 |
| Drugs used in diabetes | 0.71 | 0.66 | 0.77 |
| Vitamins | 0.7 | 0.63 | 0.79 |
| Drugs affecting bone metabolism | 0.7 | 0.78 | 0.78 |
| Antiprotozoal drugs/ drugs used in neuromuscular disorders | 0.67 | 0.58 | 0.77 |
| Beta-adrenoceptor blocking drugs/ hypertension drugs/etc | 0.67 | 0.61 | 0.73 |
| Positive inotropic drugs | 0.61 | 0.54 | 0.69 |
| Anti-arrhythmic drugs | 0.56 | 0.48 | 0.65 |
| Anticoagulants and protamine | 0.53 | 0.48 | 0.57 |

RC: relative contribution

*Table A.16: ICD-10 codes for the identification of depression*

| ICD Code | Description |
|---|---|
| F32.1 | Mild depressive episode |
| F32.2 | Moderate depressive episode |
| F32.3 | Severe depressive episode without psychotic symptoms |
| F32.8 | Severe depressive episode with psychotic symptoms |
| F32.9 | Other depressive episodes |
| F33.0 | Depressive episode, unspecified |
| F33.1 | Recurrent depressive disorder, current episode mild |
| F33.2 | Recurrent depressive disorder, current episode moderate |
| F33.3 | Recurrent depressive disorder, current episode severe without psychotic symptoms |
| F33.4 | Recurrent depressive disorder, current episode severe with psychotic symptoms |
| F33.8 | Recurrent depressive disorder, currently in remission |
| F33.9 | Other recurrent depressive disorders |
| F34.1 | Recurrent depressive disorder, unspecified |
| F38.1 | Dysthymia |

# References


1 Force USPST. Statin Use for the Primary Prevention of Cardiovascular Disease in Adults: US Preventive Services Task Force Recommendation Statement. *JAMA* 2016; **316**: 1997–2007.

2 Finnikin S, Ryan R, Marshall T. Statin initiations and QRISK2 scoring in UK general practice: A THIN database study. *British Journal of General Practice* 2017; **67**: e881–7.

3 Leopold JA, Loscalzo J. Emerging Role of Precision Medicine in Cardiovascular Disease. *Circ Res* 2018; **122**: 1302–15.

4 Pylypchuk R, Wells S, Kerr A, *et al.* Cardiovascular risk prediction in type 2 diabetes before and after widespread screening: a derivation and validation study. *Lancet* 2021; **397**: 2264–74.

5 Gal Y. Uncertainty in Deep Learning. 2016. https://mlg.eng.cam.ac.uk/yarin/thesis/thesis.pdf (accessed Aug 16, 2022).

6 Molnar C. Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. 2022 https://christophm.github.io/interpretable-ml-book (accessed Aug 16, 2022).

7 Durán JM, Jongsma KR. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *J Med Ethics* 2021; **47**: 329 LP – 335.

8 Herrett E, Gallagher AM, Bhaskaran K, *et al.* Data Resource Profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol* 2015; **44**: 827–36.

9 National INstitute for Health and Care Excellence Cardiovascular disease (NICE guideline 181). 2014. https://www.nice.org.uk/guidance/cg181 (accessed Feb 28, 2022).

10 Clark TG, Bradburn MJ, Love SB, Altman DG. Survival Analysis Part I: Basic concepts and first analyses. *Br J Cancer* 2003; **89**: 232–8.

11 Dawson D v., Blanchette DR, Pihlstrom BL. Application of Biostatistics in Dental Public Health. *Burt and Eklund's Dentistry, Dental Practice, and the Community* 2021; : 131–53.

12 Harrington D, Gail MH. Lifetime Data Anal: Editorial. Lifetime Data Anal. 2008; **14**. DOI:10.1007/s10985-008-9081-5.

13 Gunter TD, Terry NP. The emergence of national electronic health record architectures in the United States and Australia: models, costs, and questions. *J Med Internet Res* 2005; **7**: e3–e3.

14 Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet* 2012; **13**: 395–405.

15 Casey JA, Schwartz BS, Stewart WF, Adler NE. Using Electronic Health Records for Population Health Research: A Review of Methods and Applications. *Annu Rev Public Health* 2016; **37**: 61–81.

16 Wood A, Denholm R, Hollings S, *et al.* Linked electronic health records for research on a nationwide cohort of more than 54 million people in England: data resource. *BMJ* 2021; **373**: n826.

17 Goldstein BA, Navar AM, Pencina MJ. Risk Prediction With Electronic Health Records: The Importance of Model Validation and Clinical Context. *JAMA Cardiol* 2016; **1**: 976–7.

18 Galea S, Tracy M. Participation Rates in Epidemiologic Studies. *Ann Epidemiol* 2007; **17**: 643–53.
19 Hordijk-Trion M, Lenzen M, Wijns W, *et al.* Patients enrolled in coronary intervention trials are not representative of patients in clinical practice: results from the Euro Heart Survey on Coronary Revascularization. *Eur Heart J* 2006; **27**: 671–8.
20 Cowie MR, Blomster JI, Curtis LH, *et al.* Electronic health records to facilitate clinical research. *Clin Res Cardiol* 2017; **106**: 1–9.
21 Rasmussen L V. The electronic health record for translational research. *J Cardiovasc Transl Res* 2014; **7**: 607–14.
22 Ni K, Chu H, Zeng L, Li N, Zhao Y. Barriers and facilitators to data quality of electronic health records used for clinical research in China: a qualitative study. *BMJ Open* 2019; **9**: e029314.
23 Schneider A, Hommel G, Blettner M. Linear regression analysis: part 14 of a series on evaluation of scientific publications. *Dtsch Arztebl Int* 2010; **107**: 776–82.
24 Murphy KP. Machine Learning: A Probabilistic Perspective. The MIT Press, 2012.
25 Montgomery DC, Peck EA, Vining GG. Introduction to linear regression analysis. John Wiley & Sons, 2021.
26 Bishop CM, Nasrabadi NM. Pattern Recognition and Machine Learning. Springer, 2006.
27 Goodfellow I, Bengio Y, Courville A. Deep Learning. MIT press, 2016.
28 Emmert-Streib F, Yang Z, Feng H, Tripathi S, Dehmer M. An Introductory Review of Deep Learning for Prediction Models With Big Data . Frontiers in Artificial Intelligence . 2020; **3**.
29 He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770–8.
30 Szegedy C, Liu W, Jia Y, *et al.* Going Deeper With Convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1–9.
31 Yu Y, Si X, Hu C, Zhang J. A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. *Neural Comput* 2019; **31**: 1235–70.
32 Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. *Insights Imaging* 2018; **9**: 611–29.
33 Albawi S, Mohammed TA, Al-Zawi S. Understanding of a convolutional neural network. In: 2017 International Conference on Engineering and Technology (ICET). 2017: 1–6.
34 Nagi J, Ducatelle F, Caro GA Di, *et al.* Max-pooling convolutional neural networks for vision-based hand gesture recognition. In: 2011 IEEE International Conference on Signal and Image Processing Applications (ICSIPA). 2011: 342–7.
35 Sherstinsky A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D* 2020; **404**: 132306.
36 Zia T, Zahid U. Long short-term memory recurrent neural network architectures for Urdu acoustic modeling. *Int J Speech Technol* 2019; **22**: 21–30.
37 Vaswani A, Shazeer N, Parmar N, *et al.* Attention is all you need. *Adv Neural Inf Process Syst* 2017; **2017-Decem**: 5999–6009.
38 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; : 4171–86.
39 Anderson JR, Crawford J. Cognitive Psychology and Its Implications. wh freeman San Francisco, 1980.

40 Bahdanau D, Cho KH, Bengio Y. Neural machine translation by jointly learning to align and translate. In: 3rd International Conference on Learning Representations, ICLR 2015. 2015.
41 Hu D. An introductory survey on attention mechanisms in NLP problems. In: Proceedings of SAI Intelligent Systems Conference. Springer, 2019: 432–48.
42 Lee JB, Rossi RA, Kim S, Ahmed NK, Koh E. Attention models in graphs: A survey. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 2019; **13**: 1–25.
43 Wu C, Wei Y, Chu X, Weichen S, Su F, Wang L. Hierarchical attention-based multimodal fusion for video captioning. *Neurocomputing* 2018; **315**: 362–70.
44 Cheng J, Dong L, Lapata M. Long Short-Term Memory-Networks for Machine Reading. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016: 551–61.
45 Shaw P, Uszkoreit J, Vaswani A. Self-Attention with Relative Position Representations. BT - Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, . 2018; : 464–8.
46 Camburu O-M. Explaining deep neural networks. *arXiv preprint arXiv:201001496* 2020.
47 Telgarsky M. Benefits of depth in neural networks. In: Conference on learning theory. PMLR, 2016: 1517–39.
48 Yoon CH, Torrance R, Scheinerman N. Machine learning in medicine: should the pursuit of enhanced interpretability be abandoned? *J Med Ethics* 2021; : medethics-2020-107102.
49 Sha Y, Wang MD. Interpretable Predictions of Clinical Outcomes with An Attention-based Recurrent Neural Network. *ACM BCB* 2017; **2017**: 233–40.
50 Ribeiro MT, Singh S, Guestrin C. 'Why Should I Trust You?' Explaining the Predictions of Any Classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 2016; **13-17-Augu**: 1135–44.
51 Li J, Chen X, Hovy E, Jurafsky D. Visualizing and understanding neural models in NLP. *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference* 2016; : 681–91.
52 Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity checks for saliency maps. *Adv Neural Inf Process Syst* 2018; **31**.
53 Ancona M, Ceolini E, Öztireli C, Gross M. Towards better understanding of gradient-based attribution methods for Deep Neural Networks. In: International Conference on Learning Representations. 2018.
54 Sundararajan M, Najmi A. The many Shapley values for model explanation. In: International conference on machine learning. PMLR, 2020: 9269–78.
55 Pearl J, Mackenzie D. The Book of Why: The New Science of Cause and Effect, 1st edn. USA: Basic Books, Inc., 2018.
56 Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst* 2017; **30**.
57 Pearl J. Causality. Cambridge university press, 2009.
58 Koh PW, Nguyen T, Tang YS, *et al.* Concept bottleneck models. In: International Conference on Machine Learning. PMLR, 2020: 5338–48.

59 Kim B, Wattenberg M, Gilmer J, Cai C, Wexler J, Viegas F. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In: International conference on machine learning. PMLR, 2018: 2668–77.
60 Yeh C-K, Kim B, Arik S, Li C-L, Pfister T, Ravikumar P. On completeness-aware concept-based explanations in deep neural networks. *Adv Neural Inf Process Syst* 2020; **33**: 20554–65.
61 Lewis D. Counterfactuals. John Wiley & Sons, 2013.
62 Balke A, Pearl J. Probabilistic Evaluation of Counterfactual Queries. 2011.
63 Krzywinski M, Altman N. Importance of being uncertain. *Nat Methods* 2013; **10**: 809–10.
64 Kendall A, Gal Y. What uncertainties do we need in Bayesian deep learning for computer vision? *Adv Neural Inf Process Syst* 2017; **2017-Decem**: 5575–85.
65 Kim K, Lee Y-M. Understanding uncertainty in medicine: concepts and implications in medical education. *Korean J Med Educ* 2018; **30**: 181–8.
66 Dusenberry MW, Tran D, Choi E, *et al.* Analyzing the role of model uncertainty for electronic health records. In: ACM CHIL 2020 - Proceedings of the 2020 ACM Conference on Health, Inference, and Learning. 2020. DOI:10.1145/3368555.3384457.
67 Blei DM, Kucukelbir A, McAuliffe JD. Variational Inference: A Review for Statisticians. *J Am Stat Assoc* 2017; **112**: 859–77.
68 Maddox WJ, Garipov T, Izmailov, Vetrov D, Wilson AG. A simple baseline for Bayesian uncertainty in deep learning. In: Advances in Neural Information Processing Systems. 2019.
69 Gneiting T, Raftery AE. Strictly Proper Scoring Rules, Prediction, and Estimation. *J Am Stat Assoc* 2007; **102**: 359–78.
70 Bernardo JM, Smith AFM. Bayesian theory. John Wiley & Sons, 2009.
71 Zhang C, Butepage J, Kjellstrom H, Mandt S. Advances in Variational Inference. *IEEE Trans Pattern Anal Mach Intell* 2019; **41**: 2008–26.
72 Seeger M. Gaussian processes for machine learning. Int J Neural Syst. 2004; **14**. DOI:10.1142/S0129065704001899.
73 Titsias M. Variational learning of inducing variables in sparse Gaussian processes. In: Artificial Intelligence and Statistics. 2009: 567–74.
74 Wilson AG, Nickisch H. Kernel interpolation for scalable structured Gaussian processes (KISS-GP). In: 32nd International Conference on Machine Learning, ICML 2015. 2015.
75 Agency M and HPR. Clinical Practice Research Datalink - CPRD. https://www.cprd.com (accessed July 29, 2022).
76 Solares JRA, Raimondi FED, Zhu Y, *et al.* Deep learning for electronic health records: A comparative review of multiple deep neural architectures. *J Biomed Inform* 2020; **101**: 103337.
77 Li Y, Mamouei M, Salimi-khorshidi G, *et al.* Hi-BEHRT : Hierarchical Transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records. ; : 1–11.
78 Mc Cord KA, Hemkens LG. Using electronic health records for clinical trials: Where do we stand and where can we go? *CMAJ* 2019; **191**: E128–33.
79 Chisholm J. The Read clinical classification. *BMJ: British Medical Journal* 1990; **300**: 1092.
80 Joint Formulary Committee. British National Formulary. BMJ Group and Pharmaceutical Press. https://bnf.nice.org.uk/drugs/paracetamol/ (accessed Feb 27, 2020).

81 Payne, R. (Creator), Denholm R (Creator). CPRD product code lists used to define long-term preventative, high-risk, and palliative medication. 2018. DOI:10.5523/bris.k38ghxkcub622603i5wq6bwag.
82 Herbert A, Wijlaars L, Zylbersztejn A, Cromwell D, Hardelid P. Data Resource Profile: Hospital Episode Statistics Admitted Patient Care (HES APC). *Int J Epidemiol* 2017; **46**: 1093–1093i.
83 Bramer GR. International statistical classification of diseases and related health problems - Tenth revision. *World Health Statistics Quarterly* 1988; **41**: 32–6.
84 NHS. OPCS Classification of Interventions and Procedures. *NHS digital* 2016.
85 Health T, Care Information Centre All Rights Reserved S. HOSPITAL EPISODE STATISTICS: A guide to linked ONS-HES mortality data. 2015.
86 English indices of deprivation 2015 - GOV.UK. https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015 (accessed June 30, 2022).
87 HES Data Dictionary: Admitted Patient Care Admitted Patient Care (APC) Hospital Episode Statistics (HES) Data Dictionary HES Data Dictionary. 2016.
88 Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) The TRIPOD Statement. *Circulation* 2015; **131**: 211–9.
89 Rahimian F, Salimi-Khorshidi G, Payberah AH, *et al.* Predicting the risk of emergency admission with machine learning: Development and validation using linked electronic health records. *PLoS Med* 2018; **15**: e1002695–e1002695.
90 Ayala Solares JR, Diletta Raimondi FE, Zhu Y, *et al.* Deep learning for electronic health records: A comparative review of multiple deep neural architectures. *J Biomed Inform* 2020; **101**: 103337.
91 Parasrampuria S, Henry J. Hospitals' Use of Electronic Health Records Data, 2015-2017. *Hospitals (Lond)* 2015; : 2015–7.
92 of Health D, Services H, Office of the National Coordinator for Health Information Technology T. Electronic Public Health Reporting ONC Annual Meeting. 2018.
93 Ardila D, Kiraly AP, Bharadwaj S, *et al.* End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat Med* 2019; **25**: 954–61.
94 Poplin R, Varadarajan A V, Blumer K, *et al.* Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng* 2018; **2**: 158–64.
95 Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019; **25**: 44–56.
96 Esteva A, Robicquet A, Ramsundar B, *et al.* A guide to deep learning in healthcare. *Nat Med* 2019; **25**: 24–9.
97 Liang Z, Zhang G, Huang JX, Hu Q V. Deep learning for healthcare decision making with EMRs. In: 2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2014: 556–9.
98 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* 2015; **54**: 96–105.
99 Miotto R, Li L, Kidd BA, Dudley JT. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Sci Rep* 2016; **6**: 26094.

100 Cao LJ, Chua KS, Chong WK, Lee HP, Gu QM. A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine. *Neurocomputing* 2003; **55**: 321–36.
101 Nguyen P, Tran T, Wickramasinghe N, Venkatesh S. Deepr: A Convolutional Net for Medical Records. *IEEE J Biomed Health Inform* 2017; **21**: 22–30.
102 Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. In: Machine Learning for Healthcare Conference. 2016: 301–18.
103 Pham T, Tran T, Phung D, Venkatesh S. DeepCare: A Deep Dynamic Memory Model for Predictive Medicine BT - Advances in Knowledge Discovery and Data Mining. In: Bailey J, Khan L, Washio T, Dobbie G, Huang JZ, Wang R, eds. . Cham: Springer International Publishing, 2016: 30–41.
104 Choi E, Bahadori MT, Kulas JA, Schuetz A, Stewart WF, Sun J. RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. *Adv Neural Inf Process Syst* 2016; : 3512–20.
105 Kuan V, Denaxas S, Gonzalez-Izquierdo A, *et al.* 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**: e63–77.
106 Snoek J, Larochelle H, Adams RP. Practical bayesian optimization of machine learning algorithms. *Adv Neural Inf Process Syst* 2012; **25**.
107 Kingma DP, Ba JL. Adam: A method for stochastic optimization. In: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. 2015.
108 van der Maaten L. Accelerating t-SNE using tree-based algorithms. *The Journal of Machine Learning Research* 2014; **15**: 3221–45.
109 Connor R. A tale of four metrics. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2016. DOI:10.1007/978-3-319-46759-7_16.
110 Fawcett T. An introduction to ROC analysis. *Pattern Recognit Lett* 2006; **27**: 861–74.
111 Zhu M. Recall, precision and average precision. *Department of Statistics and Actuarial Science, University of Waterloo, Waterloo* 2004; **2**: 6.
112 Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ* 2017; **357**: j2099.
113 Al-Shamsi S. Performance of the Framingham coronary heart disease risk score for predicting 10-year cardiac risk in adult United Arab Emirates nationals without diabetes: a retrospective cohort study. *BMC Fam Pract* 2020; **21**: 175.
114 de la Iglesia B, Potter JF, Poulter NR, Robins MM, Skinner J. Performance of the ASSIGN cardiovascular disease risk score on a UK cohort of patients from general practice. *Heart* 2011; **97**: 491–9.
115 Yang Z, Yu Y, You C, Steinhardt J, Ma Y. Rethinking bias-variance trade-off for generalization of neural networks. *37th International Conference on Machine Learning, ICML 2020* 2020; **PartF16814**: 10698–708.
116 Batty M, Torrens PM. Modelling complexity: the limits to prediction. *Cybergeo: European Journal of Geography* 2001.
117 Li Y, Rao S, Solares JRA, *et al.* BEHRT: Transformer for Electronic Health Records. *Sci Rep* 2020; **10**: 7155.
118 Cho S-Y, Kim S-H, Kang S-H, *et al.* Pre-existing and machine learning-based models for cardiovascular risk prediction. *Sci Rep* 2021; **11**: 8886.

119 Rahimian F, Salimi-Khorshidi G, Payberah AH, *et al.* Predicting the risk of emergency admission with machine learning: Development and validation using linked electronic health records. *PLoS Med* 2018; **15**: e1002695.
120 Tiwari P, Colborn KL, Smith DE, Xing F, Ghosh D, Rosenberg MA. Assessment of a Machine Learning Model Applied to Harmonized Electronic Health Record Data for the Prediction of Incident Atrial Fibrillation. *JAMA Netw Open* 2020; **3**: e1919396–e1919396.
121 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. *BMJ* 2020; **371**. DOI:10.1136/bmj.m3919.
122 Kannel WB, D'Agostino RB, Silbershatz H, Belanger AJ, Wilson PWF, Levy D. Profile for Estimating Risk of Heart Failure. *Arch Intern Med* 1999; **159**: 1197–204.
123 Agarwal SK, Chambless LE, Ballantyne CM, *et al.* Prediction of incident heart failure in general practice: the Atherosclerosis Risk in Communities (ARIC) Study. *Circ Heart Fail* 2012; **5**: 422–9.
124 Flueckiger P, Longstreth W, Herrington D, Yeboah J. Revised Framingham Stroke Risk Score, Nontraditional Risk Markers, and Incident Stroke in a Multiethnic Cohort. *Stroke* 2018; **49**: 363–9.
125 Van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. *J Stat Softw* 2011; **45**: 1–67.
126 Hassaine A, Canoy D, Solares JRA, *et al.* Learning multimorbidity patterns from electronic health records using Non-negative Matrix Factorisation. *J Biomed Inform* 2020; **112**: 103606.
127 Read Codes V2 and Clinical Terms V3 Cross-maps - NHS Digital - Citizen Space. https://nhs-digital.citizenspace.com/uktc/crossmaps/ (accessed July 8, 2022).
128 SNOMED CT - NHS Digital. https://digital.nhs.uk/services/terminology-and-classifications/snomed-ct (accessed July 8, 2022).
129 Davidson-Pilon C. lifelines: survival analysis in Python. *J Open Source Softw* 2019; **4**: 1317.
130 Pedregosa F, Varoquaux G, Gramfort A, *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 2011; **12**.
131 Paszke A, Gross S, Massa F, *et al.* Pytorch: An imperative style, high-performance deep learning library. In: Advances in neural information processing systems. 2019: 8026–37.
132 Sáez C, Gutiérrez-Sacristán A, Kohane I, García-Gómez JM, Avillach P. EHRtemporalVariability: delineating temporal data-set shifts in electronic health records. *Gigascience* 2020; **9**. DOI:10.1093/gigascience/giaa079.
133 Rob J Hyndman, George A. Forecasting: Principles and Practice. *Principles of Optimal Design* 2014.
134 Ni C, Charoenphakdee N, Honda J, Sugiyama M. On the calibration of multiclass classification with rejection. In: Advances in Neural Information Processing Systems. 2019.
135 Yurdakul B, Naranjo J. Statistical properties of the population stability index. *Journal of Risk Model Validation* 2020; **14**. DOI:10.21314/JRMV.2020.227.
136 Moreno-Torres JG, Raeder T, Alaiz-Rodríguez R, Chawla N V., Herrera F. A unifying view on dataset shift in classification. *Pattern Recognit* 2012; **45**: 521–30.
137 Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS One* 2017; **12**: e0174944.

138 Khera R, Haimovich J, Hurley NC, *et al.* Use of Machine Learning Models to Predict Death After Acute Myocardial Infarction. *JAMA Cardiol* 2021; **6**: 633–41.
139 Dockès J, Varoquaux G, Poline J-B. Preventing dataset shift from breaking machine-learning biomarkers. *Gigascience* 2021; **10**: giab055.
140 Vock DM, Wolfson J, Bandyopadhyay S, *et al.* Adapting machine learning techniques to censored time-to-event health record data: A general-purpose approach using inverse probability of censoring weighting. *J Biomed Inform* 2016; **61**: 119–31.
141 Goldstein BA, Navar AM, Pencina MJ, Ioannidis J. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association* 2017; **24**: 198–208.
142 Banda JM, Seneviratne M, Hernandez-Boussard T, Shah NH. Advances in Electronic Phenotyping: From Rule-Based Definitions to Machine Learning Models. *Annu Rev Biomed Data Sci* 2018; **1**: 53–68.
143 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**: 301–18.
144 Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digit Med* 2021; **4**: 86.
145 Beltagy I, Peters ME, Cohan A. Longformer: The long-document transformer. *arXiv preprint arXiv:200405150* 2020.
146 Choromanski K, Likhosherstov V, Dohan D, *et al.* Rethinking Attention with Performers. In: International Conference on Learning Representations. 2021. https://openreview.net/forum?id=Ua6zuk0WRH.
147 Wang S, Li BZ, Khabsa M, Fang H, Ma H. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:200604768* 2020.
148 Pappagari R, Zelasko P, Villalba J, Carmiel Y, Dehak N. Hierarchical Transformers for Long Document Classification. *2019 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2019 - Proceedings* 2019; : 838–44.
149 Conrad N, Judge A, Tran J, *et al.* Temporal trends and patterns in heart failure incidence : a population-based study of 4 million individuals. *The Lancet* 2018; **391**: 572–80.
150 Kuan V, Denaxas S, Gonzalez-Izquierdo A, *et al.* 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**: e63–77.
151 Tran J, Norton R, Conrad N, *et al.* 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.
152 Jager KJ, Fraser SDS. The ascending rank of chronic kidney disease in the global burden of disease study. *Nephrology Dialysis Transplantation* 2017; **32**: ii121–8.
153 Centers for Disease Control and Prevention. Chronic Kidney Disease Surveillance Project - United States. CDC: Centers for Disease Control and Prevention. 2014. https://nccd.cdc.gov/ckd/Methods.aspx?Qnum=Q637 (accessed May 4, 2021).
154 Baevski A, Zhou H, Mohamed A, Auli M. wav2vec 2.0: A framework for self-supervised learning of speech representations. In: Advances in Neural Information Processing Systems. 2020.

155 Alaa AM, Bolton T, Di Angelantonio E, Rudd JHF, van der Schaar M. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. *PLoS One* 2019; **14**: 1–17.
156 Kim JOR, Jeong Y-S, Kim JH, Lee J-W, Park D, Kim H-S. Machine Learning-Based Cardiovascular Disease Prediction Model: A Cohort Study on the Korean National Health Insurance Service Health Screening Database. *Diagnostics (Basel)* 2021; **11**: 943.
157 Dziopa K, Asselbergs FW, Gratton J, Chaturvedi N, Schmidt AF. Cardiovascular risk prediction in type 2 diabetes: a comparison of 22 risk scores in primary care settings. *Diabetologia* 2022; **65**: 644–56.
158 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* 2018; **18**: 24.
159 Tarkhan A, Simon N. BigSurvSGD: Big Survival Data Analysis via Stochastic Gradient Descent. *arXiv preprint arXiv:200300116* 2020.
160 Kvamme H, Borgan Ø, Scheel I. Time-to-Event Prediction with Neural Networks and Cox Regression. *Journal of Machine Learning Research* 2019; **20**: 1–30.
161 Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng. 2010; **22**. DOI:10.1109/TKDE.2009.191.
162 Efron B. The jackknife, the bootstrap and other resampling plans. SIAM, 1982.
163 Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei L-J. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat Med* 2011; **30**: 1105–17.
164 Austin PC, Harrell Jr FE, van Klaveren D. Graphical calibration curves and the integrated calibration index (ICI) for survival models. *Stat Med* 2020; **39**: 2714–42.
165 Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ* 2016; **352**. DOI:10.1136/bmj.i6.
166 Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? *Int J Methods Psychiatr Res* 2011; **20**: 40–9.
167 Visseren FLJ, Mach F, Smulders YM, *et al.* 2021 ESC Guidelines on cardiovascular disease prevention in clinical practice: Developed by the Task Force for cardiovascular disease prevention in clinical practice with representatives of the European Society of Cardiology and 12 medical societies With the special contribution of the European Association of Preventive Cardiology (EAPC). *Eur Heart J* 2021; **42**: 3227–337.
168 Dimopoulos AC, Nikolaidou M, Caballero FF, *et al.* Machine learning methodologies versus cardiovascular risk scores, in predicting disease risk. *BMC Med Res Methodol* 2018; **18**: 179.
169 Rieke N, Hancox J, Li W, *et al.* The future of digital health with federated learning. *NPJ Digit Med* 2020; **3**. DOI:10.1038/s41746-020-00323-1.
170 Beaulieu-Jones BK, Greene CS. Semi-supervised learning of the electronic health record for phenotype stratification. *J Biomed Inform* 2016; **64**: 168–78.
171 Kwon BC, Choi MJ, Kim JT, *et al.* RetainVis: Visual Analytics with Interpretable and Interactive Recurrent Neural Networks on Electronic Medical Records. *IEEE Trans Vis Comput Graph* 2019; **25**. DOI:10.1109/TVCG.2018.2865027.
172 Smilkov D, Thorat N, Kim B, Viégas F, Wattenberg M. SmoothGrad: removing noise by adding noise. 2017. http://arxiv.org/abs/1706.03825.

173 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* 2019; **97**: 2454–63.
174 Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. In: 34th International Conference on Machine Learning, ICML 2017. 2017.
175 Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps. In: 2nd International Conference on Learning Representations, ICLR 2014 - Workshop Track Proceedings. 2014.
176 Friedrich JO, Adhikari NKJ, Beyene J. The ratio of means method as an alternative to mean differences for analyzing continuous outcome variables in meta-analysis: A simulation study. *BMC Med Res Methodol* 2008; **8**. DOI:10.1186/1471-2288-8-32.
177 Tromp J, Paniagua SMA, Lau ES, *et al.* Age dependent associations of risk factors with heart failure: pooled population based cohort study. *BMJ* 2021; **372**. DOI:10.1136/bmj.n461.
178 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* 1992; **45**: 1053–9.
179 Clayton TC, Thompson M, Meade TW. Recent respiratory infection and risk of cardiovascular disease: Case-control study through a general practice database. *Eur Heart J* 2008; **29**. DOI:10.1093/eurheartj/ehm516.
180 Jenča D, Melenovský V, Stehlik J, *et al.* Heart failure after myocardial infarction: incidence and predictors. ESC Heart Fail. 2021; **8**. DOI:10.1002/ehf2.13144.
181 Ng MKC, Celermajer DS. Glucocorticoid treatment and cardiovascular disease. Heart. 2004; **90**. DOI:10.1136/hrt.2003.031492.
182 Agabiti N, Corbo GM. COPD and bronchodilators: should the heart pay the bill for the lung? European Respiratory Journal. 2017; **49**. DOI:10.1183/13993003.00370-2017.
183 Mäenpää J, Pelkonen O. Cardiac safety of ophthalmic timolol. *Expert Opin Drug Saf* 2016. DOI:10.1080/14740338.2016.1225718.
184 Harasymowycz P, Birt C, Gooi P, *et al.* Medical Management of Glaucoma in the 21st Century from a Canadian Perspective. J Ophthalmol. 2016. DOI:10.1155/2016/6509809.
185 Abraham WT. β-blockers: The new standard of therapy for mild heart failure. Arch Intern Med. 2000. DOI:10.1001/archinte.160.9.1237.
186 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* 2000. DOI:10.1161/01.CIR.102.4.426.
187 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* 2003. DOI:10.1016/S0735-1097(03)00299-7.
188 Pass HI, Pogrebniak HW. Potential uses of prostaglandin E1 analog for cardiovascular disease [5]. Journal of Thoracic and Cardiovascular Surgery. 1994. DOI:10.1016/s0022-5223(94)70312-4.
189 Goyal Y, Feder A, Shalit U, Kim B. Explaining classifiers with causal concept effect (cace). *arXiv preprint arXiv:190707165* 2019.

190 Yang M, Kim B. Benchmarking attribution methods with relative feature importance. *arXiv preprint arXiv:190709701* 2019.
191 O\textquotesingle Shaughnessy M, Canal G, Connor M, Rozell C, Davenport M. Generative causal explanations of black-box classifiers. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, eds. Advances in Neural Information Processing Systems. Curran Associates, Inc., 2020: 5453–67.
192 Szklo M, Nieto FJ. Epidemiology: beyond the basics. Jones & Bartlett Publishers, 2014.
193 Conrad N, Judge A, Tran J, *et al.* Temporal trends and patterns in heart failure incidence: a population-based study of 4 million individuals. *The Lancet* 2018; **391**: 572–80.
194 Baevski A, Schneider S, Auli M. vq-wav2vec: Self-supervised learning of discrete speech representations. In: International Conference on Learning Representations. 2020. https://openreview.net/forum?id=rylwJxrYDS.
195 van den Oord A, Vinyals O, kavukcuoglu koray. Neural Discrete Representation Learning. In: Guyon I, Luxburg U v, Bengio S, *et al.*, eds. Advances in Neural Information Processing Systems. Curran Associates, Inc., 2017. https://proceedings.neurips.cc/paper/2017/file/7a98af17e63a0ac09ce2e96d03992fbc-Paper.pdf.
196 Lex A, Gehlenborg N, Strobelt H, Vuillemot R, Pfister H. UpSet: Visualization of Intersecting Sets. *IEEE Trans Vis Comput Graph* 2014; **20**: 1983–92.
197 Verdecchia P, Angeli F, Reboldi G. Hypertension and atrial fibrillation: Doubts and certainties from basic and clinical studies. Circ Res. 2018; **122**: 352–68.
198 Odutayo A, Wong CX, Hsiao AJ, Hopewell S, Altman DG, Emdin CA. Atrial fibrillation and risks of cardiovascular disease, renal disease, and death: systematic review and meta-analysis. *BMJ* 2016; **354**: i4482.
199 Copland E, Canoy D, Nazarzadeh M, *et al.* Antihypertensive treatment and risk of cancer: an individual participant data meta-analysis. *Lancet Oncol* 2021; **22**. DOI:10.1016/S1470-2045(21)00033-4.
200 Chang P-Y, Huang W-Y, Lin C-L, *et al.* Propranolol Reduces Cancer Risk: A Population-Based Cohort Study. *Medicine* 2015; **94**. DOI:10.1097/MD.0000000000001097.
201 Johannesdottir SA, Schmidt M, Phillips G, *et al.* Use of ß-blockers and mortality following ovarian cancer diagnosis: A population-based cohort study. *BMC Cancer* 2013; **13**. DOI:10.1186/1471-2407-13-85.
202 Štrumbelj E, Kononenko I. Explaining prediction models and individual predictions with feature contributions. *Knowl Inf Syst* 2014; **41**: 647–65.
203 Elshawi R, Al-Mallah MH, Sakr S. On the interpretability of machine learning-based model for predicting hypertension. *BMC Med Inform Decis Mak* 2019; **19**: 146.
204 Vandenbroucke JP, Broadbent A, Pearce N. Causality and causal inference in epidemiology: the need for a pluralistic approach. *Int J Epidemiol* 2016; **45**: 1776–86.
205 Snelson E, Ghahramani Z. Sparse Gaussian Processes using Pseudo-inputs. *Advances in Neural Information Processing Systems 18 (NIPS 2005)* 2005; : 1–8.
206 Burt DR, Rasmussen CE, van der Wilk M. Rates of convergence for sparse variational Gaussian process regression. In: 36th International Conference on Machine Learning, ICML 2019. 2019.

207 Bui TD, Yan J, Turner RE. A unifying framework for Gaussian process pseudo-point approximations using power expectation propagation. *Journal of Machine Learning Research* 2017; **18**.
208 Wilson AG, Hu Z, Salakhutdinov R, Xing EP. Deep kernel learning. In: Artificial intelligence and statistics. 2016: 370–8.
209 Gardner J, Pleiss G, Weinberger KQ, Bindel D, Wilson AG. Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. In: Advances in Neural Information Processing Systems. 2018: 7576–86.
210 Blundell C, Cornebise J, Kavukcuoglu K, Wierstra D. Weight uncertainty in neural networks. *32nd International Conference on Machine Learning* 2015; **2**: 1613–22.
211 Shridhar K, Laumann F, Liwicki M. A Comprehensive guide to Bayesian Convolutional Neural Network with Variational Inference. 2018.
212 Rosenblatt M. A CENTRAL LIMIT THEOREM AND A STRONG MIXING CONDITION. *Proceedings of the National Academy of Sciences* 1956; **42**. DOI:10.1073/pnas.42.1.43.
213 Yao J, Pan W, Ghosh S, Doshi-Velez F. Quality of Uncertainty Quantification for Bayesian Neural Network Inference. In: proceedings at the International Conference on Machine Learning: Workshop on Uncertainty & Robustness in Deep Learning. 2019. http://arxiv.org/abs/1906.09686.
214 Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in neural information processing systems. 2017: 6402–13.
215 Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *33rd International Conference on Machine Learning, ICML 2016* 2016; **3**: 1651–60.
216 Wang QA. Probability distribution and entropy as a measure of uncertainty. *J Phys A Math Theor* 2008; **41**. DOI:10.1088/1751-8113/41/6/065004.
217 Coughlin SS. Anxiety and Depression: Linkages with Viral Diseases. *Public Health Rev* 2012; **34**: 7.
218 Fernandez M, Colodro-Conde L, Hartvigsen J, *et al.* Chronic low back pain and the risk of depression or anxiety symptoms: insights from a longitudinal twin study. *Spine Journal* 2017; **17**. DOI:10.1016/j.spinee.2017.02.009.
219 Loshchilov I, Hutter F. Decoupled weight decay regularization. In: 7th International Conference on Learning Representations, ICLR 2019. 2019.
220 Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. In: Proceedings of the 22nd international conference on Machine learning. 2005: 625–32.