2021-04-01

Develop a comprehensive hypertension prediction model and risk score in population-based data applying conventional statistical and machine learning approaches

Chowdhury, Mohammad Ziaul Islam

UNIVERSITY OF CALGARY

Develop a comprehensive hypertension prediction model and risk score in population-based data applying conventional statistical and machine learning approaches

by

Mohammad Ziaul Islam Chowdhury

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE

DEGREE OF DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN COMMUNITY HEALTH SCIENCES

CALGARY, ALBERTA

APRIL, 2021

Abstract

Hypertension is a common medical condition and is a significant risk factor for heart attack, stroke, kidney disease, and mortality. Developing a risk prediction model for hypertension incidence incorporating its risk factors can help identify high-risk individuals who should be targeted for healthy behavioral changes or medical treatment to prevent hypertension onset. This research aims to develop a robust hypertension prediction model for the general population. More specifically, we aimed to 1) conduct a comprehensive systematic review to identify risk factors and prediction models for hypertension incidence and perform a meta-analysis to evaluate the current model's predictive performance. 2) develop a risk prediction model for incident hypertension in a Canadian cohort using a traditional modeling approach. 3) develop machine learning algorithms to predict hypertension incidence and compare their predictive performance with a traditional statistical model.

We systematically searched MEDLINE, EMBASE, Web of Science, Scopus, and the grey literature for studies predicting the risk of hypertension among the general adult population. We identified 52 studies that presented 117 models, of which 75 were developed using traditional regression-based modeling and 42 using machine learning algorithms. No studies were from Canada where a hypertension prediction model was developed or validated. Meta-analysis showed the overall pooled C-statistics 0.75 [0.73 – 0.77] for the traditional regression-based models and 0.76 [0.72 – 0.79] for the machine learning-based models.

The lack of a hypertension prediction model in a Canadian context motivated us to develop a new model. We used the data of 18,322 participants on 29 candidate variables from the large Alberta's Tomorrow Project (ATP) to develop traditional Cox proportional hazards (PH) model. Age, sex, body mass index (BMI), systolic blood pressure (SBP), diabetes, total physical activity

time, and cardiovascular disease were used as significant risk factors in the model. Our model showed good discrimination (Harrel's C-statistic 0.77) and calibration (Grønnesby and Borgan test, χ^2 statistic = 8.75, p = 0.07; calibration slope 1.006). A risk score table to estimate hypertension risks at 2-, 3-, 5-, and 6-year were derived from the model to favor the model's clinical implementation and workability.

Five machine learning algorithms were also developed to predict hypertension incidence: penalized regression Ridge, Lasso, Elastic Net (EN), random survival forest (RSF), and gradient boosting (GB). The performance of machine learning algorithms was observed, similar to the traditional Cox PH model. Average C-indexes were 0.78, 0.78, 0.78, 0.76, 0.76, for Ridge, Lasso, Elastic Net, RSF, GB, respectively. Important features associated with each machine learning algorithms were also presented.

We developed a simple yet practical prediction model to estimate the risk of incident hypertension for the Canadian population that relies on readily available variables. Our results showed little predictive performance difference between machine learning algorithms and the traditional Cox PH model in predicting hypertension incidence. Our newly developed model may help clinicians, and the general population assess their risks of new-onset hypertension and facilitate discussions on preventing this risk more effectively.

Preface

This thesis is presented in a manuscript-based format.

Chapter two is in preparation for submitting to the *Scientific Reports*.

A systematic review to identify risk factors and prediction models for hypertension incidence and a meta-analysis to evaluate model performance.

Chapter three is in preparation for submitting to the *Hypertension*.

Development of a risk prediction model for incident hypertension in a Canadian cohort using conventional regression-based modeling approach and converting into a risk score for use in daily clinical practice.

Chapter four is in preparation for submitting to the *Plos One*.

Using machine learning algorithms to predict hypertension incidence and comparing their predictive performance with a conventional statistical model in a large survival data.

The candidate designed the study with the supervisory committee. The candidate cleaned the study data, conducted the analysis, interpreted the results, and wrote each of the manuscripts.

Acknowledgments

This thesis and my PhD would not have been completed without the support of many astounding people. I am humbled by how many people have played a role in my development as a researcher. I want to express my sincere gratitude to my supervisors, Dr. Tanvir Turin Chowdhury and Dr. Hude Quan. Over the past six years, their continuous advice, support, and encouragement have made this journey meaningful. My supervisor, Dr. Tanvir Turin Chowdhury, was instrumental in my development as a researcher. Your sense of duty, commitment, and dedication has influenced me a lot. I will always be indebted to you for taking a chance on me and for your mentorship. I would also extend my deepest gratitude to my co-supervisor, Dr. Hude Quan, who was always there whenever I need something. His encouragement, direction, and financial support throughout my PhD program were invaluable.

I want to thank my supervisory committee members, Dr. Maeve O'Beirne, Dr. Alexander Ah-Chi Leung, and Dr. Khokan Sikdar. You have provided sound technical expertise throughout my PhD program, coupled with strong encouragement and pragmatic advice. I am fortunate to have had the great chance to work with you and learn from all of you. Dr. Alexander has given me many vital suggestions on this study's clinical, epidemiological, and analytical aspects.

I would also like to thank my classmates and friends at the Department of Community Health Sciences to share knowledge and experience and have lots of fun together during my study period. It was great!

Finally, I take this opportunity to express my profound gratitude to my mother (Rowshan), my wife (Tahrina), and my kids (Zaeem and Zahaa). Without their constant support,

encouragement, patience, and love, this day would never have happened. Above all, my gratitude to the Almighty Allah without his mercy, nothing is possible.

This research was supported by Western Regional Training Centre for Health Services Research studentship, Achievers in Medical Science Graduate Recruitment scholarship, and research grants from Dr. Hude Quan and Dr. Tanvir Turin Chowdhury.

Dedication

In memory of my beloved father

Table of Contents

Abstract..……ii
Preface.……...iv
Acknowledgments.……..v
Dedication……vii
Table of Contents……viii
List of Tables.…… xi
List of Figures.……...xii
List of Abbreviations……xiv

CHAPTER 1. INTRODUCTION…….……1

1.1 Brief Overview……...……..2
1.2 Hypertension and its Symptoms……...……3
1.3 Risk Factors for Hypertension……..……4
1.3.1 Modifiable Risk Factors……...…5
1.3.2 Non-modifiable Risk Factors……...6
1.4 Hypertension Consequences……....8
1.5 Hypertension Burden……...9
1.6 Hypertension Prevention: Risk Prediction Model at the Core……10
1.7 Overview of Risk Prediction Models……..11
1.7.1 Examples of Well-Known Risk Prediction Models……12
1.7.2 Methods in Risk Prediction Models……13
1.7.3 Model Validation……...19
1.7.4 Evaluating Model Performance……..21
1.7.5 Generation of Point Scoring System…….…22
1.7.6 Existing Research on Hypertension Prediction Models……..…23
1.8 Study Rationale…….…25
1.9 Research Objectives…….…..26
1.10 References……..29

CHAPTER 2. A SYSTEMATIC REVIEW TO IDENTIFY RISK FACTORS AND PREDICTION MODELS FOR HYPERTENSION INCIDENCE AND A META-ANALYSIS TO EVALUATE MODEL PERFORMANCE……...46

2.1 Abstract……..47
2.2 Introduction……49
2.3 Methods……..50
2.3.1 Data Sources and Searches…….50
2.3.2 Eligibility Criteria…….. 51
2.3.3 Study Selection……..…52
2.3.4 Data Extraction……..…….53
2.3.5 Data Analysis……..…...54
2.4 Results……..……..56
2.4.1 Study Identification and Selection……..……....56
2.4.2 Study Characteristics of Traditional Regression-based Models…….....57

2.4.3 Meta-analysis of Traditional Regression-based Models……58
2.4.4 Critical Appraisal of Traditional Regression-based Models……...59
2.4.5 Study Characteristics of Machine Learning-based Models……....59
2.4.6 Meta-analysis of Machine Learning-based Models……....60
2.4.7 Study Characteristics of Externally Validated Models……...61
2.4.8 Meta-analysis of Externally Validated Models……..61
2.5 Discussion……..62
2.6 Conclusion……67
2.7 References……..69

CHAPTER 3. DEVELOPMENT OF A RISK PREDICTION MODEL FOR INCIDENT HYPERTENSION IN A CANADIAN COHORT USING TRADITIONAL REGRESSION-BASED MODELING APPROACH AND CONVERTING INTO A RISK SCORE FOR USE IN DAILY CLINICAL PRACTICE……..135

3.1 Abstract……136
3.2 Introduction……..138
3.3 Methods……139
3.3.1 Study population……..139
3.3.2 Selection of candidate variables……...142
3.3.3 Definition of variables……..142
3.3.4 Missing values……..144
3.3.5 Statistical analysis……145
3.4 Results……..148
3.4.1 Case Study……151
3.5 Discussion……152
3.6 References……157

CHAPTER 4. USING MACHINE LEARNING ALGORITHMS TO PREDICT HYPERTENSION INCIDENCE AND COMPARING THEIR PREDICTIVE PERFORMANCE WITH A CONVENTIONAL STATISTICAL MODEL IN A LARGE SURVIVAL DATA……205

4.1 Abstract……....206
4.2 Introduction……..207
4.3 Methods……209
4.3.1 Study population……..209
4.3.2 Selection of candidate features……210
4.3.3 Definition of features……....210
4.3.4 Missing values……..212
4.3.5 Feature selection……...212
4.3.6 Machine learning models……214
4.3.6.1 Cox PH model……..214
4.3.6.2 Penalized Cox regressions (Lasso, Ridge, and EN)……215
4.3.6.3 Random survival forest……216
4.3.6.4 Boosted gradient……...217
4.3.7 Feature importance……...217

4.3.8 Statistical analysis……218
4.4 Results……..219
4.5 Discussion……221
4.6 References……225

CHAPTER 5. DISCUSSION…….....244

5.1 Overview of main findings……...245
5.1.1 Multiple prediction models exist but none in a Canadian context…….....245
5.1.2 Similar predictive performance in existing traditional and machine learning-based models identified through meta-regression……...246
5.1.3 Limitations of current models……...……247
5.1.4 New prediction model for hypertension incidence in Canadian context using large cohort data……...249
5.1.5 Overall good predictive performance of the newly developed model……..250
5.1.6 Deriving risk score from the newly developed model for clinical utility…….250
5.1.7 Developing some machine learning-based models for hypertension incidence using the same survival data……..……..…….....251
5.1.8 Similar predictive performance in newly developed machine learning models and conventional model……..252
5.2 Strengths and Limitations……253
5.2.1 Systematic review and meta-analysis…….....253
5.2.2 A new traditionally developed hypertension prediction model…….254
5.2.3 Machine learning-based hypertension prediction models…….256
5.3 Future Directions……..…257
5.3.1 External validation……....257
5.3.2 Developing a computer-assisted tool……...….258
5.3.3 Updating model using meta-modeling……..258
5.3.4 Constructing a multi-disease prediction model…….259
5.4 Conclusion……....259
5.5 References……....261

APPENDIX 1……...264

List of Tables

Table 2.1 Keywords Used to Search in MEDLINE……88

Table 2. 2 Information about existing traditional regression-based hypertension prediction models from the selected studies……89

Table 2.3 Information about existing hypertension prediction models developed using machine learning algorithms from selected studies……105

Table 2.4 Information about external validation studies of existing traditional hypertension prediction models from selected studies……...113

Table 2.5 Study quality assessment using PROBAST……116

Table S2.1 Information about existing hypertension prediction models developed using biomarkers (genetic risk score) from the selected studies……130

Table 3.1 Baseline characteristics of study participants and comparison in the derivation sample and validation sample……...168

Table 3.2 Baseline characteristics of study participants according to the status of developing hypertension or not……...173

Table 3.3 Unadjusted and adjusted hazard ratios for the risk factors of hypertension incidence.178

Table 3.4 Regression coefficients and hazard ratio's for incident hypertension……..182

Table 3.5 Calculation of point values for risk score……183

Table 3.6 Risk estimates for point totals at 2, 3, 5, and 6-year time……184

Table 3.7 Risk categories based on total points……185

Table S3.1 Missing information about different variables……200

Table S3.2 Baseline characteristics of study participants according to the missing status…..…201

Table S3.3 Test of Cox proportional-hazards assumption……204

Table 4.1 Baseline characteristics of study participants and comparison of the training and test data……...235

Table 4.2 Baseline characteristics of study participants according to the status of developing hypertension or not……...238

Table 4.3 Feature's ranked based on five different approaches……...241

Table 4.4 Top 20 features selected by the different approaches with red cells indicates commonly selected features……...242

Table S4.1 Missing information about different variables…….....243

List of Figures

Figure 2.1 PRISMA diagram for the systematic review of studies presenting hypertension prediction models developed in the general population…….81

Figure 2.2 Conventional risk factors considered by traditional regression-based models……..82

Figure 2.3 Conventional risk factors considered by machine learning-based models……83

Figure 2.4 Forest plot of traditional regression-based models with 95% prediction interval……..84

Figure 2.5 Graphical summary presenting the percentage of hypertension risk prediction studies rated by level of concern, risk of bias (ROB), and applicability for each domain…….85

Figure 2.6 Forest plot of machine regression-based models with 95% prediction interval……86

Figure 2.7 Forest plot of externally validated models with 95% prediction interval…….87

Figure S2.1 Meta-regression on the age of the participants (study participants below average age versus above average age)……118

Figure S2.2 Meta-regression on the number of risk factors considered in the model (below median versus above median)……...…119

Figure S2.3 Meta-regression on sample size considered in the model (below median versus above median)……120

Figure S2.4 Meta-regression on the ethnicity of the study participants (Whites versus Asians).121

Figure S2.5 The number of PROBAST criteria satisfied by different studies…….122

Figure S2.6 Response to different signaling questions by the number of studies……123

Figure S2.7 Meta-regression on the age of the participants (study participants below average age versus above average age)……....124

Figure S2.8 Meta-regression on the number of risk factors considered in the model (below median versus above median)……...…125

Figure S2.9 Meta-regression on sample size considered in the model (below median versus above median)……126

Figure S2.10 Meta-regression on the ethnicity of the study participants (Whites versus Asians).127

Figure S2.11 Meta-regression on the ethnicity of the study participants (Whites versus Asians).128

Figure S2.12 Forest plot of models primarily developed using genetic risk factors/biomarkers with a 95% prediction interval…….129

Figure 3.1 Grønnesby and Borgan (GB) goodness-of-fit test of the risk prediction model for incident hypertension in the validation sample……163

Figure 3.2 Arjas like plots to compare observed and expected events in five quantiles of the linear predictor in the validation sample……164

Figure 3.3 Calibration plot where expected probabilities (predicted probabilities from the model) are plotted against observed outcome probabilities (calculated by Kaplan-Meier estimates)…..165

Figure 3.4 Smooth dashed lines represent predicted survival probabilities, and vertical capped lines represent Kaplan–Meier estimates with 95% confidence intervals. Three prognosis groups are plotted: the "Good" group (green lines), the "Intermediate" group (navy blue lines), and the "Poor" group (red lines)……...166

Figure 3.5 Histogram of the prognostic index in the derivation and validation datasets…….167

Figure S3.1 Plot to test the proportionality assumption of "Total physical activity time" variable……..186

Figure S3.2 Plot to test the proportionality assumption of "Diabetes" variable……187

Figure S3.3 Plot to test the proportionality assumption of "Age" variable……..188

Figure S3.4 Plot to test the proportionality assumption of "Systolic blood pressure" variable….189

Figure S3.5 Plot to test the proportionality assumption of "Cardiovascular disease" variable…190

Figure S3.6 Plot to test the proportionality assumption of "Body mass index" variable……..191

Figure S3.7 Plot to test the proportionality assumption of the "Sex" variable……192

Figure S3.8 Plot to test the proportionality assumption of "Age by Body mass index" interaction variable……..…193

Figure S3.9 Plot to test the proportionality assumption of "Age by Systolic blood pressure" interaction variable……...…194

Figure S3.10 Plot to test the proportionality assumption of "Age by Total physical activity time" interaction variable……...195

Figure S3.11 Plot to test the proportionality assumption of "Age by Sex" interaction variable..196

Figure S3.12 Plot to test the proportionality assumption of "Sex by Systolic blood pressure" interaction variable……...197

Figure S3.13 Plot to test the proportionality assumption of "Sex by Cardiovascular disease" interaction variable……...198

Figure S3.14 Traditional risk factors considered by conventional regression-based models……199

Figure 4.1 Features ranked according to their importance by the different models……..233

Figure 4.2 Boxplots showing the spread of values of the C-index produced by the different models……..234

List of Abbreviations

Abbreviation	Definition
ACC	American college of cardiology
AHA	American heart association
ANN	Artificial neural network
AOBP	Automated office blood pressure
APTT	Activated partial thromboplastin time
AROC	Area under the receiver operating characteristic curve
ATP	Alberta's tomorrow project
AUC	Area under the curve
BMI	Body mass index
BN	Bayesian network
BP	Back-propagation
BP	Blood pressure
CART	Classification and regression trees
CDHQ-I	Canadian diet history questionnaire I
CFS	Correlation-based feature subset selection
CHD	Coronary heart disease
CHF	Cumulative hazard function
CHREB	Conjoint health research ethics board
CPTP	Canadian partnership for tomorrow project
CVD	Cardiovascular disease
DALYs	Disability-adjusted life years
DASH	Dietary approaches to stop hypertension
DBP	Diastolic blood pressure
DIMR	Data Integration, Measurement & Reporting
DM	Diabetes mellitus
EN	Elastic net
EPV	Events per variable
FBG	Fasting blood glucose
FHRS	Framingham hypertension risk model
FINDRISC	Finnish diabetes risk score
GB	Gradient boosting
GB	Grønnesby and Borgan
GRS	Genetic risk score
HbA1c	Hemoglobin A1C
HCT	Hematocrit
HDL	High-density lipoprotein

HKSJ	Hartung-Knapp-Sidik-Jonkman
HL	Hosmer-Lemeshow
HR	Heart rate
HTN	Hypertension
ICD	International classification of diseases
IDH	Isolated diastolic hypertension
ISH	Isolated systolic hypertension
KoGES	Korean genome epidemiology study
LB	LogitBoost
LC	Lymphocyte count
LDL	Low-density lipoproteins
LOCF	Last observation carried forward
LOOCV	Leave-one-out cross-validation
LR	Logistic regression
LRM	Logistic regression model
LWB	Locally weighted naïve Bayes
M1	Myocardial infarction
MICE	Multiple imputation by chained equations
MAR	Missing at random
MARS	Multivariate adaptive regression splines
MDR	Multifactor-dimensionality reduction
MeSH	Medical subject headings
MET	Metabolic equivalent
NB	Naive Bayes
NGC	Neutrophil granulocyte count
NR	Not reported
O/E	Observed/Expected
PANS	Physical activity and nutrition survey
PFAA	Plasma free amino acid
PFO	Population, prognostic factors (or models of interest), and outcome
PGC	Plasma glucose concentration
PRISMA	Preferred reporting items for systematic reviews and meta-analyses
PROBAST	Prediction model risk of bias assessment tool
PYTPAQ	Past-year total physical activity questionnaire
RBF	Radial basis function
RCT	Randomized controlled trial
REML	Restricted maximum likelihood
RMSE	Root mean square error
ROB	Risk of bias
ROC	Receiver operating characteristic

RSF	Random survival forest
RTF	Random tree forest
SBP	Systolic blood pressure
SE	Standard error
SES	Statistically equivalent signature
SVM	Support vector machine
TC	Total cholesterol
TG	Triglycerides
TIA	Transient ischemic attack
TIC	Theil inequality coefficient
UHLQ	Updated health and lifestyle questionnaire
UK	United Kingdom
UKPDS	United Kingdom prospective diabetes study
US	United States
VIF	Variance inflation factor
WBC	White blood cell count
WC	Waist circumference
WHO	World health organization
WHR	Waist hip ratio
WHtR	Waist-to-height ratio
YLL	Years of life lost

CHAPTER 1. INTRODUCTION

1.1 Brief Overview

Hypertension is a common medical condition, affecting 1 in 5 Canadians[1], and represents a major modifiable risk factor for several fatal diseases, including heart attack, stroke, kidney disease, and mortality[2]. Hypertension is responsible for 13% of global deaths[3]. Prevention and control of hypertension are considered a major public health and primary care concern[4]. Hypertension outset can be prevented or delayed with lifestyle modification[5], drug treatment[6], or both. Primary prevention strategies are most likely to be effective when targeted to individuals at the highest risk. Population health research is increasingly integrating the precision health paradigm with a focused approach toward such targeted intervention[7], thus informing whom to target, what to target, where to target, and how to target personalized preventive initiatives[8]. Evidence suggests that the risk for hypertension progression depends on several factors, such as age, body mass index (BMI), waist-hip ratio, systolic blood pressure (BP), smoking, diabetes, family history, and level of physical inactivity[9]. Combining known risk factors into a multivariable model for risk classification would help identify high-risk individuals who should be targeted for healthy behavioral changes or medical treatment to prevent hypertension development[10–12].

Hypertension risk assessment is an important mainstay for preventive efforts against the condition. Several hypertension prediction models have been developed[4,13–18], but their performance in accurately forecasting incident hypertension, reflected in the models' predictive ability, varies. Based on the underlying population characteristics and data from which they are derived, each model has its strengths and weaknesses. Notably, efforts are needed to improve risk prediction to inform individual risk, clinical care, and policymaking.

Despite their advantages, the application of hypertension prediction models in clinical practice is rare. This is mainly due to the model's complexity, lack of enough validation and impact studies

to make the models trustworthy, and inadequate understanding of the models and their predicted probabilities among health professionals and patients. A properly developed accurate hypertension prediction model, which is easy to use and has multiple validation and impact studies, should be used in clinical settings. It supplements clinical information used in decision-making.

1.2 Hypertension and Its Symptoms

Hypertension (or high BP) refers to a condition where long-term high pressure in the arterial system results in damaged blood vessels, creating health problems**.** When the heart pumps blood more, arteries become narrower, resulting in a higher BP. BP is measured in millimeters of mercury (mm Hg) and is recorded as two numbers (first, the systolic number and then the diastolic number) generally written one above the other. The top refers to the maximum pressure in blood vessels, is called the systolic blood pressure (SBP), and reflects the peak pressure within the artery when the heart contracts or beats. The number at the bottom refers to the minimum pressure in blood vessels in between heartbeats. It is called diastolic blood pressure (DBP) and coincides with the heart muscle's relaxation. Normal adult BP is defined as an SBP of 120 mm Hg and a DBP of 80 mm Hg. Hypertension is generally defined as SBP $\geq$ 140 mm Hg and DBP $\geq$ 90 mm Hg[19], but a lower threshold of SBP $\geq$ 130 mm Hg and DBP $\geq$ 80 mm Hg has been proposed recently[20]. Nevertheless, recent medical guidelines from the American College of Cardiology/American Heart Association Task Force suggests categorizing BP as normal, elevated, or stage 1 or 2 hypertension to prevent and treat high BP[20]. The range for different categories of BP are: SBP < 120 mm Hg and DBP < 80 mm Hg (normal BP), SBP 120-129 mm Hg and DBP < 80 mm Hg (elevated BP), SBP 130-139 mm Hg or DBP 80-89 mm Hg (stage 1 hypertension), and SBP $\geq$ 140 mm Hg or DBP $\geq$ 90 mm Hg (stage 2 hypertension)[20]. An average of $\geq$ 2 careful readings taken on $\geq$ 2 occasions was used to determine the above-recommended BP measurements. According to Hypertension Canada's

2017 guideline, for using Automated Office Blood Pressure (AOBP), a mean SBP ≥ 135 mm Hg or DBP ≥ 85 mm Hg; for non-AOBP, a mean SBP ≥ 140 mm Hg or DBP ≥ 90 mm Hg; for ambulatory BP monitoring, a mean SBP ≥ 135 mm Hg or DBP ≥ 85 mm Hg; and for home BP monitoring, a mean SBP ≥ 135 mm Hg or the DBP ≥ 85 mm Hg is considered as high BP[21]. Currently, the cutoffs used by Hypertension Canada to define hypertension are different from U.S. guidelines.

High BP generally develops over many years and can affect anyone, but it becomes more common as people get older. The lifetime risk for developing hypertension in older adults has been estimated to be 90%[22]. People can have high BP for years without any signs or symptoms, and most hypertensive people have no symptoms at all[19]. Occasionally, people with high BP may experience headaches, shortness of breath, dizziness, chest pain, heart palpitations, and nosebleeds. Nevertheless, these signs and symptoms are not specific and usually do not occur until high BP has reached a severe or life-threatening stage[19]. High BP is a serious warning sign and can be a silent killer. BP readings should be checked in regular doctor's appointments. Uncontrolled and undiagnosed high BP can lead to serious health problems, including heart attack, stroke, kidney disease, blindness, ruptured blood vessels, and cognitive impairment.

1.3 Risk Factors for Hypertension

Hypertension has been identified as a multi-factorial trait resulting from environmental and biological factors[9,23]. Multiple factors may cause and increase the risk of hypertension, including physical, hereditary, or behavioral. Broadly, these risk factors belong to two major categories, modifiable and non-modifiable. Conditions that can be altered or controlled by making specific lifestyle changes are modifiable risk factors. In contrast, non-modifiable risk factors consist of those conditions that a person cannot change or control. Having one of these risk factors will not

necessarily lead to the development of hypertension; however, the presence of more risk factors in a person will increase the likelihood of developing hypertension.

1.3.1 Modifiable Risk Factors

Lack of Physical Activity. Body movement generated by the skeletal muscles' contraction that eventually raises energy expenditure over resting levels is described as physical activity[24,25]. Physical activity includes routine daily tasks such as occupational tasks, commuting, household activities, and planned and repetitive movements to improve and maintain health[24,25]. Inadequate physical activity increases the risk of high BP. Although precise mechanisms by which physical activity reduces BP and prevents hypertension are not clear[25], several studies have demonstrated a positive effect of physical activity on the risk of developing hypertension[26]. Current guidelines suggest increasing physical activity as a crucial lifestyle modification to prevent hypertension[27,28]. Lack of physical activity is also responsible for the increased risk of being overweight.

Overweight or Obesity. Excess weight generates an additional strain on the heart and circulatory system. Generally, more blood is required to supply oxygen and nutrients to the tissues of an overweight person. This increased volume of blood circulation through the blood vessels creates extra pressure on the artery walls. Maintaining healthy body weight is vital for hypertension prevention[29–31].

An Unhealthy Diet. Having good healthy nutrition from different sources is crucial for health. As indicated by the Dietary Approaches to Stop Hypertension (DASH) trials, a diet that emphasizes vegetables, fruits, and low-fat dietary products; includes whole grains, fish, poultry, and nuts; and is reduced in fat, red meat, sweets, and sugar-containing beverages substantially lower BP in both hypertensive and normotensive individuals[32,33]. Excess salt (sodium) in the food can cause the body to hold fluid, which eventually raises BP. Reduction of salt intake is often recommended as

a critical measure to prevent hypertension[34]. Foods that contain high potassium are important for managing and controlling high BP because potassium reduces sodium impacts. The significance of potassium intake in controlling BP has been demonstrated in numerous epidemiological and clinical studies, and an inverse association between potassium intake and high BP has been identified[35,36]. Many DASH foods serve as natural sources of potassium. The reduction of sodium intake, combined with the DASH diet, is highly recommended for hypertension prevention[33], and making healthy food choices can help lower BP.

Too Much Alcohol Consumption. Regular, excess (more than two drinks per day) alcohol consumption can lead to hypertension because it activates the body's adrenergic nervous system, resulting in constriction of blood vessels and a concurrent increase in blood flow and heart rate. A positive association between alcohol consumption and high BP was observed in many studies[37,38]. Alcohol consumption should not exceed 14 and 9 standard drinks per week for men and women, respectively, to prevent hypertension[21].

Stress. Increased stress can cause a temporary but considerable rise in BP. Besides, excess stress can contribute to poor diet, physical inactivity, and using tobacco or drinking excess alcohol that eventually increases BP. Stress does not cause hypertension directly, but it can affect its development[39,40].

Smoking and Tobacco Use. Smoking or tobacco use can also temporarily increase BP. However, tobacco chemicals can harm the lining of artery walls and cause the arteries to narrow, increasing BP. Secondhand smoke (exposure to other people's smoke) also can increase BP. Epidemiological studies on healthy subjects, hypertensive subjects, and diabetic and renal patients have demonstrated that smokers have higher BP than nonsmokers[41,42].

1.3.2 Non-modifiable Risk Factors

Age. Age is a predisposing factor for hypertension due to the wear and tear the body undergoes over time (i.e., making it more vulnerable to chronic illness). With increasing age, the body is exposed to various strains and stressors and free radicals generated in the body, which accelerate the breakdown of cell and organ functions[43]. Our blood vessels slowly lose some of their elastic quality with age, potentially leading to increased BP. Women are as likely as men to develop high BP between the ages of 45-64 years. For individuals younger than 45 years, however, the disease affects more men than women. For people 65 years or older, high BP affects more women than men[44]. However, the risk for prehypertension and high BP is increasing in children and teens, possibly due to the rising overweight.

Sex/Gender. Research has revealed that men have a higher prevalence of hypertension than women, particularly men younger than 65, who consistently have higher hypertension levels than women of the same age group[45,46]. According to one study among 18- to 29-year-old white adults, only 1.5% of women but over 5% of men reported hypertension (for black women and men, the proportions were 4% and 10%, respectively)[47]. In all World Health Organization (WHO) regions, including Canada, men have a higher prevalence of hypertension than women[48]. Such observed gender differences in hypertension are due to both biological (sex hormones, chromosomal differences, and other biological sex differences that are protective against hypertension in women) and behavioral factors (high BMI, smoking, and physical activity)[45,49].

Race/Ethnicity. High BP is more common among blacks than people of any other racial background and often occurs in blacks at an earlier age than whites[44,50]. It also tends to be more severe in blacks, which is even less likely to achieve target BP goals with treatment[51].

Family History. A family history of high BP raises the risk of developing high BP. High BP tends to run in families, as family members share similar genes, predisposing a person to high BP. There

is also the possibility that people with a family history of high BP share common environments and other relevant factors like behaviors and lifestyles that can increase their risk of hypertension[22,52].

Certain Underlying Conditions and Medications. Certain underlying conditions can also cause and increase the risk of high BP. This type of high BP, called secondary hypertension, makes up only a tiny fraction (5% to 10%) of hypertensive cases[53,54]. Several conditions can lead to secondary hypertension, including chronic kidney disease, obstructive sleep apnea, tumors, coarctation of the aorta, other disorders of the adrenal gland, pregnancy, thyroid dysfunction, and Cushing syndrome. Also, certain medications that people need to take to manage different diseases and conditions, such as birth control pills, cold remedies, decongestants, over-the-counter pain relievers, and some prescription drugs including non-steroidal anti-inflammatory drugs, can lead to secondary hypertension[54].

1.4 Hypertension Consequences

Individuals with hypertension are at higher risk for the development of not only life-changing but also possibly life-threatening conditions[55]. Left uncontrolled or undetected, high BP can lead to dangerous health complications and poor quality of life. Vascular damage produced by high BP generally starts small and then gradually builds over time.

Blood, which supplies nutrients and oxygen to vital organs and tissue, is carried throughout the body by blood vessels and major arteries. When BP becomes high, it begins to damage artery walls. Typically, damage in artery walls starts as small tears. When these tears start forming, bad cholesterol starts to attach itself to the tears while flowing through the vessels. Over time, more and more cholesterol builds up, and as a result, arteries become narrow with reduced blood flow.

When there is insufficient blood flow, tissue or organ damage can occur, such as heart attack and stroke[55].

There are many potentially devastating complications of hypertension. In the heart, uncontrolled high BP can cause several symptoms and signs such as chest pain, irregular heartbeat, coronary artery disease, enlarged left heart, heart attack, and heart failure[56]. In the brain (which depends on a nourishing blood supply to work properly and survive), high BP can cause several problems, including transient ischemic attack (TIA), stroke, dementia, and cognitive impairment[56]. Kidneys need healthy blood vessels to filter excess fluid and waste from the blood. High BP can damage blood vessels in and around the kidneys resulting in kidney disease and kidney failure. High BP can also damage blood vessels in the eyes, causing vision difficulties, such as distorted vision, blurred vision, and complete vision loss[56]. Further, high BP can also be responsible for sexual dysfunction because of blockages to the blood vessels that lead to the sexual organs.

1.5 Hypertension Burden

High BP has long been recognized as a significant health burden that affects all segments of the population. Globally, hypertension causes 17.8% (9.4 million) of deaths each year and 7% of the disease burden, making it a leading risk factor for global mortality and disease burden[48,57]. The global prevalence of hypertension in adults aged 18 years and over was around 22% in 2014[48]. The age-standardized prevalence of hypertension was 24.1% in men and 20.1% in women in 2015[58]. Hypertension is believed to be responsible for roughly 50% of deaths due to stroke and heart disease[19]. According to randomized trials and epidemiological studies, a BP reduction of 10 mm Hg systolic or 5 mm Hg diastolic is associated with a 22% reduction in coronary heart disease events, 41% reduction in stroke events, and a 41%–46% reduction in cardiometabolic mortality[59–61]. Hypertension prevalence among adults aged 25 and over is highest in Africa (30%) and lowest

in America (18%)[48]. Overall, countries with high incomes have a lower prevalence of hypertension (35%) than other countries (40%)[19,62,63].

The prevalence of hypertension in Canadian adults was 22.6% in 2012-13 and affected more than 6 million Canadians[1]. Self-reported hypertension prevalence has also increased in Canada roughly two-fold over nearly two decades[1]. An estimate shows that if Canadians live an average lifespan, over 90% will develop hypertension[64]. In Canada, hypertension is considered among the top risk factors for death, years of life lost (YLL), and disability-adjusted life years (DALYs)[65]. In Alberta, the prevalence of hypertension among adults was 21% in 2010, with a projected increase to 27% by 2020[66].

According to studies, hypertension-related disease costs US$ 370 billion globally, accounts for approximately 10% of all expenditures in healthcare, and, if indirect costs such as welfare losses due to premature death are included, the costs could be nearly 20 times greater[67]. Over ten years, hypertension can cost about US$ 1,000 billion globally in health spending[67]. In Canada, hypertension cost a total of $13.9 billion in direct healthcare spending in 2010, and projected costs are estimated to be $20.5 billion in 2020[66]. Hypertension results in over 20 million physician visits annually in Canada[68]. There were over 85 million antihypertensive drug prescriptions in Canada in 2014, with a cost of $2 billion[1]. In Alberta, the estimated cost associated with hypertension was $1.42 billion in 2010; however, a projected increase to $2.8 billion in 2020 is also reported[66].

1.6 Hypertension Prevention: Risk Prediction Model at the Core

Due to the high prevalence and global burden of hypertension, prevention and control strategies need to be a top priority. Prevention of hypertension creates an opportunity to halt and prevent the continuing costly cycle of hypertension management and its associated complications[69]. Hypertension can be prevented by the complementary application of strategies

that target the general population and individuals and groups at higher risk for hypertension. The need for early identification of at-risk individuals who could benefit from preventive interventions has led to a growing interest in hypertension risk prediction. To identify individuals apparently free of hypertension but at risk, health professionals need reliable tools to implement preventive strategies effectively. Prediction of disease outcome through modeling is a tool that can provide reasonable estimates about the future course of an illness, serve as an important adjunct in clinical practice, and help clinicians deliver better care to avoid adverse events.

1.7 Overview of Risk Prediction Models

One priority of health and clinical research is identifying people at higher risk of developing an adverse health outcome targeted for early preventative strategies and treatment[70]. For instance, individuals who are healthy but are found to have an increased risk of developing hypertension could be recommended to change their lifestyle and behaviors (e.g., physical activity, dietary pattern, alcohol consumption, smoking, etc.) to reduce their risk. Prediction modeling can play a vital role in identifying high-risk individuals. Prediction models can be used to estimate the risk of future occurrence of a health condition in an individual by utilizing different underlying demographic and clinical characteristics called risk factors that are believed to be associated with the health outcome of interest. Prediction models help predict the chance of experiencing a health outcome by an individual with a given set of risk factors.

A clinical prediction model has many practical uses, such as detecting or screening high-risk subjects for asymptomatic disease (which helps to prevent developing diseases with early interventions); predicting disease (which helps facilitate patient-doctor communication based on more objective information); and assisting in medical decision-making, as well as assisting patients in making an informed choice regarding their treatment (which helps patients make better

decisions, leading to better outcomes for them)[71,72]. Prediction models also can assist healthcare services with planning and quality management.

1.7.1 Examples of Well-Known Risk Prediction Models

Various models have been developed that mathematically combine multiple predictors to estimate the risk of the future occurrence of different health outcomes in asymptomatic subjects in the population. The majority of models predict the occurrence of a specific disease. A well-known example is the Framingham risk score, one of the most widely used prediction tools that predict an individual's 10-year cardiovascular disease (CVD) risk. This gender-specific risk score was first developed based on age, sex, low-density lipoproteins (LDL) cholesterol, high-density lipoprotein (HDL) cholesterol, BP, diabetes, and smoking to estimate the 10-year risk for coronary heart disease (CHD) using data from the Framingham Heart Study[73]. Subsequently, other disease outcomes, such as general CVD and individual CVD events (coronary, cerebrovascular, and peripheral arterial disease and heart failure), were added in the modified version of the Framingham Risk Score[74]. The United Kingdom Prospective Diabetes Study (UKPDS) Risk Engine is a widely used type 2 diabetes-specific risk tool that delivers risk estimates for coronary heart disease and stroke. Several risk factors: current age, sex, ethnicity, smoking status, presence or absence of atrial fibrillation, levels of HbA1c, SBP, total cholesterol, and HDL cholesterol were considered in this model while calculating risk, using data from the U.K. Prospective Diabetes Study[75,76]. Gail et al. presented a risk prediction model for developing breast cancer that combines information on age, age at first live birth, family history, age at menarche, breast biopsy number, and menopause to provide the probability of developing breast cancer in healthy women[77]. The diabetes risk score, also known as the Finnish Diabetes Risk Score (FINDRISC), is a prediction tool to identify patients at risk of developing diabetes. FINDRISC uses age, BMI, physical activity,

vegetable and fruit intake, medical treatment of hypertension, history of hyperglycemia, and family history to determine the risk of developing diabetes[78]. Numerous other prediction models have been used in many different areas, including public health, clinical practice, diagnostic work-up (test ordering, starting treatment), therapeutic decision-making (surgical decision making, the intensity of treatment, delaying treatment), and research (inclusion and confounding adjustment in a randomized control trial).

1.7.2 Methods in Risk Prediction Models

While specific details may vary between clinical risk prediction models, the goals and processes of developing prediction models are mostly similar. A research question or objective is defined first, and relevant data are collected from the study population, usually longitudinal cohort data. The collected data should contain information on everyone's intended outcome status, demographics, health status, relevant risk factors for the outcome, and any other relevant aspects of the study question. The selection of candidate variables as potential predictors for analysis is based on clinical and statistical viability from all available variables. A predictive model is derived using an appropriate modeling strategy from the chosen candidate variables, and its utility is internally validated.

The conventional approach to developing prediction models is to build a single model from a dataset of individuals with known outcomes and then apply the developed model to predict future individuals' outcomes[79]. The choice of model to be fitted often depends on the nature of the endpoint. Regression methods, such as logistic regression (for binary endpoint/outcome) and Cox regression (for time-to-event endpoint/outcomes), are the most frequently used algorithms to fit prediction models. Many risk prediction models have been developed using logistic regression to identify individuals at high risk for type 2 diabetes[78,80], breast cancer[81], CVD in type 2 diabetes

patients[82,83], chronic kidney disease[84,85], etc. Many risk prediction models also have been developed using the Cox regression algorithm to assess general CVD and individual CVD events risk[74], the absolute risk of CHD[75] and stroke[76] in people with type 2 diabetes, and predicting the risk of breast cancer[77]. Over the last few years, machine learning algorithms achieved significant successes across a broad range of fields because of their advantages, such as their ability to model nonlinear relations and the accuracy of their overall predictions[86]. Decision trees, random forest, penalized regression models, neural networks, and support vector machines are examples of machine learning algorithms[87]. However, machine-learning algorithms sometimes struggle with reliable probabilistic estimation and interpretability[88,89]. Moreover, in clinical applications, machine-learning methods often demonstrate mixed performance[90–94]. Once the modeling approach is defined and the data are collected, the prediction model can be fitted to the data using statistical software.

Most fundamental steps are common in all prediction modeling despite their variations in the modeling process. We outline here, in brief, some necessary steps of prediction modeling regardless of their kind.

1. **Identify the appropriate data source and format a cohort**. For developing prediction models, generally, longitudinal data are used where there is follow-up information. This follow-up data provides information on participants who are free of the outcome at baseline, but after a specific follow-up time, either have or have not developed the outcome. When there is no follow-up information, like ours, one approach can be linking the cohort data with a data source from where follow-up information can be captured. Our study did that by linking a population-based prospective cohort Alberta's tomorrow project (ATP)[95] data with two other data sources from Alberta's administrative health data--hospital

discharge abstract data and physician/practitioner claims data. Administrative health data was linked to ATP data using encrypted personal health numbers common to all data sources. Once the data source is identified, a cohort is formatted using a set of inclusion-exclusion criteria as per the study's requirements. For our study, the cohort consists of participants enrolled in ATP, adults aged 35-69 years at enrollment, free of hypertension at baseline, and consented to have their data linked with Alberta's administrative health data.

2. **Assess the required sample size**. Prediction models should be developed to reflect the patterns existing in the underlying data accurately[96]. A small sample or small dataset often leads to inaccuracy in the model. If the sample is too small, analysis results will have wide confidence intervals, low statistical power, low precision, and biased results. There are various sample size formulae available, but no consensus on the best approach[97,98]. Events per variable (EPV) defined as the ratio of the number of individuals with the outcome event to the number of candidate variables (precisely, the number of regression coefficients), is a frequently used approach to determine the sample size in the predictive modeling[99]. It is recommended, if a variable selection is performed, the number of regression coefficients should refer to the initial set of variables before variable selection[100]. Different simulation studies had suggested a minimum EPV of 5 to 20 to provide reliable results when prediction models are developed using logistic and Cox regression[101–105]. An EPV of 10 is often used as the thumb rule and is widely recommended for multivariable logistic and Cox regression models[99]. However, these EPV recommendations primarily emphasize the regression coefficients' precision and accuracy instead of predictive ability measures. Ogundimu et al.[99] suggested considering an EPV $\geq$ 20 when a dataset includes low-prevalence binary

variables. Their suggestion was based on the regression coefficients' stability and precision and their effect on the models' predictive performance (e.g., the C-statistic, D-statistic, and R^2) using Cox regression. Since our cohort (sample) included all available incident hypertension cases within the study period, the cohort (sample) size is already maximized. However, to ensure that our cohort is sufficiently large for our model building purpose, we applied the EPV $\geq$ 20 rule.

3. **Select candidate variables**. Before commencing the analysis, a list of all available potential candidate variables needs to be compiled. These candidate variables are generally selected based on a literature search, variables used in the past, and discussion with content experts. In addition to those, we also considered the following set of criteria for selecting candidate variables in our analysis[106–108]:
    - clinical availability in a timely and cost-effective manner at the time when a patient visited a physician or clinic/hospital, such that availability does not require a time-consuming or costly procedure.
    - whether the variable was relevant in predicting the outcome and
    - whether the variable is likely to add substantial prognostic information beyond what other variables provide.
4. **Deal with missing data.** Missing data values is a common phenomenon, and conclusions drawn from the data can be heavily affected by missing values[109]. Missing data creates several problems, including reduced statistical power, biased estimated parameters, reduced sample representation, and complicacy in the analysis, which leads to invalid conclusions[110]. Among the many reasons for missing data, nonresponse and dropouts are most common. Missing data can also occur because of improper data collection or mistakes

made in data entry. It is imperative to know why the data are missing to handle the remaining data properly. Missing data can be dealt with using different approaches such as removing the missing values, imputing them, or modeling them. The most common and easiest way of dealing with the missing data is to omit the missing cases and perform the analysis on the remaining data; a technique called listwise deletion or complete case (or available case) analysis. To fill in or impute missing values is another approach to dealing with the missing data. In this technique, the missing cases are replaced with an estimated value calculated based on other available relevant information. Different ways of imputation extend from very simple to quite complex. The simplest form of imputation is to substitute each missing value with the observed values' mean for the corresponding variable. The last observation carried forward (LOCF) is another imputation approach that has been used widely. In LOCF, all the missing cases are replaced by the last observed value[111]. Fitting a regression model to the observed cases and then using that model to predict the missing cases is another imputation approach that gives a better result. Multiple imputation, the soundest strategy for handling missing data consists of replacing the missing values with a set of plausible values that accommodate both the natural variability and the correct values' uncertainty, rather than just substituting a single value for each piece of missing data. This technique predicts the missing values by utilizing the existing information from other variables[112]. Then, the missing values are substituted by the predicted values, and a complete dataset is created. This complete dataset is called an imputed dataset. Multiply imputed datasets are created by iterating this process repeatedly. These multiply imputed datasets are then analyzed by applying the standard statistical procedures for complete data, producing multiple results. Subsequently, a single overall

analysis result is created by combining these multiple results. We used multiple imputation in our study to impute the missing values.

5. **Assess collinearity.** Collinearity, a statistical phenomenon where two or more predictor variables in a prediction model are highly correlated or associated. As a result, it is hard to get reasonable estimates of their distinct effects on the outcome variable. Collinearity increases the coefficients' standard errors and makes some variables statistically insignificant when they should be significant. Although collinearity does not bias coefficients and reduces the model's predictive power or reliability as a whole, it does make the coefficients unstable. The variance inflation factor (VIF) is one common way to measure collinearity, which evaluates how much the variance of an estimated regression coefficient increases if variables are correlated. If the variables are not correlated, the VIFs will all be 1. From the list of candidate variables, those highly correlated should be excluded before starting model building. Collinearity among the variables was tested in our study using the VIF with a threshold of 2.5[113].
6. **Perform variable selection.** Regardless of the modeling technique used, one needs to apply appropriate variable selection methods during the model building stage. Selecting relevant variables for inclusion in a model is often considered the most critical and challenging part of model building[114]. Variable selection is a process where a subset of relevant variables from a large amount of data is selected to filter the dataset down to the smallest possible subset of accurate variables. It is imperative to identify the relevant variables from a dataset and remove less significant variables that contribute to the outcome to achieve the prediction model's better accuracy. The variable selection offers enhanced model performance by mitigating the risk of overfitting, improved computational speed

and time, decreased computational requirements, and more straightforward interpretability. There are different ways of selecting variables for a final model. However, there is no consensus on which method is best[114]. The standard variable selection method includes univariate analysis followed by multivariable analysis based on p-values, forward selection, backward elimination, and stepwise selection. In the machine learning domain, variable selection is called feature selection, a core concept that massively impacts machine learning algorithms' performance. Feature selection methods can be classified into three categories: filter, wrapper, and embedded methods. We employed both variable selection and feature selection methods in our model building process.

7. **Apply appropriate modeling methods.** Fitting the correct model depends on the nature of the outcome and the study's objective. Both traditional regression-based models or newly emerging machine learning-based models can be applied to develop a prediction model. Within regression-based models, Cox proportional hazard model is most frequently used for survival data (for time-to-event outcomes), which we used in our study. Due to their remarkable success in achieving improved predictive accuracy and comparing their predictive performance with traditional regression-based models, we also developed some machine learning algorithms. This study's machine learning algorithms include random survival forest, gradient boosting, and penalized regression models such as lasso, ridge, and elastic net.

1.7.3 Model Validation

There are two primary components of prediction modeling: model development and model validation. A model can be validated either internally (using the same data or data source) or externally (using new data from a different data source)[115]. The purpose of model validation is to

demonstrate that the model is accurate for the intended population (dataset) for whom the model was developed and performs well in other populations (datasets) that were not used to create the model[70]. In the split-sample method, one procedure commonly employed in prediction modeling, the dataset is split into two sections (often in a 2:1 ratio); one is used for model derivation and the second for internal validation[116]. However, for certain datasets, this method is often limited by small study power and more significant variability[117]. Also, randomly splitting the data does not guarantee that the divided data represents the target population, which could be a bias source, limiting the model's generalizability to other populations[117]. 'K-fold cross-validation' and 'bootstrapping' are two popular methods that improve the split-sample method and produce better results regarding bias and variability[117]. K-fold cross-validation and bootstrapping are also better when the sample size is small and when external validation is not readily available. K-fold cross-validation starts with randomly partitioning the original sample into k equal size subsamples. Only one subsample out of these k subsamples is kept as the validation data to test the model. The remaining k-1 subsamples are utilized as training data to derive the model. A total of k times (the folds) this process is replicated, with each k subsample used only once as the validation data. Finally, a single estimate is produced by averaging (or otherwise combining) the k results from the folds. K-fold cross-validation has the significant advantage that all observations are utilized to derive and validate the model, with each observation used only once for validation. As a result, this process has less chance to succumb to a particular biased division of the data.

On the other hand, bootstrapping involves taking random samples with replacement from the data and creating separate sub-cohorts for model selection and validation[117]. This process often occurs hundreds of times, each time producing a model for parameter estimation. Despite having some advantages compared to other methods, like attaining minimum variance, bootstrapping is

more complex to analyze and interpret due to the methods used and the amount of computation required. Studies have suggested that no particular performance difference exists between the two methods for prediction models. The procedures mentioned above for model validation pertain to internal validation, which does not examine its generalizability. The model's generalizability can be established by applying the model to entirely new data collected from an appropriate (representative) patient population not used in the development process. Most studies evaluating prediction models focus on internal validity instead of external validity[118]. Internal validation does not guarantee generalizability, and thus external validation is necessary before implementing prediction models into clinical practice[119].

1.7.4 Evaluating Model Performance

There are different methods and metrics to assess the performance of a prediction model. For binary and survival outcomes, the most commonly used measures include the Brier score to indicate overall model performance, the concordance statistic (also known as the C-statistic) for discriminative ability, and goodness-of-fit statistics for calibration[120]. The model's overall performance is quantified by considering the distance between the actual outcome and the predicted outcome. The Brier score is used to calculate the model's overall performance and is measured by calculating the squared differences between actual binary outcomes and predictions calculated by the model[120]. The range of values that the Brier score of a model can take lies between 0 and 0.25, with 0 indicating a perfect model and 0.25 showing a non-informative model with only a 50% incidence of the outcome[120]. Discrimination is defined as the model's ability to distinguish between participants who do or do not experience the event of interest (disease outcome such as hypertension). A C-statistic (which equals the area under the receiver operating characteristic [ROC] curve [AUC] for binary outcomes) is commonly employed for this purpose. A C-statistic

value refers to the probability that a randomly selected subject who experienced the outcome will have a higher predicted probability of having the outcome than a randomly selected subject who did not experience the event[121]. The C-statistic can range from 0.5 to 1.00, with higher values indicating better predictive models. A C-statistic of 0.5 suggests the model's performance in predicting an outcome is no better than the random chance. At the same time, a C-statistics of 1 indicates the model perfectly distinguishes those who will experience a particular outcome and those who will not.

The agreement between observed outcomes and predictions made by the model is calibration[120]. Model calibration measures the predictions' validity and determines whether the predictions based on the risk prediction model align with what is observed within the study cohort. A calibration plot is a method that visually inspects calibration and presents a plot for predicted against expected probabilities. It also uses the Hosmer-Lemeshow test to assess calibration. In a calibration plot, predictions are plotted on the x-axis and the observed outcome on the y-axis. In the y-axis, the plot contains only 0 and 1 values for binary outcomes. Different smoothing techniques (e.g., the loess algorithm) can be employed to estimate the observed probabilities of the outcome with respect to the predicted probabilities. Perfect predictions should be on the 45° line suggesting that predicted risks are correct[120]. An alternative assessment of calibration is to categorize predicted risk and assess whether the event rate corresponds to the average predicted risk in each risk group. The Hosmer-Lemeshow goodness-of-fit-test plots a graphical illustration to assess whether the observed event rate corresponds to the expected event rate in the model population subgroups.

1.7.5 Generation of Point Scoring System

In practice, the predicted probability of an outcome calculated by the model needs to be presented in a simplified way to be easily used[122]. Multivariable prediction models are relatively complex, and the computations using the prediction model can be tedious[123]. The points scoring system simplifies the tedious calculation of prediction models by assigning integer points to a given risk factor so that clinicians can easily approximate risk by summing integer points based on each risk factor's presence/absence. The points scoring system is generally formulated around categories[123]. To aid in interpreting risk estimates, tables of comparative risks are also often provided[123]. These comparative risks can motivate patients to change risk factors to reduce their chance of developing hypertension. There are different ways to build a point scoring system. Point scoring can be done by transforming the regression coefficient or relative risk (odds ratio or hazard ratio) for each predictor to integers[122]. We applied a point scoring system proposed by Sullivan et al.[123] to develop a point (scoring) system that can be used clinically to estimate an individual's risk of developing hypertension without a calculator or computer.

1.7.6 Existing Research on Hypertension Prediction Models

Like other health areas, risk prediction models are also common in hypertension, which estimates the probability that a currently healthy individual with specific risk factors will develop hypertension in the future within a specified time. A thorough review was performed to identify scientific publications that tested and developed clinical risk prediction tools for hypertension. This process was augmented by reference snowballing[124], where relevant papers were reviewed for both articles the authors cited and articles citing that paper. This process was repeated for every paper considered relevant. Among the identified hypertension prediction models, the most important ones are discussed here briefly. Pearson et al.[125] developed the first hypertension prediction model, known as the Johns Hopkins multiple risk equations, based in the USA in 1990.

A Cox proportional hazards regression model containing age, SBP at baseline, paternal history of hypertension, and BMI predicted hypertension. Parikh et al.[16] developed a hypertension incidence prediction score in 2008, commonly known as the Framingham hypertension risk score. Age, sex, BMI, SBP and DBP, cigarette smoking, and parental hypertension were used to predict hypertension. Scores were developed for predicting the 1-, 2-, and 4-year risk for new-onset hypertension. Paynter et al.[17] developed a series of models based on clinical characteristics and blood biomarkers. A prospective cohort of normotensive women aged 45 and older from the Women's Health Study was used to develop the logistic regression models to predict incident hypertension. Kivimaki et al.[126,127] created two models known as the Whitehall II risk scores and Whitehall II repeat measures risk score based on the British population. Among the risk factors, age, sex, parental hypertension, SBP, DBP, BMI, and cigarette smoking were considered in model building. Bozorgmanesh et al.[128] developed a point-score system for predicting incident hypertension by converting Weibull regression coefficients of predictors to integer values in an Iranian population. Among women, family history of premature CVD, waist circumference, SBP, and DBP were predictive of hypertension, whereas, among men, smoking, SBP, and DBP were identified as predictors. Chien et al.[13] developed point-based prediction models for new-onset hypertension for ethnic Chinese based on clinical and biochemical variables, including sex, age, BMI, SBP, DBP, white blood count, fasting glucose, and uric acid. Lim et al.[18] developed a hypertension incidence prediction model in a middle-aged Korean population. They used the same risk factors that were used for creating the Framingham hypertension risk score. Fava et al.[14] aggregated genetic information obtained from many markers into a single genetic risk score to see to what extent genetics can predict the incidence of future hypertension or cardiovascular events. Still, they did not find any improvement in the prediction of incident hypertension using genetic

information outside that provided by traditional risk factors such as sex, age, obesity, diabetes mellitus, family history of hypertension, smoking status, etc. Otsuka et al.[4] developed a risk prediction model in a Japanese male population to estimate the 4-year risk of incident hypertension. They used age, SBP, DBP, BMI, parental history of hypertension, current smoking status, and excessive alcohol intake as their model predictors.

Most of the studies defined hypertension as either SBP ≥ 140 mm Hg or DBP ≥ 90 mm Hg or the use of antihypertensive drugs. While developing different hypertension prediction models, participants were followed-up for 3 to 30 years. Most prediction models were built using traditional risk factors, and only a few with genetic risk factors. The most commonly used risk factors included in different models were age, SBP, DBP, BMI, gender, and parental history of hypertension. In recent times, genetic risk factors are incorporated increasingly as model predictors. However, the genetic risk factors inclusion does not improve the model's performance significantly in most cases[129].

1.8 Study Rationale

The increasing availability and richness of datasets create more opportunities for developing and deploying clinical risk prediction models. Several prediction models (or risk scores) have been developed over the past decades to predict a person's chance of developing hypertension. Such prediction helps identify individuals at risk of hypertension so that primary prevention strategies can be targeted. However, the identification of such at-risk individuals remains a challenge[130]. Multivariable hypertension risk prediction models have been used in different countries to serve that purpose[131]. These prediction models were constructed considering various risk factors for hypertension using data from diverse populations. Each population has its probability of getting the disease, and each population may have a different distribution of risk factors, which may weigh

differently in determining the disease[132]. Prediction models are determined by an equation that includes risk factors (e.g., age, BMI) and risk coefficients (multiplying factors) that attribute an etiological weight to single factors[132]. These elements change according to the type of population, particularly when very different cultures are compared (i.e., European and Asian countries). Due to the differences in the risk factors prevalence and incident hypertension between populations, a prediction model's performance can differ substantially by population. As a result, the prediction model's accuracy is often acceptable for that index population and is not necessarily generalizable to populations other than that for which the model was developed[132]. Our review identified hypertension prediction models developed in different countries, but none have been developed in the Canadian context so far to the best of our knowledge. As such, developing a hypertension prediction model using one of Canada's largest cohort studies will provide local clinicians and health care providers assistance in clinical decision-making, planning, and proper management of healthcare services regarding hypertension.

Accurate and reliable identification of individuals at high risk of developing hypertension allows for interventions that may help prevent hypertension and related cardiovascular complications. Inaccurate risk estimation can lead to failure to identify individuals who are at risk of developing hypertension. Misclassification can lead to ineffective interventions and unnecessary exposure to treatment in patients at low risk and missed opportunities to intervene in those most susceptible to developing hypertension. To see whether the prediction model's accuracy can be improved using machine learning algorithms, we will establish several machine learning algorithms to predict hypertension incidence and compare their predictive performance with traditional statistical models developed earlier.

1.9 Research Objectives

This research aims to develop a robust hypertension prediction model for the general population using Alberta's Tomorrow Project (ATP) cohort data.

The specific objectives are:

Objective 1: Conduct a comprehensive systematic review to identify risk factors and prediction models for hypertension incidence and perform a meta-analysis to evaluate the current model's predictive performance.

Objective 2: Develop a risk prediction model for incident hypertension in a Canadian cohort using a traditional modeling approach and converting it into a risk score for daily clinical practice use.

Objective 3: Develop machine learning algorithms to predict hypertension incidence and compare their predictive performance with a traditional statistical model in a large survival data.

This study's three specific objectives have been achieved as follows. The systematic review provided information on all past hypertension prediction models and the variables considered in developing the model. The meta-analysis provided an overall predictive performance of existing models and helped compare existing traditional regression-based models' predictive performance with machine learning-based models. Linked administrative health data provided information on outcomes, and ATP data supplied variables to build a new traditional prediction model using Cox proportional hazard modeling. Machine learning algorithms developed in a survival context using the same data sources provided an alternative class of prediction models on which predictive performances can be compared with the traditionally developed model.

Chapter 1 summarizes the background information on hypertension and its risk factors, the consequences of hypertension, hypertension burden, an overview of the risk prediction models,

and their role in hypertension preventions. Chapters 2, 3, and 4 are the main body of the dissertation and represent three papers. Chapter 2 describes a systematic review to identify existing models of hypertension prediction and associated risk factors. It also provides a meta-analysis to evaluate and compare the predictive performance of existing hypertension prediction models. Chapter 3 develops a new hypertension prediction model applying a traditional statistical modeling approach using a large retrospective Canadian cohort data. It also created a risk score from the developed model to facilitate clinical use. Chapter 4 develops machine learning models for hypertension prediction and compare their predictive performance with the traditionally developed model in chapter 3. Chapter 5 summarizes the study's main findings, strengths and limitations, and future research directions.

1.10 References

1. Padwal R.S., Bienek A, McAlister FA, Campbell NRC. Epidemiology of Hypertension in Canada: An Update. *Can J Cardiol*. Published online 2016. doi:10.1016/j.cjca.2015.07.734
2. CDC. High Blood Pressure Fact Sheet. *Div Hear Dis Stroke Prev*. Published online 2016.
3. Mendis S, Puska P, Norrving B. Global atlas on cardiovascular disease prevention and control. *World Heal Organ*. Published online 2011.
4. Otsuka T, Kachi Y, Takada H, et al. Development of a risk prediction model for incident hypertension in a working-age Japanese male population. *Hypertens Res*. Published online 2015. doi:10.1038/hr.2014.159
5. Grossman DC, Bibbins-Domingo K, Curry SJ, et al. Behavioral counseling to promote a healthful diet and physical activity for cardiovascular disease prevention in adults without cardiovascular risk factors: U.S. preventive services task force recommendation statement. *JAMA - J Am Med Assoc*. Published online 2017. doi:10.1001/jama.2017.7171
6. Julius S, Nesbitt SD, Egan BM, et al. Feasibility of Treating Prehypertension with an Angiotensin-Receptor Blocker. *N Engl J Med*. Published online 2006. doi:10.1056/nejmoa060838
7. Khoury MJ, Iademarco MF, Riley WT. Precision Public Health for the Era of Precision Medicine. *Am J Prev Med*. Published online 2016. doi:10.1016/j.amepre.2015.08.031
8. Chowdhury MZI, Turin TC. Precision health through prediction modelling: Factors to consider before implementing a prediction model in clinical practice. *J Prim Health Care*. 2020;12(1):3-9. doi:10.1071/HC19087
9. Pazoki R, Dehghan A, Evangelou E, et al. Genetic predisposition to high blood pressure

and lifestyle factors: Associations with midlife blood pressure levels and cardiovascular events. *Circulation*. Published online 2018. doi:10.1161/CIRCULATIONAHA.117.030898

10. Usher-Smith JA, Silarova B, Schuit E, Moons KGM, Griffin SJ. Impact of provision of cardiovascular disease risk estimates to healthcare professionals and patients: a systematic review. *BMJ Open*. Published online 2015. doi:10.1136/bmjopen-2015-008717
11. Lopez-Gonzalez AA, Aguilo A, Frontera M, et al. Effectiveness of the Heart Age tool for improving modifiable cardiovascular risk factors in a Southern European population: A randomized trial. *Eur J Prev Cardiol*. Published online 2015. doi:10.1177/2047487313518479
12. Chowdhury MZI, Naeem I, Quan H, et al. Summarising and synthesising regression coefficients through systematic review and meta-analysis for improving hypertension prediction using metamodelling: Protocol. *BMJ Open*. 2020;10(4). doi:10.1136/bmjopen-2019-036388
13. Chien KL, Hsu HC, Su TC, et al. Prediction models for the risk of new-onset hypertension in ethnic Chinese in Taiwan. *J Hum Hypertens*. Published online 2011. doi:10.1038/jhh.2010.63
14. Fava C, Sjögren M, Montagnana M, et al. Prediction of blood pressure changes over time and incidence of hypertension by a genetic risk score in swedes. *Hypertension*. Published online 2013. doi:10.1161/HYPERTENSIONAHA.112.202655
15. Kshirsagar A V., Chiu Y lin, Bomback AS, et al. A hypertension risk score for middle-aged and older adults. *J Clin Hypertens*. Published online 2010. doi:10.1111/j.1751-7176.2010.00343.x

16. Parikh NI, Pencina MJ, Wang TJ, et al. A risk score for predicting near-term incidence of hypertension: The Framingham Heart Study. *Ann Intern Med*. Published online 2008. doi:10.7326/0003-4819-148-2-200801150-00005

17. Paynter NP, Cook NR, Everett BM, Sesso H.D., Buring J.E., Ridker PM. Prediction of Incident Hypertension Risk in Women with Currently Normal Blood Pressure. *Am J Med*. Published online 2009. doi:10.1016/j.amjmed.2008.10.034

18. Lim NK, Son KH, Lee KS, Park HY, Cho MC. Predicting the Risk of Incident Hypertension in a Korean Middle-Aged Population: Korean Genome and Epidemiology Study. *J Clin Hypertens*. Published online 2013. doi:10.1111/jch.12080

19. WHO. *A Global Brief on Hypertension; World Silent Killer, Globa Health Crisis*.; 2013.

20. Whelton PK, Carey RM, Aronow WS, et al. *2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults a Report of the American College of Cardiology/American Heart Association Task Force on Clinical Pr*. Vol 71.; 2018. doi:10.1161/HYP.0000000000000065

21. Leung AA, Daskalopoulou SS, Dasgupta K, et al. Hypertension Canada's 2017 Guidelines for Diagnosis, Risk Assessment, Prevention, and Treatment of Hypertension in Adults. *Can J Cardiol*. Published online 2017. doi:10.1016/j.cjca.2017.03.005

22. Vasan RS, Beiser A, Seshadri S, et al. Residual lifetime risk for developing hypertension in middle-aged women and men: The Framingham Heart Study. *J Am Med Assoc*. Published online 2002. doi:10.1001/jama.287.8.1003

23. Li JJ, Fang CH, Hui RT. Is hypertension an inflammatory disease? *Med Hypotheses*. 2005;64(2):236-240. doi:10.1016/j.mehy.2004.06.017

24. Caspersen CJ, Powell KE, Christenson GM. Physical activity, exercuse and physical fitness: definitions and distinctions for health-related research. *Public Heal Rep*. Published online 1985.

25. Diaz KM, Shimbo D. Physical activity and the prevention of hypertension. *Curr Hypertens Rep*. Published online 2013. doi:10.1007/s11906-013-0386-8

26. Warburton DER, Charlesworth S, Ivey A, Nettlefold L, Bredin SSD. A systematic review of the evidence for Canada's Physical Activity Guidelines for Adults. *Int J Behav Nutr Phys Act*. Published online 2010. doi:10.1186/1479-5868-7-39

27. Chobanian A V., Bakris GL, Black HR, et al. Seventh report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure. *Hypertension*. Published online 2003. doi:10.1161/01.HYP.0000107251.49515.c2

28. Brook RD, Appel LJ, Rubenfire M, et al. Beyond medications and diet: Alternative approaches to lowering blood pressure: A scientific statement from the american heart association. *Hypertension*. Published online 2013. doi:10.1161/HYP.0b013e318293645f

29. Stevens VJ. Weight loss intervention in phase 1 of the Trials of Hypertension Prevention. The TOHP Collaborative Research Group. *Arch Intern Med*. Published online 1993. doi:10.1001/archinte.153.7.849

30. Hypertension Prevention Trial Research Group. The Hypertension Prevention Trial: Three-Year Effects of Dietary Changes on Blood Pressure. *Arch Intern Med*. Published online 1990. doi:10.1001/archinte.1990.00390130131021

31. Cutler JA. Effects of weight loss and sodium reduction intervention on blood pressure and hypertension incidence in overweight people with high-normal blood pressure: The trials of hypertension prevention, phase II. *Arch Intern Med*. Published online 1997.

doi:10.1001/archinte.157.6.657

32. Appel LJ, Moore TJ, Obarzanek E, et al. A Clinical Trial of the Effects of Dietary Patterns on Blood Pressure. *N Engl J Med*. Published online 1997. doi:10.1056/nejm199704173361601

33. Sacks FM, Svetkey LP, Vollmer WM, et al. Effects on Blood Pressure of Reduced Dietary Sodium and the Dietary Approaches to Stop Hypertension (DASH) Diet. *N Engl J Med*. Published online 2001. doi:10.1056/nejm200101043440101

34. Appel LJ, Brands MW, Daniels SR, Karanja N, Elmer PJ, Sacks FM. Dietary approaches to prevent and treat hypertension: A scientific statement from the American Heart Association. *Hypertension*. Published online 2006. doi:10.1161/01.HYP.0000202568.01167.B6

35. He FJ, MacGregor GA. Fortnightly review: Beneficial effects of potassium. *BMJ*. Published online 2001. doi:10.1136/BMJ.323.7311.497

36. Houston MC, Harper KJ. Potassium, magnesium, and calcium: their role in both the cause and treatment of hypertension. *J Clin Hypertens (Greenwich)*. Published online 2008. doi:10.1111/j.1751-7176.2008.08575.x

37. Puddey IB, Beilin L.J. Alcohol is bad for blood pressure. In: *Clinical and Experimental Pharmacology and Physiology*. ; 2006. doi:10.1111/j.1440-1681.2006.04452.x

38. Puddey IB, Beilin LJ, Vandongen R, Rouse IL, Rogers P. Evidence for a direct effect of alcohol consumption on blood pressure in normotensive men: A randomized controlled trial. *Hypertension*. Published online 1985. doi:10.1161/01.HYP.7.5.707

39. Kulkarni S, O'Farrell I, Erasi M, Kochar MS. Stress and hypertension. *Wis Med J*. Published online 1998. doi:10.1177/003693307301800413

40. Lucini D, Riva S, Pizzinelli P, Pagani M. Stress management at the worksite: Reversal of symptoms profile and cardiovascular dysregulation. *Hypertension*. Published online 2007. doi:10.1161/01.HYP.0000255034.42285.58

41. Virdis A, Giannarelli C, Fritsch Neves M, Taddei S, Ghiadoni L. Cigarette Smoking and Hypertension. *Curr Pharm Des*. Published online 2010. doi:10.2174/138161210792062920

42. Talukder MAH, Johnson WM, Varadharaj S, et al. Chronic cigarette smoking causes hypertension, increased oxidative stress, impaired NO bioavailability, endothelial dysfunction, and cardiac remodeling in mice. *Am J Physiol - Hear Circ Physiol*. Published online 2011. doi:10.1152/ajpheart.00868.2010

43. Buttar H.S., Li T, Ravi N. Prevention of cardiovascular diseases: Role of exercise, dietary interventions, obesity and smoking cessation. *Exp Clin Cardiol*. Published online 2005.

44. Mozaffarian D, Benjamin EJ, Go AS, et al. Heart disease and stroke statistics-2016 update a report from the American Heart Association. *Circulation*. Published online 2016. doi:10.1161/CIR.0000000000000350

45. Everett B, Zajacova A. Gender differences in hypertension and hypertension awareness among young adults. *Biodemography Soc Biol*. Published online 2015. doi:10.1080/19485565.2014.929488

46. Choi HM, Kim HC, Kang DR. Sex differences in hypertension prevalence and control: Analysis of the 2010-2014 Korea national health and nutrition examination survey. *PLoS One*. Published online 2017. doi:10.1371/journal.pone.0178334

47. Cutler JA, Sorlie PD, Wolz M, Thom T, Fields L.E., Roccella E.J. Trends in hypertension prevalence, awareness, treatment, and control rates in United States adults between 1988-

1994 and 1999-2004. *Hypertension*. Published online 2008. doi:10.1161/HYPERTENSIONAHA.108.113357

48. World Health Organization. *GLOBAL STATUS REPORT on Noncommunicable Diseases 2014 - "Attaining the Nine Global Noncommunicable Diseases Targets; a Shared Responsibility"*; 2014.

49. Sandberg K, Ji H. Sex differences in primary hypertension. *Biol Sex Differ*. Published online 2012. doi:10.1186/2042-6410-3-7

50. Nwankwo T, Yoon SS u., Burt V, Gu Q. Hypertension among adults in the United States: National Health and Nutrition Examination Survey, 2011-2012. *NCHS Data Brief*. Published online 2013.

51. Ferdinand KC, Armani AM. The management of hypertension in African Americans. *Crit Pathw Cardiol*. Published online 2007. doi:10.1097/HPC.0b013e318053da59

52. Know Your Risk for High Blood Pressure | cdc.gov. Accessed January 21, 2021. https://www.cdc.gov/bloodpressure/risk_factors.htm

53. Vongpatanasin W. Resistant hypertension: A review of diagnosis and management. *JAMA - J Am Med Assoc*. Published online 2014. doi:10.1001/jama.2014.5180

54. Charles L, Triscott J, Dobbs B. Secondary Hypertension: Discovering the Underlying Cause. *Am Fam Physician*. Published online 2017.

55. The Effects of Hypertension on the Body. Accessed January 2, 2021. https://www.healthline.com/health/high-blood-pressure-hypertension/effect-on-body

56. Mayo Clinic Staff. High blood pressure dangers: hypertension’s effects on your body. *Mayo Clin*. Published online 2016.

57. Lim SS, Vos T, Flaxman AD, et al. A comparative risk assessment of burden of disease

and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990-2010: A systematic analysis for the Global Burden of Disease Study 2010. *Lancet*. Published online 2012. doi:10.1016/S0140-6736(12)61766-8

58. Zhou B, Bentham J, Di Cesare M, et al. Worldwide trends in blood pressure from 1975 to 2015: a pooled analysis of 1479 population-based measurement studies with 19·1 million participants. *Lancet*. Published online 2017. doi:10.1016/S0140-6736(16)31919-5

59. Law MR, Morris JK, Wald NJ. 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*. Published online 2009. doi:10.1136/bmj.b1665

60. Grossman E. Blood pressure: The lower, the better: The con side. *Diabetes Care*. Published online 2011. doi:10.2337/dc11-s245

61. M.D. C, J.E. B, N. B, M. E, G.A. S, G. D. The contributions of risk factor trends to cardiometabolic mortality decline in 26 industrialized countries. *Int J Epidemiol*. Published online 2013.

62. Forouzanfar MH, Liu P, Roth GA, et al. Global burden of hypertension and systolic blood pressure of at least 110 to 115mmHg, 1990-2015. *JAMA - J Am Med Assoc*. Published online 2017. doi:10.1001/jama.2016.19043

63. GHO | By category. *WHO*.

64. Joffres MR, Campbell NRC, Manns B, Tu K. Estimate of the benefits of a population-based reduction in dietary sodium additives on hypertension and its related health care costs in Canada. *Can J Cardiol*. Published online 2007. doi:10.1016/S0828-282X(07)70780-8

65. Global Burden of Disease (GBD) | Institute for Health Metrics and Evaluation. Accessed January 21, 2021. http://www.healthdata.org/gbd/visualizations/ gbd-arrow-diagram

66. Weaver CG, Clement FM, Campbell NRC, et al. Healthcare Costs Attributable to Hypertension: Canadian Population-Based Cohort Study. *Hypertension*. Published online 2015. doi:10.1161/HYPERTENSIONAHA.115.05702

67. Gaziano TA, Bitton A, Anand S, Weinstein MC. The global cost of nonoptimal blood pressure. *J Hypertens*. Published online 2009. doi:10.1097/HJH.0b013e32832a9ba3

68. Hemmelgarn BR, Chen G, Walker R, et al. Trends in antihypertensive drug prescriptions and physician visits in Canada between 1996 and 2006. *Can J Cardiol*. Published online 2008. doi:10.1016/S0828-282X(08)70627-5

69. Whelton PK, He J, Appel LJ, et al. Primary prevention of hypertension: Clinical and public health advisory from the National High Blood Pressure Education Program. *J Am Med Assoc*. Published online 2002. doi:10.1001/jama.288.15.1882

70. Ahmed I, Debray T.P., Moons KG, Riley RD. Developing and validating risk prediction models in an individual participant data meta-analysis. *BMC Med Res Methodol*. Published online 2014. doi:10.1186/1471-2288-14-3

71. Lee YH, Bang H, Kim DJ. How to establish clinical prediction models. *Endocrinol Metab*. Published online 2016. doi:10.3803/EnM.2016.31.1.38

72. Carson AP, Lewis CE, Jacobs DR, et al. Evaluating the Framingham hypertension risk prediction model in young adults: The Coronary Artery risk Development in Young Adults (CARDIA) study. *Hypertension*. Published online 2013. doi:10.1161/HYPERTENSIONAHA.113.01539

73. Wilson PWF, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB.

Prediction of coronary heart disease using risk factor categories. *Circulation*. Published online 1998. doi:10.1161/01.CIR.97.18.1837

74. D'Agostino RB, Vasan RS, Pencina MJ, et al. General cardiovascular risk profile for use in primary care: The Framingham heart study. *Circulation*. Published online 2008. doi:10.1161/CIRCULATIONAHA.107.699579

75. Stevens RJ, Kothari V, Adler AI, Stratton IM, Holman RR. The UKPDS risk engine: A model for the risk of coronary heart disease in type II diabetes (UKPDS 56). *Clin Sci*. Published online 2001. doi:10.1042/CS20000335

76. Kothari V, Stevens RJ, Adler AI, et al. UKPDS 60: Risk of stroke in type 2 diabetes estimated by the U.K. Prospective Diabetes Study risk engine. *Stroke*. Published online 2002. doi:10.1161/01.STR.0000020091.07144.C7

77. Gail MH, Brinton LA, Byar DP, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J Natl Cancer Inst*. Published online 1989. doi:10.1093/jnci/81.24.1879

78. Lindström J, Tuomilehto J. The diabetes risk score: A practical tool to predict type 2 diabetes risk. *Diabetes Care*. Published online 2003. doi:10.2337/diacare.26.3.725

79. Visweswaran S, Angus DC, Hsieh M, Weissfeld L, Yealy D, Cooper GF. Learning patient-specific predictive models from clinical data. *J Biomed Inform*. Published online 2010. doi:10.1016/j.jbi.2010.04.009

80. Wilson PWF, Meigs JB, Sullivan L, Fox CS, Nathan DM, D'Agostino RB. Prediction of incident diabetes mellitus in middle-aged adults: The framingham offspring study. *Arch Intern Med*. Published online 2007. doi:10.1001/archinte.167.10.1068

81. Barlow WE, White E, Ballard-Barbash R, et al. Prospective breast cancer risk prediction

model for women undergoing screening mammography. *J Natl Cancer Inst*. Published online 2006. doi:10.1093/jnci/djj331

82. Davis WA, Knuiman MW, Davis TME. An Australian cardiovascular risk equation for type 2 diabetes: The Fremantle diabetes study. *Intern Med J*. Published online 2010. doi:10.1111/j.1445-5994.2009.01958.x

83. McGorrian C, Yusuf S, Islam S, et al. Estimating modifiable coronary heart disease risk in multiple regions of the world: The INTERHEART Modifiable Risk Score. *Eur Heart J*. Published online 2011. doi:10.1093/eurheartj/ehq448

84. Kshirsagar A V., Bang H, Bomback AS, et al. A simple algorithm to predict incident kidney disease. *Arch Intern Med*. Published online 2008. doi:10.1001/archinte.168.22.2466

85. Bang H, Vupputuri S, Shoham DA, et al. SCreening for Occult REnal Disease (SCORED): A simple prediction model for chronic kidney disease. *Arch Intern Med*. Published online 2007. doi:10.1001/archinte.167.4.374

86. Wang P, Li Y, Reddy CK. Machine Learning for Survival Analysis: A Survey. *arXiv*. Published online 2017.

87. Hastie T, Tibshirani R, Friedman J. *Elements of Statistical Learning 2nd Ed.*; 2009.

88. Van Hoorde K, Van Huffel S, Timmerman D, Bourne T, Van Calster B. A spline-based tool to assess and visualize the calibration of multiclass risk predictions. *J Biomed Inform*. Published online 2015. doi:10.1016/j.jbi.2014.12.016

89. Kruppa J, Liu Y, Biau G, et al. Probability estimation with machine learning methods for dichotomous and multicategory outcome: Theory. *Biometrical J*. Published online 2014. doi:10.1002/bimj.201300068

90. Desai RJ, Wang S V., Vaduganathan M, Evers T, Schneeweiss S. Comparison of Machine Learning Methods With Traditional Models for Use of Administrative Claims With Electronic Medical Records to Predict Heart Failure Outcomes. *JAMA Netw open*. 2020;3(1):e1918962. doi:10.1001/jamanetworkopen.2019.18962

91. Austin PC, Tu J V., Ho JE, Levy D, Lee DS. Using methods from the data-mining and machine-learning literature for disease classification and prediction: A case study examining classification of heart failure subtypes. *J Clin Epidemiol*. Published online 2013. doi:10.1016/j.jclinepi.2012.11.008

92. Tollenaar N, van der Heijden PGM. Which method predicts recidivism best?: A comparison of statistical, machine learning and data mining predictive models. *J R Stat Soc Ser A Stat Soc*. Published online 2013. doi:10.1111/j.1467-985X.2012.01056.x

93. Song X, Mitnitski A, Cox J, Rockwood K. Comparison of machine learning techniques with classical statistical models in predicting health outcomes. *Stud Health Technol Inform*. Published online 2004. doi:10.3233/978-1-60750-949-3-736

94. Frizzell JD, Liang L, Schulte PJ, et al. Prediction of 30-day all-cause readmissions in patients hospitalized for heart failure: Comparison of machine learning and other statistical approaches. *JAMA Cardiol*. Published online 2017. doi:10.1001/jamacardio.2016.3956

95. Survey Questions Asked - Alberta's Tomorrow Project. Accessed January 4, 2021. https://myatpresearch.ca/survey-questions/

96. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*. Published online 1996. doi:10.1002/(SICI)1097-0258(19960229)15:4<361::AID-

SIM168>3.0.CO;2-4

97. Demidenko E. Sample size determination for logistic regression revisited. *Stat Med*. Published online 2007. doi:10.1002/sim.2771

98. Chowdhury MZI, Sikdar KC, Turin TC. Sample Size Calculation in Clinical Studies: Some Common Scenarios. *Int J Stat Med Res*. Published online 2017. doi:10.6000/1929-6029.2017.06.04.3

99. Ogundimu EO, Altman DG, Collins GS. Adequate sample size for developing prediction models is not simply related to events per variable. *J Clin Epidemiol*. Published online 2016. doi:10.1016/j.jclinepi.2016.02.031

100. Pavlou M, Ambler G, Seaman SR, et al. How to develop a more accurate risk prediction model when there are few events. *BMJ*. Published online 2015. doi:10.1136/bmj.h3868

101. Concato J, Peduzzi P, Holford TR, Feinstein AR. Importance of events per independent variable in proportional hazards analysis I. Background, goals, and general strategy. *J Clin Epidemiol*. Published online 1995. doi:10.1016/0895-4356(95)00510-2

102. Harrell FE, Lee KL, Matchar DB, Reichert TA. Regression models for prognostic prediction: Advantages, problems, and suggested solutions. *Cancer Treat Rep*. Published online 1985.

103. Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis II. Accuracy and precision of regression estimates. *J Clin Epidemiol*. Published online 1995. doi:10.1016/0895-4356(95)00048-8

104. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstem AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol*. Published

online 1996. doi:10.1016/S0895-4356(96)00236-3

105. Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. *Am J Epidemiol*. Published online 2007. doi:10.1093/aje/kwk052

106. Gasparini G, Pozza F, Harris AL. Evaluating the potential usefulness of new prognostic and predictive indicators on node-negative breast cancer patients. *J Natl Cancer Inst*. Published online 1993. doi:10.1093/jnci/85.15.1206

107. Simon R, Altman DG. Statistical aspects of prognostic factor studies in oncology. *Br J Cancer*. Published online 1994. doi:10.1038/bjc.1994.192

108. Henderson IC, Patek AJ. The relationship between prognostic and predictive factors in the management of breast cancer. *Breast Cancer Res Treat*. Published online 1998. doi:10.1023/A:1006141703224

109. Graham JW. Missing data analysis: Making it work in the real world. *Annu Rev Psychol*. Published online 2009. doi:10.1146/annurev.psych.58.110405.085530

110. Kang H. The prevention and handling of the missing data. *Korean J Anesthesiol*. Published online 2013. doi:10.4097/kjae.2013.64.5.402

111. Hamer RM, Simpson PM. Last observation carried forward versus mixed models in the analysis of psychiatric clinical trials. *Am J Psychiatry*. Published online 2009. doi:10.1176/appi.ajp.2009.09040458

112. Sinharay S, Stern HS, Russell D. The use of multiple imputation for the analysis of missing data. *Psychol Methods*. Published online 2001. doi:10.1037/1082-989x.6.4.317

113. Midi H, Sarkar SK, Rana S. Collinearity diagnostics of binary logistic regression model. *J Interdiscip Math*. 2010;13(3):253-267. doi:10.1080/09720502.2010.10700699

114. Chowdhury MZI, Turin TC. Variable selection strategies and its importance in clinical

prediction modelling. *Fam Med Community Heal*. Published online 2020. doi:10.1136/fmch-2019-000262

115. Royston P, Moons KGM, Altman DG, Vergouwe Y. Prognosis and prognostic research: Developing a prognostic model. *BMJ*. Published online 2009. doi:10.1136/bmj.b604
116. Snee RD. Validation of Regression Models: Methods and Examples. *Technometrics*. Published online 1977. doi:10.1080/00401706.1977.10489581
117. Arlot S, Celisse A. A survey of cross-validation procedures for model selection. *Stat Surv*. Published online 2010. doi:10.1214/09-SS054
118. Steckler A, McLeroy KR. The importance of external validity. *Am J Public Health*. Published online 2008. doi:10.2105/AJPH.2007.126847
119. Bleeker SE, Moll HA, Steyerberg EW, et al. External validation is necessary in prediction research: A clinical example. *J Clin Epidemiol*. Published online 2003. doi:10.1016/S0895-4356(03)00207-5
120. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology*. Published online 2010. doi:10.1097/EDE.0b013e3181c30fb2
121. Austin PC, Steyerberg EW. Interpreting the concordance statistic of a logistic regression model: Relation to the variance and odds ratio of a continuous explanatory variable. *BMC Med Res Methodol*. Published online 2012. doi:10.1186/1471-2288-12-82
122. Han K, Song K, Choi BW. How to develop, validate, and compare clinical prediction models involving radiological parameters: Study design and statistical methods. *Korean J Radiol*. Published online 2016. doi:10.3348/kjr.2016.17.3.339
123. Sullivan LM, Massaro JM, D'Agostino RB. Presentation of multivariate data for clinical

use: The Framingham Study risk score functions. *Stat Med*. Published online 2004. doi:10.1002/sim.1742

124. Greenhalgh T, Peacock R. Effectiveness and efficiency of search methods in systematic reviews of complex evidence: Audit of primary sources. *Br Med J*. Published online 2005. doi:10.1136/bmj.38636.593461.68

125. Pearson TA, LaCroix AZ, Mead LA, Liang KY. The prediction of midlife coronary heart disease and hypertension in young adults: The Johns Hopkins multiple risk equations. *Am J Prev Med*. Published online 1990. doi:10.1016/s0749-3797(19)30122-9

126. Kivimäki M, Batty GD, Singh-Manoux A, et al. Validating the framingham hypertension risk score: Results from the whitehall II study. *Hypertension*. Published online 2009. doi:10.1161/HYPERTENSIONAHA.109.132373

127. Kivimäki M, Tabak AG, Batty G.D., et al. Incremental predictive value of adding past blood pressure measurements to the framingham hypertension risk equation: The whitehall II study. *Hypertension*. Published online 2010. doi:10.1161/HYPERTENSIONAHA.109.144220

128. Bozorgmanesh M, Hadaegh F, Mehrabi Y, Azizi F. A point-score system superior to blood pressure measures alone for predicting incident hypertension: Tehran Lipid and Glucose Study. *J Hypertens*. Published online 2011. doi:10.1097/HJH.0b013e328348fdb2

129. Sun D, Liu J, Xiao L, et al. Recent development of risk-prediction models for incident hypertension: An updated systematic review. *PLoS One*. Published online 2017. doi:10.1371/journal.pone.0187240

130. Brownrigg JRW, De Lusignan S, McGovern A, et al. Peripheral neuropathy and the risk of cardiovascular events in type 2 diabetes mellitus. *Heart*. Published online 2014.

doi:10.1136/heartjnl-2014-305657

131. Echouffo-Tcheugui JB, Batty GD, Kivimäki M, Kengne AP. Risk Models to Predict Hypertension: A Systematic Review. *PLoS One*. Published online 2013. doi:10.1371/journal.pone.0067370

132. Giampaoli S, Palmieri L, Mattiello A, Panico S. Definition of high risk individuals to optimise strategies for primary prevention of cardiovascular diseases. *Nutr Metab Cardiovasc Dis*. Published online 2005. doi:10.1016/j.numecd.2004.12.001

CHAPTER 2. A SYSTEMATIC REVIEW TO IDENTIFY RISK FACTORS AND PREDICTION MODELS FOR HYPERTENSION INCIDENCE AND A META-ANALYSIS TO EVALUATE MODEL PERFORMANCE

2.1 Abstract

Introduction

Hypertension is a common medical condition and is a significant risk factor for heart attack, stroke, kidney disease, and mortality. Developing a risk prediction model for hypertension incidence incorporating its risk factors can help identify high-risk individuals who should be targeted for healthy behavioral changes or medical treatment to prevent hypertension development. We plan to perform a systematic review and meta-analysis to identify existing hypertension risk prediction models and associated risk factors and evaluate the models' predictive performance.

Methods and Analysis

We systematically searched MEDLINE, EMBASE, Web of Science, Scopus, and the grey literature for studies predicting the risk of hypertension among the general adult population. The search was based on two key concepts: hypertension and risk prediction. Summary statistics from the individual studies were the C-statistic, and a random-effects meta-analysis was used to obtain pooled estimates. The predictive performance of pooled estimates was compared between traditional regression-based models and machine learning-based models. Heterogeneity was assessed using meta-regression, and study quality was assessed using the PROBAST (Prediction model Risk Of Bias ASsessment Tool) checklist.

Results

Of 14,778 articles, 52 articles were finally selected for systematic review, and 32 were selected for meta-analysis. The overall pooled C-statistics was 0.75 [0.73 – 0.77] for the traditional regression-based models and 0.76 [0.72 – 0.79] for the machine learning-based models. High heterogeneity in C-statistic was observed. The age ($p = 0.011$), and sex ($p = 0.044$) of the

participants and the number of risk factors considered in the model ($p = 0.001$) were identified as a source of heterogeneity in traditional regression-based models. Only a few studies were externally validated, and the risk of bias (ROB) and applicability was a concern in many studies.

Conclusion

Many models with acceptable-to-good predictive performance were identified; however, significant differences were not observed in overall predictive performance. More external validation of models and impact studies to implement the hypertension risk prediction model in clinical practice is required.

Key Words: Hypertension, Risk, Prediction Model, Systematic Review, Meta-analysis

2.2 Introduction

Hypertension, or high blood pressure, is a common medical condition affecting about 1 in 4 people[1] and is a significant risk factor for heart attack, stroke, kidney disease, and mortality[2]. Hypertension has been linked to 13% of deaths globally[3] and is a significant health burden that affects all population segments. Considering the high prevalence and global burden, hypertension prevention, and control strategies need to be a top priority. Preventing hypertension creates an opportunity to halt the continuing costly cycle of hypertension management and its associated complications[4]. Hypertension can be prevented by applying strategies that target the general population and individuals and groups at higher risk for hypertension. The need for early identification of at-risk individuals who could benefit from preventive interventions has led to a growing interest in hypertension risk prediction.

Many risk factors such as age, sex, body mass index, waist-hip ratio, blood pressure, smoking, family history, and level of physical inactivity significantly contribute to developing hypertension[5]. Modeling can help identify important risk factors contributing to hypertension and provide reasonable estimates about future hypertension risk[6]. Predicting the risk of developing hypertension through modeling would help identify high-risk individuals who should be targeted for healthy behavioral changes and medical treatment to prevent hypertension[7–9].

Many prediction models have been developed to predict the risk of hypertension in the general population over the years. The predictive ability of these multiple models varies due to their lack of consistency in estimating risk. To evaluate the different models' predictive performance properly, it is recommended that the same data be used[10]. Such evidence, however, is uncommon and therefore not realistic. Instead, through a systematic review and subsequent meta-analysis, a pooled synthesis of performance measures of different models produced in

multiple studies can be compared and measured[11]. This methodology would provide a detailed overview of these models' predictive ability and allow the models' performance measures based on the reported data to be explored quantitatively[11]. With this in mind, we aimed to 1) systematically review the literature to identify hypertension risk prediction models that have been applied to the general adult population and the risk factors that were considered in those models; 2) characterize the study populations in which these models were derived and validated; and 3) assess the predictive performance and quality of these prediction models to better inform the selection of models for clinical implementation.

Two prior studies systematically analyzed hypertension risk prediction models in adults[12,13]. Both studies performed a narrative synthesis of the evidence to summarize the existing knowledge and performance of hypertension prediction models. In addition, a systematic review was also carried out on prediction models to classify children at an elevated risk of developing hypertension[14]. One of the prior studies performed a meta-analysis without assessing heterogeneity[9]. Our review differs from previous studies and contributes to information on the prediction of hypertension risk and the identification of associated risk factors in the following ways: 1) we synthesized performance of the prediction models through meta-analysis and explored potential sources of heterogeneity; 2) we compared the performance of the prediction models developed using traditional statistical regression-based models and more recent machine learning-based models; 3) we provide a thorough evaluation of the quality of the studies among traditionally developed regression-based models; and 4) we describe several additional models that have recently been derived.

2.3 Methods

2.3.1 Data Sources and Searches

We searched MEDLINE, EMBASE, Web of Science, and Scopus (each from inception to March 2020) to identify studies for predicting the risk of incident hypertension in the general adult population. Google Scholar and ProQuest (theses and dissertations) were searched for grey literature. Additionally, we explored the reference lists of all relevant articles. The search strategy focused on two key concepts: hypertension and risk prediction. We used proper free-text words and Medical Subject Headings (MeSH) terms to identify all relevant studies for each key concept. Certain text words were truncated, or wildcards were used when required. The Boolean operators "AND", "OR", and "NOT" were used to combine the words and MeSH terms. A detailed search strategy for MEDLINE is provided in Table 2.1.

2.3.2 Eligibility Criteria

Only original studies were included in this review. This excludes reviews, editorials, commentaries, and letters to the editor. Although risk prediction models are generally developed using a cohort-based study design with follow-up information, we considered all types of study designs, anticipating that machine learning-based models may use other types of study design. Studies written in languages other than English and French were also excluded. The Population, Prognostic Factors (or models of interest), and Outcome (PFO)[15] framework was used to outline eligibility criteria.

Population

The study population consists of people free of hypertension at baseline and those around which hypertension risk prediction models were developed. No restrictions were imposed on the geographic region, time, or gender of the study participants. Nevertheless, models developed only on the adult population were considered, as outcome essential hypertension is expected in adults.

Prognostic Factors (or models of interest)

We only considered studies where risk prediction models for hypertension in the general adult population were developed. Studies that focused solely on the added predictive value of new risk factors to an existing prediction model, studies presenting a prediction model developed in patients with previous hypertension, or studies that derived risk prediction tools other than score-type tools (e.g., risk charts) were not considered. Further, we did not consider studies that only assessed bivariate association between predictors and hypertension incidence. Instead, we focused on those studies where risk prediction models for hypertension were built incorporating risk factors that demonstrated significant prognostic contribution in predicting incident hypertension. When a model was assessed on more than one external population, information from all reported models was considered. However, when the model was presented both in a derivation and validation cohort, only data from the validation cohort were considered for meta-analysis.

Outcome

Our outcome of interest, hypertension, was primarily defined as systolic blood pressure (SBP) ≥ 140 mm Hg, a diastolic blood pressure (DBP) ≥ 90 mm Hg or taking antihypertensive medication. Modifications on the definition of hypertension include the 2017 American College of Cardiology (ACC)/American Heart Association (AHA) Hypertension Guideline's report where SBP ≥130 mm Hg, DBP ≥80 mm Hg, or taking antihypertensive medication was recommended[16]. Nevertheless, we considered all definitions of hypertension to capture the maximum number of studies.

2.3.3 Study Selection

Two reviewers independently identified eligible articles using a two-step process. First, all searched articles were exported to EndNote (Clarivate Analytics) (a software program for managing bibliographies, citations, and references) to remove duplicates. Next, the title and

abstracts of non-duplicated records were screened by two reviewers. Studies retained (based on eligibility criteria) during this stage of screening went to a full-text screening. Full-text articles were further screened for eligibility by the same two reviewers independently. Lastly, articles containing extractable data on hypertension prediction models and hypertension risk factors were selected for data extraction. Inter-rater reliability (Kappa coefficient) was estimated to measure agreement between the independent reviewers. Any disagreement between reviewers was resolved through consensus.

2.3.4 Data Extraction

Two reviewers independently extracted data from each study using standardized forms. We classified the identified models into two categories: models developed using a traditional regression-based approach and models developed using machine learning algorithms. Separate data extraction sheets were used for each model type and included study name, the location where the model was developed/location of data used for the model developed and participants' ethnicity, study design used, sample size, age, and gender of the study participants, risk factors included in the model, number of events and total participants, an outcome considered, the definition used for hypertension, duration of follow-up, modeling method used, measures of discrimination and calibration of the prediction model, and the validation of the prediction model (Table 2.2, Table 2.3). In a separate form (Table 2.4), information about the externally validated hypertension risk prediction models was extracted, including: study name/model validated, the total number of validation studies, location of the validation study, follow-up period, number of events, and total participants, definition of outcome and discrimination and calibration of the model. We also extracted information about risk factors, particularly how many times a specific risk factor was considered in the models (Figure 2.2, Figure 2.3). Each reviewer assessed study quality according

to the Prediction model Risk Of Bias ASsessment Tool (PROBAST) checklist[17,18] (Table 2.5). The PROBAST is designed to assess the risk of bias and concerns regarding diagnostic and prognostic prediction model studies' applicability. The PROBAST contains 20 questions under four domains: participants, predictors, outcome, and analysis, facilitating judgment of risk of bias and applicability. The overall risk of bias of the prediction models was judged as "low", "high", or "unclear" and overall applicability of the prediction models was considered as "low concern", "high concern", and "unclear" according to the PROBAST checklist[17,18].

2.3.5 Data Analysis

We summarized the number of studies identified and those excluded (with the reason for exclusion) and included in the systematic review and subsequent meta-analysis using the PRISMA flow diagram[19] (Figure 2.1). In data synthesis, we performed a meta-analysis both on the traditional regression type's prediction modeling (e.g., logistic regression model and Cox proportional hazard regression model) and a more complicated modeling strategy (e.g., machine learning tools). We synthesized the performance measure of hypertension risk prediction models through meta-analysis. Discrimination (the model's ability to distinguish between patients developing and not developing hypertension) and calibration (the model's accuracy of predicted probabilities of hypertension risk) are the two most common statistical measures of predictive performance. Discrimination is commonly quantified by the concordance (C) statistic, also known as the area under the receiver operating characteristics (ROC) curve. Conversely, calibration is quantified by different measures, and different studies often report different calibration measures. This leads to difficulty in synthesizing calibration measures through meta-analysis. Recent guidelines recommend summarizing the total O (observed)/E (expected) ratio, which provides a rough estimate of overall model calibration[20]. In this review, we performed a meta-analysis on the C-

statistic or AUC (area under the receiver operating characteristic curve) only to evaluate the models' predictive performance and provide a comprehensive summary of the models' predictive ability. We did not undertake a meta-analysis of the total O/E ratio due to the unavailability of relevant data.

Summary statistics, also known as the effect measure, comprised the C-statistic or AUC of the hypertension risk prediction models from the individual studies. To summarize the predictive performance measures (e.g., C-statistic) of a model and determine the existence of unexplained heterogeneity in these measures, random-effects meta-analysis has been recommended[20]. Random-effects meta-analysis assumes that a model's 'true' performance is normally distributed within and across studies[21]; however, the C-statistic distributions, for example, are often skewed across studies in settings with considerable variability in the predictor effects[22]. Normality can be massively improved using the C-statistic logit transformation and is, therefore, more appropriate and recommended when pooling C-statistics[22,23]. Consequently, we logit transformed the C-statistics, performed pooling, and then back-transformed the results to the original scale for interpretation. We used a random-effects meta-analysis with REML estimation and Hartung-Knapp-Sidik-Jonkman (HKSJ) confidence interval (CI) to obtain the pooled weighted average of the logit C-statistic[20]. Forest plots were generated to show the pooled C-statistic together with the 95% CI, 95% approximate prediction interval for the summary C-statistic, the author's name, publication year, and study weights. In studies that only provided a C-statistic but no measure of its variance or confidence intervals, the standard error (SE) and 95% CI of the logit C-statistic (AUC) was calculated using the formula:

$$SE\ (AUC) = \sqrt{\frac{1 + \frac{\left(\frac{N}{2} - 1\right)(1 - c)}{2 - c} + \frac{\left(\frac{N}{2} - 1\right)c}{1 + c}}{c(1 - c)O(N - O)}}$$

where $N =$ the number of patients and $O =$ the total number of observed events (hypertension) and $N - O =$ the total number of non-events[20]. When the confidence intervals of the C-statistics were available, standard errors (SE's) of the logit C-statistics were derived from the CIs as follows: $[(logit(c_{ub}) - logit(c_{lb}))/(2\ X\ 1.96)]^2$, where c_{ub} and c_{lb} are the upper and lower bound of the 95% CI of the C-statistic, respectively[20]. The presence of heterogeneity (mostly due to differences in the study setting, participants, and methodology) was assessed using Cochran's Q statistic and quantified with the I^2 statistic. A p-value of less than 0.05 was considered statistically significant heterogeneity and was categorized as low, moderate, and high when the I^2 values were below 25%, between 25% and 75%, and above 75%, respectively[24]. Sources of heterogeneity were further explored using meta-regression and stratified analyses according to modeling type and study characteristics (sex of the participants, age of the participants, number of risk factors considered in the model, sample size considered in the model, and ethnicity of the study participants). Calculation of 95% approximate prediction intervals to illustrate the extent of between study heterogeneity is also recommended for meta-analysis of performance measures (e.g., C-statistic)[22,23]. We calculated 95% prediction intervals to provide a likely range of performance of a prediction model in a new population and setting. We did not assess publication bias by any statistical tests or funnel plot asymmetry. We used Stata version 16.1 (StataCorp LP, College Station, TX, USA) to perform statistical analysis using the following commands: meta, metan and metareg.

2.4 Results

2.4.1 Study Identification and Selection

We identified 14,730 articles through our electronic database search and an additional 48 articles through our grey literature search. After removing duplicates, 12,268 titles and abstracts were screened for eligibility, and from there, 119 articles were selected for full-text screening. After assessing full-texts, 52 articles were finally selected for the systematic review. Within the chosen final studies, 32 studies provided sufficient information for synthesis through a meta-analysis. The detailed study selection process is summarized in Figure 2.1. Agreement between reviewers on the initial screening and final articles eligible for inclusion in the systematic review was good ($\kappa = 0.81$, and $\kappa = 0.89$, respectively). We classified the identified prediction models into two categories based on the methodology used to develop the model: traditional regression-based models and machine learning-based models. A total of 117 models were identified from the finally selected articles predicting the risk of hypertension in the general adult population, of which 75 were developed using traditional regression-based modeling and 42 using machine learning tools.

2.4.2 Study Characteristics of Traditional Regression-based Models

Study characteristics of traditional regression-based models are presented in Table 2.2. A total of 573,268 participants were used to develop 75 traditional models in 34 studies. Models were mostly developed either in white Caucasian or Asian populations. Two studies considered only male participants, one study considered only female participants, and the remaining studies considered both to develop the models. The number of risk factors considered to create the models ranged from 1 to 19, with a median of 7 risk factors per model. Age was the most common risk factor considered in 61 models, followed by BMI (32 models), DBP (28 models), SBP (27 models), and sex (21 models). The distribution of the conventional risk factors considered in the different models is presented in Figure 2.2. Duration of follow-up time (mean/median/total) considered to develop the models varied between 1.6 years to 30 years. The age of the study participants ranged

from 15 to 90 years. SBP ≥140 mm Hg, DBP ≥90 mm Hg, or use of antihypertensive medication was the standard definition used to define hypertension in almost all the studies, except one study where SBP ≥ 130 mm Hg, DBP≥ 80 mm Hg, or use of any antihypertensive drug was used. Logistic regression was the most used methodology to develop the model (15 studies), followed by Cox proportional-hazards regression (11 studies) and Weibull regression (6 studies). Calibration of the prediction model was reported by 15 studies, mostly using the Hosmer-Lemeshow test. However, the majority of them (19 studies) did not report calibration measures. Discrimination was assessed using the C-statistic (or AUC) and reported by almost all studies with values ranging from 0.57 to 0.97. Only one model was externally validated by the same study when they developed the model.

2.4.3 Meta-analysis of Traditional Regression-based Models

The overall pooled C-statistics of the traditional regression-based models was 0.75 [0.73 – 0.77] (after back transformation to the original scale) with high heterogeneity in the discriminative performance of these models ($I^2 = 99.3$, Cochran Q-statistic $p < 0.001$). Stratified pooled results by modeling type showed pooled C-statistics were 0.73 [0.69 – 0.77], 0.77 [0.74 – 0.81], 0.73 [0.69 – 0.78], and 0.77 [0.75 – 0.79] for Cox, logistic, repeated Poisson, and Weibull respectively (Figure 2.4). The heterogeneity was still observed to be high within the different types of models (Figure 2.4). The 95% approximate prediction interval for the overall C-statistics was from 0.63 to 0.84, which indicates an expected performance range of the considered models in a new population.

To explore possible sources of heterogeneity in the overall pooled C-statistics, we performed a meta-regression. We initially considered the following potential sources of heterogeneity as follows: the definition of hypertension used (the cut-off level used to define hypertension), sex of the participants in included studies (categorized as female-only, male-only,

and both male and female), age of the participants (study participants below average age versus above average age), number of risk factors considered in the model (below median versus above median), sample size considered in the model (below median versus above median), and ethnicity of the study participants (Whites versus Asians). However, we excluded the definition of hypertension as a heterogeneity source, as almost all studies had the same definition of hypertension. Meta-regression identified the participants' sex, that is being male compared to female ($p = 0.044$), participants' age ($p = 0.011$), and the number of risk factors considered in the model ($p = 0.001$) as potential sources of high heterogeneity in the C-statistic. Sex of the participants' when both male and female compared to female-only ($p = 0.351$), sample size considered in the model ($p = 0.395$), and ethnicity of the study participants ($p = 0.899$) did not explain the observed heterogeneity in the C-statistic of these models (Figure S2.1 - S2.4).

2.4.4 Critical Appraisal of Traditional Regression-based Models

We assessed study quality using the PROBAST checklist. A detailed assessment of the risk of bias (ROB) and applicability is presented in Table 2.5 and Figure 2.5. Overall, ROB was "low" in 19 studies, "high" in 5 studies, and "unclear" in 10 studies. Overall applicability was "low concern" in 12 studies, "high concern" in 21 studies, and "unclear concern" in 1 study. Within the ROB domains, the "low" risk of bias was observed in most of the domains except the "analysis" domain, where a large portion of studies (more than 30%) was "unclear" (Figure 2.5). Similarly, within the applicability domains, the "participants" domain seems to be a concern, as a large portion of studies (more than 30%) were at "high concern" or "unclear concern" (Figure 2.5). We also presented the different PROBAST signaling questions' distribution of responses by the various studies in Supplementary Figures S2.5 and S2.6.

2.4.5 Study Characteristics of Machine Learning-based Models

Study characteristics of machine learning-based models are presented in Table 2.3. A total of 1,211,093 participants were used to develop 42 machine learning-based models in 20 studies. Models were basically developed either in white Caucasian or Asian populations. The number of risk factors/features considered to create the model ranged from 2 to 169, with a median of 7 risk factors per model. Age was the most common risk factor considered in 25 models, followed by sex/gender (8 models), BMI (7 models), DBP (6 models), smoking (6 models), and parental history of hypertension (6 models). The distribution of the conventional risk factors considered in machine learning models is presented in Figure 2.3. Hypertension was predominantly defined using SBP ≥ 140 mm Hg, DBP ≥ 90 mm Hg, or antihypertensive medication. Artificial neural network (ANN) was the most common method used to develop the models. Different studies reported different performance measures, and accuracy and AUC/C-statistic were the two most commonly reported measures. Most of the studies did not report calibration measures. In studies that reported discrimination, the AUC (or C-statistic) values range from 0.64 to 0.93.

2.4.6 Meta-analysis of Machine Learning-based Models

The overall pooled C-statistics of the machine learning-based models was 0.76 [0.72 – 0.79] (after back transformation to the original scale) with high heterogeneity in the discriminative performance of these models ($I^2 = 99.9$, Cochran Q-statistic $p < 0.001$). Similar to traditional regression-based models, we did not perform stratified pooled results by modeling type due to diversity in the modeling method. The 95% approximate prediction interval for the overall C-statistics was from 0.63 to 0.84, which indicates an expected performance range of the considered models in a new population, as well as large variability of the models' performance across studies.

We explored possible sources of heterogeneity in the overall pooled C-statistics through meta-regression. As before, we considered sex of the participants (categorized as female-only,

male-only, and both male and female), age of the participants (study participants below average age versus above average age), number of risk factors considered in the model (below median versus above median), sample size considered in the model (below median versus above median), and ethnicity of the study participants (Whites versus Asians) as potential sources of heterogeneity. However, meta-regression did not identify any of age of the participants (p = 0.358), the number of risk factors considered in the model (p = 0.812), sex of the participants, that is being male compared to female (p = 0.886) and both male and female compared to female-only (p = 0.787), sample size considered in the model (p = 0.577), or ethnicity of the study participants (p = 0.326) as the potential source of high heterogeneity in the C-statistic (Figure S2.78 - S2.102).

2.4.7 Study Characteristics of Externally Validated Models

Only four models were externally validated in a different population. Detailed characteristics of the studies that validated these four models are presented in Table 2.4. The Framingham hypertension risk model (FHRS) is the only validated model in more than one external population. The FHRS[25] model was validated by eight different studies in diverse populations. A total of 122,348 participants from 8 studies was used to validate the FHRS model. Study participants had an age range of 18 to 84 years with follow-up time (mean/median/total) from 1.6 years to 25 years. Almost all studies reported performance measures of the FHRS. The Hosmer-Lemeshow test was used to report calibration, while the C-statistic (or AUC) was used to report discrimination. The values of the reported C-statistic ranged from 0.54 to 0.84. Models by Lim et al.[26], Völzke et al.[27], and Kanegae et al.[28] were validated only once in an external population by the same authors. Within these three models, performances were best for the model by Kanegae et al.[28], with a C-statistic of 0.85 [0.76 – 0.91].

2.4.8 Meta-analysis of Externally Validated Models

The pooled C-statistic of the FHRS[25] model was 0.75 [0.68 – 0.80] (after back transformation to the original scale) with high heterogeneity in the discriminative performance of this model ($I^2 = 99.6$, Cochran Q-statistic p < 0.001) . The 95% approximate prediction interval for the C-statistic in the FHRS[25] was from 0.47 to 0.91, which indicates an expected performance range of the FHRS model in a new population, as well as large variability of the model's performance across studies. As the other three models were externally validated only once, pooling their performance measure was irrelevant.

We explored possible sources of heterogeneity in the pooled C-statistics through meta-regression considering the age of the participants (study participants below average age versus above average age), sample size considered in the model (below median versus above median), and ethnicity of the study participants (Whites versus Asians). Only ethnicity of the study participants (p = 0.044) was identified as a source of high heterogeneity in the C-statistic of the FHRS model[25] (Figure S2.11).

2.5 Discussion

This review systematically identified the models used to predict the risk of developing incident hypertension, the risk factors considered to develop the models, synthesized, and compared the predictive performance, and evaluated the included studies' quality. We classified identified models into two categories--traditional regression-based models and machine learning-based models--and assessed each category separately. This categorization assumed that there are inherent differences in these two types of models' developmental methods in computation, complexity, interpretability, and accuracy.

The models we identified mainly were comprised of Caucasian (American/European) or Asian populations. There was no model derived from African populations and only one[29] from

Latin American populations. Considering racial/ethnic groups are particularly susceptible to hypertension (e.g., people of African descent[30]), studies should incorporate subjects from different ethnic backgrounds to build hypertension risk prediction models.

The majority of the models developed considered conventional risk factors for hypertension, although there were considerable variations in the number of risk factors considered by the different models. The most frequently used risk factors included age, BMI, SBP, DBP, sex/gender, etc., which are readily available in clinical practice. Genetic risk factors/biomarkers often contribute significantly to developing hypertension, and models were developed to consider both conventional risk factors and biomarkers. In addition, there were models where biomarkers were used primarily in model building. Information about models developed using biomarkers (e.g., genetic risk scores) is presented in Table S2.1. Biomarkers are often considered very important for increasing the predictive performance of models. However, the pooled predictive performance (C-statistic) of the models that considered biomarkers primarily was 0.76 [0.71 – 0.80] (after back transformation to the original scale) (Figure S2.12) and did not show an overall improvement in the models' predictive performance. Adding genetic factors/biomarkers in the model has disadvantages. The models become less suitable for daily clinical practice, as information on those biomarkers often is not readily available and interpreting the models becomes difficult. Patients also could not use the model for the self-assessment of their risk due to a lack of instant information on biomarkers.

The pooled analysis identified the overall predictive performance of the traditional regression-based models was good (C-statistic 0.75) but with high heterogeneity. The participants' age, sex and the number of risk factors considered in the model were detected as possible sources of heterogeneity. Stratified analysis by modeling methodology (e.g., logistic, Cox) within

traditional regression-based models did not show much difference in predictive performance (C-statistic was from 0.73 to 0.77), and heterogeneity was still observed within the modeling methodology.

The reliability and acceptability of a prediction model largely depend on how well it performs in a validation cohort outside of the derivation cohort where the model was developed. Internal validation of prediction models often is not enough for generalizability, and external validation is necessary before implementing prediction models in clinical practice. The models we identified in our search were mostly internally validated. Only four models[25–28] were found to be externally validated, and only one had multiple validations. The FHRS[25] was the only model validated in eight different populations and had good/accepted pooled predictive performance. This model has potential applicability in a new population, as the model was validated in a diverse population; thus, its performance can be trusted. However, since the FHRS[25] showed high heterogeneity in its predictive performance, and ethnicity served as a source of heterogeneity, and the model was built predominantly in a White population, we need to be cautious in its application in an entirely different population. Models that have only single or no validation need external validation, preferably by a different group of investigators, to guarantee the model's generalizability to a different population.

Only eight models[25,31–37] were converted into a risk score after model development. For a prediction model to be useful in clinical practice, it is crucial that its end-users (clinicians and patients) easily comprehend how the model works and can adequately communicate its results with each other. Presenting the risk derived from the model through scoring instead of a complex mathematical formula may facilitate the use of prediction models and subsequently improve the

uptake of prediction models in clinical practice. We recommend incorporating risk scoring in hypertension risk prediction modeling.

Recently, increased emphasis has been put on using machine learning tools in clinical research, particularly precision medicine. Since machine learning tools are more recent, advanced, and have the reputation of producing more accurate predictive performance, our assumption was models developed using these tools might show better predictive performance than the traditional regression-based models. However, we did not notice much difference in predictive performance between these two categories of models (C-statistic 0.76 versus 0.75). A few machine learning-based models (e.g., models by Huang et al.[38], Sakr et al.[39], and Ye et al.[40]) showed excellent discriminative performance; however, none of these models has ever been externally validated in a different population. In fact, none of the machine learning-based models have been externally validated, an imperative criterion for the generalizability of any prediction model. Consequently, the performance of those models in a new setting/population is quite uncertain. We also noticed high heterogeneity in the predictive performance (C-statistic) of these models. Meta-regression using potential sources of heterogeneity failed to identify the real source of heterogeneity. One possible source could be the difference in methodology used to develop the machine learning-based models. We could not explore this potential source due to the diverse methods considered in different models. We did not notice higher expected variability in machine learning-based models' future predictive performance compared to traditional regression-based models, as the 95% prediction interval for machine learning-based models was similar to traditional regression-based models (0.63 to 0.84).

We also did not find any studies that assessed the impact of adopting hypertension risk prediction models in clinical settings. Ideally, a prediction model should have an impact study to

evaluate whether the model improves clinical decision-making and patient health outcomes[6,41]. Impact studies also help identify factors (ease of use, acceptability) that can affect implementation in routine care[6].

The risk of bias (ROB) was "high", or 'unclear" in a large portion of studies. This is mostly due to the "analysis" domain of ROB, where many studies failed to meet the criteria. Overall, the applicability of the models was "high concern" or "unclear concern" in many studies, and this is mostly due to the "participants" aspect. Several models were developed in a specific population, making the models less applicable to the general adult population.

One of our study's strengths is the extent of the systematic search, which includes four different databases, grey literature, and extensive use of the reference lists of the identified studies. To the best of our knowledge, this is the first study where a meta-analysis of predictive performance, together with assessment of heterogeneity, comparison of the predictive performance of traditional regression based-models and machine learning-based models, and a detailed critical appraisal of studies in hypertension risk prediction models has been performed. Nevertheless, our study also has limitations. We excluded non-English and non-French publications. While it is widely perceived that the English language is the primary language of science, the choice of scientific results in a particular language can incorporate language bias and may lead to incorrect conclusions[42]. We were only able to use C-statistics to compare the model performance, which could be insensitive to distinguish a model's ability to correctly stratify patients into clinically relevant risk groups[42,43]. A meta-analysis of calibration measures (e.g., O/E ratio) along with C-statistics could provide a comprehensive summary of the performance of these models[20]. Failing to assess publication bias amongst the studies is another potential limitation of this study. Recent guidelines[20] did not emphasize the need to assess publication bias for prediction model

performance, which encouraged us not to do so. Although studies have considered publication bias in a similar scenario before, we believe existing traditional publication bias assessment tools (e.g., funnel plot, Egger's test, Begg's test) are more appropriate for studies assessing statistically significant results (e.g., RCT) than studies assessing predictive performance (e.g., C-statistic) of the prognostic models. Instead, we assessed ROB using the PROBAST checklist. We also could not appraise studies that use machine learning algorithms to predict hypertension. Although most of the PROBAST signaling questions also apply to appraise machine learning algorithms, additional signaling questions are recommended to add due to differences in data analysis methods for machine learning algorithms and regression-based models[17,18]. Machine learning algorithms use different variable selection strategies, different estimation techniques for variable– outcome estimations, and different ways to adjust for overfitting[17,18]. When additional questions are added to the PROBAST, these questions need to be appropriately phrased, and specific guidance on assessing these signaling questions also needs to be provided[17,18]. Considering these additional works, we refrain from appraising studies considered machine learning algorithms.

2.6 Conclusion

In this review, we attempted to provide a comprehensive evaluation of hypertension risk prediction models. We identified many models with acceptable-to-good predictive performance. We did not notice significant differences in the predictive performance of traditional regression-based models and machine learning-based models. Including genetic risk factors/biomarkers also did not show much improvement in the models' predictive performance. The quality of the studies was reasonable, with areas where further improvement is needed. Only a few of the multiple models developed had been externally validated, which is a concern. Also, there is a lack of impact studies. Models with external validation and impact studies are required to implement a prediction

model in a clinical practice guideline. A model with accurate prediction is not beneficial if it is not generalizable to a different population or does not improve clinical decision-making and patient health outcomes.

2.7 References

1. Mills KT, Bundy JD, Kelly TN, et al. Global Disparities of Hypertension Prevalence and Control: A Systematic Analysis of Population-Based Studies From 90 Countries. *Circulation*. Published online 2016. doi:10.1161/CIRCULATIONAHA.115.018912
2. CDC. High Blood Pressure Fact Sheet. *Div Hear Dis Stroke Prev*. Published online 2016.
3. Mendis S, Puska P, Norrving B. Global atlas on cardiovascular disease prevention and control. *World Heal Organ*. Published online 2011.
4. Whelton PK, He J, Appel LJ, et al. Primary prevention of hypertension: Clinical and public health advisory from the National High Blood Pressure Education Program. *J Am Med Assoc*. Published online 2002. doi:10.1001/jama.288.15.1882
5. Vasan RS, Larson MG, Leip EP, Kannel WB, Levy D. Assessment of frequency of progression to hypertension in non-hypertensive participants in the Framingham Heart Study: A cohort study. *Lancet*. Published online 2001. doi:10.1016/S0140-6736(01)06710-1
6. Chowdhury MZI, Turin TC. Precision health through prediction modelling: Factors to consider before implementing a prediction model in clinical practice. *J Prim Health Care*. 2020;12(1):3-9. doi:10.1071/HC19087
7. Usher-Smith JA, Silarova B, Schuit E, Moons KGM, Griffin SJ. Impact of provision of cardiovascular disease risk estimates to healthcare professionals and patients: a systematic review. *BMJ Open*. Published online 2015. doi:10.1136/bmjopen-2015-008717
8. Lopez-Gonzalez AA, Aguilo A, Frontera M, et al. Effectiveness of the Heart Age tool for improving modifiable cardiovascular risk factors in a Southern European population: A randomized trial. *Eur J Prev Cardiol*. Published online 2015.

doi:10.1177/2047487313518479

9. Chowdhury MZI, Naeem I, Quan H, et al. Summarising and synthesising regression coefficients through systematic review and meta-analysis for improving hypertension prediction using metamodelling: Protocol. *BMJ Open*. 2020;10(4). doi:10.1136/bmjopen-2019-036388

10. Meads C, Ahmed I, Riley RD. A systematic review of breast cancer incidence risk prediction models with meta-analysis of their performance. *Breast Cancer Res Treat*. 2012;132(2):365-377. doi:10.1007/s10549-011-1818-2

11. Chowdhury MZI, Yeasmin F, Rabi DM, Ronksley PE, Turin TC. Prognostic tools for cardiovascular disease in patients with type 2 diabetes: A systematic review and meta-analysis of C-statistics. *J Diabetes Complications*. 2019;33(1):98-111. doi:10.1016/j.jdiacomp.2018.10.010

12. Echouffo-Tcheugui JB, Batty GD, Kivimäki M, Kengne AP. Risk Models to Predict Hypertension: A Systematic Review. *PLoS One*. 2013;8(7). doi:10.1371/journal.pone.0067370

13. Sun D, Liu J, Xiao L, et al. Recent development of risk-prediction models for incident hypertension: An updated systematic review. *PLoS One*. 2017;12(10):1-19. doi:10.1371/journal.pone.0187240

14. Hamoen M, De Kroon MLA, Welten M, et al. Childhood predictionmodels for hypertension later in life: A systematic review. *J Hypertens*. 2019;37(5):865-877. doi:10.1097/HJH.0000000000001970

15. Munn Z, Stern C, Aromataris E, Lockwood C, Jordan Z. What kind of systematic review should i conduct? A proposed typology and guidance for systematic reviewers in the

medical and health sciences. *BMC Med Res Methodol*. Published online 2018. doi:10.1186/s12874-017-0468-4

16. Whelton PK, Carey RM, Aronow WS, et al. *2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults a Report of the American College of Cardiology/American Heart Association Task Force on Clinical Pr*. Vol 71.; 2018. doi:10.1161/HYP.0000000000000065
17. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: A tool to assess risk of bias and applicability of prediction model studies: Explanation and elaboration. *Ann Intern Med*. Published online 2019. doi:10.7326/M18-1377
18. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. Published online 2019. doi:10.7326/M18-1376
19. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *PLoS Med*. Published online 2009. doi:10.1371/journal.pmed.1000100
20. Debray TPA, Damen JAAG, Snell KIE, et al. A guide to systematic review and meta-analysis of prediction model performance. *BMJ*. Published online 2017. doi:10.1136/bmj.i6460
21. Riley RD, Higgins JPT, Deeks JJ. Interpretation of random effects meta-analyses. *Bmj*. 2011;342(7804):964-967. doi:10.1136/bmj.d549
22. Snell KIE, Ensor J, Debray TPA, Moons KGM, Riley RD. Meta-analysis of prediction model performance across multiple studies: Which scale helps ensure between-study

normality for the C-statistic and calibration measures? *Stat Methods Med Res*. 2018;27(11):3505-3522. doi:10.1177/0962280217705678

23. Debray TPA, Damen JAAG, Riley RD, et al. A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes. *Stat Methods Med Res*. 2019;28(9):2768-2786. doi:10.1177/0962280218785504

24. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *Br Med J*. Published online 2003. doi:10.1136/bmj.327.7414.557

25. Parikh NI, Pencina MJ, Wang TJ, et al. A risk score for predicting near-term incidence of hypertension: The Framingham Heart Study. *Ann Intern Med*. Published online 2008. doi:10.7326/0003-4819-148-2-200801150-00005

26. Lim NK, Son KH, Lee KS, Park HY, Cho MC. Predicting the Risk of Incident Hypertension in a Korean Middle-Aged Population: Korean Genome and Epidemiology Study. *J Clin Hypertens*. 2013;15(5):344-349. doi:10.1111/jch.12080

27. Völzke H, Fung G, Ittermann T, et al. A new, accurate predictive model for incident hypertension. *J Hypertens*. 2013;31(11):2142-2150. doi:10.1097/HJH.0b013e328364a16d

28. Kanegae H, Oikawa T, Suzuki K, Okawara Y, Kario K. Developing and validating a new precise risk-prediction model for new-onset hypertension: The Jichi Genki hypertension prediction model (JG model). *J Clin Hypertens*. 2018;20(5):880-890. doi:10.1111/jch.13270

29. Syllos DH, Calsavara VF, Bensenor IM, Lotufo PA. Validating the Framingham Hypertension Risk Score: A 4-year follow-up from the Brazilian Longitudinal Study of the Adult Health (ELSA-Brasil). *J Clin Hypertens*. 2020;22(5):850-856. doi:10.1111/jch.13855

30. Lackland DT. Racial differences in hypertension: Implications for high blood pressure management. *Am J Med Sci*. 2014;348(2):135-138. doi:10.1097/MAJ.0000000000000308

31. Otsuka T, Kachi Y, Takada H, et al. Development of a risk prediction model for incident hypertension in a working-age Japanese male population. *Hypertens Res*. 2015;38(6):419-425. doi:10.1038/hr.2014.159

32. Chien KL, Hsu HC, Su TC, et al. Prediction models for the risk of new-onset hypertension in ethnic Chinese in Taiwan. *J Hum Hypertens*. 2011;25(5):294-303. doi:10.1038/jhh.2010.63

33. Bozorgmanesh M, Hadaegh F, Mehrabi Y, Azizi F. A point-score system superior to blood pressure measures alone for predicting incident hypertension: Tehran Lipid and Glucose Study. *J Hypertens*. 2011;29(8):1486-1493. doi:10.1097/HJH.0b013e328348fdb2

34. Kadomatsu Y, Tsukamoto M, Sasakabe T, et al. A risk score predicting new incidence of hypertension in Japan. *J Hum Hypertens*. 2019;33(10):748-755. doi:10.1038/s41371-019-0226-7

35. Wang B, Liu Y, Sun X, et al. Prediction model and assessment of probability of incident hypertension: the Rural Chinese Cohort Study. *J Hum Hypertens*. Published online 2020. doi:10.1038/s41371-020-0314-8

36. Díaz-Gutiérrez J, Ruiz-Estigarribia L, Bes-Rastrollo M, Ruiz-Canela M, Martin-Moreno JM, Martínez-González MA. The role of lifestyle behaviour on the risk of hypertension in the SUN cohort: The hypertension preventive score. *Prev Med (Baltim)*. 2019;123(October 2018):171-178. doi:10.1016/j.ypmed.2019.03.026

37. Sathish T, Kannan S, Sarma PS, Razum O, Thrift AG, Thankappan KR. A Risk Score to Predict Hypertension in Primary Care Settings in Rural India. *Asia-Pacific J Public Heal*.

2016;28:26S-31S. doi:10.1177/1010539515604701

38. Huang S, Xu Y, Yue L, et al. Evaluating the risk of hypertension using an artificial neural network method in rural residents over the age of 35 years in a Chinese area. *Hypertens Res*. 2010;33(7):722-726. doi:10.1038/hr.2010.73

39. Sakr S, Elshawi R, Ahmed A, et al. Using machine learning on cardiorespiratory fitness data for predicting hypertension: The Henry Ford exercise testing (FIT) Project. *PLoS One*. 2018;13(4):1-18. doi:10.1371/journal.pone.0195344

40. Ye C, Fu T, Hao S, et al. Prediction of incident hypertension within the next year: Prospective study using statewide electronic health records and machine learning. *J Med Internet Res*. 2018;20(1). doi:10.2196/jmir.9268

41. Kappen TH, van Klei WA, van Wolfswinkel L, Kalkman CJ, Vergouwe Y, Moons KGM. Evaluating the impact of prediction models: lessons learned, challenges, and recommendations. *Diagnostic Progn Res*. Published online 2018. doi:10.1186/s41512-018-0033-6

42. Chowdhury MZI, Yeasmin F, Rabi DM, Ronksley PE, Turin TC. Predicting the risk of stroke among patients with type 2 diabetes: A systematic review and meta-analysis of C-statistics. *BMJ Open*. 2019;9(8):1-22. doi:10.1136/bmjopen-2018-025579

43. Cook NR, Ridker PM. Advances in measuring the effect of individual predictors of cardiovascular risk: The role of reclassification measures. *Ann Intern Med*. Published online 2009. doi:10.7326/0003-4819-150-11-200906020-00007

44. Pearson TA, LaCroix AZ, Mead LA, Liang KY. The prediction of midlife coronary heart disease and hypertension in young adults: The Johns Hopkins multiple risk equations. *Am J Prev Med*. 1990;6(2 SUPPL.):23-28. doi:10.1016/s0749-3797(19)30122-9

45. Paynter NP, Cook NR, Everett BM, Sesso HD, Buring JE, Ridker PM. Prediction of Incident Hypertension Risk in Women with Currently Normal Blood Pressure. *Am J Med*. 2009;122(5):464-471. doi:10.1016/j.amjmed.2008.10.034

46. Kivimäki M, Batty GD, Singh-Manoux A, et al. Validating the framingham hypertension risk score: Results from the whitehall II study. *Hypertension*. 2009;54(3):496-501. doi:10.1161/HYPERTENSIONAHA.109.132373

47. Kivimäki M, Tabak AG, Batty GD, et al. Incremental predictive value of adding past blood pressure measurements to the framingham hypertension risk equation: The whitehall II study. *Hypertension*. 2010;55(4):1058-1062. doi:10.1161/HYPERTENSIONAHA.109.144220

48. Kshirsagar A V., Chiu Y lin, Bomback AS, et al. A hypertension risk score for middle-aged and older adults. *J Clin Hypertens*. 2010;12(10):800-808. doi:10.1111/j.1751-7176.2010.00343.x

49. Fava C, Sjögren M, Montagnana M, et al. Prediction of blood pressure changes over time and incidence of hypertension by a genetic risk score in swedes. *Hypertension*. 2013;61(2):319-326. doi:10.1161/HYPERTENSIONAHA.112.202655

50. Choi YH, Chowdhury R, Swaminathan B. Prediction of hypertension based on the genetic analysis of longitudinal phenotypes: A comparison of different modeling approaches for the binary trait of hypertension. *BMC Proc*. 2014;8(Suppl 1):8-13. doi:10.1186/1753-6561-8-S1-S78

51. Lim NK, Lee JY, Lee JY, Park HY, Cho MC. The role of genetic risk score in predicting the risk of hypertension in the Korean population: Korean genome and epidemiology study. *PLoS One*. 2015;10(6):1-11. doi:10.1371/journal.pone.0131603

52. Asgari S, Khalili D, Mehrabi Y, Kazempour-Ardebili S, Azizi F, Hadaegh F. Incidence and risk factors of isolated systolic and diastolic hypertension: a 10 year follow-up of the Tehran Lipids and Glucose Study. *Blood Press*. 2016;25(3):177-183. doi:10.3109/08037051.2015.1116221

53. Lee JW, Lim NK, Baek TH, Park SH, Park HY. Anthropometric indices as predictors of hypertension among men and women aged 40-69 years in the Korean population: The Korean Genome and Epidemiology Study. *BMC Public Health*. 2015;15(1):1-7. doi:10.1186/s12889-015-1471-5

54. Lee BJ, Kim JY. A comparison of the predictive power of anthropometric indices for hypertension and hypotension risk. *PLoS One*. 2014;9(1). doi:10.1371/journal.pone.0084897

55. Chen Y, Wang C, Liu Y, et al. Incident hypertension and its prediction model in a prospective northern urban Han Chinese cohort study. *J Hum Hypertens*. 2016;30(12):794-800. doi:10.1038/jhh.2016.23

56. Wang Y, Ma Z, Xu C, Wang Z, Yang X. Prediction of transfer among multiple states of blood pressure based on Markov model: An 18-year cohort study. *J Hypertens*. 2018;36(7):1506-1513. doi:10.1097/HJH.0000000000001722

57. Niiranen TJ, Havulinna AS, Langén VL, Salomaa V, Jula AM. Prediction of Blood Pressure and Blood Pressure Change With a Genetic Risk Score. *J Clin Hypertens*. 2016;18(3):181-186. doi:10.1111/jch.12702

58. Yeh CJ, Pan WH, Jong YS, Kuo YY, Lo CH. Incidence and predictors of isolated systolic hypertension and isolated diastolic hypertension in Taiwan. *J Formos Med Assoc*. 2001;100(10):668-675.

59. Xu F, Zhu J, Sun N, et al. Development and validation of prediction models for hypertension risks in rural Chinese populations. *J Glob Health*. 2019;9(2). doi:10.7189/jogh.09.020601

60. Wang A, An N, Chen G, Li L, Alterovitz G. Predicting hypertension without measurement: A non-invasive, questionnaire-based approach. *Expert Syst Appl*. 2015;42(21):7601-7609. doi:10.1016/j.eswa.2015.06.012

61. Muntner P, Woodward M, Mann DM, et al. Comparison of the framingham heart study hypertension model with blood pressure alone in the prediction of risk of hypertension: The multi-ethnic study of atherosclerosis. *Hypertension*. 2010;55(6):1339-1345. doi:10.1161/HYPERTENSIONAHA.109.149609

62. Ture M, Kurt I, Turhan Kurum A, Ozdamar K. Comparing classification techniques for predicting essential hypertension. *Expert Syst Appl*. 2005;29(3):583-588. doi:10.1016/j.eswa.2005.04.014

63. Yamakado M, Nagao K, Imaizumi A, et al. Plasma Free Amino Acid Profiles Predict Four-Year Risk of Developing Diabetes, Metabolic Syndrome, Dyslipidemia, and Hypertension in Japanese Population. *Sci Rep*. 2015;5(November 2014):1-12. doi:10.1038/srep11918

64. Qi Y, Zhao H, Wang Y, et al. Replication of the top 10 most significant polymorphisms from a large blood pressure genome-wide association study of northeastern Han Chinese East Asians. *Hypertens Res*. 2014;37(2):134-138. doi:10.1038/hr.2013.132

65. Lu X, Huang J, Wang L, et al. Genetic predisposition to higher blood pressure increases risk of incident hypertension and cardiovascular diseases in Chinese. *Hypertension*. 2015;66(4):786-792. doi:10.1161/HYPERTENSIONAHA.115.05961

66. Zhang W, Wang L, Chen Y, Tang F, Xue F, Zhang C. Identification of hypertension predictors and application to hypertension prediction in an urban Han Chinese population: A longitudinal study, 2005-2010. *Prev Chronic Dis*. 2015;12(10):1-10. doi:10.5888/pcd12.150192

67. Falk CT. Risk factors for coronary artery disease and the use of neural networks to predict the presence or absence of high blood pressure. *BMC Genet*. 2003;4 Suppl 1:1-6. doi:10.1186/1471-2156-4-s1-s67

68. Farran B, Channanath AM, Behbehani K, Thanaraj TA. Predictive models to assess risk of type 2 diabetes, hypertension and comorbidity: Machine-learning algorithms and validation using national health data from Kuwait-a cohort study. *BMJ Open*. 2013;3(5):1-10. doi:10.1136/bmjopen-2012-002457

69. Kwong EWY, Wu H, Pang GKH. A prediction model of blood pressure for telemedicine. *Health Informatics J*. 2018;24(3):227-244. doi:10.1177/1460458216663025

70. Polak S, Mendyk A. Artificial neural networks based Internet hypertension prediction tool development and validation. *Appl Soft Comput J*. 2008;8(1):734-739. doi:10.1016/j.asoc.2007.06.001

71. Priyadarshini R, Barik RK, Dubey H. DeepFog: Fog computing-based deep neural architecture for prediction of stress types, diabetes and hypertension attacks. *Computation*. 2018;6(4). doi:10.3390/computation6040062

72. Tayefi M, Esmaeili H, Saberi Karimian M, et al. The application of a decision tree to establish the parameters associated with hypertension. *Comput Methods Programs Biomed*. 2017;139:83-91. doi:10.1016/j.cmpb.2016.10.020

73. Wu TH, Pang GKH, Kwong EWY. Predicting systolic blood pressure using machine

learning. *2014 7th Int Conf Inf Autom Sustain "Sharpening Futur with Sustain Technol ICIAfS 2014*. Published online 2014:1-6. doi:10.1109/ICIAFS.2014.7069529

74. Wu TH, Kwong EWY, Pang GKH. Bio-medical application on predicting systolic blood pressure using neural networks. *Proc - 2015 IEEE 1st Int Conf Big Data Comput Serv Appl BigDataService 2015*. Published online 2015:456-461. doi:10.1109/BigDataService.2015.54

75. Zhang B, Wei Z, Ren J, Cheng Y, Zheng Z. An Empirical Study on Predicting Blood Pressure Using Classification and Regression Trees. *IEEE Access*. 2018;6(January):21758-21768. doi:10.1109/ACCESS.2017.2787980

76. Zhao Q, Wang L, Yang W, et al. Interactions among genetic variants from contractile pathway of vascular smooth muscle cell in essential hypertension susceptibility of Chinese Han population. *Pharmacogenet Genomics*. 2008;18(6):459-466. doi:10.1097/FPC.0b013e3282f97fb2

77. Zhao H, Qi Y, Wang Y, et al. Interactive contribution of serine/threonine kinase 39 gene multiple polymorphisms to hypertension among northeastern Han Chinese. *Sci Rep*. 2014;4:1-7. doi:10.1038/srep05116

78. Zheng L, Sun Z, Zhang X, et al. Predictive value for the rural chinese population of the framingham hypertension risk model: Results: from liaoning province. *Am J Hypertens*. 2014;27(3):409-414. doi:10.1093/ajh/hpt229

79. Carson AP, Lewis CE, Jacobs DR, et al. Evaluating the Framingham hypertension risk prediction model in young adults: The Coronary Artery risk Development in Young Adults (CARDIA) study. *Hypertension*. 2013;62(6):1015-1020. doi:10.1161/HYPERTENSIONAHA.113.01539

80. Lim NK, Lee JW, Park HY. Validation of the Korean genome epidemiology study risk score to predict incident hypertension in a large nationwide Korean cohort. *Circ J*. 2016;80(7):1578-1582. doi:10.1253/circj.CJ-15-1334

81. Wang YL, Qi Y, Bai JN, et al. Tag polymorphisms of solute carrier family 12 member 3 gene modify the risk of hypertension in northeastern Han Chinese. *J Hum Hypertens*. 2014;28(8):504-509. doi:10.1038/jhh.2013.134

Figure 2.1

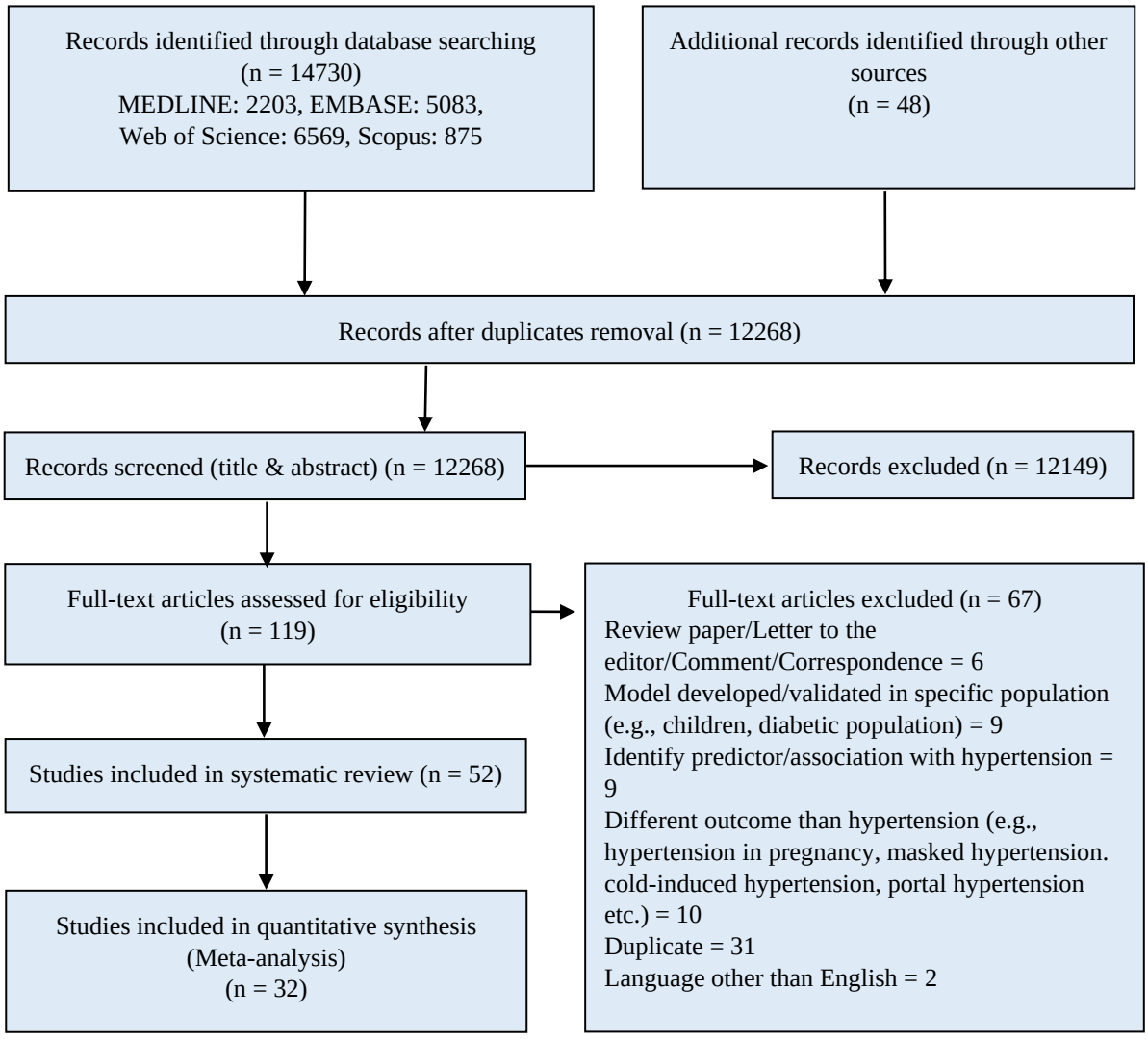

Figure 2.1 PRISMA diagram for the systematic review of studies presenting hypertension prediction models developed in the general population

Figure 2.2

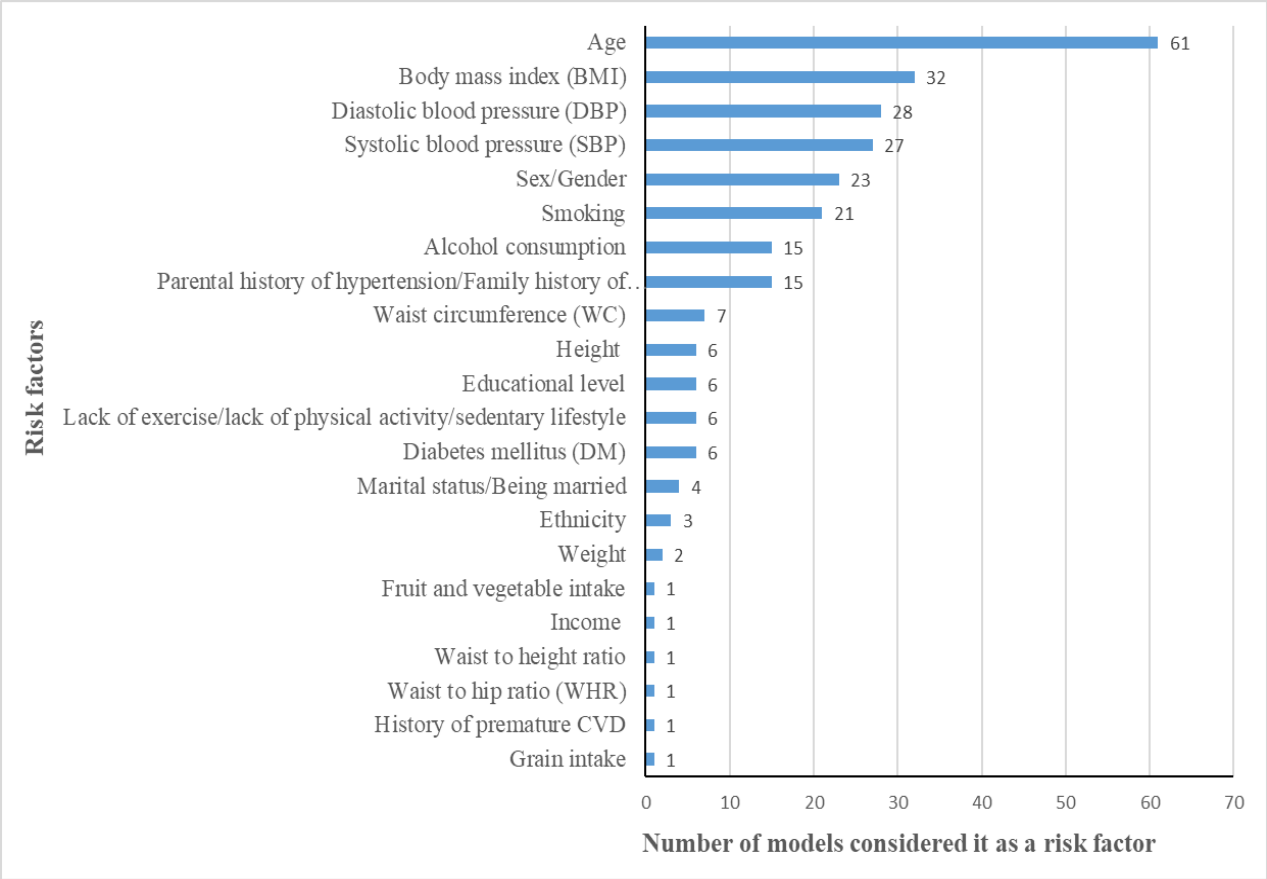

Figure 2.2 Conventional risk factors considered by traditional regression-based models.

Figure 2.3

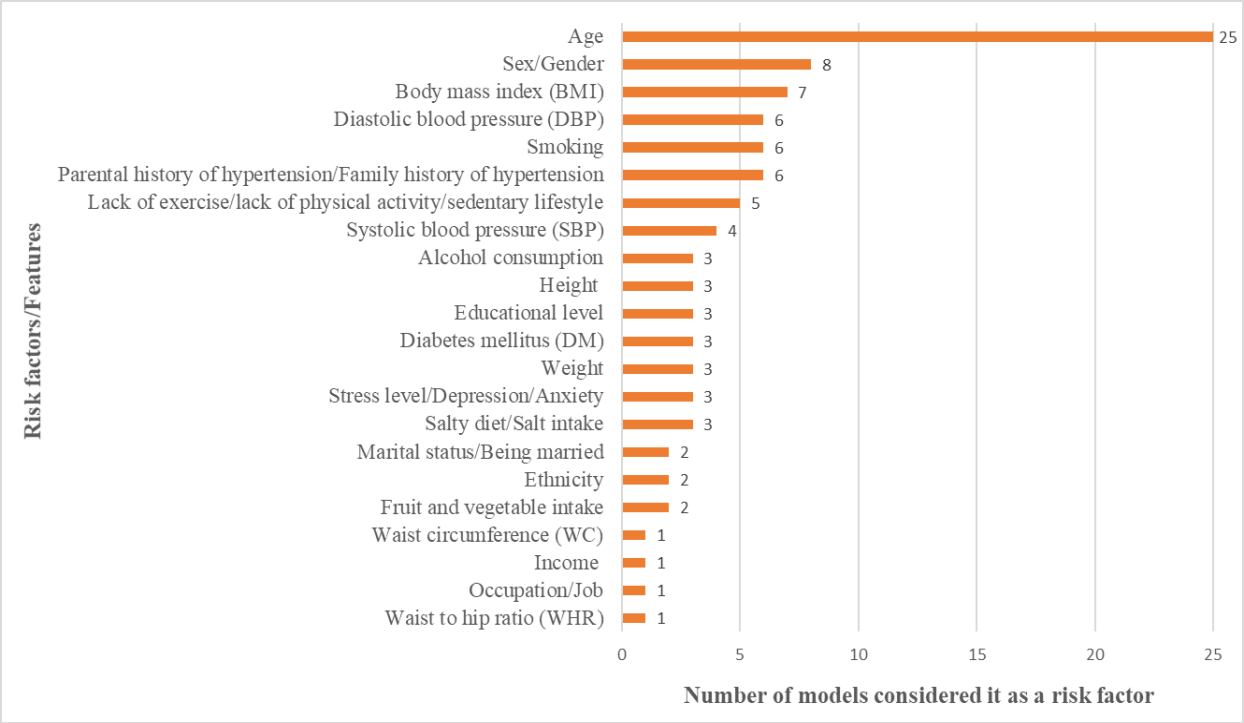

Figure 2.3 Conventional risk factors considered by machine learning-based models.

Figure 2.4

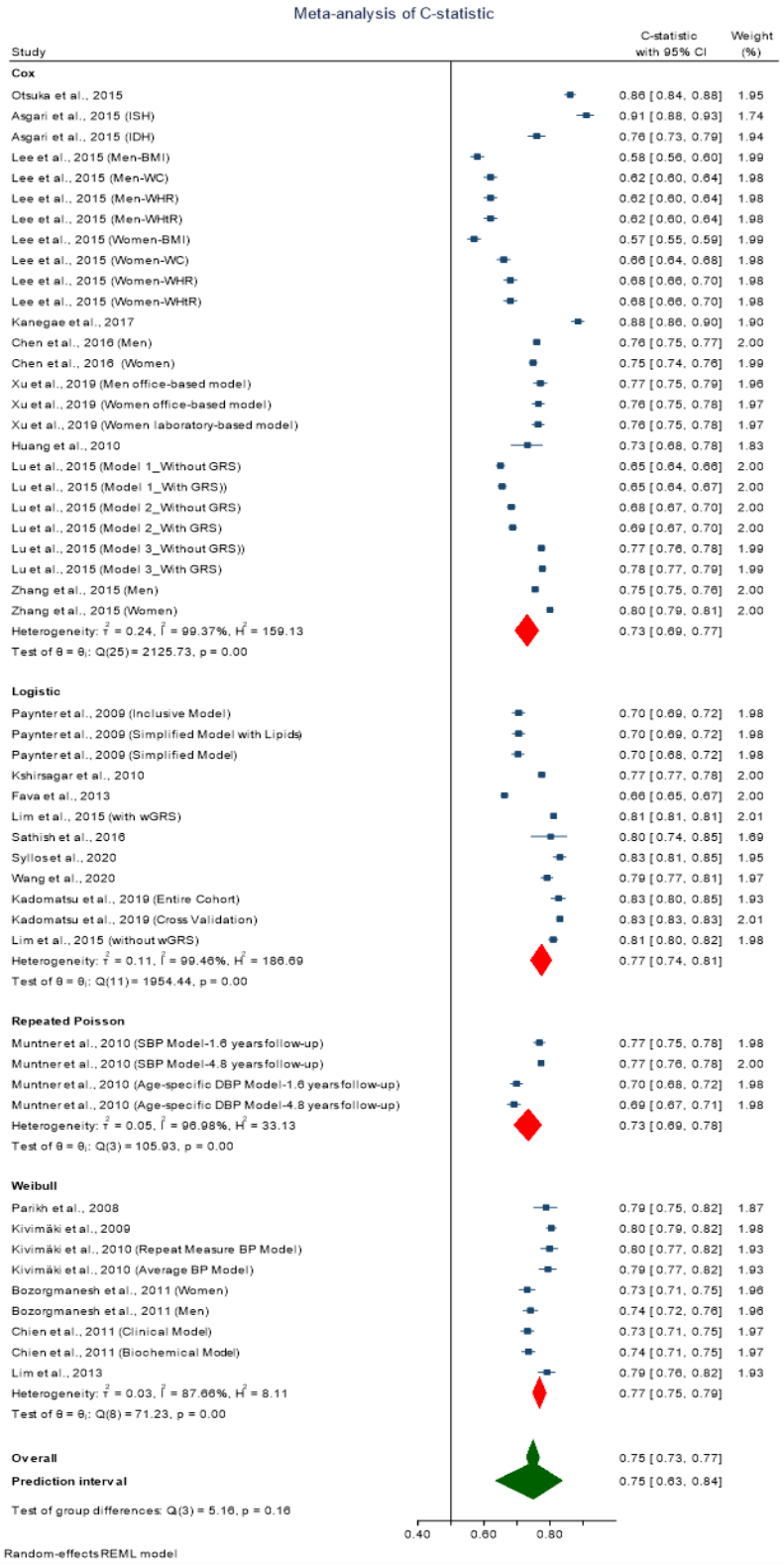

Figure 2.4 Forest plot of traditional regression-based models with 95% prediction interval.

Figure 2.5

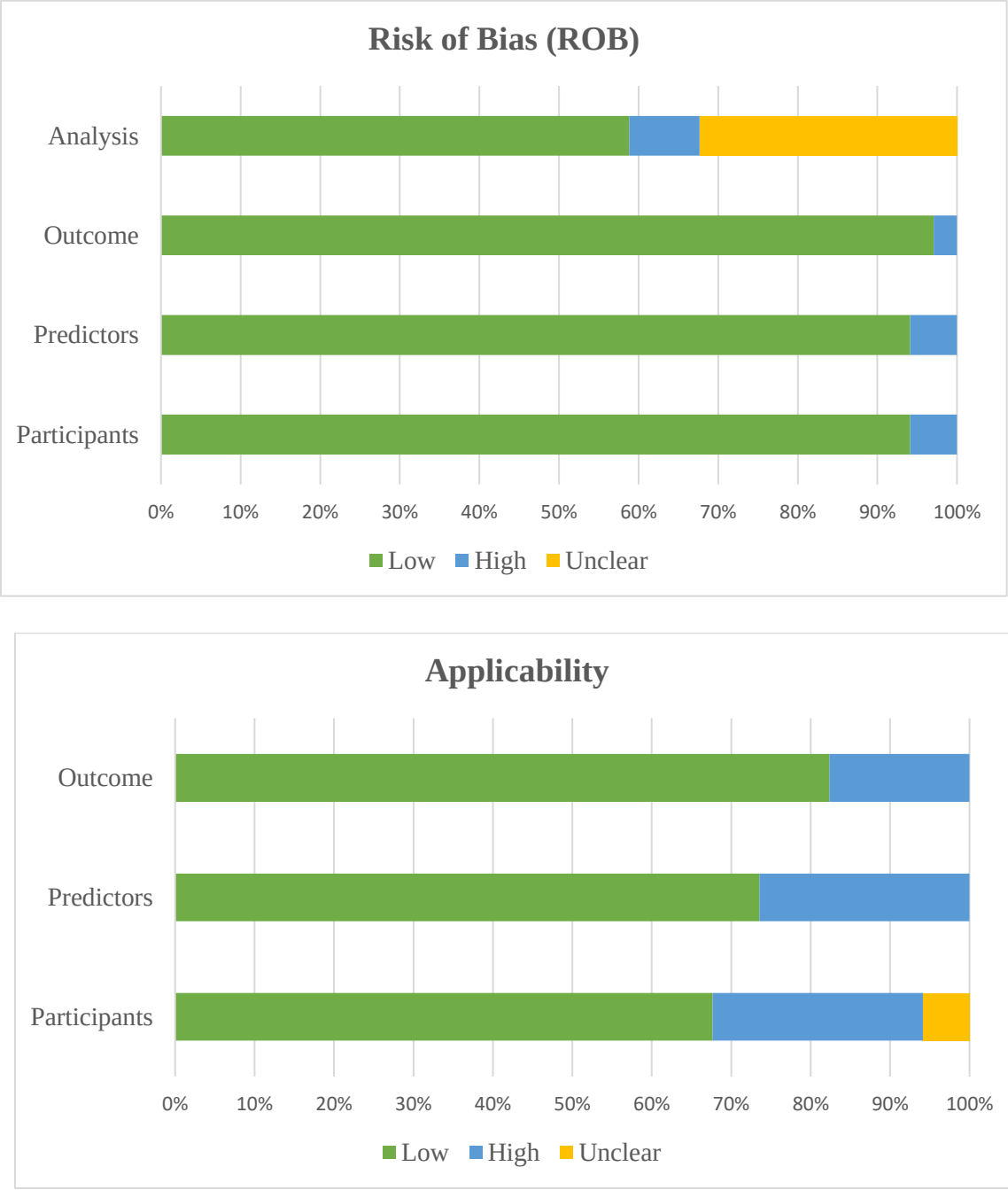

Figure 2.5 Graphical summary presenting the percentage of hypertension risk prediction studies rated by level of concern, risk of bias (ROB), and applicability for each domain.

Figure 2.6

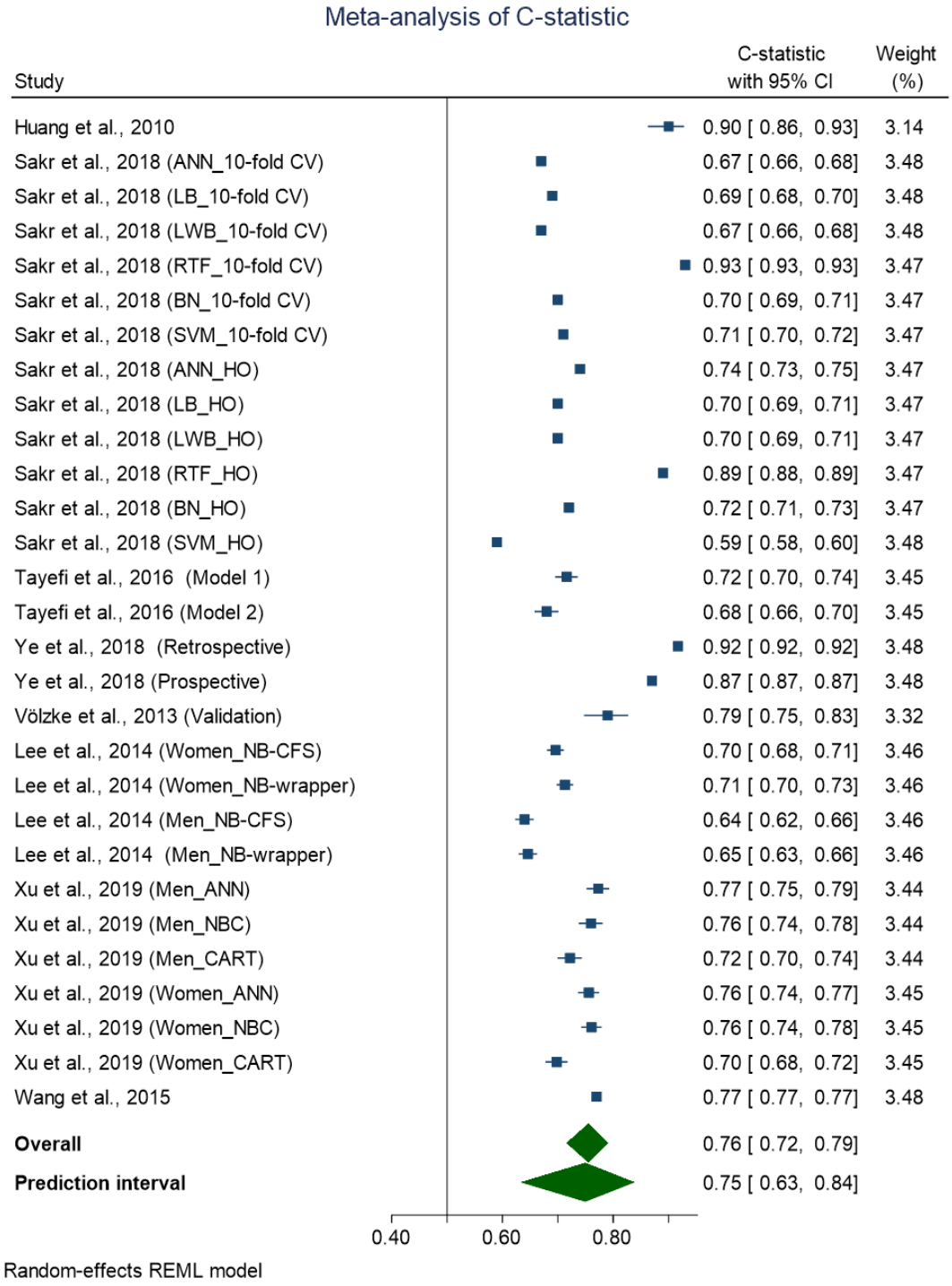

Figure 2.6 Forest plot of machine regression-based models with 95% prediction interval.

Figure 2.7

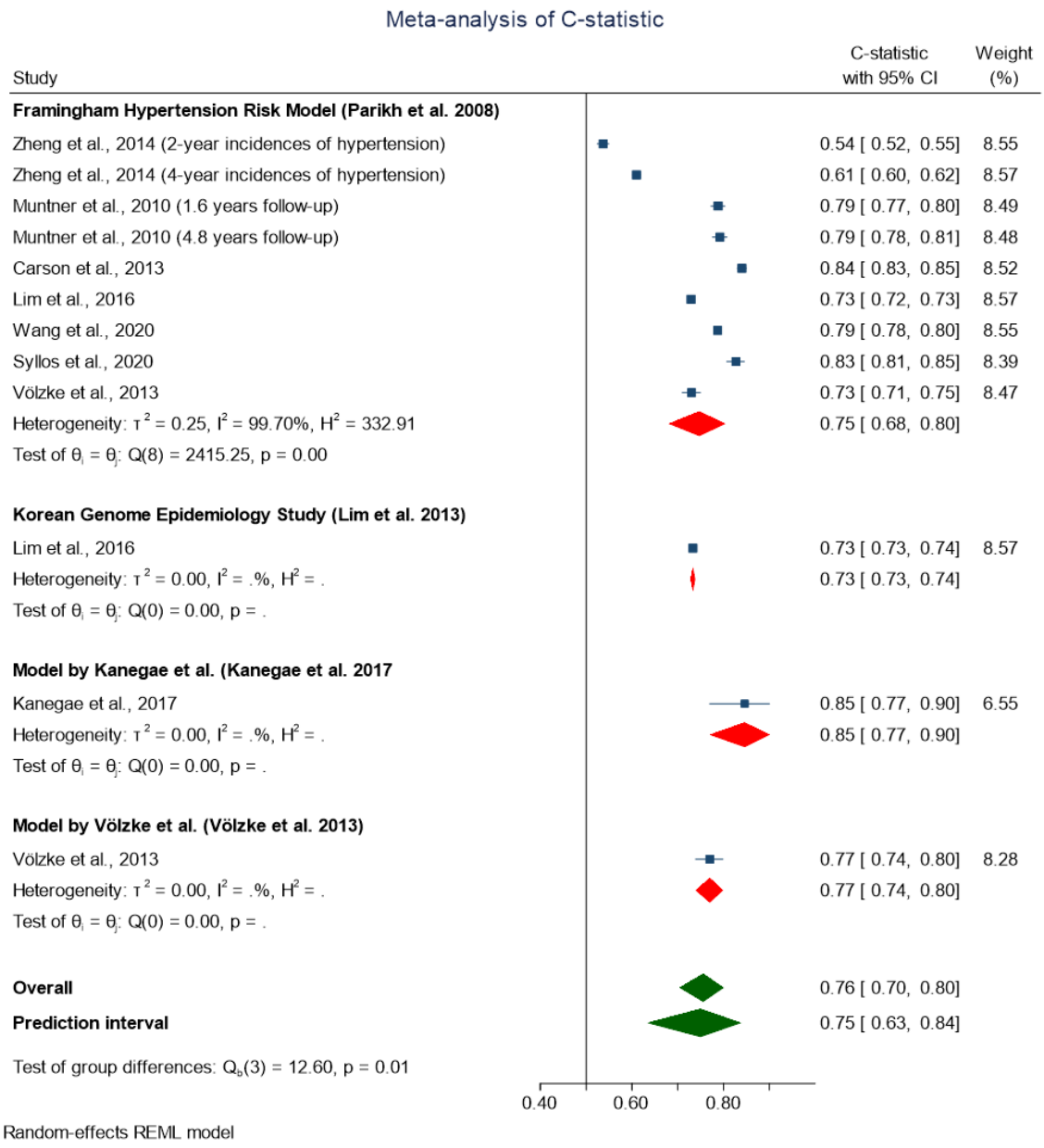

Figure 2.7 Forest plot of externally validated models with 95% prediction interval.

Table 2.1 Keywords Used to Search in MEDLINE

KEYWORDS
1. prediction model*.mp 2. risk function*.mp 3. risk prediction*.mp 4. risk table*.mp 5. predictive model*.mp 6. exp "Predictive Value of Tests"/ 7. risk chart*.mp 8. risk equation*.mp 9. risk engine*.mp 10. risk calculat*.mp 11. risk score*.mp 12. prediction tool*.mp 13. prediction rule*.mp 14. risk model*.mp 15. prognostic tool*.mp 16. prognostic model*.mp 17. exp Risk Assessment/ 18. risk algorithm*.mp 19. risk ind*.mp 20. prediction algorithm*.mp 21. (hypertension adj2 (risk score or risk model or prediction model or risk prediction model or risk assessment)).mp 22. (high blood pressure adj2 (risk score or risk model or prediction model or risk prediction model or risk assessment)).mp 23. OR/1-22 24. validation.mp. 25. exp Validation Studies/ 26. validate*.mp 27. OR/24-26 28. 23 AND 27 29. exp Hypertension/ 30. hypertens*.mp 31. (high adj2 blood pressure).mp 32. high blood pressure.mp 33. elevated blood pressure.mp 34. blood pressure.mp 35. OR/29-34 36. 28 AND 35

Table 2.2 Information about existing traditional regression-based hypertension prediction models from the selected studies

Study	Location Model Developed/ Ethnicity	Study Design	Age	Gender	Risk Factors Included	Events (n)/Total Participants (N)	Definition of Outcome Predicted/ Hypertension	Duration of Follow-up	Modeling Method	Discrimination	Calibration	Model Validation: Internal or External
Pearson et al.[44] 1990	USA/ Mixed, mainly Whites	Prospective cohort	≤ 25 years	Male only	Age, SBP at baseline, paternal history of hypertension, and BMI	114/1130	Self-reported use of blood pressure-lowering medications	30 years	Cox proportional-hazards regression	NR	NR	NR
Parikh et al.[25] 2008	USA/ Mainly Whites	Prospective cohort	20-69 years	Both male and female	Age, sex, SBP, DBP, BMI, parental hypertension, and cigarette smoking	796/1717	SBP ≥ 140 mmHg or DBP ≥ 90 mmHg or use of BP-lowering medications	Median 3.8 years	Weibull regression	C-statistic = 0.788 [0.733–0.803]	HL Chi-square = 4.35 (p = 0.88)	Internal, apparent
Paynter et al.[45] 2009	USA/ Whites and Blacks	Prospective cohort	45-64 years	Female only	Inclusive Model: Age, ethnicity, BMI, total grain intake, SBP, DBP, apolipoprotein B, lipoprotein (a), and C-reactive protein. Simplified Model with Lipids: Age, BMI, SBP, DBP, ethnicity, and total to HDL-cholesterol ratio Simplified Model: Age, BMI, ethnicity, SBP, and DBP	Derivation cohort: 1935/9427 Validation cohort: 1068/5395	Self-report or SBP ≥ 140 mmHg or DBP ≥ 90 mmHg	8 years	Logistic regression	Inclusive Model: C-statistic = 0.705 Simplified Model with Lipids: C-statistic = 0.705 Simplified Model: C-statistic = 0.703	Inclusive Model: HL Chi-square = 24.6 (p = 0.002) Simplified Model with Lipids: HL Chi-square = 20.7 (p = 0.008) Simplified Model: HL Chi-square = 12.3 (p = 0.140)	Internal, split-sample 2:1

Kivimäki et al.[46] 2009	England/ Mainly Whites	Prospective cohort	35-68 years	Both male and female	Age, sex, SBP, DBP, BMI, parental hypertension, and cigarette smoking	1258/8207	SBP ≥ 140 mmHg or DBP ≥ 90 mmHg or use of BP-lowering medications	Median 5.6 years	Weibull regression	C-statistic = 0.804	HL Chi-square = 14.3 (p = 0.88)	Internal, split-sample 6:4
Kivimäki et al.[47] 2010	England/ Mainly Whites	Prospective cohort	36-68 years	Both male and female	Repeat Measure BP Model: Age, sex, BMI, parental hypertension, repeat measures of BP, and cigarette smoking Average BP Model: Age, sex, BMI, parental hypertension, average BP, and cigarette smoking	Derivation cohort: 614/4135 Validation cohort: 438/2785	SBP ≥ 140 mmHg or DBP ≥ 90 mmHg or use of antihypertensive medications	Median 5.8 years	Weibull regression	Repeat Measure BP Model: C-statistic = 0.799 Average BP Model: C-statistic = 0.794	Repeat Measure BP Model: HL Chi-square = 6.5 Average BP Model: NR	Internal, split-sample 6:4
Kshirsagar et al.[48] 2010	USA/ Mixed but mainly Whites	Prospective cohort	45-64 years	Both male and female	Age, level of SBP or DBP, smoking, family history of hypertension, diabetes mellitus, BMI, female sex, and lack of exercise	3795/11,407 (7610 for derivation sample and 3692 for the validation sample)	SBP ≥ 140 mmHg or DBP ≥ 90 mmHg or reported use of BP-lowering medications	Up to 9 years	Logistic regression	AUC = 0.742 (3years), 0.750 (6 years), 0.791 (9 years), and 0.775 (ever)	NR	Internal, split-sample 2:1
Bozorgmanesh et al.,[33] 2011	Iran/ Asians	Prospective cohort	≥ 20 years	Both male and female	For Women: age, waist circumference, DBP, SBP, and family history of premature CVD For Men: age, DBP, SBP, and smoking	805/4656	SBP ≥ 140 mmHg or DBP ≥ 90 mmHg or reported use of BP lowering medications	6 years	Weibull regression	C-statistic = 0.731 [0.706-0.755] for women C-statistic = 0.741 [0.719-0.763] for men	HL Chi-square = 7.8 (p = 0.554) for women HL Chi-square = 8.8 (p = 0.452) for men	NR

Chien et al.[32] 2011	Taiwan/ Chinese	Prospective cohort	≥ 35 years	Both male and female	Clinical Model: Age, gender, BMI, SBP, and DBP Biochemical Model: Age, gender, BMI, SBP, DBP, white blood count, fasting glucose, uric acid	1029/2506	SBP ≥ 140 mmHg or DBP ≥ 90 mmHg or reported use of BP-lowering medications	Median 6.15 years	Weibull regression	Clinical Model: AUC = 0.732 [0.712 - 0.752] (point based, AUC = 0.737 (coefficient based) Biochemical Model: AUC = 0.735 [0.715 - 0.755] (point based), AUC = 0.74 (coefficient based)	Clinical Model: HL Chi-square = 8.3, p = 0.40 (point based), 10.9, p = 0.21 (coefficient based) Biochemical Model: HL Chi-square = 13.2, p = 0.11 (point based), 6.4, p = 0.60 (coefficient based)	Internal, fivefold cross-validation
Fava et al.[49] 2013	Sweden/ Whites	Prospective cohort	Middle -aged	Both male and female	Age, sex, age, sex times age, heart rate, obesity (BMI.30 kg/m2), diabetes, hypertriglyceridemia, prehypertension, family history of hypertension, sedentary in spare time, problematic alcohol behavior, married or living as a couple,	NR/10,781	SBP ≥ 140 mmHg or DBP ≥ 90 mmHg or reported use of BP-lowering medications	Over average 23-years	Logistic regression	AUC = 0.662 [0.651-0.672]	NR	NR

					high-level non-manual work, smoking							
Lim et al.[26] 2013	Korea/ Asians	Prospective cohort	40–69 years	Both male and female	Age, sex, smoking, SBP, DBP, parental hypertension, BMI	819/4747 Derivation cohort: 483/2840 Validation cohort: 336/1907	SBP ≥ 140 mmHg or DBP ≥ 90 mmHg or reported use of BP lowering medications	4 years	Weibull regression	AROC = 0.791 [0.766 - 0.817]	H-L Chi-square = 4.17 (p = 0.8415)	Internal, split-sample 6:4
Choi et al.[50] 2014	USA/ Mexicans	Prospective cohort	NR	Both male and female	Age, gender, smoke, age x gender, Rs10510257 (AA), Rs10510257 (AG), Rs1047115 (GT)	NR/443	SBP >140 mm Hg, DBP >90 mm Hg, or use of antihypertensive medication	NR	Generalized estimating equations for marginal model and logistic random effect model for conditional model	Marginal model: AUC = 0.839 (with SNPs), 0.826 (without SNPs) Conditional model: AUC = 0.973 (with SNPs), 0.973 (without SNPs)	NR	NR
Lim et al.[51] 2015	Korean/ Asians	Prospective cohort	40-69 years	Both male and female	Traditional variables: age, gender, SBP, current smoking status, family history of hypertension, BMI, and one genetic variable (cGRS or wGRS derived from the 4 SNPs): rs995322, rs17249754,	NR/5632	SBP ≥140 mm Hg or DBP ≥90 mm Hg or use of antihypertensive medication	4-year	Logistic regression	Derivation cohort: C-statistic = 0.810 [0.796–0.824] (model without wGRS, C-statistic = 0.811 [0.797–0.825] (model with	HL Chi-square = 6.916 (model without wGRS), HL Chi-square = 5.711 (model with wGRS)	Internal validation, fivefold cross-validation

					rs1378942, rs12945290					wGRS); Validation cohort: Mean C-statistic = 0.811 [0.809-0.816]		
Otsuka et al.[31] 2015	Japan/ Asians	Prospective cohort	19–63 years	Male only	Age, BMI, SBP and DBP, current smoking status, excessive alcohol intake, parental history of hypertension	1633/15,025	SBP ≥140 mm Hg or DBP ≥90 mm Hg or use of antihypertensive medication	Median 4 years	Cox proportional-hazards regression	Validation cohort: C-statistic = 0.861 [0.844-0.877] (model), C-statistic = 0.858 [0.840-0.876] (score)	Validation cohort: HL Chi-square = 15.2 (p = 0.085) (model), HL Chi-square = 9.30 (p = 0.41) (score)	Internal validation, split sample 4:1
Asgari et al.[52] 2015	Iran/ Asians	Prospective cohort	≥ 20 years	Both male and female	ISH: Age, SBP, BMI, 2 hours post-challenge plasma glucose IDH: Age, DBP, waist circumference, marital status, gender, HDL-C	ISH: 235/4574 IDH: 470/4809	Isolated systolic hypertension (ISH): SBP ≥ 140 mmHg and DBP < 90 mmHg Isolated diastolic hypertension (IDH): SBP <140 mmHg and DBP ≥ 90 mmHg	ISH: Median 9.57 years, IDH: Median 9.62 years	Cox proportional-hazards regression	ISH: C-statistic = 0.91, IDH: C-statistic = 0.76	NR	NR
Sathish et al.[37] 2016	India/ Asians	Prospective cohort	15-64 years	Both male and female	Age, sex, years of schooling, daily intake of fruits or vegetables, current smoking, alcohol use, BP, prehypertension, central obesity,	70/297	SBP ≥140 mm Hg or DBP ≥90 mm Hg or use of antihypertensive medication	Mean 7.1 years	Logistic regression	AUC = 0.802 [0.748-0.856]	Hosmer-Lemeshow p = 0.940	NR

					history of high blood glucose							
Lee et al.[53] 2015	Korea/ Asians	Prospective cohort	40-69 years	Both male and female	BMI, waist circumference, waist-to-hip ratio, waist-to-height ratio	Men: 384/2128 Women: 374/2326	SBP ≥140 mm Hg or DBP ≥90 mm Hg or use of antihypertensive medication	4 years	Cox proportional-hazards regression	Men: AROC = 0.58 [0.56-0.60] (BMI), 0.62 [0.60-0.64] (WC, WHR, WHtR) Women: AROC = 0.57 [0.55-0.59] (BMI), 0.66 [0.64-0.68] (WC), 0.68 [0.66-0.70] (WHR, WHtR)	NR	NR
Lee et al.[54] 2014	Korea/ Asians	Cross-sectional	21-85 years	Both male and female	Women: Height, age, neckC, axillaryC, ribC, waistC, pelvicC, rib_hip, waist_hip, pelvic_hip,	NR/12,789	SBP ≥ 140 mmHg and/or DBP ≥ 90 mmHg or physician-diagnosed hypertension	NR	Logistic regression	Women: AUC = 0.713 (LR-CFS), 0.721 (LR-	NR	Internal, 10-fold cross-validation

					rib_pelvic, axillary_rib, chest_rib, axillary_chest, forehead_neck (CFS), height, weight, BMI, age, chestC, forehead_hip, waist_hip, chest_pelvic, waist_pelvic, axillary_waist, forehead_rib, neck_axillary (LR-wrapper) Men: Age, foreheadC, neckC, axillaryC, chestC, ribC, waistC, pelvicC, hipC, rib_hip, waist_hip, rib_pelvic, waist_pelvic, chest_waist, forehead_rib, chest_rib, axillary_chest, forehead_neck (CFS), height, foreheadC, neckC, axillaryC, ribC, pelvicC, forehead_hip, chest_hip, rib_hip, pelvic_hip, forehead_waist, axillary_waist, rib_waist, neck_rib, axillary_rib,					wrapper) Men: AUC = 0.637 (LR-CFS), 0.652 (LR-wrapper)		

					chest_rib, forehead_axillary, forehead_neck, WHtR (LR-wrapper)							
Kanegae et al.[28] 2017	Japan/ Asians	Prospective cohort	18-83 years	Both male and female	Age, sex, BMI, SBP, DBP, low-density lipoprotein cholesterol, uric acid, proteinuria, current smoking, alcohol intake, eating rate, DBP by age, and BMI by age	7402/63,495	SBP/DBP ≥ 140/90 mm Hg and/or the initiation of antihypertensive medications with self-reported hypertension	Mean 3.4 years	Cox proportional-hazards regression	C-statistic = 0.885 [0.865-0.903]	Greenwood-Nam-D'Agostino $\chi 2$ statistic = 13.6)	External validation
Chen et al.[55] 2016	China/ Asians	Prospective cohort	Average age 41.73 years (men), 39.49 years (women)	Both male and female	Men: Age, BMI, SBP, DBP, gamma-glutamyl transferase, fasting blood glucose, drinking, age x BMI, age x DBP Women: Age, BMI, SBP, DBP, fasting blood glucose, total cholesterol, neutrophil granulocyte, drinking, smoking	2021 (men), 764 (women) 7537 (men), 4960 (women)	First occurrence at any follow-up medical check-up of SBP > 140 mm Hg or DBP > 90 mm Hg or of the person taking antihypertensive medication	Median 4.0 years	Cox proportional-hazards regression	Derivation: AUC = 0.761 [0.752–0.771] (men), 0.753 [0.741–0.765] (women) Validation: AUC = 0.760 [0.751–0.770] (men), 0.749 [0.737–0.761] (women)	NR	Internal, 10-fold cross-validation

Díaz-Gutiérrez et al.[36] 2019	Spain/ Spanish	Prospective cohort	Age presented according to the number of healthy lifestyle factors	Both male and female	No smoking, moderate-to-high physical activity, Mediterranean diet adherence, healthy BMI, moderate alcohol intake, and no binge drinking	1406/14057	SBP ≥ 130 mmHg, DBP≥ 80 mmHg, or use of any antihypertensive drug	Median 10.2 years	Cox regression	NR	NR	NR
Wang et al.[56] 2018	China/ Asians	Longitudinal	18-90 years	Both male and female	Age, sex, education, marriage, smoking, drinking, BMI, energy, carbo, fat, protein	882/5265 (derivation) NR/1597 (validation)	Taking antihypertensive drugs or SBP at least 140 mmHg or DBP at least 90 mmHg	Average follow-up of 8.05 ± 5.27 years	Multistate Markov model	NR	NR	Temporal validation, same data but in a later time
Niiranen et al.[57] 2016	Finland/ Whites	Prospective cohort	≥ 30 years	Both male and female	Model 1: GRS Model 2: Model 1 + age + sex Model 3: Model 2 + smoking, diabetes, education, hyper-cholesterolemia, leisure-time exercise, and BMI	NR/2045	BP ≥ 140/90 mm Hg and/or antihypertensive medication	11 years	Multiple linear and logistic regression	C-index = 0.731 (Model 1) C-index = 0.733 (Model 3)	NR	NR
Yeh et al.[58] 2001	Taiwan/ Chinese	Prospective cohort	≥ 20 years	Both male and female	Age, DM, and fibrinogen concentration (Men) Age and APTT (activated partial thromboplastin time) (Women)	88/2374	SBP ≥140 mm Hg or DBP ≥90 mm Hg	Average 3.23 years	Cox regression	NR	NR	NR

Syllos et al.[29] 2020	Brazil/ South Americans	Prospective cohort	35-74 years	Both male and female	Age, sex, educational level, parental history of hypertension, leisure-time physical activity, BMI, neck circumference, smoking, SBP, DBP	1088/8027; Derivation: 4825 Validation: 3202	SBP ≥ 140 mm Hg, DBP ≥ 90 mm Hg or the use of blood pressure-lowering medications	4 years	Logistic regression	AUC = 0.830 [0.810 - 0.849]	H-L Chi-square = 8.22, p = 0.41	Internal, split sample 6:4 ratio
Wang et al.[35] 2020	China/ Asians	Prospective cohort	≥ 18 years	Both male and female	Age, parental hypertension, SBP, DBP, BMI, and age by BMI	1658/9034	SBP ≥ 140 mm Hg, DBP ≥ 90 mm Hg or the use of blood pressure-lowering medications	Median 6 years	Logistic regression	C-index = 0.795 [0.7733–0.810] (Training set), C-index = 0.7914 [0.773–0.809] (Testing set)	H–L Chi-square = 7.747, P = 0.459 (Training set) H–L Chi-square = 14.366, P = 0.073 (Testing set)	Internal, Bootstrap validation

Xu et al.[59] 2019	China/ Asians	Prospective cohort	35-74 years	Both male and female	M1 Model: Age, SBP, DBP, hypertension parental history, WC, interaction item of age with WC, and interaction item of age with DBP W1 Model: Age, SBP, DBP, WC, fruit and vegetable intake, hypertension parental history, interaction item of age with WC, and interaction of age with DBP were included in W1 model	1036/4796 (Training)	SBP ≥ 140 mm Hg and/or DBP ≥ 90 mm Hg, and/or a diagnosis of hypertension by a physician and currently receiving anti-hypertension treatment	6 years	Cox regression	Testing Set Men: AUC=0.771 [0.750-0.791] (M1) Testing Set Women: AUC = 0.765 [0.746-0.783] (W1), 0.764 [0.746-0.783] (W2)	Testing Set Men: Modified Nam-D'Agostino test Chi-square = 6.305, p=0.708 (M1) Testing Set women: Modified Nam-D'Agostino test Chi-square = 6.783, p = 0.147(W1); 7.404, p = 0.115 (W2)	Internal, 10-fold cross-validation in training data and external in the testing data

Kadomatsu et al.[34] 2019	Japan/ Asians	Prospective cohort	Mean age 51.3 years	Both male and female	Age, sex, BMI, current smoking habit, ethanol consumption, presence of DM, parental hypertension history, SBP, DBP	324/3936	SBP ≥ 140 mm Hg, DBP ≥ 90 mm Hg, or use of antihypertensive medication	Median 5 years	Logistic regression	AUC = 0.826 [0.804-0.848] (Entire cohort validation) Median AUC = 0.83 [0.828-0.832] (Cross-validation)	H–L Chi-square = 7.06, p = 0.53, (Entire cohort validation); H–L Chi-square = 12.2 (Cross-validation)	Internal, split-sample cross-validation 6:4 ratio
Wang et al.[60] 2015	USA/ Multi-ethnic	Telephone-based health survey	≥ 18 years	Both male and female	Exercise, diabetes, hyperlipemia, age, marriage, education, income, weight, height, sex, smoke, drink	NR/308,711	NR	NR	Logistic regression	Accuracy, sensitivity, specificity, and AUC. AUC = 0.74±0.001 (logistic), Accuracy = 71.96% (logistic)	NR	Internal, split sample 7:3 ratio

Muntner et al.[61] 2010	USA/ Multi-ethnic (Whites, Blacks, Hispanics, and Asians –primarily of Chinese descent)	NR	45-84 years	Both male and female	SBP-alone model (7 SBP categories) Age-specific categories of DBP model (20 categories)	849/3013	The first study visit, subsequent to baseline, at which SBP ≥ 140 mm Hg and/or DBP ≥ 90 mm Hg and/or the initiation of antihypertensive medication	Median of 1.6 years and 4.8 years	Repeated-measures Poisson regression model	SBP model: C-statistic = 0.768 [0.751 - 0.785] (1.6 years follow-up), 0.773 [0.775 - 0.791] (4.8 years follow-up Age-specific DBP Model: C-statistic = 0.699 [0.681 - 0.717] (1.6 years follow-up), 0.691 [0.671 - 0.711] (4.8 years follow-up)	NR	NR

Ture et al.[62] 2005	Turkey/ Europeans	Retrospective	Average 48.2 years (hypertension) 46.5 (control)	Both male and female	Age, sex, family history of hypertension, smoking habits, lipoprotein (a), triglyceride, uric acid, total cholesterol, and BMI	694 (452 patients with hypertension and 242 controls)	Average of 3 or more DBP measurements on at least 3 subsequent visits is ≥ 90 mmHg or when the average of multiple SBP readings on 3 or more subsequent visits is consistently ≥ 140 mmHg	NR	Four statistical algorithms (logistic regression analysis, Flexible discriminant analysis, multivariate additive regression splines (degree 1), multivariate additive regression splines (degree 2)	Sensitivity, specificity, and predictive rate (PR)	NR	Internal, split sample 3:1 ratio
Yamakado et al.[63] 2015	Japan/ Asians	Prospective cohort	≥ 20 years	Both male and female	PFAA Index 1: Leucine, alanine, tyrosine, asparagine, tryptophan, and glycine; PFAA Index 2: Isoleucine, alanine, tyrosine, phenylalanine, methionine, and histidine	424/2637	SBP ≥ 140 mm Hg or DBP ≥ 90 mm Hg or use of antihypertensive medication	4 years	Logistic regression	NR	NR	Internal, leave-one-out cross-validation (LOOCV) and validation in a cohort dataset

Qi et al.[64] 2014	China/ Asians	Case-control	Case cohort: 64.48 ± 8.53 years Control: 64.23 ± 10.13 years	Both male and female	rs17030613, rs16849225, rs1173766, rs11066280, rs35444, rs880315, rs16998073, rs11191548, rs17249754	Patients: NR/1009 Controls = NR/756	SBP ≥ 140 mm Hg or DBP ≥ 90 mm Hg or use of antihypertensive medication	NR	Logistic regression	NR	NR	NR
Lu et al.[65] 2015	China/ Asians	Prospective cohort	35-74 years	Both male and female	Model1: GRS+ (age, sex, and BMI); Model2: GRS +Model1 + smoking, drinking, pulse rate, and education Model3: GRS+ Model2 + SBP and DBP	2559/7724	SBP ≥ 140 mm Hg or DBP ≥ 90 mm Hg or use of antihypertensive medication	Mean 7.9 years	Logistic regression and Cox proportional-hazards regression	Model1: C-statistic = 0.650 [0.637-0.663] (without GRS), 0.655 [0.642-0.668] (with GRS) Model 2: C-statistic = 0.683 [0.670-0.695] (without GRS), 0.687 [0.675-0.700] (with GRS) Model 3: C-statistic = 0.774 [0.763-0.785] (without	NR	NR

										GRS), 0.777 [0.766-0.787] (with GRS)		
Zhang et al.[66] 2015	China/ Asians	Prospective cohort	18-88 years	Both male and female	Five latent factors extracted from 11 biomarkers (BMI, SBP, DBP, FBG, TG, HDL-C, Hb, HCT, WBC, LC, NGC): inflammatory factor, blood viscidity factor, insulin resistance factor, blood pressure factor, lipid resistance factor, and age	3793/17,471	SBP ≥ 140 mm Hg or DBP ≥ 90 mm Hg or use of antihypertensive medication	5 years	Cox proportional-hazards regression	Derivation cohort: AUC = 0.755 [0.746-0.763] (men), AUC = 0.801 [0.792-0.810] (women) Validation cohort: AUC = 0.755 [0.746-0.763] (men), AUC = 0.800 [0.791-0.810] (women)	NR	Internal, 10-fold cross-validation

Table 2.3 Information about existing hypertension prediction models developed using machine learning algorithms from selected studies

Study	Location of Data used for Model Developed	Sample Size	Risk Factors Included	Outcome Considered	Definition of Outcome Predicted	Modeling Method Used	Performance Measure
Falk CT[67] 2003	USA	300 records for training and 300 for validating	Seven input values: sex; age; total cholesterol; fasting glucose; fasting HDL; fasting triglycerides; body mass index (BMI)	High blood pressure	SBP > 140 mm Hg or DBP > 90 mm Hg	Two neural network programs: NNdriver and SNNS	Classification success rate. Training: 91%-98%, (Strategy 1), 70%-87% (Strategy 2); Validation: 59% (Strategy 1), 63% (Strategy 2)
Farran et al.[68] 2013	Kuwait	10,632 (6759 hypertensive and 3873 non-hypertensive)	BMI, age, ethnicity, and diagnosis for diabetes	Incident hypertension, type 2 diabetes, and comorbidity	NR	Logistic regression (LR), k-nearest neighbors, support vector machines, and multifactor dimensionality reduction (MDR)	Classification accuracy: 90% (hypertension)
Huang et al.[38] 2010	China	Training: 2438, Validation: 616	High educational level, predominantly sedentary work, positive family history of HTN, overweight, dysarteriotony, alcohol intake, salty diet, more vegetable and fruit intake, meat consumption, and regular physical exercise	Hypertension	Average SBP or DBP > 139 mmHg or > 89 mmHg, respectively	Logistic regression model (LRM) and artificial neural network (ANN) model (back-propagated delta rule networks)	AUC: 0.900 ± 0.014 (ANN model) AUC: 0.732 ± 0.026 (LRM)

Kwong et al.[69] 2018	NR	498	Age, BMI, exercise level, alcohol consumption level, smoking status, stress level, and salt intake level	Systolic blood pressure (SBP)	BP readings > 140 mmHg	Two artificial neural networks (ANN): Back-propagation (BP) neural network and radial basis function (RBF) neural network validate the prediction system	Average Accuracy, BP ANN: 94.28% (male), 93.74% (female) RBF ANN: 91.06% (male), 90.44% (female)
Polak et al.[70] 2008	USA	159,989 records	High blood cholesterol, number of cigarettes smoked now, age, weight, height, sex	Hypertension	NR	Artificial neural network (ANN): Around 250 architectures of backpropagation (BP) and fuzzy networks	Classification rate and AUROC, different values for different Nets architecture
Priyadarshini et al.[71] 2018	USA	NR	SBP, DBP, total cholesterol (TC), high-density lipoprotein (HDL), low-density lipoprotein (LDL), plasma glucose concentration (PGC), and heart rate (HR)	Hypertension attack	DBP or SBP > 90 mm Hg or > 120 mm Hg, respectively, for at least two measuring instances	Deep neural network model	Confusion/performance matrix formed out of four evaluating parameters: accuracy 88%, precision 92%, recall 82%, and F1 score 76% (average value over 20 iterations)

Sakr et al.[39] 2018	USA	23,095	Age, METS, resting systolic blood pressure, peak diastolic blood pressure, resting diastolic blood pressure, HX coronary artery disease, the reason for the test, history of diabetes, percentage HR achieved, race, history of hyperlipidemia, Aspirin use, hypertension response	Hypertension	NR	Six machine learning techniques: LogitBoost (LB), Bayesian network classifier (BN), locally weighted naïve Bayes (LWB), artificial neural network (ANN), support vector machine (SVM), and random tree forest (RTF)	AUC, F-Score, Sensitivity, Specificity, Precision, and RMSE. AUC (0.93), F-Score (86.70%), Sensitivity (69,96%) and Specificity (91.71%) for RTF model in 10-fold cross-validation AUC (0.88), Sensitivity (74.30%), Precision (73.50%), and F-Score (73.90%) for RTF model in holdout method
Tayefi et al.[72] 2016	Iran	9078	Age, gender, BMI, marital status, level of education, occupation status, depression and anxiety status, physical activity level, smoking status, LDL, triglyceride, total cholesterol, fasting blood glucose, uric acid, and hs-CRP in Model 1 Age, gender, white blood cell, red blood cell, hemoglobin, hematocrit, mean corpuscular volume, mean corpuscular hemoglobin, platelets, red cell distribution width and platelet distribution width in Model 2	Hypertension	SBP of 140 mm Hg, DBP of 90 mm Hg, and/or current use of antihypertensive drugs	Decision tree	Accuracy, sensitivity, specificity, and area under the ROC curve (AUC): For Model 1, the values are 73%, 63%, 77% and 0.72, respectively, and for Model 2 were 70%, 61%, 74% and 0.68, respectively
Wu et al.[73] 2015	USA	75 females and 165 males	Age, gender, serum cholesterol, fasting blood sugar and electrocardiographic signal, heart rate	Systolic blood pressure	SBP and DBP > 140 mm Hg and 90 mm Hg, respectively	Two neural network algorithms: back-propagation neural network and radial basis function network	The absolute difference (error) between the real value and predicted values

Wu et al.[74] 2016	NR	498	Age, BMI, gender, exercise level, alcohol consumption, stress level, salt intake level, smoke status, cholesterol, and blood glucose	Systolic blood pressure	SBP > 140 mm Hg	Two artificial neural networks: back-propagation neural network and radial basis function neural network	The average prediction errors (absolute difference between the predicted value and measured value): 51.9% for men and 52.5% for women (backpropagation neural network) 51.8% for men and 49.9% for women (radial basis function network)
Ye et al.[40] 2018	USA	823,627 (training cohort/retrospective cohort), 680,810 (validation cohort/prospective cohort)	Total 169 features: 2 demographic features, 14 socioeconomic characteristics, 30 diagnostic diseases, 6 laboratory tests, 98 medication prescriptions, and 19 clinical utilization measures	Incident essential hypertension	International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM) diagnosis codes from category 401	A supervised machine learning and data mining tool, XGBoost	AUC = 0.917 (retrospective cohort) AUC = 0.870 (prospective cohort)
Zhang et al.[75] 2017	NR	Data collected from CM400 monitor. A total of 15,628,501 sets of valid characteristic attributes data	Seven input features: right atrium (AVR), left atrium (AVL), anterior atrium (AVF), photoplethysmography (PPG), oxygen saturation (SPO2), pulse transit time (PTT), heart rate (HR)	Blood pressure	NR	CART (classification and regression tree) model	Four evaluation indexes: accuracy rate, root mean square error (RMSE), deviation rate, and the Theil inequality coefficient (TIC)

Völzke et al.[27] 2013	Germany	Training set: 803 Validation set: 802 External validation cohort: 2887	Age, mean arterial pressure, rs16998073, serum glucose, and urinary albumin concentrations, the interaction between age and serum glucose, interaction between rs16998073 and urinary albumin concentrations	Incident hypertension	SBP ≥ 140 mmHg and DBP ≥ 90 mmHg	Bayesian network	Training set: AUC = 0.78 [0.74-0.82] Validation set: AUC = 0.79 [0.75-0.83] External validation set: AUC = 0.77 [0.74-0.80] Training set: HL Chi-square = 11.82 (p = 0.16) Validation set: HL Chi-square = 11.65 (p = 0.17) External validation set: H-L Chi-square = 40.6 (p < 0.01)
Lee et al.[54] 2014	Korea	12,789	Women: Height, age, neckC, axillaryC, ribC, waistC, pelvicC, rib_hip, waist_hip, pelvic_hip, rib_pelvic, axillary_rib, chest_rib, axillary_chest, forehead_neck (CFS), height, ge, foreheadC, eckC, hipC, axillary_hip, axillary_pelvic, chest_pelvic, chest_rib (NB-wrapper) Men: Age, foreheadC, neckC, axillaryC, chestC, RibC, waistC, pelvicC, hipC, rib_hip, waist_hip, rib_pelvic, waist_pelvic, chest_waist, forehead_rib, chest_rib, axillary_chest, forehead_neck (CFS), height, age, foreheadC, neckC, axillaryC, hipC, rib_hip, pelvic_hip, neck_pelvic, waist_pelvic, chest_waist, chest_rib, neck_chest, forehead_neck (NB-wrapper)	Hypertension and hypotension	SBP ≥ 140 mmHg and/or DBP ≥ 90 mmHg or physician-diagnosed hypertension	Naive Bayes algorithm (NB)	Women: AUC = 0.696 (NB-CFS), 0.713 (NB-wrapper) Men: AUC = 0.64 (NB-CFS), 0.646 (NB-wrapper)

Xu et al.[59] 2019	China	4796	M1 Model: Age, SBP, DBP, hypertension parental history, WC, interaction item of age with WC, and interaction item of age with DBP W1 Model: Age, SBP, DBP, WC, fruit and vegetable intake, hypertension parental history, interaction item of age with WC, and interaction of age with DBP	Hypertension	SBP ≥ 140 mm Hg and/or DBP ≥ 90 mm Hg and/or a diagnosis of hypertension by a physician and currently receiving anti-hypertension treatment	Artificial neural network (ANN), naive Bayes classifier (NBC), and classification and regression tree (CART)	Testing Set Men: AUC= 0.773 [0.752-0.793] (ANN), 0.760 [0.738-0.781] (NBC), 0.722 [0.699-0.743] (CART) Testing Set Women: AUC = 0.756 [0.737-0.775] (ANN), 0.761 [0.742-0.779] (NBC), 0.698 [0.677-0.717] (CART) Testing Set Men: Modified Nam-D'Agostino test Chi-square = 29.274, p = 0.0006 (ANN); 82.269, p < 0.00001 (NBC); 5.249, p =0.072 (CART) Testing Set women: Modified Nam-D'Agostino test Chi-square = 4.744, p = 0.314 (ANN); 189.754, p < 0.00001 (NBC); 19.733, p = 0.00005 (CART)
Wang et al.[60] 2015	USA	308,711	Exercise, diabetes, hyperlipemia, age, marriage, education, income, weight, height, sex, smoke, drink	Hypertension	NR	Multi-layer perception neural network	Accuracy, sensitivity, specificity, and AUC. Average AUC = 0.77 with h vary from 8 to 11 (neural network); Accuracy = 72% (neural network)

Ture et al.[62] 2005	Turkey	694	Age, sex, family history of hypertension, smoking habits, lipoprotein (a), triglyceride, uric acid, total cholesterol, and BMI	Essential hypertension	The average of 3 or more DBP measurements on at least 3 subsequent visits is ≥ 90 mmHg, or when the average of multiple SBP readings on 3 or more subsequent visits is consistently ≥ 140 mmHg	Three decision trees (Chi-squared automatic interaction detector. Classification and regression tree, quick, unbiased, efficient statistical tree); two neural networks (multi-layer perceptron, radial basis function)	Sensitivity, specificity, and predictive rate (PR). Values not reported.
Zhao et al.[76] 2008	China/ Asians	Total: 4759 (2411 hypertensive and 2,348 age-matched and sex-matched healthy controls)	MDR Model: 4-locus model consisted of the SNP KCNMB1-rs11739136, RGS2-rs34717272, PRKG1-rs1881597, and MYLK-rs36025624; CART Model: RGS2, PRKG1, KCNMB1, and MYLK	Hypertension CHECK	Average SBP ≥ 150 mm Hg, an average DBP ≥ 95 mm Hg, or current use of antihypertensive medication	Multifactor-dimensionality reduction (MDR) and classification and regression trees (CART)	MDR Model: Accuracy = 52.98%, cross-validation consistency = 9.7
Wang et al.[60] 2014	China/ Asians	1009 hypertensive patients and 756 normotensive controls	Genes	Hypertension	Mean SBP ≥ 140 mmHg and/or DBP ≥ 90 mmHg on two occasions and/or the current usage of antihypertensive drug treatment	Multifactor dimensionality reduction (MDR) model	The best MDR model testing accuracy = 0.6331, cross-validation consistency = 10

Zhao et al.[77] 2014	China/ Asians	1009 hypertensive patients and 756 normotensive controls	The best MDR model included rs5804 and BMI	Hypertension	Mean SBP of at least 140 mm Hg or a mean DBP of at least 90 mm Hg or the current intake of antihypertensive drugs	Multifactor dimensionality reduction (MDR) model	The best MDR model: testing accuracy of 0.7309 and a maximum cross-validation consistency of 10 ($P < 0.001$)

Table 2.4 Information about external validation studies of existing traditional hypertension prediction models from selected studies

Study Name/Prediction Model Validated	Total Number of Validation Studies	Validation Study	Location/Ethnicity	Age	Follow-up Period	Events (n)/Total Participants (N)	Outcome Definition	Calibration	Discrimination
Parikh et al.[25] 2008/Framingham Hypertension Risk Model (FHRS)	8	Zheng et al.[78] 2014	China/Asians	≥ 35 years	Median 4.8 years	8675/24,434	Average SBP ≥140 mm Hg, and/or DBP ≥ 90 mm Hg, and/or use of antihypertensive medications within 2 weeks before the follow-up examination	H–L Chi-square test = 2,287.7 (P < 0.0001), 2-year incidence of hypertension H–L Chi-square test = 8,227.1 (P < 0.0001), 4-year incidence of hypertension	C statistics = 0.537 [0 .524–0.550], 2-year incidences of hypertension C statistics = 0.610 [0.602–0.618], 4-year incidences of hypertension
		Muntner et al.[61] 2010	USA/Multiethnic (Whites, Blacks, Hispanics, and Asians–primarily of Chinese descent)	45-84 years	Median of 1.6 years and 4.8 years	849/3013	The first study visit, subsequent to baseline, at which SBP ≥ 140 mm Hg and/or DBP ≥ 90 mm Hg and/or the initiation of antihypertensive medication	H-L goodness of fit Chi-square: p < 0.001	C-statistic = 0.788 [0.773 - 0.804] (1.6 years follow-up) C-statistic = 0.792 [0.775-0.807] (4.8 years follow-up)

		Carson et al.[79] 2013	USA/Whites and Blacks	18-30 years	25 years	1179/4388	First study examination in which SBP ≥ 140 mm Hg or DBP ≥ 90 mm Hg or initiated treatment with antihypertensive medications	Modified H–L goodness of fit $\chi 2 = 249.4$; $P < 0.001$	C-index = 0.84 [0.83–0.85]
		Lim et al.[80] 2016	Korea/Asians	40–69 years	4 years	13005/69,918	SBP ≥ 140 mmHg or DBP ≥ 90 mmHg on health examination or a record with hypertensive disease codes (I10–I13) and prescription of one of the antihypertensive agents	H-L Chi-square p < 0.001	AROC = 0.729
		Kivimäki et al.[46] 2009	England/Mainly Whites	35-68 years	Median 5.6 years	NR/5472	SBP ≥ 140 mmHg or DBP ≥ 90 mmHg or use of blood pressure-lowering medications	H-L Chi-square = 11.5	C-statistic = 0.803
		Wang et al.[35] 2020	China/Asians	≥ 18 years	Median 6 years	1658/9034	SBP ≥ 140 mm Hg, DBP ≥ 90 mm Hg, or the use of blood pressure-lowering medications	NR	AUC = 0.787 [0.778–0.795]
		Syllos et al.[29] 2020	Brazil/South Americans	35-74 years	4 years	1088/8027; Derivation: 4825 Validation: 3202	SBP ≥ 140 mm Hg, DBP ≥ 90 mm Hg, or the use of blood pressure-lowering medications	H-L Chi-square = 3.78, p = 0.876	AUC = 0.827 [0.808 - 0.847]

		Völzke et al.[27] 2013	Denmark/Whites	20-79 years	5.4 ± 0.2 years	434/2887	SBP ≥ 140 mmHg and DBP ≥ 90 mmHg	Validation dataset: H-L Chi-square = 11.26 (p = 0.19) External validation dataset: H-L Chi-square = 203.34 (p < 0.001)	Validation dataset: AUC = 0.77 [0.73 – 0.82] External validation dataset: AUC = 0.73 [0.71-0.75]
Lim et al.[26] 2013/ Korean Genome Epidemiology Study (KoGES)	1	Lim et al.[80] 2016	Korea/Asians	40–69 years	4-year	13,005/69,918	SBP ≥ 140 mmHg or DBP ≥ 90 mmHg on health examination, or a record with hypertensive disease codes (I10–I13) and prescription of one of the antihypertensive agents	H-L Chi-square p = 0.062	AROC = 0.733
Völzke et al.[27] 2013	1	Völzke et al.[27] 2013	Denmark/Whites	20-79 years	5.4 ± 0.2 years	434/2887	SBP ≥ 140 mmHg and DBP ≥ 90 mmHg	H-L Chi-square = 40.6 (p < 0.001)	AUC = 0.77 [0.74 – 0.80]
Kanegae et al.[28] 2017	1	Kanegae et al.[28] 2017	Japan/Asians	18-89 years	Mean 2.4 years	NR/14,168	SBP/DBP ≥ 140/90 mm Hg and/or the initiation of antihypertensive medications with self-reported hypertension	Greenwood-Nam-D'Agostino $\chi 2$ statistic = 8.7	C-statistic = 0.846 [0.775-0.905]

Table 2.5 Study quality assessment using PROBAST

Study	Risk of Bias (ROB)				Applicability			Overall	
	Participants	Predictors	Outcome	Analysis	Participants	Predictors	Outcome	ROB	Applicability
Pearson et al.[44] (1990)	-	+	+	?	-	+	+	-	-
Parikh et al.[25] (2008)	+	+	+	+	-	+	+	+	-
Paynter et al.[45] (2009)	+	+	+	+	-	+	+	+	-
Kivimaki et al.[46] (2009)	+	+	+	+	-	+	+	+	-
Kivimaki et al.[47] (2010)	+	+	+	+	-	+	+	+	-
Kshirsagar et al.[48] (2010)	+	+	+	+	+	+	+	+	+
Bozorgmanesh et al.[33] (2011)	+	+	+	+	+	+	+	+	+
Chien et al.[32] (2011)	+	+	+	+	+	+	+	+	+
Fava et al.[49] (2013)	+	+	+	?	+	+	-	?	-
Lim et al.[26] (2013)	+	+	+	?	+	+	+	?	+
Choi et al.[50] (2014)	+	+	+	?	?	-	+	?	-
Lim et al.[51] (2015)	+	+	+	+	+	-	+	+	-
Otsuka et al.[31] (2015)	+	+	+	?	-	+	+	?	-
Asgari et al.[52] (2016)	+	-	+	+	+	+	-	-	-
Sathish et al.[37] (2016)	+	+	+	-	+	+	+	-	+
Lee et al.[53] (2015)	+	+	+	?	+	+	+	?	+
Lee et al.[54] (2014)	+	+	+	+	+	-	+	+	-
Kanegae et al.[28] (2017)	+	+	+	+	-	+	+	+	-
Chen et al.[55] (2016)	+	+	+	+	+	+	+	+	+
Diaz-Gutierrez et al.[36] (2019)	+	+	+	+	-	+	+	+	-
Wang et al.[56] (2018)	+	+	+	+	+	+	-	+	-
Niiranen et al.[57] (2016)	+	+	+	+	+	-	+	+	-

Yeh et al.[58] (2001)	+	+	+	?	+	-	-	?	-
Syllos et al.[29] (2020)	+	+	+	+	+	+	+	+	+
Wang et al.[35] (2020)	+	+	+	+	+	+	+	+	+
Xu et al.[59] (2019)	+	+	+	-	+	+	+	-	-
Kadomatsu et al.[34] (2019)	+	+	+	?	+	+	+	?	+
Wang et al.[60] (2015)	+	+	+	+	+	+	+	+	+
Muntner et al.[61] (2010)	+	+	+	+	+	+	+	+	+
Ture et al.[62] (2005)	+	+	+	+	?	+	+	+	?
Yamakado et al.[63] (2015)	-	-	-	-	+	-	-	-	-
Qi et al.[64] (2014)	+	+	+	?	-	-	-	?	-
Lu et al.[65] (2015)	+	+	+	?	+	-	+	?	-
Zhang et al.[66] (2015)	+	+	+	?	+	-	+	?	-

Figure S2.1

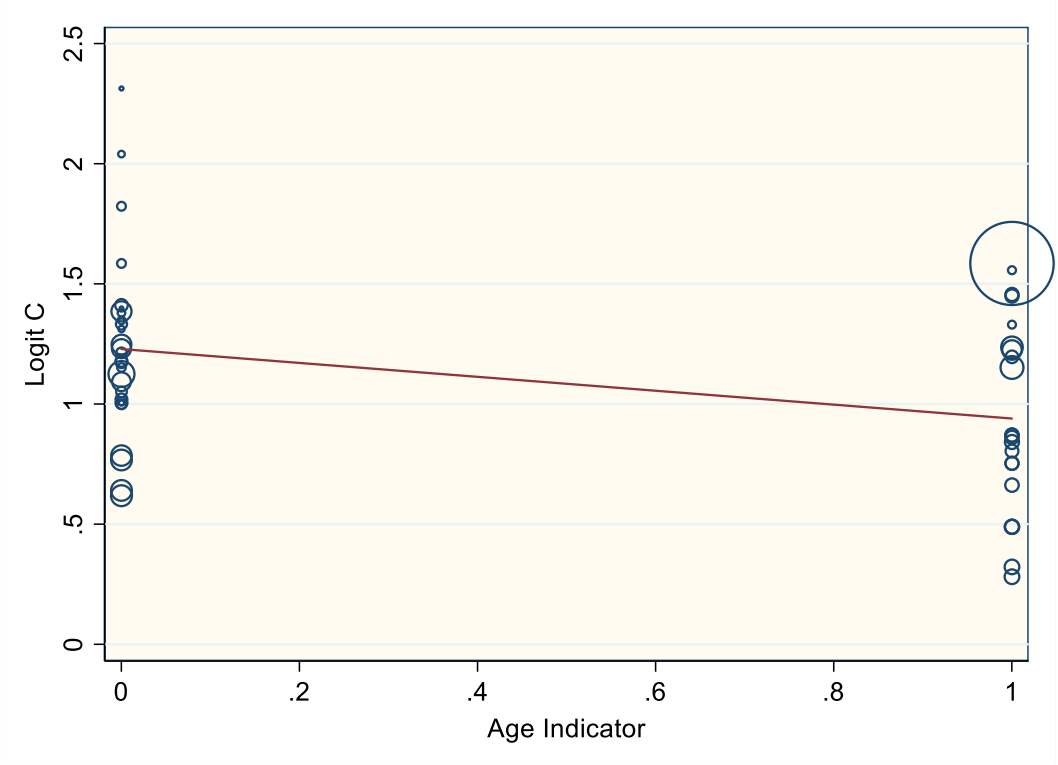

Figure S2.1 Meta-regression on the age of the participants (study participants below average age versus above average age).

Figure S2.2

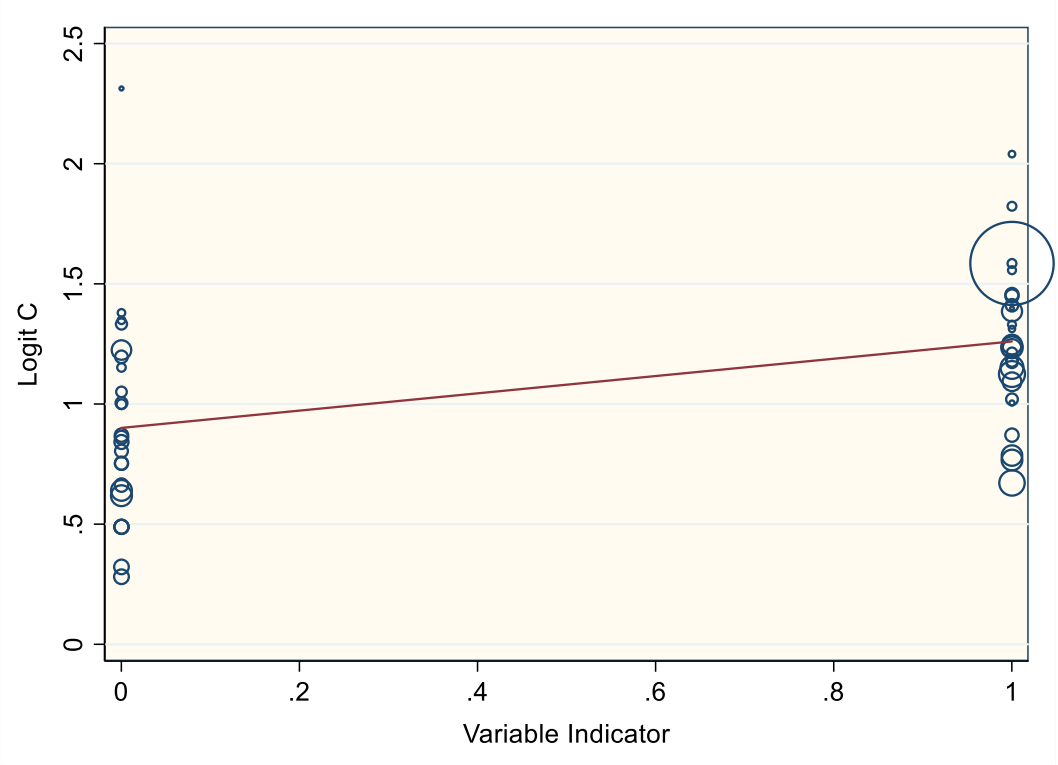

Figure S2.2 Meta-regression on the number of risk factors considered in the model (below median versus above median).

Figure S2.3

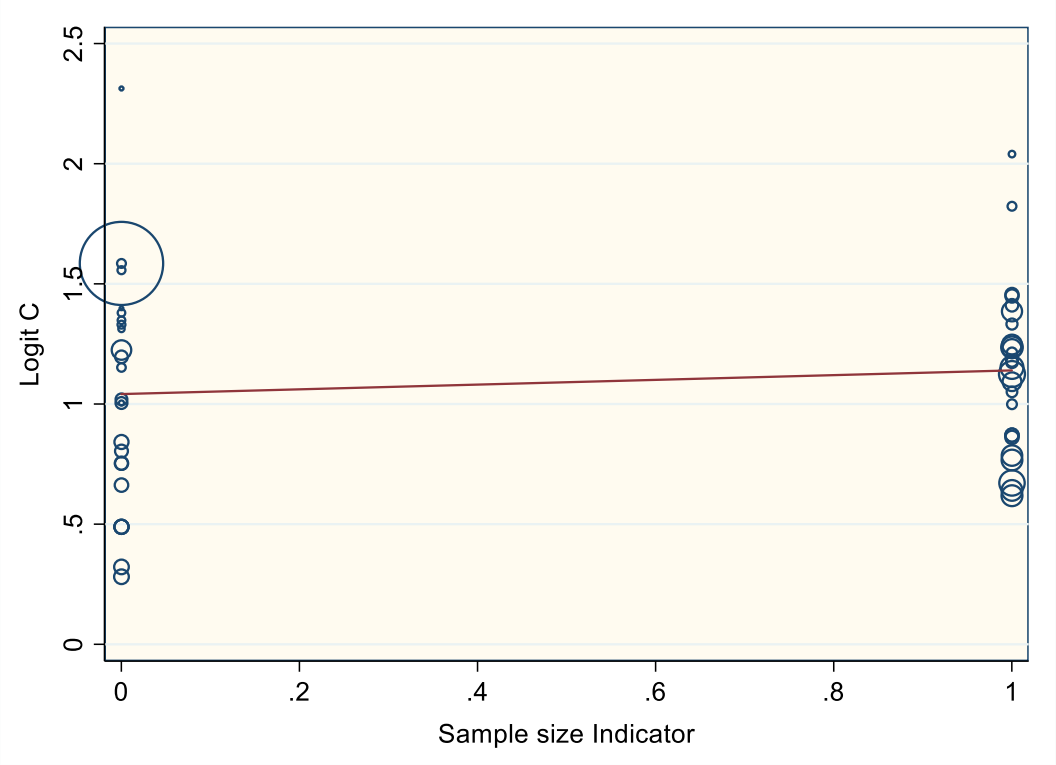

Figure S2.3 Meta-regression on sample size considered in the model (below median versus above median).

Figure S2.4

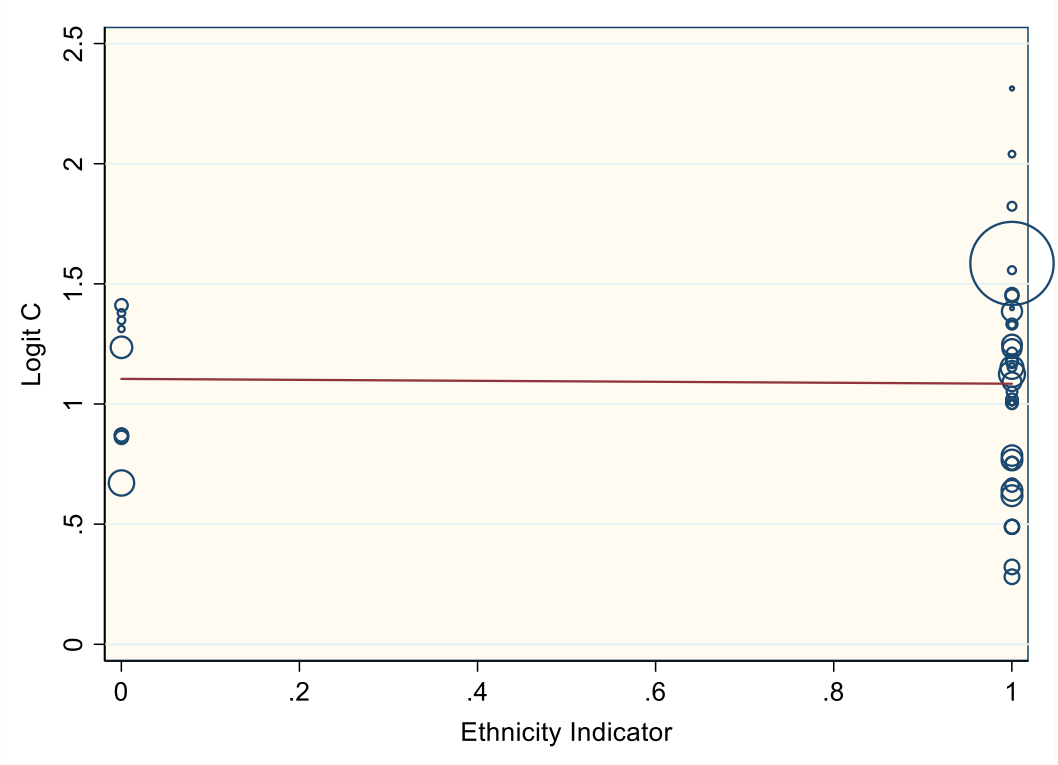

Figure S2.4 Meta-regression on the ethnicity of the study participants (Whites versus Asians).

Figure S2.5

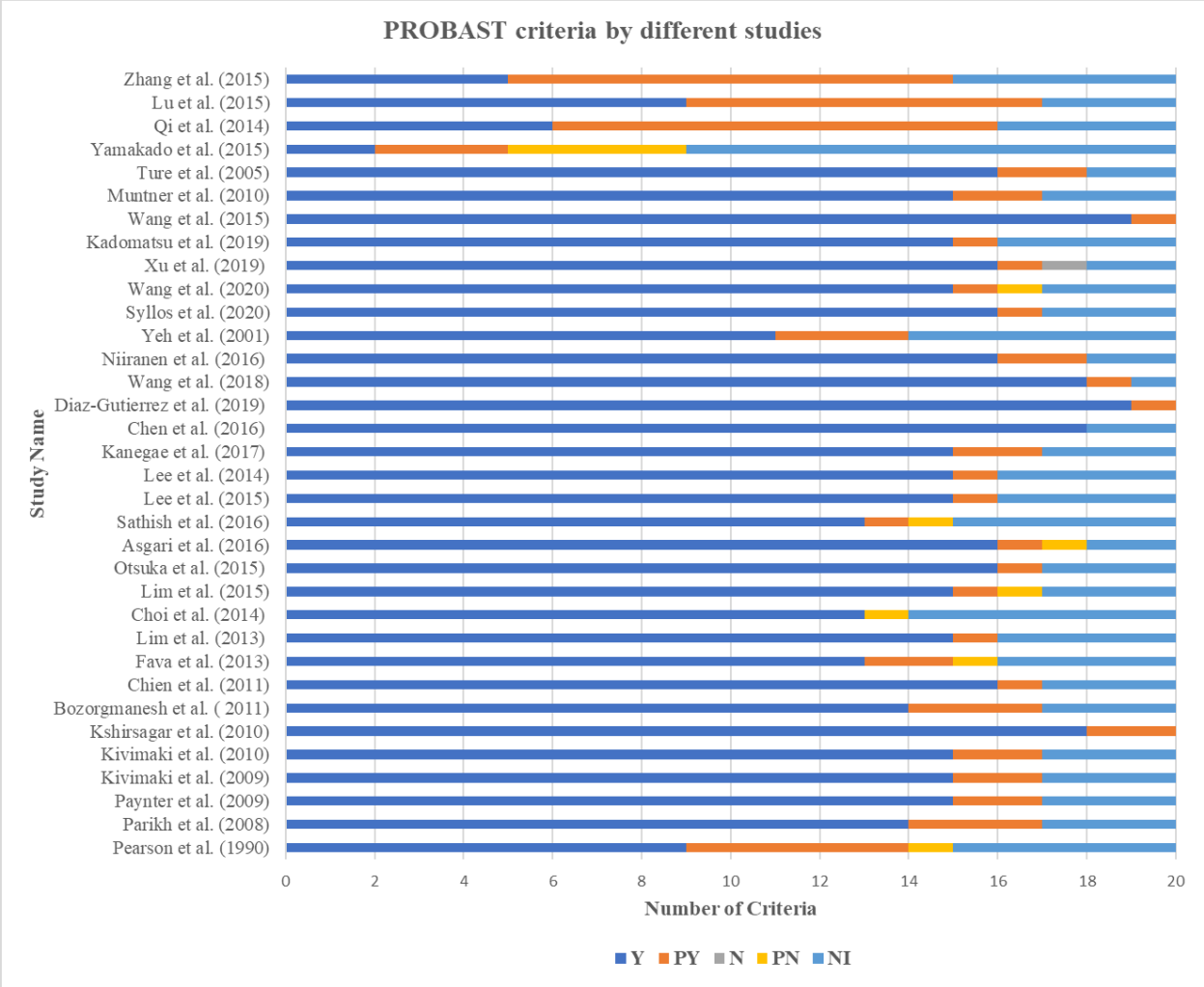

Figure S2.5 The number of PROBAST criteria satisfied by different studies.

Figure S2.6

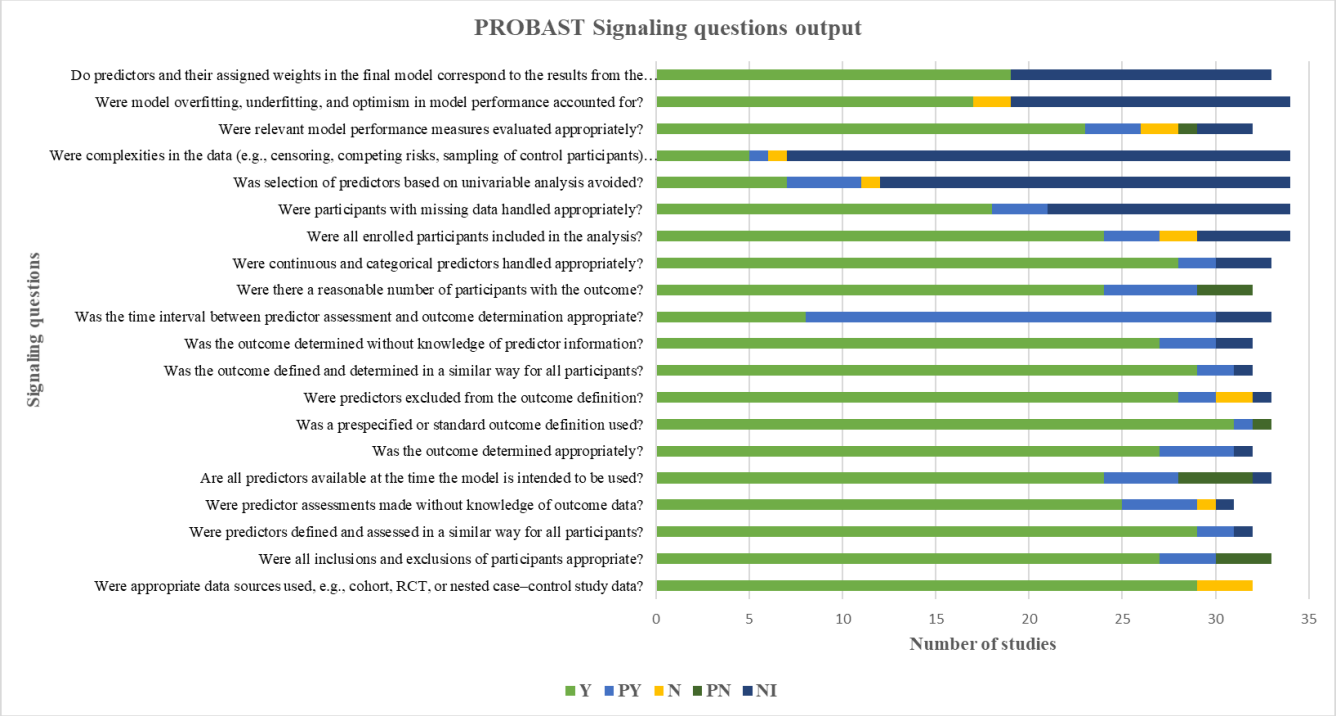

Figure S2.6 Response to different signaling questions by the number of studies.

Figure S2.7

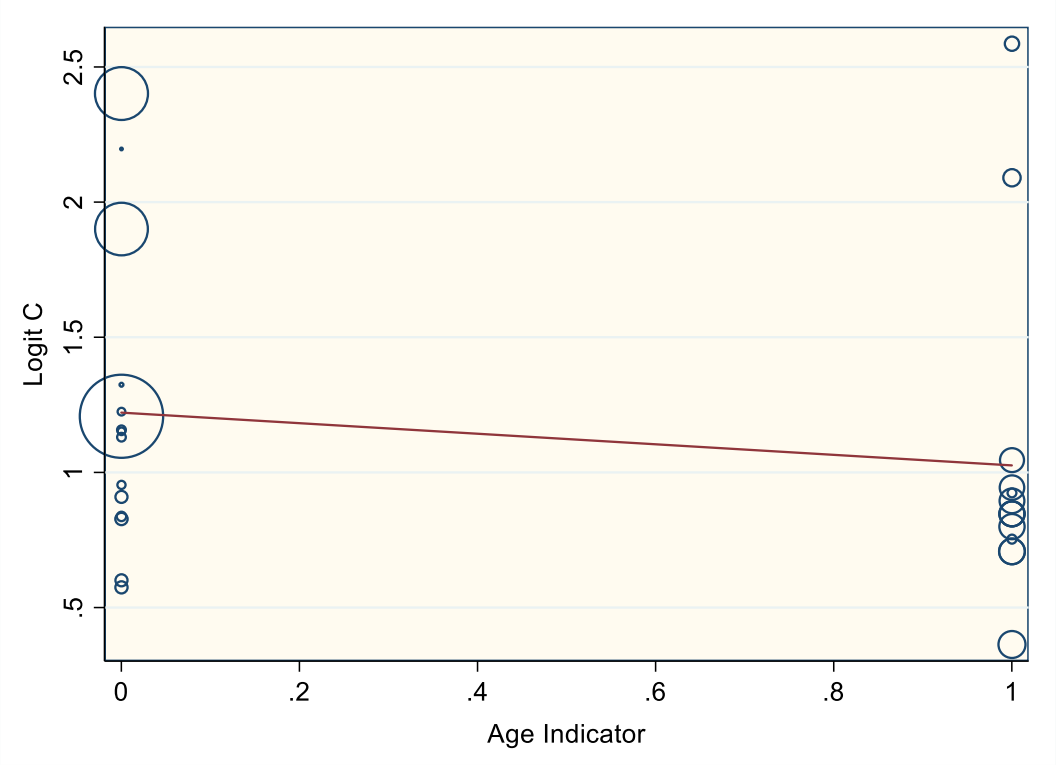

Figure S2.7 Meta-regression on the age of the participants (study participants below average age versus above average age).

Figure S2.8

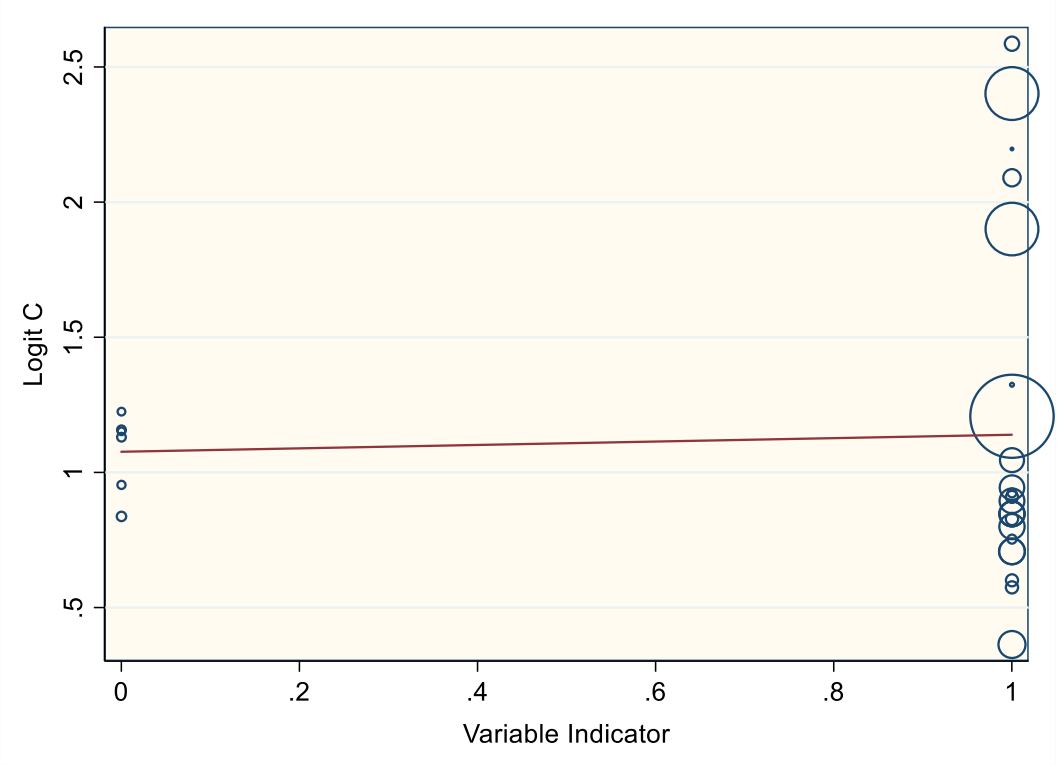

Figure S2.8 Meta-regression on the number of risk factors considered in the model (below median versus above median).

Figure S2.9

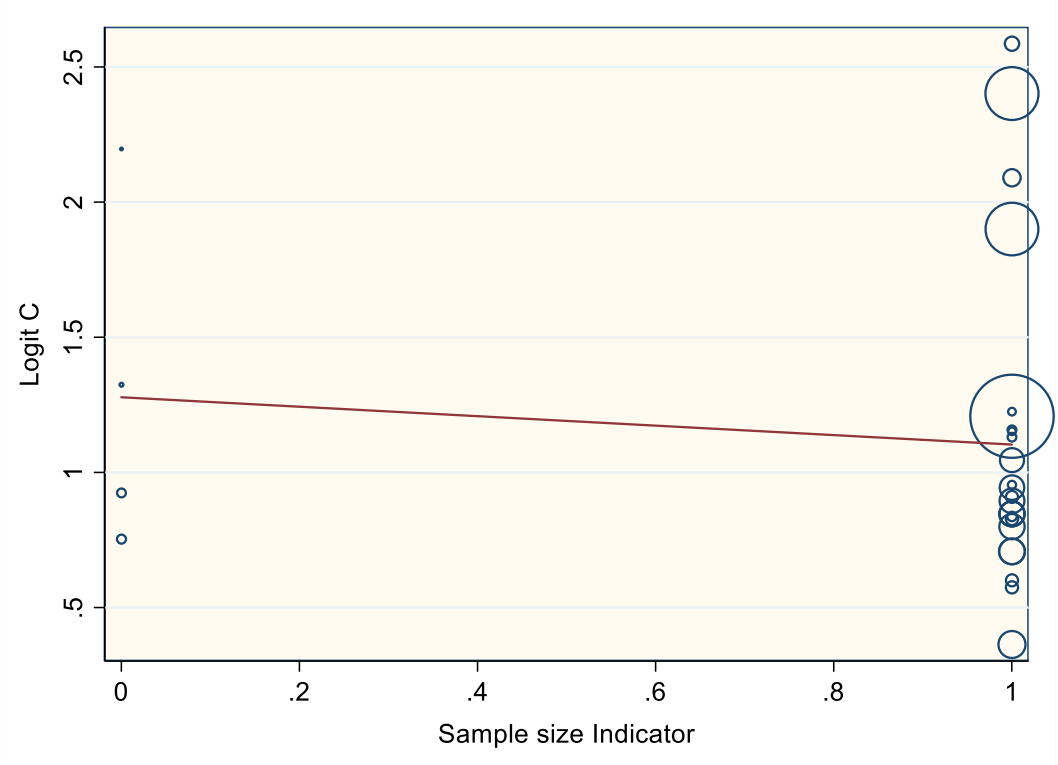

Figure S2.9 Meta-regression on sample size considered in the model (below median versus above median).

Figure S2.10

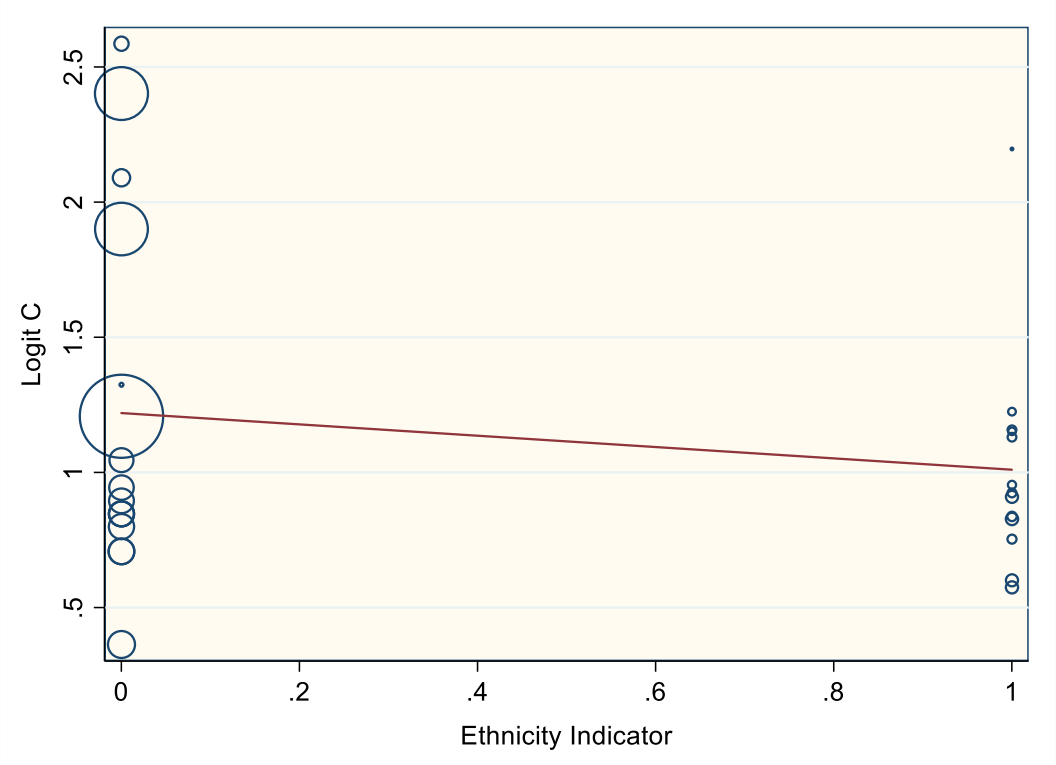

Figure S2.10 Meta-regression on the ethnicity of the study participants (Whites versus Asians).

Figure S2.11

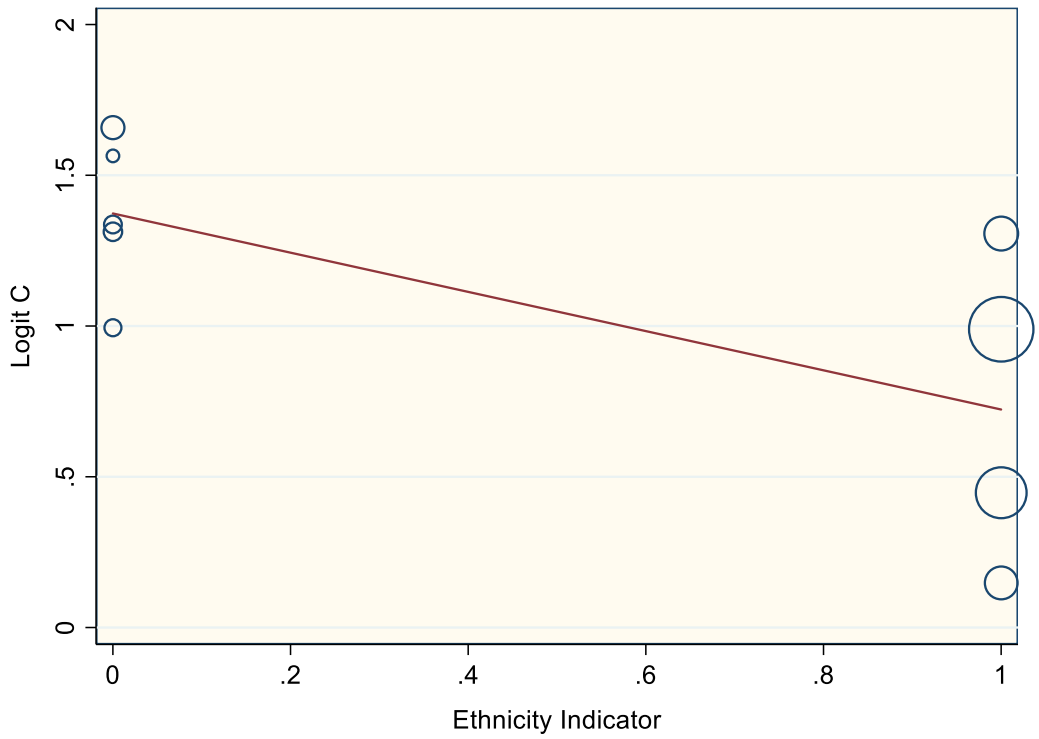

Figure S2.11 Meta-regression on the ethnicity of the study participants (Whites versus Asians).

Figure S2.12

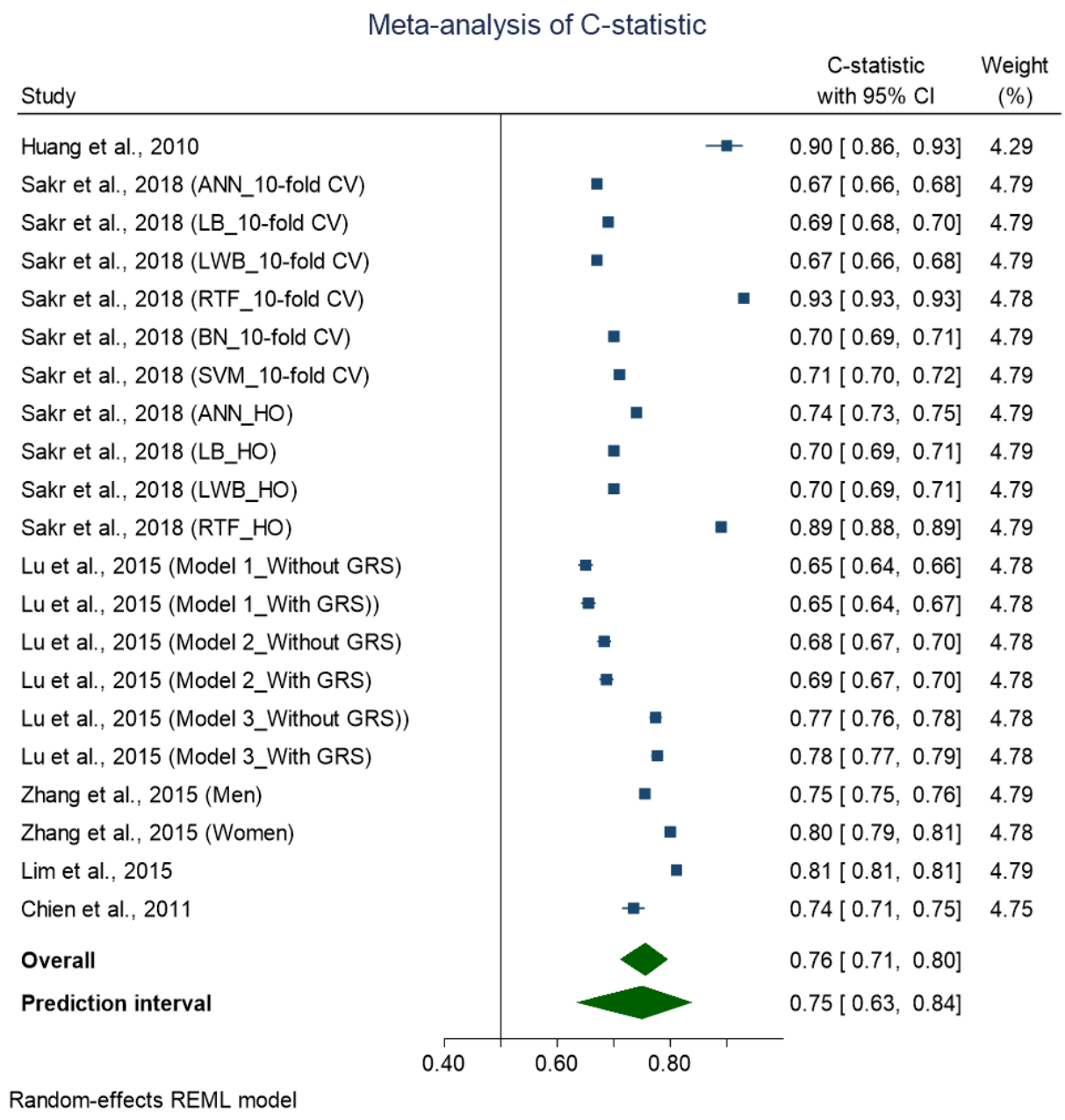

Figure S2.12 Forest plot of models primarily developed using genetic risk factors/biomarkers with a 95% prediction interval.

Table S2.1 Information about existing hypertension prediction models developed using biomarkers (genetic risk score) from the selected studies

Study	Location Model Developed/ Ethnicity	Study Design	Age	Gender	Risk Factors Included	Events (n)/Total participants (N)	Definition of Outcome Predicted/Hypertension	Duration of Follow-up	Modeling Method	Discrimination	Calibration	Model Validation: internal or external
Yamakado et al.[63] 2015	Japan/Asians	Prospective cohort	≥ 20 years	Both male and female	PFAA Index 1: Leucine, alanine, tyrosine, asparagine, tryptophan, and glycine PFAA Index 2: Isoleucine, alanine, tyrosine, phenylalanine, methionine, and histidine	424/2637	SBP ≥ 140 mm Hg or DBP ≥ 90 mm Hg or use of antihypertensive medication	4 years	Logistic regression	NR	NR	Internal, leave-one-out cross-validation (LOOCV) and External, the independent validation dataset
Qi et al.[64] 2014	China/Asians	Case-control	Case cohort: 64.48 ± 8.53 years; Control: 64.23 ± 10.13 years	Both male and female	rs17030613, rs16849225, rs1173766, rs11066280, rs35444, rs880315, rs16998073, rs11191548, rs17249754	Patients: NR/1009, Controls = NR/756	SBP ≥1 40 mm Hg or DBP ≥ 90 mm Hg or use of antihypertensive medication	NR	Logistic regression	NR	NR	NR

Lu et al.[65] 2015	China/Asians	Prospective cohort	35-74 years	Both male and female	Model1: GRS+ (Age, sex, and BMI) Model2: GRS +Model 1+smoking, drinking, pulse rate, and education Model3: GRS+ Model2 + SBP and DBP	2559/7724	SBP ≥ 140 mm Hg or DBP ≥ 90 mm Hg or use of antihypertensive medication	Mean 7.9 years	Logistic regression and Cox proportional- hazards regression	Model1: C-statistic =0.650 [0.637-0.663] (without GRS), 0.655 [0.642-0.668] (with GRS) Model 2: C-statistic = 0.683 [0.670-0.695] (without GRS), 0.687 [0.675-0.700] (with GRS) Model 3: C-statistic = 0.774 [0.763-0.785] (without GRS), 0.777 [0.766-0.787] (with GRS)	NR	NR

Zhang et al.[66] 2015	China/Asians	Prospective cohort	18-88 years	Both male and female	Five latent factors extracted from 11 biomarkers (BMI, SBP, DBP, FBG, TG, HDL-C, Hb, HCT, WBC, LC, NGC): inflammatory factor, blood viscidity factor, insulin resistance factor, blood pressure factor, and lipid resistance factor, and age	3793/17,471	SBP ≥140 mm Hg or DBP ≥90 mm Hg or use of antihypertensive medication	5 years	Cox proportional- hazards regression	Derivation cohort: AUC = 0.755 [0.746-0.763] (men), AUC = 0.801 [0.792-0.810] (women) Validation cohort: AUC = 0.755 [0.746-0.763] (men), AUC = 0.800 [0.791-0.810] (women)	NR	Internal, 10-fold cross-validation
Zhao et al.[76] 2008	China/Asians	Case-control	35–74 years	Both male and female	MDR Model: 4-locus model consisted of the SNP KCNMB1-rs11739136, RGS2-rs34717272, PRKG1-rs1881597, and MYLK-rs36025624; CART Model: RGS2, PRKG1, KCNMB1, and MYLK genes	Total: 4759 (2411 hypertensive and 2348 age-matched and sex-matched healthy controls)	Average SBP ≥ 150 mm Hg, an average DBP ≥ 95 mm Hg, or current use of antihypertensive medication	NR	Multifactor-dimensionality reduction (MDR) and classification and regression trees (CART)	MDR Model: Accuracy = 52.98%, cross-validation consistency = 9.7	NR	Internal, 10-fold cross-validation
Wang et al.[81] 2014	China/Asians	Case-control	Average 64.48 ± 8.53 years (cases), 64.23 ± 10.13 years	Both male and female	The best MDR model included rs5804 and BMI	1009 hypertensive patients and 756 normotensive controls	Mean SBP ≥ 140 mmHg and/or DBP ≥ 90 mmHg on two occasions and/or the current usage of	NR	Multifactor dimensionality reduction (MDR) model	The best MDR model testing accuracy = 0.6331, cross-validation consistency = 10	NR	Internal, 10-fold cross-validation

			(control)				antihypertensive drug treatment					
Zhao et al.[77] 2014	China/Asians	Case-control	Average 64.48 ± 8.53 years (cases), 64.23 ± 10.13 years (control)	Both male and female	The overall best model includes three-locus rs6749447, rs35929607, and rs3754777	1009 hypertensive patients and 756 normotensive controls	Mean SBP of at least 140 mm Hg or a mean DBP of at least 90 mm Hg or the current intake of antihypertensive drugs	NR	Multifactor dimensionality reduction (MDR) model	The best MDR model: testing accuracy of 0.7309 and a maximum cross-validation consistency of 10 ($P < 0.001$)	NR	Internal, 10-fold cross-validation
Niiranen et al.[57] 2016	Finland/Whites	Prospective cohort	≥ 30 years	Both male and female	Model 1: GRS Model 2: Model 1 + age + sex Model 3: Model 2 + smoking, diabetes, education, hypercholesterolemia, leisure-time exercise, and BMI	NR/2045	BP ≥ 140/90 mm Hg and/or antihypertensive medication	11 years	Multiple linear and logistic regression	C-index = 0.731 (Model 1)	NR	NR
Choi et al.[50] 2014	USA/Mexicans	Prospective cohort	NR	Both male and female	Age, gender, smoke, age × gender, Rs10510257 (AA), Rs10510257 (AG), Rs1047115 (GT)	NR/443	SBP >140 mm Hg, DBP >90 mm Hg, or use of antihypertensive medication	NR	Generalized estimating equations for Marginal model and logistic random effect model for Conditional model	Marginal model: AUC = 0.839 (with SNPs); Conditional model: AUC = 0.973 (with SNPs)	NR	NR

Lim et al.[51] 2015	Korean/ Asians	Prospective cohort	40-69 years	Both male and female	Traditional variables: age, gender, SBP, current smoking status, family history of hypertension, BMI, and one genetic variable (cGRS or wGRS derived from the 4 SNPs): rs995322, rs17249754, rs1378942, rs12945290	NR/5632	SBP ≥ 140 mm Hg or DBP ≥ 90 mm Hg or use of antihypertensive medication	4 years	Logistic regression	Derivation cohort: C-statistic = 0.810 [0.796–0.824] (model without wGRS, C-statistic = 0.811 [0.797–0.825] (model with wGRS) Validation cohort: Mean C-statistic = 0.811 [0.809-0.816]	HL Chi-square = 6.916 (model without wGRS), HL Chi-square = 5.711 (model with wGRS)	Internal validation, fivefold cross-validation
Chien et al.[32] 2011	Taiwan/Chinese	Prospective cohort	≥ 35 years	Both male and female	Biochemical Model: Age, gender, BMI, SBP and DBP, white blood count, fasting glucose, uric acid	1029/2506	SBP ≥ 140 mmHg or DBP ≥ 90 mmHg or reported use of BP-lowering medications	Median 6.15 years	Weibull regression	Biochemical Model: AUC = 0.735 [0.715 - 0.755] (point based), AUC = 0.74 (coefficient based)	Biochemical Model: HL Chi-square = 13.2, p = 0.11 (point based), 6.4, p = 0.60 (coefficient based)	Internal, fivefold cross-validation

CHAPTER 3. DEVELOPMENT OF A RISK PREDICTION MODEL FOR INCIDENT HYPERTENSION IN A CANADIAN COHORT USING TRADITIONAL REGRESSION-BASED MODELING APPROACH AND CONVERTING INTO A RISK SCORE FOR USE IN DAILY CLINICAL PRACTICE

3.1 Abstract

Background

Identifying high-risk individuals for targeted intervention may prevent or delay hypertension onset and may facilitate cost-effective approaches to management. We aimed to develop a hypertension risk prediction model and subsequent risk sore among the Canadian population using measures readily available in a primary care setting.

Methods

Eighteen thousand three hundred twenty-two participants aged 35-69 years without hypertension at baseline from a Canadian cohort were followed (median follow-up 5.80 years) for hypertension incidence, and 625 new hypertension cases were reported. The sample was randomly divided into derivation and validation sets at a 2:1 ratio. The model was developed in the derivation sample using a Cox proportional hazard model, and the model's performance was assessed in the validation sample. A risk score table was finally derived, incorporating regression coefficients from the Cox model.

Results

On the multivariable Cox model, age, BMI, SBP, diabetes, total physical activity time, and cardiovascular disease were identified as significant risk factors ($p < 0.05$) of hypertension incidence. The variable sex was forced to enter the final model. Some interaction terms were also identified as significant but were excluded due to their lack of incremental predictive capacity. Our model showed good discrimination (Harrel's C-statistic 0.77) and calibration (Grønnesby and Borgan test, χ^2 statistic = 8.75, p = 0.07; calibration slope 1.006) in the validation set. Points associated with each variable were created, and risk estimates for point totals at 2-, 3-, 5-, and 6-year time were derived from favoring the risk model's clinical implementation and workability.

Discussion

We developed a simple yet practical prediction model to estimate the risk of incident hypertension for the Canadian population that relies on readily available variables. This model may help clinicians and the general population assess their risks of new-onset hypertension and facilitate discussions on modifying this risk most effectively.

3.2 Introduction

Hypertension, which affects 1 in 5 Canadians[1], is a common medical condition and is a significant risk factor for several fatal diseases and mortality[2]. Hypertension prevention and blood pressure management in hypertensive patients is considered a major public health and primary care concern[3]. Current population health research integrates precision public health methodology, a more focused approach towards targeted intervention by identifying people at greater risk[4,5]. Screening people at greater risk of hypertension opens the possibility to promote individualized preventive initiatives because we will have the idea of who to target, what to target, where to target, and how to target. Evidence suggests that the risk of progression to hypertension depends on several factors. Combining these risk factors into a multivariable model for risk stratification would help identify high-risk individuals who should be targeted to prevent hypertension development[6,7]. Consequently, hypertension risk assessment becomes a vital mainstay for preventive efforts within the precision medicine approach.

A risk prediction model is a statistical tool for estimating the probability that a currently healthy individual with specific risk factors will develop a future condition within a particular time[8]. Over the past decades, many prediction models have been developed in different populations to predict incident hypertension[9–16], but their performance in accurately forecasting incident hypertension varies. Each model has its inherent strengths and weaknesses based on the underlying population characteristics and data from which they were derived. Prediction models cannot be directly transported from one type of population to another[17–19]. Often, models developed in one population show poor performance when applied to a different population due to differences in case-mix[17].

Prediction models for the risk of incident hypertension that directly address the Canadian population have not yet been established to the best of our knowledge. To fill this study gap, we intend to create and internally validate a simple and practical risk prediction model for incident hypertension in the Canadian general adult population. We also derived the point-based risk score from the developed model to facilitate clinical practice use for decision-making.

3.3 Methods

3.3.1 Study population

The study subjects are from Alberta's Tomorrow Project (ATP) cohort data. ATP is a province-wide prospective cohort study and consists of Alberta's residents, aged 35-69 years, without any history of cancer, other than non-melanoma skin cancer[20]. ATP is a part of a pan-Canadian initiative to investigate the causes and prevention of cancer and chronic diseases. Launched in 2000, ATP is Alberta's largest longitudinal population health cohort from the general population. It contains baseline and longitudinal information on socio-demographic characteristics, personal and family history of the disease, medication use, lifestyle and health behavior, environmental exposures, and physical measures. ATP joined the Canadian Partnership for Tomorrow Project (CPTP) in 2008[21]. ATP had three baseline questionnaires: Canadian Diet History Questionnaire-I (CDHQ-I), Health and Lifestyle Questionnaire (HLQ), and the Past-Year Total Physical Activity Questionnaire (PYTPAQ), and two follow-up questionnaires: Survey 2004 and Survey 2008, during the period 2001-2008. When ATP merged with CPTP, participants were asked to complete two versions of questionnaires: The Updated Health and Lifestyle Questionnaire (UHLQ), along with the Physical Activity and Nutrition Survey (PANS) or the CORE questionnaire[22]. As both questionnaires contained very similar information, participants completed

either UHLQ/PANS or CORE. UHLQ/PANS or CORE questionnaires were more elaborate and captured more information about the participants than the other questionnaires.

The recruitment of participants in ATP was done in two phases[23]. In Phase I (2000-08), participants were recruited using a two-stage telephone-based random digit dialing method[20]. Eight waves of telephone-based random digit dialing (RDD) using Alberta's regional health authority boundaries as the sampling frame was used to recruit participants[20]. Participants were identified using a 2-stage method. In the first stage, a household was identified, and in the second stage, one or two eligible adults within the identified household were selected for participation[20]. Participants selected a second time from the same household were excluded to avoid repetition[23]. In Phase I, 29,878 participants were recruited with a response rate of 49%[23].

In Phase II (2009-15), when ATP joined with the Canadian Partnership for Tomorrow Project (CPTP)–an alliance of five cohorts across Canada (British Columbia, Alberta, Ontario, Quebec, and Atlantic Canada), ATP-CPTP recruitment began using a volunteer sampling method[23]. Existing ATP participants (Phase I participants) were invited to join CPTP and requested to visit study centers for physical measurements and blood and urine contributions[23]. Fifteen thousand one hundred sixty-two participants from Phase I (approximately 50%) agreed to join CPTP, of which about 60% visited Study Centres[23]. Due to ATP's pledge to enroll roughly 40,000 participants to CPTP from Alberta, more participants were recruited. Nevertheless, the process for selecting potential participants in CPTP varies between jurisdictions. It includes a random selection from population-based data, purchase of mailing lists for specific geographic areas, RDD, and word of mouth[24]. Telephone-based RDD was initially used to recruit new ATP-CPTP participants in 2009 but was soon replaced by volunteer sampling due to the low response rate and increasing cost[23]. To promote volunteer recruitment, further communication and

advocating strategies were employed, such as marketing, advertising, media coverage, information booths at community events, corporate presentations, Ambassador Program, and articles[23]. In Phase II, 22,932 participants were recruited through volunteer sampling.

An invitation package was sent to the eligible participants (in both phases) that includes a cover letter, a study information booklet, an explicit consent to participate in the ATP and allow data linkage, and a self-administered ATP questionnaire[23]. Those who completed the ATP questionnaire and agreed to data linkage were considered as ATP participants. By March 2015, 52,810 Alberta residents had signed up for the ATP and decided to have their data linked to healthcare databases, with 38,094 of them agreeing to participate in the CPTP as well[23]. Of the total 52,810 ATP participants, 29,878 completed HLQ, 25,955 completed CDHQ, 25,889 completed PYTPAQ, 8,540 completed Survey 2004, 20,107 completed Survey 2008, 12,395 completed UHLQ, 12,402 completed PNAS and 25,677 completed the CORE questionnaire[23].

ATP was built to represent Alberta's general population with no history of cancer other than non-melanoma skin cancer. To see how different ATP participants are from the rest of the Alberta population and compare their characteristics, a study was conducted where the ATP cohort was compared with the Alberta-specific subsets of Canadian Community Health Survey (CCHS) participants in the same age group (35-69 years). Corresponding to the two ATP recruitment phases, two different cycles of CCHS were used to make the comparison fair. ATP participants were older, had more women, were more likely to be obese and less likely to smoke, ate more fruits and vegetables, and were more physically active than CCHS participants[23].

For this study, we used data from the CORE questionnaire. Our study cohort consists of 25,359 participants who consented to have their data linked with Alberta's administrative health data. Linking with administrative health data was primarily done due to lack of follow-up data in

ATP when accessed, necessary to determine hypertension incidence. A detailed description of data linkage is provided in Appendix 1. We excluded 6,996 participants from the analysis who had hypertension at baseline and consequently did not meet eligibility criteria (free of hypertension at baseline). We also excluded 41 participants who responded to hypertension status questions at baseline as "don't know" or "missing". Eighteen thousand three hundred twenty-two participants remained after exclusion and were finally included in the analysis. This study's ethics was approved by the Conjoint Health Research Ethics Board (CHREB) at the University of Calgary.

3.3.2 Selection of candidate variables

Before commencing the analysis, we compiled a list of available potential candidate variables. We determine the possible candidate variables for inclusion in model development based on a literature search, variables that have been used in the past, and discussion with content experts. For this study, we considered 29 candidate variables for inclusion in the model. Given our model's intended clinical application, we deliberately did not consider any genetic risk factors/biomarkers as potential candidate variables. Inclusion of the genetic risk factors in the model can reduce the model's usability due to a lack of readily available information.

3.3.3 Definition of variables

The outcome incident hypertension was determined from linked administrative health data using a coding algorithm. We used the relevant ICD-9 and ICD-10 codes (ICD-9-CM codes: 401.x, 402.x, 403.x, 404.x, and 405.x; ICD-10-CA/CCI codes: I10.x, I11.x, I12.x, I13.x, and I15.x) and a validated hypertension case definition (two physician claims within two years or one hospital discharge for hypertension) to define hypertension incidence[25].

The study participants' age was categorized into four groups: 35 to less than 45, 45 to less than 55, 55 to less than 65, and greater than or equal to 65 years. Body mass index (BMI) was

classified into four groups: underweight (< 18.5 kg/m2), normal (18.5 – 24.99 kg/m2), overweight (25.0 – 29.99 kg/m2), and obese (≥ 30.0 kg/m2). Waist circumference was classified as normal (≤ 102 cm for male and ≤ 88 cm for female) and substantially increased risk of metabolic complications (> 102 cm for male and > 88 cm for female) groups. The waist-hip ratio was categorized as normal (< 0.9 for male and < 0.85 for female) and abdominal obesity (≥ 0.9 for male and ≥ 0.85 for female). BMI waist ratio was categorized into four quartiles. Body fat percentage (BFP) was categorized as normal (< 25.0 for male and < 35.0 for female) and obese (≥ 25.0 for male and ≥ 35.0 for female). Diastolic blood pressure (DBP) was categorized into three groups: < 80 mm Hg, 80 – 89 mm Hg, and ≥ 90 mm Hg. Systolic blood pressure (SBP) was categorized into four groups: <120 mm Hg, 120 – 129 mm Hg, 130 – 139 mm Hg, and ≥ 140 mm Hg. Marital status was categorized into three groups: married and/or living with a partner, single who never married, and others (divorced, widowed, separated). Total household income was categorized into four groups: < $49,999, $50,000 - $99,999, $100,000 - $199,999, and ≥ $200,000. The highest education level completed was categorized into three groups: high school or below (none, elementary school, high school, trade, technical or vocational school, apprenticeship training or technical CEGEP), diploma but below bachelor's degree (diploma from a community college, pre-university CEGEP or non-university certificate, university certificate below bachelor's level), and bachelor's degree or above (bachelor's degree, graduate degree (MSc, MBA, MD, PhD, etc.). Ethnicity was categorized into six groups: Aboriginal, Asian (South Asian, East Asian, Southeast Asian, Filipino, West Asian, Arab), White, Latin American Hispanic, Black, and other (Jewish and others). Diabetes was categorized as "yes" or "no" based on the response to the question "Has a doctor ever told you that you had diabetes?". Cardiovascular disease was categorized as "yes" if any stroke, myocardial infarction, angina, arrhythmia, coronary heart

disease, coronary artery disease, heart disease, and heart failure was present and as ‘no” if absent. Depression was categorized as “yes” or “no” based on the response to the question “Has a doctor ever told you that you had depression?”. Family history of hypertension was categorized as “yes” if any first-degree relative is diagnosed with hypertension, otherwise “no”. Smoking status was categorized as: never, former, and current. Ever smoked was categorized as “yes” or “no” based on the response of the question “Have you smoked at least 100 cigarettes in your life?”. Alcohol consumption was categorized into five groups: never, ≤ 1 time a week, 2 to 3 times a week, 4 to 5 times a week, and ≥ 6 times a week. Working status was categorized into four groups: full-time, part-time, other (looking after a home, disable/sick, student, unpaid/voluntary), and unemployed. Total sleep time was categorized into four groups: ≤ 5 hours (short sleep duration), 6 to 7 hours, 8 hours, and ≥ 9 hours (long sleep duration). Total physical activity time was categorized as: light (< 450 MET minutes/week), moderate (450 – 900 MET minutes/week), and vigorous (> 900 MET minutes/week). Total sitting time was derived as the sum of the sitting times on weekdays and weekends. Physical activity was categorized as: low (first quartile of physical activity time and fourth quartile of sitting time), moderate (second and third quartile of physical activity time and sitting time), and high (fourth quartile of physical activity and first quartile of sitting time). Vegetable and fruit consumption was categorized as low (less than 5 servings of vegetable and fruit), moderate (less than 5 servings of vegetable but more than 5 servings of fruit OR more than 5 servings of vegetable but less than 5 servings of fruits), and high (5 or more servings of vegetable and fruit). Job schedule was categorized as regular daytime shift and other (evening shift, night shift, rotating shift, split shift, irregular shift, or on-call).

3.3.4 **Missing values**

Our dataset has missing values on several candidate variables ranging from 0 to 26%. Information on missing values for different candidate variables is presented in the supplementary table (Table S3.1). We used multiple imputation techniques to impute the missing values due to their several advantages[26]. Multiple imputation is considered the soundest strategy for handling missing data. This technique predicts the missing values by utilizing the existing information from other available variables[27] and then substitute the missing values with the predicted values to create a complete dataset. An assumption associated with multiple imputations needs to satisfy before applying multiple imputations. Missing at random (MAR) "when the probability that the responses are missing depends on the set of observed responses but is not related to the specific missing values that are expected to be obtained"[26] assumption is assessed before applying multiple imputations in our study. Multiple imputation by chained equations (MICE) was used to impute the missing values using Stata's "ice" command[28].

3.3.5 Statistical analysis

At first, we imputed the missing values using multiple imputation. However, before imputing the missing values, the required assumption (MAR) for performing multiple imputation was checked. We compared the study characteristics of those with missing with those without missing information using appropriate tests (unpaired t-test or the $\chi 2$-test). Continuous variables were expressed as the mean (SE), and categorical variables were expressed as numbers (percentage of the total). We randomly split subjects into two sets: the derivation set, which included 67% (two-thirds) of the sample (n = 12,233), and the validation set, which included the remaining 33% (one-third) (n = 6,089). The two groups' baseline characteristics were compared using the unpaired t-test or the $\chi 2$-test, as appropriate. We developed a risk prediction model from the derivation data using the multivariable Cox proportional hazards model and assessed the goodness of fit using the

validation data. Continuous variables remained continuous in the model developed and categorized only for deriving risk scores.

Collinearity among the variables was tested using the variance inflation factor (VIF) with a threshold of 2.5[29]. From the list of candidate variables, those that were highly correlated were excluded based on VIF before applying the model.

In the derivation set, the univariate Cox proportional hazards model was applied first to screen the variables for a significant association ($p < 0.20$) with hypertension incidence. Variables identified as significant in univariate association were later put into a multivariable Cox proportional hazards model to determine ultimate significant risk factors ($p < 0.05$) of incident hypertension. In the multivariable model, age, sex, body mass index, SBP, diabetes, CVD, total physical activity time, depression, waist-hip ratio, residence, highest education level completed, working status, total household income, family history of hypertension, smoking status, total sleep time, vegetable and fruit consumption, and job schedule were used as explanatory variables. The following interaction terms were also tested in the model with significant variables identified in the multivariable Cox model: age by BMI, age by SBP, age by diabetes, age by CVD, age by total physical activity time, age by sex, BMI by sex, SBP by sex, diabetes by sex, CVD by sex, and total physical activity time by sex. During the model development process, proportional hazard assumption associated with the Cox model was also tested. There are several methods for verifying proportionality assumption, and we tested the proportionality assumption by using the Schoenfeld and scaled Schoenfeld residuals. We tested the proportionality of the model as a whole and proportionality for each predictor. We also obtained the graph of the scaled Schoenfeld assumption. A non-significant p-value (> 0.05) or a horizontal line in the graph indicates no violation of the proportionality assumption.

The following general equation was used to calculate the risk of incident hypertension within time t:

$$Probability = 1 - S_0(t)^{\exp\left(\sum_{i=1}^{p}\beta_i X_i - \sum_{i=1}^{p}\beta_i \overline{X}_i\right)} \quad (1)$$

Where $S_0(t)$ is the baseline survival function, assuming all variables are represented by average values at follow-up time t; β_i is the estimated regression coefficient of the ith variable; X_i is the value of the ith variable; $\overline{X}_i$ is the corresponding mean, and p denotes the number of variables.

In the validation data, the model's predictive performance was assessed. Model discrimination was evaluated using Harrell's C-statistic[30]. Harrel's C-statistic indicates the proportion of all pairs of subjects that can be ordered such that the subject who survived longer will have the higher predicted survival time than the subjects who survived shorter, assuming that these subject pairs are selected at random. Calibration was assessed using the Grønnesby and Borgan (GB) test[31]. The GB test is an overall goodness-of-fit test for the Cox proportional hazards model and is based on martingale residuals. In the GB test, the observations are divided into K groups according to their estimated risk score, an approach similar to Hosmer and Lemeshow goodness-of-fit for logistic regression[32]. Brier score was calculated at different time points, and a calibration plot was also used for assessing calibration. In a calibration plot, expected probabilities (predicted probabilities from the model) are plotted against observed outcome probabilities (calculated by Kaplan-Meier estimates). Arjas like plots were used for assessing goodness of fit graphically[33]. We have also produced a histogram of the prognostic index (a linear predictor of the Cox model) to show the prognostic index distribution in the derivation and validation data set. The histogram will demonstrate the log relative hazard's general level and indicate the spread and outliers[34]. We also assessed calibration using the approach proposed by Royston P[35], where observed (Kaplan–Meier) and predicted survival probabilities compared in some prognostic

groups derived by placing cut points on the prognostic index. We defined three risk groups (good, intermediate, and poor) from the 25th and 75th centiles of the prognostic index in the derivation dataset based on events.

The predicted probability calculated by the model needs to be presented in a simplified way so that it can be easily used in clinical practice. The mathematical form of prediction models is relatively complex, and the computations using the prediction model can be tedious[36]. The points scoring system simplifies the tedious calculation of prediction models by assigning integer points to a given risk factor so that clinicians can easily approximate risk by summing integer points based on each risk factor's presence/absence. The points scoring system is generally formulated around categories[36]. We constructed the risk score utilizing the regression coefficients of our Cox model according to the method proposed by Sullivan et al. [36]. To facilitate the calculation of risk score, continuous variables considered in the model development were divided into categories as discussed before.

All statistical tests were two-sided. All statistical analyses were performed using Stata (Version 15.1; Stata Corporation, College Station, Texas 77845, USA).

3.4 Results

Baseline characteristics of the study participants are presented in Table 3.1 and Table 3.2. In Table 3.1, the study participants' characteristics are compared between the derivation sample and validation sample, while in Table 3.2, characteristics are compared according to the status of developing hypertension. Table 3.1 shows no significant difference ($p < 0.05$) in study characteristics between the derivation sample and validation sample except BMI waist ratio. Two quartiles (quartile 1, $p = 0.009$ and quartile 4, $p = 0.046$) in BMI waist ratio showed a significant difference between the derivation sample and validation sample. During the median 5.8-year

follow-up, 625 (3.41%) participants newly developed hypertension. In Table 3.2, most of the study characteristics were significantly different between those who developed hypertension and those who did not. Some study characteristics, however, were not significantly different and this includes first three quartiles of BMI waist ratio ($p = 0.485$, $p = 0.433$, and $p = 0.118$ respectively), marital status ($p = 0.146$), residence ($p = 0.146$), ethnicity ($p = 0.349$), depression ($p = 0.179$), family history of hypertension ($p = 0.061$), alcohol consumption ($p = 0.189$), total physical activity time ($p = 0.825$), and physical activity ($p = 0.707$). Overall, the study participants' mean age was 50.99 years, and participation of females (68.55%) in the studies were higher than the males (31.45%).

From the list of candidate variables, six (ever smoked, hip circumference, body fat percentage, BMI waist ratio, waist circumference, diastolic blood pressure.) were excluded from the model building due to their high collinearity (threshold VIF > 2.5) with other variables. Comparing the study characteristics between the missing and those are imputed is presented in the supplementary table (Table S3.2).

In the derivation sample, most of the candidate variables used in our study were identified as significant ($p < 0.20$) risk factors of incident hypertension according to the univariate Cox proportional hazard model (Table 3.3). Variables not significantly associated with incident hypertension included total sitting time, ethnicity, physical activity, alcohol consumption, and marital status and were excluded from the multivariable model. The multivariable Cox model indicated that age, BMI, SBP, diabetes, total physical activity time, and cardiovascular disease were independent significant ($p < 0.05$) risk factors of incident hypertension (Table 3.3). We forced sex into the model, considering its clinical importance. The inclusion of sex in the final model changed some of the variables' significance levels, but we deliberately overlooked it. When the interaction terms were included in the model with the variables in the multivariable Cox model,

age by sex, age by BMI, age by SBP, age by total physical activity time, sex by SBP, and sex by CVD showed significant association with incident hypertension (Table 3.4). However, the inclusion of these interaction terms did not improve the models' discriminative performance. The models with and without interaction terms were virtually identical regarding their Harrel's C-statistics value (0.77 and 0.77, respectively) and statistical significance (p = 0.64). Consequently, the interaction terms were excluded from the finally selected model. The model with only main effects was used in subsequent analyses to construct a simpler and more user-friendly risk estimation equation and risk score. A global test for Cox proportional hazards assumption indicated no violation of assumptions (p = 0.72) (Supplementary Table S3.3 and Figure S3.1 – Figure S3.13). The baseline survival function at median follow-up time 5.80-years ≈ 6-years, $S_0(6)$ was (0.977). In the derivation sample, the model's discriminative performance (Harrel's C-statistic) was 0.77.

When we applied our derived model in the validation sample, the model's discriminative performance was good (Harrel's C-statistic 0.77). The results of the GB test indicated an acceptable calibration of the risk prediction model (χ^2 statistic 8.75, p = 0.07, Figure 3.1). To compare the observed and expected events in each group based on risk score, Arjas like plots are also presented (Figure 3.2). A calibration plot of our prediction model at a time of 6-years was also presented in Figure 3.3. A calibration slope of 1.006 indicates that predicted probabilities do not vary enough[37]. Figure 3.4 represents the calibration of our model in the derivation and validation datasets. The calibration of the model looks good in each dataset. The predictions in the validation dataset are good for both "Good" and "Intermediate" risk groups where survival and predicted probabilities are quite similar, except slightly higher predictions between 6- and 14-years time intervals for the "Intermediate" group. The predictions in the "Poor" group are consistent with the

survival up to year six and somewhat high later; that is, survival tends to be worse than predicted. Due to fewer validation data events, the confidence intervals tend to be wider in validation data than in the derivation data. Figure 3.5 presents the prognostic index histogram in derivation and validation data, and no obvious irregularities and outliers were detected. Brier score calculated at 4-year, 5-year, 6-year, and 7-year time points are 0.018, 0.021, 0.026, and 0.029, respectively indicating accurate predictions.

Finally, from the developed model, a simple and practical risk score was created to calculate the risk of incident hypertension at different times (2-year, 3-year, 5-year, and 6-year) Table 3.5. The constant for the points system or the number of regression units that will correspond to one point was set as the risk associated with a 5-year increase in age. To score a continuous variable, the range of possible values of the variable was divided into appropriate categories to enable the allocation of points to the selected categories. To determine the reference values for the open-ended categories (e.g., < or >), we used the 1st percentile and the 99th percentile of that variable to minimize the influence of extreme values. The points were initially computed as a decimal value, but later rounded to the nearest integer for facile calculation. The approximate risk of incident hypertension was then estimated via summation of the points awarded to each of the items. We attach the risks associated with each point total using the Cox regression equation (Table 3.6). Finally, we created risk categories according to the total points. In our model, the maximum total point is 40, and the minimum is -2. For simple interpretation in a clinical setting, we categorize estimated risk into three categories and presented in Table 3.7.

3.4.1 Case Study

A 50-year-old male with BMI 28.5, SBP 135, diabetic, no CVD, and moderate physical activity (850 MET minutes/week).

Risk Factor	Value	Points
Age	50	2
Sex	Male	0
BMI	28.5	3
SBP	135	10
Diabetes status	Yes	4
CVD status	No	0
Physical activity	Moderate (850 MET minutes/week)	-1
Point Total		**18**
The estimate of Risk (6-year)		**7.31**

The risk estimate based on our newly developed Cox model is computed as follows:

$$\sum_{i=1}^{7} \beta_i X_i = 0.02768(50) + 0.08722(0) + 0.05147(28.5) + 0.04629(135) + 0.57066(1) + 1.08710(0) - 0.00003(850) = 9.645205$$

$$\sum_{i=1}^{7} \beta_i \bar{X}_i = 0.02768(50.94) + 0.08722(0.3142) + 0.05147(26.48) + 0.04629(119.75) + 0.57066(0.041) + 1.08710(0.021) - 0.00003(3157.97) = 8.2950638$$

$$\hat{p} = 1 - S_0(t)^{\exp(\sum_{i=1}^{7} \beta_i X_i - \sum_{i=1}^{7} \beta_i \bar{X}_i)} = 1 - 0.977^{\exp(9.645205 - 8.2950638)} = 0.085$$

The points system gives a 6-year estimate of the risk of 7 percent, employing the Cox model straight gives 8 percent.

3.5 Discussion

In this large prospective cohort study, we developed a simple model to predict the risk of developing hypertension incidence in the general Canadian adult population. The variables included in our model (age, sex, SBP, BMI, diabetes, cardiovascular disease, and total physical activity time) are routinely and easily assessed in the primary-care clinical setting. Our prediction model for hypertension risk had very good discrimination and calibration for both the derivation and validation samples, suggesting that this model has good performance and may perform well when applied to a different Canadian population. Also, a risk score table was derived for clinical

implementation and workability of the developed model. Derived point-based score where points assigned to each variable is easy to administer by health care professionals and the general population and can guide clinical counseling and decision making.

The predictive performance of our model was similar to other studies. Although prediction models' performance varies considerably across studies, our recent meta-analysis on the predictive performance of hypertension risk prediction models indicates an overall pooled C-statistic of 0.75 [95% CI: 0.73 – 0.77], which justifies our model's good predictive performance. Framingham hypertension risk score[16], the most validated hypertension risk prediction model, had a C-statistic of 0.78, similar to our model. Our model's calibration was also right on several performance measures.

Most of the variables included in our final model are consistent with other previous studies (Supplementary Figure S3.14). The variable sex was not identified as a significant factor in our model, but we forced it into the model considering its clinical implication[38]. Diabetes and CVD were the two significant risk factors in our model, often excluded by many studies. Individuals who have diabetes or CVD have a higher risk of developing hypertension than those free of these conditions. Our risk prediction model aimed to identify the risk factors for hypertension in general adults but excluding people with diabetes and CVD would limit our results' generalizability. To develop a risk prediction model applicable to as many individuals as possible, we considered diabetes and CVD subjects in model building. Smoking, alcohol consumption, and family history of hypertension are common risk factors used in the past hypertension risk prediction models (Supplementary Figure S3.14). In our study, these risk factors were not identified as significant. Their inclusion in the model also did not change the model's discriminative performance (Harrel's C-statistic remains the same as 0.77). We identified total physical activity time significantly

contributes to our model. This finding is significant because exercise is considered a preventive factor for hypertension incidence supported by scientific evidence[39]. Moreover, it is a highly modifiable lifestyle factor, and physical activity changes can modify the status of hypertension incidence.

We assessed interaction effects in our model, and several of the interaction terms were identified as significant. However, inclusion of interaction terms in the model did not improve the model's predictive performance. Our focus was on generating a simple and user-friendly risk scoring algorithm avoiding complexity. As a result, the interaction terms were excluded from the model in final considerations.

To our knowledge, this is the first hypertension risk prediction model developed explicitly in a Canadian population. The model was created using a large sample size, and the estimates from our prediction models were found to be stable, as demonstrated in the internal validation. Further, consideration of many candidate variables in model building is also a strength of this study. In contrast to most studies, where models were developed in complete cases, excluding those with missing values, we imputed missing values in our study. This approach prevented information loss, maximized information utilization, and made the results robust.

Our study has several limitations. Study participants were middle-aged and elderly Canadians. Prevention strategies are likely to be more effective if the young population can be targeted. Nevertheless, our study participants' age range will likely have minimal impact on our study's generalizability, as the people diagnosed with hypertension are generally $\geq$ 35 years of age[40]. At baseline, we excluded participants with self-reported hypertension, which can potentially lead to misclassification of hypertension status. Determining hypertension status with objective blood pressure measurement rather than relying on self-reported alone could better assemble the

cohort and avoid potential misclassification. The incidence rate of hypertension in our study was relatively low compared to what is reported for the general Alberta population[41]. There can be several potential reasons for that. The characteristics of the study participants in ATP may be different from the general Alberta population. For example, female participation in ATP data was more than double the male participation (69% vs. 31%), and the hypertension incidence rate in Alberta was much lower in females than the males in study age groups[41]. A potential selection bias also may lead to a lower incidence rate of hypertension in our study. A selection bias is an error associated with recruiting study participants or factors affecting the study participation and usually occurs when selecting participants is not random[42]. The participants in ATP were mainly selected using the volunteer sampling method[23]. Those who decided to join the study (i.e., who self-select into the survey) may have a different characteristic (e.g., healthier) than the non-participants. Due to the longitudinal nature of the study, there can also be a loss of study participants during follow-up. Participants who were lost to follow-up (e.g., due to emigration out of the province) may be more likely to develop hypertension. Our study ascertained outcome hypertension from a linked administrative health data (the hospital discharge abstract or physician claims data source) due to a lack of follow-up information in ATP. There is a possibility that the outcome ascertainment was incomplete. People who did not have a healthcare encounter after cohort enrollment (e.g., did not visit a family physician/general practitioner or were not admitted to the hospital during the study period) were missed and can potentially lead to a lower hypertension incidence. Competing risks occur when individuals experience one or more outcomes that compete with the outcome of interest[43]. It hinders the observation of the event of interest or modifies the chance that this event occurs. In our context, death could be a competing risk because if a person dies, it hinders the observation of hypertension, and the person who dies may also have a higher risk of hypertension.

We did not account for competing risks in our study because the expected event (death) rate is low as the cohort was healthy and relatively young at inception with a short follow-up time. We did not include genetic risk factors or biomarkers in our model. The inclusion of genetic risk factors in the model has the potential of improving risk prediction. However, our recent meta-analysis on hypertension risk prediction models and previous studies[12] did not show any differences in discriminative performance (pooled C-statistic was 0.76 for models developed using genetic risk factors/biomarkers). In addition, the inclusion of genetic risk factors in the model may decrease the prediction model's application in routine clinical practice. Sodium intake is an important dietary factor for the risk of incident hypertension; however, in our study, sodium intake data were not available. We could not perform an external validation of our model, essential for any prediction model's generalizability. Therefore, further validation of our model in other populations, particularly in another Canadian jurisdiction, is warranted.

In conclusion, we have developed a simple yet practical prediction model to estimate the risk of incident hypertension for the Canadian population. Risk assessment tools are believed to be convenient in motivating high-risk individuals for future health problems to modify their lifestyles to decrease their risks. Once the model is validated via external validation studies, it can help identify individuals at higher risk of hypertension, increase health consciousness, motivate individuals to improve their lifestyles and prevent or delay the onset of hypertension.

3.6 References

1. Padwal RS, Bienek A, McAlister FA, Campbell NRC. Epidemiology of Hypertension in Canada: An Update. *Can J Cardiol*. 2016;32(5):687-694. doi:10.1016/j.cjca.2015.07.734
2. Bromfield S, Muntner P. High blood pressure: The leading global burden of disease risk factor and the need for worldwide prevention programs. *Curr Hypertens Rep*. 2013;15(3):134-136. doi:10.1007/s11906-013-0340-9
3. Nerenberg KA, Zarnke KB, Leung AA, et al. Hypertension Canada's 2018 Guidelines for Diagnosis, Risk Assessment, Prevention, and Treatment of Hypertension in Adults and Children. *Can J Cardiol*. Published online 2018. doi:10.1016/j.cjca.2018.02.022
4. Khoury MJ, Iademarco MF, Riley WT. Precision Public Health for the Era of Precision Medicine. *Am J Prev Med*. Published online 2016. doi:10.1016/j.amepre.2015.08.031
5. Chowdhury MZI, Turin TC. Precision health through prediction modelling: Factors to consider before implementing a prediction model in clinical practice. *J Prim Health Care*. 2020;12(1):3-9. doi:10.1071/HC19087
6. Usher-Smith JA, Silarova B, Schuit E, Moons KGM, Griffin SJ. Impact of provision of cardiovascular disease risk estimates to healthcare professionals and patients: a systematic review. *BMJ Open*. Published online 2015. doi:10.1136/bmjopen-2015-008717
7. Lopez-Gonzalez AA, Aguilo A, Frontera M, et al. Effectiveness of the Heart Age tool for improving modifiable cardiovascular risk factors in a Southern European population: A randomized trial. *Eur J Prev Cardiol*. Published online 2015. doi:10.1177/2047487313518479
8. Chowdhury MZI, Yeasmin F, Rabi DM, Ronksley PE, Turin TC. Predicting the risk of stroke among patients with type 2 diabetes: A systematic review and meta-analysis of C-

statistics. *BMJ Open*. 2019;9(8). doi:10.1136/bmjopen-2018-025579

9. Kanegae H, Oikawa T, Suzuki K, Okawara Y, Kario K. Developing and validating a new precise risk-prediction model for new-onset hypertension: The Jichi Genki hypertension prediction model (JG model). *J Clin Hypertens*. 2018;20(5):880-890. doi:10.1111/jch.13270
10. Otsuka T, Kachi Y, Takada H, et al. Development of a risk prediction model for incident hypertension in a working-age Japanese male population. *Hypertens Res*. 2015;38(6):419-425. doi:10.1038/hr.2014.159
11. Lim NK, Son KH, Lee KS, Park HY, Cho MC. Predicting the Risk of Incident Hypertension in a Korean Middle-Aged Population: Korean Genome and Epidemiology Study. *J Clin Hypertens*. 2013;15(5):344-349. doi:10.1111/jch.12080
12. Paynter NP, Cook NR, Everett BM, Sesso HD, Buring JE, Ridker PM. Prediction of Incident Hypertension Risk in Women with Currently Normal Blood Pressure. *Am J Med*. 2009;122(5):464-471. doi:10.1016/j.amjmed.2008.10.034
13. Wang B, Liu Y, Sun X, et al. Prediction model and assessment of probability of incident hypertension: the Rural Chinese Cohort Study. *J Hum Hypertens*. Published online 2020. doi:10.1038/s41371-020-0314-8
14. Kadomatsu Y, Tsukamoto M, Sasakabe T, et al. A risk score predicting new incidence of hypertension in Japan. *J Hum Hypertens*. 2019;33(10):748-755. doi:10.1038/s41371-019-0226-7
15. Chien KL, Hsu HC, Su TC, et al. Prediction models for the risk of new-onset hypertension in ethnic Chinese in Taiwan. *J Hum Hypertens*. 2011;25(5):294-303. doi:10.1038/jhh.2010.63

16. Parikh NI, Pencina MJ, Wang TJ, et al. A risk score for predicting near-term incidence of hypertension: The Framingham Heart Study. *Ann Intern Med*. Published online 2008. doi:10.7326/0003-4819-148-2-200801150-00005

17. Moons KGM, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. *Ann Intern Med*. 2015;162(1):W1-W73. doi:10.7326/M14-0698

18. Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: Validating a prognostic model. *BMJ*. Published online 2009. doi:10.1136/bmj.b605

19. Altman DG, Royston P. What do we mean by validating a prognostic model? In: *Statistics in Medicine*. ; 2000. doi:10.1002/(SICI)1097-0258(20000229)19:4<453::AID-SIM350>3.0.CO;2-5

20. Robson PJ, Solbak NM, Haig TR, et al. Design, methods and demographics from phase I of Alberta's Tomorrow Project cohort: a prospective cohort profile. *C Open*. 2016;4(3):E515-E527. doi:10.9778/cmajo.20160005

21. Summary Data Tables | Alberta's Tomorrow Project. Accessed December 15, 2020. http://myatp.ca/for-researchers/summary-data-tables

22. Survey Questions Asked - Alberta's Tomorrow Project. Accessed January 4, 2021. https://myatpresearch.ca/survey-questions/

23. Ye M, Robson PJ, Eurich DT, Vena JE, Xu JY, Johnson JA. Cohort profile: Alberta's Tomorrow Project. *Int J Epidemiol*. 2017;46(4):1097-1098l. doi:10.1093/ije/dyw256

24. Borugian MJ, Robson P, Fortier I, et al. The Canadian Partnership for Tomorrow Project: Building a pan-Canadian research platform for disease prevention. *Cmaj*. 2010;182(11):1197-1201. doi:10.1503/cmaj.091540

25. Quan H, Khan N, Hemmelgarn BR, et al. Validation of a case definition to define hypertension using administrative data. *Hypertension*. Published online 2009. doi:10.1161/HYPERTENSIONAHA.109.139279

26. Kang H. The prevention and handling of the missing data. *Korean J Anesthesiol*. 2013;64(5):402-406. doi:10.4097/kjae.2013.64.5.402

27. Sinharay S, Stern HS, Russell D. The use of multiple imputation for the analysis of missing data. *Psychol Methods*. 2001;6(3):317-329. doi:10.1037/1082-989x.6.4.317

28. Royston P, White IR. Journal of Statistical Software Multiple Imputation by Chained Equations (MICE): Implementation in Stata. *J Stat Softw*. 2011;45(4):1-20. http://www.jstatsoft.org/

29. Midi H, Sarkar SK, Rana S. Collinearity diagnostics of binary logistic regression model. *J Interdiscip Math*. 2010;13(3):253-267. doi:10.1080/09720502.2010.10700699

30. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the Yield of Medical Tests. *JAMA J Am Med Assoc*. 1982;247(18):2543-2546. doi:10.1001/jama.1982.03320430047030

31. Grønnesby JK, Borgan Ø. A Method for Checking Regression Models in Survival Analysis Based on the Risk Score. *Lifetime Data Anal*. Published online 1996. doi:10.1007/bf00127305

32. Hosmer DW, Lemeshow S. Goodness of fit tests for the multiple logistic regression model. *Commun Stat - Theory Methods*. 1980;9(10):1043-1069. doi:10.1080/03610928008827941

33. Arjas E. A graphical method for assessing goodness of fit in Cox's proportional hazards model. *J Am Stat Assoc*. 1988;83(401):204-212. doi:10.1080/01621459.1988.10478588

34. Royston P, Altman DG. External validation of a Cox prognostic model: Principles and methods. *BMC Med Res Methodol*. Published online 2013. doi:10.1186/1471-2288-13-33

35. Royston P. Tools for checking calibration of a Cox model in external validation: Prediction of population-averaged survival curves based on risk groups. *Stata J*. 2015;15(1):275-291. doi:10.1177/1536867x1501500116

36. Sullivan LM, Massaro JM, D'Agostino RB. Presentation of multivariate data for clinical use: The Framingham Study risk score functions. *Stat Med*. 2004;23(10):1631-1660. doi:10.1002/sim.1742

37. Stevens RJ, Poppe KK. Validation of clinical prediction models: what does the "calibration slope" really measure? *J Clin Epidemiol*. 2020;118:93-99. doi:10.1016/j.jclinepi.2019.09.016

38. Ramirez LA, Sullivan JC. Sex differences in hypertension: Where we have been and where we are going. *Am J Hypertens*. 2018;31(12):1247-1254. doi:10.1093/ajh/hpy148

39. Kshirsagar A V., Chiu Y lin, Bomback AS, et al. A hypertension risk score for middle-aged and older adults. *J Clin Hypertens*. 2010;12(10):800-808. doi:10.1111/j.1751-7176.2010.00343.x

40. Hajjar I, Kotchen TA. Trends in Prevalence, Awareness, Treatment, and Control of Hypertension in the United States, 1988-2000. *J Am Med Assoc*. Published online 2003. doi:10.1001/jama.290.2.199

41. Interactive Health Data Application - Display Results. Accessed March 29, 2021. http://www.ahw.gov.ab.ca/IHDA_Retrieval/selectSubCategoryParameters.do

42. Tripepi G, Jager KJ, Dekker FW, Zoccali C. Selection bias and information bias in clinical research. *Nephron - Clin Pract*. 2010;115(2). doi:10.1159/000312871

43. Noordzij M, Leffondré K, Van Stralen KJ, Zoccali C, Dekker FW, Jager KJ. When do we need competing risks methods for survival analysis in nephrology? *Nephrol Dial Transplant*. 2013;28(11):2670-2677. doi:10.1093/ndt/gft355

Figure 3.1

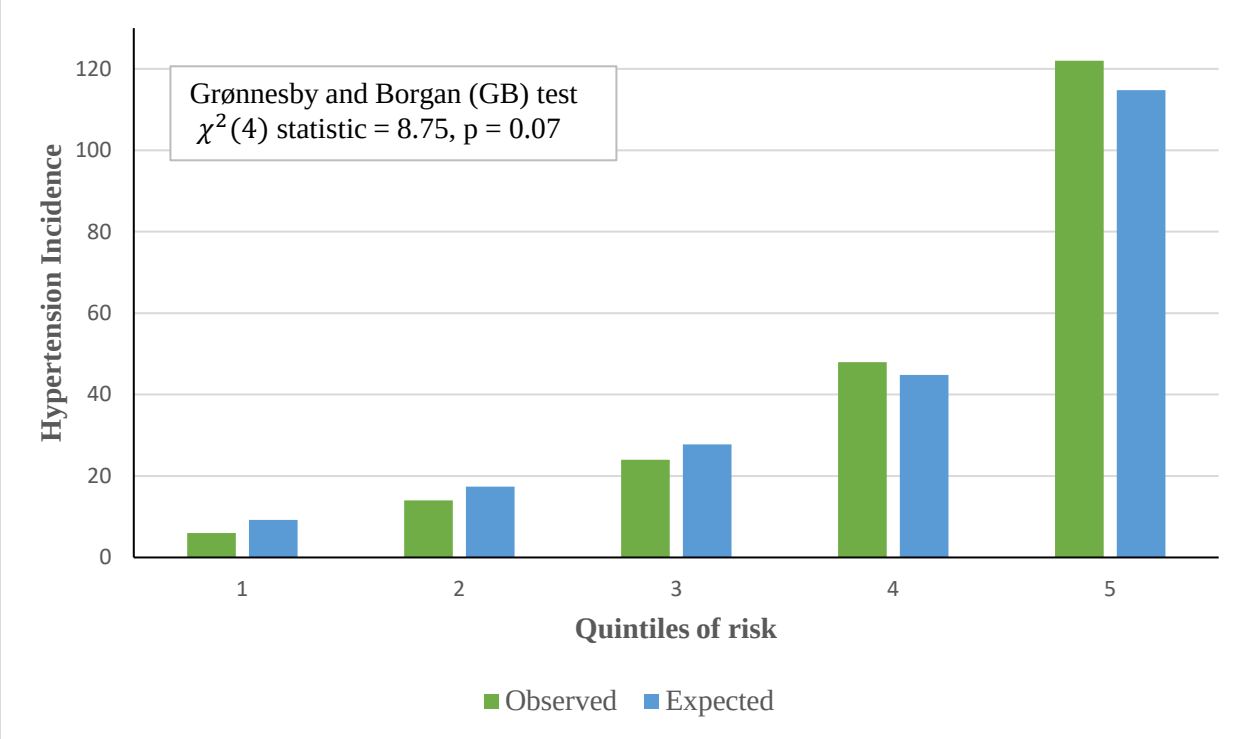

Figure 3.1 Grønnesby and Borgan (GB) goodness-of-fit test of the risk prediction model for incident hypertension in the validation sample.

Figure 3.2

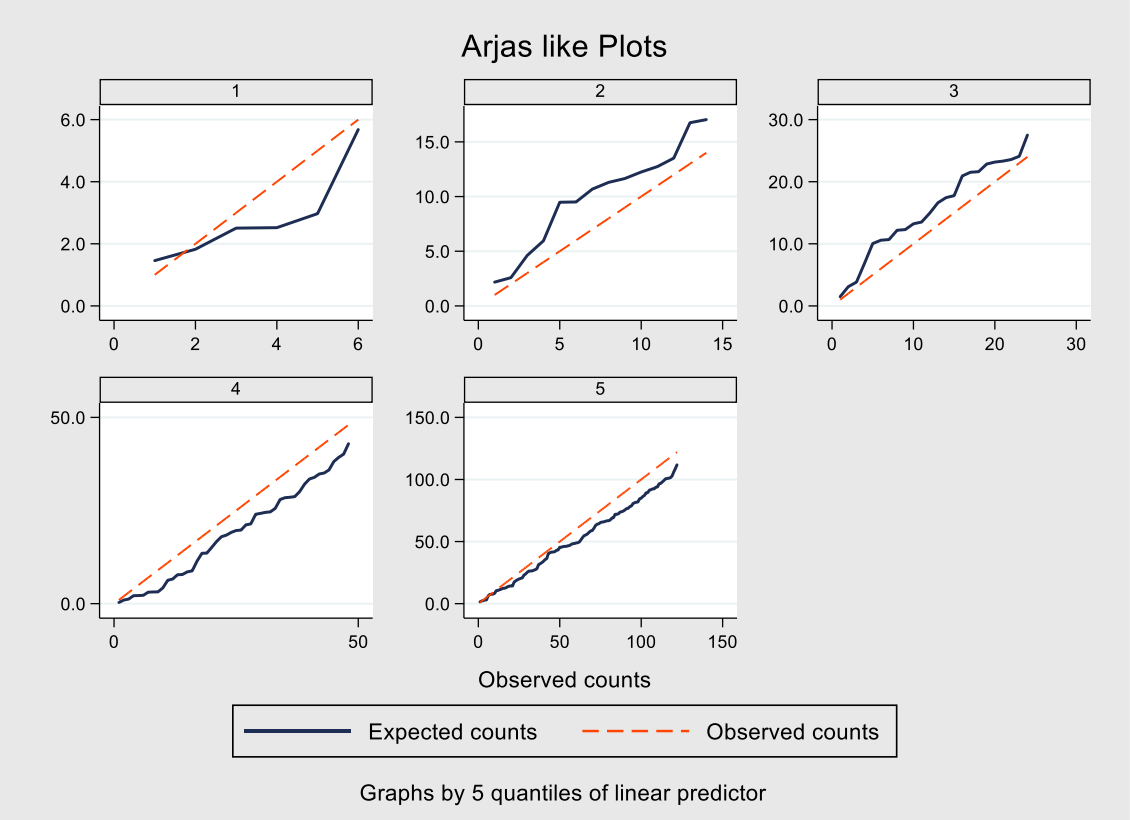

Figure 3.2 Arjas like plots to compare observed and expected events in five quantiles of the linear predictor in the validation sample.

Figure 3.3

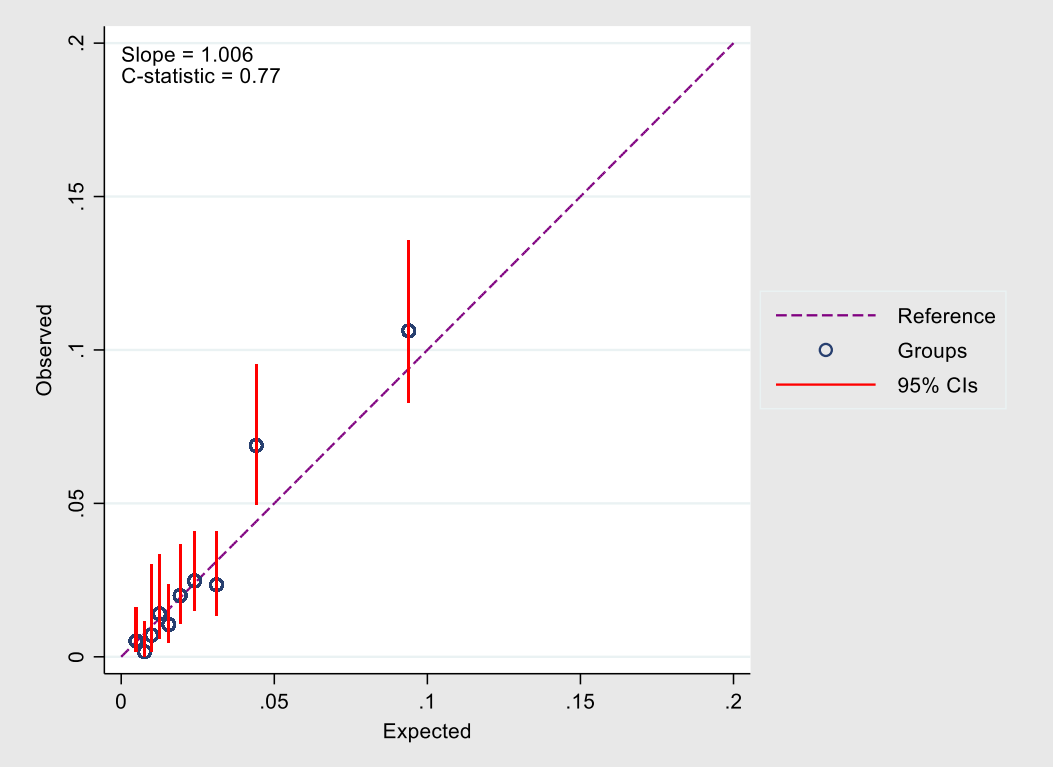

Figure 3.3 Calibration plot where expected probabilities (predicted probabilities from the model) are plotted against observed outcome probabilities (calculated by Kaplan-Meier estimates).

Figure 3.4

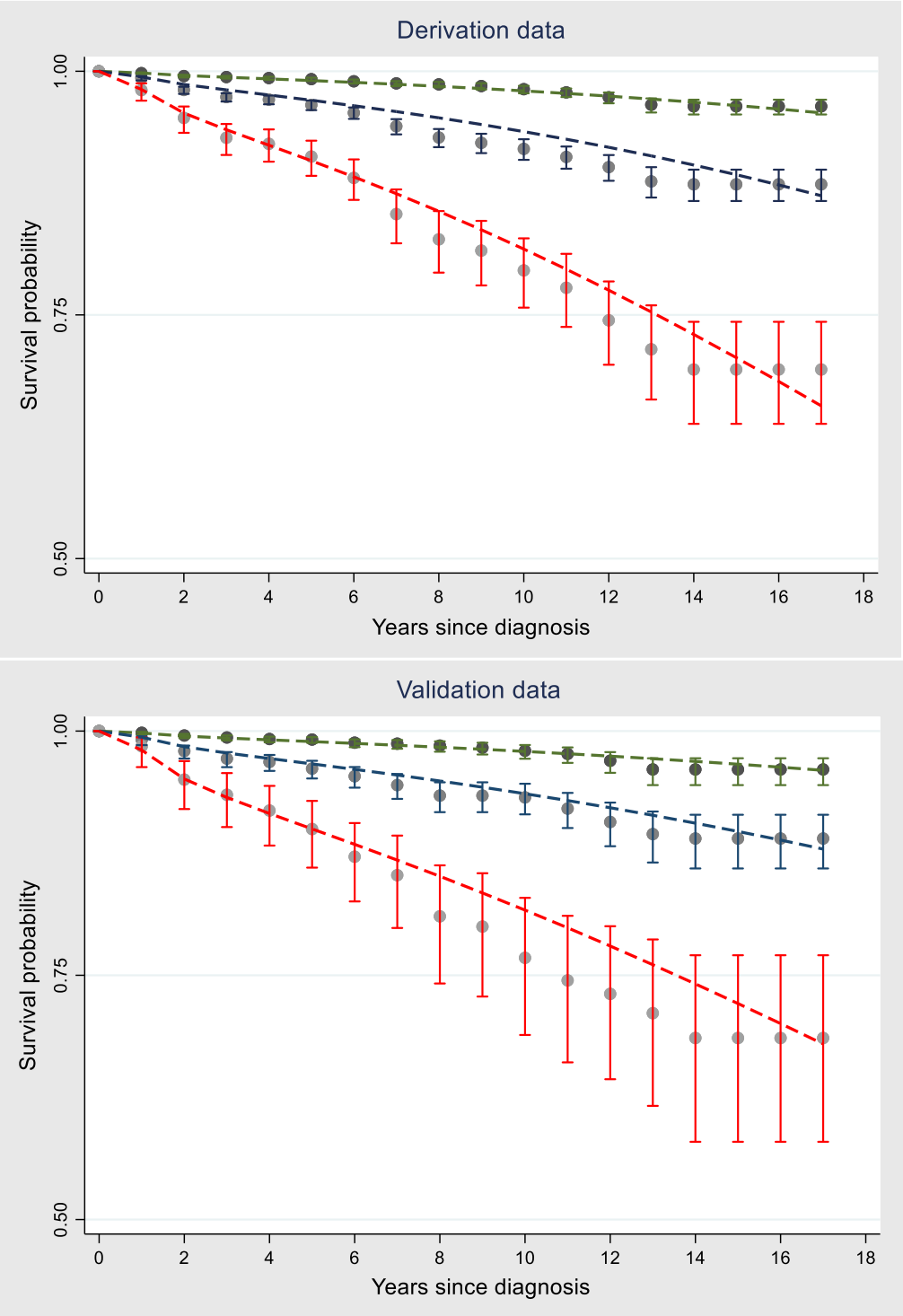

Figure 3.4 Smooth dashed lines represent predicted survival probabilities, and vertical capped lines represent Kaplan–Meier estimates with 95% confidence intervals. Three prognosis groups are plotted: the "Good" group (green lines), the "Intermediate" group (navy blue lines), and the "Poor" group (red lines).

Figure 3.5

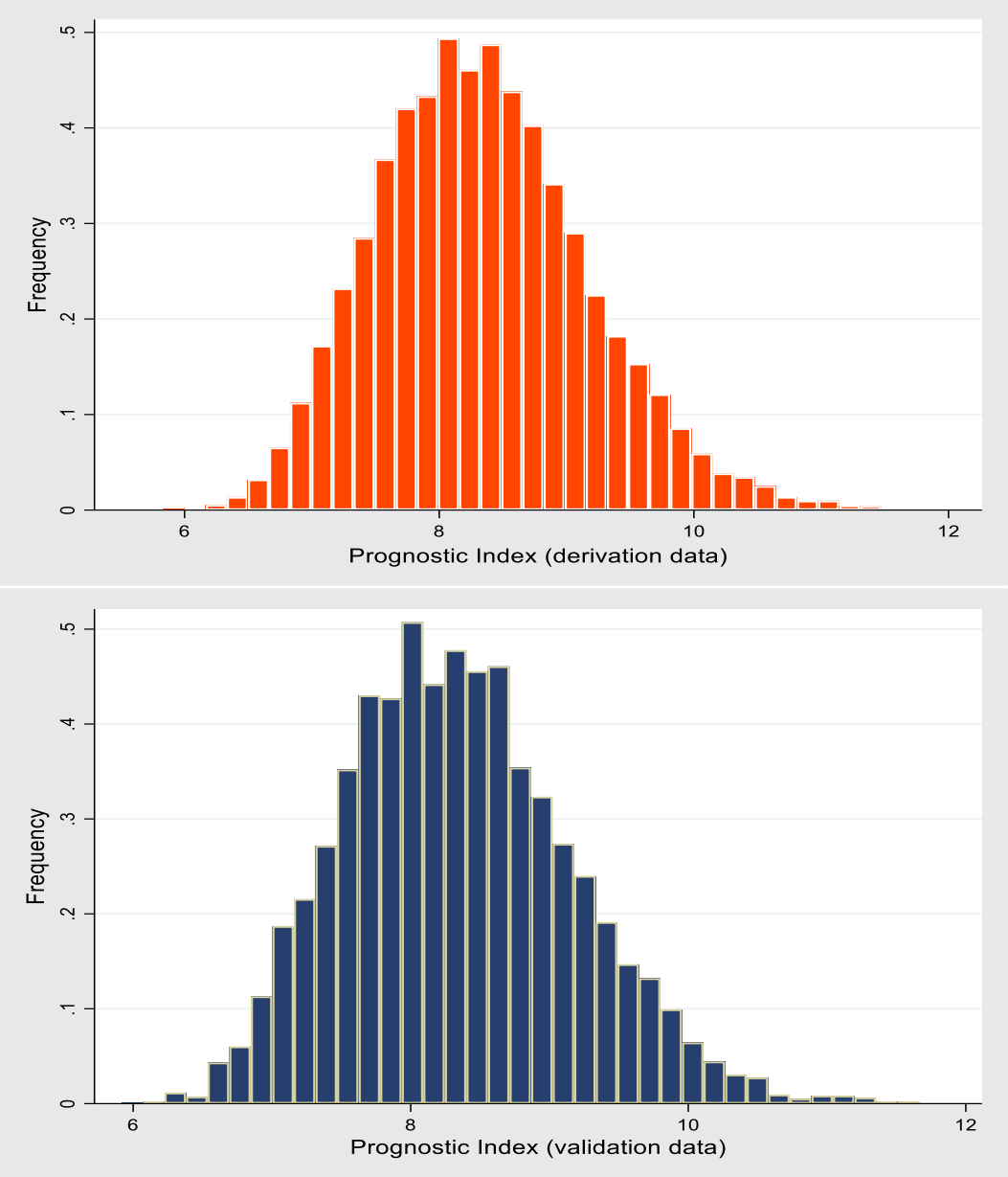

Figure 3.5 Histogram of the prognostic index in the derivation and validation datasets.

Table 3.1 Baseline characteristics of study participants and comparison in the derivation sample and validation sample

Socio-demographic characteristics of groups					
Variable	**Categories**	**All participants (18,322)**	**Derivation sample (n = 12,233)**	**Validation sample (n = 6,089)**	**P-value**
Age, years, mean (SE)		50.99 (9.20)	50.94 (9.19)	51.07 (9.24)	0.377
Age, years, n (%)	35 to less than 45	5,556 (30.32)	3,723 (30.43)	1,833 (30.10)	0.275
	45 to less than 55	6,188 (33.77)	4,169 (34.08)	2,019 (33.16)	
	55 to less than 65	5,190 (28.33)	3,410 (27.88)	1,780 (29.23)	
	≥ 65	1,388 (7.58)	931 (7.61)	457 (7.51)	
Sex, n (%)	Male	5,763 (31.45)	3,844 (31.42)	1,919 (31.52)	0.899
	Female	12,559 (68.55)	8,389 (68.58)	4,170 (68.48)	
Body Mass Index, kg/m2, mean (SE)		26.45 (4.90)	26.48 (4.94)	26.39 (4.81)	
Body Mass Index, kg/m2, n (%)	Underweight (< 18.5)	177 (0.97)	122 (1.00)	55 (0.90)	0.847
	Normal (18.5 – 24.99)	7,781 (42.47)	5,185 (42.39)	2,596 (42.63)	
	Overweight (25.0 – 29.99)	6,971 (38.05)	4,645 (37.97)	2,326 (38.20)	
	Obese (≥ 30.0)	3,393 (18.52)	2,281 (18.65)	1,112 (18.26)	
BMI Waist Ratio, mean (SE)		0.28 (0.03)	0.28 (0.03)	0.28 (0.03)	0.277
BMI Waist Ratio in Quartiles, mean (SE)	Quartile 1	0.25 (0.01)	0.25 (0.01)	0.25 (0.01)	0.009
	Quartile 2	0.27 (0.01)	0.27 (0.01)	0.27 (0.01)	0.818
	Quartile 3	0.29 (0.01)	0.29 (0 .01)	0.29 (0.01)	0.251
	Quartile 4	0.32 (0.02)	0.32 (0.02)	0.32 (0.02)	0.046
Hip Circumference, mean (SE)		104.85 (10.04)	104.91 (10.13)	104.73 (9.86)	0.250
Waist Circumference, mean (SE)		92.40 (13.18)	92.50 (13.29)	92.20 (12.95)	0.146
Waist Circumference, n (%)	Normal (≤ 102 cm for male and ≤ 88 cm for female)	10,319 (56.32)	6,854 (56.03)	3,465 (56.91)	0.260
	Substantially increased risk of metabolic complications (> 102 cm for male and > 88 cm for female)	8,003 (43.68)	5,379 (43.97)	2,624 (43.09)	
Waist Hip Ratio, mean (SE)		0.91 (0.07)	0.91 (0.07)	0.91 (0.07)	0.882
Waist Hip Ratio, n (%)	Normal (< 0.9 for male and < 0.85 for female)	4,556 (24.87)	3,056 (24.98)	1,500 (24.63)	0.609

	Abdominal obesity (≥ 0.9 for male and ≥ 0.85 for female)	13,766 (75.13)	9,177 (75.02)	4,589 (75.37)	
Body Fat Percentage, mean (SE)		31.89 (8.62)	31.93 (8.59)	31.82 (8.68)	0.411
Body Fat Percentage, n (%)	Normal (< 25.0 for male and < 35.0 for female)	9,386 (51.23)	6,258 (51.16)	3,128 (51.37)	0.784
	Obese (≥ 25.0 for male and ≥ 35.0 for female)	8,936 (48.77)	5,975 (48.84)	2,961 (48.63)	
Diastolic Blood Pressure, mean (SE)		72.95 (9.35)	72.93 (9.35)	72.97 (9.34)	0.787
Diastolic Blood Pressure, mmHg, n (%)	< 80	14,002 (76.42)	9,373 (76.62)	4,629 (76.02)	0.533
	80 – 89	3,467 (18.92)	2,287 (18.70)	1,180 (19.38)	
	≥ 90	853 (4.66)	573 (4.68)	280 (4.60)	
Systolic Blood Pressure, mean (SE)		119.81 (13.73)	119.75 (13.73)	119.92 (13.71)	0.446
Systolic Blood Pressure, mmHg, n (%)	< 120	9,561 (52.18)	6,398 (52.30)	3,163 (51.95)	0.245
	120 – 129	4,561 (24.89)	3,024 (24.72)	1,537 (25.24)	
	130 – 139	2,717 (14.83)	1,846 (15.09)	871 (14.30)	
	≥ 140	1,483 (8.09)	965 (7.89)	518 (8.51)	
Marital Status, n (%)	Married and/or living with a partner	14,458 (78.91)	9,659 (78.96)	4,799 (78.81)	0.226
	Single, never married	1,180 (6.44)	763 (6.24)	417 (6.85)	
	Other (divorced, widowed, separated)	2,684 (14.65)	1,811 (14.80)	873 (14.34)	
Residence, n (%)	Urban	15,272 (83.35)	10,180 (83.22)	5,092 (83.63)	0.484
	Rural	3,050 (16.65)	2,053 (16.78)	997 (16.37)	
Total Household Income, n (%)	< $49,999	2,855 (15.58)	1,904 (15.56)	951 (15.62)	0.416
	$50,000 - $99,999	5,889 (32.14)	3,902 (31.90)	1,987 (32.63)	
	$100,000 - $199,999	7,149 (39.02)	4,823 (39.43)	2,326 (38.20)	
	≥ $200,000	2,429 (13.26)	1,604 (13.11)	825 (13.55)	
Highest Education Level Completed, n (%)	High school or below (none, elementary school, high school, trade, technical or vocational school, apprenticeship training or technical CEGEP)	6,161 (33.63)	4,073 (33.30)	2,088 (34.29)	0.310
	Diploma but below bachelor's degree (diploma from a community college,	4,928 (26.90	3,288 (26.88)	1,640 (26.93)	

	pre-university CEGEP or non-university certificate, university certificate below bachelor's level)				
	Bachelor’s degree or above (bachelor's degree, graduate degree (MSc, MBA, MD, PhD, etc.))	7,233 (39.48)	4,872 (39.83)	2,361 (38.77)	
Ethnicity, n (%)	Aboriginal	68 (0.37)	49 (0.40)	19 (0.31)	0.316
	Asian (South Asian, East Asian, Southeast Asian, Filipino, West Asian, Arab)	827 (4.51)	545 (4.46)	282 (4.63)	
	White	16,895 (92.21)	11,274 (92.16)	5,621 (92.31)	
	Latin American Hispanic	162 (0.88)	121 (0.99)	41 (0.67)	
	Black	97 (0.53)	63 (0.52)	34 (0.56)	
	Other (Jewish and others)	273 (1.49)	181 (1.48)	92 (1.51)	
Diabetes, n (%)		735 (4.01)	502 (4.10)	233 (3.83)	0.368
Cardiovascular Disease, n (%)		377 (2.06)	257 (2.10)	120 (1.97)	0.559
Depression, n (%)		2,013 (10.99)	1,366 (11.17)	647 (10.63)	0.270
Family History of Hypertension, n (%)		10,946 (59.74)	7,266 (59.40)	3,680 (60.44)	0.176
Smoking Status, n (%)	Never	10,116 (55.21)	6,739 (55.09)	3,377 (55.46)	0.763
	Former	6,763 (36.91)	4,537 (37.09)	2,226 (36.56)	
	Current	1,443 (7.88)	957 (7.82)	486 (7.98)	
Ever Smoked, n (%)		8,206 (44.79)	5,494 (44.91)	2,712 (44.54)	0.633
Alcohol Consumption, n (%)	Never	1,293 (7.06)	869 (7.10)	424 (6.96)	0.855
	≤ 1 time a week	9,644 (52.64)	6,415 (52.44)	3,229 (53.03)	
	2 to 3 times a week	3,807 (20.78)	2,535 (20.72)	1,272 (20.89)	
	4 to 5 times a week	1,993 (10.88)	1,340 (10.95)	653 (10.72)	
	≥ 6 times a week	1,585 (8.65)	1,074 (8.78)	511 (8.39)	
Working Status, n (%)	Full time	10,281 (56.11)	6,836 (55.88)	3,445 (56.58)	0.065
	Part time	3,719 (20.30)	2,543 (20.79)	1,176 (19.31)	
	Other (looking after home, disable/sick, student, unpaid/voluntary)	3,974 (21.69)	2,614 (21.37)	1,360 (22.34)	
	Unemployed	348 (1.90)	240 (1.96)	108 (1.77)	
Total Sleep Time, n (%)	≤ 5 hours (short sleep duration)	1,191 (6.50)	804 (6.57)	387 6.36	0.257

	6 hours	3,739 (20.41)	2,441 (19.95)	1,298 (21.32)	
	7 hours (reference)	7,042 (38.43)	4,747 (38.80)	2,295 (37.69)	
	8 hours	5,111 (27.90)	3,414 (27.91)	1,697 (27.87)	
	≥ 9 hours (long sleep duration)	1,239 (6.76)	827 (6.76)	412 (6.77)	
Total Physical Activity Time, mean (SE)		3158.53 (2869.02)	3157.97 (2853.36)	3159.66 (2900.45)	0.970
Total Physical Activity Time, n (%)	Light (< 450 MET minutes/week)	1,668 (9.10)	1,096 (8.96)	572 (9.39)	0.530
	Moderate (450 – 900 MET minutes/week)	2,067 (11.28)	1,394 (11.40)	673 (11.05)	
	Vigorous (> 900 MET minutes/week)	14,587 (79.61)	9,743 (79.65)	4,844 (79.55)	
Total Sitting Time, mean (SE)		2487.77 (1174.02)	2495.39 (1176.80)	2472.48 (1168.35)	0.214
Physical Activity, n (%)	Low (first quartile of physical activity time and fourth quartile of sitting time)	1,691 (9.23)	1,157 (9.46)	534 (8.77)	0.280
	Moderate (second and third quartile of physical activity time and sitting time)	14,479 (79.03)	9,653 (78.91)	4,826 (79.26)	
	High (fourth quartile of physical activity and first quartile of sitting time)	2,152 (11.75)	1,423 (11.63)	729 (11.97)	
Vegetable and Fruit Consumption, n (%)	Low consumption (less than 5 servings of vegetable and fruit)	15,273 (83.36)	10,182 (83.23)	5,091 (83.61)	0.620
	Moderate consumption (less than 5 servings of vegetable but more than 5 servings of fruit OR more than 5 servings of vegetable but less than 5 servings of fruits	2,529 (13.80)	1,694 (13.85)	835 (13.71)	
	High consumption (5 or more servings of vegetable and fruit)	520 (2.84)	357 (2.92)	163 (2.68)	
Job Schedule, n (%)	Regular daytime shift	11,920 (65.06)	7,985 (65.27)	3,935 (64.62)	0.385

	Other (evening shift, night shift, rotating shift, split shift, irregular shift, or on call)	6,402 (34.94)	4,248 (34.73)	2,154 (35.38)	

Table 3.2 Baseline characteristics of study participants according to the status of developing hypertension or not

Socio-demographic characteristics of groups					
Variable	**Categories**	**All participants (18,322)**	**Participants who has developed hypertension (n = 625)**	**Participants who did not develop hypertension (n = 17,697)**	**P-value**
Age, years, mean (SE)		50.99 (0.07)	53.99 (0.35)	50.88 (0.07)	< 0.001
Age, years, n (%)	35 to less than 45	5556 (30.32)	107 (17.12)	5449 (30.79)	< 0.001
	45 to less than 55	6188 (33.77)	213 (34.08)	5975 (33.76)	
	55 to less than 65	5190 (28.33)	226 (36.16)	4964 (28.05)	
	≥ 65	1388 (7.58)	79 (12.64)	1309 (7.39)	
Sex, n (%)	Male	5763 (31.45)	250 (40)	5513 (31.15)	< 0.001
	Female	12559 (68.55)	375 (60)	12184 (68.85)	
Body Mass Index, kg/m2, mean (SE)		26.45 (0.04)	28.63 (0.21)	26.38 (0.04)	
Body Mass Index, kg/m2, n (%)	Underweight (< 18.5)	179 (0.97)	3 (0.48)	199 (1.12)	< 0.001
	Normal (18.5 – 24.99)	7819 (42.68)	148 (23.62)	7642 (43.18)	
	Overweight (25.0 – 29.99)	6876 (37.53)	271 (43.37)	6501 (36.73)	
	Obese (≥ 30.0)	3448 (18.82)	203 (32.53)	3355 (18.96)	
BMI Waist Ratio, mean (SE)		0.28 (0.0002)	0.2893 (0.0013)	0.2831 (0.0002)	< 0.001
BMI Waist Ratio in Quartiles, mean (SE)	Quartile 1	0.25 (0.0002)	0.25 (0.0009)	0.25 (0.0002)	0.485
	Quartile 2	0.27 (0.0001)	0.27 (0.0004)	0.27 (0.0001)	0.433
	Quartile 3	0.29 (0.0001)	0.29 (0.0005)	0.29 (0.0001)	0.118
	Quartile 4	0.32 (0.0003)	0.33 (0.0016)	0.32 (0.0003)	0.017
Hip Circumference, mean (SE)		104.85 (0.08)	108.25 (0.44)	104.78 (0.08)	< 0.001
Waist Circumference, mean (SE)		92.38 (0.10)	100.60 (0.60)	92.21 (0.10)	<0.001
Waist Circumference, n (%)	Normal (≤ 102 cm for male and ≤ 88 cm for female)	10188 (55.60)	201 (32.11)	9987 (56.43)	< 0.001
	Substantially increased risk of metabolic complications (> 102 cm for male and > 88 cm for female)	8134 (44.40)	424 (67.89)	7710 (43.57)	
Waist Hip Ratio, mean (SE)		0.9093 (0.0006)	0.9363 (0.0033)	0.9085 (0.0006)	< 0.001

Waist Hip Ratio, n (%)	Normal (< 0.9 for male and < 0.85 for female)	4466 (24.38)	101 (16.08)	4366 (24.67)	< 0.001
	Abdominal obesity (≥ 0.9 for male and ≥ 0.85 for female)	13856 (75.62)	524 (83.92)	13331 (75.33)	
Body Fat Percentage, mean (SE)		31.86 (0.07)	34.67 (0.37)	31.84 (0.07)	< 0.001
Body Fat Percentage, n (%)	Normal (< 25.0 for male and < 35.0 for female)	9425 (51.44)	179 (28.59)	9246 (52.25)	< 0.001
	Obese (≥ 25.0 for male and ≥ 35.0 for female)	8897 (48.56)	446 (71.40)	8451 (47.75)	
Diastolic Blood Pressure, mean (SE)		72.96 (0.08)	78.43 (0.47)	72.78 (0.08)	< 0.001
Diastolic Blood Pressure, mmHg, n (%)	< 80	13977 (76.28)	344 (55.05)	13633 (77.03)	< 0.001
	80 – 89	3482 (19.00)	184 (29.44)	3298 (18.63)	
	≥ 90	863 (4.71)	97 (15.51)	766 (4.33)	
Systolic Blood Pressure, mean (SE)		119.71 (0.11)	132.36 (0.67)	119.40 (0.12)	< 0.001
Systolic Blood Pressure, mmHg, n (%)	< 120	9600 (52.40)	129 (20.69)	9471 (53.52)	< 0.001
	120 – 129	4585 (25.03)	139 (22.25)	4446 (25.12)	
	130 – 139	2684 (14.65)	176 (28.23)	2508 (14.17)	
	≥ 140	1453 (7.93)	180 (28.83)	1272 (7.19)	
Marital status, n (%)	Married and/or living with a partner	14457 (78.91)	488 (78.08)	13969 (78.94)	0.146
	Single, never married	1180 (6.44)	32 (5.12)	1148 (6.49)	
	Other (divorced, widowed, separated)	2685 (14.65)	105 (16.8)	2580 (14.57)	
Residence, n (%)	Urban	15272 (83.35)	428 (68.48)	14844 (83.88)	0.146
	Rural	3050 (16.65)	197 (31.52)	2853 (16.12)	
Total Household Income, n (%)	< $49,999	2800 (15.28)	178 (28.56)	2627 (14.84)	< 0.001
	$50,000 - $99,999	5912 (32.27)	229 (36.68)	5690 (32.15)	
	$100,000 - $199,999	7174 (39.16)	177 (28.27)	6986 (39.48)	
	≥ $200,000	2436 (13.29)	41 (6.49)	2394 (13.52)	
Highest Education Level Completed, n (%)	High school or below (none, elementary school, high school, trade, technical or vocational school, apprenticeship training or technical CEGEP)	6164 (33.64)	309 (49.35)	5854 (33.08)	< 0.001

	Diploma but below bachelor's degree (diploma from a community college, pre-university CEGEP or non-university certificate, university certificate below bachelor's level)	4926 (26.89)	163 (26.15)	4764 (26.92)	
	Bachelor's degree or above (bachelor's degree, graduate degree (MSc, MBA, MD, PhD, etc.))	7232 (39.47)	153 (24.49)	7079 (40.0)	
Ethnicity, n (%)	Aboriginal	68 (0.37)	1 (0.16)	67 (0.38)	0.349
	Asian (South Asian, East Asian, Southeast Asian, Filipino, West Asian, Arab)	827 (4.51)	21 (3.4)	806 (4.55)	
	White	16894 (92.21)	588 (94.03)	16307 (92.14)	
	Latin American Hispanic	162 (0.89)	2 (0.32)	160 (0.9)	
	Black	97 (0.53)	2 (0.33)	95 (0.54)	
	Other (Jewish and others)	273 (1.49)	11 (1.76)	262 (1.48)	
Diabetes, n (%)		735 (4.01)	58 (9.28)	677 (3.83)	< 0.001
Cardiovascular Disease, n (%)		377 (2.06)	40 (6.4)	337 (1.9)	< 0.001
Depression, n (%)		2011 (10.98)	79 (12.64)	1932 (10.92)	0.179
Family History of Hypertension, n (%)		10946 (59.74)	396 (63.36)	10550 (59.61)	0.061
Smoking Status, n (%)	Never	10107 (55.16)	290 (46.37)	9823 (55.51)	< 0.001
	Former	6773 (36.97)	276 (44.15)	6491 (36.68)	
	Current	1442 (7.87)	59 (9.48)	1383 (7.81)	
Ever Smoked, n (%)		8215 (44.84)	335 (53.63)	7874 (44.49)	< 0.001
Alcohol Consumption, n (%)	Never	1279 (6.98)	56 (8.97)	1224 (6.92)	0.189
	≤ 1 time a week	9642 (52.63)	341 (54.52)	9307 (52.59)	
	2 to 3 times a week	3820 (20.85)	123 (19.77)	3689 (20.85)	
	4 to 5 times a week	1988 (10.85)	55 (8.74)	1938 (10.95)	
	≥ 6 times a week	1593 (8.69)	50 (8.0)	1539 (8.69)	
Working Status, n (%)	Full time	11449 (62.49)	352 (56.29)	11057 (62.48)	< 0.001
	Part time	4596 (25.09)	182 (29.19)	4422 (24.99)	
	Other (looking after home, disable/sick, student, unpaid/voluntary)	1857 (10.13)	83 (13.23)	1803 (10.18)	

	Unemployed	420 (2.29)	8 (1.28)	415 (2.35)	
Total Sleep Time, n (%)	≤ 5 hours (short sleep duration)	1192 (6.51)	47 (7.49)	1147 (6.48)	< 0.001
	6 hours	3732 (20.37)	127 (20.33)	3604 (20.37)	
	7 hours (reference)	7048 (38.46)	200 (32.02)	6847 (38.69)	
	8 hours	5115 (27.92)	185 (29.66)	4929 (27.85)	
	≥ 9 hours (long sleep duration)	1235 (6.74)	66 (10.49)	1170 (6.61)	
Total Physical Activity Time, mean (SE)		3159.83 (21.43)	3183.97 (126.52)	3157.58 (21.68)	0.825
Total Physical Activity Time, n (%)	Light (< 450 MET minutes/week)	1,668 (9.10)	84 (13.44)	1,584 (8.95)	0.001
	Moderate (450 – 900 MET minutes/week)	2,067 (11.28)	69 (11.04)	1,998 (11.29)	
	Vigorous (> 900 MET minutes/week)	14,587 (79.61)	472 (75.52)	14,115 (79.76)	
Total Sitting Time, mean (SE)		2488.53 (8.92)	2389.16 (49.14)	2490.98 (9.38)	0.043
Physical Activity, n (%)	Low (first quartile of physical activity time and fourth quartile of sitting time)	1685 (9.19)	59 (9.47)	1678 (9.48)	0.707
	Moderate (second and third quartile of physical activity time and sitting time)	14478 (79.02)	488 (78.12)	13957 (78.87)	
	High (fourth quartile of physical activity and first quartile of sitting time)	2159 (11.78)	78 (12.40)	2062 (11.65)	
Vegetable and Fruit Consumption, n (%)	Low consumption (less than 5 servings of vegetable and fruit)	15264 (83.31)	544 (87.05)	14721 (83.18)	0.024
	Moderate consumption (less than 5 servings of vegetable but more than 5 servings of fruit OR more than 5 servings of vegetable but less than 5 servings of fruits	2536 (13.84)	68 (10.84)	2469 (13.95)	

	High consumption (5 or more servings of vegetable and fruit)	522 (2.85)	13 (2.11)	507(2.87)	
Job Schedule, n (%)	Regular daytime shift	12866 (70.22)	385 (61.59)	12452 (70.36)	< 0.001
	Other (evening shift, night shift, rotating shift, split shift, irregular shift, or on call)	5456 (29.78)	240 (38.41)	5245 (29.64)	

Table 3.3 Unadjusted and adjusted hazard ratios for the risk factors of hypertension incidence

Unadjusted and adjusted hazard ratios and 95% confidence intervals for the risk factors of hypertension incidence							
Variable		**Unadjusted Hazard Ratio (95% CI)**	**P-value**		**Adjusted Hazard Ratio (95% CI)**	**P-value**	
Age, years		1.05 (1.03 - 1.06)	< 0.001		1.02 (1.01 - 1.03)	0.002	
Sex	Male	Reference			Reference		
	Female	0.68 (0.56 - 0.82)	< 0.001		1.01 (0.80 - 1.28)	0.923	
Body Mass Index, kg/m2		1.07 (1.06 - 1.09)	< 0.001		1.05 (1.03 - 1.07)	< 0.001	
BMI Waist Ratio,		1894.98 (93.43 - 38435.67)	< 0.001		-	-	
Hip Circumference, cm		1.03 (1.02 - 1.04)	< 0.001		-	-	
Waist Circumference, cm		1.04 (1.03 - 1.05)	< 0.001		-	-	
Waist Hip Ratio		41.81 (12.45 - 140.43)	< 0.001		0.94 (0.22 - 4.04)	0.930	
Body Fat Percentage, percentage		1.03 (1.02 - 1.04)	< 0.001		-	-	
Diastolic Blood Pressure, mmHg		1.06 (1.05 - 1.07)	< 0.001		-	-	
Systolic Blood Pressure, mmHg		1.05 (1.05 - 1.06)	< 0.001		1.05 (1.04 - 1.05)	< 0.001	
Marital Status	Married or living with a partner	Reference		0.145*	-	-	
	Single, never married	1.02 (0.66 - 1.58)	0.913		-	-	
	Other (divorced, widowed, separated)	1.29 (1.00 - 1.66)	0.050		-	-	
Residence	Urban	Reference			Reference		
	Rural	1.37 (1.11 - 1.71)	0.004		1.08 (0.86 - 1.35)	0.500	
Total Household Income,	< $49,999	Reference		< 0.001*	Reference		0.060*
	$50,000 - $99,999	0.65 (0.51 - 0.83)	0.001		0.80 (0.62 - 1.04)	0.090	
	$100,000 - $199,999	0.51 (0.39 - 0.65)	< 0.001		0.75 (0.57 - 0.99)	0.048	
	≥ $200,000	0.34 (0.22 - 0.52)	< 0.001		0.56 (0.36 - 0.88)	0.012	

Highest Education Level Completed	High school or below (none, elementary school, high school, trade, technical or vocational school, apprenticeship training or technical CEGEP)	Reference		< 0.001*	Reference		0.250*
	Diploma but below bachelor's degree (diploma from a community college, pre-university CEGEP or non-university certificate, university certificate below bachelor's level)	0.79 (0.63 - 0.99)	0.050		1.01 (0.79 - 1.28)	0.952	
	Bachelor’s degree or above (bachelor's degree, graduate degree (MSc, MBA, MD, PhD, etc.))	0.54 (0.43 - 0.69)	< 0.001		0.82 (0.63 - 1.06)	0.128	
Ethnicity	Aboriginal	0.49 (0.07 - 3.50)	0.478	0.532*	-	-	
	Asian (South Asian, East Asian, Southeast Asian, Filipino, West Asian, Arab)	1.17 (0.71 - 1.93)	0.543		-	-	
	White	Reference			-	-	
	Latin American Hispanic	0.33 (0.05 - 2.36)	0.270		-	-	
	Black	0.62 (0.09 - 4.41)	0.632		-	-	
	Other (Jewish and others)	1.61 (0.80 - 3.25)	0.182		-	-	
Diabetes	No	Reference			Reference		
	Yes	2.10 (1.48 - 2.98)	< 0.001		1.71 (1.19 - 2.46)	0.004	
Cardiovascular Disease	No	Reference			Reference		
	Yes	3.14 (2.13 - 4.64)	< 0.001		2.81 (1.89 - 4.19)	< 0.001	
Depression	No	Reference			Reference		
	Yes	1.08 (0.79 - 1.46)	0.640		0.97 (0.71 - 1.33)	0.874	
Family History of Hypertension	No	Reference			Reference		
	Yes	1.14 (0.93 - 1.39)	0.202		1.13 (0.93 - 1.39)	0.225	
Smoking Status	Never	Reference		0.031*	Reference		0.759*
	Former	1.31 (1.07 - 1.61)	0.009		1.07 (0.87 - 1.32)	0.536	
	Current	1.23 (0.87 - 1.74)	0.250		1.11 (0.78 - 1.58)	0.565	
Ever Smoked	No	Reference			-	-	
	Yes	1.29 (1.07 - 1.57)	0.009		-	-	
	Never	Reference		0.249*	-	-	

Alcohol Consumption	≤ 1 time a week	0.74 (0.53 - 1.04)	0.085		-	-	
	2 to 3 times a week	0.86 (0.59 - 1.24)	0.414		-	-	
	4 to 5 times a week	0.72 (0.47 - 1.10)	0.130		-	-	
	≥ 6 times a week	0.63 (0.40 - 1.01)	0.058		-	-	
Working Status	Full time	Reference		< 0.001*	Reference		0.294*
	Part time	0.89 (0.68 - 1.18)	0.426		0.83 (0.62 - 1.12)	0.232	
	Other (looking after home, disable/sick, student, unpaid/voluntary)	1.63 (1.32 - 2.03)	< 0.001		0.96 (0.71 - 1.30)	0.807	
	Unemployed	0.53 (0.20 - 1.41)	0.202		0.45 (0.16 - 1.23)	0.120	
Total Sleep Time, hours	≤ 5 hours (short sleep duration)	1.60 (1.11 - 2.31)	0.012	0.006*	1.03 (0.70 - 1.51)	0.882	0.178*
	6 hours	1.42 (1.08 - 1.85)	0.011		0.77 (0.53 - 1.12)	0.173	
	7 hours (reference)	Reference			Reference		
	8 hours	1.17 (0.91 - 1.51)	0.220		0.85 (0.59 - 1.24)	0.408	
	≥ 9 hours (long sleep duration)	1.70 (1.19 - 2.43)	0.003		1.07 (0.68 - 1.68)	0.781	
Total Physical Activity Time, minutes/week		0.99 (0.99 - 1.00)	0.144		0.99 (0.99993 - 0.999997)	0.033	
Total Sitting Time, minutes/week		1.00 (0.99 - 1.01)	0.660		-	-	
Physical Activity, quartiles	Low (first quartile of physical activity time and fourth quartile of sitting time)	Reference		0.738*	-	-	
	Moderate (second and third quartile of physical activity time and sitting time)	0.88 (0.64 - 1.21)	0.437		-	-	
	High (fourth quartile of physical activity and first quartile of sitting time)	0.90 (0.60 - 1.35)	0.613		-	-	
Vegetable and Fruit Consumption, servings	Low consumption (less than 5 servings of vegetable and fruit)	Reference		0.408*	Reference		0.494*
	Moderate consumption (less than 5 servings of vegetable but more than 5 servings of fruit OR more than 5	0.81 (0.59 - 1.11)	0.191		0.97 (0.70 - 1.33)	0.832	

	servings of vegetable but less than 5 servings of fruits						
	High consumption (5 or more servings of vegetable and fruit)	0.89 (0.48 - 1.67)	0.725		1.45 (0.77 - 2.74)	0.249	
Job Schedule	Regular daytime shift	Reference			Reference		
	Other (evening shift, night shift, rotating shift, split shift, irregular shift, or on call)	1.42 (1.17 - 1.73)	< 0.001		1.15 (0.91 - 1.46)	0.229	

* overall effect for categorical variables with multiple categories

Table 3.4 Regression coefficients and hazard ratio's for incident hypertension

Variable	**Simplified model without interaction terms**				**The model with interaction terms**			
	β	Standard Error (SE)	Hazard Ratio (HR)	95 % CI	β	Standard Error (SE)	Hazard Ratio (HR)	95 % CI
Age	0.02768	0.00562	1.02807	1.02-1.04	0.18825	0.05158	1.20714	1.09-1.34
Sex*	0.08722	0.10411	1.09113	0.89-1.34	- 2.75995	1.02372	0.06329	0.01-0.47
Body Mass Index (BMI)	0.05147	0.00857	1.05282	1.04-1.07	0.13194	0.04638	1.14104	1.04-1.25
Systolic Blood Pressure (SBP)	0.04629	0.00309	1.04738	1.04-1.05	0.08233	0.01898	1.08581	1.05-1.13
Diabetes	0.57066	0.18200	1.76943	1.24-2.53	0.62335	0.18262	1.86517	1.30-2.67
Cardiovascular Disease (CVD)	1.08710	0.20085	2.96566	2.00-4.39	1.43281	0.24367	4.19044	2.60-6.76
Total Physical Activity Time	- 0.00003	0.00002	0.99997	0.99-1.00	0.00024	0.00010	1.00024	1.00-1.00
Age by Sex	-	-	-	-	0.01516	0.01133	1.01527	0.99-1.04
Age by BMI	-	-	-	-	- 0.00157	0.00088	0.99843	0.99-1.00
Age by SBP	-	-	-	-	- 0.00084	0.00035	0.99916	0.99-0.99
Age by Total physical activity time	-	-	-	-	- 0.00001	0.000002	0.99999	0.99-0.99
Sex by SBP	-	-	-	-	0.01583	0.00638	1.01596	1.00-1.03
Sex by CVD	-	-	-	-	- 0.96267	0.45499	0.38187	0.16-0.93

* male is the reference category

Table 3.5 Calculation of point values for risk score

Variable	$\boldsymbol{\beta}$	**Categories**	**Reference Value** ($\boldsymbol{W}$)	$\boldsymbol{\beta\,(W\ -\ W_{REF})}$	$\boldsymbol{Points = \frac{\beta\,(W\ -\ W_{REF})}{B}}$
Age	0.02768	35 to less than 45 *	39.5 (W_{REF})	0	0
		45 to less than 55	49.5	0.2768	2
		55 to less than 65	59.5	0.5536	4
		65 to less than 75	69.5	0.8304	6
Sex	0.08722	Male *	0 (W_{REF})	0	0
		Female	1	0.0872	1
Body Mass Index ‡	0.05147	< 18.5 *	18.5 (W_{REF})	0	0
		18.5 to less than 25.0	21.75	0.1673	1
		25.0 to less than 30.0	27.5	0.4632	3
		≥ 30.0	36.35	0.9187	7
Systolic Blood Pressure †	0.04629	< 120 *	106 (W_{REF})	0	0
		120 to less than 130	125	0.8795	6
		130 to less than 140	135	1.3424	10
		≥ 140	148	1.9442	14
Diabetes	0.57066	No *	0 (W_{REF})	0	0
		Yes	1	0.5707	4
Cardiovascular Disease	1.08710	No *	0 (W_{REF})	0	0
		Yes	1	1.0871	8
Physical Activity Total**	- 0.00003	Light (< 450 MET minutes/week)	274.5 (W_{REF})	0	0
		Moderate (450 – 900 MET minutes/week)	675	- 0.0120	-1
		Vigorous (> 900 MET minutes/week)	7209	- 0.2080	-2

* Reference Category

The age range in the sample is 35 – 70.

‡ The range of body mass index is 12.5 – 64.9. To determine the reference values for the first and last categories, we use the 1st percentile (18.5) and the 99th percentile (42.7) to minimize extreme values' influence.

**The range of physical activity total is from 33 MET minutes/week to 19,278 MET minutes/week. To determine the reference values for the first and last categories, we use the 1st percentile (99) and the 99th percentile (13,518) to minimize extreme values' influence.

† The range of systolic blood pressures is 76 – 205. To determine the reference values for the first and last categories, we use the 1st percentile (92) and the 99th percentile (156) to minimize extreme values' influence.

The constant for the points system or the number of regression units will correspond to one point. Here, we let B reflect the increase in risk associated with a 5-year increase in age:

$$B = 5(0.02768) = 0.1384$$

Table 3.6 Risk estimates for point totals at 2, 3, 5, and 6-year time

2-year risk (%)		3-year risk (%)		5-year risk (%)		6-year risk (%)	
Point total	**Estimate of risk**	**Point total**	**Estimate of risk**	**Point total**	**Estimate of risk**	**Point total**	**Estimate of risk**
-2	0.27	-2	0.30	-2	0.39	-2	0.48
-1	0.31	-1	0.35	-1	0.45	-1	0.55
0	0.35	0	0.40	0	0.52	0	0.63
1	0.40	1	0.46	1	0.60	1	0.72
2	0.46	2	0.53	2	0.68	2	0.83
3	0.53	3	0.61	3	0.79	3	0.95
4	0.61	4	0.70	4	0.90	4	1.09
5	0.70	5	0.80	5	1.04	5	1.25
6	0.81	6	0.92	6	1.19	6	1.43
7	0.93	7	1.05	7	1.36	7	1.64
8	1.06	8	1.21	8	1.56	8	1.88
9	1.22	9	1.38	9	1.79	9	2.16
10	1.40	10	1.59	10	2.06	10	2.48
11	1.60	11	1.82	11	2.36	11	2.84
12	1.84	12	2.09	12	2.71	12	3.25
13	2.11	13	2.40	13	3.10	13	3.73
14	2.42	14	2.75	14	3.55	14	4.27
15	2.77	15	3.15	15	4.07	15	4.89
16	3.18	16	3.61	16	4.66	16	5.59
17	3.64	17	4.13	17	5.33	17	6.40
18	4.17	18	4.73	18	6.10	18	7.31
19	4.78	19	5.41	19	6.97	19	8.35
20	5.47	20	6.19	20	7.96	20	9.53
21	6.25	21	7.08	21	9.09	21	10.86
22	7.15	22	8.08	22	10.37	22	12.37
23	8.16	23	9.23	23	11.81	23	14.07
24	9.32	24	10.52	24	13.44	24	15.98
25	10.62	25	11.98	25	15.28	25	18.13
26	12.10	26	13.64	26	17.34	26	20.52
27	13.77	27	15.50	27	19.64	27	23.19
28	15.64	28	17.58	28	22.21	28	26.14
29	17.74	29	19.91	29	25.05	29	29.39
30	20.10	30	22.51	30	28.19	30	32.94
31	22.71	31	25.39	31	31.64	31	36.80
32	25.61	32	28.56	32	35.39	32	40.96
33	28.81	33	32.04	33	39.45	33	45.41
34	32.31	34	35.83	34	43.79	34	50.10
35	36.12	35	39.92	35	48.40	35	54.99
36	40.23	36	44.29	36	53.23	36	60.02
37	44.63	37	48.93	37	58.22	37	65.10
38	49.28	38	53.78	38	63.29	38	70.15
39	54.14	39	58.78	39	68.37	39	75.06
40	59.15	40	63.86	40	73.33	40	79.70

We determine the risks that are associated with each point in total. The first step is to select the point totals' theoretical range based on the point system computed earlier. In our point system, the theoretical range of point totals is −2 to 40. We then attached a risk estimate to each point total using the Cox regression equation.

Table 3.7 Risk categories based on total points

Total Score	**Risk Category (based on 5-years estimated risk)**
< 22 (< 10% estimated risk)	Low risk
22 - 27 (10 - 20% estimated risk)	Intermediate risk
> 27 (> 20% estimated risk)	High risk

Figure S3.1

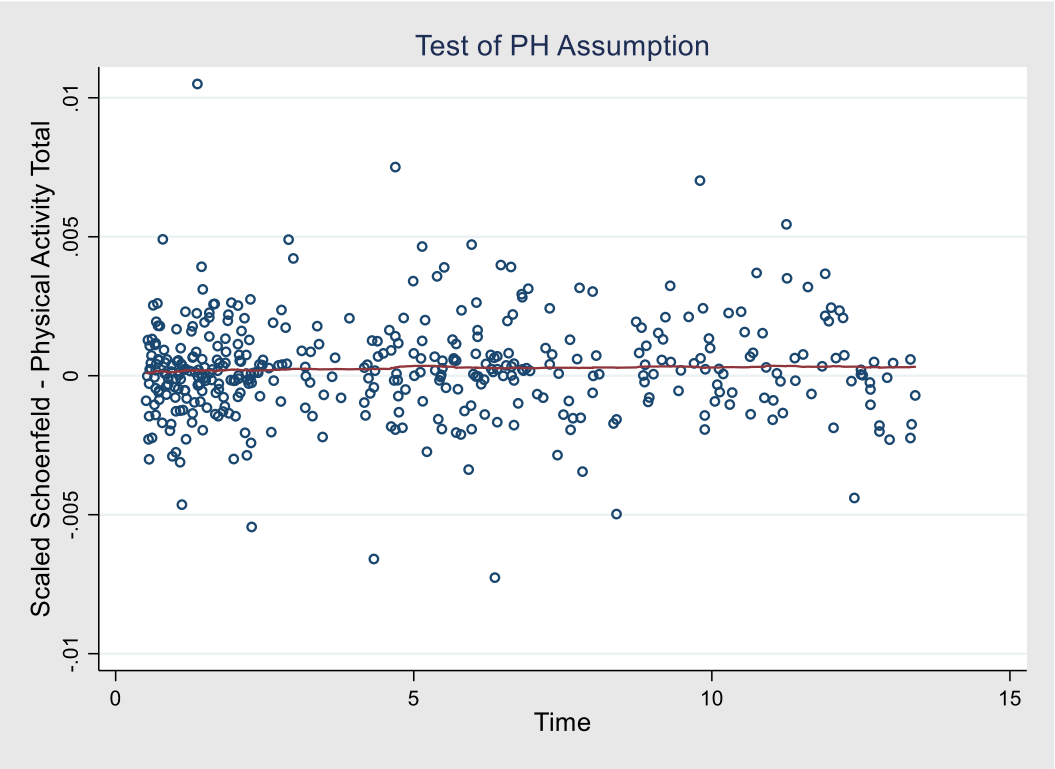

Figure S3.1 Plot to test the proportionality assumption of "Total physical activity time" variable.

Figure S3.2

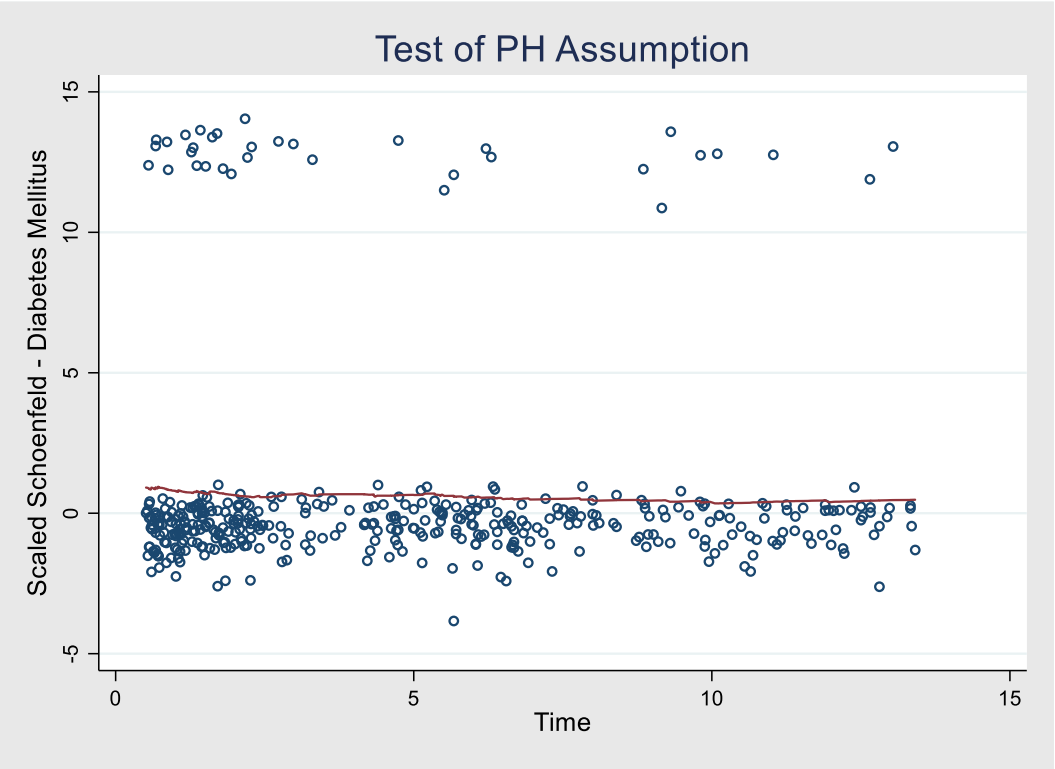

Figure S3.2 Plot to test the proportionality assumption of "Diabetes" variable.

Figure S3.3

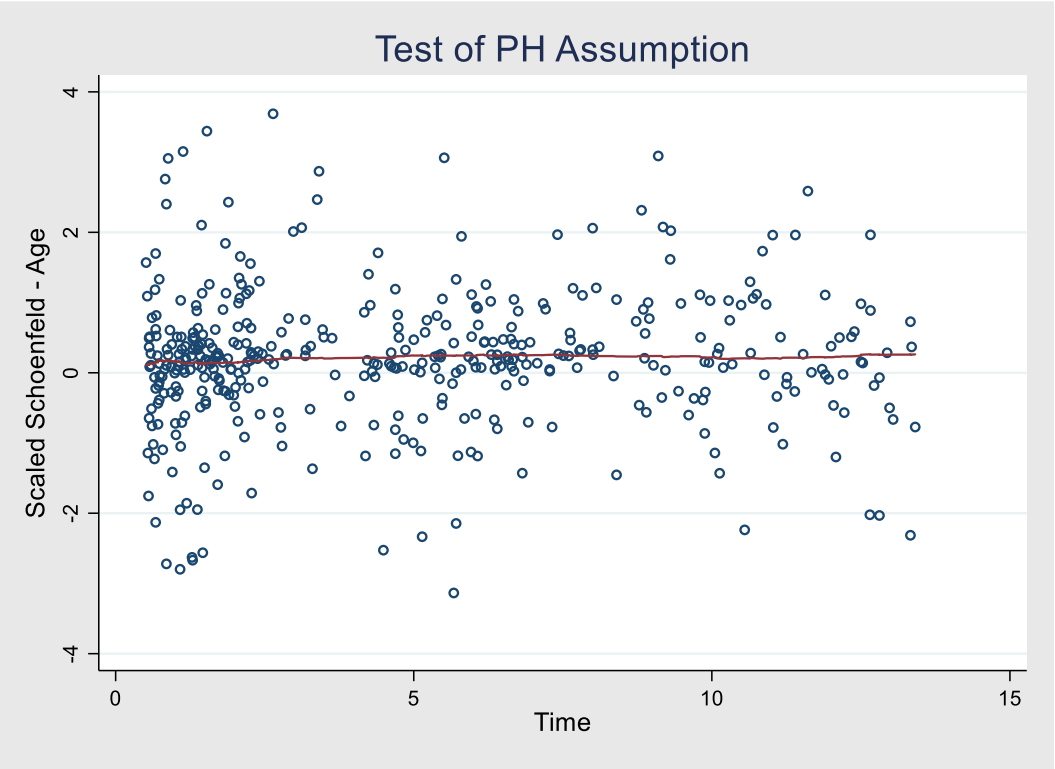

Figure S3.3 Plot to test the proportionality assumption of "Age" variable.

Figure S3.4

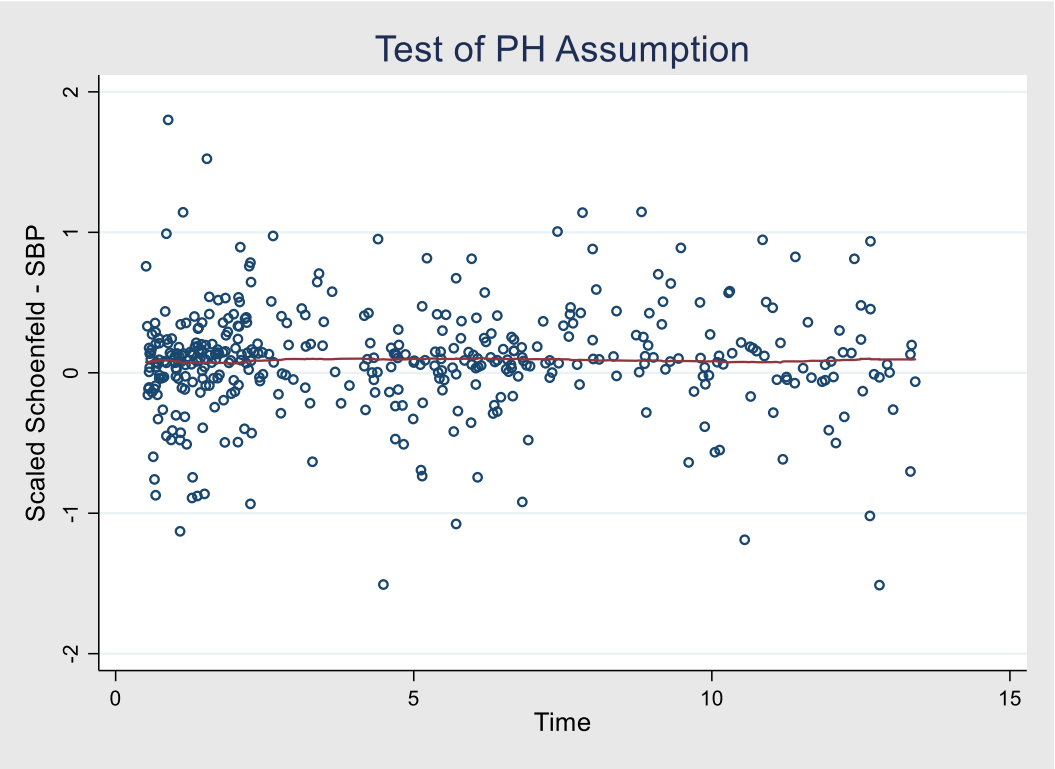

Figure S3.4 Plot to test the proportionality assumption of "Systolic blood pressure" variable.

Figure S3.5

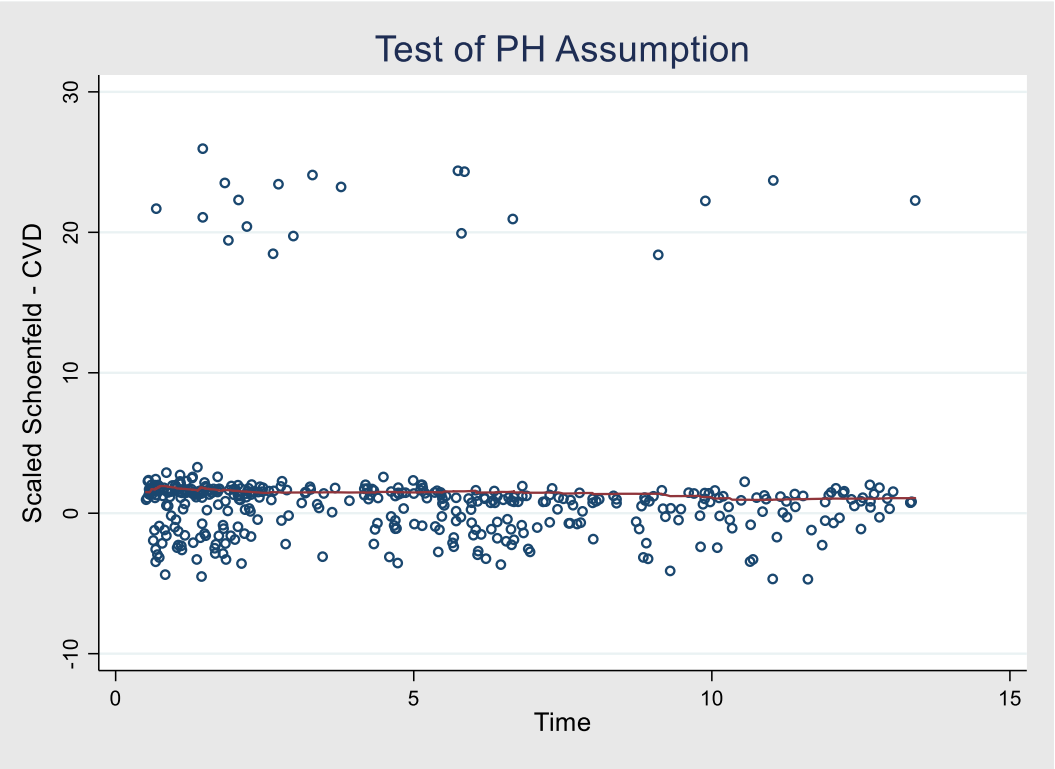

Figure S3.5 Plot to test the proportionality assumption of "Cardiovascular disease" variable.

Figure S3.6

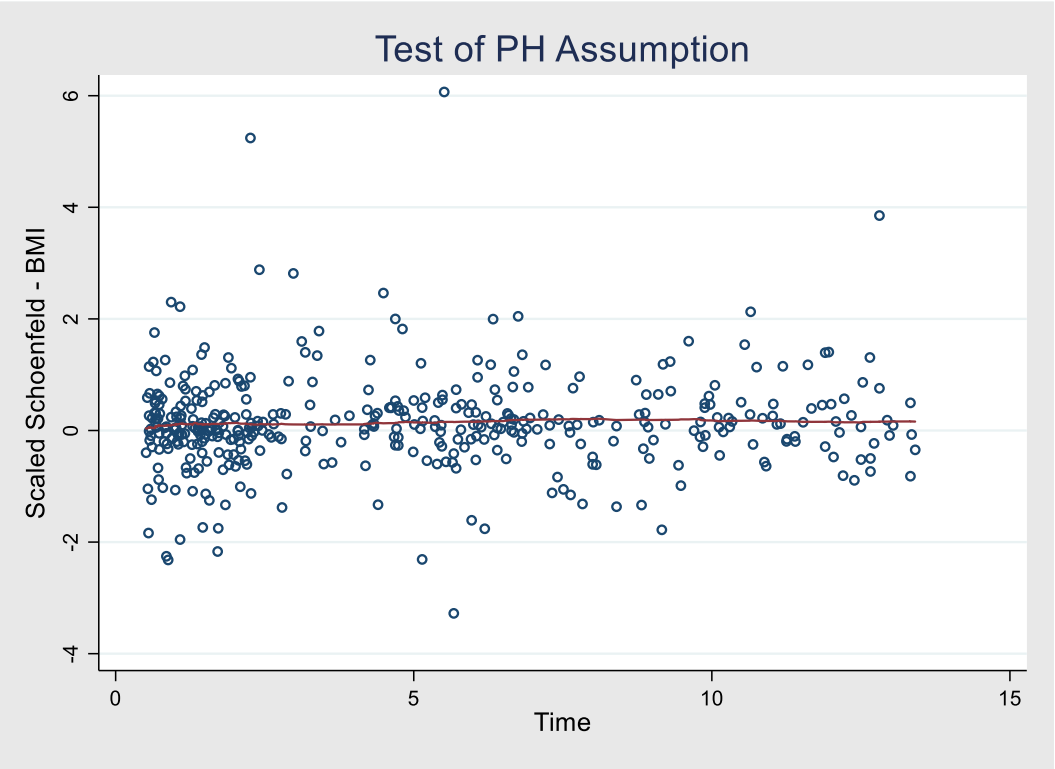

Figure S3.6 Plot to test the proportionality assumption of "Body mass index" variable.

Figure S3.7

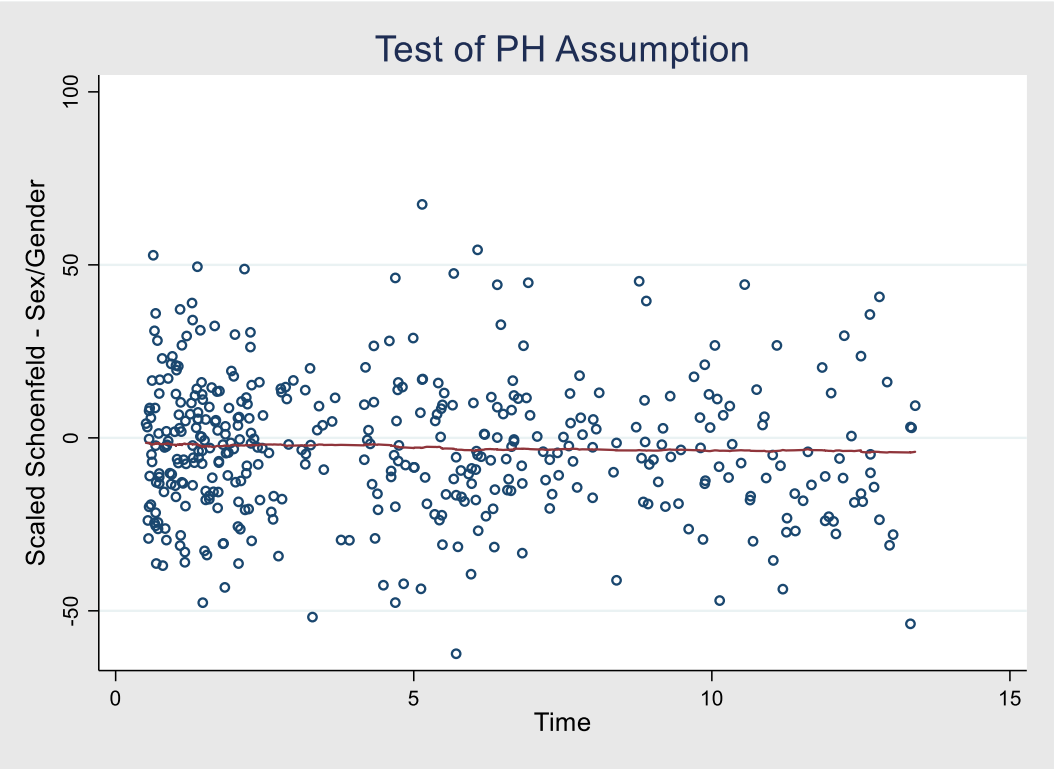

Figure S3.7 Plot to test the proportionality assumption of the "Sex" variable.

Figure S3.8

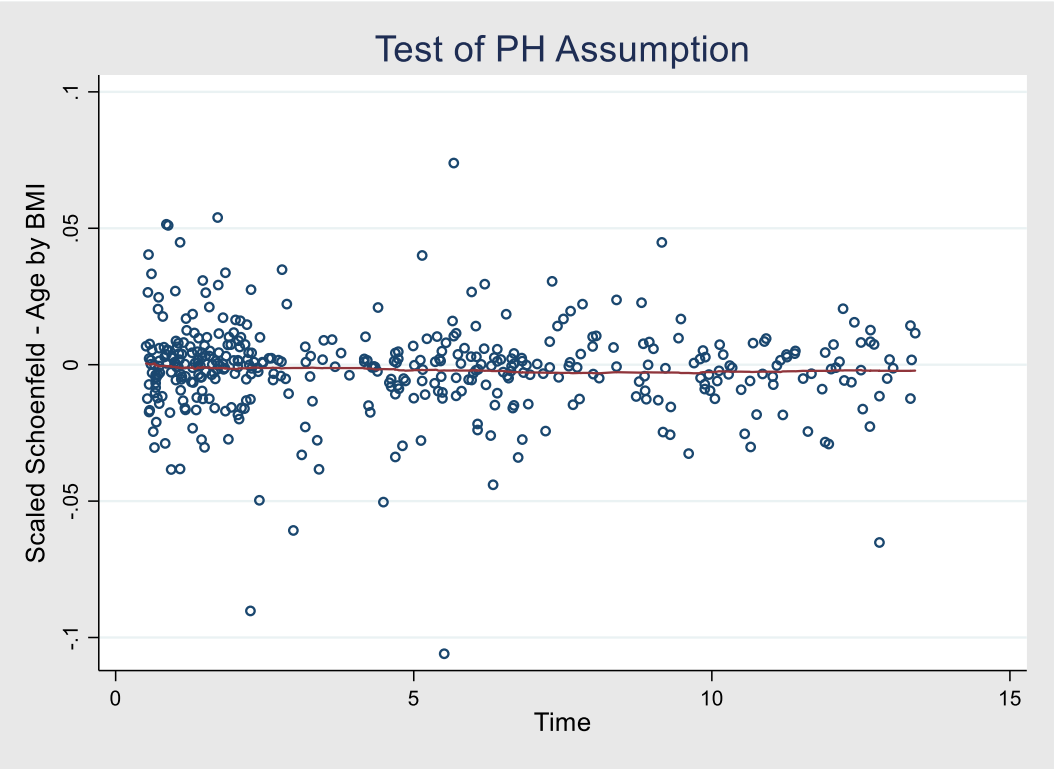

Figure S3.8 Plot to test the proportionality assumption of "Age by Body mass index" interaction variable.

Figure S3.9

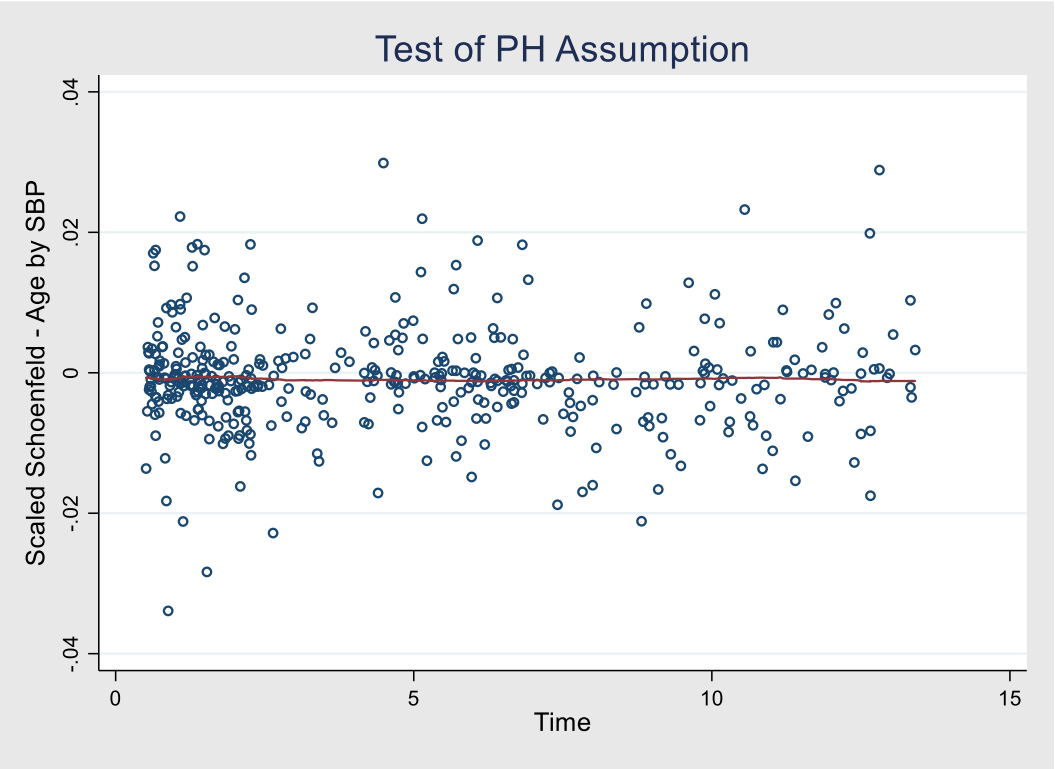

Figure S3.9 Plot to test the proportionality assumption of "Age by Systolic blood pressure" interaction variable.

Figure S3.10

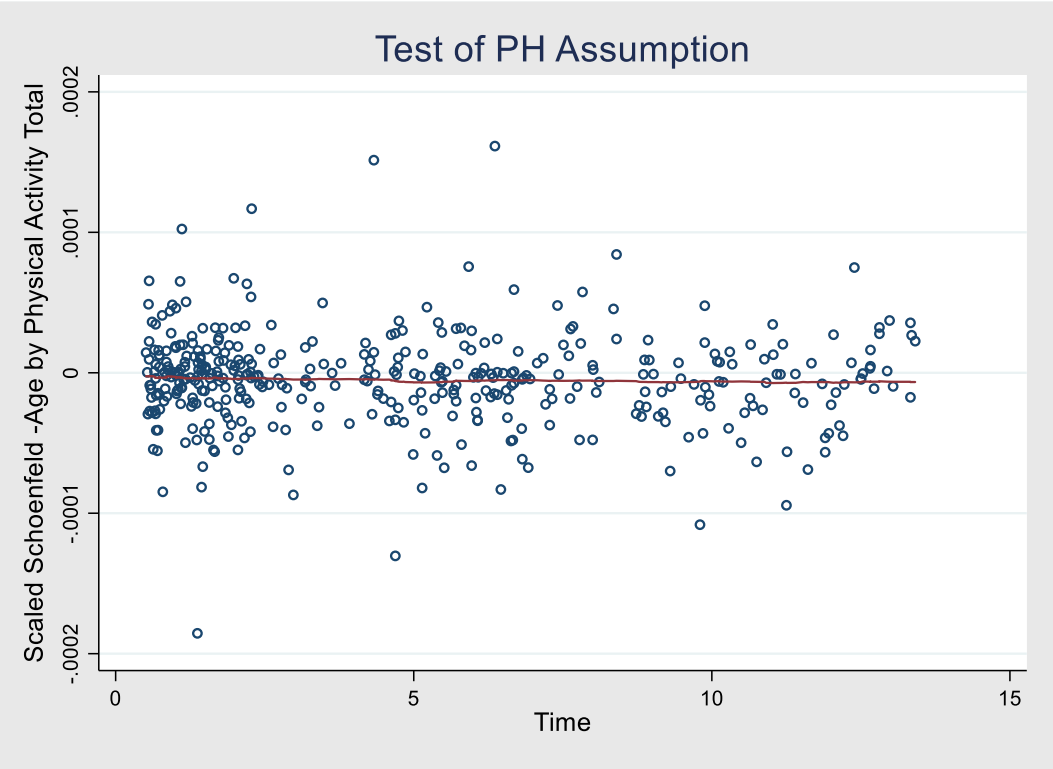

Figure S3.10 Plot to test the proportionality assumption of "Age by Total physical activity time" interaction variable.

Figure S3.11

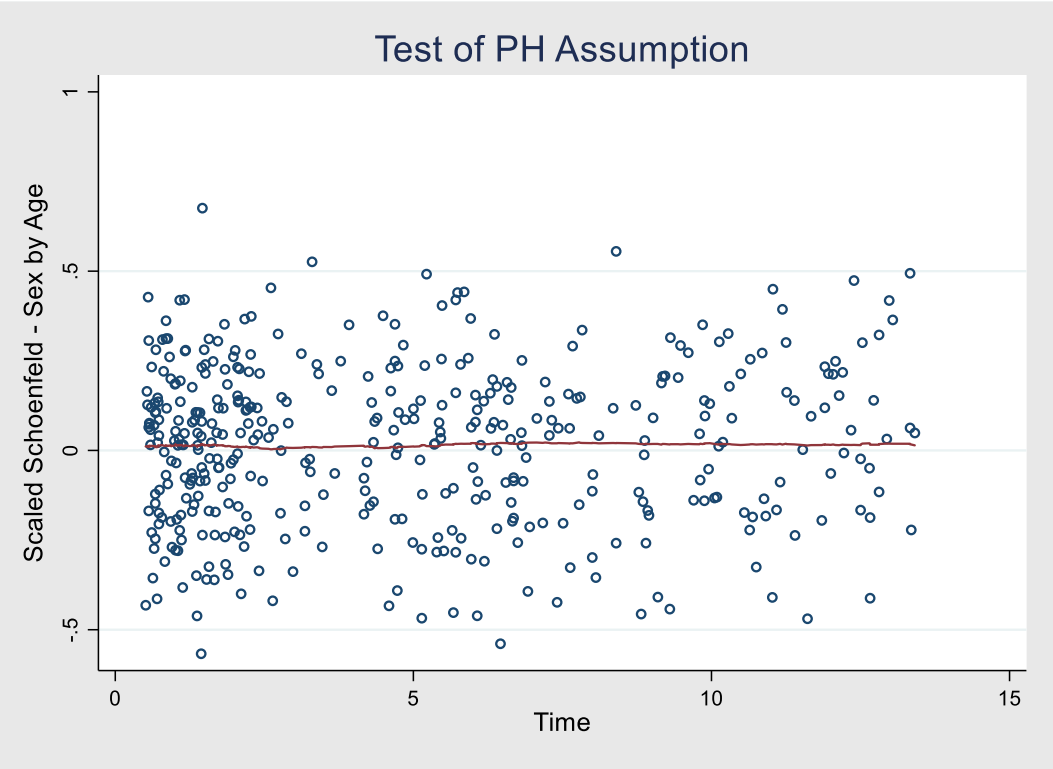

Figure S3.11 Plot to test the proportionality assumption of "Age by Sex" interaction variable.

Figure S3.12

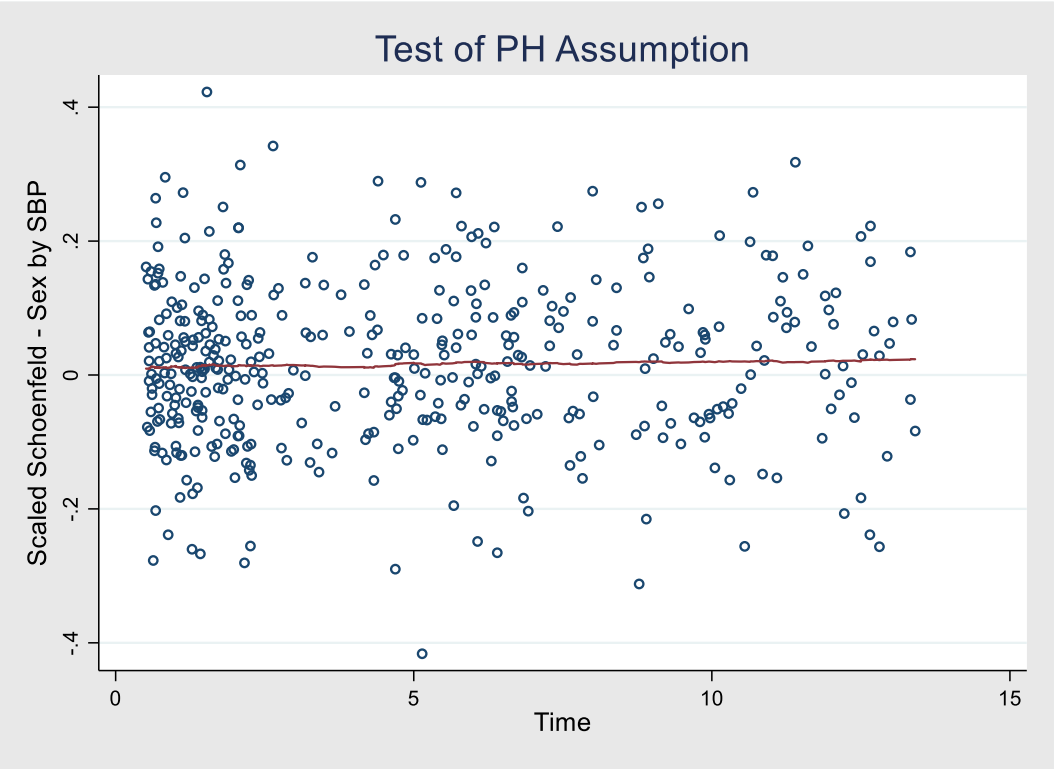

Figure S3.12 Plot to test the proportionality assumption of "Sex by Systolic blood pressure" interaction variable.

Figure S3.13

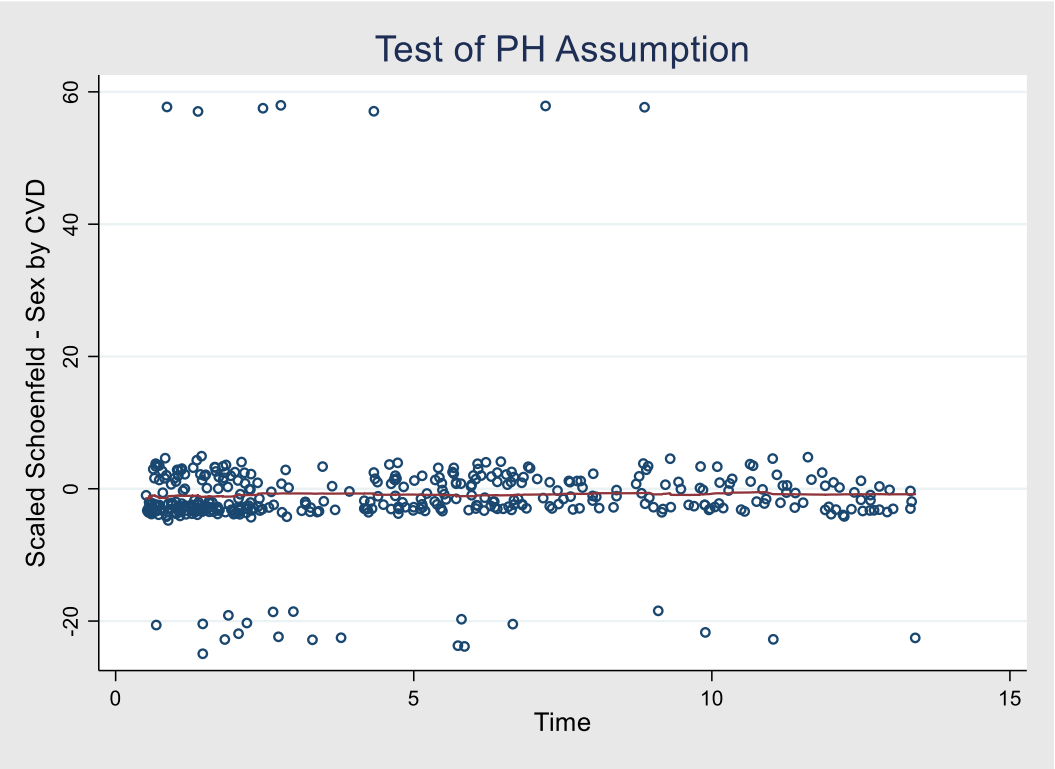

Figure S3.13 Plot to test the proportionality assumption of "Sex by Cardiovascular disease" interaction variable.

Figure S3.14

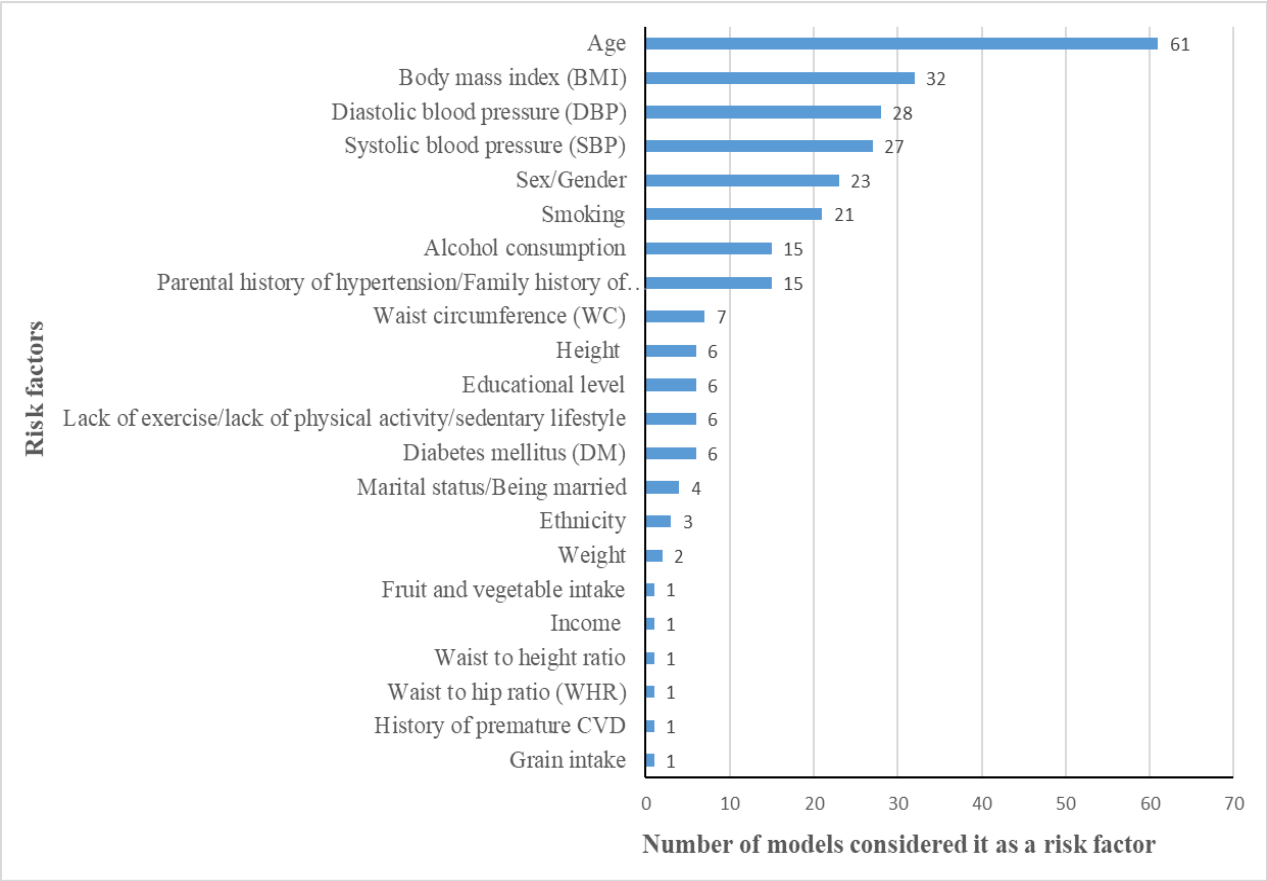

Figure S3.14 Traditional risk factors considered by conventional regression-based models.

Table S3.1 Missing information about different variables

Variables	Missing	Total	Percent Missing
Total Physical Activity Time	520	18,322	2.84
Total Sitting Time	1,421	18,322	7.76
Depression	16	18,322	0.09
Diabetes	8	18,322	0.04
Waist Hip Ratio	4,686	18,322	25.58
Sex	0	18,322	0.00
Age	0	18,322	0.00
Residence	0	18,322	0.00
Family History of Hypertension	0	18,322	0.00
Diastolic Blood Pressure	4,283	18,322	23.38
Systolic Blood Pressure	4,283	18,322	23.38
Ethnicity	23	18,322	0.13
Cardiovascular Disease	0	18,322	0.00
Highest Education Level Completed	11	18,322	0.06
Working Status	0	18,322	0.00
Vegetable and Fruit Consumption	266	18,322	1.45
Physical Activity	1,846	18,322	10.08
Total Household Income	1,402	18,322	7.65
Alcohol Consumption	846	18,322	4.62
Total Sleep Time	239	18,322	1.30
Smoking Status	45	18,322	0.25
Job Schedule	4,303	18,322	23.49
Marital Status	7	18,322	0.04
Body Mass Index	4,260	18,322	23.25
BMI Waist Ratio	4,718	18,322	25.75
Ever Smoked	41	18,322	0.22
Body Fat Percentage	4,471	18,322	24.40
Hip Circumference	4,564	18,322	24.91
Waist Circumference	4,769	18,322	26.03

Table S3.2 Baseline characteristics of study participants according to the missing status

Socio-demographic characteristics of groups				
Variable		**Observations (without missing values)**	**Observations (imputed missing values)**	**P-value**
Age, years, mean (SE)		50.99 (9.20)	-	-
Sex, n (%)	Male	5,763 (31.45)	-	-
	Female	12,559 (68.55)	-	-
Body Mass Index, kg/m2, mean (SE)		26.40 (4.78)	26.62 (5.27)	0.009
BMI Waist Ratio, mean (SE)		0.28 (0.03)	0.28 (0.03)	< 0.001
Hip Circumference, mean (SE)		104.80 (9.92)	104.99 (10.41)	0.257
Waist Circumference, mean (SE)		92.38 (13.14)	92.44 (13.28)	0.785
Waist Hip Ratio, mean (SD)		0.91 (0.07)	0.91 (0.07)	0.100
Body Fat Percentage, mean (SE)		31.90 (8.56)	31.86 (8.79)	0.795
Diastolic Blood Pressure, mmHg, mean (SE)		72.87 (9.36)	73.22 (9.29)	0.032
Systolic Blood Pressure, mmHg, mean (SE)		119.63 (13.71)	120.41 (13.78)	0.001
Marital Status, n (%)	Married and/or living with a partner	14,451 (78.90)	7 (100.00)	0.392
	Single, never married	1,180 (6.44)	0 (0.00)	
	Other (divorced, widowed, separated)	2,684 (14.65)	0 (0.00)	
Residence, n (%)	Urban	15,272 (83.35)	-	-
	Rural	3,050 (16.65)	-	
Total Household Income, n (%)	< $49,999	2,562 (15.14)	293 (20.90)	< 0.001
	$50,000 - $99,999	5,427 (32.07)	462 (32.95)	
	$100,000 - $199,999	6,649 (39.30)	500 (35.66)	
	≥ $200,000	2,282 (13.49)	147 (10.49)	
Highest Education Level Completed, n (%)	High school or below (none, elementary school, high school,	6,158 (33.63)	3 (27.27)	0.769

	trade, technical or vocational school, apprenticeship training or technical CEGEP)			
	Diploma but below bachelor's degree (diploma from a community college, pre-university CEGEP or non-university certificate, university certificate below bachelor's level)	4,924 (26.89)	4 (36.36)	
	Bachelor's degree or above (bachelor's degree, graduate degree (MSc, MBA, MD, PhD, etc.))	7,229 (39.48)	4 (36.36)	
Ethnicity, n (%)	Aboriginal	68 (0.37)	0 (0.00)	0.978
	Asian (South Asian, East Asian, Southeast Asian, Filipino, West Asian, Arab)	826 (4.51)	1 (4.35)	
	White	16,873 (92.21)	22 (95.65)	
	Latin American Hispanic	162 (0.89)	0 (0.00)	
	Black	97 (0.53)	0 (0.00)	
	Other (Jewish and others)	273 (1.49)	0 (0.00)	
Diabetes, n (%)		735 (4.01)	0 (0.00)	0.563
Cardiovascular Disease, n (%)		377 (2.06)	-	-
Depression, n (%)		2,009 (10.97)	4 (25.00)	0.073
Family History of Hypertension, n (%)		10,946 (59.74)	-	-
Smoking Status, n (%)	Never	10,084 (55.17)	32 (71.11)	0.028
	Former	6,755 (36.96)	8 (17.78)	
	Current	1,438 (7.87)	5 (11.11)	
Ever Smoked, n (%)		8,197 (44.84)	9 (21.95)	0.003
Alcohol Consumption, n (%)	Never	1,210 (6.92)	83 (9.81)	0.002
	≤ 1 time a week	9,177 (52.51)	467 (55.20)	
	2 to 3 times a week	3,653 (20.90)	154 (18.20)	
	4 to 5 times a week	1,909 (10.92)	84 (9.93)	
	≥ 6 times a week	1,527 (8.74)	58 (6.86)	
Working Status, n (%)	Full time	10,281 (56.11)	-	-
	Part time	3,719 (20.30)	-	

	Other (looking after home, disable/sick, student, unpaid/voluntary)6697	3,974 (21.69)	-	
	Unemployed	348 (1.90)	-	
Total Sleep Time, n (%)	≤ 5 hours (short sleep duration)	1,179 (6.52)	12 (5.02)	0.533
	6 hours	3,685 (20.38)	54 (22.59)	
	7 hours (reference)	6,955 (38.46)	87 (36.40)	
	8 hours	5,046 (27.90)	65 (27.20)	
	≥ 9 hours (long sleep duration)	1,218 (6.74)	21 (8.79)	
Total Physical Activity Time, mean (SE)		3168.50 (2866.67)	2817.41 (2930.63)	0.006
Total Sitting Time, mean (SE)		2493.30 (1174.18)	2422.02 (1170.47)	0.028
Physical Activity, n (%)	Low (first quartile of physical activity time and fourth quartile of sitting time)	1,449 (8.79)	242 (13.11)	< 0.001
	Moderate (second and third quartile of physical activity time and sitting time)	13,050 (79.21)	1,429 (77.41)	
	High (fourth quartile of physical activity and first quartile of sitting time)	1,977 (12.00)	175 (9.48)	
Vegetable and Fruit Consumption, n (%)	Low consumption (less than 5 servings of vegetable and fruit)	15,031 (83.25)	242 (90.98)	0.004
	Moderate consumption (less than 5 servings of vegetable but more than 5 servings of fruit OR more than 5 servings of vegetable but less than 5 servings of fruits	2,509 (13.90)	20 (7.52)	
	High consumption (5 or more servings of vegetable and fruit)	516 (2.86)	4 (1.50)	
Job Schedule, n (%)	Regular daytime shift	10,918 (77.88)	1,002 (23.29)	< 0.001
	Other (evening shift, night shift, rotating shift, split shift, irregular shift, or on call)	3,101 (22.12)	3,301 (76.71)	

Table S3.3 Test of Cox proportional-hazards assumption

Variable	rho	χ^2	Degrees of freedom (df)	P-value
Sex	-0.06572	1.61	1	0.2049
Total Physical Activity Time	0.04143	0.54	1	0.4631
Diabetes	-0.03620	0.54	1	0.4611
Age	0.04250	0.67	1	0.4121
SBP	0.00164	0.00	1	0.9731
CVD	-0.05012	1.03	1	0.3109
BMI	0.05692	1.15	1	0.2826
Age by BMI	-0.06566	1.59	1	0.2080
Age by SBP	-0.01090	0.05	1	0.8167
Age by Total Physical Activity Time	-0.04543	0.65	1	0.4208
Age by Sex	0.04340	0.74	1	0.3906
Sex by SBP	0.03560	0.47	1	0.4952
Sex by CVD	0.00310	0.00	1	0.9501
Global Test		9.66	13	0.7216

CHAPTER 4. USING MACHINE LEARNING ALGORITHMS TO PREDICT HYPERTENSION INCIDENCE AND COMPARING THEIR PREDICTIVE PERFORMANCE WITH A CONVENTIONAL STATISTICAL MODEL IN A LARGE SURVIVAL DATA

4.1 Abstract

Risk prediction models are frequently used to identify individuals who are at risk of developing hypertension. This study evaluates different machine learning algorithms and compares their predictive performance with the conventional Cox proportional hazard (PH) model to predict hypertension incidence in survival data. We used the data of 18,322 participants on 24 candidate features from the large Alberta's Tomorrow Project (ATP) to develop different prediction models.

Feature selection methods included two filter-based: a univariate Cox p-value and C-index; two embedded-based: random survival forest and least absolute shrinkage and selection operator (Lasso); and one constraint-based: the statistically equivalent signature (SES), to select the top features. Five machine learning algorithms were developed to predict hypertension incidence: penalized regression Ridge, Lasso, Elastic Net (EN), random survival forest (RSF), and gradient boosting (GB), along with the conventional Cox proportional hazards (PH) model. The predictive performance of the models was assessed using C-index. The performance of machine learning algorithms was observed, similar to the conventional Cox PH model. Average C-indexes were 0.78, 0.78, 0.78, 0.76, 0.76, and 0.77 for Ridge, Lasso, Elastic Net, RSF, GB and Cox PH, respectively. Important features associated with each model were also presented.

Our study findings demonstrate little predictive performance difference between machine learning algorithms and the conventional Cox PH regression model in predicting hypertension incidence.

4.2 Introduction

Hypertension has long been documented as a substantial health burden that affects all segments of the population. Globally, hypertension causes 17.8% (9.4 million) of deaths every year and 7% of disease burden, making it one of the most significant risk factors for global mortality and disease burden[1,2]. Individuals with hypertension are at higher risk for developing not only life-changing, but also possibly life-threatening conditions[3]. Left uncontrolled or undetected, high blood pressure (BP) can lead to dangerous health complications and poor life quality. Due to the high prevalence and global burden of hypertension, early detection and prevention strategies need to be a top priority.

One of the priorities of health and clinical research is to identify people at higher risk of developing an adverse health outcome such as hypertension so they can be targeted for early preventative strategies and treatment[4]. Multiple factors may cause and increase the risk of hypertension, including physical, hereditary, or behavioral. Individuals who are healthy but are found to have a high risk of developing hypertension could be recommended to change their lifestyle and behaviors (e.g., physical activity, dietary pattern, alcohol consumption, smoking, etc.) to reduce their risk. Prediction modeling can play a vital role in identifying high-risk individuals. Prediction models can be used to estimate the risk of future occurrence of a health condition in an individual by utilizing different underlying demographic and clinical characteristics called risk factors that are believed to be associated with the health outcome of interest. Prediction models help predict the chance of experiencing a health outcome by an individual with a given set of risk factors.

Various models have been developed that mathematically combine multiple risk factors to estimate the risk of hypertension in asymptomatic subjects in the population. While specific details

may vary between clinical risk prediction models, the goals and processes of developing prediction models are mostly similar. From all available variables, candidate variables are selected based on clinical and statistical viability. A predictive model is derived using an appropriate modeling strategy from the chosen candidate variables, and its utility is internally validated.

The regression-based methodology is the conventional approach for developing prediction models. Logistic regression (for binary endpoint/outcome) and Cox regression (for time-to-event endpoint/outcomes) are the most frequently used algorithms for conventional regression-based prediction models. Machine learning algorithms recently emerged as a popular modeling approach that offers an alternative class of models with more computational flexibility[5]. Over the last few years, machine learning algorithms achieved significant successes across a broad range of fields due to their superiority, such as their ability to model nonlinear relations and the accuracy of their overall predictions[6]. Decision trees, random forest, penalized regression models, neural networks, and support vector machines are examples of machine learning algorithms[7].

The vast majority of developed hypertension risk prediction models are conventional regression-based models[8–17]. Machine learning-based models also exist in the hypertension prediction domain[18,19,28,29,20–27]. Machine learning algorithms sometimes struggle with reliable probabilistic estimation and interpretability[30,31]. Moreover, in clinical applications, machine learning algorithms often produce mixed results in predictive performance compared with conventional regression models[32–36]. Among the models where machine learning algorithms were used to predict hypertension, data were mostly cross-sectional. Models were built without considering or utilizing survival information where time is an inherent part of model building. Due to the lack of survival data utilization in predicting hypertension in the machine learning domain, it is unclear how machine learning-based models will perform in predicting hypertension in

survival data. A formal comparison in predictive performance between conventional regression-based hypertension prediction models and machine learning-based models in a survival setting is also absent. There is also a scarcity of comparisons using the same dataset. This motivated us to assess and compare machine learning algorithms' predictive performance with conventional regression-based models in a survival setting.

In this study, we investigated and compared five machine learning algorithms' performance with the conventional Cox PH regression model to predict the risk of developing hypertension using Alberta's Tomorrow Project cohort data.

4.3 Methods

4.3.1 Study population

The data used in this study are from Alberta's Tomorrow Project (ATP) cohort data, which is Alberta's largest longitudinal population health cohort and contains data for more than 55,000 adults from the general population aged 35-69 years. ATP contains baseline and longitudinal information on socio-demographic characteristics, personal and family history of the disease, medication use, lifestyle and health behavior, environmental exposures, and physical measures. ATP launched in 2000, and in 2008 joined the Canadian Partnership for Tomorrow Project (CPTP)[37]. ATP has several questionnaires, and this study uses data from the CORE questionnaire. A more detailed description of ATP data is provided in Chapter 3. Our study cohort consists of 25,359 participants between 35-69 years of age at enrolment. Eligible subjects were free of hypertension at baseline and consented to have their data linked with Alberta's administrative health data. Linking with administrative health data was primarily done due to the lack of follow-up data in ATP, which was necessary to determine hypertension incidence. A detailed description of data linkage is provided in Appendix 1. We excluded 6,996 participants from the analysis who

had hypertension at baseline and consequently did not meet eligibility criteria (free of hypertension at baseline). We also excluded 41 participants who responded to hypertension status questions at baseline as “don’t know” or “missing”. Eighteen thousand three hundred twenty-two participants remained after exclusion and were finally included in the analysis. The Conjoint Health Research Ethics Board (CHREB) at the University of Calgary granted ethical approval for this study.

4.3.2 Selection of candidate features

We compiled a list of available potential candidate features before launching the analysis. We determined the possible candidate features for model development based on a literature search, features used in the past, and discussion with content experts. We initially considered 24 candidate features for the model development process. Given our model’s intended clinical application, we deliberately did not consider any genetic risk factors/biomarkers as potential candidate features.

4.3.3 Definition of features

The outcome incident hypertension was determined from linked administrative health data using a coding algorithm. We used the relevant ICD-9 and ICD-10 codes (ICD-9-CM codes: 401.x, 402.x, 403.x, 404.x, and 405.x; ICD-10-CA/CCI codes: I10.x, I11.x, I12.x, I13.x, and I15.x) and a validated hypertension case definition (two physician claims within two years or one hospital discharge for hypertension) to define hypertension incidence[38].

The age of the study participants, body mass index (BMI), the waist-hip ratio, diastolic blood pressure (DBP), systolic blood pressure (SBP), total physical activity time (total MET minutes/week), and total sitting time (the sum of the sitting times on weekdays and weekends) were all considered as continuous features. The remaining features were categorical. The sex of the participants was either male or female. The residence was either urban or rural. Marital status was categorized into three groups: married and/or living with a partner, single who never married,

and others (divorced, widowed, separated). Total household income was categorized into four groups: < $49,999, $50,000-$99,999, $100,000-$199,999, and ≥ $200,000. The highest education level completed was categorized into three groups: high school or below (none, elementary school, high school, trade, technical or vocational school, apprenticeship training or technical CEGEP), diploma but below bachelor's degree (diploma from a community college, pre-university CEGEP or non-university certificate, university certificate below bachelor's level), and bachelor's degree or above (bachelor's degree, graduate degree [MSc, MBA, MD, PhD, etc.]). Ethnicity was categorized into six groups: Aboriginal, Asian (South Asian, East Asian, Southeast Asian, Filipino, West Asian, Arab), White, Latin American Hispanic, Black, and other (Jewish and others). Diabetes was categorized as "yes" or "no" based on the response to the question "Has a doctor ever told you that you had diabetes?". Cardiovascular disease was categorized as "yes" if any stroke, myocardial infarction, angina, arrhythmia, coronary heart disease, coronary artery disease, heart disease, and heart failure was present and as 'no" if absent. Depression was categorized as "yes" or "no" based on the response to the question "Has a doctor ever told you that you had depression?". Family history of hypertension was categorized as "yes" if any first-degree relative was diagnosed with hypertension, otherwise "no". Smoking status was categorized as: never, former, and current. Alcohol consumption was categorized into five groups: never, ≤ 1 time a week, 2 to 3 times a week, 4 to 5 times a week, and ≥ 6 times a week. Working status was categorized into four groups: full-time, part-time, other (looking after a home, disable/sick, student, unpaid/voluntary), and unemployed. Total sleep time was categorized into four groups: ≤ 5 hours (short sleep duration), 6 to 7 hours, 8 hours, and ≥ 9 hours (long sleep duration). Physical activity was categorized as: low (first quartile of physical activity time and fourth quartile of sitting time), moderate (second and third quartile of physical activity time and sitting time), and high

(fourth quartile of physical activity and first quartile of sitting time). Vegetable and fruit consumption was categorized as low (less than 5 servings of vegetable and fruit), moderate (less than 5 servings of vegetable but more than 5 servings of fruit OR more than 5 servings of vegetable but less than 5 servings of fruits), and high (5 or more servings of vegetable and fruit). Job schedule was categorized as regular daytime shift and other (evening shift, night shift, rotating shift, split shift, irregular shift, or on-call).

4.3.4 Missing values

Our dataset has missing values on several candidate features ranging from 0 to 26%. Information on missing values for different candidate features is presented in the supplementary table (Table S4.1). To impute the missing data, we have used the multiple imputations by chained equations method[39,40].

4.3.5 Feature selection

Modern-day datasets are rich in information with data collected on many features, making the data high dimensional. Such high-dimensional datasets create computational difficulty and complicate the interpretability of a prediction model. Feature selection is a process where a subset of relevant features from a large amount of data is selected to filter the dataset down to the smallest possible subset of accurate features. It is imperative to identify the relevant features from a dataset and remove less significant features that have a minimal contribution to the outcome to achieve better prediction model accuracy. Feature selection is one of the core concepts in machine learning that massively impacts a model's performance. Feature selection offers enhanced model performance by mitigating the risk of overfitting, improved computational speed and time, decreased computational requirements, and easier interpretability of the model.

Feature selection methods can be classified into three categories: filter, wrapper, and embedded methods[41]. Filter methods use feature ranking techniques as the main criteria for feature selection[41]. An appropriate ranking criterion is applied to score the features, and features that are below a specified threshold are eliminated. Filter methods serve as a preprocess to rank the features in which the highest-ranked features are selected. Wrapper methods use the performance of the model as the feature selection criterion[41]. The model is wrapped in a search algorithm that will find a subset of the features that give the highest model performance. Embedded methods integrate the selection of features as part of the model building process[41].

There are different ways to assign numerical scores within filter methods so that features can be ordered based on their relevance. This study used two popular variants in the survival analysis setting: a univariate Cox p-value and C-index[42]. A univariate Cox model is separately applied for each feature, and p-values are obtained[43]. These p-values are used as importance scores. The C-index calculation is performed for each feature without fitting a survival model. The resulting C-index is used as a score for that feature[42]. Features are ordered according to their C-index, and a higher C-index indicates more importance. This study also employed two popular embedded methods of feature selection: RSF and Lasso. Both are machine learning approaches for building prediction models but also perform feature selection. Variable importance in RSF is calculated using a prediction error approach involving noising up the feature by randomly permuting its value[44]. A feature's variable importance is the difference between a prediction error when a feature is noised and a prediction error in the original feature[44]. The Lasso method shrinks the regression model's coefficients as part of penalization, and the features left after the shrinkage process are selected for model building. In Lasso, prediction/fitting errors are minimized using the objective functions, and the features with near-zero regression coefficients are eliminated[45]. We also

employed the statistically equivalent signature (SES)[46], a constraint-based method for feature selection that tries to identify multiple subsets of predictive features whose performance is statistically equivalent[47]. The signature here implies minimal sets of features with maximum predictive power. The primary purpose of running the SES algorithm is to select variables as important according to the increasing p-value.

4.3.6 Machine learning models

Modeling survival analysis (time-to-event data) requires specialized methods to handle unique challenges such as censoring, truncation, time-varying features, and effects. Censoring, where the event of interest is not observed due to time constraint or lost to follow-up during the study period, is challenging, and survival analysis provides different mechanisms to deal with such problems. Application of typically used statistical or machine learning approaches to analyze survival data is impractical, as regular statistical and supervised machine learning algorithms do not inherently handle censored data. Statistical methods for handling survival data are well established. Several machine learning algorithms have also been developed and adapted to work with survival analysis data, effectively addressing complex challenges associated with survival data.

This study developed five machine learning algorithms, namely RSF, boosted gradient, penalized Lasso, penalized Ridge, and penalized EN. We provide a brief description of these models below. Although it is not a machine learning algorithm, the Cox PH model is included here as a conventional regression-based model (baseline) against which we compared the machine learning-based models.

4.3.6.1 Cox PH model

The Cox PH model is considered the standard model for analyzing survival data[43]. The Cox PH model is semi-parametric (since the baseline hazard function, $h_0(t)$, is unspecified) and evaluates the effects of observed risk factors simultaneously on the time to an event of interest (e.g., diagnosis of a disease). It is the most frequently used method for modeling an individual's survival, given their baseline data.

The Cox PH model is stated by the hazard function, which is the risk of an event occurring at time *t*. The formula for the Cox PH model is

$$h\left(t, X_1, X_2, \ldots X_p\right) = h_0(t)\exp\left(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p\right)$$

where $h\left(t, X_1, X_2, \ldots X_p\right)$ is the expected or predicted hazard at time t for a subject with covariate values, $X_1, X_2, \ldots X_p$, $h_0(t)$ is the baseline hazard when all the covariates equal to zero, exp is the exponential function, X_i is the i^{th} covariate in the model, and β_i is the regression coefficient for the i^{th} covariate, X_i.

The Cox PH model does not assume a particular distribution for the survival times. The baseline hazard function is also unspecified (no assumptions about the shape of the function), which can take any form and only a function of time (i.e., no covariates involved). However, the model is limited by some strict assumptions, such as the proportional hazards, and violation of these assumptions will end up in completely misleading results. The regression coefficients in the Cox PH model are estimated by maximizing the partial likelihood.

4.3.6.2 Penalized Cox regressions (Lasso, Ridge, and EN)

When applied to high-dimensional data (the number of features in the data is almost equal to or even exceeds the number of observations), the basic Cox model does not generalize well. The model may perform poorly and provide inaccurate results due to overfitting (which occurs when a model is tailored to a specific dataset and cannot generalize to other datasets)[48]. Overfitting can be

prevented through regularization, a process of introducing additional information into the model. Several different penalty functions have been developed and introduced in prediction models to identify the most relevant features of the outcome in high-dimensional data. Such a model is called a penalized model, which adds penalty functions to restrict the features. This restriction reduces or shrinks the coefficient values toward zero to ensure that the model has less impact on the less relevant features.

The two most commonly used regularizers are the L1 penalty and L2 penalty. In the L1 penalty (also known as Lasso), the sum of the coefficients' absolute value is penalized, and feature selection and regression coefficient estimation are simultaneously performed. The L1 penalty yields sparse models (models with a smaller number of features) that are more easily interpreted[7]. In the L2 penalty (also known as Ridge regression), the sum of squared coefficients is penalized. Unlike Lasso, Ridge regression cannot produce a sparse model, as any of the coefficients never become precisely zero, and hence none are eliminated. For the same reason, Ridge regression also cannot perform variable selection. Lasso suffers from some limitations because it cannot select more features than the number of observations, and in cases where there are correlated features, it tends to choose only one from a group without discrimination[49]. Lasso feature selection can be too data-dependent and therefore unstable.

EN emerged from Lasso criticism and provided a solution by combining the Ridge regression and Lasso penalties to get the best of both worlds. EN is a linear combination of the L1 and L2 penalties and can perform feature selection and deal with the correlation between the features simultaneously[49]. Unlike Lasso, EN can be useful when the number of features is larger than the number of observations.

4.3.6.3 Random survival forest

The random forest[50] is an ensemble method specifically designed to make predictions using tree-structured models. RSF[51] is an extension of the original Breiman's random forest[50] to censored survival data using a forest of survival trees for prediction. In an RSF, many bootstrap samples are randomly drawn from the given dataset, and for each sample, a survival tree is built by randomly selecting features. Each node is split based on randomly selected candidate features in an RSF to maximize the child nodes' survival difference. Using the non-parametric Nelson-Aalen (NA) estimator, the ensemble Cumulative Hazard Function (CHF) of the bootstrapped samples is calculated by taking the CHF average of each tree[6]. Randomization in RSFs reduces the correlation among the trees and thus improves the predictive performance. RSF offers many advantages: the ability to model complex, nonlinear data, handle high-dimensional data, identify interactions, and naturally impute missing data, and has become a popular and powerful tool for survival prediction[52].

4.3.6.4 Boosted gradient

The idea behind boosting is to add new models to the ensemble sequentially. At each iteration, a new weak, base-learner model (where the error rate is only a little better than random guess) is trained concerning the error (residuals) of the whole ensemble learned so far and improved the remaining error iteratively. Once it reaches a stage where errors cannot be improved, the process can be stopped. Algorithmically, a loss function is minimized such that loss becomes its minimum. GB[53] identifies the shortcomings of weak learners by using gradients in the loss function.

4.3.7 Feature importance

It is crucial to communicate machine learning algorithms' findings to an audience who may not be familiar with such algorithms. Just presenting the algorithm's predictive performance is

often not enough. Somehow, we need to attribute the predictions to the input data elements that contribute to model accuracy. Feature importance is a tool that refers to a class of techniques for assigning scores to input features according to their usefulness at predicting a target feature. The relative scores can indicate which features are most relevant to the target and which are not. Feature importance helps with interpreting and explaining machine learning algorithms by illustrating the predictive power of the dataset's features.

Function for computing the importance of features in RSF, GB, and Cox PH models is based on Breiman's permutation method[50], where each feature is randomly permuted at a time, and the associated reduction in predictive performance is calculated. For the penalized models, the standardized regression coefficients' magnitude was used to rank order the features according to their importance[54]. To ensure comparable rank-ordering across all models, the importance metrics' absolute values for all the features were scaled to unit norm[55].

4.3.8 Statistical analysis

We first imputed the missing values. We then randomly split subjects into two sets: the training set, which included 67% (two-thirds) of the sample (n = 12,233), and the testing set, which included the remaining 33% (one-third) (n = 6,089). The two groups' baseline characteristics were compared using the unpaired t-test or the $\chi 2$-test, as appropriate. We developed risk prediction models from the training data and assessed the models' performance using the testing data. Continuous features remained continuous in the model development. Five different feature selection methods were employed to derive the most accurate risk prediction model for all the machine learning and conventional regression models. Features were first ranked according to their importance/scores/p-values. Based on the features' ranking, the top 20 features by each of

the methods were selected. Due to the variations in selected top 20 features by different methods, features that are common in all the methods are finally considered in model building.

Five machine learning algorithms and the conventional Cox PH model were developed in the training set. Machine learning algorithms have hyper-parameters that need to be selected to optimize model performance. We carried on tuning these hyper-parameters automatically within a 10-fold nested cross-validation loop. Hyper-parameter values were chosen by applying 20 random iterations in the inner loop, and model performance was assessed in the outer loop. This ensured the repetition of model selection steps for each pair of training and test data. The number of random variables for splitting and the minimal number of events in the terminal nodes was tuned when building the RSF. We fitted a Cox PH model as a base learner for GB models. The number of boosting iterations and the regression coefficients were tuned in GB. For the penalized models, parameter lambda was tuned, and the best value was chosen based on 10-fold cross-validation. The models' predictive performance was evaluated using the concordance index (C-index)[56], which measures the proportion of pairs in which observation with higher survival time has the higher probability of survival as predicted by the model. The whole process was iterated 10 times by sampling the original data with replacement.

Moreover, the training data features were ranked according to their relative contribution to the prediction of hypertension incidence using various variable importance metrics. The analyses were conducted using several packages[40,54,57–62] of R software v 3.6.2.

4.4 Results

We presented the baseline characteristics of the study participants in Table 4.1 and Table 4.2. In Table 4.1, the study participants' characteristics are compared between training data and test data, while in Table 4.2, characteristics are compared according to the status of developing

hypertension. In Table 4.1, no significant difference ($p < 0.05$) in study characteristics was observed between training data and test data. During the median 5.8-year follow-up, 625 (3.41%) participants newly developed hypertension. In Table 4.2, most of the study characteristics were significantly different between those who developed hypertension and those who did not. Some study characteristics, however, were not significantly different, including marital status ($p = 0.146$), residence ($p = 0.146$), ethnicity ($p = 0.349$), depression ($p = 0.179$), family history of hypertension ($p = 0.061$), alcohol consumption ($p = 0.189$), total physical activity time ($p = 0.825$), and physical activity ($p = 0.707$). Overall, the study participants' mean age was 50.99 years, and the participation of females (68.55%) in the studies were higher than the males (31.45%).

Table 4.3 presents feature rankings of all 24 candidate features, and Table 4.4 shows the top 20 features based on five different methods. Due to the differences in the ranking by different methods, the top 20 selected features are not the same. To avoid any less relevant features in the model building process, we chose features common in the top 20 selected features by different methods. Fourteen features were identified as common in all top 20 features and were included in the final model building process (Table 4.4, red-colored cells). These included SBP, DBP, BMI, waist-hip ratio, diabetes, cardiovascular disease, age, job schedule, working status, total household income, residence, highest education level completed, family history of hypertension, and sex.

Figure 4.1 describes the relative importance of features concerning the prediction of hypertension incidence by six different model building approaches. The waist-hip ratio was selected as the top feature by Ridge regression and GB. In contrast, cardiovascular disease was selected as the top feature by Lasso regression and EN regression. In comparison, SBP was selected as the top feature by the Cox PH model and RSF. The waist-hip ratio, cardiovascular disease, diabetes, SBP, age, and BMI have been deemed the most important features considered

by most modeling approaches. However, there are also variations in the rank ordering of important features across the investigated models.

Figure 4.2 describes the predictive accuracy of different models. There were negligible differences in the accuracy of machine learning and conventional regression-based Cox models. The average C-index for the machine learning algorithms Ridge, Lasso, EN, RSF, and GB was 0.78, 0.78, 0.78, 0.76, and 0.76, respectively. In comparison, the conventional regression-based Cox PH model's average C-index was 0.77.

4.5 Discussion

In this study, we examined the predictive accuracy of machine learning algorithms and compared their performance with the conventional regression-based Cox PH model to predict hypertension incidence. The predictive accuracy of the machine learning algorithms and the Cox PH model was good[63], as the C-index was well over 0.70 in every case. Our findings suggest that the machine learning algorithm's predictive accuracy is similar to the regression-based Cox PH model. These findings are consistent with our recent systematic review and meta-analysis, where no evidence of machine learning algorithms' superior predictive performance over conventional regression-based models was observed. According to our meta-analysis, the overall pooled C-statistics of the machine learning-based algorithms was 0.76 [0.71 – 0.80], compared with an overall pooled C-statistic of 0.75 [0.73 – 0.77] in the traditional regression-based models.

In the past, several machine learning algorithms were developed for predicting hypertension[18,19,28,29,20–27]. Most of those algorithms used cross-sectional data and did not predict hypertension incidence. Some of the models used longitudinal data but did not incorporate time into their model. Only two models predicted the incidence of hypertension, considering survival data using machine learning algorithms[21,64]. Ye et al.[21] used XGBoost, and Völzke et al.[64] used the

Bayesian network to build their model for predicting incident hypertension. However, neither study compared their model performance with conventional regression-based models. There have been only two studies[27,29] where both conventional regression-based models and machine learning-based models were developed simultaneously. Huang et al.[29] and Farran et al.[27] both created machine learning algorithms along with a conventional logistic regression model. Huang et al.[29] used AUC to assess their models' performance and found the artificial neural network's AUC (0.90±0.01) much higher than the logistic regression model's AUC (0.73±0.03). Farran et al.[27] used classification accuracy to assess their models' performance and found logistic regression had relatively similar accuracy (82.4) to other machine learning algorithms (82.4±0.6 for support vector machines, 80.0±0.8 for the k-Nearest neighbors, and 80.9 for multifactor dimensionality reduction). Nevertheless, none of the studies considered survival data in their modeling.

We employed feature selection methods before model building and selected the top 20 features by five different methods. We noticed considerable variations in the top 20 features and adopted a strategy where features common in all top 20 features were included in model building. We believe selecting common features made our model robust.

The relative importance of the features in predicting hypertension incidence revealed that waist-hip ratio, cardiovascular disease, diabetes, SBP, age, and BMI are the essential features. There are apparent discrepancies in a feature's importance by different methods. DBP was identified as an important feature by RSF and GB. However, negligible importance was assigned for it in the penalized models. Perhaps this is due to its high collinearity with SBP, and penalized models tend to eliminate correlated features. Cardiovascular disease and diabetes were the two critical features identified in our study for predicting hypertension incidence, often avoided by

most studies. This is because participants with cardiovascular disease and diabetes are often excluded from the model building process in those studies.

This study's unique strength is comparing machine learning algorithms with the conventional regression-based Cox model to predict hypertension incidence using survival data. To the best of our knowledge, this is the first time a comparison between machine learning algorithms and conventional regression models has been performed to predict hypertension incidence in survival data. Using a large cohort data and considering many features is also a significant strength of this study. Notwithstanding the strengths, this study also has some limitations. The incidence rate of hypertension in our study was relatively low compared to what is reported for the general Alberta population[65]. There can be several potential reasons for that. The characteristics of the study participants in ATP may be different from the general Alberta population. For example, female participation in ATP data was more than double the male participation (69% vs. 31%), and the hypertension incidence rate in Alberta was much lower in females than the males in study age groups[65]. A potential selection bias also may lead to a lower incidence rate of hypertension in our study. A selection bias is an error associated with recruiting study participants or factors affecting the study participation and usually occurs when selecting participants is not random[66]. The participants in ATP were mainly selected using the volunteer sampling method[67]. Those who decided to join the study (i.e., who self-select into the survey) may have a different characteristic (e.g., healthier) than the non-participants. Due to the longitudinal nature of the study, there can also be a loss of study participants during follow-up. Participants who were lost to follow-up (e.g., due to emigration out of the province) may be more likely to develop hypertension. Our study ascertained outcome hypertension from a linked administrative health data (the hospital discharge abstract or physician claims data source) due to a lack of follow-

up information in ATP. There is a possibility that the outcome ascertainment was incomplete. People who did not have a healthcare encounter after cohort enrollment (e.g., did not visit a family physician/general practitioner or were not admitted to the hospital during the study period) were missed and can potentially lead to a lower hypertension incidence. We only compared C-index to evaluate the models' predictive performance. Although we tried to compare all the models with a standard performance measure, and C-index is the most commonly used predictive measure, considering other performance measures such as the Brier score could better compare the models' performance. We could not evaluate our models' performance in an external cohort, which is essential for any prediction model's generalizability. Considering additional machine learning algorithms such as artificial neural networks and survival support vector machines could make the comparison more sophisticated.

In conclusion, we developed several machine learning algorithms for predicting hypertension incidence using survival data. We compared machine learning algorithms' performance with conventional Cox PH regression models, and a negligible difference in predictive performance was observed. Based on this study's findings, conventional regression-based models are comparable to machine learning algorithms to provide good predictive accuracy in a moderate dataset with a reasonable number of features.

4.6 References

1. Lim SS, Vos T, Flaxman AD, et al. A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990-2010: A systematic analysis for the Global Burden of Disease Study 2010. *Lancet*. Published online 2012. doi:10.1016/S0140-6736(12)61766-8
2. World Health Organization. GLOBAL STATUS REPORT on noncommunicable diseases 2014 - "Attaining the nine global noncommunicable diseases targets; a shared responsibility" *WHO Libr Cat Data*. Published online 2014.
3. The Effects of Hypertension on the Body. Accessed January 2, 2021. https://www.healthline.com/health/high-blood-pressure-hypertension/effect-on-body
4. Ahmed I, Debray TP, Moons KG, Riley RD. Developing and validating risk prediction models in an individual participant data meta-analysis. *BMC Med Res Methodol*. Published online 2014. doi:10.1186/1471-2288-14-3
5. Steyerberg EW, van der Ploeg T, Van Calster B. Risk prediction with machine learning and regression methods. *Biometrical J*. Published online 2014. doi:10.1002/bimj.201300297
6. Wang P, Li Y, Reddy CK. Machine Learning for Survival Analysis: A Survey. *arXiv*. Published online 2017.
7. Hastie T, Tibshirani R, Friedman J. *Elements of Statistical Learning 2nd Ed.*; 2009.
8. Framingham T, Study H. Article Annals of Internal Medicine A Risk Score for Predicting Near-Term Incidence of Hypertension : 2017;(2).
9. Kanegae H, Oikawa T, Suzuki K, Okawara Y, Kario K. Developing and validating a new precise risk-prediction model for new-onset hypertension: The Jichi Genki hypertension

prediction model (JG model). *J Clin Hypertens*. 2018;20(5):880-890. doi:10.1111/jch.13270

10. Chen Y, Wang C, Liu Y, et al. Incident hypertension and its prediction model in a prospective northern urban Han Chinese cohort study. *J Hum Hypertens*. 2016;30(12):794-800. doi:10.1038/jhh.2016.23
11. Lim NK, Son KH, Lee KS, Park HY, Cho MC. Predicting the Risk of Incident Hypertension in a Korean Middle-Aged Population: Korean Genome and Epidemiology Study. *J Clin Hypertens*. 2013;15(5):344-349. doi:10.1111/jch.12080
12. Pearson TA, LaCroix AZ, Mead LA, Liang KY. The prediction of midlife coronary heart disease and hypertension in young adults: The Johns Hopkins multiple risk equations. *Am J Prev Med*. 1990;6(2 SUPPL.):23-28. doi:10.1016/s0749-3797(19)30122-9
13. Paynter NP, Cook NR, Everett BM, Sesso HD, Buring JE, Ridker PM. Prediction of Incident Hypertension Risk in Women with Currently Normal Blood Pressure. *Am J Med*. 2009;122(5):464-471. doi:10.1016/j.amjmed.2008.10.034
14. Zhang W, Wang L, Chen Y, Tang F, Xue F, Zhang C. Identification of hypertension predictors and application to hypertension prediction in an urban Han Chinese population: A longitudinal study, 2005-2010. *Prev Chronic Dis*. 2015;12(10):1-10. doi:10.5888/pcd12.150192
15. Wang B, Liu Y, Sun X, et al. Prediction model and assessment of probability of incident hypertension: the Rural Chinese Cohort Study. *J Hum Hypertens*. Published online 2020. doi:10.1038/s41371-020-0314-8
16. Otsuka T, Kachi Y, Takada H, et al. Development of a risk prediction model for incident hypertension in a working-age Japanese male population. *Hypertens Res*. 2015;38(6):419-

425. doi:10.1038/hr.2014.159

17. Kadomatsu Y, Tsukamoto M, Sasakabe T, et al. A risk score predicting new incidence of hypertension in Japan. *J Hum Hypertens*. 2019;33(10):748-755. doi:10.1038/s41371-019-0226-7

18. Sakr S, Elshawi R, Ahmed A, et al. Using machine learning on cardiorespiratory fitness data for predicting hypertension: The Henry Ford exercise testing (FIT) Project. *PLoS One*. 2018;13(4):1-18. doi:10.1371/journal.pone.0195344

19. Kwong EWY, Wu H, Pang GKH. A prediction model of blood pressure for telemedicine. *Health Informatics J*. 2018;24(3):227-244. doi:10.1177/1460458216663025

20. Falk CT. Risk factors for coronary artery disease and the use of neural networks to predict the presence or absence of high blood pressure. *BMC Genet*. 2003;4 Suppl 1:1-6. doi:10.1186/1471-2156-4-s1-s67

21. Ye C, Fu T, Hao S, et al. Prediction of incident hypertension within the next year: Prospective study using statewide electronic health records and machine learning. *J Med Internet Res*. 2018;20(1). doi:10.2196/jmir.9268

22. Priyadarshini R, Barik RK, Dubey H. DeepFog: Fog computing-based deep neural architecture for prediction of stress types, diabetes and hypertension attacks. *Computation*. 2018;6(4). doi:10.3390/computation6040062

23. Wu TH, Kwong EWY, Pang GKH. Bio-medical application on predicting systolic blood pressure using neural networks. *Proc - 2015 IEEE 1st Int Conf Big Data Comput Serv Appl BigDataService 2015*. Published online 2015:456-461. doi:10.1109/BigDataService.2015.54

24. Wu TH, Pang GKH, Kwong EWY. Predicting systolic blood pressure using machine

learning. *2014 7th Int Conf Inf Autom Sustain "Sharpening Futur with Sustain Technol ICIAfS 2014*. Published online 2014:1-6. doi:10.1109/ICIAFS.2014.7069529

25. Tayefi M, Esmaeili H, Saberi Karimian M, et al. The application of a decision tree to establish the parameters associated with hypertension. *Comput Methods Programs Biomed*. 2017;139:83-91. doi:10.1016/j.cmpb.2016.10.020

26. Zhang B, Wei Z, Ren J, Cheng Y, Zheng Z. An Empirical Study on Predicting Blood Pressure Using Classification and Regression Trees. *IEEE Access*. 2018;6(January):21758-21768. doi:10.1109/ACCESS.2017.2787980

27. Farran B, Channanath AM, Behbehani K, Thanaraj TA. Predictive models to assess risk of type 2 diabetes, hypertension and comorbidity: Machine-learning algorithms and validation using national health data from Kuwait-a cohort study. *BMJ Open*. 2013;3(5):1-10. doi:10.1136/bmjopen-2012-002457

28. Polak S, Mendyk A. Artificial neural networks based Internet hypertension prediction tool development and validation. *Appl Soft Comput J*. 2008;8(1):734-739. doi:10.1016/j.asoc.2007.06.001

29. Huang S, Xu Y, Yue L, et al. Evaluating the risk of hypertension using an artificial neural network method in rural residents over the age of 35 years in a Chinese area. *Hypertens Res*. 2010;33(7):722-726. doi:10.1038/hr.2010.73

30. Kruppa J, Liu Y, Biau G, et al. Probability estimation with machine learning methods for dichotomous and multicategory outcome: Theory. *Biometrical J*. Published online 2014. doi:10.1002/bimj.201300068

31. Van Hoorde K, Van Huffel S, Timmerman D, Bourne T, Van Calster B. A spline-based tool to assess and visualize the calibration of multiclass risk predictions. *J Biomed Inform*.

Published online 2015. doi:10.1016/j.jbi.2014.12.016

32. Desai RJ, Wang S V., Vaduganathan M, Evers T, Schneeweiss S. Comparison of Machine Learning Methods With Traditional Models for Use of Administrative Claims With Electronic Medical Records to Predict Heart Failure Outcomes. *JAMA Netw open*. 2020;3(1):e1918962. doi:10.1001/jamanetworkopen.2019.18962

33. Austin PC, Tu J V., Ho JE, Levy D, Lee DS. Using methods from the data-mining and machine-learning literature for disease classification and prediction: A case study examining classification of heart failure subtypes. *J Clin Epidemiol*. Published online 2013. doi:10.1016/j.jclinepi.2012.11.008

34. Tollenaar N, van der Heijden PGM. Which method predicts recidivism best?: A comparison of statistical, machine learning and data mining predictive models. *J R Stat Soc Ser A Stat Soc*. Published online 2013. doi:10.1111/j.1467-985X.2012.01056.x

35. Song X, Mitnitski A, Cox J, Rockwood K. Comparison of machine learning techniques with classical statistical models in predicting health outcomes. *Stud Health Technol Inform*. Published online 2004. doi:10.3233/978-1-60750-949-3-736

36. Frizzell JD, Liang L, Schulte PJ, et al. Prediction of 30-day all-cause readmissions in patients hospitalized for heart failure: Comparison of machine learning and other statistical approaches. *JAMA Cardiol*. Published online 2017. doi:10.1001/jamacardio.2016.3956

37. Summary Data Tables | Alberta's Tomorrow Project. Accessed December 15, 2020. http://myatp.ca/for-researchers/summary-data-tables

38. Quan H, Khan N, Hemmelgarn BR, et al. Validation of a case definition to define hypertension using administrative data. *Hypertension*. Published online 2009.

doi:10.1161/HYPERTENSIONAHA.109.139279

39. Van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Stat Med*. Published online 1999. doi:10.1002/(SICI)1097-0258(19990330)18:6<681::AID-SIM71>3.0.CO;2-R

40. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. *J Stat Softw*. Published online 2011. doi:10.18637/jss.v045.i03

41. Chandrashekar G, Sahin F. A survey on feature selection methods. *Comput Electr Eng*. Published online 2014. doi:10.1016/j.compeleceng.2013.11.024

42. Lang M, Kotthaus H, Marwedel P, Weihs C, Rahnenführer J, Bischl B. Automatic model selection for high-dimensional survival analysis. *J Stat Comput Simul*. Published online 2015. doi:10.1080/00949655.2014.929131

43. Cox DR. Regression Models and Life-Tables. *J R Stat Soc Ser B*. Published online 1972. doi:10.1111/j.2517-6161.1972.tb00899.x

44. Ishwaran H. Variable importance in binary regression trees and forests. *Electron J Stat*. Published online 2007. doi:10.1214/07-EJS039

45. Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K. Sparsity and smoothness via the fused lasso. *J R Stat Soc Ser B Stat Methodol*. Published online 2005. doi:10.1111/j.1467-9868.2005.00490.x

46. Tsamardinos I, Brown LE, Aliferis CF. The max-min hill-climbing Bayesian network structure learning algorithm. *Mach Learn*. Published online 2006. doi:10.1007/s10994-006-6889-7

47. Lagani V, Athineou G, Farcomeni A, Tsagris M, Tsamardinos I. Feature selection with the r package mxm: Discovering statistically equivalent feature subsets. *J Stat Softw*.

Published online 2017. doi:10.18637/JSS.V080.I07

48. van Houwelingen HC, Putter H. *Dynamic Prediction in Clinical Survival Analysis*.; 2011. doi:10.1201/b11311

49. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Ser B Stat Methodol*. Published online 2005. doi:10.1111/j.1467-9868.2005.00503.x

50. Breiman L. Random forests. *Mach Learn*. Published online 2001. doi:10.1023/A:1010933404324

51. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *Ann Appl Stat*. Published online 2008. doi:10.1214/08-AOAS169

52. Spooner A, Chen E, Sowmya A, et al. A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. *Sci Rep*. Published online 2020. doi:10.1038/s41598-020-77220-w

53. Friedman JH. Stochastic gradient boosting. *Comput Stat Data Anal*. Published online 2002. doi:10.1016/S0167-9473(01)00065-2

54. Max A, Wing J, Weston S, et al. Package ‘ caret ’ R. Published online 2020:223.

55. Zihni E, Madai VI, Livne M, et al. Opening the black box of artificial intelligence for clinical decision support: A study predicting stroke outcome. *PLoS One*. Published online 2020. doi:10.1371/journal.pone.0231166

56. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the Yield of Medical Tests. *JAMA J Am Med Assoc*. Published online 1982. doi:10.1001/jama.1982.03320430047030

57. Tsagris M, Papadovasilakis Z, Lakiotaki K, Tsamardinos I. Efficient feature selection on gene expression data: Which algorithm to use? *bioRxiv*. 2018;33(2):1-39.

doi:10.1101/431734

58. Jerome A, Hastie T, Tibshirani R, Tay K, Simon N. Package ‘ glmnet ’ R topics documented : Published online 2020.

59. Learning TM, Interface D, Bsd L, Url L, Paramhelpers D, Suggests XML. *Package ‘ Mlr . ’*; 2020.

60. Lumley T, S- R, Elizabeth A, Cynthia C, Therneau MTM. Package ‘ survival .’ Published online 2020.

61. Greenwell B, Boehmke B, Cunningham J. Package “gbm” - Generalized Boosted Regression Models. *CRAN Repos*. Published online 2019:39. https://cran.r-project.org/web/packages/gbm/gbm.pdf%0Ahttps://github.com/gbm-developers/gbm

62. Boosting TM, Matrix I. *Package ‘ Mboost . ’*; 2020.

63. Hosmer DW, Lemeshow S, Sturdivant RX. *Applied Logistic Regression: Third Edition*.; 2013. doi:10.1002/9781118548387

64. Völzke H, Fung G, Ittermann T, et al. A new, accurate predictive model for incident hypertension. *J Hypertens*. Published online 2013. doi:10.1097/HJH.0b013e328364a16d

65. Interactive Health Data Application - Display Results. Accessed March 29, 2021. http://www.ahw.gov.ab.ca/IHDA_Retrieval/selectSubCategoryParameters.do

66. Tripepi G, Jager KJ, Dekker FW, Zoccali C. Selection bias and information bias in clinical research. *Nephron - Clin Pract*. 2010;115(2). doi:10.1159/000312871

67. Ye M, Robson PJ, Eurich DT, Vena JE, Xu JY, Johnson JA. Cohort profile: Alberta’s Tomorrow Project. *Int J Epidemiol*. 2017;46(4):1097-1098l. doi:10.1093/ije/dyw256

Figure 4.1 Features ranked according to their importance by the different models

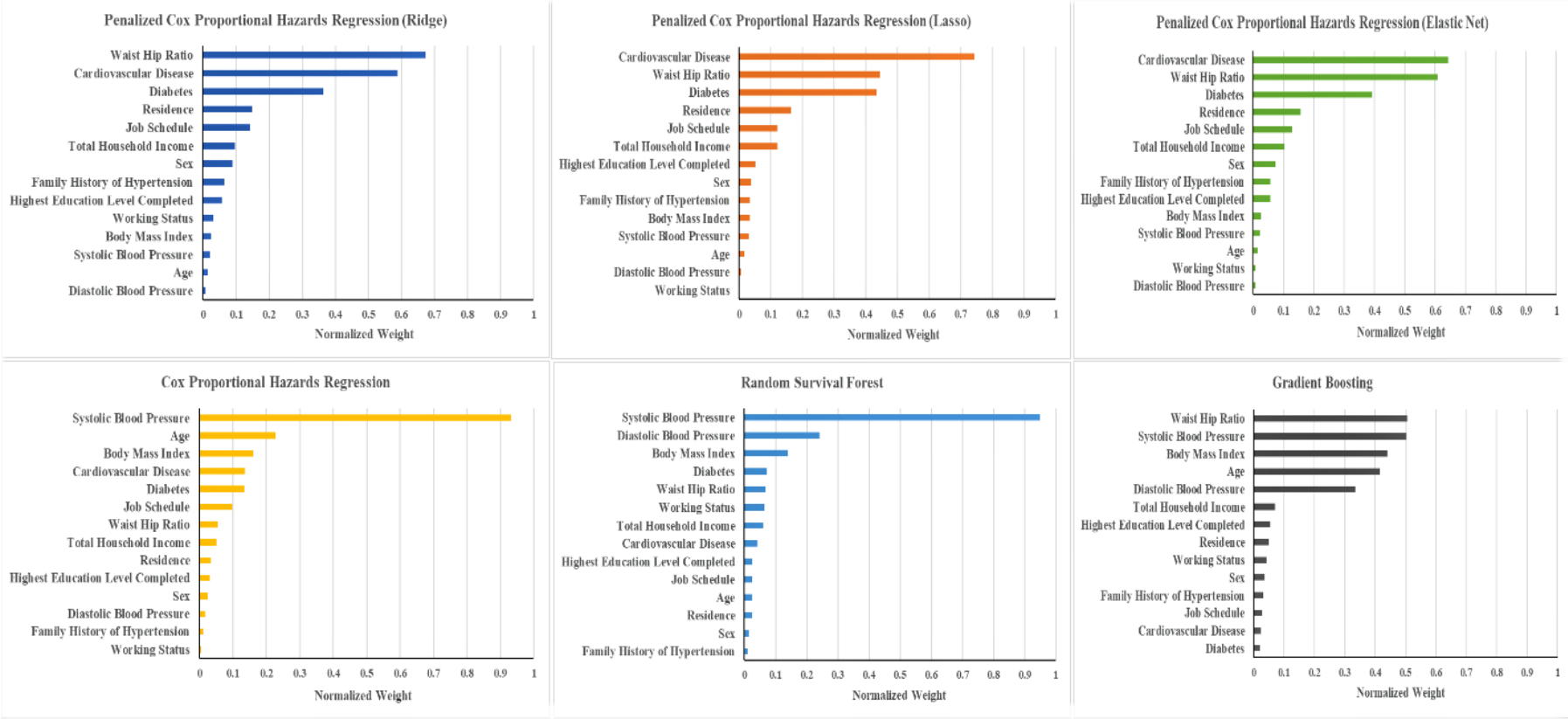

Figure 4.2 Boxplots showing the spread of values of the C-index produced by the different models

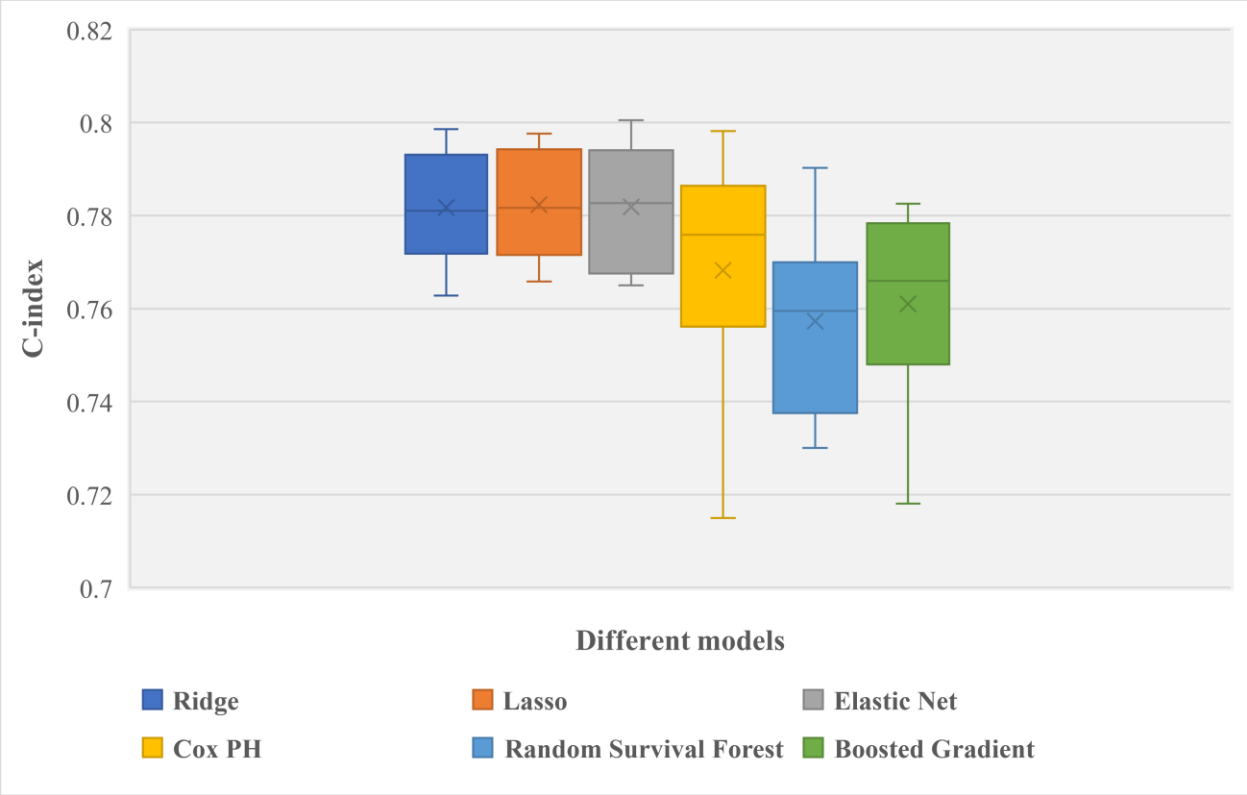

Table 4.1 Baseline characteristics of study participants and comparison of the training and test data

Socio-demographic characteristics of groups					
Variable	**Categories**	**All participants (18,322)**	**Derivation sample (n = 12,233)**	**Validation sample (n = 6,089)**	**P-value**
Age, years, mean (SE)		50.99 (9.20)	50.94 (9.19)	51.07 (9.24)	0.377
Sex, n (%)	Male	5,763 (31.45)	3844 (31.42)	1919 (31.52)	0.899
	Female	12,559 (68.55)	8389 (68.58)	4170 (68.48)	
Body Mass Index, kg/m2, mean (SE)		26.45 (4.90)	26.48 (4.94)	26.39 (4.81)	
Waist-Hip Ratio, mean (SE)		0.91 (0.07)	0.91 (0.07)	0.91 (0.07)	0.882
Diastolic Blood Pressure, mean (SE)		72.95 (9.35)	72.93 (9.35)	72.97 (9.34)	0.787
Systolic Blood Pressure, mean (SE)		119.81 (13.73)	119.75 (13.73)	119.92 (13.71)	0.446
Marital Status, n (%)	Married and/or living with a partner	14,458 (78.91)	9659 (78.96)	4799 (78.81)	0.226
	Single, never married	1180 (6.44)	763 (6.24)	417 (6.85)	
	Other (divorced, widowed, separated)	2684 (14.65)	1811 (14.80)	873 (14.34)	
Residence, n (%)	Urban	15,272 (83.35)	10,180 (83.22)	5092 (83.63)	0.484
	Rural	3050 (16.65)	2053 (16.78)	997 (16.37)	
Total Household Income, n (%)	< $49,999	2855 (15.58)	1904 (15.56)	951 (15.62)	0.416
	$50,000 - $99,999	5889 (32.14)	3902 (31.90)	1987 (32.63)	
	$100,000 - $199,999	7149 (39.02)	4823 (39.43)	2326 (38.20)	
	≥ $200,000	2429 (13.26)	1604 (13.11)	825 (13.55)	
Highest Education Level Completed, n (%)	High school or below (none, elementary school, high school, trade, technical or vocational school, apprenticeship training or technical CEGEP)	6161 (33.63)	4073 (33.30)	2088 (34.29)	0.310
	Diploma but below bachelor's degree (diploma from a community college, pre-university CEGEP or non-university certificate,	4928 (26.90	3288 (26.88)	1640 (26.93)	

	university certificate below bachelor's level)				
	Bachelor's degree or above (bachelor's degree, graduate degree (MSc, MBA, MD, PhD, etc.))	7233 (39.48)	4872 (39.83)	2361 (38.77)	
Ethnicity, n (%)	Aboriginal	68 (0.37)	49 (0.40)	19 (0.31)	0.316
	Asian (South Asian, East Asian, South East Asian, Filipino, West Asian, Arab)	827 (4.51)	545 (4.46)	282 (4.63)	
	White	16,895 (92.21)	11,274 (92.16)	5621 (92.31)	
	Latin American Hispanic	162 (0.88)	121 (0.99)	41 (0.67)	
	Black	97 (0.53)	63 (0.52)	34 (0.56)	
	Other (Jewish and others)	273 (1.49)	181 (1.48)	92 (1.51)	
Diabetes, n (%)		735 (4.01)	502 (4.10)	233 (3.83)	0.368
Cardiovascular Disease, n (%)		377 (2.06)	257 (2.10)	120 (1.97)	0.559
Depression, n (%)		2013 (10.99)	1366 (11.17)	647 (10.63)	0.270
Family History of Hypertension, n (%)		10,946 (59.74)	7266 (59.40)	3680 (60.44)	0.176
Smoking Status, n (%)	Never	10,116 (55.21)	6739 (55.09)	3377 (55.46)	0.763
	Former	6763 (36.91)	4537 (37.09)	2226 (36.56)	
	Current	1443 (7.88)	957 (7.82)	486 (7.98)	
Alcohol Consumption, n (%)	Never	1293 (7.06)	869 (7.10)	424 (6.96)	0.855
	≤ 1 time a week	9644 (52.64)	6415 (52.44)	3229 (53.03)	
	2 to 3 times a week	3807 (20.78)	2535 (20.72)	1272 (20.89)	
	4 to 5 times a week	1993 (10.88)	1340 (10.95)	653 (10.72)	
	≥ 6 times a week	1585 (8.65)	1074 (8.78)	511 (8.39)	
Working Status, n (%)	Full time	10,281 (56.11)	6836 (55.88)	3445 (56.58)	0.065
	Part time	3719 (20.30)	2543 (20.79)	1176 (19.31)	
	Other (looking after home, disable/sick, student, unpaid/voluntary)	3974 (21.69)	2614 (21.37)	1360 (22.34)	
	Unemployed	348 (1.90)	240 (1.96)	108 (1.77)	
Total Sleep Time, n (%)	≤ 5 hours (short sleep duration)	1191 (6.50)	804 (6.57)	387 6.36	0.257
	6 hours	3739 (20.41)	2441 (19.95)	1298 (21.32)	
	7 hours (reference)	7042 (38.43)	4747 (38.80)	2295 (37.69)	
	8 hours	5111 (27.90)	3414 (27.91)	1697 (27.87)	

	≥ 9 hours (long sleep duration)	1239 (6.76)	827 (6.76)	412 (6.77)	
Total Physical Activity Time, mean (SE)		3158.53 (2869.02)	3157.97 (2853.36)	3159.66 (2900.45)	0.970
Total Sitting Time, mean (SE)		2487.77 (1174.02)	2495.39 (1176.80)	2472.48 (1168.35)	0.214
Physical Activity, n (%)	Low (first quartile of physical activity time and fourth quartile of sitting time)	1691 (9.23)	1157 (9.46)	534 (8.77)	0.280
	Moderate (second and third quartile of physical activity time and sitting time)	14,479 (79.03)	9653 (78.91)	4826 (79.26)	
	High (fourth quartile of physical activity and first quartile of sitting time)	2152 (11.75)	1423 (11.63)	729 (11.97)	
Vegetable and Fruit Consumption, n (%)	Low consumption (less than 5 servings of vegetable and fruit)	15,273 (83.36)	10,182 (83.23)	5091 (83.61)	0.620
	Moderate consumption (less than 5 servings of vegetable but more than 5 servings of fruit OR more than 5 servings of vegetable but less than 5 servings of fruits)	2529 (13.80)	1694 (13.85)	835 (13.71)	
	High consumption (5 or more servings of vegetable and fruit)	520 (2.84)	357 (2.92)	163 (2.68)	
Job Schedule, n (%)	Regular daytime shift	11,920 (65.06)	7985 (65.27)	3935 (64.62)	0.385
	Other (evening shift, night shift, rotating shift, split shift, irregular shift, or on call)	6402 (34.94)	4248 (34.73)	2154 (35.38)	

Table 4.2 Baseline characteristics of study participants according to the status of developing hypertension or not

Socio-demographic characteristics of groups					
Variable	**Categories**	**All participants (18,322)**	**Participants who developed hypertension (n = 625)**	**Participants who did not develop hypertension (n = 17,697)**	**P-value**
Age, years, mean (SE)		50.99 (0.07)	53.99 (0.35)	50.88 (0.07)	< 0.001
Sex, n (%)	Male	5763 (31.45)	250 (40)	5513 (31.15)	< 0.001
	Female	12,559 (68.55)	375 (60)	12,184 (68.85)	
Body Mass Index, kg/m2, mean (SE)		26.45 (0.04)	28.63 (0.21)	26.38 (0.04)	
Waist Hip Ratio, mean (SE)		0.9093 (0.0006)	0.9363 (0.0033)	0.9085 (0.0006)	< 0.001
Diastolic Blood Pressure, mean (SE)		72.96 (0.08)	78.43 (0.47)	72.78 (0.08)	< 0.001
Systolic Blood Pressure, mean (SE)		119.71 (0.11)	132.36 (0.67)	119.40 (0.12)	< 0.001
Marital status, n (%)	Married and/or living with a partner	14,457 (78.91)	488 (78.08)	13,969 (78.94)	0.146
	Single, never married	1180 (6.44)	32 (5.12)	1148 (6.49)	
	Other (divorced, widowed, separated)	2685 (14.65)	105 (16.8)	2580 (14.57)	
Residence, n (%)	Urban	15,272 (83.35)	428 (68.48)	14,844 (83.88)	0.146
	Rural	3050 (16.65)	197 (31.52)	2853 (16.12)	
Total Household Income, n (%)	< $49,999	2800 (15.28)	178 (28.56)	2627 (14.84)	< 0.001
	$50,000 - $99,999	5912 (32.27)	229 (36.68)	5690 (32.15)	
	$100,000 - $199,999	7174 (39.16)	177 (28.27)	6986 (39.48)	
	≥ $200,000	2436 (13.29)	41 (6.49)	2394 (13.52)	
Highest Education Level Completed, n (%)	High school or below (none, elementary school, high school, trade, technical or vocational school, apprenticeship training or technical CEGEP)	6164 (33.64)	309 (49.35)	5854 (33.08)	< 0.001
	Diploma but below bachelor's degree (diploma from a community college, pre-university CEGEP or non-university certificate,	4926 (26.89)	163 (26.15)	4764 (26.92)	

	university certificate below bachelor's level)				
	Bachelor's degree or above (bachelor's degree, graduate degree (MSc, MBA, MD, PhD, etc.))	7232 (39.47)	153 (24.49)	7079 (40.0)	
Ethnicity, n (%)	Aboriginal	68 (0.37)	1 (0.16)	67 (0.38)	0.349
	Asian (South Asian, East Asian, South East Asian, Filipino, West Asian, Arab)	827 (4.51)	21 (3.4)	806 (4.55)	
	White	16,894 (92.21)	588 (94.03)	16,307 (92.14)	
	Latin American Hispanic	162 (0.89)	2 (0.32)	160 (0.9)	
	Black	97 (0.53)	2 (0.33)	95 (0.54)	
	Other (Jewish and others)	273 (1.49)	11 (1.76)	262 (1.48)	
Diabetes, n (%)		735 (4.01)	58 (9.28)	677 (3.83)	< 0.001
Cardiovascular Disease, n (%)		377 (2.06)	40 (6.4)	337 (1.9)	< 0.001
Depression, n (%)		2011 (10.98)	79 (12.64)	1932 (10.92)	0.179
Family History of Hypertension, n (%)		10,946 (59.74)	396 (63.36)	10,550 (59.61)	0.061
Smoking Status, n (%)	Never	10,107 (55.16)	290 (46.37)	9823 (55.51)	< 0.001
	Former	6773 (36.97)	276 (44.15)	6491 (36.68)	
	Current	1442 (7.87)	59 (9.48)	1383 (7.81)	
Alcohol Consumption, n (%)	Never	1279 (6.98)	56 (8.97)	1224 (6.92)	0.189
	≤ 1 time a week	9642 (52.63)	341 (54.52)	9307 (52.59)	
	2 to 3 times a week	3820 (20.85)	123 (19.77)	3689 (20.85)	
	4 to 5 times a week	1988 (10.85)	55 (8.74)	1938 (10.95)	
	≥ 6 times a week	1593 (8.69)	50 (8.0)	1539 (8.69)	
Working Status, n (%)	Full time	11,449 (62.49)	352 (56.29)	11,057 (62.48)	< 0.001
	Part time	4596 (25.09)	182 (29.19)	4422 (24.99)	
	Other (looking after home, disable/sick, student, unpaid/voluntary)	1857 (10.13)	83 (13.23)	1803 (10.18)	
	Unemployed	420 (2.29)	8 (1.28)	415 (2.35)	
Total Sleep Time, n (%)	≤ 5 hours (short sleep duration)	1192 (6.51)	47 (7.49)	1147 (6.48)	< 0.001
	6 hours	3732 (20.37)	127 (20.33)	3604 (20.37)	
	7 hours (reference)	7048 (38.46)	200 (32.02)	6847 (38.69)	
	8 hours	5115 (27.92)	185 (29.66)	4929 (27.85)	

	≥ 9 hours (long sleep duration)	1235 (6.74)	66 (10.49)	1170 (6.61)	
Total Physical Activity Time, mean (SE)		3159.83 (21.43)	3183.97 (126.52)	3157.58 (21.68)	0.825
Total Sitting Time, mean (SE)		2488.53 (8.92)	2389.16 (49.14)	2490.98 (9.38)	0.043
Physical Activity, n (%)	Low (first quartile of physical activity time and fourth quartile of sitting time)	1685 (9.19)	59 (9.47)	1678 (9.48)	0.707
	Moderate (second and third quartile of physical activity time and sitting time)	14,478 (79.02)	488 (78.12)	13,957 (78.87)	
	High (fourth quartile of physical activity and first quartile of sitting time)	2159 (11.78)	78 (12.40)	2062 (11.65)	
Vegetable and Fruit Consumption, n (%)	Low consumption (less than 5 servings of vegetable and fruit)	15,264 (83.31)	544 (87.05)	14,721 (83.18)	0.024
	Moderate consumption (less than 5 servings of vegetable but more than 5 servings of fruit OR more than 5 servings of vegetable but less than 5 servings of fruits)	2536 (13.84)	68 (10.84)	2469 (13.95)	
	High consumption (5 or more servings of vegetable and fruit)	522 (2.85)	13 (2.11)	507(2.87)	
Job Schedule, n (%)	Regular daytime shift	12,866 (70.22)	385 (61.59)	12,452 (70.36)	< 0.001
	Other (evening shift, night shift, rotating shift, split shift, irregular shift, or on call)	5456 (29.78)	240 (38.41)	5245 (29.64)	

Table 4.3 Feature's ranked based on five different approaches

Feature	Ranking based on Random Survival Forest Relative Importance	Ranking based on Statistical Equivalent Signature	Ranking based on Harrel's C-Index/Somers' Dxy Rank Correlation	Ranking based on Lasso Cox Coefficients/Variable Importance	Ranking based on Univariate Cox p-values
Systolic Blood Pressure	1	1	1	13	1
Diastolic Blood Pressure	2	20	2	15	5
Body Mass Index	3	2	3	11	3
Waist-Hip Ratio	4	11	5	1	4
Diabetes	5	5	14	3	10
Cardiovascular Disease	6	3	16	2	9
Age	7	4	4	14	2
Job Schedule	8	6	6	4	7
Working Status	9	8	7	19	8
Total Household Income,	10	7	9	6	6
Residence	11	13	10	5	12
Total Sleep Time	12	9	11	22	15
Highest Education Level Completed	13	12	8	10	11
Family History of Hypertension	14	17	18	12	16
Physical Activity, quartiles	15	19	22	21	23
Smoking Status	16	14	12	23	14
Total Physical Activity Time	17	24	15	16	17
Depression,	18	21	21	9	24
Ethnicity	19	10	24	18	21
Sex	20	18	13	8	13
Total Sitting Time	21	22	23	17	22
Alcohol Consumption	22	16	17	7	19
Marital Status	23	15	20	24	20
Vegetable and Fruit Consumption	24	23	19	20	18

Table 4.4 Top 20 features selected by the different approaches with red cells indicates commonly selected features

Top 20 Features				
Random Survival Forest Relative Importance	**Statistical Equivalent Signature**	**Harrel's C-Index/Somers' Dxy Rank Correlation**	**Lasso Cox Coefficients/Variable Importance Feature**	**Univariate Cox p-values**
Systolic Blood Pressure	Systolic Blood Pressure	Systolic Blood Pressure	Waist-Hip Ratio	Systolic Blood Pressure
Diastolic Blood Pressure	Body Mass Index	Diastolic Blood Pressure	Cardiovascular Disease	Age
Body Mass Index	Cardiovascular Disease	Body Mass Index	Diabetes	Body Mass Index
Waist-Hip Ratio	Age	Age	Job Schedule	Waist-Hip Ratio
Diabetes	Diabetes	Waist-Hip Ratio	Residence	Diastolic Blood Pressure
Cardiovascular Disease	Job Schedule	Job Schedule	Total Household Income	Total Household Income
Age	Total Household Income	Working Status	Alcohol Consumption	Job Schedule
Job Schedule	Working Status	Highest Education Level Completed	Sex	Working Status
Working Status	Total Sleep Time	Total Household Income	Depression	Cardiovascular Disease
Total Household Income	Ethnicity	Residence	Highest Education Level Completed	Diabetes
Residence	Waist-Hip Ratio	Total Sleep Time	Body Mass Index	Highest Education Level Completed
Total Sleep Time	Highest Education Level Completed	Smoking Status	Family History of Hypertension	Residence
Highest Education Level Completed	Residence	Sex	Systolic Blood Pressure	Sex
Family History of Hypertension	Smoking Status	Diabetes	Age	Smoking Status
Physical Activity, quartiles	Marital Status	Total Physical Activity Time	Diastolic Blood Pressure	Total Sleep Time
Smoking Status	Alcohol Consumption	Cardiovascular Disease	Total Physical Activity Time	Family History of Hypertension
Total Physical Activity Time	Family History of Hypertension	Alcohol Consumption	Total Sitting Time	Total Physical Activity Time
Depression	Sex	Family History of Hypertension	Ethnicity	Vegetable and Fruit Consumption
Ethnicity	Physical Activity, quartiles	Vegetable and Fruit Consumption	Working Status	Alcohol Consumption
Sex	Diastolic Blood Pressure	Marital Status	Vegetable and Fruit Consumption	Marital Status

Table S4.1 Missing information about different variables

Variables	Missing	Total	Percent Missing
Total Physical Activity Time	520	18,322	2.84
Total Sitting Time	1,421	18,322	7.76
Depression	16	18,322	0.09
Diabetes	8	18,322	0.04
Waist Hip Ratio	4,686	18,322	25.58
Sex	0	18,322	0.00
Age	0	18,322	0.00
Residence	0	18,322	0.00
Family History of Hypertension	0	18,322	0.00
Diastolic Blood Pressure	4,283	18,322	23.38
Systolic Blood Pressure	4,283	18,322	23.38
Ethnicity	23	18,322	0.13
Cardiovascular Disease	0	18,322	0.00
Highest Education Level Completed	11	18,322	0.06
Working Status	0	18,322	0.00
Vegetable and Fruit Consumption	266	18,322	1.45
Physical Activity	1,846	18,322	10.08
Total Household Income	1,402	18,322	7.65
Alcohol Consumption	846	18,322	4.62
Total Sleep Time	239	18,322	1.30
Smoking Status	45	18,322	0.25
Job Schedule	4,303	18,322	23.49
Marital Status	7	18,322	0.04
Body Mass Index	4,260	18,322	23.25
BMI Waist Ratio	4,718	18,322	25.75
Ever Smoked	41	18,322	0.22
Body Fat Percentage	4,471	18,322	24.40
Hip Circumference	4,564	18,322	24.91
Waist Circumference	4,769	18,322	26.03

CHAPTER 5. DISCUSSION

5.1 Overview of main findings

This study's overall objective was developing a prediction tool that is informative for patients and clinicians, providing a quantifiable and readily interpretable metric of an individual's risk for developing hypertension. Providing this information will aid patients in making treatment decisions and clinicians in providing treatment recommendations to patients. To achieve this goal, we searched the literature to explore the existing knowledge, incorporated knowledge that we gained in building a new prediction model and attempted to improve the model's predictive accuracy by applying some new analytical tools.

We presented below the main findings of this study.

5.1.1 Multiple prediction models exist but none in a Canadian context

The development of a risk prediction model often begins with a systematic review of the literature to identify existing models and their nature and get an idea about the model's set of candidate variables[1]. Performing a systematic review helped us identify existing hypertension prediction models, providing a comprehensive summary of these models and a list of risk factors considered in the model development. We identified 52 studies that presented 117 models predicting the risk of hypertension in the general adult population by searching four databases and grey literature. Of the models, 75 were developed using traditional regression-based modeling in 34 studies, and 42 using machine learning algorithms in 20 studies. Models were mostly developed either in white Caucasian or Asian populations. Continent-wise, the highest 28 studies developed models using the Asian population, followed by 14 using North American, 8 using European, and 1 using the South American population. No studies were from Africa and Oceania. Country-wise both USA and China had the highest 14 studies each. Among other countries, five studies were from Korea, four from Japan, three from Iran, two from England and Turkey, and one each from

Sweden, India, Spain, Finland, Germany, Kuwait, and Brazil. No studies from Canada were identified where a hypertension risk prediction model was developed or validated.

The number of variables/risk factors considered to create the models ranged from 1 to 19 in traditional regression-based models and from 2 to 169 in machine learning algorithms. However, the median risk factors per model were seven, both in regression-based and machine learning algorithms. Age was the most common risk factor, considered in 86 models, followed by BMI (39 models), DBP (34 models), SBP (31 models), and sex (29 models). Diabetes and cardiovascular disease (CVD) are the two important risk factors for hypertension, excluded by most studies. Individuals who have diabetes or CVD are expected to develop hypertension more than those free of these conditions. Most of the models excluded participants who were with diabetes or CVD during model building. If the intention is to build a model for the general adult population, excluding people with diabetes and CVD would limit the models' generalizability.

5.1.2 Similar predictive performance in existing traditional and machine learning-based models identified through meta-regression

Performing a meta-analysis helped us synthesize the evidence of existing hypertension prediction models' overall predictive performance. The meta-analysis of model discrimination, which was typically assessed using the C-statistic (also known as the area under the receiver operating characteristic curve), has provided us information about the model's predictive performance. We did not perform a meta-analysis of the total O/E ratio, a rough measure of overall model calibration, due to the unavailability of relevant data. We classified identified models into two categories--traditional regression-based models and machine learning-based models—due to their inherent differences and assessed each category separately. The traditional regression-based modeling approach is still dominating in predicting hypertension.

The overall pooled C-statistics of the traditional regression-based models and the machine learning-based models were almost similar (0.75 versus 0.76). The 95% approximate prediction interval for the overall C-statistics was also observed similar (0.63-0.84). In both categories, high heterogeneity in models' discriminative performance was observed. Stratified analysis by modeling methodology (e.g., logistic, Cox) within traditional regression-based models did not show much difference in predictive performance, and heterogeneity was still there within different modeling methods. A similar stratified pooled analysis within machine learning-based models was not performed due to diversity in machine learning algorithms' modeling method. Meta-regression, based on various study characteristics, was performed to identify potential heterogeneity sources. The participants' age, sex, and the number of risk factors considered in the model were determined C-statistic's potential sources of high heterogeneity in traditional regression-based models. However, the sources of heterogeneity were left unidentified in machine learning algorithms.

Machine learning algorithms are renowned for providing more accurate predictive performance. As such, we assumed models developed using machine learning algorithms would demonstrate better predictive performance than the traditional regression-based models. However, our meta-analysis did not support the evidence of a difference in predictive performance between these two categories of models.

5.1.3 Limitations of current models

The quality of the studies assessed by PROBAST[2,3] identified many of the studies failed to meet the criteria under the "analysis" domain of risk of bias. Consequently, the risk of bias was observed as "high" or 'unclear" in a large portion of studies. Due to lack of fulfilling the "participants" criteria properly, overall, the applicability of the models was also observed as "high

concern" or "unclear concern" in many studies. Several models were developed focusing on a specific population, making them inappropriate for the general adult population.

We identified many hypertension prediction models to serve; however, only four were externally validated, and only one had multiple validations. External validity establishes the generalizability of a prediction model. Generally, the accuracy of a prediction model degrades from the sample in which the model was first developed to subsequent application. For a prediction model to be generalizable, its accuracy needs to be reproducible and transportable. A prediction model that cannot predict outcomes accurately in a new sample is useless. Clinicians did not find confidence and trust to use prediction models in their practice that are not well validated. Despite its importance being recognized, external validation of prediction models is not common, which has primarily contributed to the failure to translate hypertension prediction models into clinical practice.

For a prediction model to be useful in clinical practice, it is crucial that its end-users (clinicians and patients) easily comprehend how the model works and can adequately communicate its results with each other. Models developed can be converted into a risk score to serve this purpose and simplify the tedious calculation of prediction models. We identified only eight models that were converted into a risk score after model development. A risk score needs to be provided when the models are developed to aid in interpreting risk estimates.

Studies assessing the impact of adopting hypertension risk prediction models in clinical settings was also absent. A prediction model with an impact study to evaluate whether the model improves clinical decision-making and patient health outcomes is ideal but lacks reality. Impact studies can help identify factors (ease of use, acceptability) that can affect the implementation of prediction models in clinical practice.

5.1.4 New prediction model for hypertension incidence in Canadian context using large cohort data

The lack of a hypertension prediction model in a Canadian context motivated us to develop a new model. We developed a new hypertension incidence prediction model using large Canadian ATP cohort data. To obtain follow-up information, ATP data was linked to Alberta's administrative health data. Eighteen thousand three hundred twenty-two participants aged 35-69 years without hypertension at baseline from ATP were followed (median follow-up 5.80 years) for hypertension incidence, and 625 new hypertension cases were identified. The sample was randomly divided into derivation and validation sets at a 2:1 ratio. The model was developed in the derivation sample. We used the standard Cox PH model to create the model. While developing the new model, we followed the necessary steps required to build a prediction model properly.

We identified a large set of candidate variables based on literature search and expert opinion. A total of 29 candidate variables were compiled available in ATP data. We dealt with missing values of the variables by substituting imputed values produced by the multiple imputation techniques. On a couple of variables, missing values were up to 26%. Complete case analysis, instead of substituting missing values, would reduce our sample size to one fourth. Collinearity among the risk factors was assessed using VIF, and highly correlated variables were removed before model building to obtain stable estimates. The linearity of the continuous variables was evaluated using fractional polynomial, and no issues were detected. Cox proportionality assumption was assessed to check violation of assumptions. Only the variables identified as significant in univariate association at $p < 0.20$ were further considered from the set of candidate variables. Significant variables identified in univariate associations were put in a multivariable model, and variables significant at $p < 0.05$ were regarded as final risk factors. Nevertheless, we

forced the variable sex into the model due to its clinical relevance with hypertension despite being statistically insignificant. Within finally selected variables, potential interaction was assessed. Several interaction terms were identified as significant. However, the inclusion of those interaction terms did not improve the predictive ability of the model significantly. Consequently, we dropped the interaction terms from the final model. Our final model consists of age, BMI, SBP, diabetes, total physical activity time, cardiovascular disease, and sex.

5.1.5 Overall good predictive performance of the newly developed model

We assessed the predictive performance of the newly developed model using various measures in the validation data. When we applied our derived model in the validation data, the model's discriminative performance was good, as assessed by Harrel's C-statistic 0.77. The GB test results indicated a good calibration of the model (χ^2 statistic 8.75, p = 0.07). The model's calibration was also presented graphically using Arjas like plot and calibration plot and was observed decent. These plots helped assess calibration visually by comparing the observed and expected events in each group based on the risk score. A calibration slope of 1.006 indicated that predicted probabilities do not vary enough. The prognostic index histogram in derivation and validation data also did not reveal obvious irregularities and outliers. Brier scores calculated at 4-year, 5-year, 6-year, and 7-year time points were: 0.018, 0.021, 0.026, and 0.029, respectively, indicated accurate predictions.

5.1.6 Deriving risk score from the newly developed model for clinical utility

To facilitate the use of our newly developed model in clinical practice, a user-friendly and straightforward risk score from the developed model was created to calculate the risk of incident hypertension at different times (2-year, 3-year, 5-year, and 6-year). An algorithm[4] was followed to prepare the point scoring system. The process involved several steps and started organizing the

risk factors into categories and determining each variable's baseline category and reference values. It was then determined how far each category was from the reference category in regression units, and a base constant (the number of regression units that reflects one point in the point scoring system) was set. Next, the number of points for each category of a variable was determined. It was computed by dividing how far each category was from the reference category in regression units by the base constant. Then the created final points were rounded to the nearest integers. Finally, risk categories were created according to the total score, and patients were classified according to their total score into different risk categories.

5.1.7 Developing some machine learning-based models for hypertension incidence using the same survival data

Machine learning algorithms, an alternative class of models, emerged as a popular modeling approach and, due to their superiority, achieved significant successes across a broad range of fields. Machine learning algorithms have a reputation for delivering better accuracy in predicting outcomes. Due to the lack of use of survival data in predicting hypertension in the machine learning domain, it was unclear how machine learning-based models will perform predicting hypertension in survival data. A formal comparison in predictive performance between conventional regression-based hypertension prediction models and machine learning-based models in a survival setting was also absent. There was also a scarcity of comparisons using the same dataset. These motivated us to develop machine learning algorithms and compare their predictive performance with conventional regression-based models in a survival setting.

The same ATP data were used to develop machine learning algorithms. Missing values were imputed using multiple imputations as before. Before creating the machine learning models, we first selected candidate features and then employed five feature selection methods to choose

the top 20 features. Feature selection methods included two filter-based: a univariate Cox p-value and C-index; two embedded-based: random survival forest and least absolute shrinkage and selection operator (Lasso); and one constraint-based: the statistically equivalent signature (SES). Due to considerable variations in the top 20 features, we adopted a strategy to choose only those features common in all top 20 features. Fourteen features were identified as common and were included in the final model building process. Hyper-parameters of different machine learning algorithms were tuned automatically within a 10-fold nested cross-validation loop.

Five machine learning algorithms were developed to predict hypertension incidence: penalized regression Ridge, Lasso, Elastic Net (EN), random survival forest (RSF), and gradient boosting (GB), along with the conventional Cox proportional hazards (PH) model. Moreover, the training data features were ranked according to their relative contribution to the prediction of hypertension incidence using various variable importance metrics.

5.1.8 Similar predictive performance in newly developed machine learning models and conventional model

Fourteen common features used in the model building included SBP, DBP, BMI, waist-hip ratio, diabetes, cardiovascular disease, age, job schedule, working status, total household income, residence, highest education level completed, family history of hypertension, and sex. The predictive performance of the models was assessed using C-index. A negligible difference in the predictive accuracy between machine learning and conventional regression-based Cox models was observed. The average C-index for the machine learning algorithms Ridge, Lasso, EN, RSF, and GB was 0.78, 0.78, 0.78, 0.76, and 0.76, respectively. In comparison, the conventional regression-based Cox PH model's average C-index was 0.77.

Regarding feature importance, the waist-hip ratio was selected as the top feature by Ridge regression and GB. In contrast, cardiovascular disease was selected as the top feature by Lasso regression and EN regression; meanwhile, SBP was selected as the top feature by the Cox PH model and RSF. Waist-hip ratio, cardiovascular disease, diabetes, SBP, age, and BMI have been deemed the most important features considered by most modeling approaches. Nevertheless, there were also variations in the rank ordering of important features across the investigated models.

This study's findings have shown that conventional regression-based models are comparable to machine learning algorithms to provide good predictive accuracy in hypertension prediction in a moderate dataset with a reasonable number of features.

5.2 Strengths and Limitations

This study's overall goal was to develop a comprehensive hypertension risk prediction model in a Canadian context. The three specific objectives associated with the overall goal were: performing a systematic review and meta-analysis on hypertension prediction models, developing a new hypertension prediction model applying a traditional regression modeling approach, and developing machine learning algorithms for predicting hypertension risk, and compare their performance with the traditionally developed model. Each of these specific objectives has been reflected as a separate study and has been accomplished with some pros and cons. We discuss the strengths and limitations of each below one by one.

5.2.1 Systematic review and meta-analysis

One of our systematic review's strengths was the extent of the systematic search, which included four different databases, grey literature, and extensive use of the reference lists of the identified studies. Accordingly, there was little chance that any relevant studies would have been missed. This study was also unique in several ways to the best of our knowledge. This was the first

study in which a meta-analysis was carried out to synthesize the predictive performance of the hypertension risk prediction models along with the heterogeneity assessment. Comparing the overall predictive performance of traditional regression-based models and machine learning-based models in predicting hypertension was also exclusive. Moreover, performing a detailed critical appraisal of studies in hypertension risk prediction models was also exceptional.

Nevertheless, there were also limitations to the study. We excluded non-English and non-French publications. While it is widely perceived that the English language is the primary language of science, the choice of scientific results in a particular language can incorporate language bias and may lead to incorrect conclusions[5]. We could only use C-statistics to compare the model performance, which could be insensitive to distinguish a model's ability to stratify patients into clinically relevant risk groups correctly[5,6]. A meta-analysis of calibration measures (e.g., O/E ratio) along with C-statistics could provide a comprehensive summary of the performance of these models[7]. Failing to assess publication bias amongst the studies is another potential limitation of this study. Recent guidelines[7] did not emphasize the need to assess publication bias for prediction model performance, which encouraged us not to do so. Instead, we assessed ROB using the PROBAST[2,3] checklist.

5.2.2 A new traditionally developed hypertension prediction model

To our knowledge, this was the first hypertension risk prediction model developed explicitly in a Canadian population. Using a large sample size to create the model was a significant strength of this study. This ensured the stability of the prediction model estimates. Further, consideration of many candidate variables in the model building process was also a strength of this study. In contrast to most studies, where models were developed in complete cases excluding those

with missing values, we imputed missing values in our study. This approach prevented information loss, maximized information utilization, and made the results robust.

Our study had several limitations. Study participants were middle-aged and elderly Canadian. Prevention strategies are likely to be more effective if the young population can be targeted. Still, our study participants' age range will likely have minimal impact on our study's generalizability, as the people diagnosed with hypertension are generally ≥ 35 years of age[8]. At baseline, we excluded participants with self-reported hypertension, which can potentially lead to misclassification of hypertension status. Determining hypertension status with objective blood pressure measurement rather than relying on self-reported alone could better assemble the cohort and avoid potential misclassification. The incidence rate of hypertension in our study was relatively low compared to what is reported for the general Alberta population[9]. There can be several potential reasons for that. The characteristics of the study participants in ATP may be different from the general Alberta population. For example, female participation in ATP data was more than double the male participation (69% vs. 31%), and the hypertension incidence rate in Alberta was much lower in females than the males in study age groups[9]. A potential selection bias also may lead to a lower incidence rate of hypertension in our study. A selection bias is an error associated with recruiting study participants or factors affecting the study participation and usually occurs when selecting participants is not random[10]. The participants in ATP were mainly selected using the volunteer sampling method[11]. Those who decided to join the study (i.e., who self-select into the survey) may have a different characteristic (e.g., healthier) than the non-participants. Due to the longitudinal nature of the study, there can also be a loss of study participants during follow-up. Participants who were lost to follow-up (e.g., due to emigration out of the province) may be more likely to develop hypertension. Our study ascertained outcome hypertension from a linked

administrative health data (the hospital discharge abstract or physician claims data source) due to a lack of follow-up information in ATP. There is a possibility that the outcome ascertainment was incomplete. People who did not have a healthcare encounter after cohort enrollment (e.g., did not visit a family physician/general practitioner or were not admitted to the hospital during the study period) were missed and can potentially lead to a lower hypertension incidence. Competing risks occur when individuals experience one or more outcomes that compete with the outcome of interest[12]. It hinders the observation of the event of interest or modifies the chance that this event occurs. In our context, death could be a competing risk because if a person dies, it hinders the observation of hypertension, and the person who dies may also have a higher risk of hypertension. We did not account for competing risks in our study because the expected event (death) rate is low as the cohort was healthy and relatively young at inception with a short follow-up time. We did not include genetic risk factors or biomarkers in our model. The inclusion of genetic risk factors in the model had the potential of improving risk prediction. Nevertheless, our performed meta-analysis and previous studies[13] did not show any differences in discriminative performance when genetic risk factors were included in the model. Besides, genetic risk factors in the model may decrease the prediction model's application in routine clinical practice. Salt intake, a key dietary factor for the risk of incident hypertension; however, data on salt intake were not available in our study. We could not perform an external validation of our model, essential for any prediction model's generalizability. Therefore, further validation of our model in other populations, particularly in another Canadian jurisdiction, is warranted.

5.2.3 Machine learning-based hypertension prediction models

This study's unique strength was comparing machine learning algorithms with the conventional regression-based Cox model to predict hypertension incidence using survival data.

Comparing machine learning algorithms with traditional regression models to predict hypertension incidence using survival data was the first time to the best of our knowledge. The utilization of extensive cohort data and consideration of many features is also this study's significant strengths.

Notwithstanding the strengths, this study also had some limitations. As outlined earlier, a lower incidence rate of hypertension and failure to handle potential reasons associated with the lower incidence rate can be considered a limitation of this study. We only compared C-index to evaluate models' predictive performance. Although we intended to compare all the models with a standard performance measure and C-index is the standard and most used predictive measure, considering other performance measures such as the Brier score could make the comparison more comprehensive. We could not evaluate our models' performance in an external cohort, which is essential for prediction models' generalizability. Consideration of additional machine learning algorithms such as artificial neural networks and survival support vector machines could make the comparison more elaborate.

5.3 Future Directions

Based on this study's findings, there are a few directions that would be worth further investigation.

5.3.1 External validation

The reliability and acceptability of a prediction model largely depend on how well it performs in a validation cohort outside of the derivation cohort where the model was developed. Internal validation of prediction models is often not sufficient for generalizability, and external validation is necessary before implementing prediction models in clinical practice. External validation requires data collected from a similar group of patients in a different setting. It aims to address a prediction model's accuracy and performance in patients from a different but plausibly

related population. External validation of our newly developed hypertension prediction model needs to be performed in an external dataset to assess its performance for generalizability. The Canadian Partnership for Tomorrow Project (CPTP)[14], a Canada-wide prospective cohort study, can be a potential data source for this purpose.

5.3.2 Developing a computer-assisted tool

For a prediction model to be helpful in clinical practice, it is crucial that its end-users (clinicians and patients) easily comprehend how the model works and can adequately communicate its results with each other. A typical representation of a predictive model is non-intuitive and requires an alternative presentation that is discernable so that its users can easily understand it. The development of a computerized electronic interactive version of the risk score is one such possibility. A web-based version of the risk score that is easily downloadable to a computer or mobile phone and can be accessed by physicians and non-physician health workers can quickly identify those at high risk of hypertension. Such a tool would be handy and designed to support clinicians for quick and consistent estimation of hypertension risk in the general population. We can develop such a computerized automatic tool for our hypertension prediction model that can smoothly be adopted in routine clinical practice.

5.3.3 Updating model using meta-modeling

Understanding and quantifying the already reported estimates opens the possibility of incorporating those models' performance characteristics into a newly developed hypertension prediction model to improve hypertension's overall prediction. Using a meta-model updating technique, we can enhance the newly developed hypertension prediction model by incorporating parameters derived from the existing hypertension prediction models. The meta-model updating approach works from the 'middle ground' in which current prediction models that may be relevant

for the population and endpoint of interest are used and revised to suit the new population[15]. The updated model is then based on both the new and existing data, further improving its performance in the new population. There are different approaches for updating a prediction model considering the latest data: regression coefficients updating, meta-model updating, and dynamic updating[15], and any of them can be employed. The application of the meta-model updating technique in prediction research is still in its early stage[16,17]. However, those who have applied the concept have found it very successful in accurate outcome prediction[18].

5.3.4 Constructing a multi-disease prediction model

Abnormalities in physiological indicators may indicate not only a single disease but also multiple diseases. Therefore, determining the common risk factors and developing a prediction model for multiple diseases (e.g., hypertension and hyperlipidemia) can be more important than doing so for only a single disease. A two-phase analysis procedure to simultaneously predict multiple diseases can be applied. In the first phase, individual risk factors for each disease will be selected and combined to determine the common risk factors for both diseases using voting principles. In the second phase, a statistical tool (e.g., the multivariate adaptive regression splines [MARS][19] method or multivariate logistic regression) can be applied to construct a multi-disease predictive model.

5.4 Conclusion

This study's overall objective was to develop a comprehensive hypertension risk prediction model in a Canadian context. We split the overall objective into three pieces to achieve our goal smoothly. Through the systematic review, we identified the existing hypertension prediction models, how they were developed, the risk factors considered in different models, and how the predictive accuracy varies in various types of models. These findings eventually helped us identify

a gap in the hypertension risk prediction models specific to the Canadian context. To fill these gaps, we developed a new hypertension incidence prediction model using extensive population-based Canadian data. The systematic review also helped us figure out the lack of machine learning models predicting hypertension incidence in survival context and a formal comparison with traditional regression-based models. These further motivated us to develop machine learning models for predicting hypertension incidence. We recognized no significant difference in the newly developed traditional Cox PH model and machine learning models' predictive performance. Consequently, we recommended proceeding with the traditional regression-based Cox PH model due to its easier interpretability. We converted it into a risk score to facilitate its use in the clinical setting. After successfully validating the model, this model can be implemented in daily clinical practice to support decision-making.

5.5 References

1. Pavlou M, Ambler G, Seaman SR, et al. How to develop a more accurate risk prediction model when there are few events. *BMJ*. Published online 2015. doi:10.1136/bmj.h3868
2. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. Published online 2019. doi:10.7326/M18-1376
3. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: A tool to assess risk of bias and applicability of prediction model studies: Explanation and elaboration. *Ann Intern Med*. Published online 2019. doi:10.7326/M18-1377
4. Sullivan LM, Massaro JM, D'Agostino RB. Presentation of multivariate data for clinical use: The Framingham Study risk score functions. *Stat Med*. 2004;23(10):1631-1660. doi:10.1002/sim.1742
5. Chowdhury MZI, Yeasmin F, Rabi DM, Ronksley PE, Turin TC. Predicting the risk of stroke among patients with type 2 diabetes: A systematic review and meta-analysis of C-statistics. *BMJ Open*. 2019;9(8):1-22. doi:10.1136/bmjopen-2018-025579
6. Cook NR, Ridker PM. Advances in measuring the effect of individual predictors of cardiovascular risk: The role of reclassification measures. *Ann Intern Med*. Published online 2009. doi:10.7326/0003-4819-150-11-200906020-00007
7. Debray TPA, Damen JAAG, Snell KIE, et al. A guide to systematic review and meta-analysis of prediction model performance. *BMJ*. Published online 2017. doi:10.1136/bmj.i6460
8. Hajjar I, Kotchen TA. Trends in Prevalence, Awareness, Treatment, and Control of Hypertension in the United States, 1988-2000. *J Am Med Assoc*. Published online 2003.

doi:10.1001/jama.290.2.199

9. Interactive Health Data Application - Display Results. Accessed March 29, 2021. http://www.ahw.gov.ab.ca/IHDA_Retrieval/selectSubCategoryParameters.do

10. Tripepi G, Jager KJ, Dekker FW, Zoccali C. Selection bias and information bias in clinical research. *Nephron - Clin Pract*. 2010;115(2). doi:10.1159/000312871

11. Ye M, Robson PJ, Eurich DT, Vena JE, Xu JY, Johnson JA. Cohort profile: Alberta's Tomorrow Project. *Int J Epidemiol*. 2017;46(4):1097-1098l. doi:10.1093/ije/dyw256

12. Noordzij M, Leffondré K, Van Stralen KJ, Zoccali C, Dekker FW, Jager KJ. When do we need competing risks methods for survival analysis in nephrology? *Nephrol Dial Transplant*. 2013;28(11):2670-2677. doi:10.1093/ndt/gft355

13. Paynter NP, Cook NR, Everett BM, Sesso HD, Buring JE, Ridker PM. Prediction of Incident Hypertension Risk in Women with Currently Normal Blood Pressure. *Am J Med*. 2009;122(5):464-471. doi:10.1016/j.amjmed.2008.10.034

14. Dummer TJB, Awadalla P, Boileau C, et al. The Canadian Partnership for Tomorrow Project: A pan-Canadian platform for research on chronic disease prevention. *CMAJ*. Published online 2018. doi:10.1503/cmaj.170292

15. Su TL, Jaki T, Hickey GL, Buchan I, Sperrin M. A review of statistical updating methods for clinical prediction models. *Stat Methods Med Res*. Published online 2018. doi:10.1177/0962280215626466

16. Debray TPA, Koffijberg H, Vergouwe Y, Moons KGM, Steyerberg EW. Aggregating published prediction models with individual participant data: A comparison of different approaches. *Stat Med*. Published online 2012. doi:10.1002/sim.5412

17. Hemingway H, Riley RD, Altman DG. Ten steps towards improving prognosis research.

BMJ. Published online 2010. doi:10.1136/bmj.b4184

18. Janssen KJM, Moons KGM, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. *J Clin Epidemiol*. Published online 2008. doi:10.1016/j.jclinepi.2007.04.018

19. Friedman JH. Multivariate Adaptive Regression Splines. *Ann Stat*. Published online 1991. doi:10.1214/aos/1176347963

APPENDIX 1.

Data Linkage

Individual datasets are often limited in scope, consequently limiting their utility in comprehensively addressing important questions[1]. Linking data from multiple sources can overcome some of the limitations[1]. Different information that is believed to be related to the same person or event can be connected through the data linkage techniques. Data linkage contains pairing observations from two or more files and identifying the pairs belonging to the same entity[2]. Collecting information on the same person from two datasets is a common form of linkage. Among many other advantages, data linkage allows the passive follow-up of study participants and improved measurement of risk factors and outcomes[1]. Data linkage from multiple sources is challenging because linkage errors can arise from multiple sources and privacy and confidentiality issues. To perform the data linkage, we first need to determine its necessity, confirm the data availability and check whether a unique identifier exists. If a unique identifier exists, such as a personal health number (PHN), linking is a simple operation. When a unique identifier is absent, linking is done by combining a range of identifiers, such as date of birth, name, address, etc. There are two main types/methods of data linkage algorithms: deterministic and probabilistic. The choice of method depends on many interacting factors, such as time, resources, the research question, and the quantity and quality of the variables available to link in the dataset[1].

Deterministic Linkage

The deterministic linkage can be of different types starting from a simple connection of two or more datasets with a single reliable and stable identifier to a sophisticated stepwise algorithmic linkage. A single identifier or linkage key is used in the deterministic linkage technique to join two or more datasets. Deterministic linkage requires a high degree of certainty, which can

be achieved if there is a unique identifier such as a PHN. The PHN uniquely identifies an individual across datasets. If this unique identifier exists in all datasets to be linked, it can connect an individual's records across those datasets. As deterministic linkage is based on exact matches, variables used in deterministic linkage need to be accurate, robust, stable over time, and complete. Examples of such variables are sex, date of birth, and first name and last name. Alternatively, a linkage key can be created using a combination of attributes such as last name, first name, sex, and date of birth, which can be used to match records with the same linkage key value[3]. This linkage key is known as a derived linkage key or statistical linkage key (SLK). Generally, most SLKs are constructed from last name, first name, sex, and full date of birth.

Stepwise deterministic record linkage, a more sophisticated form of deterministic linkage, is developed in response to variations that often exist in the attributes used in creating the linkage keys for deterministic linkage[4]. Auxiliary information on the datasets is used in stepwise deterministic linkage to provide a platform from which variation in the reported linkage key or SLK information can be captured[4]. This differs from simple deterministic linkage that relies on an exact, one-to-one character matching of linkage keys across two or more datasets. "Rules-based linkage" is another form of deterministic linkage where a set of rules are used to categorize pairs of records as matches or non-matches. Despite being more flexible than using a linkage key, rules-based linkage development is labor-intensive and overly reliant on the data sets to be linked[3].

Probabilistic Linkage

Probabilistic linkage is generally applied in the absence of a unique identifier or statistical linkage keys or when the linking variables or identifiers are not accurate, stable, or complete to perform the deterministic linkage. Attaining a sufficiently comparable value to unique identification using several identifying variables is the key in linking in the probabilistic linkage.

Individually, each of these variables serves as a partial identifier, but, in combination, they provide a reasonably accurate match for the intended purpose of linking datasets.

When errors exist in linking variables, the probabilistic linkage has a higher capacity to link and can provide better linkage than deterministic methods[5,6]. The deterministic approach's limitations include not considering certain identifiers or certain values having more discriminatory power than others. Probabilistic approaches have been developed to address these issues to evaluate 1) each identifier's discriminative ability and 2) the possibility that two records are a correct match based on whether they agree with the different identifiers.

In our study, the three data sources were linked through deterministic linkage using unique encrypted health numbers common to all three data sources. Data from the ATP cohort was used to define baseline predictors/variables. Data from hospital discharge abstract data and physician claims data were linked to identify diagnosed hypertension cases, our study's outcome. We then linked the diagnosed hypertension cases with the ATP cohort data to obtain follow-up information about the ATP participants who developed hypertension.

The ATP has performed the data linkage for us. The ATP retrieved data from external sources such as Alberta's administrative health data through DIMR (Data Integration, Measurement & Reporting) and then linked it before releasing it to us for further analysis.

Cohort Formation

The cohort was derived from the ATP cohort data. The cohort included all participants between 35-69 years of age at enrollment. This age range of the study participants will likely have minimal impact on our study's generalizability, as most of the people diagnosed with hypertension are ≥ 35 years of age[7]. Eligible subjects are free of hypertension at baseline and consented to have their data linked with Alberta's administrative health data.

Outcome: Hypertension Incidence

Our proposed study's outcome is the incidence of hypertension, which was determined from administrative health data. We used a coding algorithm to define diagnosed hypertension in administrative health data that refers to individuals who have a diagnostic code for hypertension in either the hospital discharge abstract or physician claims data source. The following steps were taken to define diagnosed hypertension:

Step 1. We initially identified patients with diagnosed hypertension using hospital discharge data and physician claims data. The relevant ICD-9 and ICD-10 codes (ICD-9-CM codes: 401.x, 402.x, 403.x, 404.x, and 405.x; ICD-10-CA/CCI codes: I10.x, I11.x, I12.x, I13.x, and I15.x) in the ≤ 25 coding fields for diagnosis in the hospital discharge data, and ≤ 3 fields in the physician claims data was used. We then applied the following validated hypertension case definition to these sources: two physician claims within 2 years or one hospital discharge for hypertension[8]. Incident cases were defined as any patient having diagnosed hypertension as defined above but not previously identified as such.

Step 2. We identified the first encounter when an individual meets the hypertension case definition algorithm in the study period and exclude subsequent encounters in the study years. After exclusion, each patient had one index diagnosis date. We did not consider those events related to patients with pregnancy-induced hypertension, defined as females with a hypertension diagnostic code and a physician service claim or hospital discharge record within five months (indicating an obstetrical event) as a hypertension outcome.

Step 3. Finally, we linked the index diagnosis with the ATP data (those ATP participants who consented to link their data with administrative health data).

References

1. Dusetzina SB, Tyree S, Meyer A-M, Meyer A, Green L, Carpenter WR. An overview of record linkage methods. *Link Data Heal Serv Res A Framew Instr Guid*. Published online 2014.

2. Valliant R, Scheuren F, Winglee M, Valliant R, Scheuren F, Valliant SA; R. *A Case Study in Record Linkage*.; 2005.

3. Christen P, Goiser K. Quality and complexity measures for data linkage and deduplication. *Stud Comput Intell*. Published online 2007. doi:10.1007/978-3-540-44918-8_6

4. Islander TS. *National Best Practice Guidelines*.; 2012.

5. NCSIMG. Statistical Data Linkage in Community Services Data Collections. 2004;(May):88.

6. Bloomrosen M, Detmer D. Advancing the Framework: Use of Health Data-A Report of a Working Conference of the American Medical Informatics Association. *J Am Med Informatics Assoc*. Published online 2008. doi:10.1197/JAMIA.M2905

7. Hajjar I, Kotchen TA. Trends in Prevalence, Awareness, Treatment, and Control of Hypertension in the United States, 1988-2000. *J Am Med Assoc*. Published online 2003. doi:10.1001/JAMA.290.2.199

8. Quan H, Khan N, Hemmelgarn BR, et al. Validation of a case definition to define hypertension using administrative data. *Hypertension*. Published online 2009. doi:10.1161/HYPERTENSION.109.139279