2020-12-22

Personalized survival prediction of cardiovascular disease among hypertensive patients: a machine learning approach based on health administrative data

Feng, Yuanchao

UNIVERSITY OF CALGARY

Personalized survival prediction of cardiovascular disease among hypertensive patients: a machine learning approach based on health administrative data

by

Yuanchao Feng

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF SCIENCE

GRADUATE PROGRAM IN COMMUNITY HEALTH SCIENCES

CALGARY, ALBERTA

DECEMBER, 2020

Abstract

Background: Cardiovascular disease (CVD) kills approximately 17 million people globally every year, and they mainly exhibit as myocardial infarctions (MI), heart failure (HF) and stroke. Hypertension is the leading risk factor for premature death from CVD. Available routinely collected administrative health data with demographic features, comorbidities information, clinical laboratory test values and medication usage results can be used to perform biostatistics analysis aimed at highlights and correlations otherwise undetectable by medical doctors. Machine learning algorithms can predict patients' survival by using information recorded in their medical records.

Objective: To compare performance of four machine learning approaches on personalized survival prediction of CVD outcomes among newly diagnosed hypertensive patients.

Method: Hypertension cohort, CVD outcomes, and covariates were defined using validated case definitions applied to inpatient and outpatient administrative health databases. We analyzed a cohort of 11863 CVD events among 259,873 newly diagnosed hypertensive patients from April 1, 2009 to March 31, 2015 and had at least one-year follow-up. The dataset was divided into a training dataset (70% of the original data set) for generation of plausible models under machine learning algorithms, and a test dataset (30% of the original dataset) to test the performance of the models. We applied linear multi-task logistic regression (LMTLR), neural multi-task logistic regression (NMTLR), random survival forest (RSF) and Cox proportional hazard (CoxPH) models to both predict number of CVD outcomes in each survival time point and predict individual survival probability curve. The predictive performance was evaluated by root mean squared error (RMSE), mean absolute error (MAE), concordance index ($C - index$) and Brier score.

Results: Our results show that the RSF model has the lowest RMSE value at 33.94 and lowest MAE value at 28.37, which means it has the better performance to predict the number of CVD events in any time point during the follow-up period. Although all fours models exhibited a high discrimination and calibration ability when used for predicting individual survival probability curve, NMTLR model has the highest $C - index$ at 0.8149 and lowest Brier score at 0.0242 for the individual survival prediction.

Conclusions: This is the first personalized survival prediction for CVD among hypertensive patients using administrative data. The four models tested in this analysis (LMTLR, NMTLR, RSF, CoxPH) exhibited a similar discrimination and calibration ability in predicting the survival of hypertension patients. In the test dataset, RSF has better performance for population-based survival prediction while the NMTLR had better discrimination and calibration for individual-based survival prediction.

Acknowledgements

The research work reported in this thesis was carried out in the Department of Community Health Sciences at the University of Calgary under the supervision of Dr. Hude Quan and Dr. Robin Walker.

Firstly, I would like to express my sincere gratitude to my supervisor, Dr. Quan for his excellent supervision, invaluable guidance, constant encouragement, careful and patient assistance and friendly help throughout my whole MSc program. His deep love and perception of science, his persistent endeavour for pursuing the truth, and his consistent efforts at achieving perfection have always inspired and helped me carry out this research project. Secondly, I would like to recognize and thank my Co-supervisor Dr. Robin Walker. She has been instrumental in guiding me and providing feedback on this project. She gave many valuable suggestions and comments on the draft of my thesis. Certainly, the thesis cannot be accomplished without her supervision. They raise me up to more than I can be.

Thanks are extended to my committee members including Dr. Alexander Ah-Chi Leung and Dr. Xuewen Lu, each of whom provided valuable guidance throughout this research. Zhiying Liang who supported and provided valuable help and advice with data access and linking of administrative datasets.

Dedication

For

My wife Ye Lu, my daughter Isabelle, my parents and other family relatives,

and for others who have taught, encouraged and supported me over the past years.

Table of Contents

List of Tables

List of Figures and Illustrations

List of Symbols, Abbreviations and Nomenclature

Symbol	Definition
CVD	Cardiovascular disease
MI	Myocardial infarction
HF	Heart failure
LMTLR	Linear multi-task logistic regression
NMTLR	Neural multi-task logistic regression
RSF	Random survival forest
CoxPH	Cox proportional hazard
RMSE	Root mean squared error
MAE	Mean absolute error
C-index	Concordance index
DBP	Diastolic blood pressure
SBP	Systolic blood pressure
HbA1c	Glycated hemoglobin
LDL	Low-density lipoprotein
DAD	Discharge abstract database
CIHI	Canadian Institute for Health Information
PHN	Personal health number
ICD	International Classification of Disease
WHO	World Health Organization
HOST	Hypertension Outcome and Surveillance Team
PIN	Pharmaceutical information network
AHCIP	Alberta Health Care Insurance Plan
ICD-9-CM	ICD-9 Clinical Modification
ICD-10-CA	ICD-10 Canadian Modification
AHS	Alberta Health Services
EHR	Electronic Health Record
PPMSs	Pharmacy Practice Management Systems
DIN	Drug identification number
ATC	Anatomical therapeutic chemical
K-M	Kaplan–Meier
CART	Classification and Regression Trees
SVM	Support Vector Machine
OOB	Out-of-bag
CHF	Cumulative hazard function
CI	Concordance index
eGFR	Estimated glomerular filtration rate
AUC	Area under the receiver operation characteristic (ROC) curve
SD	Standard deviation

Chapter One: **Introduction**

1.1 Problem and significance

Blood pressure is the force of the blood pushing against the artery walls, which depends on the work being done by the heart and the resistance of the blood vessels. 'Normal' blood pressure is considered 120 over 80 mm of mercury (mmHg) or less. Hypertension Canada defines hypertension (known as high blood pressure) as a blood pressure 140 over 90 mmHg or higher for adults (130 over 80 or higher for diabetic chronic kidney disease) [1]. Hypertension is a highly prevalent condition affecting more than 20% of the Canadian adult population [2], and is the number one reason for primary care visits in Canada [3]. The health resource utilization associated with the management of hypertension is estimated to consume over 10% of all health care spending [4]. In fact, managing hypertension requires financial resources comparable to those used in diagnosis and management of stroke, MI and all other ischemic heart diseases combined [5]. Elevated blood pressure is a major etiological factor for cardiovascular disease (CVD), such as stroke and coronary heart disease, despite the fact that it can be effectively controlled through pharmacotherapy or with lifestyle changes in physical activity, diet, sodium intake, alcohol use, weight management, and smoking status [6].

Health care providers aim to provide timely and appropriate treatment for patients with hypertension to reduce the risk of CVD. However, this is not always the case due to the following reasons: (1) unawareness of the condition/diagnosis by the patient or physician; (2) unawareness of the risks associated with uncontrolled hypertension by the patient or physician; (3) non-adherence to treatment by the patient; (4) therapeutic inertia by the physician; and (5) the presence of treatment-resistant forms of hypertension, which

will lead to serious negative results, such as stroke and MI [7, 8]. Moreover, patients may miss the suitable treatment timepoint and develop worse outcomes due to treatment delays. This not only influences the life experience of the patients but also potentially wastes healthcare resources. For example, increased blood pressure is strongly associated with an increased risk of death [9]. As hypertension is an identified preventable risk factor, timely and accurate treatment is essential to reduce the risk of CVD and death among hypertensive patients [10].

The CVD risks attributed to hypertension can be reduced by blood pressure treatment and control. The use of antihypertensive drugs reduces the risk of subsequent CVD [11], including a 35-40% stroke risk reduction and a 20-25% reduction in the risk of MI and HF [12, 13]. All large randomized controlled trials have consistently shown that antihypertensive therapy can reduce the risk of stroke, regardless of the drug category [14, 15]. However, most epidemiological studies on hypertension focus on population level and few studies centre on individual patients [16-18]. Accurate individual estimates of hypertension outcomes and demographic risk factors are critical to health services planning and prioritization.

It is well known that estimating and reporting an individual's global cardiovascular risk (e.g., Framingham Risk Score) can help to improve risk perception, strengthen communication between physicians and patients regarding healthcare and healthy behaviors, which may potentially reduce adverse consequences for patients [19-21]. An examination of approximately 13 million individuals with uncontrolled blood pressure revealed that 36.2% of them are neither aware of their hypertension nor taking antihypertensive medications [22]. At the same time, more individuals are actively

interested in their own health; and more physicians are interested in incorporating the wealth of existing patients' information for their treatment [23]. Therefore, individual cardiovascular risk assessment and survival prediction can be acted as an effective way to engage individuals in conversations to improve healthy behaviors to lower their blood pressure and reduce the risk of CVD outcomes.

Precision medicine which takes into account the characteristics of each patient's disease treatment and prevention, will allow doctors and researchers to provide more accurate treatment and prevention strategies for a patient with the specific disease. Different from the traditional "one size fits all" approach, a strategy developed for the average person without considering individual differences, precision medicine focuses on personalized treatment.

Accurate survival prediction is desirable to provide prognosis and treatment decisions for hypertension patients when they seek healthcare services. However, current prediction models are based on aggregate data for population-based predictions opposed to individual-level predictions. The population-based prediction models are not suitable for survival prediction at the individual level, due to the fact that when a health risk is incorrectly made for an individual based on the aggregated data of a certain group, it may lead to the rise of reasoning failure (known as ecological fallacy in epidemiology) [24]. Therefore, it would be useful for healthcare providers to have an accurate and applicable method to provide individual survival rate distribution and probability survival time for each unique patient. Unfortunately, many current survival analysis methods (e.g., CoxPH, K-M) cannot accurately answer such questions. The CoxPH model with risk scores can only provide a relative survival measure but not the probability of survival time.

Conventional survival analysis methods (i.e., K-M) are not focused on specific patients but an entire population. Since all these estimation models are applied on the average population, they are not designed to conduct individual patient estimation since these models do not include patient-specific information such as age, treatment administered, or comorbidities. Moreover, socio-economic status, level of education, occupation, as well as ethnic groups are main factors that lead to health inequity, which may result in disparities for individual hypertensive patients [25]. Patient-specific information (e.g., age, sex, comorbidities) causes patient heterogeneity which leads to the need to provide individual-level survival prediction models for hypertensive patients. Individual-level prediction models among hypertensive patients can incorporate important risk factors (e.g., age, sex) into model development to avoid patient heterogeneity. Individual level modelling has the potential to answer key policy questions with respect to reducing the burden of hypertension in Canada by improving individual health behaviours. Thus, it is paramount to provide individual-level prediction for hypertensive patients by directly and correctly incorporating these important factors explicitly in the prognostic models to reduce the future burden of hypertension in Canada by improving individual health behaviours.

1.2 Research purpose

The digitalization of health data is rapidly changing our lives. The emergence of big health data is changing health services delivery, which is also becoming a critical source of data for achieving precision medicine and population health. As current technological capabilities continue to develop, our analytical capacity must adapt. Big data is attracting many professions, from healthcare providers to health system administrators,

toward precision medicine. Precision medicine proposes that medical decisions, treatments, and clinical practices must be tailored to individual patients. Precision medicine is about applying appropriate statistical modelling on available clinical and biological data to make personalized outcome predictions for patients. When a physician is armed with precision medicine technology, they will obtain more accurate and precise information about their patients' predicted outcome(s). This project aims to lay the groundwork to provide precision medicine for individual hypertensive patients, which will ideally improve their treatment in Alberta. The overall goal of this study is to develop a personalized CVD survival predictive model for individual hypertension patients to improve the quality of care delivered for Alberta residents. Progress will be made in the following areas:

1. Develop personalized CVD survival prediction models for newly diagnosed hypertension patients by using information collected in administrative health data;

2. Compare the performance of different models and choose the better one for personalized survival prediction;

3. Develop tools that can be applied to clinical decision-making with the aim to reduce the morbidity and mortality of cardiovascular diseases.

The long-term objective of this research to provide accurate personalized risk prediction for patients with newly diagnosed hypertension, inform treatment, engage patients and providers, and ultimately reduce the global risk of CVD.

Chapter Two: **Background**

2.1 Hypertension and CVD

2.1.1 Definitions

The delivery of blood from the heart to the rest parts of the body through circulation is essential for human life. This process ensures the delivery of oxygen and nutrients to vital tissues and organs. The cardiovascular system includes the heart and all blood vessel networks and diseases or disorders related to the system are termed CVD. High blood pressure, known as hypertension, is a long-term medical condition in which arterial blood pressure continues to rise and can lead to a variety of CVD, such as coronary heart disease, stroke, congestive heart failure, peripheral arterial disease or peripheral vascular diseases, and MI.

The definition of high blood pressure (hypertension) is based on systolic blood pressure (SBP) and diastolic blood pressure (DBP) levels [26]. SBP is the maximum pressure generated following contraction of the left ventricle which ejects blood out of the heart into the aorta and peripheral arteries. DBP is the lowest pressure in the artery and coincides with relaxation of the left ventricle. Most international expert panels have defined hypertension using a SBP/DBP diagnostic threshold of ≥140/90 mmHg [27], though some have proposed a lower threshold of ≥130/80 mmHg [28]. Antihypertensive therapy is commonly recommended for patients with macrovascular target organ damage or other independent cardiovascular risk factors when SBP is higher than 140 mmHg or DBP is higher than 90 mmHg, and for patients without macrovascular target organ damage when SBP is higher than 160 mmHg or DBP is higher than 100 mmHg [29]. High blood pressure leads to functional and structural changes in the endothelium and the vascular wall

by exerting a force on the arterial walls. These changes include thickening of the arterial wall through proliferation of the vascular media smooth muscle cells and collagen deposition (fibrosis), as well as the dilation and elongation of blood vessels [30].

Hypertension is called the “silent killer” because it can damage essential body organs - such as the brain, heart, kidney and eyes - without symptoms until a devastating clinical event occurs, such as stroke or MI. In the heart, hypertension causes arteriosclerosis and a subsequent narrowing of coronary arteries, which eventually leads to cardiac ischemia, manifested as angina, MI, or arrhythmia. Because hypertension opposes the contractile force of the left ventricle, the workload of the left ventricle increases, leading to compensatory ventricular hypertrophy and, when compensation is no longer possible, to HF [31]. Hypertension can lead to ischemic or hemorrhagic stroke by impaired blood flow or bleeding, respectively, and result in serious neurological impairment. Furthermore, the cardiovascular complications related to hypertension can also result in death in some situations.

2.1.2 Relationship between hypertension and CVD

Hypertension is considered to be one of the most important modifiable risk factors for the development of CVD. The Framingham study indicated that 91% of HF, 84% of strokes, and 70% of MI occurred in hypertensive individuals. [32] It has been shown that treatment of hypertension is associated with a decrease in mortality and mobility for cardiovascular events, however, hypertension is commonly undiagnosed and often untreated. [33] Canada’s healthcare system consists of 13 provincial and territorial health insurance plans that provide universal healthcare coverage to Canadian citizens, permanent

residents and certain temporary residents. Hypertension Canada, a national non-for-profit organization, has made tremendous efforts to improve hypertension treatment and control nationwide. Quan et al. examined rates of cardiovascular events and associated mortality in 3.5 million Canadian adults with newly diagnosed hypertension for a period of 12 years [8]. The crude incidence of composite CVD hospitalization in this cohort without previous cardiovascular disease was 19.3 per 1000 person-years: 6.9 per 1000 person-years for stroke, 8.4 per 1000 person-years for MI, and 8.5 per 1000 person-years for HF. The results of this study indicated that the highest incidence cases of CVD outcomes among hypertensive Canadians are stroke, MI and HF. Quan et al. also found that the hypertension individuals that are elderly, male, reside in rural or low-income areas and those with comorbidities tend to have the poorest outcomes, which provides strong evidence that patient-specific information can influence individual CVD survival prediction among hypertensive patients.

2.1.3 Importance of survival prediction for patients with hypertension

Analyzing Canadian data, Robitaille et al. reported that all-cause mortality was 2-4 times higher among patients with hypertension than among those without hypertension in the 20-49 year-old age strata, despite substantial improvements in hypertension treatment patterns and control rates over this time period in Canada. Due to the high prevalence of CVD among hypertensive patients, identifying high-risk patients and prevention strategies for CVD are important steps toward improving management and reducing future health problems. Predicting the disease development process of individual hypertension patients has been shown feasible to improve the management of hypertension

among Canadians by accurate blood pressure measurement and timely health care [34]. Individual survival prediction using big data and machine learning methods can provide opportunities to assess hypertension prevalence and risk, to diagnose and gauge the severity of hypertension and to estimate the risks of subsequent complication, thereby offering the opportunity for timely treatment [35-37].

2.1.4 CVD survival prediction among patients with hypertension

As discussed above, CVD is a very common outcome among patients with hypertension. Certain risk factors, such as age, sex/gender, genetics, lifestyles, and some biochemistry parameters of the body have been found to be contributing factors to the progression of CVD among hypertensive patients [38]. Unlike non-modifiable risk factors (age, sex, and family history), modifiable cardiovascular risk factors, on the other hand, may potentially be mitigated with intervention and treatment. These include high blood pressure, smoking, diabetes, and elevated cholesterol.

Hypertension, which is a well-known risk factor for CVD, may develop cardiovascular disease, especially when uncontrolled. Higher blood pressure makes the heart work harder to pump blood out of the heart to blood vessels, which results in the wall of heart thickens at the beginning and then gradually thins, and the heart muscle finally becomes too thin to contract powerfully, causing uncontrolled pressure.

Certain biochemistry parameters, such as serum lipid panel, fasting blood glucose, estimated glomerular filtration rates (eGFR), cholesterol levels, glycated hemoglobin (HbA1c), and comorbidities are modifiable risk factors for CVD. For example, low-density lipoproteins (LDL) assists fat to build up on the artery wall. The deposited fat will shrink

the space of the artery which eventually can lead to high blood pressure. As the LDL cholesterol level increases the risk of CVD also rises.

Age and sex are typically uncontrollable risk factors for CVD. As we age, many physiological processes decline progressively and the structure of the heart and blood vessels changes. Specifically, aging results in endothelial dysfunction, which leads to death and decline of heart endothelial cells. A clinical study has shown that endothelial dysfunction can lead to high blood pressure and increases the risk for development of atherosclerosis and stroke [39]. Sex is also a critical risk factor for CVD, as shown in research that men are more likely to have heart and blood vessel problems than women [40].

The impact of these risk factor(s) can be cumulative or enhance the development of CVD, which thus should be considered interactively or in an integrative way to study their influence on CVD. Therefore, it would be meaningful to build up a system or model considering all potential risks. Then, effective risk prediction or evaluation can be made, and suggestions and treatments can be used to prevent or delay the disease onset or progression.

Several algorithms and models have been developed for predication of cardiovascular events based on an individual's CVD risk factors [41-43]. Researchers and clinicians usually apply these models to compare the survival time of two populations or to test significant risk factors that affect survival, such as the Framingham Heart Study, the Lipid Research Clinic Follow-up Cohort (The Cardiovascular Life Expectancy Model), the Dundee Cohort, the British Regional Heart Study, the Munster Cohort and the Systematic Coronary Risk Evaluation (SCORE) [44-47]. These models are not designed for the task

of predicting survival time for individual patients. Moreover, potential factors, such as geography, social, culture and genetic uniqueness, will influence the accuracy of models to be generalize to other target populations. In addition, as these models work with the hazard function instead of the survival function, they might not provide calibrated predictions on survival rates for individuals. Recently, Manuel et al. [48] developed a Cardiovascular Disease Population Risk Tool (CVDPoRT) for individual risk prediction using Canadian Community Health Surveys data. The model can be used for both individual CVD risk assessment and population CVD risk assessment. However, this model only focused on the risk assessment rather than the survival prediction, which therefore cannot be used for patient's survival time prediction.

2.2 Administrative health data for hypertension and CVD surveillance

2.2.1 What are administrative health data

Understanding the nature and distribution of hypertension and its complications (such as CVD) in a population is important to control the burden from them.. Routinely collected health-related data from healthcare services generates so-called administrative health data, which are frequently used for hypertension surveillance and research. The administration of healthcare services provides a wide variety of information that is routinely collected and covers large segments of the population: administrative health data. These databases commonly record hospitalizations and ambulatory visits of individual patients, and often list specific diagnoses and procedures. These data are collected for three primary purposes: 1) administration of healthcare services; 2) enrollment into health insurance plans, and; 3) reimbursement for services.

These data are commonly used for health services research because they are readily available, inexpensive to access, and provide information for large populations. Administrative health data are accumulated in a distributed manner over a prolonged period of time, allowing researchers to follow patients through time. As a result, these data capture a large number of individuals with a wide range of demographic information, including race and geographical area of residence. There are many administrative databases in Canada, for example, registry data, physician claims, hospital discharge abstract databases, ambulatory care, laboratory data, Alberta Blue Cross pharmacy claims, pharmaceutical information network (PIN) data, and so on [49]. As the health administrative databases are maintained primarily for financial and administrative purpose, some of the databases are audited to ensure their accuracy for physician remuneration and resource allocation, but not necessarily for clinical correctness. For example, the national Discharge Abstract Dataset (DAD) is audited by the Canadian Institute for Health Information (CIHI), and physician claims in Alberta is audited by Alberta Health for fee-for-service billing. The administrative health databases available in Alberta share some similar qualities with each patient having a unique personal health number (PHN), which enables linkages between databases. Given the above, administrative data can be utilized to develop, analyze and interpret CVD survival analysis for patients with hypertension as well as survival model development, and report CVD risk behaviors, risk factors, morbidity, mortality and estimate incidence and prevalence.

2.2.2 Content and coding of administrative health data

Administrative data uses a consistent set of unique identifiers, permitting researchers or health service providers to build histories of patients across files. The completeness, accuracy and validity of administrative data are important elements to understand in order to make informed decisions around healthcare policies, assessing clinical care, healthcare delivery and healthcare utilization [50]. The administrative system in healthcare sites routinely collects information from patients' medical chart for administrative purposes. In Canada, some administrative health data use International Classification of Diseases (ICD) to record diseases. The ICD coding system was developed by the World Health Organization (WHO) and published its first iteration in 1949. It allows the classification of diseases as well as various signs, symptoms, abnormal findings, cause of injury and/or health outcomes. The purpose of WHO's ICD regulations was to promote international comparability in the classification and collection of health data. Since then new versions have been published periodically, with each iteration attempting to improve the accuracy and comprehensiveness of the previous version. Both ICD-9 and ICD-10 are widely used in the administrative health data system. One of the major characteristics of ICD-10 is that it contains an alphanumeric classification system with a code size ranging from 3 to 6 characters starting with a letter followed by the minimum 2 digits, to represent various diseases, conditions and health outcomes. Generally speaking, the first three digits of an ICD-10 code represent a broader condition or related group of conditions, with the remaining digits representing more detailed information. It is far more detailed (there are a total of 12,420 codes in ICD-10 compared to 6,969 in ICD-9) permitting richer and more precise capture of clinical information. ICD-11 has been under development since 2014

and was officially released in 2018, which will better reflect the modern understanding of diseases. There are also minor regional changes or additions to the basic ICD coding system with its broad changes from each coding iteration. Canada uses ICD-10-CA version of ICD-10 for DAD, which includes enhancements that are relative to the Canadian healthcare system and was nationally adopted in 2002 [51].

2.2.3 Definition and validation of administrative health data for CVD and hypertension

Prior to using administrative data, it is imperative that the validity and accuracy of coding is deemed adequate for its intended use in terms of the conditions and procedures being measured. The accuracy of a coding definition depends on many factors including the codes chosen, and the population studied. Lack of specificity, misclassification of diagnosis resulting from human error in recording the diagnosis, as well as the range of codes, may affect the data quality when using ICD codes to study the disease of interest, which is consequently problematic. It is common to compare administrative data with other types of data, such as a reference or "gold standard" level of data (such as patient medical chart), to clarify definitions of disease, especially when selecting a cohort of patients through a diagnostic ICD code.

For hypertension and CVD case definitions, measuring the extent and impact of these conditions has been difficult due to the discrepancies in clinical definitions, resulting in large differences in estimates. Differences in coding algorithms can remarkably affect the reported incidence rates for a given disease. Quan et al., identified patients with hypertension in the administrative databases using the corresponding ICD-9 and ICD-10-CA codes (ICD-9 codes:401.x, 402.x, 403.x, 404.x, and 405.x; ICD-10-CA codes: I10.x,

I11.x,I12.x, I13.x, and I15.x) in the ≤25 coding fields for diagnosis in the DAD, ≤3 fields in physician claims data, and ≤10 fields in National Ambulatory Care Reporting System database. The administrative data coding algorithm to identify hypertension of "2 claims within 2 years or 1 hospitalization" had relatively high validity with sensitivity of 75%, specificity of 94%, positive predictive value of 81%, and negative predictive value of 92%. These results provide evidence that administrative data can be used as a relatively valid data source to define cases of hypertension for surveillance and research purposes. [52] Tu et al., identified hospital visits for stroke using the ICD-10-CA codes I60.x, I61.x, I63.x (excluding I63.6 cerebral infarction due to central venous thrombosis), I64, H34.1, and G45.x (excluding G45.4 transient global amnesia), H34.0 after 2002 and 362.3, 430, 431, 434.x,436, and 435.x ICD-9 codes before 2002. The study found that the use of physician claims data in addition to hospitalization data improved the sensitivity of the coding algorithm. The coding algorithm of "2 physician claims within 1 year or 1 hospitalization" had a sensitivity of 68.0% (95% confidence interval [CI], 60.5%-75.5%) and specificity of 98.9% (95% CI, 98.6%-99.2%) for the identification of stroke patients. [53] Quach et al., studied more than 70 different ICD codes for defining HF in 25 published studies and found that ICD-9 code 428 and ICD-10-CA code I50 were the most commonly used ones to define HF in DAD. [54] Metcalfe, et al., studied the case definition for acute MI in administrative databases. The study recommended that ICD-9-CM 410 or ICD-10 I21-I22 in the primary diagnosis coding field should be used to define AMI with sensitivity (ICD-9-CM 410: 83.3%; ICD-10 I21-I22: 82.8%) and PPV (ICD-9-CM 410: 82.8%; ICD-10 I21-I22: 82.2%), separately [55].

2.2.4 HOST Project

The Hypertension Outcome and Surveillance Team (HOST), a national team composed of clinicians and researchers, utilizes rich Alberta population level databases that are a valuable resource for studying and monitoring research and surveillance of hypertension, management and outcomes [52, 56-58]. It aims to enhance our understanding of the outcomes associated with hypertension and inform intervention strategies to improve hypertension management and outcomes. The HOST project uses data on individually health behavioural risk factors from routinely collected large population-based health administrative data, to develop and validate an individual-based risk prediction model for hypertensive patients. The HOST project extracted several administrative health datasets in Alberta including registry data, vital statistics data, physician claims, hospital DAD, ambulatory care, laboratory data, Alberta Blue Cross pharmacy claims, and PIN data. The datasets have been cross-linked through a unique and anonymous identifier.

The registry data contains demographic information for all Albertans with Alberta Health Care Insurance Plan (AHCIP) coverage, which is commonly used as a denominator for Alberta population counts. Several registry tables are available from Alberta Health that report registration data at month end, quarter end, and fiscal year end. The fiscal year end file is considered to be the most accurate snapshot of each registrant as of March 31 each year, which is useful for consistently reporting recipient date of birth, sex, and postal code at fiscal year-end (rather than using health services event data which may vary). The registry data contains personal health number, migration indicators, date of birth and death, sex and postal code information.

All live births and deaths within Alberta must be registered with Alberta Vital Statistics. The mother's date of birth, sex, place of birth and demographic information will be available from the vital statistics birth data. The vital statistics death data provides date and place of death, cause of death, and demographic information including age, sex and residence information. ICD-10-CA is used for coding the underlying cause of death. Alberta Health adds the identifier to Vital Statistics data to permit linkage to other data sources.

The Physician Claims dataset is collected as part of a routinely submitted fee-for-service physician's remuneration system to the AHCIP. The data collected includes information on provider, patient and services. Providers submit up to 3 ICD-9 diagnosis codes for each patient visit. Alberta physicians submit their services to Albertan residents who are registered under the AHCIP.

The DAD contains all acute care inpatient visit information, from time of admission until discharge or death. Clinical visit information for all patients discharged from Alberta hospitals is abstracted and recorded. DAD includes administrative, clinical and diagnostic data, capturing up to 25 diagnoses, which are coded in ICD-9, ICD-9 Clinical Modification (ICD-9-CM) or after 2002, ICD-10 Clinical Modification (i.e. ICD-10-CA).

The Ambulatory care dataset includes emergency department visits and day procedures. Ambulatory care is facility-based outpatient medical and/or surgical care that is provided in publicly funded clinics, day surgery, and emergency department settings. Patient demographics, admission/discharge/transfer information, provider information, coded diagnoses and intervention information can be found in the Ambulatory care dataset.

Multiple laboratory information systems are used within Alberta to place laboratory orders, track workflows and report laboratory and pathology results. All laboratory data utilized in the HOST project is released by Alberta Health Services (AHS) and includes clinical chemistry, toxicology, hematology, serology, urinalysis, and immunology. The inclusion of laboratory data in this study allows the development of highly accurate predictive models, providing the opportunity to move towards precision medicine. For example, the cardiovascular risk can be evaluated by the presence of severe hypercholesterolemia (LDL cholesterol >5.0 mmol/L) and diabetes (fasting blood glucose ≥7.0 mmol/L and/or glycated hemoglobin [HbA1C] ≥6.5%) [59].

Alberta Blue Cross pharmacy claims covers people ≥ 65 years of age, their dependents, and persons on assistance. It provides information on prescription drug dispensing under supplementary health benefit plans, private or semi-private hospital accommodation and claims for extended health services, such as ambulance services, clinical psychological services, home nursing care, prosthetic and orthotic benefits. This database provides information such as claim type/subtype, date of claimed service, date of payment, information about recipients, claim details, prescriber and pharmacy identifiers from Alberta Blue Cross pharmacy claims.

PIN data records each time a prescription is dispensed to a patient. PIN is a web-enabled application within the Alberta Netcare Electronic Health Record (EHR) that provides access to a patient's active and previous dispensed medications. It has real-time access through integrated clinical system EHR, includes Electronic Medical Records, and Pharmacy Practice Management Systems. PIN dispenses data that was been extracted from the EHR and contains rich information, such as dispensing data, drug identification number

(DIN) and anatomical therapeutic chemical (ATC) code. Moreover, it provides a report of an individual's drug dispensed summary, and a detailed listing of medications dispensed for a specific period. All of those HOST project data can be linked together by the PHN which is the unique resident identification for residents in Alberta.

2.3 Traditional survival analysis models

2.3.1 What is survival analysis

Survival analysis aims to examine and model the time of an event of interest, which has been widely used in medicine. For instance, in clinical study, it can be defined as a set of statistical methods looking at elapsed time between the beginning of a study and the occurrence of one or more events of interest. It is worth to note that survival analysis is different from survival prediction. Survival analysis models the probability distribution of one phenomenon within a population, which is typically used to discover the prognostic factors (such as risk factors) or estimate the median survival time of this population. In contrast, the goal of survival prediction is to accurately predict the remaining time to this outcome for each individual within a population. In the present study, we modeled the survival prediction of personalized hypertensive patients to develop CVD outcomes.

2.3.2 Failure (event) and censoring

Survival prediction is similar to regression as both involve models that regress the covariates to estimate the value of a dependent real-valued response variables - here, the variable is “time to event” (e.g., death, disease recurrence). In many medical studies, the event of interest for many individuals (death, disease recurrence) might not have occurred

within the fixed period of study. In addition, other subjects could move out of town or decide to drop out any time. Here we know only the date of the final visit, which provides a lower value on the survival time. We refer to the time recorded as the "event time", whether it is the true survival time, or just the time of the last visit (censoring time). Such datasets are considered censored, as shown in Figure 2.1.

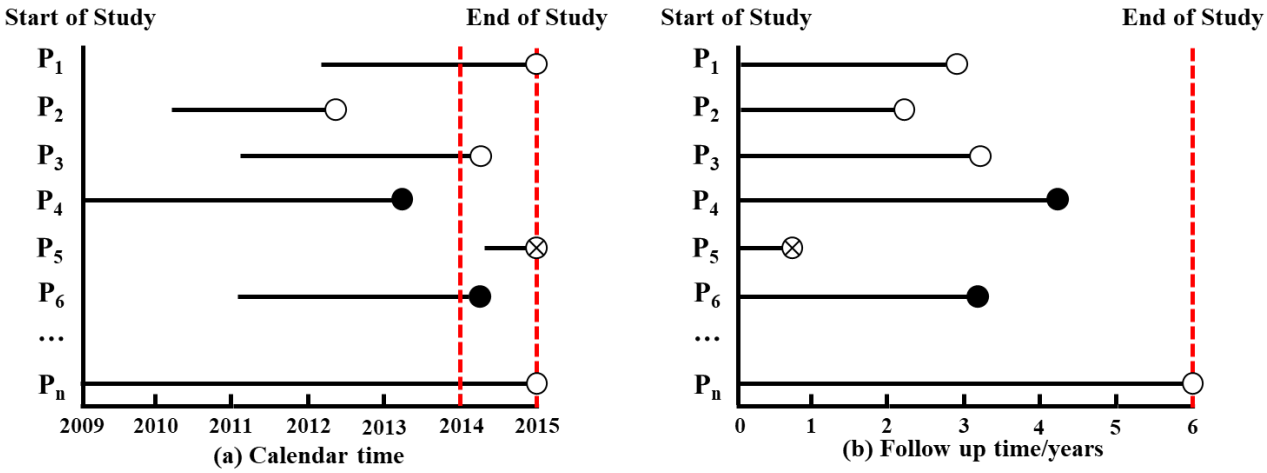


Figure 2.1 An illustration to demonstrate the survival analysis.

Note: P = patient; ○ = censored patient; ● = uncensored patient; ⊗ = patients who have outcome less than one-year follow-up.

For example, patient one is still alive at the end of the study, which means we only know that he/she lived until the end of the study without information on the actual survival time. If a study participant drops off during the study period, for example patient two and patient three, we also do not know the outcome and alive time of that study subject. If the follow-up fails, standard regression method will not be able to handle "partial label information". Survival prediction algorithms can model from a cohort including such

censored data by developing an effective model. The survival dataset of this model contains a set of patient descriptions as well as two other labels: one is time, corresponding to the time from diagnosis to a final data point (either death, CVD, or time of last follow-up), and the other is the status, which indicates whether the patient has outcome by the final date of the study.

2.3.3 Basic functions of survival analysis

Survival analysis traditionally focuses on analyzing prognostic factor and/or modelling the survival distribution of a population. The heterogeneity of patients coupled with the need to provide probabilistic estimates at several time points has motivated the creation of several survival time distribution and survival analysis. There are some well-established methods on modelling such survival analysis and survival distribution, which are designed to deal with various tasks and provide the survival time T of a population, which provides the values for features $\vec{x}_i = [x_i^{(1)}, \dots, x_i^{(k)}]$ (i is for the individual, k represents the number of covariates) for each member of a historical patients cohort, as well as the actual time of the "event" which is either disease (uncensored) or the last visit (censored), with 0 or 1 serving as the indicator respectively for event and censoring. In most of the survival dataset, $\vec{x}$ is a vector of feature values describing a patient, using information that are available when that patient entered the study, for example, the patient's age, sex/gender, comorbidities information when the patient was first diagnosed with the disease or started the treatment. Additionally, it is assumed that each patient has a outcome time, d_i, and a censoring time, c_i, and $\delta_i = 1$ if $d_i \leq c_i$ or $\delta_i = 0$ if $d_i > c_i$ (δ indicates the outcome).

In most of these models, a basic quantity of interest is the survival function S(t) = P (T $\geqslant$ t), which is the probability that an individual within the population will survive longer than time t [60]. Given the survival times of a set of individuals, we can plot the proportion of surviving individuals against time, as a way to visualize S(t) (as shown in Figure 2.2). The plot of this empirical survival distribution is called the K-M curve. Empirical distribution cannot handle censoring. In the absence of censoring they are the same. This is closely related to the hazard function h(t), which describes the instantaneous rate of failure at time t and is also known as the force of mortality, the conditional failure rate, or the instantaneous death rate [61]. The hazard function is not the chance or probability of the event of interest, but instead it is the event rate at time t conditional on survival until time t. The mathematically hazard function and its relationship with survival function are defined as below:

$$\mathrm{h(t)} = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \text{ and } S(t) = \exp\big(-H(t)\big) = \exp\left(-\int_0^t \mathrm{h(u)}\, du\right.$$

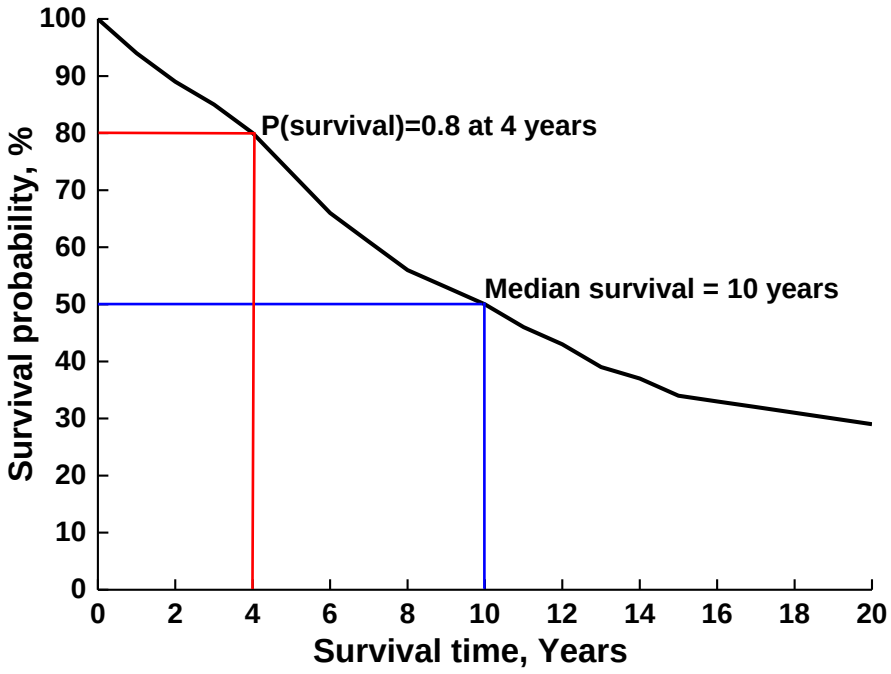


Figure 2.2 An example of survival curve.

2.3.4 Statistical methods for survival analysis

Statistical models in survival analysis can be classified into three categories: parametric models, semi-parametric models, and non-parametric models. Among all statistical methods, non-parametric methods themselves do not deal with multivariate regression analysis, while parametric methods reply on distribution selection. In contrast, semi-parametric methods should be the better approach, because it is distribution free and can handle multivariate regression problems. Therefore, the most widely-used semi-parametric statistical method, CoxPH model, was utilized in this project. However, in many existing machine learning methods for survival analysis, non-parametric methods are integrated with standard machine learning methods to deal with censored data and predict

survival. Table 2.1 provides a comprehensive summary of different statistical methods for survival analysis [62].

Table 2.1 Summary of different types of statistical methods for survival analysis.

Types	Advantage	Disadvantage	Specific methods
Non-parametric	Very straightforward and efficient if unknown the suitable theoretical distribution.	Complex to interpret the results.	Kaplan-Meier Life table
Semi-parametric	No need to know the distribution of survival times. Hazard ratios are well understood.	Unknow outcome distribution	CoxPH Regularized Cox
Parametric	Easy to interpret. Will get more accurate results when the survival time follows the particular distribution.	It may be inconsistent and can give sub-optimal results if the distribution assumption can not meet.	Linear regression Accelerated failure time

2.3.5 Non-parametric survival analysis (K-M model)

The survival function and hazard function can be calculated by researchers to help them understand patient outcome(s). The survival function gives the probability that a person survives longer than a specific time [63]. K-M survival analysis is widely used for patient survival prediction for specific risk factors [64]. It is advantageous in the estimation of the survival distribution, survival function, $S(t)$ that can be used to calculate the survived probability of patients up to a certain point in time. The estimate is the product of a series of estimated conditional probabilities by the formula:

$$S(t) = \prod_{t_i<t}(1 - \frac{d_i}{n_i})$$

where with d_i the number of events and the $n_i$total individuals at risk at time t_i.

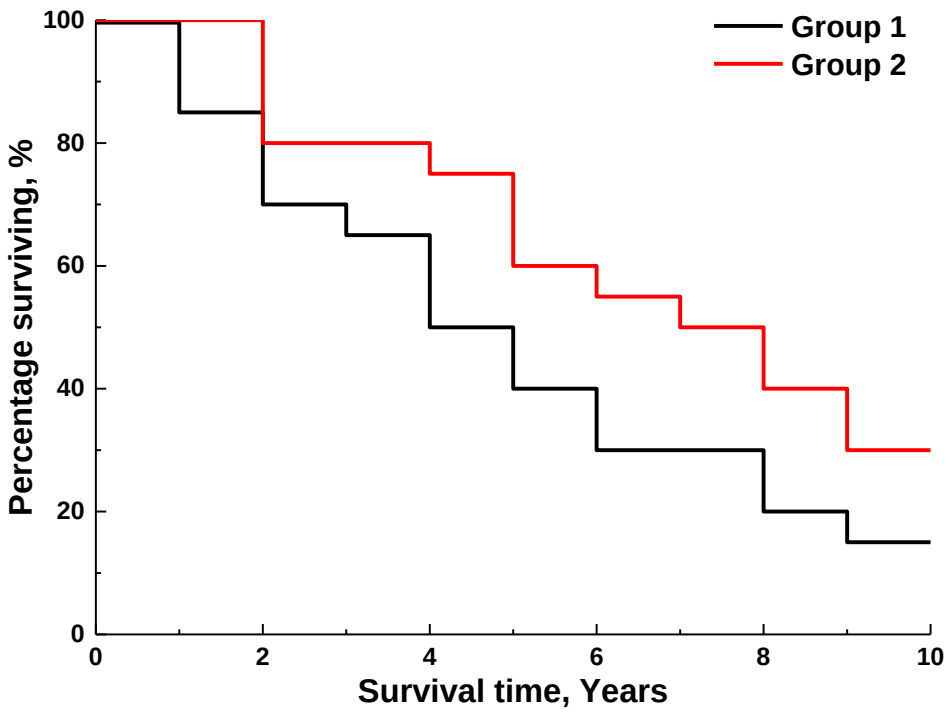


Figure 2.3 K-M curve shown survival probability to time *t*.

K-M curves that comes from K-M survival analysis can be used to compare survival curves by splitting the cohorts of patients into two or more groups using a single categorical variable, as seen in Figure 2.3 [65]. Each point on the curve [t, S(t)], shows the survival probability S(t) $\in$ [0, 1] for each time $t \geqslant 0$.

As one of the standard methods in survival analysis, K-M analysis can provide survivor tables or a survivor curve. One advantage is that this model effectively incorporates censored observations since the estimator only requires the information on the size of the risk set and number of events at each distinct time. It is worth to note that the final product of K-M estimator is not a survival time but a survival distribution, which can be used to better understand the disease itself by investigating whether some specific

feature made a difference or if a treatment's influence on the outcome. K-M cannot provide individual survival prediction and is different from this study's objective (individual level prediction).

2.3.6 Semi-parametric survival analysis, hazard function and hazard ratio

CoxPH model is the most popular method that incorporates time to fit the survival time of a population [66]. Unlike the K-M model, the CoxPH model does use the subject features and works with the hazard function instead of survival function to investigate the relationship between independent study variables and survival time. The mathematical form of the hazard function is:

$$H(t, \vec{x}_i) = h_0(t) \times e^{\sum \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \cdots + \beta_k x_i^{(k)}} = h_0(t) e^{\beta\, \vec{x}_i^{(k)}}$$

Where $h_0(t)$ is the baseline hazard function, which can be an arbitrary non-negative function of time, $\vec{x}_i = [x_i^{(1)}, \ldots, x_i^{(k)}]$ is considered as potential k predictors of the event for a set of variables (X), and $(\beta_1,\ \beta_2 \cdots \beta_k)$ is the coefficient vector. i represents for the individual patients. The Cox model is a semi-parametric model since it does not specify the form of baseline hazard function $h_0(t)$. Based on this assumption, the hazard ratio does not depend on the baseline hazard function, which can be expressed as follows for two individuals:

$$\frac{H(t, \vec{x}_1)}{H(t, \vec{x}_2)} = \frac{h_0(t) e^{\beta\, \vec{x}_1^{(k)}}}{h_0(t) e^{\beta\, \vec{x}_2^{(k)}}} = e^{\beta\, (\vec{x}_1^{(k)} - \vec{x}_2^{(k)})}$$

Since the hazard ratio is a constant, and all the subjects share the same baseline hazard function, the Cox model is a proportional hazard model. Based on this assumption the survival function is given by:

$$S(t, \vec{x}_i) = S_0(t) \times e^{\sum \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \cdots + \beta_n x_i^{(k)}} = S_0(t) e^{\beta\ \vec{x}_i^{(k)}}$$

where $S_0(t)$ is the baseline survival function.

One of the benefits of the CoxPH mode is that we can leave the baseline function unspecified and research the hazard ratio between two groups if we know the baseline function is consistent. Moreover, we can estimate the variable parameter β by maximum partial likelihood estimation without knowledge of the baseline hazard $h_0(t)$. For this reason, this model is not designed to produce the survival time for an individual patient as the baseline hazard function is unknown.

2.3.7 Parametric survival analysis models

Parametric methods for estimating the survival probability are efficient and accurate when survival times follow a particular distribution. Unlike the CoxPH model, in parametric methods, a complete likelihood function can be solved directly, and the parameters can be estimated using maximum-likelihood estimation. Choosing an appropriate theoretical distribution to approximate the success curve is a critical component of the parametric methods. However, since the distribution is not known priori, one has to fit different models to the data in the suitable possible manner. Table 2.2 shows the probability density function, survival, and hazard functions of commonly used distributions for parametric survival regression [67, 68].

Table 2.2 Survival density function, survival and hazard functions for commonly used distributions.

Distribution	PDF $f(t)$	Survival function $S(t)$	Hazard function $h(t)$
Exponential	$\lambda e^{-\lambda t}$	$e^{-\lambda t}$	λ
Weibull	$\lambda k(t)^{k-1}e^{-\lambda t^k}$	$e^{-\lambda t^k}$	$\lambda k t^{k-1}$
Gamma	$\frac{\lambda^k t^{k-1}e^{-\lambda t}}{\Gamma(k)}$	$1-\frac{\int_0^t \sigma^{k-1}e^{-\sigma}d\sigma}{\Gamma(k)}$	$\frac{\lambda^k t^{k-1}e^{-\lambda t}}{(1-\frac{\int_0^t \sigma^{k-1}e^{-\sigma}d\sigma}{\Gamma(k)})\Gamma(k)}$
Log-logistic	$\lambda p t^{k-1}(1+\lambda t^k)^{-2}$	$\frac{1}{1+\lambda k t}$	$\frac{\lambda k t^{k-1}}{1+\lambda k t}$
Log-normal	$\frac{1}{\sqrt{2\pi}\sigma t}e^{-\frac{(\ln(t)-\mu)^2}{2\sigma^2}}$	$e^{\mu+\frac{\sigma^2}{2}}$	$e^{2\mu+\sigma^2}(e^{\sigma^2}-1)$

2.3.8 Limitation of traditional survival analysis

Although there are many estimation methods and models, few predict survival time and risk factors of individual patients. Moreover, it is not easy to choose a predictive method to directly model survival function and individual patient's survival distribution. One of the main challenges in this context is the presence of instances whose event outcomes become unobservable after a certain time point or in some instances do not experience any event during the monitoring period [69]. This is called censoring which will influence the accuracy of prediction results. Most of the present survival models focus on the population-based survival results analysis but not individual survival time prediction and risk prediction. In order to overcome this challenge, new technologies and methods need to be developed to determine the personalized survival prediction for precision medicine.

2.4 Machine learning method on survival analysis/prediction systems

2.4.1 What is machine learning

Machine learning is a scientific study of models that are relatively new to health research community. Machine learning is the process by which a computer performs a specific task without being explicitly programmed, relying on patterns and inference instead [62]. It is devoted to the methodology and algorithms for predictive modeling with the objective of making the better possible linear or non-linear predictions [70]. Machine learning was initiated in the late 1970s by Breiman et al. who proposed the methodology of Classification and Regression Trees and later proposed the methods of Bagging and Random Forests [71-73]. The Neural Networks, Boosting and Support Vector Machine were developed in the 1990s with the development of the computer science algorithms [74]. These methods can be used to study the relationship between an outcome variable and several features that could potentially affect the outcome.

Generally, machine learning models can be broadly categorized as two categories: supervised and unsupervised methods [74]. Supervised methods are used when there is a known outcome, and a model is being trained to predict the outcome. This requires at least a sample of data that includes the outcome, so a model can be trained to accurately predict. Unsupervised learning, on the other hand, does not necessarily have an outcome. Often, unsupervised learning methods are used to identify clusters in the features space, differentiating groups or clusters of observations based on a number of features.

This study will be looking at binary classification: where a CVD outcome is either present or absent. This is a supervised machine learning approach, and there are a variety of machine learning methods, ranging from simple to complex. The following summarizes

the common machine learning methods that have been implemented in the published literatures and the algorithms that will be used in this study.

2.4.2 Machine learning method on survival analysis and prediction systems

In the past several years machine learning techniques have achieved significant success in various areas. This is partly due to the advantages of machine learning techniques, such as its ability to model the non-linear relationships and the quality of overall predictions made. In survival analysis, the main challenge of machine learning methods is the difficulty to appropriately deal with censored information and the time estimation of the model. Machine learning method can optimize specific loss function or performance measures by modifying the common regression loss functions to handle censoring data for survival analysis in real-world data [75].

2.4.3 CoxPH Model

CoxPH model has been widely used to learn from uncensored survival data for survival analysis. It was introduced to model the hazard function of the i^{th} subject with the help of machine learning technology by fitting with each individual's feature vector. Generally, the form of the CoxPH model is shown as above when we talk about tradition survival analysis (Section 2.3.6). As the baseline hazard function $h_0(t)$ in CoxPH model is not specified, it is not possible to fit the model using the standard likelihood function. In other words, the hazard function $h_0(t)$ is nuisance function, while the coefficients β are the parameters of interest in the model. Cox proposed a partial likelihood which depends only on the parameter of interest β and is free of the nuisance parameters to estimate the

coefficients. The hazard function refers to the probability that a person with covariate vector X fails at time t on the condition that it survives until time t, which can be expressed by $h(t,\vec{x}_i)dt$ with $dt \to 0$. Let $J(J < N)$ be the total number of events of interest that occurred during the observation period for N persons, and $T_1 < T_2 < \cdots < T_J$ is the distinct ordered time to event of interest. Without considering the ties, let X_i^J be the corresponding covariate vector for the subject who fails at T_j and R_j be the set of risk subjects at T_j. Thus, conditional on the fact that the event occurs at T_j, the individual probability corresponding to covariate X_j can be formulated as

$$\frac{h(T_j, X_i^J)}{\sum_{i \in R_j} h(T_j, X_i^J)}$$

and the partial likelihood is the product of the probability of each subject; referring to the Cox assumption and the presence of the censoring, the partial likelihood is defined as

$$L(\beta) = \prod_{j=1}^{N} \left[\frac{\exp(X_i^J \beta)}{\sum_{i \in R_j} \exp(X_i^J \beta)} \right]^{\delta_j}$$

It should be noted that $j = 1, 2, 3 \cdots N$. The coefficient vector $\hat{\beta}$ is estimated by maximizing this partial likelihood, or equivalently, minimizing the negative log-partial likelihood to improve efficiency.

$$\log(L(\beta)) = -\sum_{j=1}^{N} \delta_j \left\{ X_i^J \beta - log \left[\sum_{i \in R_j} \exp(X_i^J \beta) \right] \right\}$$

The maximum partial likelihood estimator can be used along with numerical Newton-Raphson method to iteratively find an estimator $\hat{\beta}$ which minimizes the negative $\log\big(L(\beta)\big)$.

2.4.4 Multi-task machine learning method used in personalized risk prediction model development

Multi-task logistic regression is a recently developed method from the machine learning community to produce a patient-specific survival function, which has shown a satisfied result. Unlike the CoxPH model, MTLR does not have an explicit assumption or restriction about the hazard function or the shape of survival curves. Each patient has its own specific survival curves which can intersect with other patients' curves.

Consider a simpler classification task of predicting whether an individual will survive for more than t months. A common approach for this classification task is the logistic regression model, where the probability of surviving no less than t months is modeled as:

$$P_{\vec{\beta}}(T \geq t|\vec{x}) = \frac{1}{1 + \exp\left(\vec{\beta}\vec{x} + b\right)}$$

The parameter vector $\vec{\beta}$ describes the effect of how the features $\vec{x}$ affect the chance of survival. This task corresponds to a specific time point on the K-M curve, which attempts to discriminate those who survive against those who have died, based on the features $\vec{x}$.

MTLR first discretizes the continuous time axis into M time points $\tau = (t_1, t_2, \dots, t_m)$, for example, τ could be 10 year intervals from 1 year up to 10 years. We

can then transform the survival function prediction task into a sequence of binary classification tasks, by constructing initially a logistic regression model for each time point.

$$P_{\vec{\beta}_i}(T \geq t_i | \vec{x}) = \frac{1}{1 + \exp(\vec{\beta}_i \vec{x} + b_i)} \; 1 \leq i \leq m$$

where $\vec{\beta}_i$ and b_i are time-specific parameter vector and thresholds. The input features $\vec{x}$ stay the same for all these classification tasks, but the binary labels $y_i = [T \geq t_i]$ can change depending on the threshold t_i. This particular setup allows us to answer queries about the survival probability of individual patients at each of the time snapshots t_i, getting close to the study objective of modeling a personal survival time distribution for individual patients. The use of time-specific parameter vector naturally allows us to capture the effect of time-varying covariates, similar to many dynamic regression models.

However, the outputs of these logistic regression models are not independent, as a death event at or before time t_i implies death at all subsequent time points t_j for all $j > i$. MTLR enforces the dependency of the outputs by predicting the survival status of a patient at each of the time snapshots t_i jointly instead of independently. We encode the survival time *s* of a patient as a binary sequence $y = (y_1, y_2, \ldots, y_m)$, where $y_i \in \{0, 1\}$ denotes the survival status of the patient at time t_i, so that $y_i = 0$ (no outcome event yet) for all i with $t_i < s$, and $y_i = 1$ (outcome) for all i with $t_i \geq s$. We denote such an encoding of the survival time s as $y(s)$, and let $y_i(s)$ be the value at its ith position. Here there are $m + 1$ possible legal sequences of the form $(0, 0, \ldots, 1, 1, \ldots, 1, 1, \ldots, 0, 0)$, including the sequence of all '0's and the sequence of all '1's.

$$S = \underset{y_1}{0} \cdots \underset{y_n}{0} \underset{y_{n+1}}{1} \cdots \underset{y_m}{1} \underset{y_{m+1}}{0} \cdots \underset{y_z}{0} \cdots$$

As the unnormalized probability of **y**=[$y_1, \dots, y_z$] : $f_\Theta(x, \boldsymbol{y}) = \sum_{i=1}^{z} y_i \times (\vec{\beta}_i \vec{x} + b_i)$, so the probability of observing the survival status sequence $y = (y_1, y_2, \dots, y_m)$ can be represented by the following generalization of the logistic regression model:

$$f_\Theta(\vec{x}, S) = \sum_{i=1}^{n} 0 \times \left(\vec{\beta}_i \vec{x} + b_i\right) + \sum_{i=n+1}^{m} 1 \times \left(\vec{\beta}_i \vec{x} + b_i\right) + \sum_{i=m+1}^{z} 0 \times \left(\vec{\beta}_i \vec{x} + b_i\right) \dots$$

By introducing the Partition function: $\mathrm{Z}(x, \Theta) = \sum_{i=1}^{z} \exp\left(f_\Theta(x, S)\right)$ into the above equation, we have probability of S:

$$P_\Theta(\boldsymbol{Y}{=}(y_1,\dots,y_z) \mid \vec{x}) = \frac{\exp\left(f_\Theta(\vec{x}, S)\right)}{\mathrm{Z}(\vec{x}, \Theta)} = \frac{\exp\left(f_\Theta(\vec{x}, S)\right)}{\sum_{i=1}^{z} \exp\left(f_\Theta(\vec{x}, S)\right)}$$

$$= \frac{exp\left(\sum_{i=1}^{z} y_i \times (\vec{\beta}_i \vec{x} + b_i)\right)}{\sum_{i=1}^{z} exp\left(\sum_{i=n+1}^{m} 1 \times \left(\vec{\beta}_i \vec{x} + b_i\right) + \cdots\right)}$$

where $\Theta = (\vec{\theta}_1, \dots, \vec{\theta}_z)$ is the vector of estimated β for each patient, and $f_\Theta(\vec{x}, z) = \sum_{i=z+1}^{m} \left(\vec{\beta}_i \vec{x} + b_i\right)$ for $0 \le z \le m$ is the score of the sequence with the event occurring in the interval (t_z, t_{z+1}) before taking the logistic transform, with the boundary case $f_\Theta(\vec{x}, m) = 0$ being the score for the sequence of all '0's. This is similar to the objective of conditional random fields for sequence labeling, where the labels at each node are scored and predicted jointly.

Therefore the log likelihood of a set of uncensored patients with survival time s_1, $s_2 \dots, y_n$and feature vectors $\vec{x}_1, \vec{x}_2 \dots, \vec{x}_n$ is:

$$\sum_{i=1}^{N} \left[\sum_{j \in D_i} y_i(s_i) \times \left(\vec{\beta}_i \vec{x} + b_i\right) - \log \sum_{j \in R_i} exp f_\Theta(\vec{x}, z) \right]$$

where R_i is the risk set at t_z, which consists of all patients whose observed times are equal to or greater than t_z, D_i contains all patients whose failure time is t_z.

Instead of directly maximizing this log likelihood, we solve the following optimization problem:

$$\min_{\Theta} \frac{C_1}{2}\sum_{j=1}^{Z} ||\vec{\beta}||^2 + \frac{C_2}{2}\sum_{j=1}^{Z-1} ||\vec{\beta}_{j+1} - \vec{\beta}_j||^2$$

$$- \sum_{i=1}^{z}\left[\sum_{j=1}^{Z} y_i(s_i) \times \left(\vec{\beta}_i\vec{x} + b_i\right)\text{-log}\sum_{k=0}^{m} exp\, f_\Theta(\vec{x}, z)\right]$$

The first regularizer over $||\vec{\beta}||^2$ is designed to reduce the chance of overfitting. The second regularizer $||\vec{\beta}_{j+1} - \vec{\beta}_j||^2$ is designed to smooth the prediction and control the model capacity. The regularization constants C_1 and C_2, which control the amount of smoothing for the model, can be estimated via cross-validation. As the above optimization problem is convex and differentiable, optimization algorithms such as Newton's method or quasi-Newton methods can be applied to solve it efficiently [76]. Since we model the survival distribution as a series of dependent prediction tasks, we call this model MTLT.

2.4.5 NMTLR model

Although the MTLR model provides similar results as the CoxPH model without a rely on the proportional hazard assumption, it is still powered by a linear transformation. Thus, in the presence of nonlinear elements in the data, it will stop yielding satisfactory performance. The NMTLR which allows the use of Neural Networks within the original MTLR design will help solve this issue. Neural networks was inspired by biological neural systems several decades ago [77]. In this approach, the simple artificial nodes denoted by "neurons" are connected based on a weighted link to form a network which simulates a

biological neural network. A neuron in this context is a computing element which consists of sets of adaptive weights and generates the output based on a certain kind of activation function. The general structure of neural networks is illustrated in Figure 2.4 [78]. Each neuron in the input layer represents a feature in the dataset. Hidden layers represent the non-linear relationship between inputs and outputs. Both the number of hidden layers and the number of neurons in each hidden layer require experiments to optimize. Neurons in the output layer represent predicted outcomes.

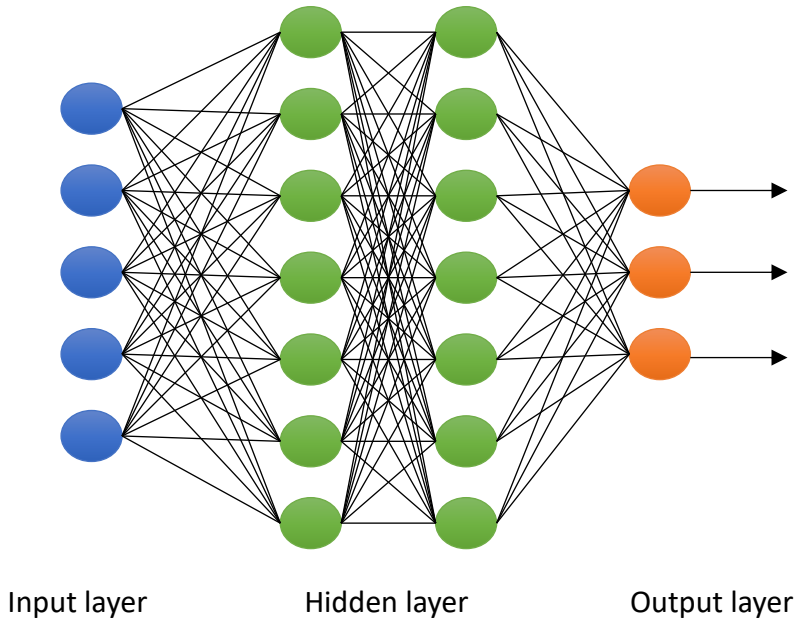


Figure 2.4 General structure of neural networks with two hidden layers transformation.

Both hidden and output layers in Figure 2.4 have activation functions. Each neuron takes the weighted sum from previous outputs, transfers it through an activation function, and outputs this value to its downstream neuron. *Sigmoid* function is commonly used to predict probabilities because outputs are in (0, 1) as illustrated in Figure 2.5 [79]. Other

common activation functions include *ReLu, softmax, tanh* etc., and experiments are required to find the suitable activation function for each hidden layer and output layer.

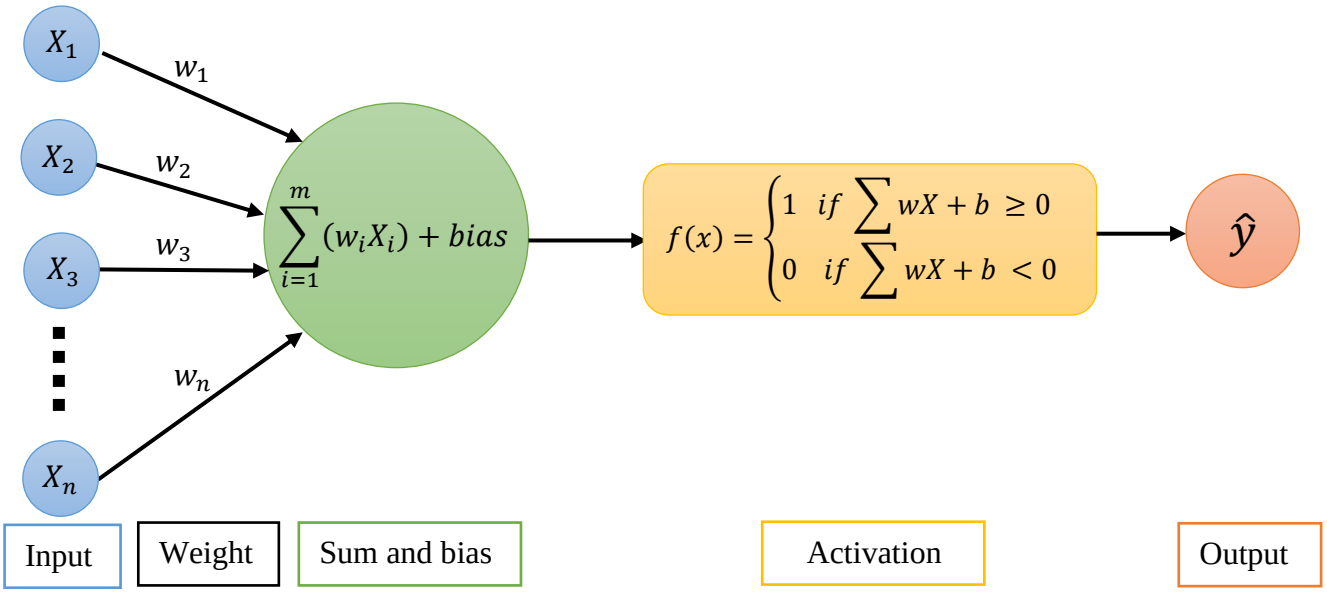


Figure 2.5 Activation function.

Minimizing the loss function is the objective of a neural networks. A common loss function in regression problems is the MSE, which minimizes the mean of the square difference between predicted and actual outcomes. Other common loss functions include mean absolute error, cross entropy function and so on. However, for survival analysis, these loss functions may require some asymmetric modifications. All neurons are fully connected with weights. All neurons are fully connected with weights. Back-propagation algorithm makes neural networks adjust weights and learn the non-linear pattern between inputs and outputs. The algorithm works as follows:

1. Initialize all weights to some random numbers.

2. Repeat until certain termination condition is met. One forward and backward propagation is called one epoch.

(a) Propagate inputs forward through the network and compute outputs.

(b) Use optimizer to minimize the loss function with respect to each weight ω_{ji} and update weight $\omega_{ji} \leftarrow \omega_{ji} + \Delta\omega_{ji}$ after each epoch, where $\Delta\omega_{ji}$ is determined by the optimizer. Common optimizers include Stochastic gradient descent, Adam etc.

Again, the most suitable optimizer is determined by the experiment. Within the optimizer, hyperparameters like learning rate, decay rate, momentum and so on, can be tuned for specific tasks as well.

2.4.6 RSF model

Random forest is an ensemble method specifically proposed to make predictions using the tree structured models [80]. It is based on a framework similar to Bagging; the main difference between random forest and bagging is that, at a certain node, rather than using all the attributes, random forest only uses a random subset of the residual attributes to select the attributes based on the splitting criterion. It is shown that randomization can reduce the correlation among the trees and thus improve the prediction performance. RSF extends the random forest method by using a forest of survival trees for prediction [81]. There are mainly four steps in RSF:

1. Draw B bootstrap samples randomly from the given dataset. This is also called out-of-bag (OOB) data because around 37% of the data is excluded in each sample.

2. For each sample, build a survival tree by randomly selecting features and split the node using the candidate feature which can maximize the survival difference between the child nodes.
3. Build the tree to the full size with a constraint that the terminal node has greater than or equal to a specific number of unique deaths.
4. Using the non-parametric Nelson-Aalen estimator, calculate the ensemble cumulative hazard function (CHF) of OOB data by taking the average of the CHF of each tree.

In RSF, survival trees are grown just as they would for classification and regression trees in random forests. The process begins at the root node, the top of the tree comprising of all data in the bootstrap sample. From the bootstrap sample, p predictor variables are chosen at random and used to split the root node, according to a predetermined survival criterion, into two daughter nodes. For each of the two daughter nodes, another set of p predictor variables are chosen at random and used to split each of the two nodes into two additional daughter nodes. This process is repeated recursively until each node has, at minimum, one unique event. This process groups together (in nodes) observations that are similar in the values of their predictor variables, and 'pushes apart' observations that are different.

2.4.7 Model performance and validation

Evaluating statistical performance of survival prediction model allow us to understand how well our predicted outcomes measure up relative to actual patient outcomes. Due to the presence of the censoring in survival data, the standard evaluation

metrics for regression such as RMSE are not suitable for measuring the performance in survival analysis [82].

The concordance index (C-index), known as the C-statistic, is a standard metric for evaluating the correctness of a survival distribution by measuring the proportion of all comparable patient pairs in which the predicted and actual survival time are ranked in the correct order [83]. As a result, it has the ability to discriminate individuals who experience and those who don't experience a dichotomous outcome such as CVD (yes/no). For example, models with excellent discrimination should assign higher probability of mortality to those patients who actually died. The mode used in this study to predict whether or not a hypertensive patient can survive (in our study, survive means the patient does not have CVD) until a specific time t, and the model predicts the probability that this hypertensive patient will not experience a CVD outcome before time t. In order to get the concordance index, we need to know the number of concordant pairs among all pairs of comparable patients. Each member in a pair is assigned a probability of CVD outcome. Next, model probabilities are compared to assess concordance with actual observed outcomes. Within each pair a better model should assign a higher probability of CVD outcome to patients experiencing the outcome (concordant pair).

Consider both the observations and prediction values of two instances, $(y_1, \hat{y}_1)$ and $(y_2, \hat{y}_2)$, where y_i and $\hat{y}_i$ represent the actual observation time and the predicted value for patient one and patient two, respectively. The concordance probability between them can be computed as

$$c - index = \Pr(\hat{y}_1 > \hat{y}_2 | y_1 \geq y_2).$$

The C-statistic ranges between 0.5 and 1, where a value close to 0.5 indicates that the model predicts death no better than chance while 1 is perfect prediction. In survival analysis the $C-index$ is normally between 0.6 and 0.7 [84]. The main advantage of the $C-index$ is that is works for any type of predicted outcomes. One disadvantage of $C-index$ is that it only measures the relative difference, instead of the actual difference between predicted and actual outcomes.

The model calibration, related to goodness of fit, is commonly assessed by the Brier score. This score is used to evaluate the accuracy of a predicted survival function at a given time t; it represents the average squared distances between the observed survival probability and the predicted survival probability and is always a number between 0 and 1, with 0 being the possible value. When we consider the binary outcome prediction with a sample of N population and for each individual patients i $(i = 1, 2, \dots, N)$, the predicted outcome at t is $\hat{y}_i(t)$ for individual i, and the actual outcome is $y_i(t)$; then, the empirical definition of the Brier score at the specific time t can be given by

$$BS(t) = \frac{1}{N}\sum_{i=1}^{N}[\hat{y}_i(t) - y_i(t)]^2$$

where the actual survival probability $\boldsymbol{y_i(t)}$ for any individual can only be 0 or 1 [85].

2.5 Summary

The CVD survival prediction model for hypertensive patients is a dynamic area in epidemiological research. A challenge to healthcare policymakers is how to develop "front-end" tools based on these survival prediction models that can be integrated successfully into routine clinical care. Survival estimation and survival analysis are very important

methods in hypertensive patients' prevention as they can be used to accelerate lifestyle changes or adopt appropriate therapeutic interventions, or both. Moreover, commonly used linear logistic regression methods are inadequate for survival prediction since they could not handle multicollinearity and non-linear functional relationships. Therefore, it is necessary to conduct further methodological research in the fields of biostatistics and epidemiology to accurately estimate the survival probability of CVD among patients with hypertension. Machine learning may support a more accurate and efficient survival prediction model for special designed population, such as hypertension patients in this study.

Chapter Three: **Research method**

3.1 Study Design and Setting

This is a longitudinal retrospective population-based cohort of hypertensive adults in Alberta, Canada (population 4.2 million) who have also been hospitalized for one or more of the following conditions: HF, stroke and MI. Alberta has a publicly funded healthcare system that provides coverage of all necessary medical services delivered in hospitals, emergency departments and physician offices for eligible Alberta residents.

3.2 Data sources and linkage

Six administrative databases were used in this study: (1) DAD; (2) physician billing claims; (3) AHCIP registry; (4) Vital statistics; (5) Alberta PIN; and (6) Alberta clinical measurements for lab data. DAD records and abstracts all clinical visit information for all patients discharged from Alberta hospitals. It includes up to 16 clinical diagnosis codes in each abstract using International Classification of Disease (ICD), 9th revision (ICD-9), ICD-9 Clinical Modification (ICD-9-CM) and up to 25 clinical diagnosis codes in ICD-10th Revision Canadian Modification (ICD-10-CA). The Physician Claims dataset is collected as part of a routinely submitted fee-for-service physician's remuneration system to the AHCIP. The data collected includes information on provider, patients and services. Providers submit, with at least one, but up to three, ICD-9 diagnosis codes for both outpatient as well as inpatient hospital services. Alberta physicians submitted their services for Albertan residents who are registered under the AHCIP. The AHCIP contains demographic information including personal health number, migration indicators, date of birth and death, sex and postal code information for all Albertans with AHCIP coverage.

The fiscal year end file in AHCIP is considered to be the most accurate snapshot of each registrant as of March 31. This file is useful for consistently reporting recipient date of birth, sex, and postal code at fiscal year-end (rather than using service event data which may vary). Vital Statistics captures all live births delivered and death occurring within Alberta. PIN data records each time a prescription is dispensed to a patient. PIN is a web-enabled application within the Alberta Netcare EHR that provides access to a patient's active and previous dispensed medications. Alberta clinical measurements for lab data is used within Alberta to place laboratory orders, track workflow and report laboratory and pathology results. Laboratory data includes clinical chemistry, toxicology, hematology, serology, urinalysis, and immunology. All six datasets were linked through a unique and anonymous identifier, as shown in Figure 3.1.

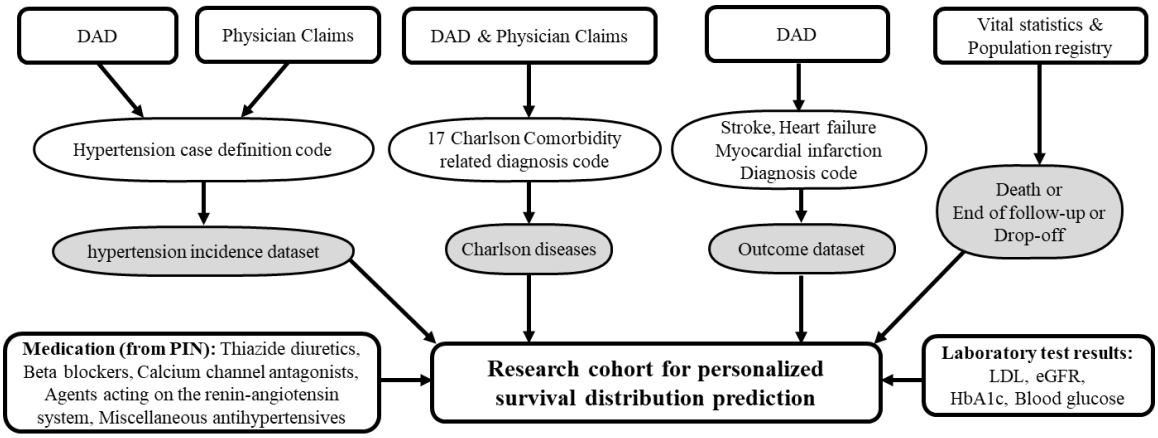


Figure 3.1 Linkage of six administrative health datasets to determine the study cohort.

Note: DAD = Discharge Abstract Database, PIN= Pharmaceutical Information Network, LDL = Low-density lipoproteins, eGFR = Estimated glomerular filtration rate, HbA1C = Glycated haemoglobin (A1c)

3.3 Study population

The study population included all newly diagnosed hypertension patients between 20 and 99 years old who were residents of Alberta. We identified hypertension cases using a validated case definition of two physician claims within two years or one hospitalization with hypertension related diagnosis codes (ICD-9-CM: 401.x, 402.x, 403.x, 404.x or 405.x; ICD-10-CA: I10.x, I11.x, I12.x, I13.x or I15.x) in the 25 coding fields for diagnosis in the hospital discharge abstract data, and 3 fields in physician claims data [8]. This case definition has a sensitivity of 75%, specificity 94%, positive predictive value 81%, and negative predictive value 92%. The first date of the hypertension diagnosis (index date) was assigned to patients for case definitions with more than one hypertension diagnosis. We excluded patients with an index date between April 1, 2006 and March 31, 2009 (three years washout period), thereby focusing on patients newly diagnosed with hypertension between April 1, 2009 and March 31, 2014. Patients were also excluded from the study if they met any of the following criteria: (1) hypertensive patients with MI, HF or stroke during the 3 years before the study or within 30 days following the diagnosis of hypertension diagnosis index date; (2) patients who have the same death date and hypertension diagnosis date; (3) patients whose follow-up period was less than 1 year; (4) patients younger than 20 or older than 99 years old; (5) patients diagnosed with hypertension in the washout period (between April 1, 2006 and March 31, 2009); (6) patients who were pregnant. Figure 3.2 displays the flow of how the study cohort was identified, and how the training and testing datasets were determined as described below in Section 3.5.3.

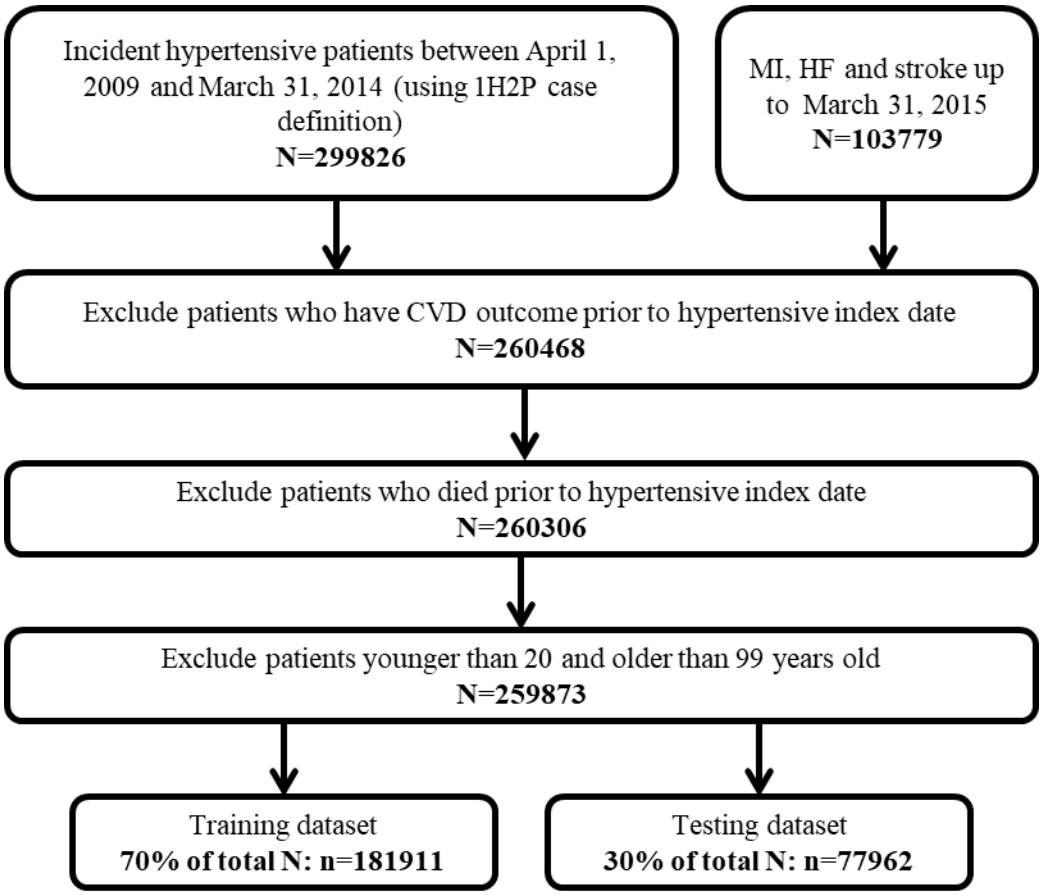


Figure 3.2 The flow for hypertension cohort identification.

Note: IH2P = 1 hospitalization or 2 physician claims within 2 years, CXD = Cardiovascular disease, MI = myocardial infarction, HF = heart failure.

3.4 Study Variables

3.4.1 Dependent variable

The outcome variable was composite CVD between April 1, 2009 and March 31, 2015 (6-year follow-up period). The composite outcome included MI, HF, and/or stroke. We used validated case definitions to define MI (ICD-10-CA: I21.x, I22.x, I25.2), HF (ICD-10-CA: I09.9, I11.0, I13.0, I13.2, I25.5, I42.0, I42.5-I42.9, I43.x, I50.x, P29.0), and

stroke (ICD-10-CA: H34.1, I60.x, I61.x, I63.x, I64.x, G45.x) and applied the ICD codes to all 25 diagnosis field in DAD [8].

The CVD outcome date was defined as the earliest date of HF, MI or stroke. We followed patients from diagnosis until CVD event(s), death or the end of the study, up to March 31, 2015. Deaths were identified from both Vital Statistics and AHCIP.

3.4.2 Independent variable

Age, sex, region of residence, fasting blood glucose, eGFR, cholesterol level, HbA1c, prescription medication dispensed related to hypertension treatment and Charlson comorbidities were investigated as potential risk factors for personalized CVD prediction model for hypertension patients. Age and sex at the hypertension index date were captured from AHCIP data. Census data was used to define rural and urban residence of patients. Charlson comorbidities were obtained from DAD and physician claims databases [86]. Patients with pre-existing MI, congestive heart failure, cerebrovascular disease were excluded from this cohort. AIDS/HIV at the time of their hypertension diagnosis were also excluded. Patient's comorbidity date was determined as the earliest date of diagnosis obtained from DAD or physician claims data using the 3-year period up to and including the hypertension index date. The number of Charlson comorbidities present in each patient was categorized into "0, 1-2, or $\geq$ 3." Absence of the Charlson comorbidities including "missing" was assigned "0." Measurements of fasting blood glucose, eGFR, cholesterol levels and HbA1c between hypertension index date and outcome date was determined through laboratory information systems. These measurements represent the most commonly recommended tests for the initial evaluation of newly diagnosed hypertension

by major clinical practice guidelines [59]. We used test results that were defined by values outside the standardized reference intervals as set by the laboratory and are most likely to affect clinical decision-making to define the abnormal test results (Blood glucose ≥ 7.0 mmol/L, eGFR <60 mL/min/1.73m^2, Cholesterol levels >3.5 mmol/L, HbA1c ≥ 6.5%) [59].

Patient medications were recorded from the ATC classification system through PIN dataset. The following categories of antihypertensive medications were specified: beta blockers (ATC codes in category C07, excluding C07AA07, C07AA12 and C07AG02); agents acting on the renin-angiotensin system (ATC codes in category C09); thiazide diuretics (ATC codes in category C03, excluding C03BA08 and C03CA01); calcium channel antagonists (ATC codes in category C08); and miscellaneous antihypertensives (ATC codes in category C02, excluding C02KX01) [87]. These medications have been shown to improve cardiovascular risk in clinical trials [88]. Respondents were categorized as using antihypertensive medication if an ATC code corresponded to the above list between the hypertension index date and outcome date.

3.5 Statistical Analysis

Analyses were conducted using SAS version 9.4, R software version 3.5.1 and Python version 3.7.6. R package “survival” and Python package “PySurvival” was used in this study. Characteristic results are present with 95% confidence interval (CI).

3.5.1 Descriptive Statistics

Descriptive statistics to summarize data on age, sex, region of residence, Charlson comorbidities, lab test results and dispensed prescriptions were used in this study to characterise newly diagnosed adults with hypertension. K-M method was used to analyse the survival data. Survival time was calculated from the hypertension index date to the date on which a patient was censored (lost follow up during the study, death or study end March 31, 2015) or occurrence of a CVD outcome. When a patient experienced more than one event of the same or different outcomes, then only the first event was counted as the incident event. This event date was considered as the CVD outcome date. Outcome event rates were calculated per 1000 person-years based on a maximum 6-year follow-up period.

3.5.2 Univariate Analysis

Univariate analysis was carried out using the K-M survival estimates to determine the difference between the composite CVD outcome and each individual CVD outcomes.

3.5.3 Machine learning models development

The study cohort was randomly divided into a training (70% of total: n=181911) and a testing (30% of total: n=77962) subset (as shown in Figure 3.2). The training dataset was used to develop the prediction models. The testing dataset was used to evaluate the performance of the models. All identified demographic and clinical characteristics (age, sex, region of residence, Charlson diseases, laboratory results and dispensed prescriptions) were initially included in the four machine learning models (LMTLR, NMTLR, RSF, CoxPH). Our primary goal in this study was to assess the performance of those four

machine learning models on survival prediction for hypertension patients with CVD outcomes and propose a CVD survival risk prediction model to improve prediction accuracy and performance of the model.

3.5.4 Construction of the machine learning models

The LMTLR model is an alternative to the Cox's proportional hazard model. It can be regarded as a series of logistic regression models built on different time intervals so as to estimate the probability that the event of interest occurring within each interval. The constructed LMTLR includes 25 features and 50 intervals in this study. The NMTLR allows the use of Neural Networks within the original MTLR design. We used the same 25 features, 100 neurons in the first hidden layer and 100 neurons in the second hidden layer, and one output neuron before input to LMTLR. The RSF is an extension of the Random Forest model that can take into account censoring individuals. For model development, we used 50 trees, with the maximum depth of 5 trees, and minimum number of leaves to be 20 leaf nodes. CoxPH is a semi-parametric model, which focuses on modeling the hazard function by assuming that its time component and feature component are proportional over time. The maximum number of iterations in the Newton optimization in this model was 600.

3.5.5 Model testing and validation.

The final survival prediction model was tested within the test dataset for those four models (LMTLR, NMTLR, RSF, CoxPH). The actual and predicted number of patients that experienced a CVD event at each time t was compared by computing the actual

survival function of the testing data, which can be obtained using the K-M estimator and compare it to the average of all predicted survival functions. RMSE and MAE was used to provide the comparison as well as the performance metrics between the actual and predicted number of hypertensive patients experiencing a CVD event at each time t.

Model accuracy was evaluated using discrimination and calibration. Discrimination was assessed using concordance index or C-index. The $C - index$ is a generalization of the area under the ROC curve (AUC) which takes into account censored data and thus can measure how well the model performs in discriminating between those who survive and those who don't, and evaluating the probability of correct prediction who of two randomly selected patients will first suffer CVD. It represents the global assessment of the model discrimination power to correctly provide a reliable ranking of the survival times based on the individual actual survival time. It ranges from 0.5 (chance) to 1 (perfect prediction). Calibration, related to goodness of fit, was assessed by the Brier score, which is defined as the mean squared difference between the predicted probability and the actual outcome in the overall follow-up period. Brier scores vary between 0 and 1, with a lower Brier score indicating a better calibrated prediction.

3.6 Ethics Approval

The dataset used in this project is de-identified and anonymized in electronic format at two different levels to ensure that the provincial confidentiality requirements are met. The available HOST database is stored on a secure, password-protected computer that can only permit to the research team. The Institutional Ethics Review Board of the University of Calgary approved this study.

Chapter Four: **Results**

4.1 Cohort characteristics

We identified 299,826 newly diagnosed Alberta hypertensive patients between April 1, 2009 and March 31, 2014. There was a total of 259,873 patients, 899,393 person-years follow-up, with 11,863 CVD events remaining (after applying exclusion criteria) during the median follow-up time of 3.5 years (inter-quartile range 2.2 to 4.8 years) until March 31, 2015. Most of patients lived in urban areas (83.6%) and 33.6% had at least one non-cardiovascular Charlson comorbidities. 60.8% of the patients had taken at least one lab test and 80.7% have received a least one medication treatment. The median age was 56.1 years, with 26.7% older than 65 years of age. These results are very close with Quan et al., who used similar methods to identify the incident hypertensive study cohort. [8] Individual Charlson comorbidities, such as chronic pulmonary disease, diabetes without chronic complications, diabetes with chronic complications, renal disease and any malignancy have proportion than 2% in our cohort. Abnormal cholesterol levels (higher than 3.5 mmol/L) had the highest rate (33.7%, 95% CI:33.5-33.9) among all laboratory tests examined. Over 67.3% of hypertensive patients (CI: 67.1-67.5) were dispensed medication acting on the renin-angiotensin system (see Table 4.1).

Table 4.1 Population characteristics that were prespecified and included in the predictor models using administrative health data.

Characteristics	**No. (%), Mean or Median (IQR)**	
	N=259873 (899392.8 person-years)	Percentage (95% CI)
Age in years (Mean±SD)	56.6 ± 14.0	
Median (Q1, Q3)	56.1 (47.2, 65.8)	

Age groups (years)		
20-49	84032	32.3 (32.2-32.5)
50-64	106526	41.0 (40.8-41.2)
65-74	42107	16.2 (16.1-16.4)
75-99	27208	10.5 (10.4-10.6)
Sex		
Female	122233	47.0 (46.8-47.2)
Male	137640	53.0 (52.8-53.2)
Region of residence		
Rural	42669	16.4 (16.3-16.6)
Urban	217204	83.6 (83.4-83.7)
Number of Charlson Comorbidities		
0	175227	67.4 (67.3-67.6)
1-2	69443	26.7 (26.6-26.9)
≥3	15203	5.9 (5.8-5.9)
Charlson comorbidities		
Peripheral vascular disease	4721	1.8 (1.8-1.9)
Dementia	5008	1.9 (1.9-2.0)
Chronic pulmonary disease	38808	14.9 (14.8-15.1)
Rheumatologic disease	4849	1.9 (1.8-1.9)
Peptic ulcer disease	4406	1.7 (1.7-1.8)
Mild liver disease	5010	1.9 (1.9-2.0)
Diabetes without chronic complications	20065	7.7 (7.6-7.8)
Diabetes with chronic complications	5247	2.0 (2.0-2.1)
Hemiplegia or paraplegia	525	0.2 (0.2-0.2)
Renal disease	5344	2.1 (2.0-2.1)
Any malignancy, including leukemia and lymphoma	17208	6.6 (6.5-6.7)
Moderate or severe liver disease	583	0.2 (0.2-0.2)
Metastatic solid tumor	2294	0.9 (0.9-0.9)
Lab test results		
LDL (>3.5 mmol/L)	87633	33.7 (33.5-33.9)
Blood glucose (≥7.0 mmol/L)	53337	20.5 (20.4-20.7)
eGFR (<60 mL/min/1.73m^2)	63697	24.5 (24.4-24.7)
HbA1c (≥6.5%)	36016	13.9 (13.7-14.0)
At least one lab test	157934	60.8 (60.6-61.0)
Medications		
Thiazide diuretics	64248	24.7 (24.6-24.9)
Beta blockers	47089	18.1 (18.0-18.3)
Calcium channel antagonists	62112	23.9 (23.7-24.1)
Agents acting on the renin-angiotensin system	174891	67.3 (67.1-67.5)
Miscellaneous antihypertensives	5245	2.0 (2.0-2.1)
≥ 1 of the medications listed above	209729	80.7 (80.6-80.9)

Follow-up years (median (Q1, Q3))	3.6 (2.2, 4.8)

Note: IRQ = Interquartile range, SD = Standard deviation, Q1 = First quartiles, Q3 = Third quartiles, CI = Confidence interval, Number of Charlson comorbidities: excluding CVD related comorbidities and HIV/AIDS.

The crude incidence of composite CVD hospitalization in this cohort without previous CVD was 13.2 per 1000 person-year: 6.1 per 1000 person-year for MI, 5.6 per 1000 person-year for HF and 3.4 per 1000 person-year for stroke, as shown in Table 4.2. The composite CVD hospitalization rate was higher for male, older patients, as well as those who have a higher number of Charlson comorbidities or resided in rural location. MI has the highest overall incidence rate with respect to HF and stroke.

Table 4.2 Hospitalization for incident cardiovascular disease (95% CI) among Albertan adults with newly diagnosed hypertension.

Outcomes	Cardiovascular diseases		Myocardial infarction		Heart failure		Stroke	
	n	Rate (95% CI)	n	Rate (95% CI)	n	Rate (95% CI)	n	Rate (95% CI)
Overall	11863	13.2 (13.0-13.4)	5566	6.1 (6.0-6.3)	5075	5.6 (5.4-5.7)	3105	3.4 (3.3-3.5)
Age groups								
20-49	1166	4.0 (3.8-4.2)	650	2.2 (2.1-2.4)	301	1.0 (0.9-1.1)	299	1.0 (0.9-1.1)
50-64	3713	9.9 (9.6-10.3)	2056	5.5 (5.2-5.7)	1204	3.2 (3.0-3.4)	918	2.4 (2.3-2.6)
65-74	2787	19.0 (18.3-19.8)	1349	9.1 (8.6-9.6)	1153	7.7 (7.3-8.2)	729	4.9 (4.5-5.2)
75-99	4197	48.5 (47.0-50.0)	1511	16.7 (15.9-17.6)	2417	26.9 (25.9-28.0)	1159	12.7 (12.0-13.5)
Sex								
Female	4989	11.6 (11.3-12.0)	1876	4.3 (4.1-4.5)	2438	5.6 (5.4-5.9)	1459	3.4 (3.2-3.5)
Male	6874	14.6 (14.3-15.0)	3690	7.8 (7.5-8.0)	2637	5.5 (5.3-5.7)	1646	3.4 (3.3-3.6)
Region of residence								
Rural	2347	15.7 (15.1-16.3)	1164	7.7 (7.3-8.1)	990	6.5 (6.1-6.9)	586	3.8 (3.5-4.2)
Urban	9516	12.7 (12.4-12.9)	4402	5.8 (5.6-6.0)	4085	5.4 (5.2-5.5)	2519	3.3 (3.2-3.4)
Number of Charlson comorbidities								
0	5909	9.6 (9.3-9.8)	2980	4.8 (4.6-5.0)	2055	3.3 (3.2-3.4)	1657	2.7 (2.5-2.8)
1	4264	18.0 (17.4-18.5)	1901	7.9 (7.5-8.3)	2023	8.4 (8.0-8.8)	1058	4.4 (4.1-4.6)
2	1690	36.8 (35.1-38.6)	685	14.5 (13.4-15.6)	997	21.2 (19.9-22.5)	390	8.1 (7.4-9.0)

4.2 Univariate K-M analysis

Figure. 4.1 shows K-M plots of the cumulative probability of being free of hospitalization composited CVD (red line), MI (purple line), HF (blue line) and stroke (green line) as a function of survival time among newly diagnosed hypertension patients. Compared with HF and stroke, MI had the lowest cumulative probability in the entire survival period .

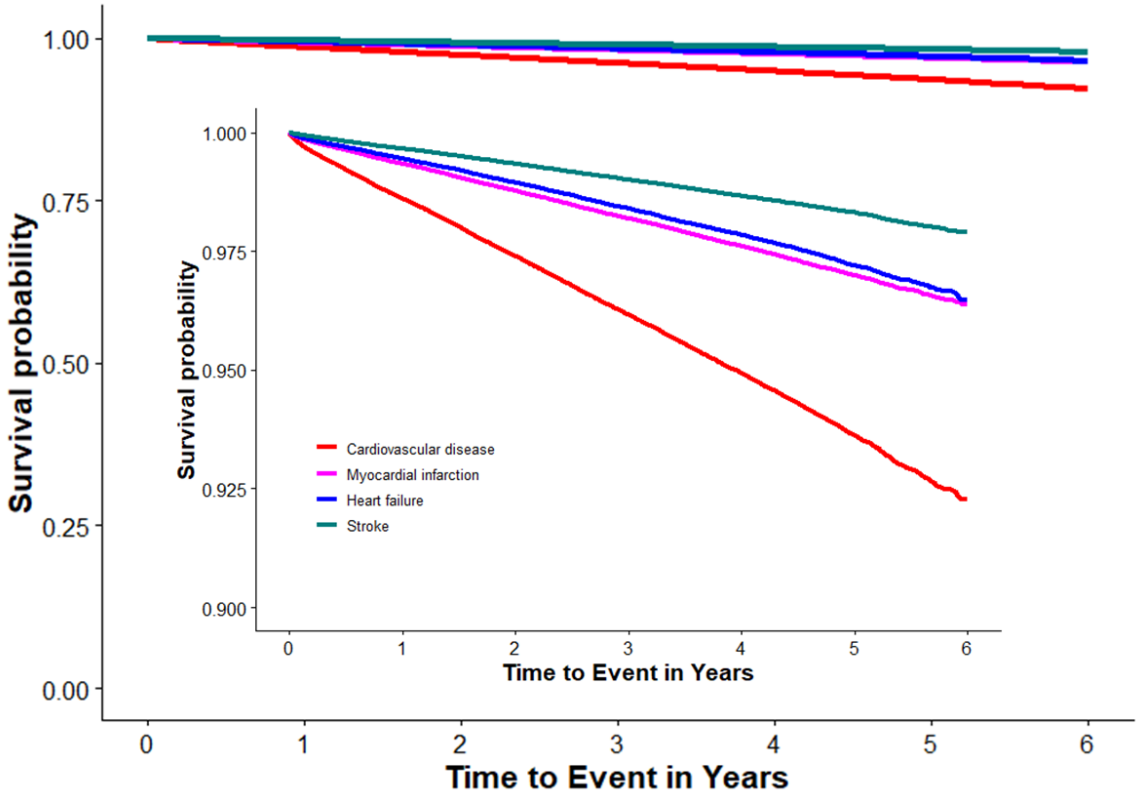


Figure 4.1 K-M estimate for incident hypertensive patients stratified by cardiovascular disease (any one of MI, HF and stroke), MI, HF and stroke.

4.3 Machine learning models performance

Table 4.3 shows a comparison of the actual and predicted number of hypertension patients experiencing CVD events at each time t for all four models. The RSF model had the lowest RMSE at 33.94 and the lowest MAE at 28.37, which means that the RSF model has the better performance for population-based prediction in our study.

Table 4.3 Comparison of predicting the number of incident hypertensive patients diagnosed with cardiovascular disease(s) using predicted survival functions in multiple models (LMTLR, NMTLR, RSF and CoxPH).

Model performance	**Models**			
	LMTLR	NMTLR	RSF	CoxPH
RMSE	508.92	143.49	**33.94**	58.55
MAE	383.63	132.54	**28.37**	49.80

Note: RMSE = root mean squared error, MAE = mean absolute error

By plotting the individual's survival function over time, we can compare the survival probability of developing CVD events among hypertensive patients. Table 4.4 shows the results when applying all four models in our hypertension cohort. Overall, the $c-index$ of all four models is higher than 0.7, while the Brier score is lower than 0.25, which indicates a strong predictive ability in the internal validation cohort. Due to the addition of two layers of neural networks before LMTLR, the NMTLR model has the highest $c-index$ of 0.8202 and lowest Brier score of 0.0243, which means that this model

has the better discrimination and calibration, and is more accurate and outperformed in predicting CVD outcomes for hypertensive patients.

Table 4.4 Summary of results measured by $C - index$ and Brier score.

Model performance	**Models**			
	LMTLR	NMTLR	RSF	Cox PH
C-index	0.7792	**0.8202**	0.8146	0.8165
Brier score	0.0350	**0.0243**	0.0343	0.0340

Chapter Five: **Discussion**

5.1 Model performance

In this study, we explored the performance of four different models in predicting the survival rate distribution of individual CVD outcomes among incident hypertension patients by using routinely collected large population-based health administrative data. Four different types of models were developed and validated by applying the same training dataset. All of the four models had high discrimination with $C - index$ higher than 0.7, and good calibration with Brier score less than 0.25.

The models were developed for both population and individual use. For population use, the model in this study could predict the number of the hypertension patients who have experienced a CVD event at each time point by using predicted survival functions. Compared with the other three models, the RSF model has the better performance for population-based prediction in our study. Moreover, in the RSF model, the RMSE and MAE are quite close with each other, which indicates that the prediction results have no large errors and remain constant during the whole 6 years follow-up period. Although we did not enforce the use of any particular variables, the RSF algorithm could find any variable most relevant to the risk of CVD outcomes in the administrative dataset, which makes this model more accurate than a regression model (LMTLR, NMTLR and CoxPH) for the population-based predictions.

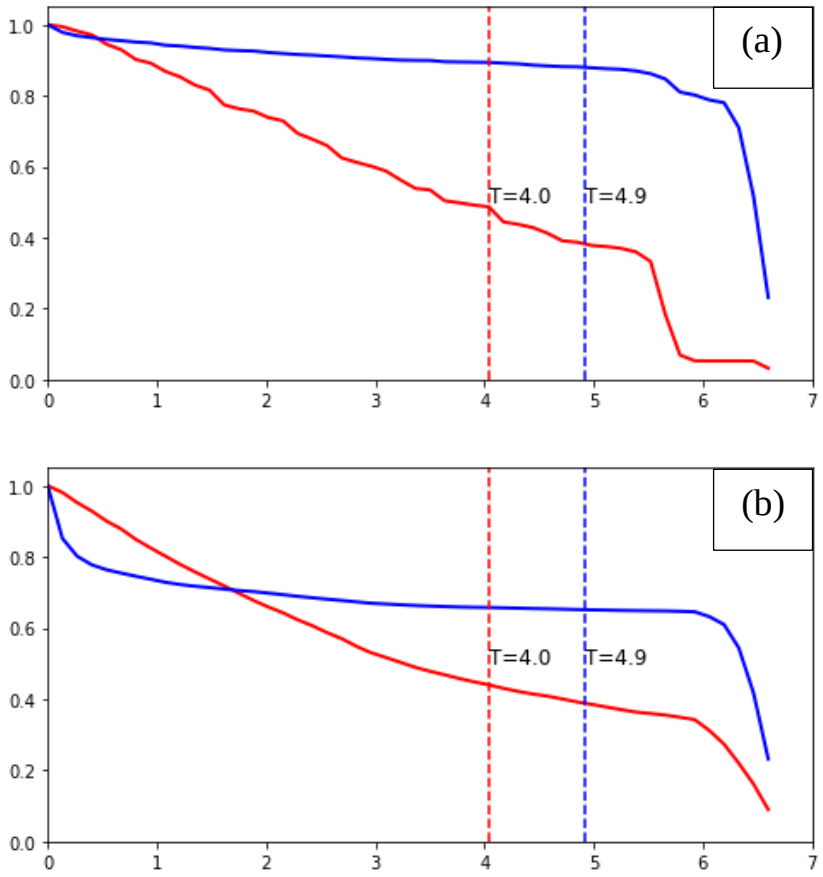
For individual use, the models can be used to predict individual CVD survival probability for hypertension patients. With the help of laboratory data and medication data in the models, clinicians and patients could utilize CVD treatment and management with more accurate interventions, which leads to precision medicine.

The NMTLR model with the highest $C - index$ and lowest Brier score means it has the better discrimination and calibration for individual survival prediction, which thus is more accurate and outperformed in predicting CVD outcomes for hypertensive patients. The performance of neural network depends on the number of parameters, network weights, the selection of an appropriate training algorithm, the type of transfer functions used, and the determination of network size. Therefore, neural networks require initialization and adjustment of many individual parameters to optimize the performance of the classification. The NMTLR model, which combines the neural network and MTLR together, was developed empirically and can better fit the training data in our study. NMTLR models the survival function by combining multiple local logistic regression modes in a dependent manner and following a two-layer neural networking procedure. It can naturally handle censored observations, thereby providing significantly better results. The combination of neural network and multiple task logistic regression in NMTLR allows the model to build a nonlinear statistical data modeling tool to deal with complex relationships and shows significantly better predictive performance than the other three models.

Figure 5.1 visualizes the LMTLR, NMTLR, RSF and CoxPH models for two patients from the test cohort. Patient one (red line) is a short survivor who lives for 4.0 years from diagnosis of CVD, while patient two (blue line) is a long survivor whose survival time is censored at 4.9 years. Patient one developed CVD after 4.0 years diagnosed as hypertension, however, patient two may have been lost to follow-up or did not develop CVD at the end of the study or death until 4.9 years after diagnosed as hypertension patient. All those four models did a good job on prediction the survival time for patient one, whose

50% survival probability corresponds to 4 years of survival. However, only NMTLR model provided the most accurate prediction for patient two who lost survival information at 4.9 years in this study. The CVD incidence rate is 13.4 per 1000 person-years in our cohort, which means we have a very high censoring rate (more than 90%). The results show that NMTLR has super ability to handle the censored individual and proves accurate prediction results. Moreover, the predicted survival curves of those two patients are intercrossed at the beginning of follow-up. This may reflect the real situation that patient two had worse health condition in the beginning, who however may be under hypertension treatment or prevention after one year of diagnose with hypertension. Patient two had a fairly flat curve in the following period, patient one however became worse in the whole follow-up period. The CoxPH model almost has no ability to handle the censored individual with a horizontal survival probability line for patient two. Figure 5.2 shows the prediction results for two patients who have CVD outcome at 1.1 years and 2.3 years, respectively. Hypertension patient one was diagnosed as CVD outcome at 1.1 survival years while hypertension patient two was diagnosed as CVD outcome at 2.3 survival years. All of those four models can discriminate the two patients' survival time versus survival probability. Patients one's survival curve was always lower than patient two but reached 50% survival probability faster than patient two. Except for NMTLR, all the other three models give poor predictions for both patients one and two. Only NMTLR can correctly predict the survival time for patient one at 1.1 years and patient two at 2.3 years at the 50% survival probability. Moreover, NMTLR also has the smoothest survival curves with different shapes for the two patients, while CoxPH model predicts survival curves of similar shapes because of the proportional hazard assumption. Actually, it is well-known that the survival curves of two

individuals never cross for a Cox model [89]. For RSF model, we observe that the survival function is monotonically decreasing and parallel. This is probably because both of these patients are in the same tree branch node during model development.



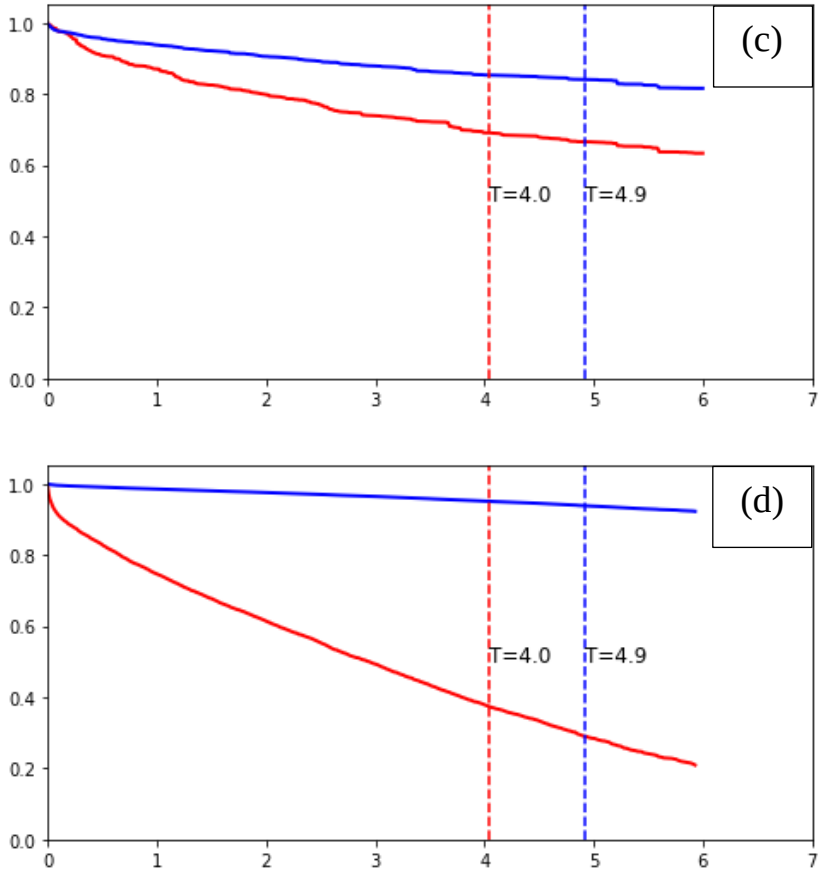
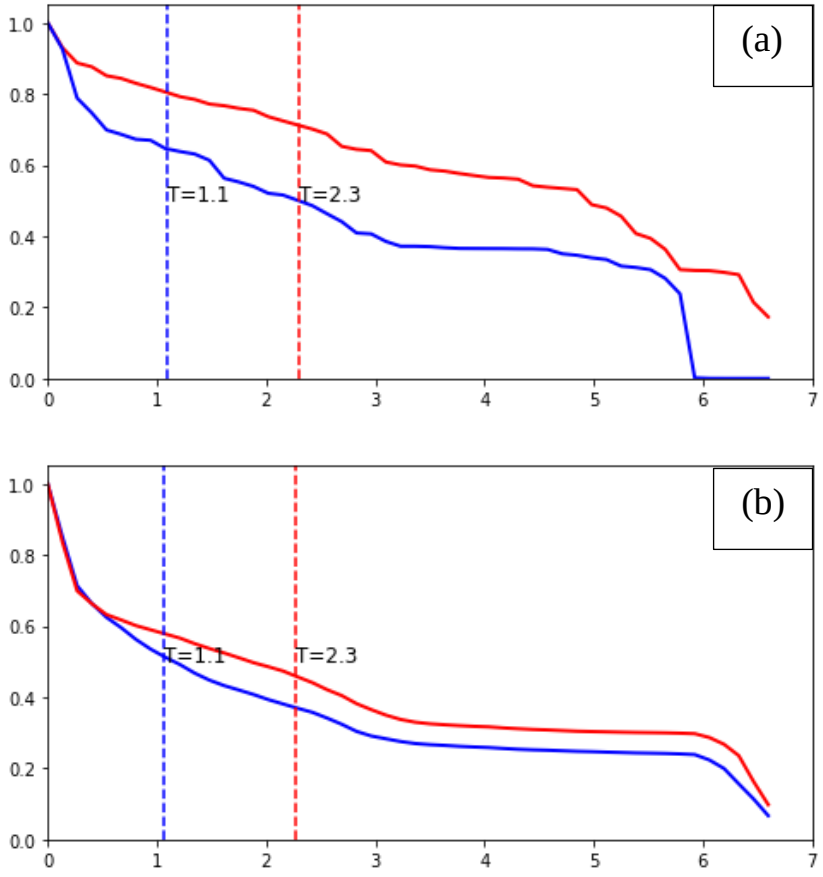


Figure 5.1 Personalized survival prediction by using (a) LMTLR model, (b) NMTLR model, (c) RSF model and (d) CoxPH model when randomly identifying two patients as an example.

Note: Patient one (red line) is a short survivor who lives for 4.0 years from diagnosis of hypertension to CVD, while patient two (blue line) is a long survivor whose survival time is censored at 4.9 years. Patient one developed CVD after 4.0 years diagnosed as hypertension, however, patient two may lost follow-up or did not develop to CVD at the end of the study or death until 4.9 years after diagnosed as hypertension patient. The bottom position of survival time text (4.0 and 4.9) on the pictures correspond to 50% survival probability horizontal location line. X axis = time in years, and Y axis = probability of event-free survival


(a)
T=1.1
T=2.3
(b)
T=1.1
T=2.3

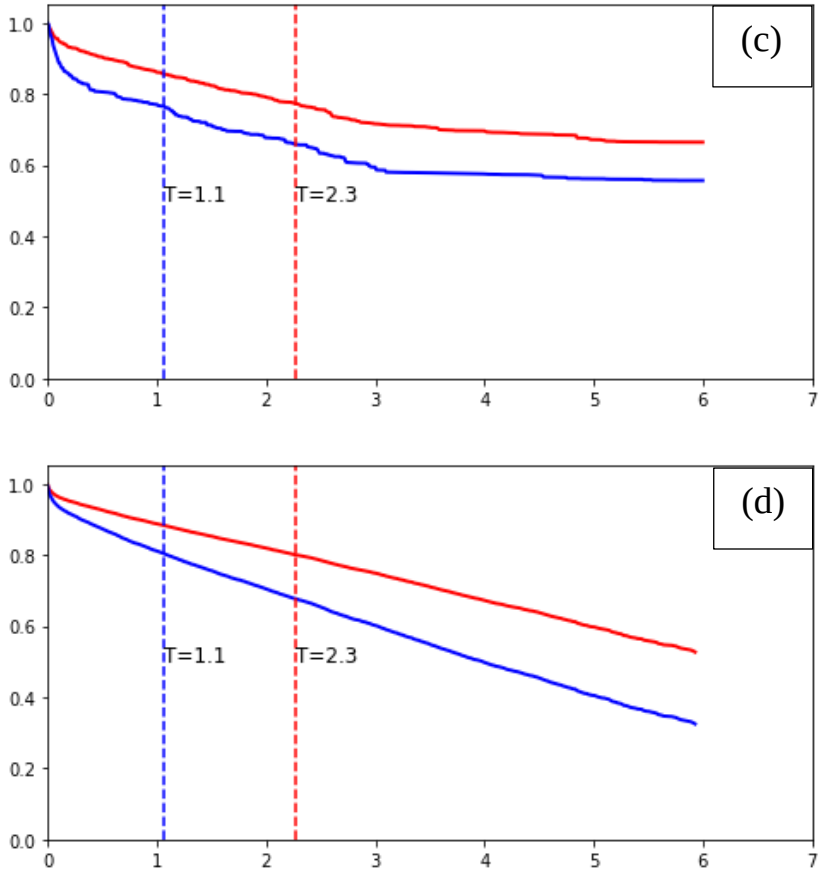


Figure. 5.2 Personalized survival prediction by using (a) LMTLR model, (b) NMTLR model, (c) RSF model and (d) CoxPH model when randomly choosing two patients who have CVD outcome at 1.1 years and 2.3 years as an example.

Note: Hypertension patient one (blue line) was diagnosed as CVD outcome at 1.1 survival years while hypertension patient two (red line) was diagnosed as CVD outcome at 2.3 survival years. The bottom position of survival time text (1.1 and 2.3) on the pictures correspond to 50% survival probability horizontal location line. X axis = time in years, and Y axis = probability of event-free survival.

5.2 Implication of models

Based on the above discussion, we know that the RSF model has the better performance on population-based survival prediction, while NMTLR has the better prediction for individual survival prediction when compared with other developed models. We can use RSF model to predict the number of events that will happen at a give survival time point. For example, if we want to know how many newly diagnosed hypertension patients will develop CVD outcomes at 1 year of follow-up period, then the RSF model is the top choice. The NMTLR model is the top one for individual survival prediction. NMTLR could be used to predict the personalized survival distribution if there is information available on patient's prediction variables, like age, sex/gender, comorbidities, etc. For example, if we know the personal characteristics of a newly diagnosed hypertensive patient, we could provide the patient with information about the 50% survival probability time to develop CVD outcomes in the future. This could help the clinician and the patient take timely action to treat and control hypertension before the situation worsens.

Yu et al., compared the MTLR and CoxPH models in different types of datasets [90]. The results showed that MTLR always have a better performance with a higher $C-index$ values. The results from Li et al., also indicated that MTLR model has better performance compared with other models in several high-dimension datasets [76]. The combination of MTLR and neural network in our study has shown better performance compared with the other three models. This means that NMTLR model has stable performance in a wide range of datasets and likely implemented in different parts of healthcare system.

The variables used in our models were: age, sex, region of residence, peripheral vascular disease, dementia, chronic pulmonary disease, rheumatologic disease, peptic ulcer disease, mild liver disease, diabetes without chronic complications, diabetes with chronic complications, hemiplegia or paraplegia, renal disease, any malignancy, moderate or severe liver disease, metastatic solid tumor, LDL, blood glucose, eGFR, HbA1c, thiazide diuretics, beta blockers, calcium channel antagonists, agents acting on the renin-angiotensin system, and miscellaneous antihypertensives. A hypertensive patient could input their predictor variables as listed above to access the survival probability of having CVD. Although most of the variables included in our models were non-modifiable variables, which means they cannot be changed once patients have this condition. However, the model itself could also provide significant guidance for hypertensive patients or clinicians for their CVD disease evaluation and management based on patients' condition. For example, diabetes is a prediction factor in our models, eating a balanced diet, regular exercise, and leading a generally healthy lifestyle can both help manage diabetes in those who already have the condition and help prevent the onset of the condition in those that don't. In people with diabetes, careful management of blood sugar levels is also very important in helping to reduce the risk of cardiovascular disease. Moreover, our models, for example NMTLR model, is a dynamic model, which means patient can input their predictors in the model and make comparison of the before and after of having a condition on their survival probability of CVD. This could be very helpful for them to know their risk of having CVD and take timely action to prevent negative outcomes.

5.3 Strengths and limitations

To the best of our knowledge, this is the first study to develop and validate models for individual CVD outcomes among newly diagnosed hypertension patients using administrative data and assess model performance within the same testing groups. The developed models include routinely collected risk factors and can be easily adapted into existing hypertension and CVD outcomes, as well as screening programs relevant to precision medicine. Large databases provide sufficient statistical power to develop and validate predictive algorithms, which in turn generate distinct survival probability for a wide range of health profiles or populations. Importantly, the research results suggest that, compared with the other three models, NMTLR has the better discrimination and calibration performance. Moreover, NMTLR also has the better performance when dealing with patients with partial information (censored). This reflects that the model, which is able to handle and utilize both censored and uncensored instances, can provide significantly better results than the model which use only fully observed (uncensored) instances.

This study is strengthened by using large, routinely collected data (clinical and demographic information) for model development. The wealth of administrative data includes information on health services utilization, comorbidities, and other clinical information (such as laboratory and medications). It has considerable potential to use administrative data to improve clinical care cross a spectrum of diseases, especially for patients with chronic diseases, such as HF, stroke and MI. Although the administrative datasets used in this project did not include modifiable risk factors of CVD (e.g., diet, lifestyles and smoking), it is important to note that despite the fact that administrative datasets do not contain modifiable risk factors, the data still does a an adequate job on

personalized survival prediction when compared with other datasets that contain a lot more clinical variables [90, 91]. These findings suggest that the application of administrative data may be used for personalized prediction models. Despite the accuracy of our findings, there are limitations to our work. First, most patients in the study were followed up for 3 years, which may not be adequate to capture all CVD outcomes, especially for those younger patients with no comorbidities. Second, we have designed this study as retrospective and in a single administrative district (province of Alberta), it will be of major interest to observe its performance in other provinces/territories and countries. There are other clinical factors that may significantly vary across geographic locations, such as severity of CVD, number of comorbidities, etc. The population and data differences in various geographic areas may require local training of models to obtain higher accuracy. However, the use of data from multiple areas helps to improve model generalizability, because models trained on one specific dataset have the risk of being over fitted with little applicability to other sites. We believe that by using data from different institutions and by performing internal and external cross-validation, we can appropriately balance the model accuracy and generalizability. Nevertheless, our models may suffer from issues such as overfitting, and only further testing with data from other geographic locations (as well as other institutions) would better clarify our model's performance. Moreover, in the effort to use institutional data to deliver precise health, some have advocated for data recalibration at different sites to maximize predictive model accuracy. The need for recalibration is not necessarily a failure of our approach to predictive modeling, because others have found that recalibrating more well-known epidemiologically derived risk scores can provide better risk estimates for specific populations as well [92].

Third, this study did not fully take into account missing data with variables been included in the model even if one patient had a single value in the dataset. This may have diminished our predictive accuracy to some extent. However, a strength of this approach is that it represents the true nature of rare administrative data with minimal transformations and no data imputations. Another consideration is that we choose to use a validated case definition to define hypertension patients, which has a high degree of sensitivity and specificity. This methodology represents a more easily deployable solution to cohort building and model development [52].

Last, another potential limitation worth discussing is the use of a black box algorithm to predict CVD among incident hypertensive patients. Our ideal goal is to understand causal pathways for diseases and outcomes or to predict their occurrence [93]. Although similar statistical algorithms can be applied to these separate modeling tasks (prediction or inference), models with good predictors of disease are not models from which we can derive understanding of the disease mechanisms. The NMTLR model has the advantage of combining neural networks and multi-task logistic regression and can be more accurate than other models. However, the fact that the variables used are not always identifiable and possible explanation can be a drawback of this methodology. We cannot actually know which variable acts as the most contributing risk factor for CVD outcomes among hypertensive patients. Even so, machine learning algorithms are still beneficial and may revolutionize our ability to deliver both precision and population health, especially with an understanding of the limitations of each algorithm. For instance, associations between clinical, demographic, or imaging characteristics and patient outcomes are not always linear. Machine learning algorithms, such as neural multi-task logistic regression,

are still able to identify nonlinear signals that can predict disease occurrence and outcomes. Such capabilities will become even more important as the depth and breadth of healthcare data grow.

5.4 Future research

The work presented in this dissertation is the first step towards using the administrative health data routinely collected in Alberta, Canada to directly provide individual CVD survival predictions for newly diagnosed hypertensive patients. In general, it is believed that NMTLR has the potential to outperform state-of-the-art methods for individual survival prediction because it has demonstrated its predictive power using the $\boldsymbol{C-index}$ and Brier Score. We also found that RSF is the better model to predict the number of patients who will have disease at a given time point when compared with other three models.

For future work, more accurate survival time point estimation is desirable. A survival time point estimation can be meaningful for both patients and clinicians, because instead of getting a survival probability curve, an accurate survival time in days can be helpful in decision making. We have considered the use of median survival time (50% survival probability) as a point estimation for survival, however, it lacks the ability to represent the general predictive ability of a model and can easily lead to a biased evaluation when $C-index$ is used as the performance evaluation metric. Although the modeled $C-index$ and Brier Score are very close with each other, the patient's individual survival probability curves have totally different patterns. Chen et al., also found that different performance metrics for evaluation of a survival prediction model may give different

conclusions in its discriminatory ability [94]. Although the Brier Score of the models are close with each other, the other performance metrics (like p-values from the log-rank test) are obviously different. Before embarking on the development of more machine learning methods for survival analysis, further research should be conducted to find a good performance evaluation metric.

A CVD survival prediction tool among hypertensive patients powered by machine learning methods for individual survival prediction can provide guidance to both patients and clinicians. In recent years, some online CVD survival prediction tools have been released, which allow users to input information on variables such as age, sex, lifestyle, etc., and provides survival predictions based on the input. However, these online prediction tools are all based on traditional statistical methods. For example, “Project Big Life” provides life expectancy, heart attack and stroke risk prediction based on traditional statistical models published by the University of Ottawa [95]. Therefore, it is crucial to build a machine learning-based platform that allows patients and clinicians to perform discussions on treatments and utilizations.

Chapter Six: **Conclusion**

Survival prediction is the task of predicting the length of time that an individual patient will survive; accurate predictions can provide doctors with better guidelines on treatments selection and future planning. This differs from the standard survival analysis, which focuses on population-based studies and attempts to discover prognostic factors and/or analyze the median survival times of different patient groups.

In this study, we demonstrated that these four models (LMTLR, NMTLR, RSF and CoxPH) exhibited similar high discrimination and calibration in predicting the survival of the hypertension patients. Machine learning algorithms using available administrative data can predict which hypertension patients are most likely to develop CVD and accurately predict the survival curve of individual patients. The NMTLR model has the better individual survival prediction, while the RSF model has the better population survival prediction. Improved prediction of outcomes has the potential to help clinicians make more meaningful treatment decisions and to assist with planning of future social and care needs, especially for those who might need more aggressive disease management. Most importantly, the described approach makes use of digital administrative data that is already routinely collected but underexploited by clinical health systems. Although machine learning methodologies have many advantages, to truly improve patient care and outcomes, methods for investigating causal relationships will remain an important part of the healthcare and biomedical armamentarium.

References

1. Leung AA, Daskalopoulou SS, Dasgupta K, McBrien K, Butalia S, Zarnke KB, et al. Hypertension Canada's 2017 guidelines for diagnosis, risk assessment, prevention, and treatment of hypertension in adults. Can J Cardiol. 2017;33(5):557-76.

2. McAlister FA, Wilkins K, Joffres M, Leenen FH, Fodor G, Gee M, et al. Changes in the rates of awareness, treatment and control of hypertension in Canada over the past two decades. CMAJ. 2011;183(9):1007-13.

3. Hemmelgarn BR, Chen G, Walker R, McAlister FA, Quan H, Tu K, et al. Trends in antihypertensive drug prescriptions and physician visits in Canada between 1996 and 2006. Can J Cardiol. 2008;24(6):507-12.

4. Campbell N, Young ER, Drouin D, Legowski B, Adams MA, Farrell J, et al. A framework for discussion on how to improve prevention, management, and control of hypertension in Canada. Canadian Journal of Cardiology. 2012;28(3):262-9.

5. Weaver CG, Clement FM, Campbell NR, James MT, Klarenbach SW, Hemmelgarn BR, et al. Healthcare costs attributable to hypertension: Canadian population-based cohort study. Hypertension. 2015;66(3):502-8.

6. Public Health Agency of Canada. tracking heart disease and stroke in Canada. Ottawa, Ontario: Public Health Agency of Canada; 2011.

7. Padwal RS, Bienek A, McAlister FA, Campbell NR, Outcomes research task force of the Canadian hypertension education epidemiology of hypertension in Canada: An Update. Can J Cardiol. 2016;32(5):687-94.

8. Quan H, Chen G, Tu K, Bartlett G, Butt DA, Campbell NR, et al. Outcomes among 3.5 million newly diagnosed hypertensive Canadians. Can J Cardiol. 2013;29(5):592-7.

9. Gaciong Z, Siński M, Lewandowski J. Blood pressure control and primary prevention of stroke: summary of the recent clinical trial data and meta-analyses. Current hypertension reports. 2013;15(6):559-74.

10. Schiffrin EL, Campbell NRC, Feldman RD, Kaczorowski J, Lewanczuk R, Padwal R, et al. Hypertension in Canada: past, present, and future. Annals of Global Health. 2016;82(2):288-99.

11. Organization WH. Prevention of Cardiovascular Disease, Pocket guidelines for assessment and management of cardiovascular risk. http://www.ish-worldcom/documents/pocketgl_english_wpr-ab pdf. 2007.

12. Collaboration BPLTT. Blood pressure-lowering treatment based on cardiovascular risk: a meta-analysis of individual patient data. The Lancet. 2014;384(9943):591-8.

13. Law M, Morris J, Wald N. Use of blood pressure lowering drugs in the prevention of cardiovascular disease: meta-analysis of 147 randomised trials in the context of expectations from prospective epidemiological studies. Bmj. 2009;338:b1665.

14. Karnes JH, Cooper-DeHoff RM. Antihypertensive medications: benefits of blood pressure lowering and hazards of metabolic effects. Expert review of cardiovascular therapy. 2009;7(6):689-702.

15. Ettehad D, Emdin CA, Kiran A, Anderson SG, Callender T, Emberson J, et al. Blood pressure lowering for prevention of cardiovascular disease and death: a systematic review and meta-analysis. 2016;387(10022):957-67.

16. Echouffo-Tcheugui JB, Batty GD, Kivimaki M, Kengne AP. Risk models to predict hypertension: a systematic review. PLoS One. 2013;8(7):e67370.

17. Khanji MY, Bicalho VV, van Waardhuizen CN, Ferket BS, Petersen SE, Hunink MG. Cardiovascular risk assessment: a systematic review of guidelines. Ann Intern Med. 2016;165(10):713-22.

18. Alabousi M, Abdullah P, Alter DA, Booth GL, Hogg W, Ko DT, et al. Cardiovascular risk factor management performance in canada and the united states: a systematic review. Can J Cardiol. 2017;33(3):393-404.

19. Grover SA, Lowensteyn I, Joseph L, Kaouache M, Marchand S, Coupal L, et al. Discussing coronary risk with patients to improve blood pressure treatment: secondary results from the CHECK-UP study. 2009;24(1):33-9.

20. Patnode CD, Evans CV, Senger CA, Redmond N, Lin JSJJ. Behavioral counseling to promote a healthful diet and physical activity for cardiovascular disease prevention in adults without known cardiovascular disease risk factors: updated evidence report and systematic review for the US Preventive Services Task Force. 2017;318(2):175-93.

21. Usher-Smith JA, Silarova B, Schuit E, Moons KG, Griffin SJJBo. Impact of provision of cardiovascular disease risk estimates to healthcare professionals and patients: a systematic review. 2015;5(10):e008717.

22. Gu Q. Hypertension among adults in the United States: National health and nutrition examination survey, 2011–2012. 2013.

23. Bobashev G. myEpi. Epidemiology of one. Frontiers in Public Health. 2014;2:97.

24. Pearce N. The ecological fallacy strikes back. BMJ Publishing Group Ltd; 2000.

25. Butler-Jones D. The chief public health officer's report on the state of public health in Canada: 2008: Public Health Agency of Canada Ottawa, ON; 2008.

26. Leung AA, Daskalopoulou SS, Dasgupta K, McBrien K, Butalia S, Zarnke KB, et al. Hypertension Canada's 2017 guidelines for diagnosis, risk assessment, prevention, and treatment of hypertension in adults. Can J Cardiol. 2017;33(5):557-76.

27. J TFftMoAHotESoHJEH. Guidelines for the management of arterial hypertension. 2007;28:1462-536.

28. Carey RM, Whelton PKJAoim. Prevention, detection, evaluation, and management of high blood pressure in adults: synopsis of the 2017 American College of Cardiology/American Heart Association hypertension guideline. 2018;168(5):351-8.

29. Nerenberg KA, Zarnke KB, Leung AA, Dasgupta K, Butalia S, McBrien K, et al. Hypertension Canada's 2018 guidelines for diagnosis, risk assessment, prevention, and treatment of hypertension in adults and children. Can J Cardiol. 2018;34(5):506-25.

30. Intengan HD, Schiffrin ELJH. Vascular remodeling in hypertension: roles of apoptosis, inflammation, and fibrosis. 2001;38(3):581-7.

31. London GMJNDT. Left ventricular alterations and end - stage renal disease. 2002;17(suppl_1):29-36.

32. Kannel WBJHR. Framingham study insights into hypertensive risk of cardiovascular disease. 1995;18(3):181-96.

33. Mancia G, Fagard R, Narkiewicz K, Redon J, Zanchetti A, Boehm M, et al. 2013 ESH/ESC guidelines for the management of arterial hypertension: the task force for the management of arterial hypertension of the European Society of Hypertension (ESH) and of the European Society of Cardiology (ESC). 2013;22(4):193-278.

34. Cloutier L, Daskalopoulou SS, Padwal RS, Lamarre-Cliche M, Bolli P, McLean D, et al. A new algorithm for the diagnosis of hypertension in Canada. Canadian Journal of Cardiology. 2015;31(5):620-30.

35. Krittanawong C, Bomback AS, Baber U, Bangalore S, Messerli FH, Tang WW. Future direction for using artificial intelligence to predict and manage hypertension. Current hypertension reports. 2018;20(9):75.

36. Johnson KW, Soto JT, Glicksberg BS, Shameer K, Miotto R, Ali M, et al. Artificial intelligence in cardiology. Journal of the American College of Cardiology. 2018;71(23):2668-79.

37. Miyazawa AA. Artificial intelligence: the future for cardiology. Heart. 2019;105(15):1214-.

38. Kjeldsen SE. Hypertension and cardiovascular risk: General aspects. Pharmacological research. 2018;129:95-9.

39. Iadecola C, Davisson RL. Hypertension and cerebrovascular dysfunction. Cell metabolism. 2008;7(6):476-84.

40. Humphries K, Izadnegadar M, Sedlak T, Saw J, Johnston N, Schenck-Gustafsson K, et al. Sex differences in cardiovascular disease–impact on care and outcomes. Frontiers in neuroendocrinology. 2017;46:46.

41. Taljaard M, Tuna M, Bennett C, Perez R, Rosella L, Tu JV, et al. Cardiovascular Disease Population Risk Tool (CVDPoRT): predictive algorithm for assessing CVD risk in the community setting. A study protocol. BMJ Open. 2014;4(10):e006701.

42. Ben-Shlomo Y, Spears M, Boustred C, May M, Anderson SG, Benjamin EJ, et al. Aortic pulse wave velocity improves cardiovascular event prediction: an individual

participant meta-analysis of prospective observational data from 17,635 subjects. Journal of the American College of Cardiology. 2014;63(7):636-46.

43. Melander O, Newton-Cheh C, Almgren P, Hedblad B, Berglund G, Engström G, et al. Novel and conventional biomarkers for prediction of incident cardiovascular events in the community. Jama. 2009;302(1):49-57.

44. Siontis GC, Tzoulaki I, Siontis KC, Ioannidis JP. Comparisons of established risk prediction models for cardiovascular disease: systematic review. Bmj. 2012;344:e3318.

45. Matheny M, McPheeters ML, Glasser A, Mercaldo N, Weaver RB, Jerome RN, et al. Systematic review of cardiovascular disease risk assessment tools. 2011.

46. Cooney MT, Dudina AL, Graham IM. Value and limitations of existing scores for the assessment of cardiovascular risk: a review for clinicians. Journal of the American College of Cardiology. 2009;54(14):1209-27.

47. Sehestedt T, Jeppesen J, Hansen TW, Wachtell K, Ibsen H, Torp-Petersen C, et al. Risk prediction is improved by adding markers of subclinical organ damage to SCORE. European heart journal. 2009;31(7):883-91.

48. Manuel DG, Tuna M, Bennett C, Hennessy D, Rosella L, Sanmartin C, et al. Development and validation of a cardiovascular disease risk-prediction model using population health surveys: the Cardiovascular Disease Population Risk Tool (CVDPoRT). CMAJ. 2018;190(29):E871-E82.

49. Jacobs P, Yim R. Using Canadian administrative databases to derive economic data for health technology assessments: Canadian Agency for Drugs and Technologies in Health Agence canadienne; 2009.

50. Cadarette SM, Wong L. An introduction to health care administrative data. The Canadian journal of hospital pharmacy. 2015;68(3):232.

51. Walker RL, Hennessy DA, Johansen H, Sambell C, Lix L, Quan H. Implementation of ICD-10 in Canada: how has it impacted coded hospital discharge data. BMC health services research. 2012;12(1):149.

52. Quan H, Khan N, Hemmelgarn BR, Tu K, Chen G, Campbell N, et al. Validation of a case definition to define hypertension using administrative data. Hypertension. 2009;54(6):1423-8.

53. Tu K, Wang M, Young J, Green D, Ivers NM, Butt D, et al. Validity of administrative data for identifying patients who have had a stroke or transient ischemic attack using EMRALD as a reference standard. Canadian Journal of Cardiology. 2013;29(11):1388-94.

54. Quach S, Blais C, Quan H. Administrative data have high variation in validity for recording heart failure. The Canadian journal of cardiology. 2010;26(8):306-12.

55. Metcalfe A, Neudam A, Forde S, Liu M, Drosler S, Quan H, et al. Case definitions for acute myocardial infarction in administrative databases and their impact on in-hospital mortality rates. Health services research. 2013;48(1):290-318.

56. Quan H, Khan N, Hemmelgarn B, Tu K, Chen G, Campbell N, et al. Hypertension Outcome and Surveillance Team of the Canadian Hypertension Education Programs: Validation of a case definition to define hypertension using administrative data. Hypertension. 2009;54(6):1423-8.

57. Chen G, Lix L, Tu K, Hemmelgarn BR, Campbell NR, McAlister FA, et al. Influence of using different databases and ‘look back’intervals to define comorbidity

profiles for patients with newly diagnosed hypertension: Implications for health services researchers. PloS one. 2016;11(9):e0162074.

58. Quan H, Smith M, Bartlett-Esquilant G, Johansen H, Tu K, Lix L. Hypertension Outcome and Surveillance Team. Mining administrative health databases to advance medical science: geographical considerations and untapped potential in Canada. Can J Cardiol. 2012;28:152-4.

59. Quan S, Chen G, Padwal RS, McAlister FA, Tran KC, Campbell NR, et al. Frequency of laboratory testing and associated abnormalities in patients with hypertension. The Journal of Clinical Hypertension. 2020.

60. Klein JP, Zhang MJ. Survival analysis, software. Encyclopedia of biostatistics. 2005;8.

61. Carpenter M. Survival analysis: a self-learning text. Taylor & Francis; 1997.

62. Wang P, Li Y, Reddy CKJapa. Machine learning for survival analysis: A survey. 2017.

63. Kleinbaum DG, Klein M. Survival analysis: Springer; 2010.

64. Lee ET, Wang J. Statistical methods for survival data analysis: John Wiley & Sons; 2003.

65. Zwiener I, Blettner M, Hommel G. Survival analysis. Deutsches Aerzteblatt Online. 2011.

66. Fox J. Cox proportional-hazards regression for survival data. An R and S-PLUS companion to applied regression. 2002;2002.

67. Kleinbaum DG, Klein M. Parametric survival models. Survival analysis: Springer; 2012. p. 289-361.

68. Bradburn MJ, Clark TG, Love S, Altman D. Survival analysis part II: multivariate data analysis–an introduction to concepts and methods. British journal of cancer. 2003;89(3):431.

69. Wang P, Li Y, Reddy CK. Machine learning for survival analysis: A survey. arXiv preprint arXiv:170804649. 2017.

70. Amaratunga D, Cabrera J, Sargsyan D, Kostis JB, Zinonos S, Kostis WJ. Uses and opportunities for machine learning in hypertension research. International Journal of Cardiology Hypertension. 2020:100027.

71. Breiman L, Stone CJ, Gins JD. New Methods for Estimating Tail Probabilities and Extreme Value Disributions. technology service corp santa monica calif; 1979.

72. Breiman L. Bagging predictors. Machine learning. 1996;24(2):123-40.

73. Breiman L. Random forests. Machine learning. 2001;45(1):5-32.

74. Alpaydin E. Introduction to machine learning: MIT press; 2020.

75. Yoo KD, Noh J, Lee H, Kim DK, Lim CS, Kim YH, et al. A machine learning approach using survival statistics to predict graft survival in kidney transplant recipients: a multicenter cohort study. Sci Rep. 2017;7(1):8904.

76. Li Y, Wang J, Ye J, Reddy CK, editors. A multi-task learning formulation for survival analysis. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016.

77. Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review. 1958;65(6):386.

78. McDonald C. Machine learning fundamentals (II): neural networks. Towards Data Science. 2017.

79. Sibi P, Jones SA, Siddarth P. Analysis of different activation functions using back propagation neural networks. Journal of theoretical and applied information technology. 2013;47(3):1264-8.

80. Liaw A, Wiener M. Classification and regression by randomForest. R news. 2002;2(3):18-22.

81. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. The annals of applied statistics. 2008;2(3):841-60.

82. Heagerty PJ, Zheng Y. Survival model predictive accuracy and ROC curves. Biometrics. 2005;61(1):92-105.

83. Steck H, Krishnapuram B, Dehing-Oberije C, Lambin P, Raykar VC, editors. On ranking in survival analysis: bounds on the concordance index. Advances in neural information processing systems; 2008.

84. Miller Jr RG. Survival analysis: John Wiley & Sons; 2011.

85. Brier GW. Verification of forecasts expressed in terms of probability. Monthly weather review. 1950;78(1):1-3.

86. Pace R, Peters T, Rahme E, Dasgupta K. Validity of health administrative database definitions for hypertension: a systematic review. Can J Cardiol. 2017;33(8):1052-9.

87. Wilkins K, Gee M, Campbell N. The difference in hypertension control between older men and women. Health reports. 2012;23(4):1C.

88. Bushnik T, Hennessy DA, McAlister FA, Manuel DG. Factors associated with hypertension control among older Canadians. Health Rep. 2018;29(6):3-10.

89. Singh R, Mukhopadhyay K. Survival analysis in clinical trials: Basics and must know areas. Perspectives in clinical research. 2011;2(4):145.

90. Yu C-N, Greiner R, Lin H-C, Baracos V, editors. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. Advances in Neural Information Processing Systems; 2011.

91. Li Y, Wang J, Ye J, Reddy CK. A multi-task learning formulation for survival analysis. 2016:1715-24.

92. Barzi F, Patel A, Gu D, Sritara P, Lam T, Rodgers A, et al. Cardiovascular risk prediction tools for populations in Asia. Journal of epidemiology and community health. 2007;61(2):115-21.

93. Mayeux R. Biomarkers: potential uses and limitations. NeuroRx. 2004;1(2):182-8.

94. Chen H-C, Kodell RL, Cheng KF, Chen JJ. Assessment of performance of survival prediction models for cancer prognosis. BMC medical research methodology. 2012;12(1):102.

95. University of Ottawa . Project Big Life 2020 [Available from: https://www.projectbiglife.ca/].

Appendix A. ICD-9 code for Charlson comorbidities

Comorbidities	ICD-9 Codes
Myocardial infarction	410.x, 412.x
Congestive heart failure	398.91, 402.01, 402.11, 402.91, 404.01, 404.03, 404.11, 404.13, 404.91, 404.93, 425.4–425.9, 428.x
Peripheral vascular disease	093.0, 437.3, 440.x, 441.x, 443.1–443.9, 47.1, 557.1, 557.9, V43.4
Cerebrovascular disease	362.34, 430.x–438.x
Dementia	290.x, 294.1, 331.2
Chronic pulmonary disease	416.8, 416.9, 490.x–505.x, 506.4, 508.1, 508.8
Rheumatic disease	446.5, 710.0–710.4, 714.0–714.2, 714.8, 725.x
Peptic ulcer disease	531.x–534.x
Mild liver disease	070.22, 070.23, 070.32, 070.33, 070.44, 070.54, 070.6, 070.9, 570.x, 571.x, 573.3, 573.4, 573.8, 573.9, V42.7
Diabetes without chronic complication	250.0–250.3, 250.8, 250.9
Diabetes with chronic complication	250.4–250.7
Hemiplegia or paraplegia	334.1, 342.x, 343.x, 344.0– 344.6, 344.9
Renal disease	403.01, 403.11, 403.91, 404.02, 404.03, 404.12, 404.13, 404.92, 404.93, 582.x, 583.0–583.7, 585.x, 586.x, 588.0, V42.0, V45.1, V56.x
Any malignancy, including lymphoma and leukemia, except malignant neoplasm of skin	140.x–172.x, 174.x–195.8, 200.x–208.x, 238.6
Moderate or severe liver disease	196.x–199.x
Metastatic solid tumor	042.x–044.x

Appendix B. ICD-10-CA code for Charlson comorbidities

Comorbidities	ICD-10-CA Codes
Myocardial infarction	I21.x, I22.x, I25.2
Congestive heart failure	I09.9, I11.0, I13.0, I13.2, I25.5, I42.0, I42.5-I42.9, I43.x, I50.x, P29.0
Peripheral vascular disease	I70.x, I71.x, I73.1, I73.8, I73.9, I77.1, I79.0, I79.2, K55.1, K55.8, K55.9, Z95.8, Z95.9
Cerebrovascular disease	G45.x, G46.x, H34.0, I60.x--I69.x
Dementia	F00.x--F03.x, F05.1, G30.x, G31.1
Chronic pulmonary disease	I27.8, I27.9, J40.x--J47.x, J60.x--J67.x, J68.4, J70.1, J70.3
Rheumatic disease	M05.x, M06.x, M31.5, M32.x--M34.x, M35.1, M35.3, M36.0
Peptic ulcer disease	K25.x--K28.x
Mild liver disease	B18.x, K70.0--K70.3, K70.9, K71.3--K71.5, K71.7, K73.x, K74.x, K76.0, K76.2--K76.4, K76.8, K76.9, Z94.4
Diabetes without chronic complication	E10.0, E10.l, E10.6, E10.8, E10.9, E11.0, E11.1, E11.6, E11.8, E11.9, E12.0, E12.1, El2.6, E12.8, El2.9, E13.0, E13.1, E13.6, E13.8, E13.9, E14.0, E14.1, E14.6, E14.8, E14.9
Diabetes with chronic complication	E10.2--E10.5, E10.7, E11.2, E11.5, E11.7, E12.2--E12.5, E12.7, E13.2--E13.5, E13.7, E14.2--E14.5, E14.7
Hemiplegia or paraplegia	G04.1, G11.4, G80.1, G80.2, G81.x, G82.x, G83.0--G83.4, G83.9
Renal disease	I12.0, I13.1, N03.2--N03.7, N05.2--N05.7, N18.x, N19.x, N25.0, Z49.0--Z49.2, Z94.0, Z99.2
Any malignancy, including lymphoma and leukemia, except malignant neoplasm of skin	C00.x--C26.x, C30.x--C34.x, C37.x--C41.x, C43.x, C45.x--C58.x, C60.x--C76.x, C81.x--C85.x, C88.x, C90.x--C97.x
Moderate or severe liver disease	I85.0, I85.9, I86.4, I98.2, K70.4, K71.1, K72.1, K72.9, K76.5, K76.6, K76.7
Metastatic solid tumor	C77.x--C80.x