Article published online: 2022-12-23



# Multivariate Sequential Analytics for Cardiovascular Disease Event Prediction

William Hsu[1] Jim Warren[1] Patricia Riddle[1]

[1] School of Computer Science, University of Auckland, Auckland, New Zealand

Address for correspondence William Hsu, PhD, School of Computer Science, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand (e-mail: whsu014@aucklanduni.ac.nz).

Methods Inf Med 2022;61:e149–e171.

## Abstract

**Background** Automated clinical decision support for risk assessment is a powerful tool in combating cardiovascular disease (CVD), enabling targeted early intervention that could avoid issues of overtreatment or undertreatment. However, current CVD risk prediction models use observations at baseline without explicitly representing patient history as a time series.

**Objective** The aim of this study is to examine whether by explicitly modelling the temporal dimension of patient history event prediction may be improved.

**Methods** This study investigates methods for multivariate sequential modelling with a particular emphasis on long short-term memory (LSTM) recurrent neural networks. Data from a CVD decision support tool is linked to routinely collected national datasets including pharmaceutical dispensing, hospitalization, laboratory test results, and deaths. The study uses a 2-year observation and a 5-year prediction window. Selected methods are applied to the linked dataset. The experiments performed focus on CVD event prediction. CVD death or hospitalization in a 5-year interval was predicted for patients with history of lipid-lowering therapy.

**Results** The results of the experiments showed temporal models are valuable for CVD event prediction over a 5-year interval. This is especially the case for LSTM, which produced the best predictive performance among all models compared achieving AUROC of 0.801 and average precision of 0.425. The non-temporal model comparator ridge classifier (RC) trained using all quarterly data or by aggregating quarterly data (averaging time-varying features) was highly competitive achieving AUROC of 0.799 and average precision of 0.420 and AUROC of 0.800 and average precision of 0.421, respectively.

**Conclusion** This study provides evidence that the use of deep temporal models particularly LSTM in clinical decision support for chronic disease would be advantageous with LSTM significantly improving on commonly used regression models such as logistic regression and Cox proportional hazards on the task of CVD event prediction.

**Keywords**

- cardiovascular disease
- event prediction
- machine learning
- deep learning

## Introduction

A powerful tool in combating cardiovascular disease (CVD) is automated clinical decision support for risk assessment. This is particularly valuable in identifying at-risk patients for initiating risk communication and management. Numerous efforts have sought to advance CVD risk prediction to better identify and manage populations at risk. These include the Systematic

**received**
March 2, 2022
**accepted after revision**
August 25, 2022

DOI https://doi.org/10.1055/s-0042-1758687.
ISSN 0026-1270.



Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

COronary Risk Evaluation (SCORE),[1,2] the Pooled cohort equations,[3] and in New Zealand the PREDICT equations.[4] Recently, research has also advanced the prediction of long-term risk of recurrent CVD events as improvements in disease management have contributed to a growing number of patients with established CVD in the community.[5] Modern risk assessment tools use statistical methods to identify vulnerable patients and quantify their level of risk.[6] For patients who are identified as high risk, an array of interventions are available to reduce the level of risk as well as to prevent an acute CVD event. These include, adopting lifestyle changes (e.g., smoking cessation, regular exercise), pharmacological therapy, and closer monitoring (e.g., more frequent risk assessments).[6] A CVD event is the prediction outcome of paramount clinical interest due to its high cost to the health care systems (hospitalizations and rehabilitation), associated disability-adjusted life years burden, and patient mortality.[7] The ability to accurately predict CVD events within a population enables targeted early intervention that could avoid issues of overtreatment or undertreatment in the population.[4]

All current CVD prediction models use predictors at baseline. The central question that the current study seeks to investigate is whether by including an observation window leading up to the baseline, thus accounting for patient history, CVD risk prediction may be improved. Additionally, in this study, we focus on lipid management. TC/HDL (total cholesterol to high density lipoprotein ratio) is a known important CVD risk factor.[8–10] In New Zealand, clinical guidelines recommend patients assessed to have a 5-year CVD risk of 15% or more to use lipid-lowering pharmacotherapy to reduce risk of CVD event or death.[6] Further, despite the strong evidence of the benefits of preventive medicine, non-adherence to medication is a long-standing challenge in health care delivery and presents a significant obstacle to patients benefiting from treatment.[11] Both international and New Zealand studies have found long-term adherence to statin (a lipid-lowering drug) to be low.[12,13] In New Zealand, adherence to statin in secondary prevention has been found to be 69 and 76% in the first year and drops down to 66% in the third year. For primary prevention, adherence to statin was found to be 63% in the first year.[14,15] A U.S. study found non-adherence to statin to be as high as 56.0% for secondary prevention patients and 56.4% for primary prevention patients.[16] Similarly, a United Kingdom-based study found patterns of discontinuation of treatment for 41% of patients who are using statin as secondary prevention and 47% of patients who are using statin as primary prevention, although many of these patients restarted their treatment following discontinuation (75 and 72%, respectively).[17] The current study hypothesizes that by integrating the temporal dynamics of TC/HDL levels and adherence to lipid-lowering therapy, the prediction of CVD risk can be improved. This hypothesis informs our cohort selection criteria which is detailed in Section Cohort Selection.

In the domain of health care, over a period of years, aided by government efforts, there has been growing uptake of electronic health record (EHR) systems. In New Zealand, government initiatives in the 1990s supported development of health IT infrastructure, including creation of a national health index (NHI), providing the sector with a unique individual patient identifier; implementing a health information privacy code; and actively encouraging the private sector to develop and sell electronic services.[18] In the United States, in the wake of the Global Financial Crisis massive growth in EHR uptake was driven by the HITECH act.[19] Of particular interest to this study are EHRs that are routinely collected. These data are often the biproduct of health care services and, in socialized health care systems such as New Zealand's, tend to have a whole-of-population coverage. When linked across various datasets, they have a longitudinal structure, allowing treatment and disease trajectories (e.g., patient's physiological changes) to be examined over time.[20]

The present resurgence of deep learning in the machine learning community is chiefly facilitated by advances in computational power, specifically graphics processing units (GPUs) and the increasing availability of enormous datasets. Many of the notable breakthroughs in the application of deep learning are in the area of computer vision and natural language processing: image classification, object detection, machine translation, and natural language generation.[21,22] A shared feature of these tasks is the use of unstructured data (images or plain text) where deep learning models' capacity for representation learning is exploited. In the domain of health care, computer vision has achieved some of the most significant successes in the application of deep learning. Here, medical image analysis, often using convolutional neural networks, has achieved levels of performance on par or exceeding human experts on a range of complex diagnosis tasks.[23–27] However, the performance gain of deep learning methods against conventional machine learning methods on structured/tabulated data, the type of data that is ubiquitous in EHRs, is less certain.[28–30]

Deep learning/neural networks (NNs) overcome some of the limitations of regression-based models. Deep learning models can jointly exploit feature interactions and hierarchy.[31] Of specific interest to this study is the class of artificial NNs called recurrent neural networks (RNNs) which are temporal models that are explicitly multivariate and sequential. In the context of risk prediction in public health, RNNs afford the opportunity for patient history to be modeled in a temporal manner, in contrast to conventional risk modelling where risk assessment is based on patient data at a specific point in time. Here, the temporal dynamic relationships between risk factors is integrated into the risk assessment. A variant of RNNs called LSTM includes an internal cell state and gated units that regulate what is inputted, retained, and outputted from the cell. LSTM was developed to overcome the problem of long-range dependencies (remembering significant events from the distant past)[32] and has the capacity to reset its internal state[33] (forget unimportant events in the past). Since its development, LSTM-based methods have proven remarkably competitive on a range of tasks.[34–39] and have been successfully applied to a range of sequential tasks in the biomedical domain.[40–42] Given the long-term nature of CVD progression and CVD

management, this study hypothesizes that LSTM will be well suited for CVD event prediction, where an observation window of patient history is integrated into the prediction task.

Vascular informatics using epidemiology and the web (VIEW) is a vascular health research program based at University of Auckland; the program includes a research stream named PREDICT.[43] For the current study, the PREDICT dataset is linked to other routinely collected national datasets including pharmaceutical dispensing, hospitalization, laboratory test results, and deaths, to investigate methods for multivariate sequential modelling in the context of CVD risk prediction. From the data linkage, features that have clinical feasibility are derived. The study focuses on a cohort with lipid management.

## Objective

This study is motivated to investigate if risk prediction performance in CVD can be improved if temporal deep learning methods are utilized, specifically in a context where structured/tabulated data are used. The model long short-term memory (LSTM) appears to be an excellent fit to the problem of chronic disease risk prediction and thus is central in our investigation. LSTM allows patient history to be explicitly modeled in a multivariate and sequential fashion, where internal mechanisms of the unit control the content of its memory. As such, LSTM should be well suited for prediction tasks where the progression and management of a disease are prolonged and long term. Of particular interest to the current study is the relevance of the temporal dynamics of lipid management. We hypothesize that patient history over time in the 2 years run up to PREDICT assessment will be informative for CVD risk prediction.

Our study compares LSTM against several model comparators. The models are selected to assess: the consequence of explicitly modelling time through the use of sequential data, the usefulness of learning long-term dependencies and "forgetting" facilitated by the LSTM units, the advantages of modelling non-linear relationships in the predictor variables and the benefits of overcoming the problem of multicollinearity for the task of CVD event prediction against traditional risk models used in clinical decision support.

## Methods

### Data Sources

PREDICT is a web-based CVD risk assessment and management decision support system developed for primary care in New Zealand. The system is integrated with general practice EHR and since its deployment in 2002 has produced a constantly growing cohort of CVD risk profiles. Through the use of encrypted NHI, the de-identified cohort is annually linked to other routinely collected databases to produce a research cohort. The PREDICT cohort and its use in improving CVD risk assessment have been described in detail previously.[4,44]

The current study links the PREDICT cohort to TestSafe (Auckland regional laboratory test results[45]) and national collections by the Ministry of Health – the pharmaceutical collection, the National Minimum Dataset (hospital events), and the Mortality Collection.[46] TestSafe is used to obtain laboratory test results of clinically relevant measures (see next section). The Pharmaceutical collection is used to obtain dispensing history of medication relevant to the management of CVD including lipid-lowering, blood pressure lowering, antiplatelets, and anticoagulants as well as dispensings of drugs used in the management of important comorbidities, e.g., insulin. The National Minimum Dataset (NMDS) is used to identify hospitalization with their dates of admission and discharge and diagnosis. The mortality collection enables the identification of patients who died during the study period and their cause of death. From these sources, history of CVD, treatment trajectories, important comorbidities as well as CVD events can be derived.

A lookup table constructed by the VIEW research team is used to identify relevant chemical names from the Pharmaceutical collection. Identified chemical names using this lookup table are grouped into three broad categories: lipid-lowering, CVD, and other. Similarly, a lookup table constructed by the VIEW research team is used to identify ICD-10 codes in the hospitalization collection that are related to CVD conditions: more, specifically, International Statistical Classification of Diseases and Related Health Problems, Tenth Revision, Australian Modification, ICD-10-AM, which was used in New Zealand from 1999 to 2019.[47] The conditions are broadly in two categories: history and outcome, with the addition of mortality. For the list of the CVD conditions and their respective categories see ►**Appendix Table 1** in Appendix. For the definitions of listed conditions see https://wiki.auckland.ac.nz/display/VIEW/Complete+Variable+Names+Index.

### Laboratory Tests

Through TestSafe, records of high-density lipoproteins (HDL), low-density lipoproteins (LDL), triglycerides (TRI), total cholesterol (TCL), cholesterol ratio (TC/HDL), serum creatinine (sCr), and glycated hemoglobin (HbA1c) are obtained.[45] TC/HDL is the ratio of TCL divided by HDL. TCL is calculated by[48]

$$TCL = HDL + LDL + 0.2 \times TRI. \quad (1)$$

sCr is a measure used to determine the health of a patient's kidney. However, an individual's sCr level can vary depending on one's sex, age, ethnicity, and body size. A more precise measure for determining an individual's kidney health is the estimated glomerular filtration rate (eGFR)[49] which is estimated for every sCr laboratory test in the TestSafe record. HbA1c measures the glucose level in an individual's blood, it is used for diabetes diagnoses and to assess long-term glucose control for patients diagnosed with diabetes.[50] Patients with kidney disease or diabetes have significantly increased CVD risk.[6]

### TestSafe Feature Construction

The measures from TestSafe are irregularly sampled. For TC/HDL, some patients might have one test over the period of 2 years, while others might have three tests in one quarter. To construct time-series from TestSafe, values from tests are

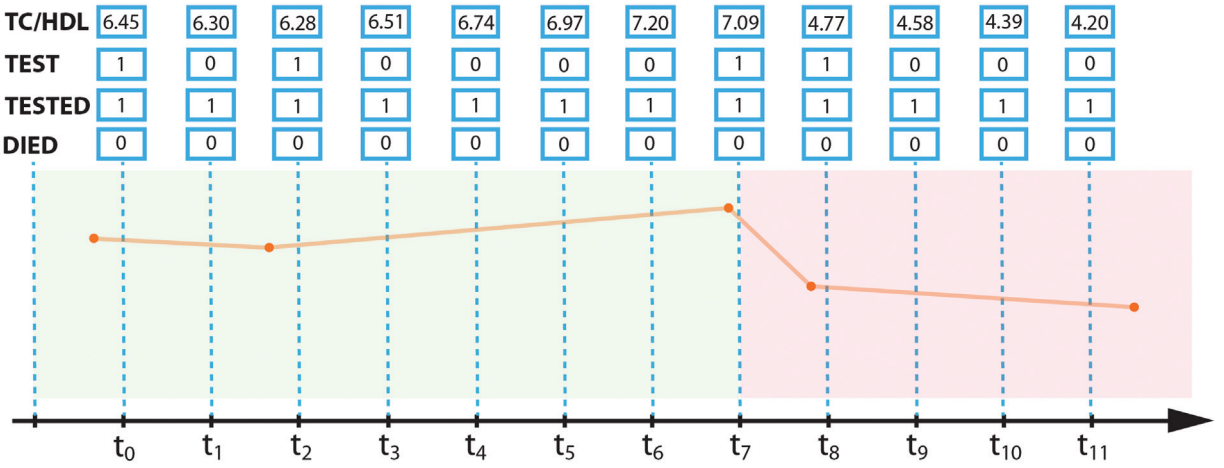


**Fig. 1** If TC/HDL test results outside the study window exist (before $t_0$ and after $t_{11}$), they are used in the interpolation. HDL, high-density lipoprotein; TC, total cholesterol.

linearly interpolated and extrapolated over the study period. The method connects a straight line between any two adjacent data points within the study window. If no feature value exists before the first and or after the last feature value, the first/last feature values are linearly extrapolated. Linear extrapolation uses the first/last value of a feature and sets all values of that feature before/after to that value. Laboratory tests generally occur intermittently within a patient's history, however, for intervals without a measure for Lipid, HbA1c, or GFR it does not mean these biometric measures cease to exist (drop to zero) in these intervals. Experiments were conducted exploring spline interpolation as a potential method for interpolating between feature values within a study window. However, the variability between when measures are taken meant spline interpolation could potentially introduce extreme values that are biologically implausible. It was decided that interpolating and extrapolating linearly offer the most parsimonious explanation of a patient's biometric trajectory without introducing extreme values. In addition, auxiliary features TEST, TESTED, and DIED are constructed, these are binary time-series indicating whether the patient had a cholesterol test in this quarter (encompassing HDL, LDL, TRI, TCL, and TC/HDL), whether the patient has ever had a cholesterol test and whether the patient has died, respectively. Using TC/HDL as an example, the rules used in constructing the cholesterol time-series are illustrated in ►Figs. 1 and 2.

►Figs. 1 and 2 show examples of TC/HDL, TEST, TESTED, and DIED time series. TC/HDL laboratory test results and their interpolated and extrapolated values are represented by orange dots and orange lines, respectively. TC/HDL values are point estimates, representing where the TC/HDL line intersects with the blue dotted line at $t_i$ . TEST, TESTED, and DIED are binary indicators. TEST values are evaluated over an interval, between $t_{i-1}$ (exclusive) and $t_i$ (inclusive). If the patient has had any cholesterol test within this interval, the value of TEST would be 1, otherwise 0. For simplicity, the above examples comprise only TC/HDL tests in the study window. TESTED indicates whether the patient has ever had a cholesterol test and DIED indicates whether the patient has died.

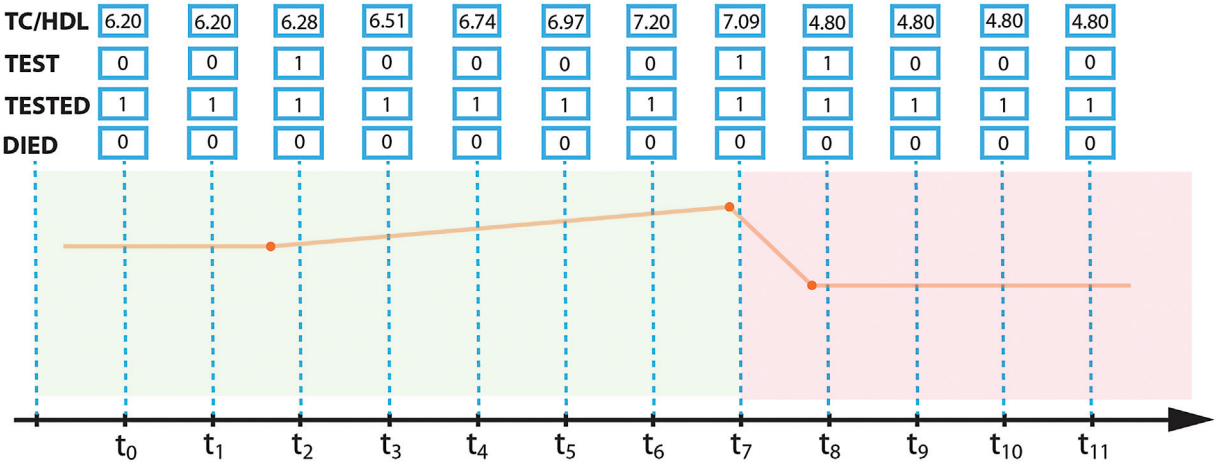


**Fig. 2** If TC/HDL laboratory test results outside the study window do not exist the TC/HDL values are extrapolated from the first test result leftward and from the last test result rightward. HDL, high-density lipoprotein; TC, total cholesterol.

Laboratory test results of eGFR and HbA1c are treated similarly in the construction of their respective time series. Additional auxiliary time series of TEST_GFR, TEST_HBA1C, TESTED_GFR, and TESTED_HBA1C are also included as time-series features.

### Pharmaceutical Dispense

A lookup table constructed by the VIEW research team is used to identify relevant categories of medications. The categories constructed by VIEW are lipid_lowering, statins, bp_lowering, antiplatelets, anticoagulants, antianginals, loop diuretics, anti-diabetes, insulin, metformin, other_oralhypos, metolazone, ppi_h2a, corticosteroid, and nonasp_nsaids. Identified chemical names using this look up table are grouped into three broad categories: lipid-lowering (comprised of lipid_lowering and statins medications), CVD (comprised of bp_lowering, antiplatelets, anticoagulants, antianginals, loopdiuretics and metolazone medications), and other (comprised of anti-diabetes, insulin, metformin, other_oralhypos, ppi_h2a, corticosteroid, nonasp_nsaids medications).

### Data Cleansing

We calculate the proportion of days covered (PDC) as a percentage for pharmaceutical features. To do so, the field DAYS_SUPPLY is used to infer the number of days covered by a specific medication. However, anomalous values need to be addressed before PDC can be calculated. For each drug, if DAYS_SUPPLY is 0 or outside a range specific for that drug (there were no missing values in this field), the value of DAYS_SUPPLY is inferred using the value of QUANTITY_DISPENSED divided by the value DAILY_DOSE if these values are available. Otherwise, the most frequently occurring QUANTITY_DISPENSED and/or DAILY_DOSE for that drug is used in the calculation. Following this inference, if the value of DAYS_SUPPLY is still outside the range for this drug we assign the most frequently occurring DAYS_SUPPLY value that is nonzero for this drug to DAYS_SUPPLY. With a few exceptions, all medications used the minimum of seven and maximum of 90 as the range for DAYS_SUPPLY.

Insulin treatment and usage pattern are not one where medication adherence can be reliably calculated from dispensing records through the variables available. In the vast majority of cases DAYS_SUPPLY is 0 and no sensible value could be derived from dividing QUANITY_DISPENSED by DAILY_DOSE, as DAILY_DOSE is not measured in pill counts but volume, e.g., mL. Additionally, insulins are covariates in our analysis, indicating the patient is managing the comorbidity of diabetes and an overall more complex health state. Therefore, it is important for the signal of insulin dispense to be kept in the data but it is not required for it to be of a value where patient adherence to insulin can be measured. All DAYS_SUPPLY of insulins are set to the most frequent nonzero QUANTITY_DISPENSED.

### Pharmaceutical Collection Feature Construction

A PDC time series for each chemical name is constructed. It is common for patients to switch treatments in the lipid-lowering category. To address this, an extra PDC time series bounded to 100, representing PDC for all lipid-lowering medication is added to the features.

Chemical names in the category of CVD and other are treated as covariates. For these chemical names, we constructed PDC time series for each name, where in the case of combined treatment we split the chemical name with the word "with" and construct a time series for each of the elements in the combined treatment.

### Hospitalization Discharge

The NMDS contains hospitalization records including variables DIAG_TYP (diagnosis type), ADM_TYP (admission type), EVSTDATE (event start date), EVENDATE (event end date), and CLIN_CD_10 (ICD-10 code). There are four relevant DIAG_TYPs in the record[51]:

A. Principal diagnosis.
B. Other relevant diagnosis.
O. Operation/procedure.
E. External cause of injury.

Each admission can have up to 99 diagnosis/procedure codes where there exists only one that is of DIAG_TYP A – principal diagnosis. With remaining codes categorized by the other DIAG_TYPs. A list of the retired and current ADM_TYPs exist in the dataset[51]:

CURRENT

AA Arranged admission
AC Acute admission
AP Elective admission of a privately funded patient
RL Psychiatric patient returned from leave of more than 10 days
WN Admitted from DHB booking system (used to be known as "waiting list")

RETIRED

ZA Arranged admission, ACC covered (retired June 30, 2004)
ZC Acute, ACC covered (retired June 30, 2004)
ZP Private, ACC covered (retired June 30, 2004)
ZW Waiting list, ACC covered (retired June 30, 2004)
WU Waiting list – urgent (code not used from August 20, 1993)

A lookup table constructed by the VIEW research team is used to identify ICD-10 codes in the NMDS that are related to CVD conditions of interest. The conditions are broadly divided in two categories: history and outcome.

### Hospitalization Discharge Feature Construction

Binary time series are constructed for all CVD conditions defined by the VIEW research team, including 21 CVD history, two CVD mortality and 18 CVD outcome categories. Patients' NMDS records prior to the observation window/study period are searched for evidence of CVD history. If there exists a clinical code mapping to any of the CVD history categories, the corresponding time series will contain 1s otherwise 0s.

All hospitalization records that fall within the study period are parsed. Any hospitalization record with a clinical code mapping to any CVD history categories will switch the time series for the categories from 0s to 1s from the time step the hospitalization event occurs and onward. Only clinical codes with DIAG_TYP A, O and E are used to identify CVD mortalities and outcomes. If there exists a clinical code with DIAG_TYP A, O or E mapping to one of the CVD mortality and/or outcome categories, the corresponding categories will be 1 in the time step(s) in which the record of the event falls.

In addition to the features constructed based on CVD conditions defined by VIEW two time series NUMBER_OF_-DAYS and ACUTE_ADM are constructed. NUMBER_OF_DAYS is of the number of days within this time step (quarter) the patient was in hospital. The equation

$$NUMBER\_OF\_DAYS = (EVENDATE - EVSTDATE) + 1 \quad (2)$$

is used to derive the value for the variable to account for day patients. ACUTE_ADM is a binary vector that has the value 1 if the event is an acute admission (holding the value of AC or ZC in ADM_TYP), otherwise 0.

## Study Design

To investigate whether patients' CVD event prediction may be improved by the inclusion of patient history a study design is formulated using each patient's PREDICT assessment as the index date, and approximately 2 years ($8 \times 90$ day quarters) prior to the index date and approximately 5 years ($20 \times 90$ day quarters) after the index date as the observation window and prediction window, respectively (►Fig. 3). An approximately 5 years interval for the prediction window is chosen because it aligns with Ministry of Health guidelines for CVD risk assessment and is underpinned by the fact that patients' CVD risk and risk management can change considerably over a longer period (e.g., 10 years), most randomized controlled trials of CVD medications are based on a period of 5 years or less and that practitioners are accustomed to this approach.[6] An approximately 2 year interval for the observation window is chosen in the interest of retaining enough samples in the dataset.

## Cohort Selection

The study cohort was selected through several exclusion criteria. First, patients having their first PREDICT assessment prior to January 01, 2007 and after December 30, 2013 are excluded as their pharmaceutical records are censored in the observation or prediction windows. Second, informed by our interest in integrating the temporal pattern of disease states, patients without all components of lipid profile (HDL, LDL, TRI, TCL, and TC/HDL) in either the observation or prediction windows are excluded. Third, informed by our interest in integrating the temporal pattern of disease management process, patients without lipid-lowering medication dispensed in the observation window with a 2 week look ahead post PREDICT assessment (to account for patients prescribed lipid-lowering medication around the time of PREDICT assessment) are excluded. Patients with infeasible data values and patients under the age of 18 are excluded. See ►Fig. 4 for the study cohort selection flowchart.

## Preprocessing

This subsection outlines the actions taken during preprocessing to address categorical variables, missing values as well as data imbalance and removing erroneous data. During preprocessing, four samples were removed from the data because the value of the variable PT_DIABETES_YR was $<0$. If a sample's PT_DBP2 value is missing the PT_DBP value is assigned to the PT_DBP2 variable (seven samples). PREDICT variables PT_RENAL which is ordinal and PT_ATRIAL_FIBRILLATION which is binary with missing values have 0 assigned to the missing values and all other values changed to the value $+1$. Missing PT_DIABETES_YR is assigned 0 (65,084 samples). Missing PT_EN_TCHDL is assigned the last TC/HDL result before PREDICT assessment from TestSafe (889 samples). SEX is encoded as a binary variable and ETHNICITY is one-hot encoded. Ethnicities MELAA (Middle Eastern, Latin American and African; comprise only 1.5% of the New Zealand population[52]) and Other are excluded due to small sample size. Ethnicities Chinese and Other Asian are combined. This resulted in five ethnicity groups: European, Māori, Pacific, Chinese/Other Asian, and Indian. Samples missing PT_SMOKING (two samples) and PT_GEN_LIPID (one sample) are removed.

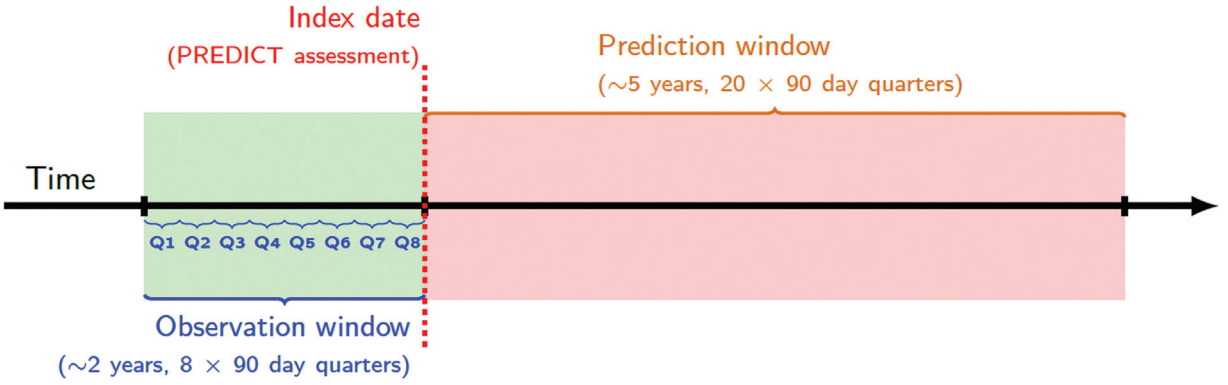


**Fig. 3** Study design showing date range from index date for the observation window (*shaded in green*) and the prediction window (*shaded in red*).

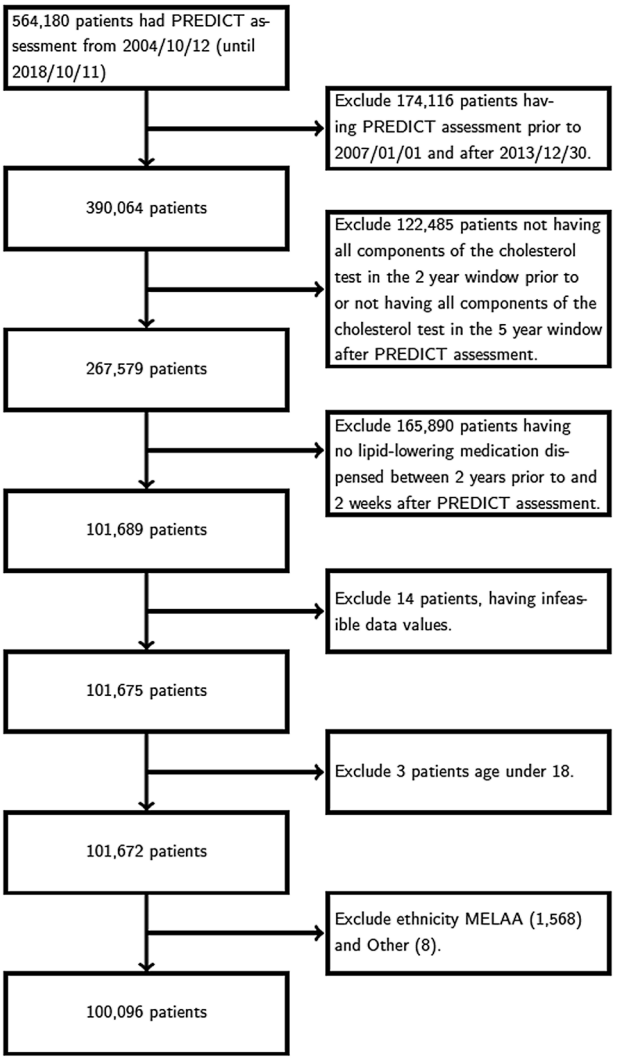


**Fig. 4** Flowchart of study cohort selection.

The above steps leaves 100,096 samples in the data. These samples are randomly shuffled, then a test set of the last 10,096 samples is set aside. Using data not in the test set, linear regression models were developed to impute missing HBA1C and eGFR values in the entire dataset using AGE, SEX, NZDEP, and ETHNICITY as predictor variables. See ►**Appendix Table 2** in Appendix for the list of PREDICT variables and their descriptions. See ►**Appendix Table 3** in Appendix for the affected variables, their conditions that require addressing, the actions taken, and the number of affected cases.

## Descriptive Statistics

Based on the study design outlined in the Study Design section and the result of the cohort selection outlined in the Cohort Selection section, quarterly time series based on 90 day quarters are constructed for each patient in the cohort using the linked data outlined in the Data Sources section. The features of the data fall into eight categories: demographic, lipid profile, lipid-lowering drugs, CVD drugs, other drugs, hospitalization, HbA1c and eGFR, and PREDICT (i.e., other clinical variables such as systolic blood pressure, diastolic blood pressure, smoking status collected at the same time of CVD risk assessment). See ►**Appendix Tables 4** to **8** in Appendix for the features' descriptive statistics. Due to commercial sensitivity of pharmaceutical data, the descriptive statistics of lipid-lowering drugs, CVD drugs, and other drugs are not shown.

## Test Data

An attribute of time series constructed through interpolation is that the gradient of slopes afford the chance for data in the observation window to peek ahead into data in the prediction window. Obviously, this is strictly illegal in the task of forecasting or prediction, because what the experiments are seeking to quantify is how well the models can perform on these tasks using only data up to the index date, hence peeking ahead constitutes cheating. To avoid this problem, separate test data are created that extrapolates from the last test value in the observation window to the end of the observation window for all the interpolated features (TestSafe tests: HDL, LDL, TRI, TCL, TC/HDL, HbA1c, and eGFR). See ►**Fig. 5** for an illustration of this treatment. In all experiments, the TestSafe features used for training are the unaltered interpolated time series, while the separate extrapolated test data are used for testing to ensure no peeking ahead occurs during testing.

## Prediction Outcome

The problem of CVD event prediction is formulated as a binary classification task; predicting event and no event. In the

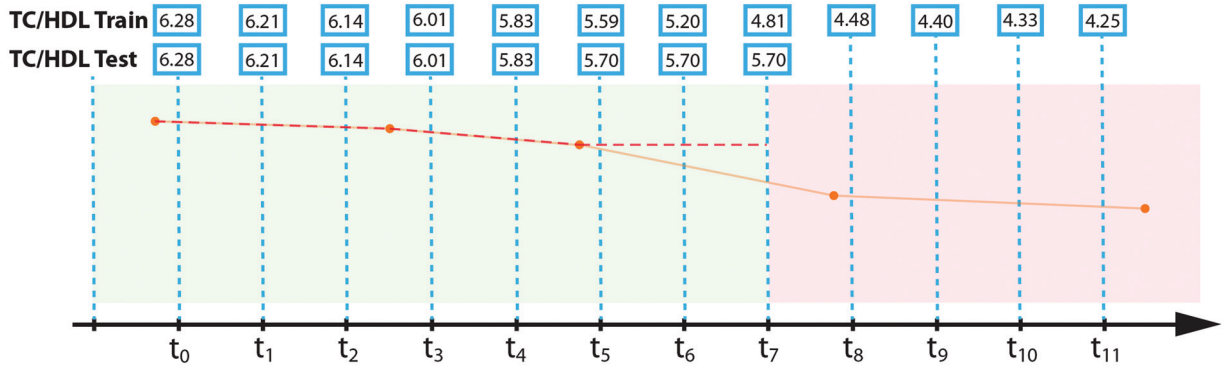


**Fig. 5** Test data are flattened beyond the last laboratory test result in the observation window to prevent looking ahead; laboratory test results beyond the observation window influencing the gradient within the observation window. Here, the dots are the laboratory test measures, the solid line is the constructed time-series and the dashed line represents the test data.

context of this study, the outcome of a CVD event (fatal or non-fatal) is defined as having an acute hospital admission with the ICD-10-AM code of the principal diagnosis matching one of the CVD mortality or outcome categories defined by VIEW (excluding atrial fibrillation, the feature OUT_ATRIAL_FIBRILLATION), or a CVD-related death without hospitalization. See ►**Appendix Table 1** in Appendix for the set of CVD categories. A PREDICT variable (PT_IMP_FATAL_CVD) is used to identify all patients who died due to CVD. This feature captures those who have CVD as a cause of death on their death certificate with or without hospitalization, as well as those without CVD recorded on their death certificate but who had a CVD hospital admission up to 28 days before their date of death. The VIEW research group refers to this as "the 28 day rule" for reclassifying non-CVD death as CVD death.[53]

Of the 100,096 patients, 25,419 patients have prior history of CVD, defined as having a hospital admission prior to their PREDICT assessment date with an ICD-10-AM code matching the "broad CVD history" category (HX_BROAD_CVD) defined by VIEW. The remaining 74,677 patients are patients without prior CVD. The proportions of each sub cohort (with or without prior CVD) having a CVD event and a fatal CVD event in their prediction window are shown in ►**Table 1**.

## Prediction Models

This study investigates the performance of LSTM against five model comparators on the task of CVD event (fatal or non-fatal) prediction. These model comparators are: simple recurrent neural network (Simple RNN), multilayer perceptron (MLP), ridge classifier (RC), logistic regression (LR), and Cox proportional hazards model (Cox).

Conventionally, with the exception of the output layer, MLP layers incorporate a non-linear activation, common among which are sigmoid, tanh, or the more recently developed rectified linear unit. It is the non-linear activation that provides the expressive power of MLP. Even with only a single hidden layer, an MLP can be universal (represent arbitrary functions) under certain technical conditions.[54] Increasing the depth of the network allows the network to represent complex functions more compactly. The hidden layer(s) of MLP can be thought of as learning nonlinear feature mapping, transforming a nonlinearly separable representation of the features to one that is linearly separable.[54,55]

Ridge regression and its classification variant RC are linear models that address the problem of multicollinearity in the predictor variables.[56] The models are part of a family of penalized regression models including Lasso[57] and Elastic Net[58] that adds a penalty to the loss. This penalty constrains and shrinks the size of the model coefficients, which has a regularization effect and prevents overfitting. For classification problems, RC first modifies binary response to −1 and 1 and then treats the task as a regression task, minimizing the penalized residual sum of squares. The sign of the regressor's prediction then represents the predicted class.[59] Ridge regression/classification has shown to be a promising modelling technique in the domain of epidemiology, particularly in high dimensional settings where the number of features is large, such as in genomic data analysis.[60,61] As a comparatively more interpretable model, it has shown to be competitive against black-box models such as support vector machines and NN.[62]

LR is a statistical method for modelling the relationship between one or more predictor variables and a dichotomous response variable of the values 1 or 0. It is a function of the odds ratio, and it models the proportion of new incidents developed within a given period of time. Cox is a statistical method for modelling the relationship between one or more predictor variables and the amount of time to pass before an occurrence of an event. It differs from LR by assessing a rate instead of a proportion. Cox regression is a function of the relative risk and it models the hazard rate, the number of new incidents per population per unit time. Although penalized LR and regularized Cox variations exist, here we are interested in the utility of LR and Cox as widely used in traditional clinical risk models[4,63,64]–i.e., without regularization – in the context of CVD event prediction. Their inclusion in the investigation provides baselines for the prediction task. The performance benefits of adding a penalty to linear models is represented in our investigation of RC.

The input datasets for LSTM and Simple RNN are explicitly sequential. The input datasets for MLP, RC, and LR are flattened across the time step dimension and concatenated. To examine the effect of multicollinearity as well as the effect of using history on RC and LR, two other input datasets are constructed. First, instead of concatenating the features across multiple time steps, an input dataset is constructed that uses the values of the last time step in the observation window (quarter 8) for features that are invariable across time (i.e., SEX, ETHNICITY, NZDEP) and the mean value of features that are variable across time (i.e., TC/HDL, LL_SIMVASTATIN, HX_BROAD_CVD). Here, an exception is AGE where the value at the 8th quarter is used. This dataset is from here on referred to as *aggregated*. Second, an input dataset is constructed using only the values of the last quarter in the observation window. This dataset is from here on referred to as *last quarter*. Due to the effect of multicollinearity only the *aggregated* and *last quarter* datasets are used to evaluate Cox.

**Table 1** Number of patients in the cohort with and without prior CVD and proportions of each respective subcohort that had a CVD event and a fatal CVD event in their prediction window

| | CVD event | Fatal CVD |
|---|---|---|
| Patients with prior CVD: 25,419 | 7,242 (approximately 28%) | 2,116 (approximately 8%) |
| Patients with no prior CVD 74,677 | 4,989 (approximately 7%) | 882 (approximately 1%) |

Abbreviation: CVD, cardiovascular disease.

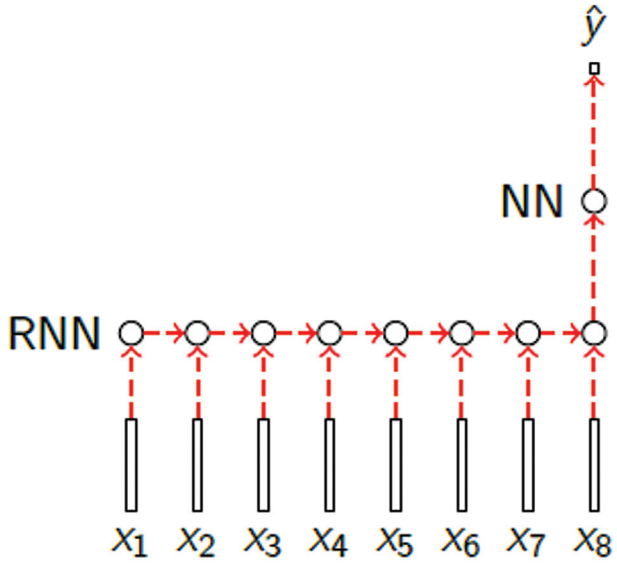


**Fig. 6** An unrolled view of RNN across the time-step dimension. Here, RNN can be a layer of Simple RNN or LSTM. NN is a layer of densely connected NN with softmax activation. $X_n$ are the inputs across $n$ timesteps. $\hat{y}$ is the output. LSTM, long short-term memory; NN, neural networks; RNN, recurrent neural network.

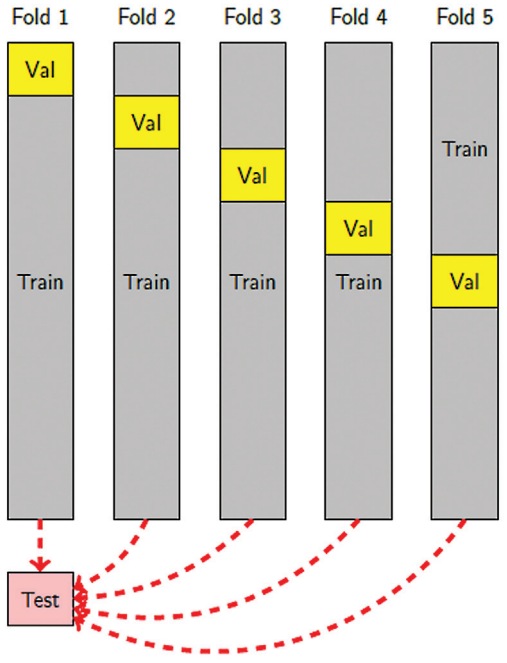


**Fig. 7** Illustration of the procedure used in splitting data into test, train, and validation sets across different folds.

All NN models used a two-unit densely connected layer with softmax activation as the output layer. The unrolled view across the time step dimension of the RNN models is shown in ►**Fig. 6**.

### Software Setup

Experiments are performed using Python 3.6.8,[65] with NN models using library Keras 2.2.4[66] with Tensorflow 1.13.1[67] backend and linear models RC and LR using library Scikit-learn 0.21.2.[59] Experiments also used R version 3.6.0, package pROC 1.16.2[68] for conducting DeLong's test and packages survival 3.2.7[69] and pec 2019.11.3[70] for Cox regression analysis. The package Autorank 1.1.1 is used for comparing models' performance as measured by average precision.[71]

### Procedures for Hyperparameter Search

This section outlines the procedures performed to search for the optimal set of hyperparameters for the LSTM, Simple RNN, and MLP models. From the entire dataset, 10,096 samples are set aside as the test set and removed from the search process. The remaining data (90,000 samples) are used in the search process. For each combination of hyperparameters, a five-fold cross validation is performed where while the proportion of data used for the train and validation sets are consistent, with 90% train (81,000 samples) and 10% validation (9,000 samples), different splits of train and validation sets are used in the experiments. See ►**Fig. 7** for a visual illustration of how the data are split into train and validation sets across the five-folds. In these experiments we use categorical cross-entropy as loss, where the validation loss is monitored and the lowest mean validation loss is used to determine the best set of hyperparameters.

For all experiments, the optimizer ADAM7[72] is used due to its capacity to adaptively adjust the learning rate during the training process and because its default hyperparameters have been shown to work on a range of problems. The ADAM optimizer is used with the default hyperparameter values outlined in the original paper.[72] These hyperparameter values are, learning rate $\alpha = 0.001$, the exponential decay rate for the first moment estimate $\beta_1 = 0.9$, the exponential decay rate for the second moment estimate $\beta_2 = 0.999$ and the small constant for numeric stability $\hat{\epsilon} = 1e-7$.[66]

See ►**Table 2** for the found optimal hyperparameters of the NN models.

For RC, hyperparameter search for the *L2* regularisation parameter and assessment of model performance on the validation set is done at the same time using the data split shown in Fold 1 in ►**Fig. 7**. Here, the values $1e^{-6}$, $1e^{-5}$, $1e^{-4}$, $1e^{-3}$, $1e^{-2}$, 0.1, 1 and 10.0 are searched. The found optimal *L2* values and their respective accuracy on the validation set are shown in ►**Table 3**, where the value of *L2* is estimated using the training samples, the accuracies reported are calculated using the validation set.

### Multicollinearity and Cox

When fitting the Cox model, several features returned a coefficient of NA: "unknown." These features were removed from the analysis to ensure predictions from the model could be made. For Cox (aggregated) seven features were removed. For Cox (last quarter) nine features were removed. See ►**Appendix Table 9** in Appendix for the removed features.

**Table 2** NN model hyperparameters for the CVD event prediction experiment

| Models | Hyperparameters |
|---|---|
| LSTM | Layers: 1 LSTM and 1 Dense<br>Units: 32 (LSTM) and 2 (Dense)<br>Batch size: 16,384<br>*L2*: $6.422e^{-2}$<br>Loss: categorical cross-entropy<br>Epochs: 200 |
| Simple RNN | Layers: 1 Simple RNN and 1 Dense<br>Units: 4 (Simple RNN) and 2 (Dense)<br>Batch size: 8,192<br>*L2*: $1.318e^{-1}$<br>Loss: categorical cross-entropy<br>Epochs: 200 |
| MLP | Layers: 3 Dense and 2 Dropout<br>Units; 32, 32, 2<br>Batch size: 64<br>Dropout rate: Layer 1 $2.500e^{-1}$<br>Layer 2 $2.500e^{-1}$<br>Loss: categorical cross-entropy<br>Epochs: 50 |

Abbreviations: CVD, cardiovascular disease; LSTM, long short-term memory; MLP, multilayer perceptron; RNN, recurrent neural network.

**Table 3** Optimal *L2* values found for ridge classifiers for CVD event prediction and their respective accuracy on the validation set

| | *L2* | Accuracy |
|---|---|---|
| RC | 1.0 | 0.886 |
| RC (aggregated) | 0.1 | 0.889 |
| RC (last quarter) | 0.1 | 0.887 |

Abbreviations: CVD, cardiovascular disease; RC, ridge classifier.

### Assess Model Performance

Once the optimal hyperparameters for each NN model have been found, the models are trained using the found hyperparameters with the data split shown in Fold 1 in ►Fig. 7. The Test set that is held aside is then used to assess model performance. To ensure fairness, all linear models RC, LR, and Cox are trained using the same training samples in Fold 1 and use the same test samples to measure model performance. For LR and Cox, the samples from the validation set are simply set aside in the process of model fitting and assessing model performance.

Taking into consideration the skewness of the classes (i.e., having a CVD event in the prediction window is much less frequent than not having one), a further set of experiments are conducted to address class imbalance. For the NN models, sample weighting that balances the two classes by making each sample inversely proportional to their class frequency in the training set is utilized. Sample weighting scales the loss function during training; here the less frequent class samples are given more weight thus contributing to greater loss.[73,74] The same weighting is applied to the classes for the RC and LR models.

The analysis uses AUROC and average precision as metrics for assessing model performance. Average precision is a summary statistic for a precision–recall (PR) curve. PR curves can provide more discerning information when the dataset is highly imbalanced.[75,76] Recall (the *x*-axis of the PR curve) is defined as

$$Recall = \frac{True\ positives}{True\ positives + False\ negatives} \quad (3)$$

and precision (the *y*-axis of PR curve) is defined as

$$Precision = \frac{True\ positives}{True\ positives + False\ positives} \quad (4)$$

In PR space, the position of $(1, 1)$ represents perfect discrimination – as opposed to $(0, 1)$ in ROC space, where closer the curve is to this point the better the discriminatory power of the model. A horizontal line at $\frac{P}{N+P}$, where $P$ and $N$ are the positive class and negative class frequencies, represent a no-skill classifier. The no-skill classifier is equivalent to the classifier always predicting the minority class.[77] Average precision is a more conservative measure than calculating the AUC with the trapezoidal rule. The average precision is formally defined as

$$Average\ precision = \sum_n (R_n - R_{n-1}) P_n \quad (5)$$

here, $P_n$ and $R_n$ are precision and recall at the $n$th threshold.[59] Our experiments use a large number of thresholds (equal to the size of the test set), so the difference is likely to be small.

DeLong's test is used to statistically compare the resulting AUROC of each model's predictions. Currently, there is no known significance test for comparing two PR curves.[78,79] To compare the performance of models in PR space, the evaluation utilizes bootstrapping to sample $100 \times 10{,}000$ dependent samples of models' predictions. From using 100 equal splits of the sampled predictions, 100 average precision scores are calculated for each model. The resulting average precision scores are evaluated using the Autorank package.[71] The Autorank package is built for conducting statistical comparison between (multiple) paired populations. The package uses the guidelines described in Demšar[80] to first assess data normality and homoscedasticity before selecting the appropriate statistical test for comparison.

Finally, to ascertain that the improvement in predictive performance as the result of integrating patient history, an ablation study using one-quarter and four-quarters observation windows is conducted using LSTM. The resulting two models' predictive performance are then compared with the LSTM trained on eight quarters of observation window.

## Results

The results of the models' performance on the test set are shown in ►Table 4. The best performing models' ROC curves and PR curves (with or without sample/class weighting) are shown in ►Figs. 8 and 9. In ►Fig. 10 details of the PR curves are shown (with the same mapping of line colors to classifiers as in ►Figs. 8 and 9).

**Table 4** Model performance on CVD event prediction

| Model | Without weighting | | With weighting | |
|---|---|---|---|---|
| | AUROC | Average precision | AUROC | Average precision |
| LSTM | 0.801 | 0.425[a] | 0.800 | 0.423 |
| Simple RNN | 0.798 | 0.402 | 0.795 | 0.418[a] |
| MLP | 0.797 | 0.415[a] | 0.798 | 0.414 |
| RC | 0.799 | 0.420[a] | 0.798 | 0.409 |
| RC (aggregated) | 0.800 | 0.421[a] | 0.798 | 0.410 |
| RC (last quarter) | 0.794 | 0.417[a] | 0.794 | 0.400 |
| LR | 0.798 | 0.411[a] | 0.798 | 0.409 |
| LR (aggregated) | 0.801 | 0.421 | 0.802 | 0.421[a] |
| LR (last quarter) | 0.797 | 0.414[a] | 0.798 | 0.413 |
| Cox (aggregated) | 0.798 | 0.417 | – | – |
| Cox (last quarter) | 0.793 | 0.411 | – | – |

Abbreviations: AUROC, area under the receiver operating characteristic; CVD, cardiovascular disease; LR, logistic regression; LSTM, long short-term memory; MLP, multilayer perceptron; RC, ridge classifier; RNN, recurrent neural network.
[a]The best performing average precision of the model.

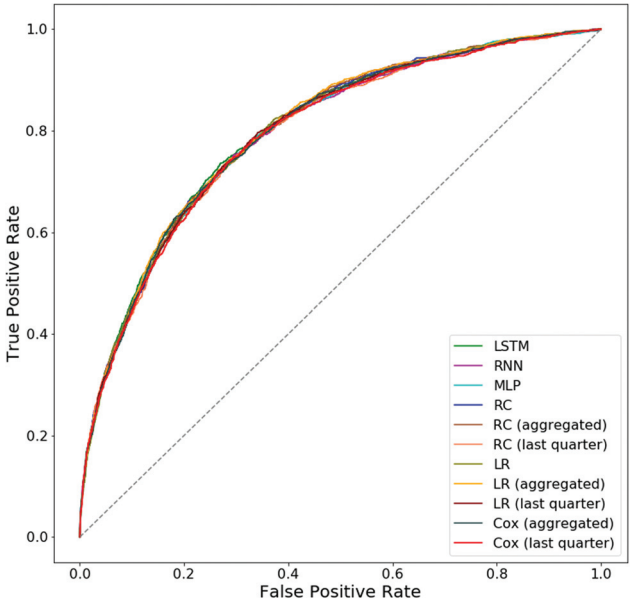


**Fig. 8** ROC curves of CVD event prediction. CVD, cardiovascular disease; ROC, receiver operating characteristic.

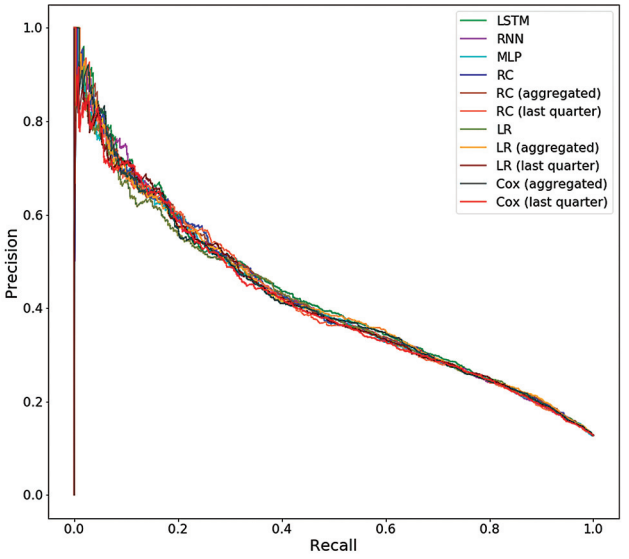


**Fig. 9** PR curves of CVD event prediction. CVD, cardiovascular disease; PR, precision recall.

The significance level of 0.05 is used for the comparison of models' AUROC. See ►**Table 5** for the results of DeLong's tests. The same significance level is used for the comparison of bootstrapped average precision scores using Autorank. The internal evaluation using Shapiro-Wilk test and Bartlett's test showed the data from all models are normal and homoscedastic. For that reason, repeated measures ANOVA and Tukey's HSD test are used to determine if a significant difference of the mean exists between the models' average precision scores and which differences are of statistical significance. See ►**Fig. 11** for the mean and 95.0% confidence interval of the models' average precision scores. The result of the analysis shows that no significant differences were found within the groups: RC (aggregated), RC, LR (aggregated), and Simple RNN; RC, LR (aggregated), Simple RNN, and Cox (aggregated); LR (aggregated), Simple RNN, Cox (aggregated), and RC (last quarter); Simple RNN, Cox (aggregated), RC (last quarter), and MLP; Cox (aggregated), RC (last quarter), MLP, and LR (last quarter); LR (last quarter), LR, and Cox (last quarter). However, all other differences are of statistical significance.

Lastly, the results of the ablation study are shown in ►**Fig. 12**.

## Discussion

The results of the CVD event prediction experiment show using average precision, LSTM is the overall leader (0.425) in this prediction task with RC (aggregated) and LR (aggregated)

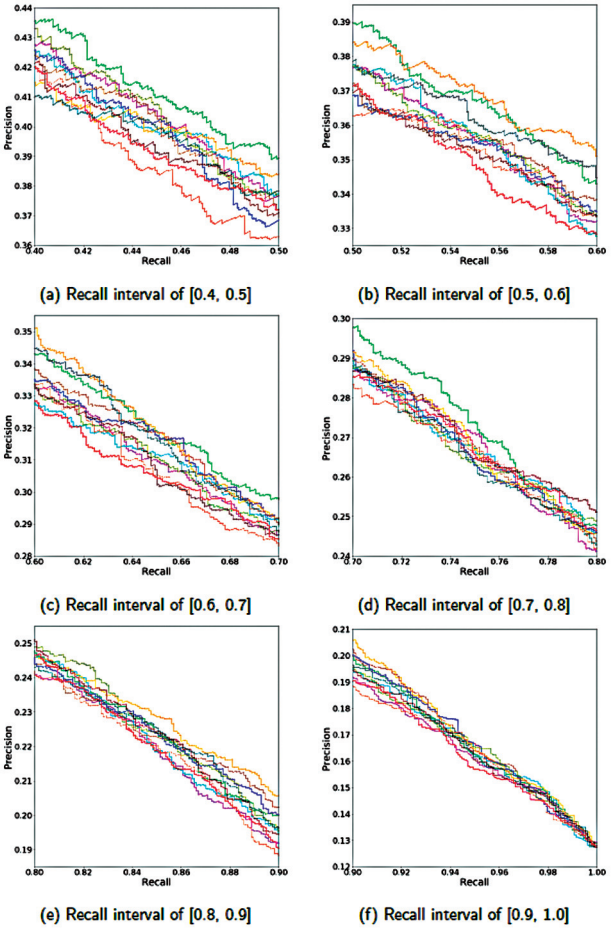


**Fig. 10** Detail plots of CVD event prediction PR curves, with the same mapping of line colors to classifiers as in Figs. 8 and 9. CVD, cardiovascular disease; PR, precision recall.

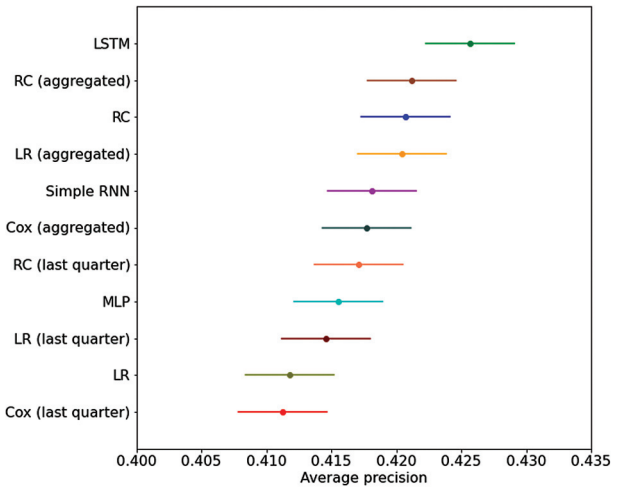


**Fig. 11** Statistical comparison of models' performances on the CVD event prediction task. The plot shows the average precision mean and 95.0% confidence intervals of the mean. Tukey's HSD test determined no significant differences exist within the groups: RC (aggregated), RC, LR (aggregated) and Simple RNN; RC, LR (aggregated), Simple RNN and Cox (aggregated); LR (aggregated), Simple RNN, Cox (aggregated) and RC (last quarter); Simple RNN, Cox (aggregated), RC (last quarter), and MLP; Cox (aggregated), RC (last quarter), MLP, and LR (last quarter); LR (last quarter), LR and Cox (last quarter). All other differences are found to be statistically significant. CVD, cardiovascular disease; HSD, honestly significant difference; LR, logistic regression; MLP, multilayer perceptron; PR, precision recall; RC, ridge classifier; RNN, recurrent neural network.

**Table 5** *p*-Values of pairwise comparison of AUROC using DeLong's test. The results are based on the best performing results of the models, where the models Simple RNN and LR (aggregated) are trained with sample/class weighting. Using significance level of 0.05, values under the Bonferroni adjusted significance level of 9.091e − 4 are highlighted

| | Simple RNN | MLP | RC | RC (aggr) | RC (last) | LR | LR (aggr) | LR (last) | Cox (aggr) | Cox (last) |
|---|---|---|---|---|---|---|---|---|---|---|
| LSTM | $1.429e^{-3}$ | $8.107e^{-2}$ | 0.2171 | 0.4420 | $3.262e^{-3}$ | 0.1638 | 0.5561 | $4.218e^{-2}$ | $5.551e^{-2}$ | $5.901e^{-4}$ |
| Simple RNN | | 0.4844 | 0.1420 | $6.848e^{-2}$ | 0.6908 | 0.2948 | $1{,}365e^{-3}$ | 0.5711 | 0.2678 | 0.4641 |
| MLP | | | 0.3744 | 0.2219 | 0.2235 | 0.7310 | $2.972e^{-2}$ | 0.8878 | 0.7782 | 0.1666 |
| RC | | | | 0.3885 | $1.602e^{-3}$ | 0.6324 | $8.400e^{-2}$ | 0.2690 | 0.6262 | $1.222e^{-2}$ |
| RC (aggr) | | | | | $3.631e^{-3}$ | 0.4285 | 0.1238 | 0.1698 | 0.2954 | $9.491e^{-3}$ |
| RC (last) | | | | | | 0.1166 | $1.054e^{-3}$ | $5.921e^{-2}$ | 0.1466 | 0.6304 |
| LR | | | | | | | $4.035e^{-2}$ | 0.5481 | 0.9589 | $4.179e^{-2}$ |
| LR (aggr) | | | | | | | | $4.848e^{-3}$ | $1.738e^{-4}$ | $9.985e^{-5}$ |
| LR (last) | | | | | | | | | 0.5733 | $1.269e^{-3}$ |
| Cox (aggr) | | | | | | | | | | $2.094e^{-2}$ |

Abbreviations: CVD, cardiovascular disease; LR, logistic regression; LSTM, long short-term memory; MLP, multilayer perceptron; PR, precision recall; RC, ridge classifier; RNN, recurrent neural network.

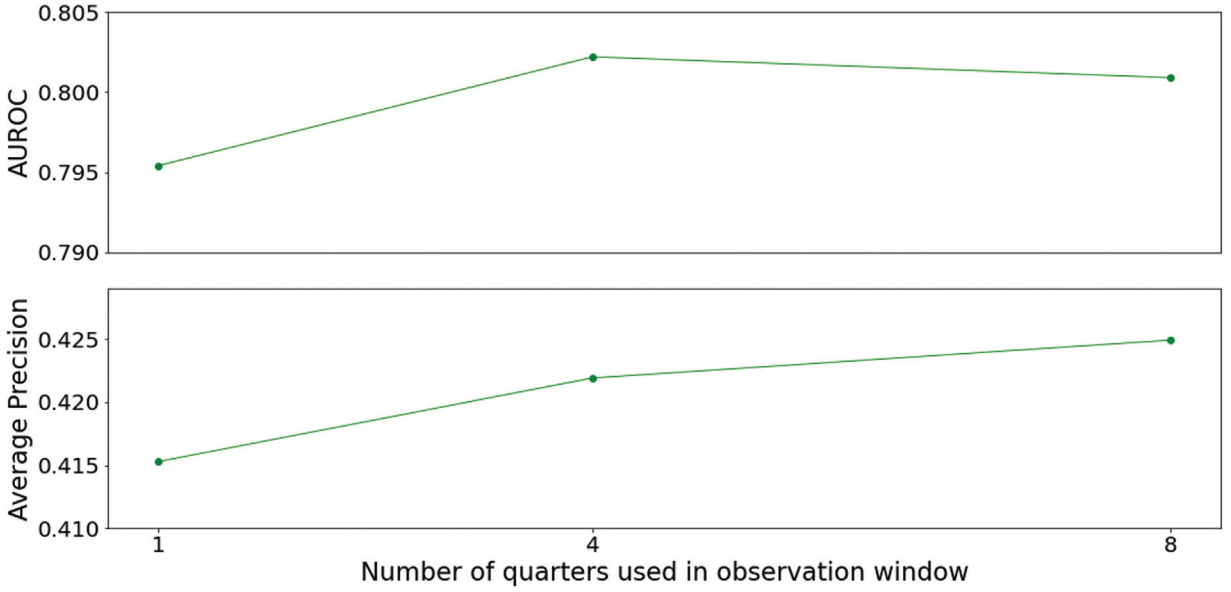


**Fig. 12** Results of ablation study.

ranked second equal (0.421). Our results confirm that PR curves provide further valuable information when the data are highly imbalanced. As an example, LR and Cox (aggregated) both achieved AUROC of 0.798. However, the same predictions achieved average precisions of 0.411 and 0.417, respectively (►**Table 4**), a substantial difference in PR space without any noticeable difference in ROC space. This discrepancy is further confirmed when visually assessing the ROC curves and PR curves plots (►**Figs. 8** and **9**). In ROC space, the curves of the models are densely packed together, virtually indistinguishable from one another, whereas there is a region in the PR space where the curves are noticeably more variable and spread out. The detail plots of the PR curves of recall in the interval of [0.4, 0.8] show there are regions where LSTM clearly dominates the other models. However, at the other end of the PR space where recall is in the interval of [0.8, 1.0], the results are much more mixed.

The statistical analysis using ANOVA and Tukey's HSD test comparing average precision scores of bootstrapped samples shows significant differences exist between groups, and that the LSTM model is determined to be significantly better than all other models at this prediction task. It appears that the capacity to retain and discard significant and unimportant events in the patient's past in addition to modelling patient history sequentially provides LSTM the predictive advantage, making it the best performing model, by a small margin, overall for this task.

The results also show that for this problem RC, RC (aggregated) and LR (aggregated) are highly competitive against the NN models. These models performed equally well as Simple RNN. Here, it can be observed that RC (aggregated) – the best performing regression-based model – achieved an average precision mean of 0.421 and 95.0% confidence interval of (0.418, 0.425) and Simple RNN achieved an average precision mean of 0.418 and 95.0% confidence interval of (0.415, 0.422). From the statistical analysis, both models are found to belong to the same group (as well as RC and LR [aggregated]) where there is no group differences that are determined to be significant. The statistical analysis also found no significant differences in the group containing MLP, Cox (aggregated), RC (last quarter), and LR (last quarter).

With the exception of LR, the worst performing linear models are the models using only features from the last quarter of the observation window. This indicates that for this task, patient history is important irrespective of whether it is explicitly sequential or in another representation. The method of aggregating data by taking the mean of features that vary across time in the observation window is the most effective treatment of data for the linear models with RC (aggregated), LR (aggregated), and Cox (aggregated) achieving the best results of each respective models. LR's relatively poor performance (i.e., of the model using 8 quarters of history) can be seen as the result of its incapacity to handle multicollinearity. The findings of this experiment suggest there are no or limited non-linear interactions between the features that the NN model could exploit.

In addition to the predictive advantage of LSTM, a surprising finding is the competitiveness of RC and RC (aggregated) in integrating patient history in a risk prediction task when using structured data. These models are by comparison much smaller than the NN models and require far less hyperparameter tuning. This result shows that the traditional regression-based approaches for risk modelling can be improved by moving toward approaches in this direction, by combining: (a) Integrating patient history by capturing more factors across more time steps instead of only using features from the last quarter before the index date; and (b) Fitting a model with regularization such as using RC so the fitted coefficients are apt to deal with multicollinearity.

Given the complexity of LSTM architecture a question regarding the results might be whether LSTM's predictive advantage is entirely due to model capacity rather than it

being a temporal model that is explicitly sequential. The LSTM model used in our experiment contained 27,714 trainable parameters. In contrast, the MLP had 47,522 trainable parameters. This shows that model capacity alone does not explain LSTM's performance. Additionally, the results of the ablation study show that by including patient history beyond just using patient data at the index date the model performance improved, while the slight dip in AUROC between using observation windows of 4 and 8 quarters (from 0.802 to 0.801) is unlikely to be significant. Further, the metric better suited for imbalanced classification – average precision – shows a monotonic increase in performance as the observation window lengthened.

Recent results in clinical risk prediction using sequential modelling typically focus on a short prediction horizon, e.g., next visit or 6 months.[81–83] In contrast, the current study adopted a 5-year prediction horizon used in an established clinical decision support system,[4] and leveraged routinely collected EHR from a diverse population level dataset to facilitate comparison. If LSTM is adopted as a model for assessing CVD risk, it will be applied at a large scale. PREDICT has been used >500k times in New Zealand. If a performance difference is statistically significant, then even if it is only moderately better, it is a meaningful difference because, at this scale, there would be many more cases where the clinician gets the right answer, instead of the wrong answer, from the model.

Two decades ago, there was a paradigm shift in CVD risk management in clinical practice from prevention based on managing individual risk factors (e.g., high blood pressure and high cholesterol) to one that is based on the combination of risk factors; a shift from focusing on relative risk to one that focuses on absolute risk.[84] Since then, many guidelines on CVD risk assessment have moved from using paper charts to computerized clinical decision support systems as the number of predictor variables have grown over the intervening years.[1–3,6,85–88] This trend is likely to continue as non-classical CVD risk factors such as socio-economic deprivation are found to be strongly associated with CVD risk.[1,4] Conventionally, Cox proportional hazard models are used for these clinical decision support systems. Recently, studies have focused on machine learning techniques to improve predictive performance.[89,90]

Like many other non-communicable diseases, the development, progression, and management of CVD are prolonged and long-term. This characteristic of the disease makes the ability to include in the analytics of CVD risk patient history in a multivariate and explicitly sequential manner a desideratum, so that the dynamic temporal interactions between the risk factors can be modeled. Until recently, sequentially modelling long-range dependency has remained computationally infeasible as shown in the case of the widely studied and used Hidden Markov Models.[91] This study demonstrates the suitability of using LSTM for sequentially modelling patient history on structured/tabulated data and a proof of concept that gains can be made using LSTM for predicting CVD event over a 5-year interval.

There are several limitations of the current study. "Long-term" in the context of CVD can mean decades. Researchers of CVD therapy have pointed to the knowledge gap that exists between the evidence from randomized clinical trails, typically only lasting a few years, and the effect of long-term medication treatment (it is common for therapy to continue for decades) in secondary prevention.[92] The study design was unable to capture the long-term (defined in the scale of decades) effect of disease progression and treatment trajectory. While preserving a useful number of cases, the data construction used in this study was only able to achieve a 7 year window to divide between observation and prediction. In the future, however, this will change as routinely collected EHRs lengthen year on year. Another limitation of the study is that LSTM like other NN models, is a class of black box models where the influence of and interactions between predictor variables cannot be readily explained. Considerable research has been performed investigating methods to interpret and explain neural models,[93,94] and some specifically for RNNs.[95,96] These methods are clearly worthy directions of future work as they hold the potential for aiding risk communication. Another possible future direction is to incorporate time information such as by using: a decay function, temporal encoding, or by combining a vector representation for time with model architecture in sequential modelling[83,97,98]; or to utilize an attention mechanism to boost model performance.[81–83,95] Lastly, the current study focused on event prediction not time-to-event estimation nor risk level prediction, which Cox proportionate hazards models facilitate. Determining if the results of the present study extend from event prediction to risk level and time-to-event estimation would be a valuable next step in making the case for widespread use of explicitly-temporal models in chronic disease decision support.

## Conclusion

The investigations performed in this study found that routinely collected health data can be leveraged to predict patients' risk of a CVD event (fatal or non-fatal). Moreover, it is observed that the LSTM model, outperformed linear additive models. For CVD event prediction, LSTM provided the best average precision, significantly outperforming all other models compared. The additive models RC (aggregated), RC and LR (aggregated) were found to be highly competitive, outperforming MLP and matching the performance of Simple RNN as measured by average precision. These results suggest for this prediction task, apart from LSTM, classical statistical models are equally performant as non-linear models. In our experiments, various inputs were examined for the linear models to quantify the potential for patient history to be used to improve their performance. These include using the full sets of features across the eight quarters of observation window, using aggregated features and using only the last quarter of the observation window. For all linear models, using aggregated data provided the best performance and RC (aggregated) was found to be the best performing linear model for the prediction task. Alongside the strength of LSTM, these findings regarding the

inputs of linear models further corroborate that history matters in the context of CVD event prediction. As routinely collected EHR continues growing, alleviating one of the primary obstacles in applying deep learning methods, this study provides incentive for LSTM to be further explored as an event prediction model in the management of CVD, where even a marginal gain can have substantial economic and social benefits.

### Conflict of Interest

None declared.

### Acknowledgment

This study is supported by the University of Auckland Doctoral Scholarship and in part by New Zealand Health Research Council program grant HRC16–609. The authors thank Kylie Chen for code checking the time series construction code and Mike Merry for facilitating network connection to the GPU machine during COVID lockdowns. Thanks to the members of the VIEW research team for their feedback on earlier drafts of the manuscript.

## References

1 Jørstad HT, Colkesen EB, Minneboo M, et al. The Systematic COronary Risk Evaluation (SCORE) in a large UK population: 10-year follow-up in the EPIC-Norfolk prospective population study. Eur J Prev Cardiol 2015;22(01):119–126
2 Conroy RM, Pyörälä K, Fitzgerald AP, et al; SCORE project group. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. Eur Heart J 2003;24(11):987–1003
3 Goff DC Jr, Lloyd-Jones DM, Bennett G, et al; American College of Cardiology/American Heart Association Task Force on Practice Guidelines. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American college of cardiology/American heart association task force on practice guidelines. Circulation 2014;129(25, suppl 2):S49–S73
4 Pylypchuk R, Wells S, Kerr A, et al. Cardiovascular disease risk prediction equations in 400 000 primary care patients in New Zealand: a derivation and validation study. Lancet 2018;391(10133):1897–1907
5 Poppe KK, Doughty RN, Wells S, et al. Developing and validating a cardiovascular risk score for patients in the community with prior cardiovascular disease. Heart 2020;106(07):506–511
6 Ministry of Health Cardiovascular Disease Risk Assessment and Management for Primary Care; 2018. Accessed July 3, 2020, at: https://www.health.govt.nz/system/files/documents/publications/cardiovascular-disease-risk-assessment-management-primary-care-feb18-v4_0.pdf
7 National Health Committee Strategic Overview: Cardiovascular Disease in New Zealand; 2013. Accessed December 3, 2020, at: https://www.moh.govt.nz/NoteBook/nbbooks.nsf/0/FAC55041FD6DBDADCC257F7F006CDC16/$file/strategic-ver-view-cardiovascular-disease-in-nz.pdf
8 Hero C, Svensson AM, Gidlund P, Gudbjörnsdottir S, Eliasson B, Eeg-Olofsson K. LDL cholesterol is not a good marker of cardiovascular risk in type 1 diabetes. Diabet Med 2016;33(03):316–323
9 Lemieux I, Lamarche B, Couillard C, et al. Total cholesterol/HDL cholesterol ratio vs LDL cholesterol/HDL cholesterol ratio as indices of ischemic heart disease risk in men: the Quebec Cardiovascular Study. Arch Intern Med 2001;161(22):2685–2692
10 Millán J, Pintó X, Muñoz A, et al. Lipoprotein ratios: physiological significance and clinical usefulness in cardiovascular prevention. Vasc Health Risk Manag 2009;5:757–765
11 Stewart RA, Kerr A. Non-adherence to medication and cardiovascular risk. N Z Med J 2011;124(1343):6–10
12 Brown MT, Bussell JK. Medication adherence: WHO cares? Mayo Clin Proc 2011;86(04):304–314
13 Mabotuwana T, Warren J, Harrison J, Kenealy T. What can primary care prescribing data tell us about individual adherence to long-term medication?-comparison to pharmacy dispensing data Pharmacoepidemiol Drug Saf 2009;18(10):956–964
14 Grey C, Jackson R, Wells S, et al. Maintenance of statin use over 3 years following acute coronary syndromes: a national data linkage study (ANZACS-QI-2). Heart 2014;100(10):770–774
15 Sigglekow F, Horsburgh S, Parkin L. Statin adherence is lower in primary than secondary prevention: a national follow-up study of new users. PLoS One 2020;15(11):e0242424
16 Ellis JJ, Erickson SR, Stevenson JG, Bernstein SJ, Stiles RA, Fendrick AM. Suboptimal statin adherence and discontinuation in primary and secondary prevention populations. J Gen Intern Med 2004;19(06):638–645
17 Vinogradova Y, Coupland C, Brindle P, Hippisley-Cox J. Discontinuation and restarting in patients on statin treatment: prospective open cohort study using a primary care database. BMJ 2016;353:i3305
18 Protti D, Bowden T. Electronic medical record adoption in New Zealand primary care physician offices. Commonw Fund 2010;96(1434):1–14
19 What is the HITECH act? HIPAA Journal. 2021 Accessed January 27, 2021, at: https://www.hipaajournal.com/what-is-the-hitech-act/
20 Jorm L. Routinely collected data as a strategic resource for research: priorities for methods and workforce. Public Heal Res Pract 2015 Sep 30;25(04):e2541540
21 Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. Nat Med 2019;25(01):24–29
22 Lundervold AS, Lundervold A. An overview of deep learning in medical imaging focusing on MRI. Z Med Phys 2019;29(02):102–127
23 Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542(7639):115–118
24 Haenssle HA, Fink C, Schneiderbauer R, et al; Reader study level-I and level-II Groups. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann Oncol 2018;29(08):1836–1842
25 Kooi T, Litjens G, van Ginneken B, et al. Large scale deep learning for computer aided detection of mammographic lesions. Med Image Anal 2017;35:303–312
26 De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med 2018;24(09):1342–1350
27 Liu Y, Gadepalli K, Norouzi M, et al. Detecting cancer metastases on gigapixel pathology images. 2017:1–13 Accessed February 22, 2022, at: http://arxiv.org/abs/1703.02442
28 Neural nets vs. regression models. Statistical Modeling, Causal Inference, and Social Science. 2019 Accessed January 10, 2021, at: https://statmodeling.stat.columbia.edu/2019/05/21/neural-nets-vs-statistical-models/
29 Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med 2018;1(March):18
30 Benedetto U, Sinha S, Lyon M, et al. Can machine learning improve mortality prediction following cardiac surgery? Eur J Cardiothorac Surg 2020;58(06):1130–1136
31 Cheng JZ, Ni D, Chou YH, et al. Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans. Sci Rep 2015;2016(06):1–13
32 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9(08):1735–1780

33 Gers FA, Schmidhuber J, Cummins F. Learning to forget: continual prediction with LSTM. Neural Comput 2000;12(10):2451–2471
34 Graves A, Jaitly N. Towards end-to-end speech recognition with recurrent neural networks. Paper presented at: 31st Int Conf Mach Learn ICML 2014. 2014;5:3771–3779
35 Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. Adv Neural Inf Process Syst 2014;4 (January):3104–3112
36 Graves A. Generating sequences with recurrent. Neural Netw 2013;•••:1–43 Accessed February 22, 2022, at: http://arxiv.-org/abs/1308.0850
37 Vinyals O, Toshev A, Bengio S, Erhan D. Show and tell: a neural image caption generator. Paper presented at: Proceedings of the IEEE conference on computer vision and pattern recognition; 2015:3156–3164
38 Venugopalan S, Rohrbach M, Darrell T, Donahue J, Saenko K, Mooney R. Sequence to Sequence – Video to Text. 2015 Accessed February 22, 2022, at: http://arxiv.org/abs/1505.00487
39 Ren M, Kiros R, Zemel RS. Image question answering: a visual semantic embedding model and a new dataset. 2015 Accessed February 22, 2022, at: http://arxiv.org/abs/1505.02074v1
40 Lipton ZC, Kale DC, Elkan C, Wetzel R. Learning to diagnose with LSTM recurrent. Neural Netw 2015;•••:1–18 Accessed February 22, 2022, at: http://arxiv.org/abs/1511.03677
41 Xu Y, Biswal S, Deshpande SR, Maher KO, Sun J. RAIM: recurrent attentive and intensive model of multimodal patient monitoring data. Paper presented at: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM 2018;18:2565–2573
42 Pham T, Tran T, Phung D, Venkatesh S. DeepCare: A Deep Dynamic Memory Model for Predictive Medicine. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). 2016; 9652 LNAI(i):30–41
43 VIEW research. The University of Auckland, Medical and Health Sciences. Accessed May 23, 2021, at: https://www.fmhs.auckland.ac.nz/en/soph/about/our-departments/epidemiology-and-biostatistics/research/view-study/research.html
44 Wells S, Riddell T, Kerr A, et al. Cohort Profile: the PREDICT cardiovascular disease cohort in New Zealand primary care (PREDICT-CVD 19). Int J Epidemiol 2017;46(01):22
45 Welcome to TestSafe. CareConnect. 2022 Accessed February 22, 2022, at: https://www.careconnect.co.nz/testsafe/
46 Collections. Ministry of Health, Manatū Hauora. 2019 Accessed May 30, 2021, at: https://www.health.govt.nz/nz-health-statistics/national-collections-and-surveys/collections
47 ICD-10-AM/ACHI/ACS Development. Ministry of Health, Manatū Hauora. 2021. Accessed August 29, 2021, at: https://www.health.govt.nz/nz-health-statistics/national-collections-and-surveys/collections
48 What your cholesterol levels mean. American Heart Assoication. 2017. Accessed March 3, 2020, at: https://www.heart.org/en/health-topics/cholesterol/about-cholesterol/what-your-cholesterol-levels-mean
49 Creatinine: what is it? National Kidney Foundation. 2019 Accessed March 3, 2020, at: https://www.kidney.org/atoz/content/what-creatinine
50 What is the HbA1c test? Health Navigator New Zealand. 2019 Accessed March 3, 2020, at: https://www.healthnavigator.org.-nz/health-a-z/h/hba1c-testing
51 National Minimum Dataset (Hospital Events) Data Dictionary version 7.9. 2018. Accessed November 23, 2022, at: https://www.health.govt.nz/system/files/documents/publications/nmds_data_dictionary_v7.9.pdf
52 Ethnic group summaries reveal New Zealand's multicultural make-up. Stats NZ. 2020. Accessed May 22, 2021, at: https://www.stats.govt.nz/news/ethnic-group-summaries-reveal-new-zealand-multicultural-make-up
53 Fatal and Non-fatal EVENTS. VIEW#Data Wikipage. 2018 Accessed December 3, 2020, at: https://wiki.auckland.ac.nz/display/VIEW/Fatal+and+Non-fatal+Events
54 Goodfellow I, Bengio Y, Courville A. Deep Learning. MIT Press; 2016. Accessed February 22, 2022, at: http://www.deeplearning-book.org
55 Grosse R. Lecture5: Multilayer Perceptrons. Intro to Neural Networks and Machine Learning. 2018. Accessed November 8, 2020, at: http://www.cs.toronto.edu/~rgrosse/courses/csc321_2018/readings/L05%20Multilayer%20Perceptrons.pdf
56 Hoerl AE, Kennard RW. Ridge regression: applications to non-orthogonal problems. Technometrics 1970;12(01):69–82
57 Tibshirani R. Regression shrinkage and selection via the Lasso. J R Stat Soc B 1996;58(01):267–288
58 Zou H, Hastie T. Regularization and Variable Selection Via the Elastic Net. JSTOR; 2005
59 Pedregosa F, Weiss R, Brucher M. Scikit-learn. Machine Learn Python 2011;12:2825–2830
60 Cule E, De Iorio M. Ridge regression in prediction problems: automatic choice of the ridge parameter. Genet Epidemiol 2013;37(07):704–714
61 De Vlaming R, Groenen PJF. The Current and Future Use of Ridge Regression for Prediction in Quantitative Genetics. BioMed Research International. 2015. Accessed February 22, 2022, at: https://www.hindawi.com/journals/bmri/2015/143712/
62 Niemann U, Boecking B, Brueggemann P, Mebus W, Mazurek B, Spiliopoulou M. Tinnitus-related distress after multimodal treatment can be characterized using a key subset of baseline variables. PLoS One 2020;15(01):e0228037
63 Smolin B, Levy Y, Sabbach-Cohen E, Levi L, Mashiach T. Predicting mortality of elderly patients acutely admitted to the Department of Internal Medicine. Int J Clin Pract 2015;69(04):501–508
64 Lanièce I, Couturier P, Dramé M, et al. Incidence and main factors associated with early unplanned hospital readmission among French medical inpatients aged 75 and over admitted through emergency units. Age Ageing 2008;37(04):416–422
65 Python language reference. Python Software Foundation. 2017. Accessed November 10, 2020, at: https://www.python.org/
66 Chollet F, et al. Keras: the Python deep learning API. 2015 Accessed November 23, 2022, at: https://keras.io
67 Abadi M, Agarwal A, Barham P, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. 2015 Accessed November 23, 2022, at: https://tensorflow.org
68 Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011;12(77):77
69 Therneau TM. A package for survival analysis in R. 2020. R package version 3.2–7, Accessed November 23, 2022, at: https://CRAN.R-project.org/package=survival
70 Mogensen UB, Ishwaran H, Gerds TA. Evaluating random forests for survival analysis using prediction error curves. J Stat Softw 2012;50(11):1–23
71 Herbold S. Autorank: a Python package for automated ranking of classifiers. J Open Source Softw 2020;5:2173
72 Kingma DP, Ba J. Adam: a method for stochastic optimization. 2014:1–15. Accessed November 23, 2022, at: https://arxiv.org/abs/1412.6980
73 The Sequential model API. Keras 2.0.6 Documentation. Accessed January 16, 2021, at: https://faroit.com/keras-docs/2.0.6/
74 sklearn.utils.class_weight.compute_class_weight. scikit learn. 2020. Accessed January 16, 2021, at: https://scikit-learn.org/stable/modules/generated/sklearn.utils.class_weight.compute_-class_weight.html
75 Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. J Big Data 2019;6(27):1–54
76 Davis J, Goadrich M. The relationship between precision -recall and ROC curves. Proc. 23 rd International Conference on Machine

Learning. 2006. Accessed February 22, 2022, at: https://www.biostat.wisc.edu/~page/rocpr.pdf
77 On ROC and precision-recall curves. Towards data science. Accessed December 20, 2020, at: https://towardsdatascience.-com/on-roc-and-precision-recall-curves-c23e9b63820c
78 Baldwin B. Comparing precision-recall curves the bayesian way? LingPipe Blog. 2010. Accessed December 21, 2020, at: https://lingpipe-blog.com/2010/01/29/comparing-precision-recall-curves-bayesian-way/
79 Statistical test for comparing precision-recall curves. Cross validated. 2020. Accessed December 21, 2020, at: https://stats.stackexchange.com/questions/499672/statistical-test-for-comparing-precision-recall-curves
80 Demšar J. Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 2006;7:1–30
81 Ma F, Chitta R, Zhou J, You Q, Sun T, Gao J. Dipole: Diagnosis Prediction in Healthcare via Attention-based Bidirectional Recurrent Neural Networks. Paper presented at: Proceedings of the 23 rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. August 13–17, 2017; Halifax, NS
82 Luo J, Ye M, Xiao C, Ma F. HiTANet: Hierarchical Time-Aware Attention Networks for Risk Prediction on Electronic Health Records. Paper presented at: Proceedings of the 26 th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. August 23–27, 2020; Virtual Conference
83 Pham TH, Yin C, Mehta L, Zhang X, Zhang P. Cardiac Complication Risk Profiling for Cancer Survivors via Multi-View Multi-Task Learning. Paper presented at: Proceedings of the 2021 IEEE International Conference on Data Mining (ICDM). December 07–11, 2021; Auckland, NZ
84 Jackson R. Guidelines on preventing cardiovascular disease in clinical practice. BMJ 2000;320(7236):659–661
85 ASCVD risk estimator. American College of Cardiology. 2018 Accessed January 22, 2021, at: https://tools.acc.org/ldl/ascvd_risk_estimator/index.html#!/calulate/estimator/estimator
86 Welcom to the QRISK 3–2018 risk calculator. ClinRisk. 2018. Accessed January 23, 2021, at: https://qrisk.org/three/index.php
87 Tunstall-Pedoe H. Cardiovascular risk and risk scores: ASSIGN, Framingham, QRISK and others: how to choose. Heart 2011;97(06):442–444
88 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(06):491–499
89 Alaa AM, Bolton T, Angelantonio ED, Rudd JHF, Van Der Schaar M. Cardiovascular Disease Risk Prediction Using Automated Machine Learning: A Prospective Study of 423, 604 UK Biobank Participants. PLOZ ONE; 2019:1–17
90 Chun M, Clarke R, Cairns BJ, et al; China Kadoorie Biobank Collaborative Group. Stroke risk prediction using machine learning: a prospective cohort study of 0.5 million Chinese adults. J Am Med Inform Assoc 2021;28(08):1719–1727
91 Lipton ZC, Berkowitz J, Elkan C. A critical review of recurrent neural networks for sequence learning. 2015 Accessed February 22, 2022, at: http://arxiv.org/abs/1506.00019
92 Rossello X, Pocock SJ, Julian DG. Long-term use of cardiovascular drugs challenges for research and for patient care. J Am Coll Cardiol 2015;66(11):1273–1285
93 Holzinger A, Biemann C, Pattichis CS, Kell DB. What do we need to build explainable AI systems for the medical domain? 2017 Accessed February 22, 2022, at: http://arxiv.org/abs/1712.09923
94 Goebel R, Chander A, Holzinger K, et al. Explainable AI : the new 42? Machine Learn Knowledge Extraction 2018;11015:295–303
95 Choi E, Bahadori MT, Kulas JA, Schuetz A, Stewart WF, Sun J. RETAIN: an interpretable predictive model for healthcare using reverse time attention mechanism. 2016 Accessed February 22, 2022, at: http://arxiv.org/abs/1608.05745
96 Ho LV, Aczon M, Ledbetter D, Wetzel R. Interpreting a recurrent neural network's predictions of ICU mortality risk. J Biomed Inform 2021;114:103672
97 Baytas IM, Xiao C, Zhang X, Wang F, Jain AK, Zhou J. Patient Subtyping via Time-Aware LSTM Networks. Paper presented at: Proceedings of the 23 rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. August 13–17, 2017; Halifax, NS
98 Kazemi SM, Goel R, Eghbali S, et al. Time2Vec: Learning a Vector Representation of Time. 2019 Accessed August 8, 2022, at: https://arxiv.org/abs/1907.05321

## Appendix A

**Appendix Table 1** VIEW CVD categories: CVD history, CVD mortality and CVD outcome, feature names under the categories and feature descriptions. Feature names prefixed with MORTALITY or OUT are used to identify outcome events (with the exception of OUT_ATRIAL_FIBRILLATION)

| VIEW CVD categories | | |
|---|---|---|
| **Category** | **Feature name** | **Description** |
| History | HX_BROAD_CVD | History of broad CVD |
| | HX_ATHERO_CVD | History of atherosclerotic CVD |
| | HX_CHD_DIAG | History of coronary heart disease (diagnoses) |
| | HX_ACS | History of acute coronary syndrome |
| | HX_MI | History of myocardial infarction |
| | HX_UNST_ANGINA | History of unstable angina |
| | HIST_ANGINA | History of angina |
| | HX_OTHER_CHD | History of other coronary disease |
| | HX_CHD_PROCS | History of coronary heart disease |
| | HX_PCI | History percutaneous coronary intervention |
| | HX_CABG | History of coronary artery bypass graft |
| | HX_OTHER_CHD_PROCS | History of other coronary procedure |
| | HX_PVD_DIAGS | History of peripheral vascular disease |
| | HX_PVD_PROCS | History of peripheral vascular procedure |
| | HX_HAEMORRHAGIC_STROKE | History of hemorrhagic stroke |
| | HX_CEVD | History of cerebral vascular disease |
| | HX_ISCHAEMIC_STROKE | History of ischemic stroke |
| | HX_TIA | History of transient ischemic attack |
| | HX_OTHER_CEVD | History of other cerebral vascular disease |
| | HX_HEART_FAILURE | History of heart failure |
| | HX_ATRIAL_FIBRILLATION | History of atrial fibrillation |
| Mortality | MORTALITY_BROAD_CVD_WITH_OTHER | Death involving broad CVD |
| | MORTALITY_OTHER_RELATED_CVD_DEATHS | Death involving other related CVD |
| Outcome | OUT_BROAD_CVD | Outcome of broad CVD |
| | OUT_ATHERO_CVD | Outcome of atherosclerotic CVD |
| | OUT_CHD | Outcome of coronary heart disease |
| | OUT_MI | Outcome of myocardial infarction |
| | OUT_ACS | Outcome of acute coronary syndrome |
| | OUT_UNST_ANGINA | Outcome of unstable angina |
| | OUT_ANGINA | Outcome of angina |
| | OUT_OTHER_CHD | Outcome of acute coronary syndrome |
| | OUT_PVD_DIAGS | Outcome of peripheral vascular disease |
| | OUT_PVD_PROCS | Outcome of peripheral vascular procedure |
| | OUT_PCI_CABG | Outcome of percutaneous coronary intervention |
| | OUT_HAEMORRHAGIC_STROKE | Outcome of hemorrhagic stroke |
| | OUT_CEVD | Outcome of cerebral vascular disease |
| | OUT_ISCHAEMIC_STROKE | Outcome of ischemic stroke |
| | OUT_TIA | Outcome of transient ischemic attack |
| | OUT_OTHER_CEVD | Outcome of other cerebral vascular disease |
| | OUT_HEART_FAILURE | Outcome of heart failure |
| | OUT_ATRIAL_FIBRILLATION | Outcome of atrial fibrillation |

**Appendix Table 2** PREDICT variables and their descriptions

| Variable name | Description |
|---|---|
| PT_SBP | Current systolic blood pressure (sitting) |
| PT_SBP2 | Previous systolic blood pressure (sitting) |
| PT_DBP | Current diastolic blood pressure (sitting) |
| PT_DBP2 | Previous diastolic blood pressure (sitting) |
| PT_SMOKING | Smoking history or current status |
| PT_EN_TCHDL | TC/HDL cholesterol result |
| PT_DIABETES | Diabetes status |
| PT_FAMILY_HISTORY | Family history of premature CVD |
| PT_GEN_LIPID<br>PT_RENAL | Diagnosed genetic lipid disorder<br>Renal disease status |
| PT_DIABETES_YR | Number of years since diabetes diagnosis |
| PT_ATRIAL_FIBRILLATION | ECG confirmed atrial fibrillation |
| PT_IMP_FATAL_CVD[a] | Improved fatal CVD using mortality record and 28 day rule |

Abbreviation: CVD, cardiovascular disease.
[a]This feature captures all patients with CVD as cause of death on their death certificate with or without hospitalization. In addition, those without CVD recorded on their death certificate but who had a CVD hospital admission up to 28 days before their date of death are included. The VIEW research group refers to this as "the 28 day rule" for reclassifying non-CVD death as CVD death.

**Appendix Table 3** Affected variables, their conditions that require addressing, the action taken, and the number of affected cases

| Variable | Condition | Action | Number of cases |
|---|---|---|---|
| PT_DIABETES_YR | $<0$ | Remove samples | 4 |
| PT_DBP2 | Missing | Assign PT_DBP value | 7 |
| PT_RENAL | Missing | Assign 0 to missing values and change all other values to value $+1$ | 65,086 |
| PT_ATRIAL_FIBRILLATION | Missing | Assign 0 to missing values and change all other values to value $+1$ | 22 |
| PT_DIABETES_YR | Missing | Assign 0 to missing values | 65,084 |
| PT_EN_TCHDL | Missing | Assign last TC/HDL result from TestSafe | 889 |
| SEX | String values | Encode as a binary variable | 100,096 |
| ETHNICITY (MELAA and Other) | Small sample size | Remove samples | MELAA (1568), Other (8) |
| ETHNICITY (Chinese and Other Asian) | Small sample size | Combined | Chinese (5,317)<br>Other Asian (3,655) |
| PT_SMOKING | Missing | Remove samples | 2 |
| PT_GEN_LIPID | Missing | Remove sample | 1 |
| ETHNICITY | String values | One-hot encoded | 100,096 |
| HBA1C | Missing | Impute using a linear model with AGE, SEX, NZDEP and ETHNICITY as predictor variables | 983 |
| EGFR | Missing | Impute using a linear model with AGE, SEX, NZDEP and ETHNICITY as predictor variables | 56 |

**Appendix Table 4** Descriptive statistics: demographic variables. Number of patients in each category

| **ID** | **100,096** | | NZDEP | |
|---|---|---|---|---|
| | | | 1 | 21,167 |
| Sex | | | 2 | 19,074 |
| Male | 56,557 (56.5%) | | 3 | 17,141 |
| Female | 43,539 (43.5%) | | 4 | 18,903 |
| Age (at index date) | | | 5 | 23,811 |
| Mean (SD) | 61.82 (11.29) | | | |
| 18–24 | 48 | | Ethnicity | |
| 25–34 | 691 | | European | 56,641 |
| 35–44 | 5,690 | | Māori | 9,977 |
| 45–54 | 20,380 | | Pacific | 14,878 |
| 55–64 | 32,885 | | Chinese/Other Asian | 8,971 |
| 65–74 | 28,261 | | Indian | 9,629 |
| 75–84 | 10,379 | | DIED (%) | 6,634 (6.6%) |
| 85+ | 1,762 | | | |

**Appendix Table 5** Descriptive statistics: cholesterols. TEST and TESTED are binary features and the statistics are the number of quarters in the entire dataset where the features contained a 1 and its relative percentage

| | |
|---|---|
| Test (%) | 885,936 (31.6%) |
| HDL mean (SD) | 1.28 (0.37) |
| LDL mean (SD) | 2.26 (0.96) |
| TRI mean (SD) | 1.74 (1.04) |
| TCL mean (SD) | 4.69 (1.13) |
| TC/HDL mean (SD) | 3.85 (1.15) |
| Tested (%) | 2,698,599 (96.3%) |

**Appendix Table 6** Descriptive statistics: hospitalization. Number of patients who had acute hospital admission within their time-series and number of patients who had hospitalizations with clinical code mapping to the specified category in their time-series

| | | | | |
|---|---|---|---|---|
| NUMBER_OF_DAYS> 0 mean (SD) | 6.37 (11.87) | | MORTALITY_BROAD_CVD _WITH_OTHER | 17,463 |
| ACUTE_ADM | 54,448 | | MORTALITY_OTHER _RELATED_CVD_DEATHS | 2,416 |
| HX_BROAD_CVD | 32,542 | | | |
| HX_ATHERO_CVD | 30,259 | | OUT_BROAD_CVD | 16,421 |
| HX_CHD_DIAGS | 23,207 | | OUT_ATHERO_CVD | 14,308 |
| HX_ACS | 16,777 | | OUT_CHD | 9,689 |
| HX_MI | 13,799 | | OUT_MI | 5,944 |
| HX_UNST_ANGINA | 6,596 | | OUT_ACS | 7,445 |
| HX_ANGINA | 8,489 | | OUT_UNST_ANGINA | 2,104 |
| HX_OTHER_CHD | 20,416 | | OUT_ANGINA | 3,300 |
| HX_CHD_PROCS | 12,771 | | OUT_OTHER_CHD | 3,539 |
| HX_PCI | 8,646 | | OUT_PVD_DIAGS | 1,537 |
| HX_CABG | 5,659 | | OUT_PVD_PROCS | 1,922 |
| HX_OTHER_CHD_PROCS | 335 | | OUT_PCI_CABG | 5,758 |
| HX_PVD_DIAGS | 5,301 | | OUT_HAEMORRHAGIC _STROKE | 521 |
| HX_PVD_PROCS | 3,551 | | | |
| HX_HAEMORRHAGIC_STROKE | 1,204 | | OUT_CEVD | 4,364 |
| HX_CEVD | 8,403 | | OUT_ISCHAEMIC_STROKE | 3,011 |
| HX_ISCHAEMIC_STROKE | 5,878 | | OUT_TIA | 1,598 |
| HX_TIA | 3,159 | | OUT_OTHER_CEVD | 50 |
| HX_OTHER_CEVD | 772 | | OUT_HEART_FAILURE | 3,096 |
| HX_HEART_FAILURE | 8,079 | | OUT_ATRIAL_FIBRILLATION | 3,288 |
| HX_ATRIAL_FIBRILLATION | 10,902 | | | |

**Appendix Table 7** Descriptive statistics: HbA1c and eGFR. TEST_HBA1C, TESTED_HBA1C, TESTED_EGFR and TESTED_EGFR are binary features and the statistics are the number of quarters in the entire dataset where the feature contained a 1 and its relative percentage

| | |
|---|---|
| HBA1C mean (SD) | 47.98 (15.20) |
| TEST_HBA1C | 819,747 (28.9%) |
| TESTED_HBA1C | 2,268,295 (80.9%) |
| EGFR mean (SD) | 77.85 (20.11) |
| TEST_EGFR | 1,041,487 (37.2%) |
| TESTED_EGFR | 2,694,767 (96.1%) |

**Appendix Table 8** Descriptive statistics: PREDICT. PT_SMOKING, PT_DIABETES, PT_FAMILY_HISTORY, PT_GEN_LIPID, PT_RENAL, PT_ATRIAL_FIBRILLATION and PT_IMP_FATAL_CVD show number of patients in each category

| | | | | |
|---|---|---|---|---|
| PT_SBP mean (SD) | 132.25 (16.99) | | PT_GEN_LIPID | |
| | | | 0 (None) | 92,492 |
| | | | 1 (Familial hypercholesterolemia) | 5,569 |
| PT_SBP2 mean (SD) | 132.57 (17.24) | | 2 (Familial defective apoB) | 20 |
| | | | 3 (Familial combined dyslipidemia) | 499 |
| PT_DBP mean (SD) | 78.70 (10.25) | | 4 (Other genetic lipid disorder) | 1,516 |
| PT_DBP2 mean (SD) | 79.07 (10.30) | | | |
| PT_SMOKING | | | | |
| 0 (Never) | 66,896 | | | |
| 1 (Quit >12 mo) | 20,162 | | | |
| 2 (Quit ≤12 mo) | 1,901 | | | |
| 3 (Up to 10/d) | 6,249 | | | |
| 4 (11–19/d) | 3,046 | | PT_RENAL | |
| 5 (20+/d) | 1,842 | | 0 (Missing value) | 64,131 |
| PT_EN_TCHDL mean (SD) | 3.90 (1.22) | | 1 (No nephropathy) | 27,585 |
| | | | 2 (Confirmed microalbuminuria) | 5,996 |
| PT_DIABETES | | | 3 (Over diabetic Nephropathy) | 1,975 |
| 0 (No diabetes) | 64,125 | | 4 (Non-diabetic nephropathy) | 409 |
| 1 (Type 1) | 1,267 | | | |
| 2 (Type 2) | 32,754 | | | |
| 3 (Type unknown) | 1,950 | | | |
| PT_FAMILY_HISTORY | 20,162 | | | |
| | | | PT_DIABETES_YR mean (SD) | 8.19 (7.30) |
| | | | PT_ATRIAL_FIBRILLATION | |
| | | | 0 (Missing value) | 21 |
| | | | 1 (None) | 95,292 |
| | | | 2 (Confirmed atrial Fibrillation) | 4,783 |
| | | | | |
| | | | | |
| | | | PT_IMP_FATAL_CVD | 2,998 |

**Appendix Table 9** Removed features for the Cox regression analysis

| | Removed features |
|---|---|
| Cox (aggregated) | ETHN_5 |
| | DIED |
| | CVD_METOLAZONE |
| | OTHER_PREDNISOLONE |
| | OTHER_CLARITHROMYCIN |
| | OTHER_VILDAGLIPTIN |
| | PT_IMP_FATAL_CVD |
| Cox (last quarter) | ETHN_5 |
| | TESTED |
| | DIED |
| | CVD_METOLAZONE |
| | CVD_HYDRALAZINE_HYDROCHLORIDE |
| | OTHER_INSULIN_ZINC_SUSPENSION |
| | OTHER_PREDNISOLONE |
| | OTHER_CLARITHROMYCIN |
| | OTHER_VILDAGLIPTIN |
| | PT_IMP_FATAL_CVD |

 