11-23-2018 2:00 PM

# Prognostic Predictive Model to Estimate the Risk of Multiple Chronic Diseases: Constructing Copulas Using Electronic Medical Record Data

Jason E. Black
*The University of Western Ontario*

Supervisor
Lizotte, Daniel
*The University of Western Ontario*

Graduate Program in Epidemiology and Biostatistics
A thesis submitted in partial fulfillment of the requirements for the degree in Master of Science

# Abstract

Introduction: Multimorbidity, the presence of two or more chronic diseases in an individual, is a pressing medical condition. Novel prevention methods are required to reduce the incidence of multimorbidity. Prognostic predictive models estimate a patient's risk of developing chronic disease. This thesis developed a single predictive model for three diseases associated with multimorbidity: diabetes, hypertension, and osteoarthritis.

Methods: Univariate logistic regression models were constructed, followed by an analysis of the dependence that existed using copulas. All analyses were based on data from the Canadian Primary Care Sentinel Surveillance Network.

Results: All univariate models were highly predictive, as demonstrated by their discrimination and calibration. Copula models revealed the dependence between each disease pair.

Discussion: By estimating the risk of multiple chronic diseases, prognostic predictive models may enable the prevention of chronic disease through identification of high-risk individuals or delivery of individualized risk assessments to inform patient and health care provider decision-making.

## Keywords

Multimorbidity, Chronic disease, Diabetes, Hypertension, Osteoarthritis, Risk prediction, Prognostic predictive model, Electronic medical records, CPCSSN

# Co-Authorship Statement

Jason Black, Dr. Daniel Lizotte, and Dr. Amanda Terry contributed to the conception and design of the work. JB was responsible for data analysis and interpretation. JB drafted the article. All authors participated in critical revision of the article.

# Acknowledgments

The past two years have been filled with an abundance of growth, both personally and professionally. It is my pleasure to acknowledge several individuals for the roles that they have played in my development.

To begin, I would like to thank the members of my thesis committee: Dr. Dan Lizotte, Dr. Amanda Terry, and Dr. Sonny Cejic. These individuals have pushed me to explore new avenues of research and apply research to where it has not previously been done before; yet, they have always brought me back to the *“so what?”,* the clinical interpretation. In doing so, these individuals have provided me with a strong foundation of knowledge and skills that I will draw on for years to come.

I would also like to thank several key mentors who have supported me greatly through my graduate career. These mentors included: Dr. Heather Maddocks, Dr. Moira Stewart, Dr. Stewart Harris, Alexandria Ratzki-Leewing, and Dr. Kelly Anderson. Each of these individuals have provided me the opportunity to further my skills and competencies and explore new and exciting areas of research. Additionally, I would like to thank the larger DELPHI team for supporting me in my research and providing perspectives that were essential to shaping this work into a relevant and timely piece of literature. I must also acknowledge both the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) and Deliver Primary Health Care Information (DELPHI) project for providing the data for this thesis.

To my colleagues and classmates, I am grateful for your fellowship through our time here. In particular, I am grateful to the members of SHET for their support through the various challenges that came my way. Every brainstorming session, Kresge night, and coffee run has been instrumental in getting me to this point.

For my friends and family, I am extremely grateful. Their support in all my endeavors, both personal and professional, has been essential to my success. I would especially like to thank parents for their endless support and encouragement.

# Table of Contents

# List of Tables

# List of Figures

# List of Appendices

# Chapter 1

## 1 Introduction

Prognostic predictive models (PPM) estimate a patient's risk for future disease development based on the patient's current predictors of disease (1,2). Potential predictors, including patient demographics, family history, lifestyle factors, medical conditions, or genetic factors (3), are used to produce risk estimates of disease. These estimates can be used by primary care practitioners (PCP) in their primary prevention activities with patients (4,5). Many chronic diseases have been accurately predicted through risk estimation using prognostic predictive models (6–8), such as cardiovascular disease (7,9). These tools have been shown to improve patients' risk perception and knowledge (10), as well as to modify PCP care, including prescribing behaviours (11). The objective of a PPM is to inform patient and PCP decision-making by providing risk estimations and identifying high-risk individuals to target with risk-reducing interventions, thereby reducing the future incidence of disease.

The occurrence of two or more chronic diseases within a single patient, or multimorbidity, is a growing concern in the health care community. Prevalence estimates vary due to inconsistent definitions of multimorbidity; levels in Canadian older adults range from 55 to 98% (12). Methods of preventing the development of multimorbidity are severely lacking (13); novel methods of prevention must be developed to reduce future incidence. One potential method that may aid in preventing future cases of multimorbidity is the use of prognostic predictive models aimed at informing patients and their care providers about multimorbidity risk.

A prognostic predictive model capable of estimating the risk of multiple diseases simultaneously would deliver a comprehensive risk assessment that could be used to identify patients at highest risk of disease in general. This model would incorporate aspects beyond risk factors for individual diseases when estimating risk; for example, prior morbidities would be included as predictors. Risk estimates produced by this model would include more than an overall risk of disease; it would present a comprehensive description of patient risk for multiple outcomes, including high-risk individual diseases, common disease pairings, and/or common clusters of chronic disease. In the past, the development of models capable of estimating risk of multiple diseases has been restricted by the lack of datasets large enough and rich enough to support the production of such estimates. Recently established large-scale electronic medical record (EMR) databases (14–17) represent a potential source of data of the size necessary to support efforts in multimorbidity risk modelling.

Primary prevention is one of the fundamental goals of primary care (18), making primary care an ideal setting to deploy such methods to reduce multimorbidity risk. Modelling risk of multiple diseases could impact clinical practice twofold. First, it would identify patients at highest risk, allowing the PCP to target these patients with risk-reducing interventions (19). Second, practitioners and patients are often tasked with determining which risk factors are of the greatest importance, as these should be targeted first. Risk modelling of multiple diseases would identify risk factors that significantly contribute to the risk of multiple diseases to maximize the impact of risk factor modification and reduce intervention-burden on patients. It is hoped that these targeted interventions will help prevent future incidence of multimorbidity.

Traditionally, risk is modelled for individual diseases under the assumption that disease develops independent of other diseases; however, we know from the study of multimorbidity that this assumption is often false. Occurrence of diseases in the same individual is not an independent event; disease processes influence the development or progression of other diseases leading to chronic diseases often occurring together (20–22). For this reason, the estimation of disease risk under the assumption that disease occurrences are independent events does not accurately reflect a patient’s true risk of disease. Therefore, an understanding of the dependence that exists between each disease is required prior to the construction of a prognostic predictive model for multiple chronic diseases. Given this, models can be built to describe the dependence between multiple diseases and estimate their risk of occurring in the same individual. The two main goals of this work are 1) to build a prognostic predictive model that both accounts for existing chronic disease and predicts the occurrence of multiple diseases simultaneously and 2) to examine the dependence that exists between three chronic diseases – diabetes, hypertension, and osteoarthritis – after adjusting for known risk factors. To achieve this, two objectives were identified: 1) to construct a univariate multivariable logistic regression model for each chronic disease, which allowed for 2) the construction of a copula that captures their joint dependence after adjusting for risk factors. The resulting model achieves both goals, since it can be used to make risk predictions and it can be used to assess dependence in disease development.

This thesis first explores chronic disease, namely multimorbidity, and the possibility of predicting multiple diseases using prognostic predictive models. Electronic medical records are discussed as a potential data source. Subsequently, the methods of developing such a model are presented, followed by the resulting model. Finally, the model that was developed for this thesis is compared to existing research into prognostic predictive models and multiple diseases. In addition, the limitations and implications of the model produced for this thesis are discussed.

# Chapter 2

# 2 Literature Review

The following literature review first describes chronic disease, particularly multimorbidity, and its impacts. It continues by presenting prognostic predictive models as a potential method of supporting the primary prevention of chronic disease and multimorbidity. It then finishes by describing two requirements of multiple disease risk estimation: advanced statistical methodologies and electronic medical record data.

## 2.1 Chronic Disease

As the leading cause of death worldwide, chronic disease represents one of the world's largest challenges (23). Disease is categorized into two distinct types: chronic and acute. Acute diseases are characterized by their short duration. Patients typically recover from their disease within a brief period of time (24); the length of this period is dependent upon the disease context. Acute diseases are most often communicable, or transmitted from person to person, and thus commonly referred to as "communicable diseases". On the other hand, chronic diseases are long-lasting in duration. Patients recover from their disease only after an extended period of time, if ever; many chronic diseases are lifelong diseases. Chronic diseases are most commonly non-communicable, or not transmitted from person to person. There are some exceptions to this, most notably HIV/AIDS (25). In the past, acute infectious diseases were the main cause of morbidity and mortality, globally. Typhoid, cholera, smallpox, tuberculosis, and diphtheria, among other infectious and parasitic diseases were common until the early twentieth century (25). Due to recent advancements in health care and the ability to treat (and often cure) infectious diseases, such as the advent of antibiotics and vaccines and improvements in housing, sanitation, water supply, and nutrition, rates of infectious diseases have severely fallen. However, due to the increased lifespan resulting from the factors mentioned, there has been a concurrent rise in the occurrence of chronic diseases. Chronic diseases such as cardiovascular disease, cancer, and COPD are among the leading causes of death for developed nations (25). One of the main focuses of today's health care efforts is the treatment and prevention of chronic diseases.

### 2.1.1 Definition

There is no single, agreed upon definition of chronic disease (26). Most definitions depend on either disease duration or disease transmission; however, these definitions will produce inconsistent classifications of disease. As described above, HIV is a communicable disease for which no cure exists that results in lifelong health effects.  Many organizations have constructed definitions of chronic disease and lists of diseases they consider chronic diseases. These definitions are often non-specific; for example, the U.S. National Center for Health Statistics defines chronic disease as "a disease lasting 3 months or longer" (27). Additionally, these definitions often disagree. The Public Health Agency of Canada considers five main groups of chronic diseases: cardiovascular diseases, cancers, chronic respiratory diseases, diabetes, and mood and anxiety disorders (28). By the U.S. National Center for Health Statistics definition, diseases such as epilepsy or chronic kidney disease would be considered chronic diseases, whereas by the Public Health Agency of Canada definition they would not.

Given this lack of a standard definition of chronic disease, it is not possible to use it as an outcome for prediction without significant additional effort. However, this thesis was not subject to this issue as it simply selected three diseases for prediction that are managed in primary care and possess lifelong clinical implications. The methodology in the thesis is general and can be applied to any outcome of interest.

### 2.1.2 Prevalence

In Canada, greater than one fifth of Canadians age twenty or older live with at least one major chronic disease (CVD, cancer, chronic respiratory disease, and diabetes) (29). In Canadians aged 65 or older, this proportion of individuals living with at least one major chronic disease grows to over 40% (29).

### 2.1.3 Risk Factors

A risk factor is some characteristic that is causally associated with a given disease; these can be environmental, genetic, behavioural, somatic, or social. Risk factors are distinct from risk markers, which are associated with risk of disease non-causally or of unproven causation (30). Risk factors must be proven to be causally associated with risk of disease through the use of epidemiologic methods. For example, smoking has been shown to impart an increased risk for lung cancer (31), making it a risk factor. Identification of risk factors is important for two reasons; it allows for 1) risk assessments in which an individual's risk factors are assessed to enable risk estimation and 2) subsequent intervention upon modifiable risk factors. Most diseases are multifactorial (i.e., their development is influenced by multiple risk factors). Indeed, it is common for individuals to have more than one risk factor. When an individual possesses multiple risk factors, these risk factors can have an additive or multiplicative impact on risk; this effect is known as interaction. When risk factors interact in an additive manner, the individual's total risk is greater or less than the sum of the component risks (32). When risk factors interact in a multiplicative manner, the individual's total risk is greater or less than the product of the component risks (32). Risk factors are often targeted with interventions aimed at reducing a person's risk of developing chronic disease.

Risk factors can be modifiable or non-modifiable. Modifiable risk factors can be changed through some intervention. Interventions often include behavioural or lifestyle changes, medical procedures, or pharmaceutical treatments. For example, smoking is a risk factor for cardiovascular disease, which can be modified by quitting smoking; studies have demonstrated that quitting smoking reduces the risk of developing cardiovascular disease (33,34). Non-modifiable risk factors cannot be changed. For example, family history is non-modifiable risk factor for cardiovascular disease. Intervention upon risk factors is the focus of primary prevention efforts, which aim to lower an individual's risk of developing disease.

Many chronic diseases share common risk factors. For example, obesity is a known risk factor for diabetes (35,36), depression (37), and osteoarthritis (38–40). Where shared risk factors exist (and are modifiable), targeting these first will have the largest impact on an individual's risk of disease overall. In patients at high risk of multiple diseases, it is preferable to target risk factors associated with multiple diseases; in this situation, the individual would be subject to fewer interventions than if each disease were intervened upon individually, thus reducing intervention burden on the patient.

### 2.1.4 Multimorbidity

Multimorbidity is an extremely common medical condition, especially among older adults (12). Although definitions vary, multimorbidity is commonly considered the presence of two or more chronic conditions within an individual (41). For example, a patient diagnosed with both asthma and diabetes has multimorbidity. Multimorbidity can be contrasted with comorbidity. Both of these terms refer to multiple chronic diseases within the same individual; however, when examining comorbidities, there is always an index disease that is the primary focus, for which its care is modified when additional morbidities are considered. Multimorbidity, on the other hand, does not prioritize one disease over another.

Despite professional agreement that multimorbidity is a pressing health issue (12), there is no standard, consistently used definition of multimorbidity (42–44). The European General Practice Research Network defines multimorbidity as “any combination of chronic disease with at least one other disease (acute or chronic) or bio-psychosocial factor (associated or not) or somatic risk factor” (45), without a list of diseases or conditions that should be considered. In contrast, the Public Health Agency of Canada considers multimorbidity to be two or more of the following diseases within the same individual: heart disease, stroke, cancer, asthma, COPD, DM, arthritis, Alzheimer’s or other dementia, mood disorder (depression), and anxiety (29). A recent systematic review by Fortin et al. examined the impact of including various numbers of diseases in a definition for multimorbidity on the prevalence of multimorbidity (42). Findings of this systematic review demonstrated that inclusion of at least 12 chronic conditions in the definition of multimorbidity resulted in stable prevalence estimates; including more conditions in the definition did not significantly alter the prevalence estimates. Fortin et al. suggest including the 12 most prevalent chronic conditions in a definition of multimorbidity.

Prevalence estimates in the literature are inconsistent due to varying definitions of multimorbidity; however, estimates of multimorbidity prevalence in older adults range from 55 to 98% (12). According to the Public Health Agency of Canada definition of multimorbidity, the prevalence of multimorbidity in 2014 was 14.8% in Canadians aged 20 and older.

Multimorbidity has impacts on both patients and their HCPs. This complex condition that imposes a huge burden on patients has been shown to reduce health-related quality of life, limit activities of daily living, and decrease self-rated health (46,47). An inverse relationship has been found between multimorbidity and health-related quality of life; as the number of multimorbid diseases increases, health-related quality of life has been found to decrease (48). Issues such as polypharmacy, fragmentation of care, and conflicting or competing health care recommendations may be faced by patients with multimorbidity (12), making treatment of these individuals complicated. The economic burden of multimorbidity is massive; in 2009, nearly 80% of health care costs in Canada were due to individuals with multimorbidity (49). Wikström et al (50) examined risk factors specifically for multimorbidity. They found that smoking, physical activity, and BMI were significant contributors to risk of multimorbidity development. Additionally, systolic blood pressure and low education contributed to risk of multimorbidity among men. Dhalwani et al (51) examined the impact of physical activity on development of multimorbidity among an older English population; they found a dose-response relationship between levels of physical activity and multimorbidity: for those at higher levels of physical activity, fewer developed multimorbidity. Dankel et al. (19) examined the impact of muscle-strengthening activities on multimorbidity risk. Those who participated in the muscle-strengthening activities had 26% lower odds of developing multimorbidity. These studies demonstrate the importance of physical activity for the prevention of multimorbidity. Recently, there has been a focus on developing strategies to prevent multimorbidity as health policy makers and health care providers recognize the importance of multimorbidity (52). In 2015, the Public Health Agency of Canada published a report that stressed the importance of addressing chronic disease from a comprehensive, holistic approach, including consideration of multimorbidity, rather than a single-disease-centred approach (53).

Three chronic diseases commonly associated with multimorbidity are diabetes mellitus, hypertension, and osteoarthritis.

### 2.1.5 Diabetes Mellitus

Diabetes mellitus (DM) is a group of metabolic disorders characterized by elevated blood glucose levels over prolonged periods (54). DM comprises two main conditions: type 1 DM, in which the pancreas is not able to produce enough insulin and type 2 DM, in which the pancreas produces insulin but the body's cells fail to respond properly. As type 2 DM progresses, failure to produce insulin may also develop (55). Type 1 DM, traditionally termed *juvenile diabetes* as its onset typically occurs before adulthood, comprises approximately 10% of cases of DM (56). The most common cause of type 1 DM is an autoimmune attack on the insulin-producing beta cells of the pancreatic islets, resulting in insulin deficiency (57). Type 2 DM is the more common of the two, comprising approximately 90% of cases of DM (56). Its onset is typically in adulthood; however, a growing proportion of younger individuals are developing type 2 DM (58). This is likely due to the increase in risk factors for DM, such as obesity, lack of physical activity, and poor diet, in youth.

Based on national survey data, the population prevalence of DM (both type 1 and type 2) is roughly 9.8% (29). Based on a Canadian study using electronic medical record data conducted by Greiver et al., the prevalence within primary care patients was 8.2% (59). When corrected using a corrected yearly contact group denominator, the population prevalence of DM was 7.6% (59).

Patients with type 1 DM must manage their glucose levels using insulin injections; this requires monitoring of blood glucose levels using repeated blood tests and administration of insulin injections (60). Patients with type 2 DM do not always require insulin. Lifestyle changes, such as proper diet and exercise, and medications (e.g., metformin), are used to manage patients with type 2 DM (60). Insulin injections may be added to treatment when the disease has progressed; however, most individuals do not initially require insulin (55).

Complications of DM are the same for both type 1 and type 2 DM and are minimized through proper control of glucose levels (61,62). DM leads to both microangiopathy and macroangiopathy, often resulting in severe complications, both acute and chronic (54). Acute complications, such as hypoglycemia, hyperglycemia, and, less commonly, diabetic coma, can occur within patients with DM. Chronic complications include diabetic nephropathy (i.e., damage to the kidney that can lead to chronic kidney failure); diabetic retinopathy (i.e., growth of poor-quality blood vessels and swelling that can result in vision loss or blindness); diabetic cardiomyopathy (i.e., damage to the heart muscle that can lead to heart failure); cardiovascular disease; and foot ulcers.

Type 1 DM is currently not preventable (60); however, type 2 DM may be delayed or prevented through the modification of risk factors (60). Non-modifiable risk factors for type 2 DM include older age (36,63); male sex (36); polycystic ovarian syndrome (PCOS) (64); psychiatric disorders such as schizophrenia (65,66), depression (67), bipolar disorder (66,68); family history of type 2 diabetes (36); air pollution (69); and low socioeconomic status (70). Modifiable risk factors for type 2 diabetes include obesity (35,36,63,70), waist circumference, lipid disorders (36), hypertension (36,63), smoking (36,63), stress (36), and low physical activity (70). The effectiveness of diet-and-exercise programs in reducing diabetes incidence through weight reduction and regular physical activity has been demonstrated in many randomized controlled trials (RCTs) (71–73). Bariatric surgery to facilitate weight loss has shown promise in preventing type 2 DM development (74). Several pharmacotherapies have been investigated for type 2 DM prevention. Metformin has been shown to significantly reduce risk of developing diabetes (71), even after medication discontinuation (75,76). Other medications have been investigated, including thiazolidinediones such as troglitazone (77), rosiglitazone (78), ramipril (79), and pioglitazone (80); alpha-glucosidase inhibitors (81); orlistat (82); and incretin-based therapies such as liraglutide (83); however, results remain indefinitive, limiting their use. Clinical practice guidelines published by Diabetes Canada (formerly the Canadian Diabetes Association) (84) recommend reduction of type 2 DM risk through a structured lifestyle modification program, including weight loss, physical activity and pharmacological therapy with metformin or acarbose, in patients with impaired glucose tolerance.

### 2.1.6 Hypertension

Hypertension is a condition in which blood pressure in the arteries is consistently elevated (85). Hypertension due to some identifiable cause, such as pregnancy, polycystic kidney disease, or medication, is referred to as “secondary hypertension”; this form of hypertension comprises only 5-10% of cases (86). Hypertension due to unknown causes is referred to as *primary* (or *essential*) *hypertension*, constituting the remaining 90-95% of cases (86,87). Hypertension does not usually cause symptoms; however, chronic high blood pressure is a known risk factor for cardiovascular disease, stroke, vision loss, and chronic kidney disease (85,88). Reduction of blood pressure through lifestyle modifications and medications reduces risk of complications (85,88). Patients diagnosed with hypertension are encouraged to reduce their blood pressure to target blood pressure recommended by various expert groups (89); the Canadian Hypertension Education Program (90) recommends a target blood pressure of less than 140/90 for the general population.

Based on the Canadian Health Measures Survey (CHMS), the national prevalence of hypertension in Canada was 22.6% in the years 2012-2013 (91). Based on a Canadian study using electronic medical record data conducted by Godwin et al., 22.8% of a primary care population had a diagnosis of hypertension as of 2012 (92). Of this primary care population, most patients (80%) diagnosed with hypertension were able to reach target blood pressure levels.

As hypertension does not result in symptoms (85), the burden of the disease itself on patients is low. However, hypertension has a large impact on individuals when considering its long-term complications; CVD, stroke, vision loss, and chronic kidney disease all have severe impacts on an individual's well-being.

Non-modifiable risk factors for hypertension include older age (93), diabetes (93,94), kidney disease (95), and sleep apnea (96). Modifiable risk factors include obesity (93,94), smoking (93), stress (97), tricyclic antidepressant use (98), and high salt intake (93,99).

Primary prevention of hypertension focuses on modification of its risk factors. For example, the DASH diet (Dietary approaches to stop hypertension) (100) aims to reduce salt consumption through modification of diet. Physical activity is commonly recommended to reduce risk of hypertension; many epidemiologic studies have demonstrated a consistent dose-response relationship between physical activity and development of hypertension, in which higher levels of physical activity were associated with lower rates of hypertension (101). Additionally, weight loss, even modest, has been shown to decrease risk of hypertension (102). Indeed, the Canadian Hypertension Education Program recommends the following (103) for the prevention of hypertension: regular physical activity; maintenance of a healthy body weight; alcohol consumption within the Canadian low-risk drinking guidelines (104); maintenance of the DASH diet (100); reduction of sodium consumption; and stress management.

### 2.1.7 Osteoarthritis

Osteoarthritis is a degenerative joint disease characterized by the breakdown of joint cartilage and the underlying bone (105). Symptoms of osteoarthritis most commonly include joint pain and stiffness (105). As the disease progresses, the severity of pain and stiffness increases, and movement patterns are typically affected. The most commonly affected joints are those of the hands, neck, lower back, hips, and knees. Typical treatment of osteoarthritis involves lifestyle modification and medications (106–108). Weight loss has been shown to reduce pain and stiffness and improve function of the joint (109). Medications, such as non-steroidal anti-inflammatory drugs, are used for treatment of pain. In cases where the impact of osteoarthritis symptoms are severe and conservative forms of treatment are ineffective, joint replacement surgery may be recommended, in which the affected joint is replaced with an artificial joint. Replacement of hip and knee joints have been shown to be clinically (110,111) and cost effective (112,113).

The population-based prevalence of osteoarthritis was estimated to be 13.0% in Canadians aged 20 and older (28). Based on a study in Canada using electronic medical record data conducted by Birtwhistle et al., the prevalence of osteoarthritis within the primary care population was 14.2% in 2012 (114).

The impact of osteoarthritis on individuals living with the disease is dependent upon the joint(s) affected and the progression of disease. Osteoarthritis of the knee and hip can limit activities such as running or walking, whereas osteoarthritis of the hand joints can impede activities such as writing or typing. Degree of disease progression and an individual's ability to manage symptoms affects how activities are affected.

Efforts surrounding osteoarthritis management typically focus on treatment of symptoms, rather than disease prevention. This is likely due to its slow, progressive nature with no clear point of disease onset. Non-modifiable risk factors for osteoarthritis include leg length inequality (115), older age (39,40,116–118), female sex (39,40,116,117), family history of osteoarthritis (116), and osteoporosis (39). The modifiable risk factors for osteoarthritis are obesity (38–40,116–119), previous joint injury (38,40,117,119), and physically intensive occupations (116). Suggested osteoarthritis prevention efforts concentrate on obesity. For example, a diet-and-exercise program aimed at weight reduction was found to be suggestive of a reduction in the incidence of knee osteoarthritis (120). There is growing interest in preventing osteoarthritis through the implementation of joint injury prevention programs (121). To date, there are no established guidelines pertaining to the prevention of osteoarthritis.

## 2.2 Prognostic Predictive Models

### 2.2.1 Overview

Prognostic predictive models can be used to estimate the risk of developing a chronic disease. In particular, the objective of prognostic predictive models (PPMs) is to inform patients and care providers about patient risk, and thereby motivate the use of interventions that prevent future disease development. There are several levels of disease prevention, each used at different stages of disease progression. Primary prevention describes the efforts taken prior to disease development to reduce risk of future disease development (122,123). Common primary prevention interventions include eating a healthy diet, quitting smoking, or exercising regularly. Secondary prevention describes diagnostic efforts after disease onset but prior to clinical manifestations (symptoms) of disease to detect the initial stages of disease, allowing for early treatment; for example, breast cancer screening is done to detect the disease earlier, when treatments are more effective (122,124). Tertiary prevention describes the actions taken to reduce the impact and ease the burden of an on-going disease (122,125). One example of a tertiary prevention intervention is physiotherapy following a joint injury to improve joint function. PPMs are one tool used to support the primary prevention of disease.

Traditional primary prevention interventions include risk management. Risk management refers to the forecasting and evaluation of patient risk and practices aimed at reducing this risk (126–128). Individual risk factor management is a specific type of risk management. According to this strategy, risk factors for chronic disease are individually assessed through risk assessments and subsequently intervened upon; however, recent research has demonstrated that risk assessments that examine a patient's global risk of disease by considering multiple risk factors simultaneously have been more effective in risk reduction (129). A patient's global risk takes into account the impact of multiple risk factors to estimate the risk that the patient will develop disease within a given time period (130). One method of estimating a patient's global risk of disease is through the use of PPMs.

Previous methods of risk estimation have relied upon anecdotal evidence and professional opinion of the practitioner and often vary between practitioners. PPMs represent an objective, evidence-based method of assessing an individual's risk of future disease development using multiple risk factors (131). These models present estimated risk in the form of a risk score, usually a numeric value where a greater value denotes greater risk (132). Alternatively, individuals are assigned to categories corresponding to varying degrees of risk. Common PPMs include the Framingham Risk Score for cardiovascular disease (9) and EuroScore to estimate risk of death after a heart operation (133).

Multivariable statistical methods are used in PPMs. A PPM estimates the risk of some outcome (future development of disease) based on a set of covariates (characteristics shown to indicate risk, or risk indicators). This is distinct from past risk estimation methods where individual risk factors were assessed independently then intervened upon. Compared to the use of a single predictor, the use of multiple predictors allows for more accurate estimation of a patient's risk (5), as multiple risk factors commonly coexist within an individual (130). Interaction between risk factors occurs when the joint effect on risk is greater than what would be expected by adding or multiplying the effects of each risk factor; these interactions can be modelled by PPMs by including an interaction term in the model (134). Characteristics of the patient, provider, or practice can be included as covariates in the model. Potential patient level covariates include patient demographics, family history, lifestyle factors, medical conditions, and genetic factors (3). Potential provider level covariates include specialization (if any), years in practice, and additional certifications. Potential practice level covariates include rurality, number of practitioners, and geographic location. Inclusion of characteristics from multiple levels requires the use of advanced methods such as multilevel modelling (135).

PPMs are distinct from etiologic research. The focus of prognostic research is to predict some outcome, whereas etiologic research aims to identify the cause of some outcome (136). Both prognostic and etiologic research use multivariable approaches; however, in etiologic research, the goal is to isolate the main causal effect of some exposure by adjusting for the effects of other confounding factors (136). In contrast, prognostic research uses a multivariable approach to estimate risk of an outcome as accurately as possible by including as much potentially predictive information as possible (5). As such, covariates included in a prognostic model do not have to be causally related to the outcome (and are often not). Risk management practices can then be used to minimize the risk among high-risk patients by intervening on factors that are known to be causal.

### 2.2.2 Conceptual Model

Two theories exist that drive the use of PPMs, one at the individual level and another at the population level. At the individual level, the theory driving the use of PPMs posits that knowledge of disease course enables patients and practitioners to make informed decisions to avoid or deter the development of disease (5). At the population level, the use of PPMs is driven by the need for identification of patients at high-risk of disease, enabling targeted interventions aimed at reducing risk within these patients (5). The following conceptual model describes the process of risk estimation via PPMs, from both the individual and population perspectives.

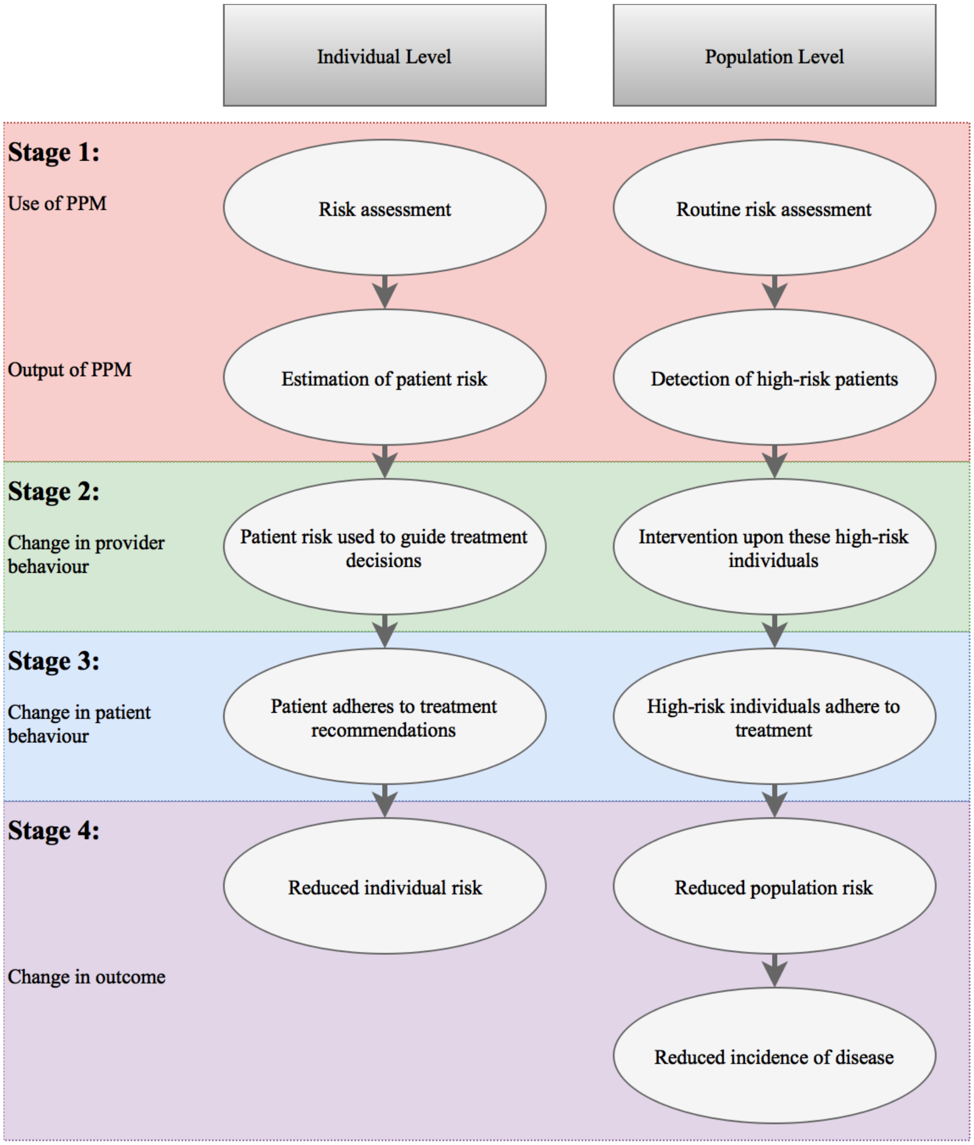
Individual Level
Population Level
Stage 1:
Use of PPM
Risk assessment
Routine risk assessment
Output of PPM
Estimation of patient risk
Detection of high-risk patients
Stage 2:
Change in provider behaviour
Patient risk used to guide treatment decisions
Intervention upon these high-risk individuals
Stage 3:
Change in patient behaviour
Patient adheres to treatment recommendations
High-risk individuals adhere to treatment
Stage 4:
Reduced individual risk
Reduced population risk
Change in outcome
Reduced incidence of disease

**Figure 1: Conceptual Model for Risk Assessment**

The four stages of risk assessment outlined in the conceptual model above will be discussed below.

#### 2.2.2.1 Stage 1: Use of Prognostic Predictive Models

Risk assessments, which commonly include a PPM, can occur in two ways. In the first, either the patient or PCP is concerned about risk of future morbidity and the PCP actively performs a risk assessment. In this case, there must be some initiation or interest in risk assessment by either the patient or PCP. Alternatively, risk assessments can run passively in the background, assessing all patients' risks, and send an alert when a patient is classified as high-risk. In this case, the risk estimation model is run regardless of patient or PCP concern about future morbidity. Integration of risk assessments into electronic medical record (EMR) systems facilitates both these strategies, as tools within an EMR are easily accessible to the PCP and are able to run in the background of the EMR. Building these tools into EMRs results in simple incorporation into clinical workflow to offer real-time recommendations.

#### 2.2.2.2 Stages 2 & 3: Change in Provider and Patient Behaviour

PPMs can be used in primary care to inform PCP decision making regarding risk reduction. Most often PPMs are used as a part of a risk assessment. Studies examining the use and impact of PPMs in primary care in Canada are rare; most PPM studies focus on the development or validation of models, but do not evaluate how current models are used (137). However, many Canadian guidelines advocate for the use of risk assessments. The Canadian Cardiovascular Society recommends cardiovascular risk assessment using every 5 years for men and women between ages 40 and 75 using PPMs such as the modified Framingham Risk Score or Cardiovascular Life Expectancy Model (138). The results of these risk assessments are used to inform decisions regarding interventions to reduce risk of cardiovascular events. Other examples include the 2010 Clinical Practice Guidelines from Osteoporosis Canada, which recommends osteoporosis and fracture risk assessment for individuals over age 50 who have experienced a fragility fracture or who have a history of falls (139). These risk assessments can inform interventions such as exercise and prevention of falls, calcium and vitamin D supplementation, or pharmacologic therapy.

Many factors limit the use of PPMs in clinical practice. A qualitative study examined barriers cited by PCPs limiting their use of PPMs in clinical practice including lack of lifestyle recommendations, legal and regulatory constraints, and lack of accuracy of risk scores (140). PCPs felt that predictive models focused more on non-modifiable risk factors, such as age, thus limiting their ability to give recommendations on more relevant, modifiable risk factors such as lifestyle. PCPs also expressed concerns over the current regulations that did not place a focus on disease prevention; no remuneration exists for time spent on risk assessment and prevention. PCPs feared that the estimated risk scores were not an accurate representation of individual patient's risks, and thus were less likely to use the PPM.

#### 2.2.2.3 Stage 4: Change in Patient Outcomes

Similar to studies looking at their use, few studies in Canada have examined the impact of PPMs on clinical outcomes, such as impact on risk, disease incidence, or physician behaviours; however, some studies exist examining their use. The National Health Service (NHS) Health Check is a health check-up for adults that includes a CVD risk assessment (141). All adults in England between the ages 40 and 74 without a pre-existing condition are invited by their PCP or local authority for a free NHS Health Check every 5 years in an effort to reduce risk of chronic disease. The introduction of the NHS Health Check was shown to be associated with significant reductions in CVD risk, as well as improvements in statin prescriptions (142). The Multidisciplinary Risk Assessment and Management Program for Patients with Diabetes Mellitus (RAMP-DM) in Hong Kong had similarly promising results, finding that a risk assessment and management program (which included a risk assessment component) was associated with fewer cardiovascular complications and lower all-cause mortality after 3 years (143). In contrast, a systematic review conducted by Brindle et al. described several studies that observed no impact from PPM risk assessment on the observed outcomes, which included predicted risk of CVD, fatal or non-fatal CVD events, risk factor levels, prescription of risk-reducing drugs, and changes in health-related behaviour (144); however, this review frequently noted that the PPM's use by PCPs was either poorly recorded or not recorded at all. Indeed, it is unclear whether the lack of impact was due to the efficacy of the PPM or the lack of use by the physician.

### 2.2.3 Settings of Prognostic Predictive Model Use

Primary care is the first point of patient contact where, most often, patients are seen prior to disease development; once patients reach secondary or tertiary levels of care they have already developed disease, thus prognostic risk estimation is of little value in these clinical settings. For this reason, primary care is an ideal setting for targeted primary prevention efforts, including the use of PPMs. PPMs can be used to inform the PCP decision-making process surrounding risk management, such as whether or not to recommend risk lowering interventions. Many guidelines recommend the use PPMs to detect high-risk individuals for primary prevention efforts (145,146). One common PPM used in primary care is the Framingham model for cardiovascular disease (CVD) (147), which estimates a patient's risk of developing CVD in the next 10 years based on risk factors including age, sex, cholesterol levels, and smoking status. PCP decisions such as whether or not to prescribe statins or to recommend lifestyle changes are informed by this model.

Additionally, PPMs are commonly used as online tools to deliver risk estimates to the general population over the internet (148,149). This method allows a larger population, beyond primary care patients, to access personalized risk estimates in a convenient manner and receive recommendations on ways to reduce risk. Online risk assessments are based on established PPMs that have been empirically developed and validated; however, these models are sometimes modified to substitute covariates that are not commonly known by individuals. For example, personal blood cholesterol within the Framingham cardiovascular risk assessment may be replaced with an average value as most individuals will not likely know their blood cholesterol levels off-hand. Given the current status of PPM use in primary care, online risk assessments are able to reach a much greater proportion of the population. The deployment of online PPMs aims to reduce the overall incidence of disease within the community by informing individuals of their personal risk and subsequently recommending risk reducing interventions.

### 2.2.4 Prognostic Predictive Model Development

The development of a PPM is a complex process involving risk factor identification, data processing, and statistical analysis. For a complete description of the processes involved in PPM development, see the Transparent Reporting of a Multivariable Prediction Model for Individual

Prognosis or Diagnosis (TRIPOD) Statement (150). Ideally, there are three stages in prognostic research: model development, validation, and clinical impact evaluation (5).

Model development requires knowledge of the relevant risk factors for the disease of interest; a review of the relevant literature will reveal risk factors for the disease. Also required for model development is a dataset from which to build the model. When developing a PPM, it is optimal to use a prospective cohort study conducted specifically to collect data to inform the PPM development (151). However, due to the high costs associated with primary data collection, development of PPMs may make use of data previously collected (retrospective data). Results from retrospective data are more prone to bias as predictor and outcome information is less systematically recorded (1). Common sources of retrospective data include previous observational studies, randomized controlled trials, and health administrative data (6,152,153).

PPMs are constructed using statistical techniques, commonly regression, to estimate risk of disease development in the future (2). The type of regression used is primarily dependent on the type of outcome that is to be predicted. For example, when continuous outcomes are to be predicted, linear regression (154) is commonly used; when binary ("dichotomous" or yes/no) outcomes are to be predicted, logistic regression (155) is commonly used; or when time-to-event outcomes are to be predicted, Cox regression (156) is commonly used. Each of these methods posit a (possibly transformed) linear relationship between the covariates and the outcome and are each considered Generalized Linear Models. Advanced methods, such as Generalized Additive Models (157), can be used to model more complex, non-linear relationships between covariates and outcomes. These methods offer the advantage of accounting for non-linearity between the covariates and the outcome, better modelling the relationships in the data; however, they sacrifice some interpretability of the model, as such models are often more difficult to understand. Disease status is generally considered a binary outcome; a patient either has the disease or they do not. For this reason, logistic regression or survival analysis methods are most commonly used to model disease outcome data for PPMs. Other non-regression methods exist to predict future disease development; these include decision trees (158), where patients pass through a series of yes/no questions to eventually classify their risk of disease development, or k-nearest neighbours (159), where a patient is classified according to the risk corresponding to those most similar to them. While these models are sometimes more easily interpreted, they are not as frequently used as regression methods.

Following the development of the model, it must be validated to assess its accuracy. Model validation consists of internal and external validation (160). Internal validation is performed on data from the same source used to construct the model, either using a held-out portion of the data or methods such as cross validation (161) or bootstrapping (162). External validation applies the model to a different population that is similar to determine its accuracy (163). At a minimum, validation looks at discrimination and calibration.  Discrimination is the ability to correctly assign higher risk to a patient who ultimately experiences the outcome compared to the patient who does not (4). Discrimination is commonly assessed via the c-statistic or an ROC curve (164). Calibration is a measure of how well a model fits the data (4); this describes how well the risk estimates the true proportion of patients that will develop disease. For patients assigned a given risk (probability), approximately the same proportion of patients should actually go on to develop disease in a model with high calibration. This can be assessed using the Hosmer-Lemeshow test (155) or a calibration plot; however, the Hosmer-Lemeshow test has been shown to be over-sensitive when dealing with large datasets (3).

The final step in the development of a PPM is to assess its clinical impact (137). Rather than simply assessing performance of the model (i.e. how well the model predicts the outcome), this stage assesses the model's impact on physician practices, patient care, and patient outcomes. Examples of impacts include modification of physician behaviours, such as prescribing patterns; modification of patient risk; and change in disease incidence. Most PPMs are internally validated at a minimum; however, few are validated on an external dataset, and even fewer have had their clinical impact assessed (165).

### 2.2.5 Prognostic Predictive Model for Multiple Diseases

Traditional PPMs estimate risk for individual diseases; however, many chronic diseases have been found to cluster in the same individuals, whether they occurred at the same time or accumulated over a period of time (20–22). As a result, patients are often burdened by multiple diseases. PPMs for multiple diseases have been poorly studied in the literature. In order to enable the development of PPMs for multiple diseases, advanced statistical methodologies and data that are numerous and highly descriptive, such as electronic medical record data, are required.

## 2.3 Methodologies for Multiple Disease Risk Estimation

To understand the methodologies required for the estimation of multiple disease risk, an understanding of the concepts related to joint distributions and dependence are first required. Subsequently, an overview of one method for the estimation of multiple disease risk is presented.

### 2.3.1 Joint and Marginal Distribution of Binary Random Variables

Consider two random variables $X_d$ and $X_h$, each of which can take the value 0 or 1. Let $X_d = 1$ represent the event that an individual develops diabetes within a 5-year period and $X_d = 0$ represent no development of diabetes within that interval. Let $X_h$ be the analogous random variable, but for hypertension.

After waiting for the 5-year period, the values $(x_d, x_h)$ are observed, which are the *realizations* of the two random variables. There are four possibilities: (0,0) if the patient develops neither disease; (0,1) if the patient develops hypertension but not diabetes; (1,0) if the patient develops diabetes but not hypertension; and (1,1) if the patient develops both diseases. Note that these four possibilities are mutually exclusive and exhaustive – the patient must fall into exactly one of these categories.

By observing many patients, the probability of observing any one of these realizations may be estimated. The four probabilities give the *joint distribution* of the two random variables. (Note, however, that since they sum to one, if three probabilities are known we can compute the fourth.) A *marginal distribution* refers to the distribution of one disease without consideration of the other; each variable has a marginal distribution. In this example, one marginal distribution would describe the probabilities of developing diabetes without considering hypertension while the other would describe the probability of developing hypertension without considering diabetes. These can be computed by *marginalizing* out the other variable – for example, the marginal probability of developing diabetes would be the probability of observing (1,0) plus the probability of observing (1,1).

Suppose the development of diabetes and the development of hypertension were truly independent of each other; that is, the occurrence of diabetes was not related to the occurrence of hypertension. In this case, the joint distribution of diabetes and hypertension would be no different from what would be expected by multiplying the marginal probabilities of diabetes and hypertension. This is the null hypothesis commonly used when analyzing contingency tables with the $\chi^2$ test or Fisher's exact test.

The $\chi^2$ test evaluates the null hypothesis that each disease is independent of the other by comparing the expected and observed disease frequencies (166). When sample sizes are small, Fisher's exact test should be used (166).

The same methods can be applied to the analysis of three diseases. In this case, the frequencies of each combination of the three diseases are considered. Again, the $\chi^2$ test and Fisher's exact test can be used to evaluate the null hypothesis that each disease is independent of the others.

Dependence among the variables describing disease development may be observed in a population because they have common risk factors. For example, higher BMI may be associated with development of diabetes and development of hypertension. If a contingency table is analyzed using a sufficiently large population with a range of BMIs, dependence will be detected between $X_d$ and $X_h$. However, it may be that for any given value of BMI (or perhaps a small range) the development of the two diseases happens independently. Assessment of dependence after stratifying on a risk factor can be achieved by the Cochrane-Mantel-Haenszel test. The Cochrane-Mantel-Haenszel test accounts for confounding variables through stratification into discrete categories; however, this method does not directly make use of continuous variables.

Instead, a two-stage approach to model the joint distribution of disease development can be used. This approach first uses univariate multivariable logistic regression to model the marginal distribution of development of each disease and then uses a copula to combine the marginal distributions together into a joint distribution. This approach has the ability to 1) account for more than two diseases at a time; 2) adjust for both categorical and continuous variables; and 3) ultimately be used in the construction of a predictive model.

A copula is a multivariate probability distribution used to describe the non-linear dependence between multiple outcomes where each univariate marginal distribution is uniquely defined (167,168). Copulas are defined by Sklar's theorem, which states that every multivariate cumulative distribution function of the variables considered can be expressed in terms of their univariate marginal distributions and a copula (169). The copula (meaning *link* in Latin) links the univariate marginal distributions together, forming the multivariate joint distribution.

### 2.3.2 Univariate Multivariable Logistic Regression

The first step in constructing a copula is the estimation of univariate multivariable logistic regression models. Univariate multivariable logistic regression seeks to understand the relationship between multiple covariates and a single binary outcome (*univariate:* one outcome; *multivariable:* multiple covariates) (155). Univariate logistic regression is based on the logistic function, which is used for modelling the probability distribution of binary data as its output only takes on values between zero and one due to its *S-shape*. The logistic function σ(*t*) is defined as follows:

$$\sigma(t) = \frac{1}{1+\ e^{-t}}$$

The logistic function applied to a linear function of several explanatory covariates gives a logistic regression function.

$$F\ (x_i) = \frac{1}{1+\ e^{-(\beta_0+\beta_i x_i)}} = \ \sigma(\beta_0 +\ \beta_i x_i)$$

Where $\beta_0$ denotes some constant, $\beta_i$ denotes some constant(s) by which the explanatory variable(s) will be multiplied, and $x_i$ denotes the explanatory variable(s). Given knowledge of all explanatory covariates and estimates $\hat{\beta}_i$ of the coefficients, the probability of experiencing the outcome is estimated by

$$\hat{F}(x_i) = \frac{1}{1+\ e^{-(\hat{\beta}_0+\hat{\beta}_i x_i)}} = \ \sigma(\hat{\beta}_0 +\ \hat{\beta}_i x_i).$$

Odds can be estimated by applying the exponential function.

Odds = $e^{\widehat{\beta}_0+\widehat{\beta}_i x_i}$

The logistic regression coefficients were estimated using maximum likelihood estimation.

### 2.3.3 Copulas

Once univariate multivariable logistic regression models have been constructed, the copula is then constructed, which ties together the univariate models. There exist many copula functions, each with its own unique properties that allow it to model different dependence structures. For example, the Frank copula (170) exhibits weak dependence in both tails. One of the most common classes of copula functions, Archimedean copulas, is described below.

#### 2.3.3.1 Archimedean copulas

Archimedean copulas encompass a variety of copula functions that can all be characterized by an explicit formula. Archimedean copulas are commonly the preferred method of dependence modelling due to their ability to model dependence in arbitrarily high dimensions with a single parameter that governs the strength of the dependence (167,168). Archimedean copulas follow the structure:

$$C(u_1, \dots, u_d; \theta) = \psi^{[-1]}(\psi(u_1;\theta) + \cdots + \psi(u_d;\theta);\theta)$$

where $\theta$ is a parameter within some parameter space $\Theta$, $\psi$ is the *generator function* (a function unique to the copula used), and $\psi^{[-1]}$ is its pseudo-inverse given as:

$$\psi^{[-1]}(t;\theta) = \begin{cases} \psi^{-1}(t;\theta) \text{ if } 0 \leq t \leq \psi(0;\theta) \\ 0 \text{ if } \psi(0;\theta) \leq t \leq \infty. \end{cases}$$

As seen, the generator function determines the copula function; $\theta$ must be estimated based on the dependence that exists between variables. Larger values of $\theta$ correspond to larger amounts of dependence between diseases.

For example, the generator function for the Gumbel copula is (171):

$$(-\log(t))^{\theta}$$

which gives the following function:

$$C_\theta(u,v) = \exp\left[-((-\log(u))^\theta + (-\log(v))^\theta)^{\frac{1}{\theta}}\right]$$

$$\text{where } \theta \in [1, \infty)$$

The Gumbel copula (171) is an asymmetric copula that exhibits greater dependence in the positive tail than in the negative tail. Other examples include the Frank copula (170), which is a symmetric copula that is used to model weak dependence and the Clayton copula (172), which is an asymmetric copula function that exhibits greater dependence in the negative tail than in the positive. Many other copula functions exist, each with its own unique properties.

Copulas have been used commonly in finance, where financial distributions often are non-normal. For example, copulas have been used extensively in the area of financial risk management (54). During a recession, investors who hold positions in riskier assets, such as real estate, may move their investments into safer alternatives, such as cash or bonds. This trend results in an asymmetric distribution, where correlations across equities are greater in the downward direction compared to the upward direction. Copulas aid by modelling the marginal distributions separately from the dependence structure. In this example, marginal models can describe the behaviour of individual investors. However, the actions of one investor are not independent of those of other investors; thus, copulas allow the modelling of the behaviour of investors while considering the actions of other investors. Copulas have also been used in the areas of engineering (55), neuroscience (56,57), and climate and weather research (58,59).

## 2.4 Electronic Medical Records

As mentioned, one potential source of data for the development of a PPM for multiple diseases is EMR data. Explained in greater detail below, these data sources contain the medical records, including diagnoses, prescriptions, treatments and laboratory results, of thousands of patients that may enable the estimation of the risk of multiple diseases.

### 2.4.1 Overview

Electronic medical records (EMRs) are software programs used to store patient information electronically. Traditionally, patient information has been stored in paper records; however, there has been a shift from using paper records for this purpose to using EMRs (173). The goal of an EMR is to support the delivery of quality care by providing accessible and structured storage of information. These digital records contain individual patient information describing demographics, medical history, medications, allergies, laboratory test results, radiology images, vital signs, patient characteristics such as height and weight, risk factor information, and billing information (174).

Data are stored within an EMR in a variety of ways. Data can be stored in a highly structured manner, such as pick-lists or drop-down menus, or highly unstructured manner, such as free-text. For example, disease information such as a diagnosis of diabetes may be included in the EMR as an entry in the billing table or problem list with the corresponding International Classification of Disease (ICD) code. Alternatively, a diagnosis of diabetes could simply be noted in the free-text narrative portion. Indeed, there are often multiple ways to store the same information within the EMR (175);  thus, data of interest may be found in multiple locations within the EMR. All data within the record have an associated date and time, allowing PCPs to look back in the record to observe changes over time.

EMRs support many functions beyond the mere storage and retrieval of information, including billing services, appointment scheduling, referral services, laboratory test requisitions, and medication prescriptions (173,174). Furthermore, EMRs often support other functions known as decision support tools. Examples of these tools include medication interaction tools, which alert PCPs to potential interactions between medications when prescribing (176); clinical guidelines, which provide easy access to evidence-based guidelines (177); and risk assessments, which estimate a patient's risk of experiencing some future outcome (178).

EMRs are commonly developed and maintained by private vendors; however, open source options, such as OSCAR (179), exist. Canada does not have one single EMR software program, as health care is managed at the provincial and territorial levels. Instead, several EMR vendors exist, each with their own EMR software, competing for PCP and hospital business. As a result, there is no single repository containing the health records of all Canadians.

### 2.4.2 Uptake of EMRs in Canada

The current rate of EMR use in Canada is more than twice that of 2009; thirty seven percent of PCPs used an EMR to store patient information in 2009, whereas 73% of PCPs reported doing so in 2015 (180). While this recent increase in uptake is promising, Canada still falls below the international average by 15% (180). Provincial rates were found to be quite variable, with Alberta at 85% adoption and New Brunswick at 40% adoption (180). Given their level of use and potential to support clinical care, the extent to which EMRs are being utilized has been examined. Only 41% of Canadian PCPs use EMRs to support quality of care decisions, such as drug interaction tools or reminders for regular care or screening tests, compared to 58% internationally (180).

### 2.4.3 Use of EMRs for Research Purposes

EMRs represent a rich source of information describing a patient's health and health care. EMR databases can be linked together to form large repositories of patient information that allow for health surveillance to inform clinical and epidemiological research, public health interventions, health care planning, and quality assurance. For example, the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) is a collection of EMR databases from across Canada that contains primary care information on more than 1.5 million patients (181). This database has been used for surveillance of chronic diseases including hypertension (92), depression (182), and diabetes (59). Information are only recorded in an EMR where deemed clinically relevant by the EMR user; information such as physical activity, occupation, ethnicity, family history or other characteristics that may be important for research purposes are often not noted in EMRs. However, an EMR is a great source of population-level data pertaining to patient characteristics including diagnoses, laboratory results, medication prescriptions, and referral patterns.

The form that data are recorded in the database greatly impacts its utility for research purposes. Data are readily analyzed when stored in a structured form that arises from the use of drop-down boxes or pick-lists. When data are stored in the form of a free-text narrative, methods such as natural language processing (183) must be used to extract information from the data. One suggested method of improving the usability of EMR databases for research is to encourage PCPs to engage in consistent and accurate coding of patient information (184).

EMR data only describe the population receiving primary care, not necessarily the general population. A study looking at the representativeness of the CPCSSN national database found that, compared to data from the 2011 census, CPCSSN patients were somewhat representative of the Canadian population (185). CPCSSN patients were roughly 4 years older on average and less likely to be male, making it is necessary to adjust for age and sex to generalize results based on CPCSSN data to the general population. When applying EMR data to a primary care population, no adjustment is necessary.

EMR data are limited by their use of diagnostic codes as proxies for health events, such as disease development. For a symptom or disease to be successfully captured within an EMR, the patient or PCP must recognize and report the symptom or disease; subsequently, the practitioner must know the proper code and record this in the EMR. Any break in this stream of events will result in failure to capture the information. This has implications for research using EMR data, where the absence of a diagnostic code is often interpreted as the absence of disease. The extent to which this impacts results depends on how well diagnoses of disease are recorded in the EMR. Diseases with more significant and clearly defined diagnostic features, such as diabetes, are better recorded within the EMR (186). The use of diagnostic codes is also problematic as diagnostic codes are not always able to fully capture the complexities of chronic diseases.

Compared to alternative sources of health information, such as health administrative data or primary data collection from observational studies, EMR data are both rich and numerous. Despite describing the health of the majority of the population in Canada, health administrative data are limited by what is captured; for example, only billing codes for the "most responsible diagnosis" are stored in health administrative databases (187). EMR databases can be used to overcome this limitation as they contain a rich history of patients' health, including past and current diagnoses, medications, laboratory results, and radiographic images (174). In Canada, EMR databases do not contain records of the entire population, whereas health administrative databases contain data wherever a patient has received care due to the remuneration methods employed in Canada; however, the data that are collected in EMR repositories such as CPCSSN are often sufficient to allow for analyses at the provincial and national levels (185). Primary data collection obtains precisely what patient characteristics are of interest using a consistent method or measure; comparatively, EMR data are only collected where clinically relevant and often do not describe the measure used (175). Primary data collection, however, requires significantly greater resources when compared to EMR data (188), where the data have been previously collected.

**Table 1: Comparison of Data Sources**

| **EMR** | **Observational Studies** | **Health Administrative Data** |
|---|---|---|
| Large sample size | Small sample size | Large sample size |
| Contain all diagnostic codes recorded in a single encounter | - | Contain only one diagnostic code per encounter |
| Collect only data deemed clinically relevant | Collect all data of interest | Collect only data deemed necessary for administration/billing |
| Measurement method unknown | Data collected using a standardized measure | Measurement method unknown |

## 2.5 Summary

The current literature demonstrates the need for novel techniques aimed at the prevention of chronic disease. In particular, multimorbidity is a pressing concern for which prevention techniques remain underdeveloped. Prognostic predictive models present an opportunity for such a technique that might allow insight into a patient’s risk of multimorbidity. Such insight might allow for targeted interventions aimed at reducing patient risk. EMRs may contain the data needed for the development of these models, as these data sources contain health information of numerous patients over time.

Chapter 3

# 3 Methods

The following chapter describes the development of a prognostic PPM for multiple chronic diseases using data from EMRs. Logistic regression and copula modelling were used in model development.

## 3.1 Data Source

Data were derived from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) database (181). Initially formed in 2008 through funding provided by the Public Health Agency of Canada (PHAC), this nation-wide database contains patient information from EMRs of primary care practices across Canada (184). The objective of this network is to enable both the surveillance of chronic disease and primary care research at a national level. CPCSSN aims to accomplish these goals by collecting clinical data that provide insight into the health of Canadians from a primary care perspective through clinical and epidemiological research.

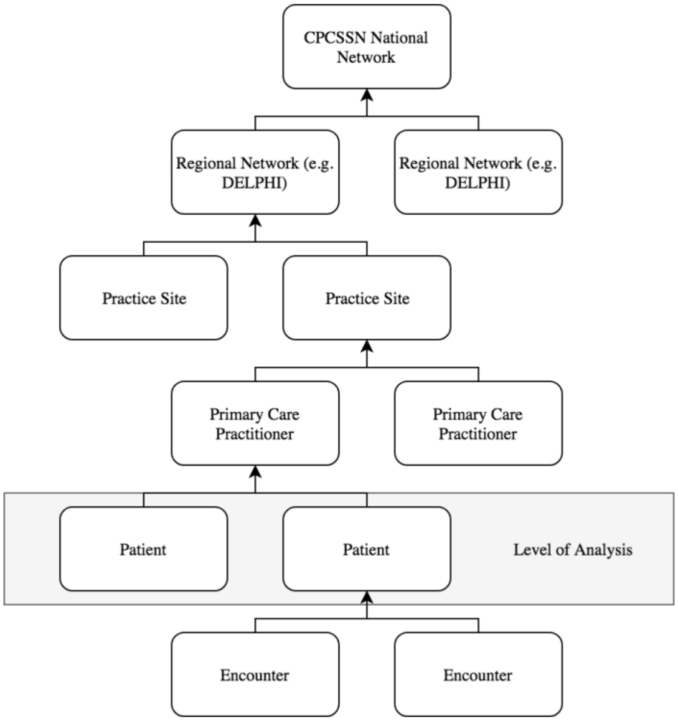
CPCSSN National Network
Regional Network (e.g. DELPHI)
Regional Network (e.g. DELPHI)
Practice Site
Practice Site
Primary Care Practitioner
Primary Care Practitioner
Patient
Patient
Level of Analysis
Encounter
Encounter

**Figure 2: CPCSSN Structure**

The CPCSSN database follows a hierarchical structure: individual patient encounters (visits with their PCP) are collected for each patient; these encounters are grouped by patients involved in these encounters; which are grouped by the PCP from which they receive care; which are grouped by the primary health care (PHC) site in which they practice; which are grouped by the network to which they contribute their data; which are then contained within the CPCSSN database. Originally involving 7 academic primary care research networks across 4 provinces (Newfoundland, Quebec, Ontario, and Alberta) (184), the CPCSSN database now involves 12 regional networks across 8 provinces and territories. Initial recruitment of practices occurred from 2008 to 2010, in which family practices (mostly those associated with academic or university sites, as these were more likely to participate in research) were recruited to contribute their data to the CPCSSN database. Following this initial period, recruitment expanded to include non-academic practices in various settings (urban, suburban, and rural). Past and ongoing patient consent was obtained via an opt-out system, in which patients who do not wish for their information to be contributed to the database may choose to opt out; all provinces operate under this system, except for Quebec, where an opt-in process is mandated by provincial law. Within these regional networks are 218 practices. Ontario, as one of the founding and most populated provinces, has the greatest number of participating practices. British Columbia and Quebec make smaller contributions as British Columbia is a relatively new network and legislative requirements in Quebec deter the process. The CPCSSN database contains records from 1189 PCPs. Data describing the nature of PCP's practice, such as profession (i.e. physician or nurse practitioner) or payment model (e.g. fee-for-service or capacitation), are unavailable for most PCPs. CPCSSN contains deidentified records of more than 1.5 million patients, making it the largest source of primary care information available in Canada.

The CPCSSN database is comprised of several data tables containing information pertaining to either the practice, provider, patient, or patient encounter. For example, the *Billing* table contains all ICD-9 codes used by the provider to submit a billing claim; these data can be used to identify diagnoses made by the provider. Note, however, that providers are limited to one diagnosis per patient encounter, thus the diagnosis recorded is known as the *most responsible diagnosis*. The *Health Condition* table contains additional diagnoses made during an encounter, regardless of whether they were billed for; this is congruent with the *problem list* used in other EMRs. From the tables contained within CPCSSN, all structured patient records can be extracted, including patient demographics, billing codes, laboratory results, prescriptions, referrals, risk factor information, medical procedures, vaccinations, and allergies. For privacy reasons, the free-text narrative where PCPs record their notes is not available in CPCSSN. Tables were linked using an identifier unique to each patient.

**Table 2: CPCSSN Data Tables**

| **Table Name** | **Contents** | **Format** | **Completeness** |
|---|---|---|---|
| Billing | Diagnoses | ICD-9 Codes | - |
| Health Condition | Diagnoses | ICD-9 Codes<br>Free text | - |
| Encounter Diagnosis | Diagnoses | ICD-9 Codes<br>Free Text | - |
| Patient | Age | Numeric | 99.8% |
| | Sex | Text | 99.9% |
| Patient Demographic | Occupation | Text | 5.1% |
| | Highest education | Text | 2.0% |
| | Housing status | Text | 4.4% |
| | Forward Sortation Area (FSA) | Text | 95.4% |
| | Language | Text | 14.1% |
| | Ethnicity | Text | 1.0% |
| Lab | Laboratory results | Numeric | - |
| Exam | Examination results | Numeric | - |
| Medication | Medication prescriptions | Text | - |
| Family History | Reported family history | Text | - |
| Risk Factor | Reported risk factors | Text | - |
| Medical Procedure | Medical procedures | Text | - |
| Referral | Referrals | Text | - |
| Vaccine | Vaccines received | Text | - |
| Allergy/Intolerance | Allergies and intolerances | Text | - |
| Disease Case | Validated cases of disease | Text | - |
| Provider | Age | Numeric | 87.8% |
| | Sex | Text | 98.3% |
| (Provider) Group Info | Group type | Text | 75.9% |
| | Payment model | Text | 3.4% |
| Site | Province | Text | 100% |

Due to the volunteer basis of practice recruitment, the CPCSSN database can be seen as a convenience sample of primary care patients across Canada. CPCSSN patients are somewhat representative of the Canadian general population (185). Provincial-level comparisons are appropriate for all included provinces, except for British Columbia and Quebec due to their low participation (185). As of 2013, CPCSSN patients were older and more likely to be female compared to the overall Canadian population as reported in census data. Research has demonstrated that this trend is typical of primary care (189–191). Compared to practitioners responding to the National Physician Survey, CPCSSN practitioners were younger, more likely to be female (51.1 vs. 44.0%), and from an academic practice (19.3% vs. 7.8%) (185).

First, the construction of a PPM requires an understanding of what risk factors are known to increase the risk of disease development. Next, a cohort of people whose risk factor status at baseline and their subsequent disease outcome are known is needed. From this cohort, multivariable models are built to describe the associations between each risk factor and the disease outcome.

## 3.2 Measures

### 3.2.1 Outcome

The following diseases were predicted simultaneously: diabetes, hypertension, and osteoarthritis. These diseases were selected based on several criteria. First, the selected diseases are among the most prevalent in Canada (53,192). Previously validated case-detecting algorithms for use with EMR data exist for these diseases (described more fully below) (193). These diseases are often diagnosed and treated in primary care (54,85,105). Each of these diseases have modifiable risk factors, some of which overlap between diseases (35,36,63,70). Finally, expert consultation revealed that risk estimations for the selected diseases, in particular their co-occurrence, would be clinically useful. In this work, recovery from disease was considered not possible; once a patient has one of the diseases, they will always have the disease.

Chronic pain and asthma were also among the most prevalent diseases; however, neither have a validated case-detecting algorithm.

One initiative of CPCSSN researchers has been to develop and validate case detecting algorithms for several chronic diseases that can be used to identify cases of disease within the database (193). In an effort to facilitate quality research, CPCSSN has created disease case-detecting algorithms for osteoarthritis, depression, hypertension, diabetes, chronic obstructive pulmonary disease (COPD), epilepsy, Parkinson’s disease, and dementia. These case-detecting algorithms are composed of information including ICD-9 codes within the billing or problem list; medication prescriptions; laboratory results; or any combination of these elements. Construction of the case-detecting algorithms was informed by published evidence and input from both primary care and specialist physicians. Subsequently, each disease case-detecting algorithm was validated by chart review. Chart review was performed by research assistants blinded to the diagnosis assigned by the algorithm. Reviewers determined the absence of disease through examination of the entire electronic medical record. Where a reviewer was uncertain, the study epidemiologist and a physician from the study team performed a chart review. The ability of the case-detecting algorithm to correctly assign diagnoses was assessed by comparing its results to those of the chart review, resulting in both sensitivity and specificity statistics. Sensitivity and specificity for all diseases were high (Appendix A). In this thesis, diagnoses of diabetes, hypertension, and osteoarthritis were identified using the case-detecting algorithms developed by CPCSSN researchers.

The use of validated disease case-detecting algorithms helps ensure that the identification of disease cases is accurate. This is especially important as inaccuracy in the identification of the disease will decrease a predictive model's performance due to incorrect estimation of the relationships between the predictors and actual disease development. This poor performance would not be revealed by internal validation because the data used for validation would be subject to the same issue of inaccuracy in disease identification as the data used for constructing the model. Often only internal validation is feasible, reinforcing the importance of using a validated case-detecting algorithm for the identification of disease cases. However, the correctness of predictors is not as crucial, since the main goal of this analysis was not etiologic research, but the prediction of future disease development. For example, a diagnosis of osteoporosis is a risk factor for osteoarthritis; however, the ICD-9 code used for osteoporosis also includes several other bone disorders. Despite this ICD-9 code not being specific to osteoporosis, it is still useful as a predictor for osteoarthritis because its presence in patient's EMR was found to be significantly associated with future development of osteoarthritis. Accordingly, caution must be taken when making causal inferences from the resulting predictive model since the model does not truly describe the impact of an osteoporosis diagnosis on risk of osteoarthritis, but rather the impact of the presence of the ICD-9 code.

As both predictor and outcome assessment was done by the PCP, blinding did not occur during outcome assessment, which may have introduced measurement bias. This issue is present in all EMR and health administrative data due to the nature of the data. However, the diseases predicted are common conditions with clearly identifiable diagnostic criteria, thus skilled PCPs should be able to diagnose cases of disease with a high degree of accuracy, limiting the influence of knowledge of predictor status at baseline on this assessment, and in turn limiting the amount of bias introduced.

### 3.2.2 Predictors

Predictors for each of the three diseases to be predicted were identified through review of the relevant literature.

**Table 3: Disease Predictors**

| Diabetes | Hypertension | Osteoarthritis |
|---|---|---|
| Hypertension (36,63)<br>Older age (36,63)<br>Lipid disorders (36)<br>Obesity (35,36,63,70)<br>Waist circumference<br>Smoking (36,63)<br>Stress (36)<br>Male sex (36)<br>Polycystic ovarian syndrome (PCOS) (64)<br>Schizophrenia (65,66)<br>Depression (67)<br>Bipolar disorder (66,68)<br>Low physical activity (70)<br>Family history of type 2 diabetes (36)<br>Air pollution (69)<br>Low socioeconomic status (70) | Older age (93)<br>Diabetes (93,94)<br>Obesity (93,94)<br>Smoking (93)<br>Stress (97)<br>Kidney disease (95)<br>Tricyclic antidepressant (TCA) use (98)<br>High salt intake (93,99)<br>Sleep apnea (96) | Osteoporosis (39)<br>Previous leg injury (38,40,117,119)<br>Leg length inequality (115)<br>Older age (39,40,116–118)<br>Obesity (38–40,116–119)<br>Female sex (39,40,116,117)<br>Family history of osteoarthritis (116)<br>Physically intensive occupations (116) |

**Table 4: Shared Risk Factors**

| | Diabetes | Hypertension | Osteoarthritis |
|---|---|---|---|
| Older age | X | X | X |
| Obesity | X | X | X |

|  |  |  |  |
|---|---|---|---|
| Smoking | X | X |  |
| Stress | X | X |  |
| Hypertension | X |  |  |
| Lipid disorders | X |  |  |
| Waist circumference | X |  |  |
| Male sex | X |  |  |
| Female sex |  |  | X |
| PCOS | X |  |  |
| Schizophrenia | X |  |  |
| Depression | X |  |  |
| Bipolar disorder | X |  |  |
| Low physical activity | X |  |  |
| Air pollution | X |  |  |
| Low socioeconomic status | X |  |  |
| Diabetes |  | X |  |
| Kidney disease |  | X |  |
| Tricyclic antidepressant use |  | X |  |
| High salt intake |  | X |  |
| Sleep apnea |  | X |  |
| Osteoporosis |  |  | X |
| Previous leg injury |  |  | X |
| Leg length inequality |  |  | X |
| Family history of type 2 diabetes | X |  |  |
| Family history of osteoarthritis |  |  | X |
| Physically intensive occupations |  |  | X |

For each predictor, an algorithm for the identification of each risk factor was developed. Where a CPCSSN validated case-detecting algorithm was available, this was used; otherwise the following process was used to identify information that described the predictor. First, the CPCSSN data dictionary (181) was examined to determine if any fields contained information describing the predictor exactly (for example, BMI was found in the exam table). Next, other methods of detecting the predictor were identified, then investigated for their presence within CPCSSN. These included diagnostic terms and ICD-9 codes; medications used specifically to treat a given condition; and laboratory test results indicative of a given condition (for example, an LDL measurement between 3.37 and 9 mmol/L was indicative of a lipid disorder). All diagnostic codes in the CPCSSN database are stored as ICD-9 codes, thus only ICD-9 codes were used. Multiple inclusion terms were used to capture all terminologies used to describe the condition; additionally, exclusion terms were used to exclude those that did not describe the condition. All methods of identifying predictors were reviewed by a PCP who was a member of the study team to ensure accuracy (for example, the PCP ensured that all medications used to identify predictors are medications only prescribed for the predictor condition). Subsequently, predictor information was compiled into predictor case-detecting algorithms that would be used to identify cases of predictor presence. Case definitions for each risk factor can be found in Appendix B.

An estimate of income was obtained using the Forward Sortation Area (FSA) available for most patients. Full postal codes are unavailable in the CPCSSN data for privacy reasons. Each patient's FSA, where available, was matched to average personal income according to the National Household Survey (NHS) conducted in 2011 (194). Similarly, rurality was based on postal code. The second digit of a postal code is used to denote whether the area is urban or rural. A zero indicates that the area is rural, while all other digits indicate urban areas (195). Where the second digit of the FSA was zero, the patient was considered to live in a rural area.

Interaction terms were considered; however, no interactions were suggested in the existing literature (196).

As suggested in TRIPOD (150), all continuous risk factors were kept in their original form, rather than binning them into categories, in order to maximize the amount of information available for each covariate.

## 3.3 Participants

Participants were drawn from the CPCSSN primary care database. The PPM that was developed for this thesis is intended to be deployed in primary care in Canada to address risk of multiple diseases in adults. All patients aged 18 or older were included in the cohort, irrespective of prior morbidities. Patients who have previously been diagnosed with all three diseases (diabetes, hypertension, and osteoarthritis) were excluded, as these patients were not at risk of developing any of these diseases. All eligible patients were considered for analysis.

## 3.4 Sample Size Considerations

The retrospective cohort made use of all available patient data; no sub-sampling of the CPCSSN database was done. The cohort was split into two partitions: one for model development and one for validation.

Often the minimum required size for each partition to be confident in risk estimations is determined by an anecdotal heuristic stating that for each predictor, there should be at least 10 events (in this case, 10 patients who develop the disease(s) of interest). This method has been commonly criticized for the lack of evidence supporting its use (197). However, no method has been agreed upon to determine the sufficient number of events per variable; thus, in order to maximize the number of events per variable, predictors were only selected for use where external evidence of an association existed.

## 3.5 Cohort Construction

From the time-stamped records, a retrospective cohort was constructed. To begin, all patients listed in the patient table were considered. Patient "recruitment" began 1 January 2009 and ended 31 December 2010 (a period of two years); patients who had any EMR entry (billing occurrence, encounter recording, encounter diagnosis, exam recording, or health condition recording) in the recruitment time period were included. For each patient, the date of the first record within the recruitment time period was considered the patient's unique start-date. At this point, predictors (including diabetes, hypertension, and osteoarthritis, as one may predict for another) were assessed using the disease and predictor case definitions (Appendix B). Patients who had been previously diagnosed with all 3 diseases were excluded, as these patients were not at risk of developing the diseases. Additionally, patients younger than 18 years of age were excluded. Any diagnoses of disease within the subsequent 5-year period were noted.

The cohort was randomly divided into development and validation datasets at approximately a 2:1 ratio: the development set for model selection and parameter estimation and the validation set for assessing discrimination and calibration of the resulting model.

## 3.6 Missing Data

As data in an EMR are collected for clinical purposes, not specifically for research use, data are often missing from the EMR because they are not relevant for patient care, despite being highly relevant for research. Data can be missing in a variety of ways. Data can be missing completely at random (MCAR), where the reason that the data are missing is independent of all other variables, observed or unobserved (198). For example, data that are artificially sub-sampled at random would be MCAR. Where data are MCAR and must be omitted from analysis, analyses have less power but remain unbiased. Data are missing at random (MAR) when their missingness can be explained by the value of observed variables (198). For example, a lab test result for cholesterol may be missing because a patient is observed to be young and have normal BMI. Where data are MAR, techniques can be used to impute missing data using strategies that minimize the amount of bias that is introduced. Data are missing not at random (MNAR) where their missingness is dependent upon some unobserved variable, including the missing variable itself (198). For example, a blood pressure measurement may not be recorded because it is within normal ranges. Analyses based on MNAR missing data will produce biased results (198). Most methods to address missing data assume data to be MAR; the validity of this assumption impacts the amount of bias introduced into analyses.

Imputation is the process of replacing missing values with plausible values. Depending on how the data are missing, different imputation methods can be used. Common examples of imputation include last observation carried forward (199), in which the missing value is replaced with the last value that was observed; mean substitution (199), in which the missing value is replaced with the mean of the characteristic's observed values; and regression (199), in which other observed characteristics of the individual are used to estimate a value for the missing value. These methods are single imputation methods, which do not account for the uncertainty in the imputed values (200). In contrast, multiple imputation can be used to replace missing values while accounting for the uncertainty in the imputations by creating multiple estimates for the missing value (200). In multiple imputation, several values are estimated for the missing value, creating multiple imputed datasets. There are several methods of multiple imputation. For this work, multiple imputation by chained equations (MICE) was used. MICE follows these steps, as described by Azul et al. (201): 1) A simple imputation, such as mean substitution, is used to complete all missing values. These imputed values should be thought of as placeholders. 2) One variable with missing data is selected as the variable of interest, *var*. The values for *var* that were originally missing are set back to missing. 3) These now missing values are imputed using regression based on all variables, including those containing placeholders. *Var* can be thought of as the dependent variable, for which the other variables serve as independent variables. Subsequent imputations using *var* as an independent variable for other variables will use these imputed values. 4) Steps 2 & 3 are repeated for each variable with missing data. Imputed values from previous cycles are used instead of the placeholders. 5) Steps 2 through 4 will be repeated a given number of times, updating the imputations each time, resulting in multiple imputed datasets.

Multiple imputation was used for this study, which produced multiple completed datasets. While a single point estimate will be presented for each statistic, in actuality, several were computed (one for each imputed dataset); these results were then combined using Rubin's rules (200) to create a single statistic whose variance has been adjusted to account for the uncertainty of deriving an estimate from multiple datasets.

## 3.7 Statistical Analysis

To facilitate the construction of a PPM for diabetes, hypertension, and osteoarthritis, an analysis of the dependence between these diseases was performed. As described above, copulas were selected to model the dependence between diseases because of their ability to account for more than two diseases, adjust for both continuous and discrete variables, and ultimately be used to construct a PPM. The steps in dependence modelling using copulas are: 1) univariate (marginal) models are constructed for each outcome and 2) copulas are used to describe the dependence between outcomes (202,203).

### 3.7.1 Univariate Multivariable Logistic Regression

To address Objective 1, univariate multivariable logistic regression models of the development of each disease were constructed. Three univariate models were produced, one for each disease to estimate its marginal distribution. For each disease, a subgroup of the development cohort who were free of the disease being predicted at baseline were included in the estimation of the univariate model. For example, a subgroup of patients who did not have diabetes at baseline were used to construct the diabetes univariate model. The $\hat{\beta}_i$ coefficient estimates of each model are presented along with 95% confidence intervals for the $\beta_i$. Internal validation assessing discrimination and calibration was performed. Discrimination refers to the ability of the model to assign a higher estimated risk to a person who ultimately experiences the outcome compared to a person who does not. For discrimination, models were assessed by calculating the area under the receiver operator characteristic curve (AUC). Calibration examines how well the model fits the data. For calibration, models were assessed by constructing calibration plots. As the dataset was extremely large, methods such as cross-validation (161) or bootstrapping (162) were not necessary; instead, discrimination and calibration were assessed using the validation set.

### 3.7.2 Analysis of Dependence

To address Objective 2, an analysis of the dependence between each outcome was conducted both with and without adjustment for risk factors, in a purely descriptive (non-predictive) framework and then in a predictive framework. Each analysis of dependence was conducted in a pairwise fashion; specifically, the dependence between diabetes and hypertension, diabetes and osteoarthritis, and hypertension and osteoarthritis was estimated. To be included in a pairwise analysis, a patient had to be free of both diseases under investigation at baseline. For example, only patients free of both diabetes and hypertension at baseline were included in any analysis of the dependence between diabetes and hypertension.

To begin, the pairwise unadjusted correlation between each outcome was measured using the $\phi$ coefficient (also known as the mean square contingency coefficient). The $\phi$ coefficient is a measure of association between two binary variables, similar to the Pearson correlation coefficient for continuous variables (204). In fact, estimating a Pearson correlation coefficient for two binary variables gives the $\phi$ coefficient (204). Pairwise $\phi$ coefficients were calculated along with the corresponding test statistic and 95% confidence interval.

Partial correlation examines the correlation that exists between variables after adjusting for the effect of other variables; again, this is measured using the $\phi$ coefficient. Partial correlations were determined for each outcome pair, adjusted for the combined risk factors for each outcome using the function *pcor* from the *ppcor* R package (205).

Subsequently, copulas were used to describe the dependence between outcomes. The choice of copula was determined by the structure of the dependence. For this study, the Frank copula (170) was selected for used based on its ability to capture weak dependence. When modelling the dependence between binary variables, the copula is defined by both *θ* and the marginal distributions (202). As such, the two-stage estimation procedure based on the composite likelihood suggested by Zhao and Joe (203) was used for the estimation of *θ*. First, the marginal models were determined using the maximum likelihood estimation procedure. This step yielded *β* estimates that were used in the second step. From these univariate models, the probabilities for the independent occurrence of each outcome were estimated, by:

$$\pi_j(\mathbf{x}) = \frac{\exp(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_j)}{1 + \exp(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_j)}$$

where $\boldsymbol{\beta}_j$ is a vector containing the *β* estimates for each outcome *j* and **x** is a matrix of covariate data. Second, estimates of *θ* were obtained, again using the maximum likelihood estimation procedure. This process made use of the bivariate conditional distributions of each outcome pair. From these, the likelihood function was constructed. By setting the derivative of the log likelihood function (known as the score function, $s_\theta$) equal to zero, *θ* was estimated.

$$s_\theta(\theta, \beta_k, \beta_l) = \sum_{i=1}^{n} \dot{C}_\theta(\bar{\pi}_{ik}, \bar{\pi}_{il}) \left( \frac{(1 - Y_{ik})(1 - Y_{il})}{C_\theta(\bar{\pi}_{ik}, \bar{\pi}_{il})} - \frac{(1 - Y_{ik})Y_{il}}{\bar{\pi}_{ik} - C_\theta(\bar{\pi}_{ik}, \bar{\pi}_{il})} - \frac{Y_{ik}(1 - Y_{il})}{\bar{\pi}_{il} - C_\theta(\bar{\pi}_{ik}, \bar{\pi}_{il})} + \frac{Y_{ik}Y_{il}}{1 - \bar{\pi}_{ik} - \bar{\pi}_{il} + C_\theta(\bar{\pi}_{ik}, \bar{\pi}_{il})} \right)$$

$$\dot{C}_\theta(u, v) = \frac{\frac{e^\theta \theta((u-1)\left(-e^{\theta v}\right) - e^{\theta(u+v)} + u e^{\theta v + \theta} - (v-1)e^{\theta u} + v e^{\theta u + \theta} - e^\theta(u+v) + u + v - 1)}{(e^\theta - 1)(-e^{\theta(u+v)} + e^{\theta u + \theta} + e^{\theta v + \theta} - e^\theta)} + \ln\left(\frac{\left(e^{-\theta u} - 1\right)\left(e^{-\theta v} - 1\right)}{e^{-\theta} - 1} + 1\right)}{\theta^2}$$

$$\bar{\pi}_{ik} = 1 - \pi_{ik}$$

$$\bar{\pi}_{il} = 1 - \pi_{il}$$

where $C_\theta$ is the copula function; $\dot{C}_\theta$ is the derivative of the copula function; $\pi_{ik}$ and $\pi_{il}$ are estimated probabilities of disease $k$ and $l$ for patient $i$ based on their univariate models, respectively; and $Y_{ik}$ and $Y_{il}$ are the observed disease outcomes for patient $i$.

A dependence structure using copulas is completely specified by its univariate multivariable models and copula, which is specified by its $\theta$ estimate. Estimates of the parameter $\theta$ were obtained for each disease pair. Bootstrapping was used to construct confidence intervals for the $\theta$ estimates; the percentile method was used (206), in which the sample means at the 2.5 and 97.5 percentiles were used to approximate the confidence interval based on one thousand bootstrapped replicates. Additionally, the following hypothesis test based on the score test was used to test the null hypothesis that the observed outcome frequencies are no different than what would be expected under independence (202). The null hypothesis was rejected if $z_{obs}$ is larger in absolute value than a critical value derived from the standard Normal distribution, denoted $N(0,1)$.

$$z_{obs} = \sum_{i=1}^{n} \frac{\dot{C}_{\theta_0}(\hat{\bar{\pi}}_{ik}, \hat{\bar{\pi}}_{il})(Y_{ik} - \hat{\pi}_{ik})(Y_{il} - \hat{\pi}_{il})}{\hat{\pi}_{ik}\hat{\pi}_{il}\hat{\bar{\pi}}_{ik}\hat{\bar{\pi}}_{il}} \Big/ \sqrt{\sum_{i=1}^{n} \frac{\dot{C}^2_{\theta_0}(\hat{\bar{\pi}}_{ik}\hat{\bar{\pi}}_{il})}{\hat{\pi}_{ik}\hat{\pi}_{il}\hat{\bar{\pi}}_{ik}\hat{\bar{\pi}}_{il}}}$$

Based on these copula models, trivariate probabilities that account for the dependence between outcomes can be estimated; that is, the probabilities of each combination of diseases will be estimated. Each trivariate probability can be described as a probability mass function.

$$p(x_1, x_2, x_3) = p(X_1 = \ x_1, X_2 = \ x_2, X_3 = \ x_3)$$

Bivariate probability mass functions can be used to describe the marginal distributions of the trivariate probability mass functions.

$$p(x_1, x_2) = \sum_{x_3 \in \{0,1\}} p(x_1, x_2, x_3)$$

Similar expressions are true for $p(x_1, x_3)$ and $p(x_2, x_3)$. Based on $\hat{p}(x_1, x_2)$, $\hat{p}(x_1, x_3)$, and $\hat{p}(x_2, x_3)$ as estimated by the copula model, trivariate probability mass functions ($\hat{p}(x_1, x_2, x_3)$) can be found that satisfy the bivariate marginal distributions. In fact, there may be many combinations of trivariate probability mass functions that satisfy this relationship; the combination with the highest entropy (207) (highest uncertainty) was chosen, as this gives the most conservative estimate. From this analysis, the trivariate probabilities can be estimated. For example, the risk of developing diabetes, hypertension, and osteoarthritis all within a 5-year window can be estimated.

# Chapter 4

# 4 Results

The following chapter first provides descriptive statistics about the study cohort. This is followed by the analysis of dependence using copulas. First, the univariate multivariable logistic regression models are presented. Next, both unadjusted and adjusted dependence analyses, including the copulas, are presented. Finally, the copulas are used to estimate the risk of multiple diseases for two simulated patients.

## 4.1 Descriptive Statistics

A cohort of 425228 adult patients who did not have comorbid diabetes, hypertension, and osteoarthritis who had received care between 1 January 2009 and 31 December 2010 were followed for 5 years. Figure 4 details the flow of patients into the cohort. The final cohort was split into a development set of 265228 patients (62%) and a validation set of 160000 patients (38%). Most patients began the period of study without morbidities (70%) (the following diseases were considered when assessing morbidities: asthma, arthritis, COPD, diabetes, cardiovascular disease, mental disorder (mood disorder and/or anxiety), Alzheimer's disease and related dementias, cancer, and stroke). The most common condition was having a lipid disorder (17.9%). The majority of patients were female (58.1%), which is typical of a primary care population (189–191). The average age of patients was 47.1 years old (standard deviation: 18.0 years). Most patients were overweight or obese (64.1%). After the 5-year period, hypertension was the most commonly acquired disease, with an incidence proportion of 9.4% and an incidence rate of 0.0818 events/person-year, followed by diabetes with an incidence proportion of 4.4% and an incidence rate of 0.0413 events/person-year. Osteoarthritis was developed by the least number of patients, with an incidence proportion of 3.0% and an incidence rate of 0.0248 events/person-year. No significant differences between the development and validation sets were observed. For a detailed description of all patient characteristics, see Table 6; note that each percentage denotes the percent of patients with the risk factor among those who had complete information for that risk factor. Each risk factor has been compared to its national prevalence from approximately 2010.

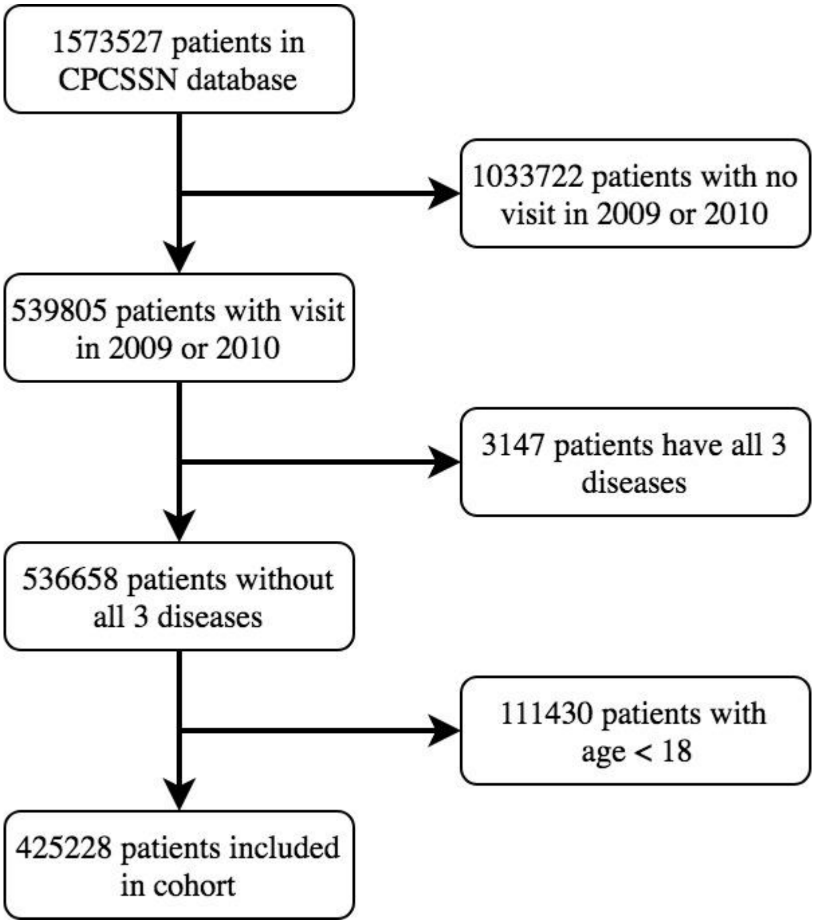
1573527 patients in CPCSSN database
1033722 patients with no visit in 2009 or 2010
539805 patients with visit in 2009 or 2010
3147 patients have all 3 diseases
536658 patients without all 3 diseases
111430 patients with age < 18
425228 patients included in cohort

**Figure 3: Cohort Construction**

**Table 5: Incidence of Diabetes, Hypertension, and Osteoarthritis**

| | Entire Cohort (*n* = 425228) | Development Set (*n* = 265228) | Validation Set (*n* = 160000) |
|---|---|---|---|
| Diabetes | | | |
| Incidence Proportion, *n* (%) | 18769<br>(4.4%) | 11677<br>(4.4%) | 7092<br>(4.4%) |
| Incidence rate, events/person-year (95% CI) | 0.0415<br>(0.0409 to 0.0421) | 0.0413<br>(0.0406 to 0.0421) | 0.0418<br>(0.0408 to 0.0428) |
| Hypertension | | | |
| Incidence Proportion, *n* (%) | 39882<br>(9.4%) | 24828<br>(9.4%) | 15054<br>(9.4%) |
| Incidence rate, events/person-year (95% CI) | 0.0818<br>(0.0810 to 0.0826) | 0.0816<br>(0.0806 to 0.0827) | 0.0820<br>(0.0807 to 0.0833) |
| Osteoarthritis | | | |
| Incidence Proportion, *n* (%) | 12803<br>(3.0%) | 7980<br>(3.0%) | 4823<br>(3.0%) |
| Incidence rate, events/person-year (95% CI) | 0.0248<br>(0.0243 to 0.0252) | 0.0248<br>(0.0242 to 0.0253) | 0.0248<br>(0.0241 to 0.0255) |

**Table 6: Descriptive Statistics**

| | Entire Cohort (*n* = 425228) | | Development Set (*n* = 265228) | | Validation Set (*n* = 160000) | | |
|---|---|---|---|---|---|---|---|
| | *n* cases | % | *n* cases | % | *n* cases | % | National Prevalence |
| Osteoarthritis | 41853 | 9.8% | 26013 | 9.8% | 15840 | 9.9% | 13.0%[a] |
| Diabetes | 28979 | 6.8% | 18140 | 6.8% | 10839 | 6.8% | 8.7%[b] |
| Hypertension | 66030 | 15.5% | 41185 | 15.5% | 24845 | 15.5% | 17.6%[c] |
| Depression | 61977 | 14.6% | 38629 | 14.6% | 23348 | 14.6% | 11.3%[d] |
| Smoking | 17844 | 63.9% | 11037 | 63.8% | 6807 | 64.2% | 13.7%[e] |
| Female Sex | 246866 | 58.1% | 153664 | 57.9% | 93202 | 58.3% | 50.4%[f] |
| Alcohol | 6467 | 1.5% | 4038 | 1.5% | 2429 | 1.5% | 2.4%[g] |
| Stress | 12636 | 3.0% | 7907 | 3.0% | 4729 | 3.0% | 22.9%[h] |
| Epilepsy | 2979 | 0.7% | 1842 | 0.7% | 1137 | 0.7% | 0.4%[i] |
| Schizophrenia | 6379 | 1.5% | 3955 | 1.5% | 2424 | 1.5% | 1.0%[j] |
| Anxiety | 30326 | 7.1% | 18894 | 7.1% | 11432 | 7.1% | > 12%[k] |
| Cancer | 17653 | 4.2% | 11139 | 4.2% | 6514 | 4.1% | 7.1%[l] |
| CVD | 23502 | 5.5% | 14730 | 5.6% | 8772 | 5.5% | 5.4%[h] |
| COPD | 7265 | 1.7% | 4515 | 1.7% | 2750 | 1.7% | 8.7%[m] |
| Rheumatoid Arthritis | 3263 | 0.8% | 2039 | 0.8% | 1224 | 0.8% | 0.9%[n] |
| Lipid Disorder | 76253 | 17.9% | 47619 | 18.0% | 28634 | 17.9% | 17.3%[o] |
| PCOS | 1154 | 0.5% | 706 | 0.5% | 448 | 0.5% | 6.5%[p] |
| CKD | 14767 | 3.5% | 9283 | 3.5% | 5484 | 3.4% | 3.1%[q] |
| TCA | 13035 | 3.1% | 8114 | 3.1% | 4921 | 3.1% | |
| Osteoporosis | 14384 | 3.4% | 8971 | 3.4% | 5413 | 3.4% | 10.0%[r] (40+) |
| Leg Injury | 12411 | 2.9% | 7808 | 2.9% | 4603 | 2.9% | |
| Family History of Osteoarthritis | 282 | 0.1% | 168 | 0.1% | 114 | 0.1% | |
| Family History of DM | 4578 | 1.1% | 2851 | 1.1% | 1727 | 1.1% | |

| | Entire Cohort (*n* = 425228) | | Development Set (*n* = 265228) | | Validation Set (*n* = 160000) | | |
|---|---|---|---|---|---|---|---|
| | *n* cases | % | *n* cases | % | *n* cases | % | National Prevalence |
| Family History of Hypertension | 2904 | 0.7% | 1817 | 0.7% | 1087 | 0.7% | |
| Lives in a rural location | 88898 | 20.9% | 55527 | 20.9% | 33371 | 20.9% | 19.0%[s] |
| Morbidity | | | | | | | |
| -1 disease* | 127781 | 30.0% | 79671 | 30.0% | 48110 | 30.1% | 38.4%[t] |
| Multimorbidity | | | | | | | |
| -2 disease* | 37679 | 8.9% | 23565 | 8.9% | 14114 | 8.8% | 14.5%[u] |
| -3 disease* | 10063 | 2.4% | 6286 | 2.4% | 3777 | 2.4% | 4.9%[u] |
| Age | | | | | | | |
| -18 to 24 | 58947 | 13.9% | 36962 | 13.9% | 21985 | 13.7% | 12.0%[v] |
| -25 to 44 | 143660 | 33.8% | 89608 | 33.8% | 54052 | 33.8% | 34.5%[v] |
| -45 to 64 | 152924 | 36.0% | 95161 | 35.9% | 57763 | 36.1% | 35.8%[v] |
| -65 and older | 69438 | 16.3% | 43330 | 16.3% | 26108 | 16.3% | 17.7%[v] |
| BMI | | | | | | | |
| -Underweight (< 18.5 kg/m$^2$) | 2694 | 1.9% | 1680 | 1.9% | 1014 | 1.9% | |
| -Normal (18.5 to 24.9 kg/m$^2$) | 48920 | 34.0% | 30541 | 34.1% | 18379 | 33.9% | 32%[w] |
| -Overweight (25 to 29.9 kg/m$^2$) | 49380 | 34.3% | 30736 | 34.3% | 18644 | 34.4% | 40%[w] |
| -Obese (> 30 kg/m$^2$) | 42834 | 29.8% | 26639 | 29.7% | 16195 | 29.9% | 27%[w] |
| Personal Income | | | | | | | |
| -Less than $30000 | 2243 | 0.6% | 1401 | 0.6% | 842 | 0.6% | |
| -$30000 to $49999 | 297791 | 73.9% | 185981 | 74.0% | 111810 | 73.8% | |
| -$50000 to $74999 | 102784 | 25.5% | 64016 | 25.5% | 38768 | 25.6% | |
| -Greater than $75000 | 7 | 0.0% | 6 | 0.0% | ** | 0.0% | |

[a] Canadian Chronic Disease Surveillance System (CCDSS) 2013/14(28)
[b] Canadian Chronic Disease Surveillance System (CCDSS) 2008(208)
[c] Canadian Community Health Survey (CCHS) 2011 (209)
[d] Canadian Community Health Survey (CCHS) 2012 (210)
[e] Canadian Tobacco Use Monitoring Survey (CTUMS) 2012 (208)
[f] Statistics Canada (211)
[g] National Population Health Survey (NPHS) 2006 (212)
[h] Canadian Community Health Survey (CCHS) 2009/10 (208)
[i] Canadian Community Health Survey (CCHS) 2010/11 (213)
[j] Public Health Agency of Canada (214)
[k] Offord et al (215)
[l] Canadian Community Health Survey (CCHS) 2015 (29)
[m] Canadian Chronic Disease Surveillance System (CCDSS) 2008 (208)
[n] Ontario Rheumatoid Arthritis administrative Database (ORAD) (216)
[o] Canadian Health Measures Survey (CHMS) 2009/10 (208)
[p] Lujan et al. (217)
[q] Canadian Health Measures Survey (CHMS) 2007/08 (218)
[r] 2009 Canadian Community Health Survey – Osteoporosis Rapid Response (219)
[s] Statistics Canada (220)
[t] Canadian Community Health Survey (CCHS) 2014 (29)
[u] Canadian Community Health Survey 2011/12 (208)
[v] Canadian Census 2012 (221)
[w] Canadian Health Measures Survey (CHMS) 2009/10 (222)

*Morbidity and multimorbidity considered the following diseases: asthma, arthritis, COPD, diabetes, cardiovascular disease, mental disorder (mood disorder and/or anxiety), Alzheimer's disease and related dementias, cancer, stroke.

**Cell counts of 5 or less have been suppressed.

## 4.2 Missing Data and Multiple Imputation

Table 7 displays the amount of missing data in fields where data were missing, such as age or BMI. All other variables were assessed under the assumption that the absence of an indication of risk factor presence signified that the risk factor was not present in the individual. However, in some cases, when compared to national averages, this assumption seemed unreasonable. For example, a diagnosis of polycystic ovarian syndrome (PCOS) was found in only 0.5% of women, whereas the national average of PCOS among women was 6.5%. As this seemed implausible, PCOS was not considered in any analyses. The same approach was used in removing alcohol use. Information regarding family history of the diseases of interest was not readily available for most patients as several networks did not collect this information; family history was removed accordingly.

**Table 7: Variables with Missing Data**

| | **Entire Cohort (*n* = 425228)** | | **Development Set (*n* = 265228)** | | **Validation Set (*n* = 160000)** | |
|---|---|---|---|---|---|---|
| | *n* missing | % | *n* missing | % | *n* missing | % |
| Smoking | 397319 | 93% | 247918 | 93% | 149401 | 93% |
| Sex | 69 | 0.02% | 44 | 0.02% | 25 | 0.02% |
| BMI | 281400 | 66% | 175632 | 66% | 105768 | 66% |
| Age | 259 | 0.061% | 167 | 0.063% | 92 | 0.058% |
| Income | 22403 | 5.27% | 13824 | 5.21% | 8579 | 5.36% |

Multiple imputation was used to account for missing data in sex, BMI, age, and income. Five iterations were used, creating five imputed datasets. Smoking was not considered in analyses because there was not sufficient information available to reliably perform imputation.

Risk factors deemed sufficiently well-recorded in the database and thus included in the models were:

**Table 8: Risk Factors Available Within CPCSSN Database**

| Osteoarthritis | Diabetes | Hypertension |
|---|---|---|
| Osteoporosis<br>Previous leg injury<br>Older age<br>Obesity<br>Female sex | Hypertension<br>Older age<br>Lipid disorders<br>Obesity<br>Male sex<br>Schizophrenia<br>Depression<br>Low socioeconomic status | Older age<br>Diabetes<br>Obesity<br>Kidney disease<br>Tricyclic antidepressant (TCA) use |

## 4.3 Univariate Results

As described in Objective 1, the following results describe the univariate multivariable logistic regression models that were constructed for diabetes, hypertension, and osteoarthritis. First, the estimated $\beta$ coefficients and odds ratios (with 95% confidence intervals) are presented followed by model validation measures such as the ROC curve, AUC, and calibration plot for each model. Note that direct comparisons of the magnitude of the estimated $\beta$ coefficients to determine their relative impact on disease risk would be inappropriate as these parameters were constructed for the purpose of prediction, rather than causal inference. However, the significance of each estimate can be considered. Of greatest importance are the model validation measures, as these provide insight into model performance.

**Table 9: Diabetes Univariate Results**

| | Reference Category/Units | $\beta$ estimate | 95% CI | Odds Ratio | 95% CI |
|---|---|---|---|---|---|
| Hypertension | No | Reference | | Reference | |
| | Yes | 0.30 | 0.26 to 0.35 | 1.35 | 1.30 to 1.42 |
| Age | (Years) | 0.04 | 0.03 to 0.04 | 1.04 | 1.03 to 1.04 |
| Lipid disorders | No | Reference | | Reference | |
| | Yes | 1.69 | 1.64 to 1.73 | 5.42 | 5.16 to 5.87 |
| BMI | (kg/m$^2$) | 0.07 | 0.07 to 0.08 | 1.07 | 1.07 to 1.08 |
| Sex | Male | Reference | | Reference | |
| | Female | -0.30 | -0.34 to -0.26 | 0.74 | 0.71 to 0.77 |
| Schizophrenia | No | Reference | | Reference | |
| | Yes | 0.63 | 0.51 to 0.75 | 1.88 | 1.67 to 2.12 |
| Depression | No | Reference | | Reference | |
| | Yes | 0.14 | 0.08 to 0.20 | 1.15 | 1.08 to 1.22 |
| Income | ($10000) | -0.89 | -1.15 to -0.64 | 0.41 | 0.32 to 0.53 |

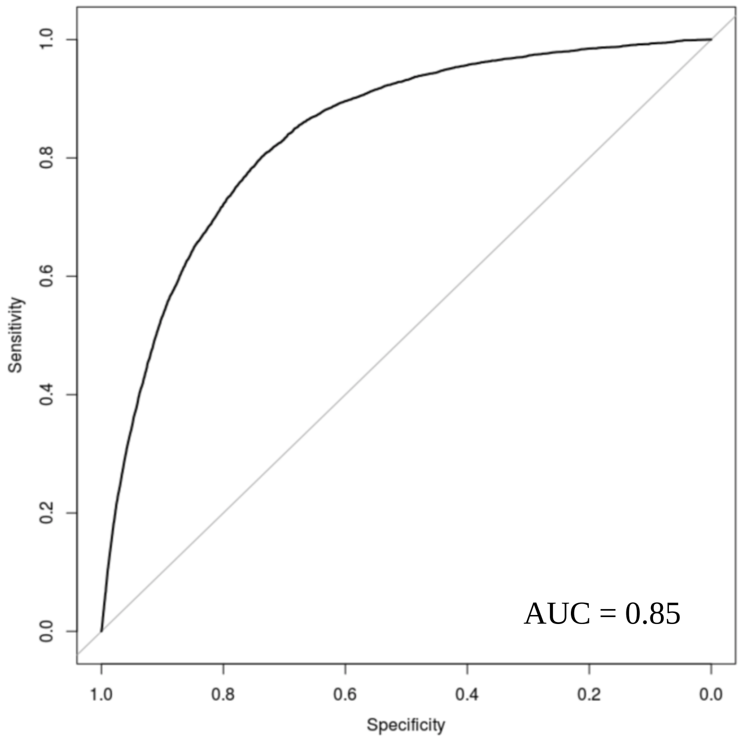

**Figure 4: ROC Curve for Diabetes**

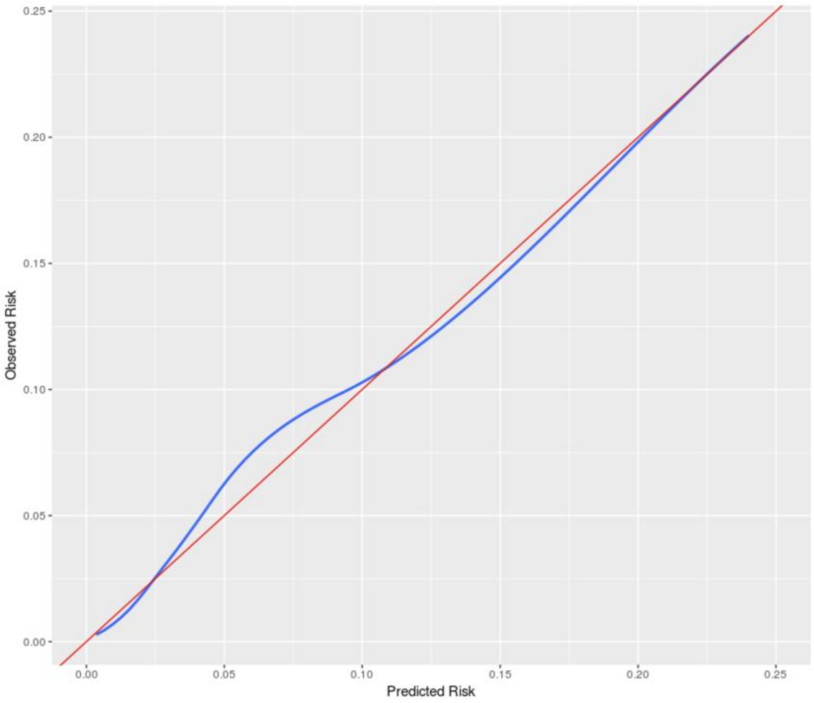

**Figure 5: Calibration Plot for Diabetes**

All estimated $\beta$ coefficients for the diabetes model were found to be significant, as expected given the large sample size. The model discriminated very well, as indicated by its ROC curve and AUC (0.8523; 0.8476 to 0.8570). It slightly overestimated risk in lower risk patients, while it estimated higher risk patients quite well (only a very slight underestimation), as depicted in its calibration plot.

**Table 10: Hypertension Univariate Results**

| | Reference | $\beta$ estimate | 95% CI | Odds Ratio | 95% CI |
|---|---|---|---|---|---|
| Diabetes | No | Reference | | Reference | |
| | Yes | 0.18 | 0.12 to 0.23 | 1.19 | 1.13to 1.26 |
| Age | (Years) | 0.07 | 0.06 to 0.07 | 1.07 | 1.06 to 1.07 |
| BMI | (kg/m$^2$) | 0.06 | 0.06 to 0.07 | 1.06 | 1.06 to 1.07 |
| Chronic Kidney Disease | No | Reference | | Reference | |
| | Yes | 0.80 | 0.74 to 0.85 | 2.22 | 2.09 to 2.35 |
| Tricyclic Antidepressant Use | No | Reference | | Reference | |
| | Yes | 0.55 | 0.49 to 0.62 | 1.74 | 1.63 to 1.86 |

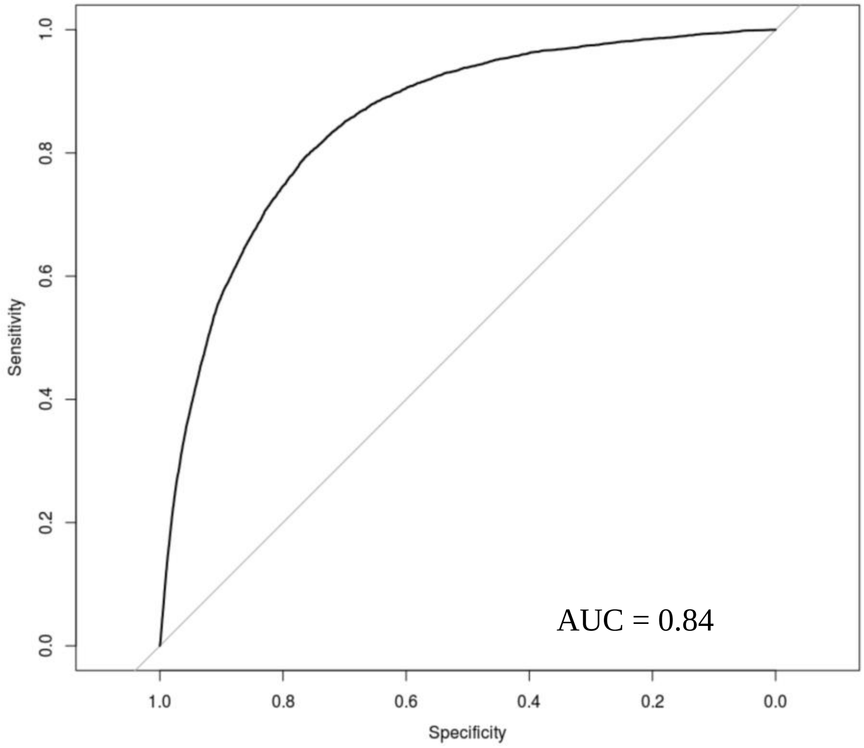

**Figure 6: ROC Curve for Hypertension**

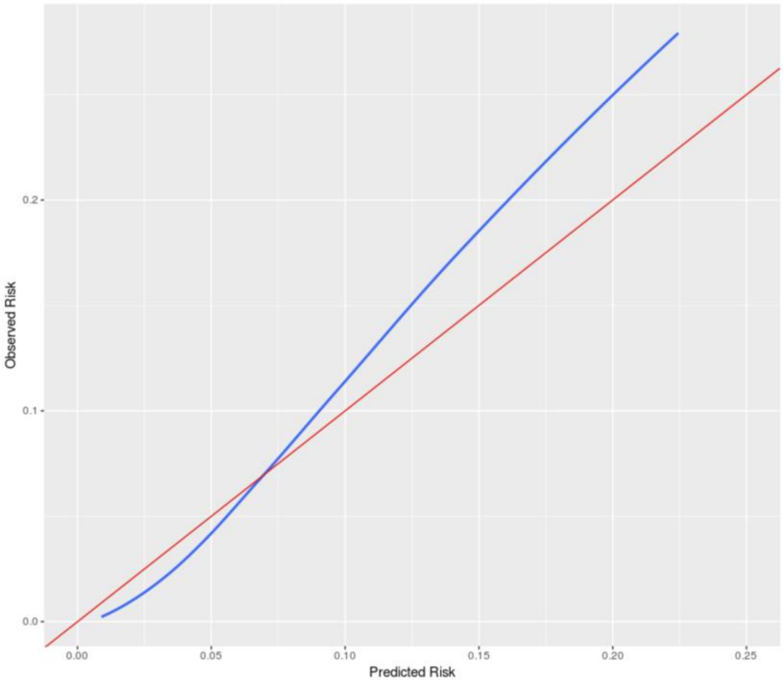

**Figure 7: Calibration Plot for Hypertension**

Again, all estimated $\beta$ coefficients from the hypertension univariate model were found to be significant. The model discrimination was high, as indicated by its ROC curve and AUC (0.8391; 0.8353 to 0.8429). It slightly underestimated risk in lower risk patients, while it overestimated risk in moderate and higher risk patients, as depicted in its calibration plot.

**Table 11: Osteoarthritis Univariate Results**

| | Reference | $\beta$ estimate | 95% CI | Odds Ratio | 95% CI |
|---|---|---|---|---|---|
| Age | (Years) | 0.06 | 0.05 to 0.06 | 1.06 | 1.05 to 1.06 |
| Sex | Male | Reference | | Reference | |
| | Female | 0.22 | 0.17 to 0.27 | 1.25 | 1.19 to 1.31 |
| BMI | (kg/m$^2$) | 0.04 | 0.03 to 0.04 | 1.04 | 1.04 to 1.05 |
| Previous Leg Injury | No | Reference | | Reference | |
| | Yes | 1.60 | 1.52 to 1.68 | 4.94 | 4.57 to 5.35 |
| Osteoporosis | No | Reference | | Reference | |
| | Yes | 0.90 | 0.83 to 0.98 | 2.47 | 2.29 to 2.66 |

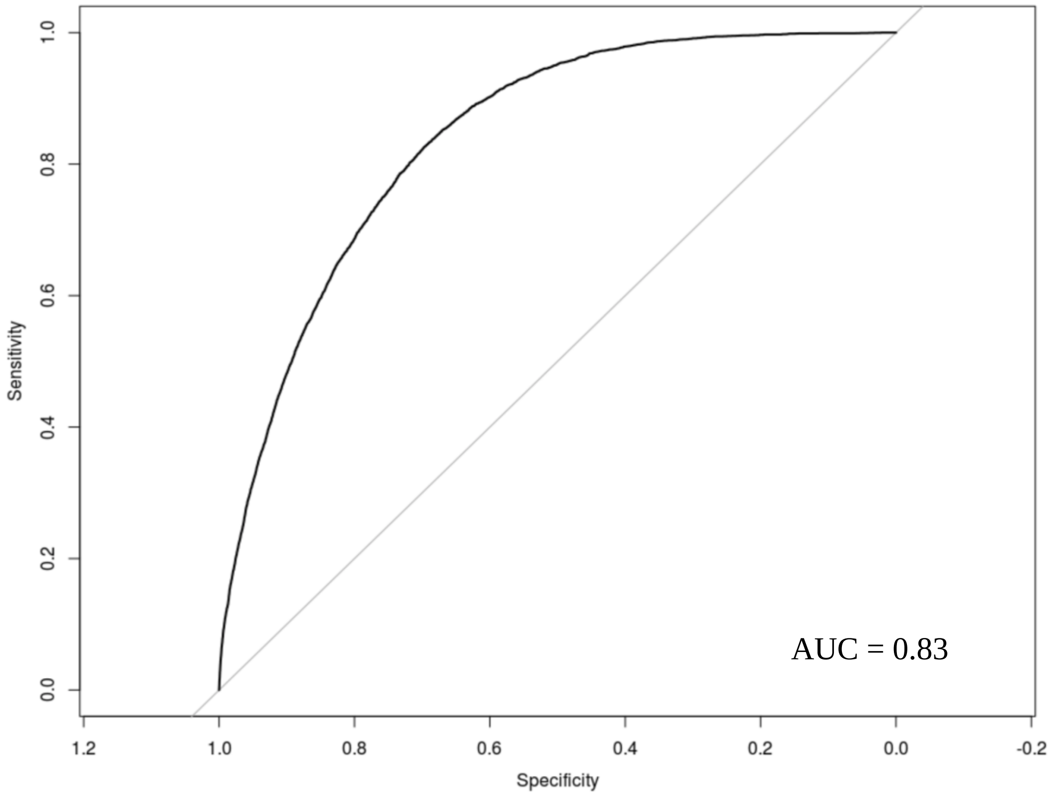

**Figure 8: ROC Curve for Osteoarthritis**

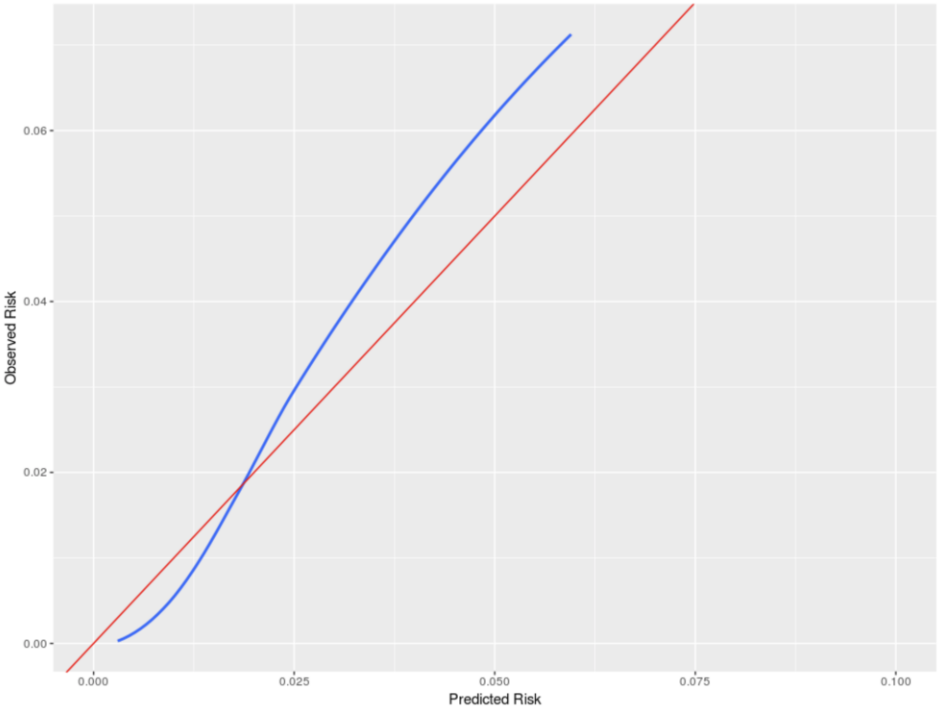

**Figure 9: Calibration Plot for Osteoarthritis**

Similar to the previous models, the estimated $\beta$ coefficients for the osteoarthritis model were found to be significant. Model discrimination was high, as indicated by its ROC curve and AUC (0.8394; 0.8342 to 0.8446). It slightly overestimated risk in most patients, with the risk of those at moderate risk overestimated the most and slight underestimation at both extremes.

## 4.4 Dependence Analysis

To measure the unadjusted correlation between outcomes, the $\phi$ coefficient was computed for each outcome pair using the *cor.test* function in R. All pairs showed positive correlation. As shown in Table 12, diabetes and hypertension were the most correlated outcomes, followed closely by hypertension and osteoarthritis. Diabetes and osteoarthritis were the least correlated.

**Table 12: ϕ Coefficients**

| | Diabetes | Hypertension | Osteoarthritis |
|---|---|---|---|
| **Diabetes** | 1 | | |
| **Hypertension** | 0.2420<br>(0.2380 to 0.2460,<br>p < 0.0001) | 1 | |
| **Osteoarthritis** | 0.0975<br>(0.0934 to 0.1016,<br>p < 0.0001) | 0.2086<br>(0.2045 to 0.2127,<br>p < 0.0001) | 1 |

Partial correlation was also computed for each outcome pair using the *pcor* function, which adjusted for the effects of all risk factors for both outcomes. For example, the partial correlation between diabetes and hypertension was adjusted for all risk factors associated with diabetes and/or hypertension. Again, all pairs were positively correlated, though the magnitudes of correlation were reduced. As seen in Table 12, partial correlation was highest between diabetes and hypertension; then hypertension and osteoarthritis; and diabetes and osteoarthritis.

**Table 13: Partial Correlation**

| | Diabetes | Hypertension | Osteoarthritis |
|---|---|---|---|
| **Diabetes** | 1 | | |
| **Hypertension** | 0.1323<br>(0.1281 to 0.1366,<br>p < 0.0001) | 1 | |
| **Osteoarthritis** | 0.0377<br>(0.0336 to 0.0419,<br>p < 0.0001) | 0.1227<br>(0.1183 to 0.1270,<br>p < 0.0001) | 1 |

For each outcome pair, a copula was constructed to describe the non-linear dependence between outcomes while adjusting for covariates using the univariate multivariable logistic regression models. The Frank copula (170)  was selected for use, given its ability to model weak dependence well. The Frank copula can be seen below:

$$C_\theta(u, v) = -\frac{1}{\theta}\ln\left(1 + \frac{(e^{-\theta u} - 1)(e^{-\theta v} - 1)}{e^{-\theta} - 1}\right)$$

Following the construction of univariate models for each outcome, estimates of the copula parameter $\theta$ were obtained for each disease pair. Results from the estimation of $\theta$ for each outcome pair are displayed in Table 13. A positive $\theta$ estimate represents a positive dependence (i.e., diseases tend to either both occur or not occur), while a negative $\theta$ estimate represents a negative dependence (i.e., patients tend to develop one disease or the other but not both). The strength of the dependence can be inferred from the magnitude of the $\theta$ estimate.

**Table 14: θ Estimates**

| | Diabetes | Hypertension | Osteoarthritis |
|---|---|---|---|
| **Diabetes** | | | |
| **Hypertension** | 1.6766<br>(1.5657 to 1.7876,<br>p < 0.0001) | | |
| **Osteoarthritis** | 0.6830<br>(0.5256 to 0.8405<br>p < 0.0001) | 1.9490<br>(1.8224 to 2.0755,<br>p < 0.0001) | |

All disease pairs exhibited a significant positive dependence after adjusting for risk factors, as demonstrated by their $\theta$ estimates greater than zero.

Based on these copula models, trivariate probabilities were estimated. The following are simulated patients whose trivariate probabilities have been estimated. For comparison, trivariate probabilities under the assumption of independence have been estimated. The ratio between these is presented for comparison purposes. Ratios greater than one indicate a higher risk based on the copula than when assuming independence.

*Example patient 1:* Fifty-nine-year-old male whose BMI is 29 kg/m$^2$, who has osteoporosis and an income of roughly $40000.

**Table 15: Trivariate Probabilities for Example Patient 1**

| P(Diabetes, Hypertension, Osteoarthritis) | Based on copula model | Based on independence assumption | Ratio |
|---|---|---|---|
| P(0,0,0) | 0.8221 | 0.8132 | 1.01 |
| P(0,0,1) | 0.0466 | 0.0529 | 0.88 |
| P(0,1,0) | 0.0907 | 0.0991 | 0.92 |
| P(1,0,0) | 0.0121 | 0.0064 | 1.88 |
| P(0,1,1) | 0.0212 | 0.0238 | 0.89 |
| P(1,0,1) | 0.0015 | 0.0015 | 0.97 |
| P(1,1,0) | 0.0049 | 0.0029 | 1.69 |
| P(1,1,1) | 0.0008 | 0.0002 | 4.25 |

*Example patient 2:* Seventy-nine-year-old woman whose BMI is 34 kg/m$^2$ with an income of roughly $35000 and free of any other risk factors.

**Table 16: Trivariate Probabilities for Example Patient 2**

| **P(Diabetes, Hypertension, Osteoarthritis)** | **Based on copula model** | **Based on independence assumption** | **Ratio** |
|---|---|---|---|
| P(0,0,0) | 0.6088 | 0.5798 | 1.05 |
| P(0,0,1) | 0.0481 | 0.0665 | 0.72 |
| P(0,1,0) | 0.2362 | 0.2633 | 0.90 |
| P(1,0,0) | 0.0466 | 0.0302 | 1.54 |
| P(0,1,1) | 0.0282 | 0.0371 | 0.76 |
| P(1,0,1) | 0.0026 | 0.0043 | 0.61 |
| P(1,1,0) | 0.0239 | 0.0169 | 1.42 |
| P(1,1,1) | 0.0055 | 0.0019 | 2.84 |

Chapter 5

# 5 Discussion

This chapter describes the key findings from the development of a PPM for multiple diseases, with further discussion and elaboration. The strengths, limitations, and implications of this work are discussed as well.

## 5.1 Overview of Results

Univariate models were constructed for diabetes, hypertension, and osteoarthritis that can be used to estimate a patient's risk of future disease development. Each model was comprised of a set of $\beta$ estimates that describe the contribution of each risk factor. All models had good predictive ability, as demonstrated by their AUCs and calibration plots. Diabetes was the best predicted outcome, with the greatest AUC (0.85) and the best calibration plot. The hypertension model had the next best performance, with an AUC of 0.84 and a good calibration plot. The osteoarthritis model had the lowest performance of the predicted diseases, with an AUC of 0.83 and a calibration plot that slightly underestimated risk in low-risk patients and slightly overestimated risk in high-risk patients.

Following the construction of univariate models for each disease, an analysis of dependence between each disease was conducted. This began with an analysis of the unadjusted correlation measured using the $\phi$ coefficient. Diabetes and hypertension were the most correlated ($\phi$ = 0.24), followed by hypertension and osteoarthritis ($\phi$ = 0.21), then diabetes and hypertension ($\phi$ = 0.10).

Next, the correlation between diseases after adjusting for the effects of relevant risk factors (partial correlation) was determined. An examination of the partial correlation between each disease pair revealed lower correlation coefficients between outcomes. This was expected, as some dependence was anticipated to be explained by risk factors. Interestingly, the correlation between hypertension and osteoarthritis ($\phi$ = 0.12) became roughly the same as that of diabetes and hypertension ($\phi$ = 0.13) after adjustment. The correlation between diabetes and osteoarthritis decreased almost to zero after adjustment ($\phi$ = 0.04), indicating that most of the observed association between these two diseases could be explained by their risk factors.

Finally, construction of copula models produced $\theta$ coefficients that describe the dependence between outcomes that existed after adjusting for relevant risk factors. The largest $\theta$ estimate was observed between hypertension and osteoarthritis ($\theta$ = 1.95), indicating that the strongest dependence exists between this pair after adjusting for all risk factors. Diabetes and hypertension had the next largest $\theta$ estimate (1.68), followed by diabetes and osteoarthritis ($\theta$ = 0.68). As observed in the trivariate probability charts, the probability of developing multiple diseases was greater when based on the copula models than when assuming independence. For example, for *example patient 1*, the probability of developing all three diseases within five years was roughly four times greater under the copula model than when assuming independence (0.0008 vs 0.0002). The smallest increase was observed in the estimated probability of developing both diabetes and osteoarthritis, which aligns with the previous correlation analyses that found the least correlation between these two diseases. These findings indicate that risk estimates made under the assumption of independence underestimate the risk of disease co-occurrence.

## 5.2 Comparisons of Univariate Models with Existing Models

Several models have been constructed in other works to individually estimate risk of each of diabetes, hypertension, and osteoarthritis. In the following, the models produced by this thesis will be compared with these existing univariate models.

### 5.2.1 Diabetes

Several models are commonly used to estimate an individual's risk of diabetes development. These include the American Diabetes Association Questionnaire (ADA) (223), hosted on the American Diabetes Association website; the Canadian Diabetes Risk Questionnaire (CANRISK) (224), hosted on the government of Canada website; the Leicester Risk Assessment (LRA) (225), found on the Diabetes UK and the UK National Health Service websites; and Australian Type 2 Diabetes Risk Assessment Tool (AUSDRISK) (226), found on The Australian Department of Health website. Compared to these tools, the model derived from the CPCSSN database includes many similar risk factors, with the notable addition of several diseases as risk factors, such as depression or schizophrenia. However, the model derived for this thesis did not include several lifestyle or environmental risk factors as these were not stored in the database. The CPCSSN database relied on EMR data, which is limited by the nature of the data collected (only clinically relevant data) when compared to data collected for purpose through questionnaires or physical examinations. The model derived for this thesis was derived from a considerably larger cohort than previous models. It performed with similar, if not superior, discrimination compared to traditional models. A comparison of the model derived for this thesis and existing models for estimating diabetes risk is displayed below.

**Table 17: Comparison of Diabetes Univariate Model with Existing Models**

| **Name of tool/study** | **Source population** | **Sample size (development set)** | **Data collection method** | **Method of analysis** | **Validity** |
|---|---|---|---|---|---|
| CPCSSN (2018) | CPCSSN primary care records | 265228 | Electronic medical records | Multivariable logistic regression | Internal validation: AUC of 0.85 |
| ADA (2009) **(223)** | NHANES (National Health and Nutrition Examination Survey) 1999-2004 | 5258 | Interviews, physical examinations, and laboratory tests | Multivariable logistic regression | Internal validation: AUC of 0.83 |
| eCANRISK (2009) **(224)** | CANRISK study | 4366 | Questionnaire | Multivariable logistic regression | Internal validation: AUC of 0.75 |
| LRA (2010) **(225)** | Random sample of UK | 6390 | Interviews, physical examinations, and laboratory tests | Multivariable logistic regression | Internal validation: AUC of 0.69<br>External validation: AUC of 0.72 |
| AUSDRISK (2010) **(226)** | Australian Diabetes, Obesity, and Lifestyle Study | 6060 | Interviews and laboratory tests | Multivariable logistic regression | Internal validation: AUC of 0.783<br>External validation: AUC of 0.66 |

**Table 18: Risk Factors Included in Diabetes Risk Estimation**

| **Risk factor** | **CPCSSN** | **ADA** | **CANRISK** | **LRA** | **AUSDRISK** |
|---|---|---|---|---|---|

| Risk factor | CPCSSN | ADA | CANRISK | LRA | AUSDRISK |
|---|---|---|---|---|---|
| Age | X | X | X | X | X |
| Sex | X | X | X | X | X |
| Diabetes in family |  | X | X | X | X |
| High blood pressure | X | X | X | X |  |
| Blood pressure medication use/history |  |  | X | X |  |
| Lipid disorder | X |  |  |  |  |
| Schizophrenia | X |  |  |  |  |
| Depression | X |  |  |  |  |
| Physical activity |  | X | X |  | X |
| Obesity (BMI) | X | X | X | X | X |
| Gestational diabetes |  | X | X |  |  |
| Waist measurement |  |  | X | X |  |
| Eats vegetables and fruits |  |  | X |  | X |
| High blood glucose history |  |  | X |  | X |
| Ethnic group/country of birth |  |  | X | X | X |
| Level of education |  |  | X |  |  |
| Income | X |  |  |  |  |
| Smoking |  |  |  |  | X |

### 5.2.2 Hypertension

When dealing hypertension and other cardiovascular diseases, focus is often placed on predicting severe events such as heart attack or stroke (9) rather than hypertension. However, some prognostic predictive models aimed at the estimation of hypertension risk exist; these include models based on the Framingham Heart Study (227); Women's Health Study (228); and data combined from the Atherosclerosis Risk in Communities and the Cardiovascular Health Study (229). These models included a wide variety of risk factors. The model derived for this thesis included the fewest risk factors, as it was not able to include several lifestyle or environmental risk factors. Despite this, it performed the best out of all models, as it had the greatest AUC. Compared to the other models considered, the model derived for this thesis had considerably more individuals in its development set. Again, a comparison of the model developed for this thesis and existing models used to estimate risk of hypertension are presented below. The following tables present a comparison of each of these models.

**Table 19: Comparison of Hypertension Univariate Model with Existing Models**

| **Name of tool/study** | **Source population** | **Sample size (development set)** | **Data collection method** | **Method of analysis** | **Validity** |
|---|---|---|---|---|---|
| CPCSSN | CPCSSN primary care records | 265228 | Electronic medical records | Multivariable logistic regression | Internal validation: AUC of 0.84 |
| Framingham Heart Study (2008) **(227)** | Population based | 1717 | Interviews, physical examinations, and laboratory tests | Multivariable Weibull regression | Internal validation: AUC of 0.79 |
| Women's Health Study (2009) **(228)** | US female health professionals | 9427 | Interviews, physical examinations, and laboratory tests | Multivariable logistic regression | Internal validation: AUC of 0.71 |
| ARIC/CHS (2010) **(229)** | Population based | 7683 | Interviews, physical examinations, and laboratory tests | Multivariable logistic regression | Internal validation: AUC of 0.75 |
| ARIC: Atherosclerosis Risk in Communities Study<br>CHS: Cardiovascular Health Study | | | | | |

**Table 20: Risk Factors Included in Hypertension Risk Estimation**

| **Risk factor** | **CPCSSN** | **Framingham Heart Study** | **Women's Health Study** | **ARIC-CHS** |
|---|---|---|---|---|
| Diabetes | X | | | X |
| Age | X | X | X | X |
| Sex | | X | X | X |
| BMI | X | X | | X |
| Chronic kidney Disease | X | | | |
| Tricyclic antidepressant | X | | | |

| Risk factor | CPCSSN | Framingham Heart Study | Women's Health Study | ARIC-CHS |
|---|---|---|---|---|
| use | | | | |
| Systolic blood pressure | | X | X | X |
| Diastolic blood pressure | | X | X | X |
| Family history of hypertension | | X | | X |
| Ethnicity | | | X | |
| Total/HDL cholesterol | | | X | |
| Lipoprotein | | | X | |
| High-sensitivity C-reactive protein | | | X | |
| Total grains | | | X | |
| Current smoker | | | | X |
| Lack of exercise | | | | X |

### 5.2.3 Osteoarthritis

Several models for the estimation of osteoarthritis risk have been developed, including the Tool for Osteoarthritis Risk Prediction (TOARP) (230); the Nottingham knee osteoarthritis risk prediction models (231); and models derived from data from the Rotterdam Study-1 (232) and the Multicenter Osteoarthritis Study (MOST) (233). While the model derived for this thesis is not specific to the location of osteoarthritis (as the case definition for osteoarthritis did not specify the affected joint), all other models were designed to predict exclusively knee osteoarthritis. Similar to hypertension, osteoarthritis predictive models made use of a wide variety of risk factors, including radiographic measures such as the Kellgren and Lawrence score. The model developed for this thesis did not use any radiographic measures; in fact, it was the simplest model while maintaining the best discrimination according to internal validation. The model developed for this work was based on a considerably larger sample of patients aged 18 and older; it did not restrict its sample to an older population at high risk of osteoarthritis in order to enable the estimation of risk among all adults. A thorough comparison of each of these models is displayed below.

**Table 21: Comparison of Osteoarthritis Univariate Model with Existing Models**

| Name of | Source | Sample size | Data | Method of | Validity |
|---|---|---|---|---|---|

| **tool/study** | **population** | **(development set)** | **collection method** | **analysis** | |
|---|---|---|---|---|---|
| CPCSSN | CPCSSN primary care records | 265228 | Electronic medical records | Multivariable logistic regression | Internal validation: AUC of 0.83 |
| Tool for Osteoarthritis Risk Prediction (TOARP) (2018) **(230)** | Population based cohort (age 45-79) | 641 | Interviews, physical examinations, and laboratory tests, including MRI | Multivariable logistic regression | Internal validation: AUC of 0.72 |
| Rotterdam Study-1 (2014) **(232)** | Population based cohort (age 55+) | 2628 | Interviews, physical examinations, and laboratory tests, including x-ray | Multivariable logistic regression | Internal validation: AUC of 0.79 |
| Multicenter Osteoarthritis Study (MOST) (2016) **(233)** | Population based cohort (age 50-79) | 3026 | Interviews, physical examinations, and laboratory tests, including x-ray | Multivariable logistic regression | Internal validation: AUC of 0.78 External validation: AUC of 0.76 |
| Nottingham knee osteoarthritis risk prediction models (2011) **(231)** | Population based cohort (age 40+) | 424 | Interviews, physical examinations, and laboratory tests, including x-ray | Multivariable logistic regression | Internal validation: AUC of 0.70 External validation: AUC of 0.6 and 0.79 |

**Table 22: Risk Factors Included in Osteoarthritis Risk Estimation**

| **Risk factor** | **CPCSSN** | **TOARP** | **Rotterdam Study-1** | **MOST** | **Nottingham** |
|---|---|---|---|---|---|
| Age | X | X | X | | X |
| Sex | X | X | X | | X |
| BMI | X | X | X | X | X |
| Previous leg injury | X | X | | | X |
| Osteoporosis | X | | | | |
| KL grade | | X | X | X | |

| Risk factor | CPCSSN | TOARP | Rotterdam Study-1 | MOST | Nottingham |
|---|---|---|---|---|---|
| Joint damage |  | X |  |  |  |
| T2 cartilage relaxation time |  | X |  |  |  |
| Genetic risk |  |  | X |  |  |
| Knee pain |  |  | X | X |  |
| Education level |  |  | X |  |  |
| Smoking |  |  | X |  |  |
| Contralateral/multiple joint osteoarthritis |  |  |  | X |  |
| Average WOMAC score |  |  |  | X |  |
| Depression |  |  |  | X |  |
| Knee misalignment |  |  |  | X |  |
| Occupation |  |  |  |  | X |
| Family history |  |  |  |  | X |
| WOMAC: Western Ontario and McMaster Universities arthritis index<br>KL: Kellgren and Lawrence |  |  |  |  |  |

## 5.3 Comparison of Dependence Analysis with Existing Dependence Analyses

While no studies have examined the dependence between diabetes, hypertension, and osteoarthritis together, several studies have looked at each pair of diseases. The findings of this thesis will be compared to the finding of other studies below.

### 5.3.1 Diabetes and Hypertension

Epidemiological and pathophysiological evidence supports an association between diabetes and hypertension beyond what would be expected due to shared risk factors (234). Evidence from epidemiologic studies found an association between blood pressure and blood glucose; this has been observed in both children (235) (where the effect of risk factors such as drugs and alcohol are minimal) and adults (236). Higher blood glucose levels have been associated with an increased risk of developing hypertension in the future. After an 18-year follow-up, a long-term Finnish study of men without hypertension found higher rates of hypertension development among those who had higher blood glucose concentrations at the outset of the study, even after adjusting for age, adiposity, alcohol consumption, and baseline blood pressure (237). Similarly, increased blood pressure is associated with an increased risk of diabetes. A study of 10000 men in Israel found systolic blood pressure to be significantly associated with the development of type 2 diabetes after five years (238). Several pathophysiological mechanisms have been proposed to explain the association between these two diseases; however, none of these definitively explain the relationship (234).

These findings align with those of the current analysis, which found an association between diabetes and hypertension that persisted after adjusting for relevant risk factors. Higher probabilities of diabetes and hypertension co-occurrence were estimated using the copula model compared to those estimated by assuming independence.

### 5.3.2 Diabetes and Osteoarthritis

A recent systematic review and meta-analysis found diabetes and osteoarthritis to be associated (239). Examination of osteoarthritis risk among 32,137 people revealed an odds ratio of 1.46 (95% confidence interval: 1.08 to 1.96) comparing people with diabetes to those without. Several studies retained a significant association after adjusting for obesity (240–242), a considerable risk factor for both diseases. A similar association was found for the risk of diabetes among people with hypertension. An odds ratio of 1.41 (95% confidence interval:1.21 to 1.65) was observed for diabetes development, comparing those with osteoarthritis to those without across a group of 1,040,175 people. Interestingly, the association between diabetes and osteoarthritis was significant in studies including hand osteoarthritis only (243,244), which highlights the metabolic and systemic nature of hand osteoarthritis. Similarly, several studies have observed the impact of metabolic syndrome (which includes diabetes) on the risk of osteoarthritis. The Japanese Research on Osteoarthritis/Osteoporosis Against Disability (ROAD) study found that the development of diseases considered components of metabolic syndrome was associated with an increased risk of knee osteoarthritis development and progression (245). In fact, the co-occurrence of obesity, diabetes, and hypertension (all of which are components of metabolic disorder) was found to increase the odds of experiencing hand osteoarthritis by a factor of 2.3 (95% confidence interval 1.3 to 3.9) (246).

In the current analysis, some dependence was observed between diabetes and osteoarthritis ($\phi$ = 0.10). However, much of this was likely due to the effect of risk factors, as the observed dependence decreased after adjusting for relevant factors ($\phi$ = 0.04). Accordingly, diabetes and osteoarthritis had the lowest $\theta$ estimate of the disease pairs that were examined, corresponding to the least dependence. As suggested by the literature, an association may exist between diabetes and hand osteoarthritis, specifically; however, diagnoses of osteoarthritis within CPCSSN were not specific to the joint(s) affected, thus sub-analyses could not be performed.

### 5.3.3 Hypertension and Osteoarthritis

Research investigating the relationship between hypertension and osteoarthritis found an association between the two diseases (247–250). A research group in Korea studied hypertension and its impact on osteoarthritis and found that while hypertension was not significantly associated with osteoarthritis generally (251), it was significantly associated with an increased risk of knee osteoarthritis (OR: 1.26, 95% confidence interval: 1.08 to 1.48) (252). Hypertension is a component of metabolic disease, which has been linked to the development of osteoarthritis (245,246), similar to diabetes. One theory hypothesizes that subchondral ischemia (inadequate blood flow to bone tissues) due to the vessel-narrowing effects of hypertension results in degradation of the joint cartilage, resulting in osteoarthritis (253–255).

Results of this thesis revealed an association between hypertension and osteoarthritis as well. Correlation assessed via the $\phi$ coefficient revealed a relationship that persisted after adjustment for relevant factors. When using the copula to estimate the trivariate probabilities, the estimated probability of the co-occurrence of hypertension and osteoarthritis was greater than the estimated probability assuming independence.

This thesis clearly demonstrated that when making estimations about the risk of multiple diseases, it is inappropriate to assume that each disease is independence of the other. Instead, models must be used that are able to capture the dependence that exists between diseases and express this when estimating risk.

### 5.3.4 Multiple Disease Risk Estimation

There has been one PPM developed for multiple diseases. Wang et al. (256) developed a model for both COPD and congestive heart failure (CHF) using the EMR data of roughly 8000 patients. Risk factors used as predictors included musculoskeletal disorders, heart arrhythmias, diabetes, tobacco use, and asthma. Rather than considering predictors individually, predictors were grouped by selecting those that best predicted the outcome, resulting in three groups: predictors for COPD and CHF, predictors of only COPD, and predictors of only CHF. The main objective of this study was to identify a set of shared predictors in addition to the development of a model that accurately predicts the development of each disease. Predictors such as osteoarthritis, back disorders, and cardiac dysrhythmias were shared by CHF and COPD; predictors such as diabetes mellitus, chronic ischemic heart disease, and acute ischemic heart disease were mainly associated with CHF; and predictors such as asthma, kidney stones, and tobacco use disorder were mainly associated with COPD. The resulting predictive model performed well, with an AUC of 0.72. Similar to the model developed for this thesis, Wang et al. used EMR data to derive their prognostic predictive model.

## 5.4 Limitations

This research has several limitations that should be considered.

The current analysis was limited by the availability of information within the EMR. Information describing key risk factors was unavailable, such as behavioural or environmental factors, as this information is not typically collected during a clinical encounter. For the univariate models, this likely resulted in an underestimation in the risk of patients who possess the uncollected risk factor. For the dependence analysis, this potentially resulted in some of the observed dependence being due to a risk factor that was not collected in the EMR. As the risk factor was not collected, it could not be adjusted for. Such a factor could act in either direction; a risk factor could increase or decrease the dependence between the diseases, thus the true dependence could be less than or greater than what was observed. However, the use of information available within the EMR has several advantages. The primary care setting is considered an ideal site to deploy models to estimate patients' risk of chronic disease as patients are commonly seen by a PCP prior to disease development. EMR data are readily available to base predictions on in primary care; that is, no additional information needs to be collected in order to support the use of a predictive model in clinical practice. The model would operate in the background of an EMR, assessing the risk of disease among patients and flagging those at increased risk. Additionally, analyses based on EMR data are not limited by poor statistical power due to small sample sizes. EMR databases often collect the records of thousands of patients, providing sufficient power to make strong conclusions.

Caution must be taken when applying the results to other settings, as the data used are likely not representative of the general population. However, as previously mentioned, primary care is an excellent setting where predictive models can be used to identify high-risk patients by estimating risk of chronic disease. Thus, deriving predictive models in the same setting that they will be used is ideal.

There are several errors that may occur that would result in a diagnosis not being recorded by the PCP, resulting in missing data. First, the PCP must correctly identify and diagnose the disease. It is possible that a disease may go undetected or undiagnosed and would not be recorded in the EMR. Second, the PCP must record the diagnosis in the EMR; diagnoses of certain diseases may carry stigma, limiting the PCP's willingness to record the diagnosis in the EMR. For example, a diagnosis of schizophrenia sometimes carries stigma; a PCP may want to be completely certain of their diagnosis before recording it in the EMR and may not record the diagnosis otherwise. Third, in CPCSSN, the diagnosis found in the Billing table corresponds to the diagnosis that is most responsible for the visit, or the most responsible diagnosis. The PCP must be sufficiently motivated to record any additional diagnoses in the Health Condition table. However, a thorough chart review was used to validate the CPCSSN disease-case algorithms that resulted in high sensitivity and specificity for these algorithms. Other conditions relied upon case definitions created for the purpose of this thesis; these case definitions have not been validated. However, all efforts to make these definitions as accurate as possible have been performed, including a review of relevant literature; a comprehensive examination of the database; and review by an expert EMR user who was a member of the research team (PCP).

There may have been some bias introduced through patterns in physician diagnosis of diabetes, hypertension, and osteoarthritis. For example, should a physician diagnose a patient with diabetes, it is likely that they will assess for related conditions, such as hypertension. In some cases, a diagnosis of hypertension would have gone undetected had the physician not diagnosed the patient with diabetes. This may have led to some dependence between these diseases being due to patterns in diagnosis.

No external validation was performed for the univariate multivariable models. This would have required access to an external data source from a similar yet distinct population. Access to such a database was unavailable. Accordingly, the univariate models can only be confidently applied to the data from which they were derived. However, the univariate models were intended to be specific to the Canadian primary care population and can be confidently applied to this setting.

The nested nature of the CPCSSN database results in clustered data, in which patients within the same group are likely more similar than those in different groups; for example, patients who receive care from the same PCP are more likely to be similar than those who receive care from a different PCP. This typically requires a methodology capable of accounting for the clustered nature of the data; however, linkages between patients and PCPs were unavailable, thus clustered analyses were not performed.

The CPCSSN case definition used to identify patients with diabetes does not separate patients by type of diabetes. As such, all diagnoses of diabetes were treated as type 2 diabetes. However, these are different diseases with distinct etiologies, each with unique risk factors. Many external factors contribute to risk of type 2 diabetes, such as obesity, diet, and smoking (257), whereas type 1 diabetes has been linked to more genetic factors (258). This likely resulted in misclassification bias. However, the amount of bias introduced was likely minimal, as type 1 diabetes makes up only 10% of all cases of diabetes, based on national statistics (259). Additionally, type 1 diabetes is usually diagnosed in childhood. Given that only incident cases of diabetes in adults (18 or above) are being considered, these cases are more likely to be type 2 where adult onset is more common. Similarly, diagnoses of osteoarthritis did not specify which joint was affected. Thus, osteoarthritis included any diagnosis of osteoarthritis, irrespective of location.

## 5.5 Implications

Although constructed for the purpose of developing a combined prognostic predictive model for diabetes, hypertension, and osteoarthritis, the univariate models developed for this thesis could be used to independently estimate a patient's risk of each disease. External validation should be performed prior to deployment; however, each model's strong internal validation in a Canadian primary care population indicates that these models would perform well in a primary care setting in Canada. For example, the univariate model for diabetes development could be used by PCPs to estimate a patient’s risk of developing diabetes in the next 5 years. This model could either operate in the background, flagging high-risk patients, or as requested by the PCP where they desire a risk estimate. The PCP can then suggest interventions aimed at reducing the patient’s risk of developing diabetes.

It is widely known that chronic diseases tend to co-occur or cluster within individuals. As chronic diseases often have similar risk factors, it is sometimes assumed that this clustering is due to their shared risk factors. However, this thesis found that chronic diseases tended to co-occur more frequently than can be explained by their risk factors. This could be a result of many factors such as patient susceptibility or shared disease processes. Irrespective of the mechanism resulting in this dependence, a thorough understanding of the dependence between diseases is necessary to enable the construction of a prognostic predictive model for the development of multiple chronic diseases. This work examined the dependence between diseases using a variety of techniques including correlation, partial correlation, and copula modelling. Based on these methods, this thesis confirms the findings of previous works that have also demonstrated dependence between diabetes, hypertension, and osteoarthritis (234,239,247–250). However, this thesis is the first to do so using a method that accounts for the non-Gaussian distribution of diseases while simultaneously adjusting for relevant risk factors. Based on this dependence analysis, trivariate probabilities can be estimated to inform patients and their PCPs about their risk of diabetes, hypertension, and osteoarthritis, including the co-occurrence of these diseases.

The availability of a prognostic predictive model capable of estimating a patient's risk of multiple diseases could impact a physician's clinical care in many ways. First, this tool may reveal a risk of the development of multiple diseases that is greater than what would be expected when estimating disease risk independently; this elevated risk due to the dependence between diseases would likely have gone undetected otherwise. Informed of this risk, physicians can suggest preventative interventions accordingly. For example, a patient's risk of developing diabetes and hypertension within the same 5-year window could be estimated by multiplying their risk of diabetes by their risk of hypertension. This method assumes that these outcomes are independent; however, this assumption is invalid for diabetes and hypertension. Instead, the copula model would produce a greater risk of developing these two diseases. The difference between these risks could be the difference between the PCP making a recommendation for preventative action or not. The most useful and effective way to convey this information must be the subject of future work.

## 5.6 Future Directions

The completion of this work enables the construction of a prognostic predictive model for diabetes, hypertension, and osteoarthritis. The current model assumes dependence between diseases does not vary between individuals, as $\theta$ is fixed for all patients after adjusting for risk factors. Further research is needed to allow $\theta$ to vary depending on the values of a patient's risk factors.

The model developed in this thesis would present risks that are adjusted for the dependence between diseases. Future work must investigate how best to present these risks in a way that is meaningful to both patients' and their PCPs. This will require specific research into how people interpret information about joint risk, as this is harder to interpret than a single disease risk. For example, does knowledge of increased risk of both diabetes development and hypertension development have a different effect compared to knowledge of increased risk of diabetes on its own.

The model can be operationalized into a tool capable of running in the background of an EMR to flag high-risk patients and/or deliver risk estimates for patients when called upon by the PCP. Future research should assess the effectiveness of this tool, ideally through a randomized controlled trial in which PCPs are randomly assigned to receive the tool for use in their clinical practice. This trial would assess outcomes such as whether PCPs make different decisions when given information about a patient's risk; whether patients are more likely to adopt a preventative change when this recommendation is supported by a risk estimate; whether patient risk is reduced after receiving a risk estimate; and whether patient outcomes are ultimately changed by receiving risk estimates. In the primary prevention of multimorbidity, this thesis takes a first step in developing a tool capable of delivering risk estimates to inform PCP decision-making.

## 5.7 Conclusion

Through the construction of univariate models for diabetes, hypertension, and osteoarthritis, and an examination of the dependence between each of these diseases, this thesis developed a prognostic predictive model for the occurrence, including the co-occurrence, of these diseases. Univariate models were able to accurately estimate patient risk, as demonstrated by their discrimination and calibration. A dependence analysis using copulas to capture the non-Gaussian distribution of each disease revealed the correlations between each disease pair. This dependence analysis enabled the estimation of the risk of developing any combination of the diseases considered. The development and implementation of this model in clinical practice will enable accurate risk estimation to inform interventions aimed at risk reduction.

# References

1. Hendriksen JMT, Geersing GJ, Moons KGM, de Groot JAH. Diagnostic and prognostic prediction models. J Thromb Haemost. 2013 Jun;11 Suppl 1:129–41.

2. Steyerberg EW. Clinical Prediction Models. New York, NY: Springer New York; 2009. (Statistics for Biology and Health).

3. Lee YH, Bang H, Kim DJ. How to Establish Clinical Prediction Models. Endocrinol Metab (Seoul, Korea). 2016 Mar;31(1):38–44.

4. Moons KGM, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart. 2012 May;98(9):683–90.

5. Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? 2009;338:b375.

6. Cadarette SM, Jaglal SB, Kreiger N, McIsaac WJ, Darlington GA, Tu J V. Development and validation of the Osteoporosis Risk Assessment Instrument to facilitate selection of women for bone densitometry. CMAJ. 2000 May 2;162(9):1289–94.

7. D'Agostino RB, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, et al. General Cardiovascular Risk Profile for Use in Primary Care. Circulation. 2008;117(6).

8. Buijsse B, Simmons RK, Griffin SJ, Schulze MB. Risk assessment tools for identifying individuals at risk of developing type 2 diabetes. Epidemiol Rev. 2011 Jan 1;33(1):46–62.

9. Lloyd-Jones DM, Wilson PWF, Larson MG, Beiser A, Leip EP, D'Agostino RB, et al. Framingham risk score and prediction of lifetime risk for coronary heart disease. Am J Cardiol. 2004 Jul 1;94(1):20–4.

10. Walker JG, Licqurish S, Chiang PPC, Pirotta M, Emery JD. Cancer risk assessment tools in primary care: a systematic review of randomized controlled trials. Ann Fam Med. 2015

Sep;13(5):480–9.

11. Hall LML, Jung RT, Leese GP. Controlled trial of effect of documented cardiovascular risk scores on prescribing. BMJ. 2003;326(7383).

12. Marengoni A, Angleman S, Melis R, Mangialasche F, Karp A, Garmen A, et al. Aging with multimorbidity: A systematic review of the literature. Ageing Res Rev. 2011;10(4):430–9.

13. Guthrie B, Payne K, Alderson P, McMurdo MET, Mercer SW. Adapting clinical guidelines to take account of multimorbidity. Bmj. 2012;345(October):e6341.

14. Canada Health Infoway. Electronic Medical Records (EMR) Progress in Canada. 2016.

15. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol. 2016 Jun;74:167–76.

16. Hsiao C-J, Hing E, Ashman J. Trends in electronic health record system use among office-based physicians: United States, 2007-2012. Natl Health Stat Report. 2014 May 20;(75):1–18.

17. Schoen C, Osborn R. The Commonwealth Fund 2012 International Health Policy Survey of Primary Care Physicians 2012 International Symposium On Health Care Policy. 2012;

18. Demaio AR, Kragelund Nielsen K, Pinkowski Tersbøl B, Kallestrup P, Meyrowitsch DW. Primary Health Care: a strategic framework for the prevention and control of chronic non-communicable disease. Glob Health Action. 2014;7:24504.

19. Dankel SJ, Loenneke JP, Loprinzi PD. Participation in muscle-strengthening activities as an alternative method for the prevention of multimorbidity. Prev Med (Baltim). 2015 Dec;81:54–7.

20. García-Olmos L, Salvador CH, Alberquilla Á, Lora D, Carmona M, García-Sagredo P, et al. Comorbidity patterns in patients with chronic diseases in general practice. PLoS One.

2012;7(2):e32141.

21. Steinman MA, Lee SJ, John Boscardin W, Miao Y, Fung KZ, Moore KL, et al. Patterns of multimorbidity in elderly veterans. J Am Geriatr Soc. 2012 Oct;60(10):1872–80.

22. Prados-Torres A, Poblador-Plou B, Calderón-Larrañaga A, Gimeno-Feliu LA, González-Rubio F, Poncel-Falcó A, et al. Multimorbidity patterns in primary care: interactions among chronic diseases using factor analysis. PLoS One. 2012;7(2):e32190.

23. World Health Organization. The top 10 causes of death: fact sheet. WHO. 2017.

24. Abrams DB, Turner JR, Baumann LC, Karel A, Collins SE, Witkiewitz K, et al. Acute Disease. In: Encyclopedia of Behavioral Medicine. New York, NY: Springer New York; 2013. p. 27–27.

25. Harris RE. Epidemiology of Chronic Disease. Jones & Bartlett Learning; 2013. p. 724.

26. Bernell S, Howard SW. Use Your Words Carefully: What Is a Chronic Disease? Front public Heal. 2016;4:159.

27. Adams PF, Kirzinger WK, Martinez M. Summary health statistics for the U.S. population: National Health Interview Survey, 2012. Vital Health Stat 10. 2013;(259):1–95.

28. Public Health Agency of Canada. Canadian Chronic Disease Indicators, Quick Stats, 2017 Edition. Ottawa, ON; 2017.

29. Centre for Chronic Disease Prevention Public Health Agency of Canada. Chronic Disease and Injury Indicator Framework: Quick Stats, 2016 Edition. Ottawa, ON; 2016.

30. Porta MS, International Epidemiological Association. A dictionary of epidemiology. Oxford University Press; 2008. 289 p.

31. Schwartz AG, Cote ML. Epidemiology of Lung Cancer. In: Advances in experimental medicine and biology. 2016. p. 21–41.

32. Knol MJ, VanderWeele TJ. Recommendations for presenting analyses of effect

modification and interaction. Int J Epidemiol. 2012 Apr;41(2):514–20.

33. Ockene JK, Kuller LH, Svendsen KH, Meilahn E. The relationship of smoking cessation to coronary heart disease and lung cancer in the Multiple Risk Factor Intervention Trial (MRFIT). Am J Public Health. 1990 Aug;80(8):954–8.

34. Clair C, Rigotti NA, Porneala B, Fox CS, D'Agostino RB, Pencina MJ, et al. Association of Smoking Cessation and Weight Change With Cardiovascular Disease Among Adults With and Without Diabetes. JAMA. 2013 Mar 13;309(10):1014.

35. Ardisson Korat A V, Willett WC, Hu FB. Diet, lifestyle, and genetic risk factors for type 2 diabetes: a review from the Nurses' Health Study, Nurses' Health Study 2, and Health Professionals' Follow-up Study. Curr Nutr Rep. 2014 Dec 1;3(4):345–54.

36. Ding D, Chong S, Jalaludin B, Comino E, Bauman AE. Risk factors of incident type 2-diabetes mellitus over a 3-year follow-up: Results from a large Australian sample. Diabetes Res Clin Pract. 2015 May;108(2):306–15.

37. Cairns KE, Yap MBH, Pilkington PD, Jorm AF. Risk and protective factors for depression that adolescents can modify: a systematic review and meta-analysis of longitudinal studies. J Affect Disord. 2014 Dec 1;169:61–75.

38. Cooper C, Inskip H, Croft P, Campbell L, Smith G, McLaren M, et al. Individual risk factors for hip osteoarthritis: Obesity, hip injury and physical activity. Am J Epidemiol. 1998 Mar;147(6):516–22.

39. Lee KM, Chung CY, Sung KH, Lee SY, Won SH, Kim TG, et al. Risk Factors for Osteoarthritis and Contributing Factors to Current Arthritic Pain in South Korean Older Adults. Yonsei Med J. 2015 Jan;56(1):124.

40. Silverwood V, Blagojevic-Bucknall M, Jinks C, Jordan JL, Protheroe J, Jordan KP. Current evidence on risk factors for knee osteoarthritis in older adults: a systematic review and meta-analysis. Osteoarthritis Cartilage. 2014 Nov 29;23(4):507–15.

41. Garin N, Koyanagi A, Chatterji S, Tyrovolas S, Olaya B, Leonardi M, et al. Global

Multimorbidity Patterns: A Cross-Sectional, Population-Based, Multi-Country Study. J Gerontol A Biol Sci Med Sci. 2016;71(2):205–14.

42. Fortin M, Stewart M, Poitras M-E, Almirall J, Maddocks H. A systematic review of prevalence studies on multimorbidity: toward a more uniform methodology. Ann Fam Med. 2012;10(2):142–51.

43. Stewart M, Fortin M, Britt HC, Harrison CM, Maddocks HL. Comparisons of multi-morbidity in family practice--issues and biases. Fam Pract. 2013 Aug 1;30(4):473–80.

44. Diederichs C, Berger K, Bartels DB. The measurement of multiple chronic diseases--a systematic review on existing multimorbidity indices. J Gerontol A Biol Sci Med Sci. 2011 Mar 1;66(3):301–11.

45. Le Reste JY, Nabbe P, Manceau B, Lygidakis C, Doerr C, Lingner H, et al. The European General Practice Research Network presents a comprehensive definition of multimorbidity in family medicine and long term care, following a systematic review of relevant literature. J Am Med Dir Assoc. 2013 May;14(5):319–25.

46. Arokiasamy P, Uttamacharya U, Jain K, Biritwum RB, Yawson AE, Wu F, et al. The impact of multimorbidity on adult physical and mental health in low- and middle-income countries: what does the study on global ageing and adult health (SAGE) reveal? BMC Med. 2015 Dec 3;13(1):178.

47. Lawson KD, Mercer SW, Wyke S, Grieve E, Guthrie B, Watt GC, et al. Double trouble: the impact of multimorbidity and deprivation on preference-weighted health related quality of life a cross sectional analysis of the Scottish Health Survey. Int J Equity Health. 2013 Aug 20;12(1):67.

48. Fortin M, Lapointe L, Hudon C, Vanasse A, Ntetu AL, Maltais D. Multimorbidity and quality of life in primary care: a systematic review. Health Qual Life Outcomes. 2004 Sep 20;2:51.

49. Thavorn K. Multimorbidity And Health Syetem Costs Among Older Adults In Ontario, Canada. Smdm; 2016.

50. Wikström K, Lindström J, Harald K, Peltonen M, Laatikainen T. Clinical and lifestyle-related risk factors for incident multimorbidity: 10-year follow-up of Finnish population-based cohorts 1982–2012. Eur J Intern Med. 2015 Apr;26(3):211–6.

51. Dhalwani NN, O'Donovan G, Zaccardi F, Hamer M, Yates T, Davies M, et al. Long terms trends of multimorbidity and association with physical activity in older English population. Int J Behav Nutr Phys Act. 2016 Dec 19;13(1):8.

52. Navickas R, Petric V-K, Feigl AB, Seychell M. Multimorbidity: What do we know? What should we do? Vol. 6, Journal of Comorbidity. 2016. 4-11 p.

53. Roberts KC, Rao DP, Bennett TL, Loukine L, Jayaraman GC. Prevalence and patterns of chronic disease multimorbidity and associated determinants in Canada. Heal Promot Chronic Dis Prev Canada. 2015 Aug;35(6):87–94.

54. Holt RIG, Cockram CS, Flyvbjerg A, Goldstein BJ. Textbook of diabetes. 5th ed. Wiley-Blackwell; 2017.

55. Gardner DG, Shoback DM, Greenspan FS (Francis S. Greenspan's basic and clinical endocrinology. 9th ed. McGraw-Hill Medical; 2011.

56. Melmed S, Polonsky KS, Larsen PR, Kronenberg H. Williams textbook of endocrinology. 13th ed. Elsevier; 2016.

57. Zaccardi F, Webb DR, Yates T, Davies MJ. Pathophysiology of type 1 and type 2 diabetes mellitus: a 90-year perspective. Postgrad Med J. 2016 Feb;92(1084):63–9.

58. Dabelea D, Mayer-Davis EJ, Saydah S, Imperatore G, Linder B, Divers J, et al. Prevalence of Type 1 and Type 2 Diabetes Among Children and Adolescents From 2001 to 2009. JAMA. 2014 May 7;311(17):1778.

59. Greiver M, Williamson T, Barber D, Birtwhistle R, Aliarzadeh B, Khan S, et al. Prevalence and epidemiology of diabetes in Canadian primary care practices: a report from the Canadian Primary Care Sentinel Surveillance Network. Can J diabetes. 2014 Jun;38(3):179–85.

60. World Health Organization. Diabetes: fact sheet. WHO. 2013.

61. Nathan DM, Cleary PA, Backlund J-YC, Genuth SM, Lachin JM, Orchard TJ, et al. Intensive Diabetes Treatment and Cardiovascular Disease in Patients with Type 1 Diabetes. N Engl J Med. 2005 Dec 22;353(25):2643–53.

62. The Diabetes Control and Complications Trial Research Group. The Effect of Intensive Diabetes Therapy on the Development and Progression of Neuropathy. Ann Intern Med. 1995 Apr 15;122(8):561–8.

63. Lindström J, Tuomilehto J. The diabetes risk score: a practical tool to predict type 2 diabetes risk. Diabetes Care. 2003 Mar;26(3):725–31.

64. Bates GW, Legro RS. Longterm management of Polycystic Ovarian Syndrome (PCOS). Mol Cell Endocrinol. 2013 Jul 5;373(1–2):91–7.

65. Dixon L, Weiden P, Delahanty J, Goldberg R, Postrado L, Lucksted A, et al. Prevalence and correlates of diabetes in national schizophrenia samples. Schizophr Bull. 2000;26(4):903–12.

66. Regenold WT, Thapar RK, Marano C, Gavirneni S, Kondapavuluru P V. Increased prevalence of type 2 diabetes mellitus among psychiatric inpatients with bipolar I affective and schizoaffective disorders independent of psychotropic drug use. J Affect Disord. 2002 Jun;70(1):19–26.

67. Semenkovich K, Brown ME, Svrakic DM, Lustman PJ. Depression in type 2 diabetes mellitus: prevalence, impact, and treatment. Drugs. 2015 Apr 8;75(6):577–87.

68. Ruzickova M, Slaney C, Garnham J, Alda M. Clinical Features of Bipolar Disorder with and without Comorbid Diabetes Mellitus. Can J Psychiatry. 2003 Aug 24;48(7):458–61.

69. Rao X, Montresor-Lopez J, Puett R, Rajagopalan S, Brook RD. Ambient air pollution: an emerging risk factor for diabetes mellitus. Curr Diab Rep. 2015 Jun;15(6):603.

70. Wilmot E, Idris I. Early onset type 2 diabetes: risk factors, clinical impact and

management. Ther Adv Chronic Dis. 2014 Nov;5(6):234–44.

71. Knowler WC, Barrett-Connor E, Fowler SE, Hamman RF, Lachin JM, Walker EA, et al. Reduction in the Incidence of Type 2 Diabetes with Lifestyle Intervention or Metformin. N Engl J Med. 2002 Feb 7;346(6):393–403.

72. Pan XR, Li GW, Hu YH, Wang JX, Yang WY, An ZX, et al. Effects of diet and exercise in preventing NIDDM in people with impaired glucose tolerance. The Da Qing IGT and Diabetes Study. Diabetes Care. 1997 Apr;20(4):537–44.

73. Tuomilehto J, Lindström J, Eriksson JG, Valle TT, Hämäläinen H, Ilanne-Parikka P, et al. Prevention of Type 2 Diabetes Mellitus by Changes in Lifestyle among Subjects with Impaired Glucose Tolerance. N Engl J Med. 2001 May 3;344(18):1343–50.

74. Carlsson LMS, Peltonen M, Ahlin S, Anveden Å, Bouchard C, Carlsson B, et al. Bariatric Surgery and Prevention of Type 2 Diabetes in Swedish Obese Subjects. N Engl J Med. 2012 Aug 23;367(8):695–704.

75. Diabetes Prevention Program Research Group. Effects of withdrawal from metformin on the development of diabetes in the diabetes prevention program. Diabetes Care. 2003 Apr;26(4):977–80.

76. Diabetes Prevention Program Research Group, Knowler WC, Fowler SE, Hamman RF, Christophi CA, Hoffman HJ, et al. 10-year follow-up of diabetes incidence and weight loss in the Diabetes Prevention Program Outcomes Study. Lancet. 2009 Nov 14;374(9702):1677–86.

77. Knowler WC, Hamman RF, Edelstein SL, Barrett-Connor E, Ehrmann DA, Walker EA, et al. Prevention of type 2 diabetes with troglitazone in the Diabetes Prevention Program. Diabetes. 2005 Apr;54(4):1150–6.

78. DREAM (Diabetes REduction Assessment with ramipril and rosiglitazone Medication) Trial Investigators, Gerstein HC, Yusuf S, Bosch J, Pogue J, Sheridan P, et al. Effect of rosiglitazone on the frequency of diabetes in patients with impaired glucose tolerance or impaired fasting glucose: a randomised controlled trial. Lancet. 2006 Sep

23;368(9541):1096–105.

79. DREAM Trial Investigators, Bosch J, Yusuf S, Gerstein HC, Pogue J, Sheridan P, et al. Effect of Ramipril on the Incidence of Diabetes. N Engl J Med. 2006 Oct 12;355(15):1551–62.

80. DeFronzo RA, Tripathy D, Schwenke DC, Banerji M, Bray GA, Buchanan TA, et al. Pioglitazone for Diabetes Prevention in Impaired Glucose Tolerance. N Engl J Med. 2011 Mar 24;364(12):1104–15.

81. Chiasson J-L, Josse RG, Gomis R, Hanefeld M, Karasik A, Laakso M, et al. Acarbose for prevention of type 2 diabetes mellitus: the STOP-NIDDM randomised trial. Lancet. 2002 Jun 15;359(9323):2072–7.

82. Torgerson JS, Hauptman J, Boldrin MN, Sjöström L. XENical in the prevention of diabetes in obese subjects (XENDOS) study: a randomized study of orlistat as an adjunct to lifestyle changes for the prevention of type 2 diabetes in obese patients. Diabetes Care. 2004 Jan;27(1):155–61.

83. Astrup A, Rössner S, Van Gaal L, Rissanen A, Niskanen L, Al Hakim M, et al. Effects of liraglutide in the treatment of obesity: a randomised, double-blind, placebo-controlled study. Lancet (London, England). 2009 Nov 7;374(9701):1606–16.

84. Canadian Diabetes Association Clinical Practice Guidelines Expert Committee;, Cheng AYY. Canadian Diabetes Association 2013 clinical practice guidelines for the prevention and management of diabetes in Canada. Can J Diabetes. 2013 Apr;37:S1–3.

85. Naish J, Syndercombe Court D. Medical sciences. Elsevier; 2015.

86. Poulter NR, Prabhakaran D, Caulfield M. Hypertension. Lancet. 2015 Aug 22;386(9995):801–12.

87. Carretero OA, Oparil S. Essential hypertension. Part I: definition and etiology. Circulation. 2000 Jan 25;101(3):329–35.

88. Lackland DT, Weber MA. Global Burden of Cardiovascular Disease and Stroke: Hypertension at the Core. Can J Cardiol. 2015 May;31(5):569–71.

89. James PA, Oparil S, Carter BL, Cushman WC, Dennison-Himmelfarb C, Handler J, et al. 2014 Evidence-Based Guideline for the Management of High Blood Pressure in Adults. JAMA. 2014 Feb 5;311(5):507.

90. Houle SKD, Padwal R, Tsuyuki RT. The 2012-2013 Canadian Hypertension Education Program (CHEP) guidelines for pharmacists. Can Pharm J / Rev des Pharm du Canada. 2013 May 21;146(3):146–50.

91. Padwal RS, Bienek A, McAlister FA, Campbell NRC. Epidemiology of Hypertension in Canada: An Update. Can J Cardiol. 2016 May;32(5):687–94.

92. Godwin M, Williamson T, Khan S, Kaczorowski J, Asghari S, Morkem R, et al. Prevalence and management of hypertension in primary care practices with electronic medical records: a report from the Canadian Primary Care Sentinel Surveillance Network. C open. 2015 Feb 25;3(1):E76-82.

93. Devi P, Rao M, Sigamani A, Faruqui A, Jose M, Gupta R, et al. Prevalence, risk factors and awareness of hypertension in India: a systematic review. J Hum Hypertens. 2013 May;27(5):281–7.

94. Doulougou B, Gomez F, Alvarado B, Guerra RO, Ylli A, Guralnik J, et al. Factors associated with hypertension prevalence, awareness, treatment and control among participants in the International Mobility in Aging Study (IMIAS). J Hum Hypertens. 2015 Apr 2;

95. Gargiulo R, Suhail F, Lerma E V. Hypertension and chronic kidney disease. Disease-a-Month. 2015 Sep;61(9):387–95.

96. Floras JS. Hypertension and sleep apnea. Can J Cardiol. 2015 May;31(7):889–97.

97. Cuffee Y, Ogedegbe C, Williams N. Psychosocial risk factors for hypertension: an update of the literature. Curr Hypertens. 2014;

98. Licht CMM, de Geus EJC, Seldenrijk A, van Hout HPJ, Zitman FG, van Dyck R, et al. Depression Is Associated With Decreased Blood Pressure, but Antidepressant Use Increases the Risk for Hypertension. Hypertension. 2009 Apr 1;53(4):631–8.

99. Yang Q, Zhang Z, Kuklina E V, Fang J, Ayala C, Hong Y, et al. Sodium intake and blood pressure among US children and adolescents. Pediatrics. 2012 Oct;130(4):611–9.

100. Saneei P, Salehi-Abargouei A, Esmaillzadeh A, Azadbakht L. Influence of Dietary Approaches to Stop Hypertension (DASH) diet on blood pressure: A systematic review and meta-analysis on randomized controlled trials. Nutr Metab Cardiovasc Dis. 2014 Dec;24(12):1253–61.

101. Diaz KM, Shimbo D. Physical Activity and the Prevention of Hypertension. Curr Hypertens Rep. 2013 Dec 20;15(6):659–68.

102. Stevens VJ, Obarzanek E, Cook NR, Lee IM, Appel LJ, Smith West D, et al. Long-term weight loss and changes in blood pressure: results of the Trials of Hypertension Prevention, phase II. Ann Intern Med. 2001 Jan 2;134(1):1–11.

103. Daskalopoulou SS, Rabi DM, Zarnke KB, Dasgupta K, Nerenberg K, Cloutier L, et al. The 2015 Canadian Hypertension Education Program Recommendations for Blood Pressure Measurement, Diagnosis, Assessment of Risk, Prevention, and Treatment of Hypertension. Can J Cardiol. 2015 May;31(5):549–68.

104. Canadian Centre on Substance Use and Addiction. Canada's Low-Risk Alcohol Drinking Guidelines. Ottawa ON; 2017.

105. Doherty M, Bijlsma JWJ (Johannes WJ, Arden N (Nigel), Hunter D (David), Dalbeth N. osteoarthritis and crystal arthropathy. 3rd ed. Oxford University Press; 2016. 507 p.

106. Filardo G, Kon E, Longo UG, Madry H, Marchettini P, Marmotti A, et al. Non-surgical treatments for the management of early osteoarthritis. Knee Surgery, Sport Traumatol Arthrosc. 2016 Jun 4;24(6):1775–85.

107. Bennell KL, Buchbinder R, Hinman RS. Physical therapies in the management of

osteoarthritis. Curr Opin Rheumatol. 2015 May;27(3):304–11.

108. McAlindon TE, Bannuru RR, Sullivan MC, Arden NK, Berenbaum F, Bierma-Zeinstra SM, et al. OARSI guidelines for the non-surgical management of knee osteoarthritis. Osteoarthr Cartil. 2014 Mar;22(3):363–88.

109. Cibulka MT, White DM, Woehrle J, Harris-Hayes M, Enseki K, Fagerson TL, et al. Hip Pain and Mobility Deficits — Hip Osteoarthritis: Clinical Practice Guidelines Linked to the International Classification of Functioning, Disability, and Healthfrom the Orthopaedic Section of the American Physical Therapy Association. J Orthop Sport Phys Ther. 2009 Apr;39(4):A1–25.

110. Santaguida PL, Hawker GA, Hudak PL, Glazier R, Mahomed NN, Kreder HJ, et al. Patient characteristics affecting the prognosis of total hip and knee joint arthroplasty: a systematic review. Can J Surg. 2008 Dec;51(6):428–36.

111. Carr AJ, Robertsson O, Graves S, Price AJ, Arden NK, Judge A, et al. Knee replacement. Lancet. 2012 Apr 7;379(9823):1331–40.

112. Jenkins PJ, Clement ND, Hamilton DF, Gaston P, Patton JT, Howie CR. Predicting the cost-effectiveness of total hip and knee replacement: A health economic analysis. Bone Joint J. 2013 Jan 1;95–B(1):115–21.

113. Daigle ME, Weinstein AM, Katz JN, Losina E. The cost-effectiveness of total joint arthroplasty: A systematic review of published literature. Best Pract Res Clin Rheumatol. 2012 Oct;26(5):649–58.

114. Birtwhistle R, Morkem R, Peat G, Williamson T, Green ME, Khan S, et al. Prevalence and management of osteoarthritis in primary care: an epidemiologic cohort study from the Canadian Primary Care Sentinel Surveillance Network. C open. 2015;3(3):E270-5.

115. Harvey WF, Yang M, Cooke TD V., Segal NA, Lane N, Lewis CE, et al. Association of Leg-Length Inequality With Knee Osteoarthritis A Cohort Study. Ann Intern Med. 2010 Mar;152(5):287-W92.

116. Leung GJ, Rainsford KD, Kean WF. Osteoarthritis of the hand I: aetiology and pathogenesis, risk factors, investigation and diagnosis. J Pharm Pharmacol. 2014 Mar;66(3):339–46.

117. Neogi T, Zhang Y. Osteoarthritis prevention. Curr Opin Rheumatol. 2011 Mar;23(2):185–91.

118. Vrezas I, Elsner G, Bolm-Audorff U, Abolmaali N, Seidler A. Case-control study of knee osteoarthritis and lifestyle factors considering their interaction with physical workload. Int Arch Occup Environ Health. 2010 Mar;83(3):291–300.

119. Vignon E, Valat J-P, Rossignol M, Avouac B, Rozenberg S, Thoumie P, et al. Osteoarthritis of the knee and hip and activity: a systematic international review and synthesis (OASIS). Joint Bone Spine. 2006 Jul;73(4):442–55.

120. Runhaar J, van Middelkoop M, Reijman M, Willemsen S, Oei EH, Vroegindeweij D, et al. Prevention of Knee Osteoarthritis in Overweight Females: The First Preventive Randomized Controlled Trial in Osteoarthritis. Am J Med. 2015 Aug;128(8):888–895.e4.

121. Emery CA, Roos EM, Verhagen E, Finch CF, Bennell KL, Story B, et al. OARSI Clinical Trials Recommendations: Design and conduct of clinical trials for primary prevention of osteoarthritis by joint injury prevention in sport and recreation. Osteoarthr Cartil. 2015 May;23(5):815–25.

122. Gordon RS, Jr. An operational classification of disease prevention. Public Health Rep. 1983;98(2):107–9.

123. Katz DL, Ali A. Preventive medicine, integrative medicine & the health of the public. Summit Integr Med Heal Public. 2009;45.

124. JL K. Breast cancer epidemiology: summary and future directions. Epidemiol Rev. 1993;15(1):256–63.

125. Loring K. Chronic disease self-management: a model for tertiary prevention. Kango Kenkyu. 1996 May 27;31(1):23–9.

126. Cox JL, Vallis TM, Pfammatter A, Szpilfogel C, Carr B, O'Neill BJ. A novel approach to cardiovascular health by optimizing risk management (ANCHOR): behavioural modification in primary care effectively reduces global risk. Can J Cardiol. 2013 Nov 1;29(11):1400–7.

127. Saini P, While D, Chantler K, Windfuhr K, Kapur N. Assessment and Management of Suicide Risk in Primary Care. Crisis. 2014 Nov;35(6):415–25.

128. Phelan EA, Mahoney JE, Voit JC, Stevens JA. Assessment and Management of Fall Risk in Primary Care Settings. Med Clin North Am. 2015 Mar;99(2):281–93.

129. Viera AJ, Sheridan SL. Global risk of coronary heart disease: assessment and application. Am Fam Physician. 2010 Aug 1;82(3):265–74.

130. Redon J. Global Cardiovascular Risk Assessment: Strengths and Limitations. High Blood Press Cardiovasc Prev. 2016 Jun 18;23(2):87–90.

131. Geisser S. Predictive Inference: An Introduction. Chapman & Hall; 1993. 264 p.

132. Kuhn M, Johnson K. Applied predictive modeling. 600 p.

133. Nashef SAM, Roques F, Michel P, Gauducheau E, Lemeshow S, Salamon R. European system for cardiac operative risk evaluation (EuroSCORE). Eur J Cardio-Thoracic Surg. 1999 Jul 1;16(1):9–13.

134. World Health Organization. Prevention of Cardiovascular Disease Guidelines for assessment and management of cardiovascular risk WHO Library Cataloguing-in-Publication Data. 2007;

135. Leyland AH (Alastair H., Goldstein H. Multilevel modelling of health statistics. Wiley; 2001. 217 p.

136. Miettinen OS (Olli S, Karp I. Epidemiological research : an introduction. Springer; 2012.

137. Moons KGM, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ. 2009 Jun

4;338:b606.

138. Anderson TJ, Gr J, Pearson GJ, Barry AR, Couture P, Dawes M, et al. 2016 Canadian Cardiovascular Society Guidelines for the Management of Dyslipidemia for the Prevention of Cardiovascular Disease in the Adult. 2016;

139. Papaioannou A, Morin S, Cheung AM, Atkinson S, Brown JP, Feldman S, et al. 2010 clinical practice guidelines for the diagnosis and management of osteoporosis in Canada: summary. CMAJ. 2010 Nov 23;182(17):1864–73.

140. Müller-Riemenschneider F, Holmberg C, Rieckmann N, Kliems H, Rufer V, Müller-Nordhorn J, et al. Barriers to Routine Risk-Score Use for Healthy Primary Care Patients. Arch Intern Med. 2010 Apr 26;170(8):719.

141. Public Health England. NHS Health Check: Best Practice Guidance. London, England; 2016.

142. Artac M, Dalton ARH, Majeed A, Car J, Millett C. Effectiveness of a national cardiovascular disease risk assessment program (NHS Health Check): Results after one year. Prev Med (Baltim). 2013 Aug 1;57(2):129–34.

143. Jiao F, Fung CSC, Wan YF, McGhee SM, Wong CKH, Dai D, et al. Long-term effects of the multidisciplinary risk assessment and management program for patients with diabetes mellitus (RAMP-DM): a population-based cohort study. Cardiovasc Diabetol. 2015 Aug 14;14(1):105.

144. Brindle P, Beswick A, Fahey T, Ebrahim S. Accuracy and impact of risk assessment in the primary prevention of cardiovascular disease: a systematic review. Heart. 2006 Dec 1;92(12):1752–9.

145. Khanji MY, Bicalho VVS, van Waardhuizen CN, Ferket BS, Petersen SE, Hunink MGM, et al. Cardiovascular Risk Assessment. Ann Intern Med. 2016 Nov 15;165(10):713.

146. Goetzel RZ, Staley P, Ogden L, Stange P, Fox J, Spangler J, et al. A Framework for Patient-Centered Health Risk Assessments - Providing Health Promotion and Disease

Prevetion Services to Medicare Beneficiaries. Centers Dis Control Prev. 2011;52.

147. Wilson PWF, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of Coronary Heart Disease Using Risk Factor Categories. Circulation. 1998;97(18).

148. O'Brien SM, Shahian DM, Filardo G, Ferraris VA, Haan CK, Rich JB, et al. The Society of Thoracic Surgeons 2008 Cardiac Surgery Risk Models: Part 2-Isolated Valve Surgery. Ann Thorac Surg. 2009;88(1 SUPPL.):S23–42.

149. Cortez S, Milbrandt M, Kaphingst K, James A, Colditz G. The readability of online breast cancer risk assessment tools. Breast Cancer Res Treat. 2015 Nov;154(1):191–9.

150. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. Ann Intern Med. 2015 Jan 6;162(1):W1.

151. Grobbee D, Hoes A. Clinical epidemiology: principles, methods, and applications for clinical research. 2009.

152. Wang LE, Shaw PA, Mathelier HM, Kimmel SE, French B. Evaluating Risk-Prediction Models Using Data From Electronic Health Records. Ann Appl Stat. 2016 Mar;10(1):286–304.

153. Steyerberg EW, Mushkudiani N, Perel P, Butcher I, Lu J, McHugh GS, et al. Predicting Outcome after Traumatic Brain Injury: Development and International Validation of Prognostic Scores Based on Admission Characteristics. Singer M, editor. PLoS Med. 2008 Aug 5;5(8):e165.

154. Montgomery DC, Peck EA. Introduction to Linear Regression Analysis. John Wiley Sons, Inc 5th Ed. 2012;672.

155. Hosmer DW, Lemeshow S, Sturdivant RX. Applied logistic regression.

156. Harrell FE. Cox Proportional Hazards Regression Model. In Springer, New York, NY; 2001. p. 465–507.

157. Wood SN. Generalized additive models : an introduction with R. Chapman & Hall/CRC; 2006. 391 p.

158. Quinlan JR. Induction of decision trees. Mach Learn. 1986 Mar;1(1):81–106.

159. Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inf Theory. 1967 Jan;13(1):21–7.

160. Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. BMJ. 2009 May 28;338:b605.

161. Kohavi R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. 1995;

162. Johnson RW. An Introduction to the Bootstrap. Teach Stat. 2001 Jun 1;23(2):49–54.

163. Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol. 2014;14:40.

164. Pencina MJ, D'Agostino RB, JB R, M G, R P, G S, et al. Evaluating Discrimination of Risk Prediction Models. JAMA. 2015 Sep 8;314(10):1063.

165. Reilly BM, Evans AT, CS R, NR P, AW W, MH W, et al. Translating Clinical Research into Clinical Practice: Impact of Using Prediction Rules To Make Decisions. Ann Intern Med. 2006 Feb 7;144(3):201.

166. Ross SM. A first course in probability. 9th ed. Pearson; 2012. 467 p.

167. Joe H. Multivariate models and dependence concepts. Chapman & Hall; 1997. 399 p. (Monographs on statistics and applied probability).

168. Trivedi PK, Zimmer DM. Copula modeling: an introduction for practitioners *. Found

Trends Econom. 2005 Jan 1;1(1):1–2.

169. A. S. Fonctions de repartition a n dimensions et leurs marges. Publ Inst Stat Univ Paris. 1959;8:229–31.

170. Frank MJ. On the simultaneous associativity of F(x, y) and x+y −F(x, y). Aequationes Math. 1979;21:194–226.

171. Gumbel EJ. Bivariate exponential distributions. J Am Stat Assoc. 1960;55:698–707.

172. Clayton DG. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. Biometrika. 1978 Apr 1;65(1):141–51.

173. Canada Health Infoway. The Emerging Benefits of Electronic Medical Record Use in Community-Based Care: Full Report - Canada Health Infoway. 2013.

174. Tu K, Widdifield J, Young J, Oud W, Ivers NM, Butt DA, et al. Are family physicians comprehensively using electronic medical records such that the data can be used for secondary purposes? A Canadian perspective. BMC Med Inform Decis Mak. 2015 Aug 13;15(1):67.

175. Terry AL, Chevendra V, Thind A, Stewart M, Marshall JN, Cejic S. Using your electronic medical record for research: a primer for avoiding pitfalls. Fam Pract. 2010 Feb;27(1):121–6.

176. Saverno KR, Hines LE, Warholak TL, Grizzle AJ, Babits L, Clark C, et al. Ability of pharmacy clinical decision-support software to alert users about clinically important drug—drug interactions. J Am Med Informatics Assoc. 2011 Jan 1;18(1):32–7.

177. Kaplan B. Evaluating informatics applications—clinical decision support systems literature review. Int J Med Inform. 2001 Nov;64(1):15–37.

178. Comer PJ, Huntly PJ. TSE risk assessments: a decision support tool. Stat Methods Med Res. 2003 Jun 2;12(3):279–91.

179. OSCAR EMR | Clinical Management System. 2017.

180. Canadian Institution for Health Information. How Canada Compares: Results from the Commonwealth Fund 2015 International Health Policy Survey of Primary Care Physicians. Ottawa; 2016.

181. Birtwhistle R V. Canadian Primary Care Sentinel Surveillance Network. Canadian Family Physician 2011.

182. Wong ST, Manca D, Barber D, Morkem R, Khan S, Kotecha J, et al. The diagnosis of depression and its treatment in Canadian primary care practices: an epidemiological study. C Open. 2014 Nov 28;2(4):E337–42.

183. Kapetanios E, Tatar D, Sacarea C. Natural language processing : semantic aspects. CRC Press; 2014.

184. Birtwhistle R, Keshavjee K, Lambert-Lanning A, Godwin M, Greiver M, Manca D, et al. Building a pan-Canadian primary care sentinel surveillance network: initial development and moving forward. J Am Board Fam Med. 2009 Jan;22(4):412–22.

185. Queenan JA, Williamson T, Khan S, Drummond N, Garies S, Morkem R, et al. Representativeness of patients and providers in the Canadian Primary Care Sentinel Surveillance Network: a cross-sectional study. C open. 2016;4(1):E28-32.

186. Barber J, Muller S, Whitehurst T, Hay E. Measuring morbidity: self-report or health care records? Fam Pract. 2010 Feb 1;27(1):25–30.

187. Lucyk K, Lu M, Sajobi T, Quan H. Administrative health data in Canada: lessons from history. BMC Med Inform Decis Mak. 2015 Aug 19;15:69.

188. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JP. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc. 2016 May 17;

189. Nie JX, Wang L, Tracy CS, Moineddin R, Upshur RE. Health care service utilization

among the elderly: findings from the Study to Understand the Chronic Condition Experience of the Elderly and the Disabled (SUCCEED project). J Eval Clin Pract. 2008 Dec;14(6):1044–9.

190. Bertakis KD, Azari R, Helms LJ, Callahan EJ, Robbins JA. Gender differences in the utilization of health care services. J Fam Pract. 2000 Feb;49(2):147–52.

191. Mustard CA, Kaufert P, Kozyrskyj A, Mayer T. Sex Differences in the Use of Health Care Services. N Engl J Med. 1998 Jun 4;338(23):1678–83.

192. Touchie C, Medical Education Advisor C. Medical Council Of Canada Incidence and Prevalence of Diseases and Other Health Related Issues in Canada: Secondary Study for MCC Blueprint Project. 2013;

193. Williamson T, Green ME, Birtwhistle R, Khan S, Garies S, Wong ST, et al. Validating the 8 CPCSSN Case Definitions for Chronic Disease Surveillance in a Primary Care Database of Electronic Health Records. Ann Fam Med. 2014 Jul 14;12(4):367–72.

194. Statistics Canada. National Household Survey Profile, 2011. 2013.

195. Statistics Canada. How Postal Codes Map to Geographic Areas. Geostatistics Canada. 2007;

196. E. Harrell F. Regression Modeling Strategies: With Applications to Linear Models, Logistic .... 2001;

197. van Smeden M, de Groot JAH, Moons KGM, Collins GS, Altman DG, Eijkemans MJC, et al. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. BMC Med Res Methodol. 2016 Dec 24;16(1):163.

198. Graham JW. Missing Data Theory. In: Missing Data: Analysis and Design. New York, NY: Springer New York; 2012. p. 3–46.

199. Engels JM, Diehr P. Imputation of missing longitudinal data: A comparison of methods. J Clin Epidemiol. 2003 Oct 1;56(10):968–76.

200. Rubin DB, Wiley J, New York Chichester Brisbane Toronto Singapore S. Multiple Imputation for Nonresponse in Surveys. 1987.

201. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res. 2011 Mar;20(1):40–9.

202. Genest C, Nikoloulopoulos AK, Rivest L-P, Fortin M. Predicting dependent binary outcomes through logistic regressions and meta-elliptical copulas. Brazilian J Probab Stat. 2013 Aug;27(3):265–84.

203. Zhao Y, Joe H. Composite Likelihood Estimation in Multivariate Data Analysis. Vol. 33, The Canadian Journal of Statistics / La Revue Canadienne de Statistique. Statistical Society of Canada; 2005. p. 335–56.

204. Cramer H. Mathematical Methods of Statistics. Princet Univ Press. 1946;282.

205. Kim S. Partial and Semi-Partial (Part) Correlation. 2015;

206. Efron B TR. An Introduction to the Bootstrap. Chapman & Hall. 1994.

207. Gray RM. Entropy and Information Theory. New York. Springer; 2009. 409 p.

208. Centre for Chronic Disease Prevention Public Health Agency of Canada. Chronic Disease Indicator Framework, 2013 Edition. 2013;

209. Government of Canada Statistics Canada. High blood pressure, 2011. 2013;

210. Government of Canada Public Health Agency of Canada. What is Depression? 2016;

211. Government of Canada Statistics Canada. Female Population, 2010. 2015;

212. Bulloch A, Lavorato D, Williams J, Patten S. Alcohol consumption and major depression in the general population: the critical important of dependence. Depress Anxiety. 2012 Dec;29(12):1058–64.

213. Government of Canada Statistics Canada. Epilepsy in Canada: Prevalence. 2016;

214. Government of Canada Public Health Agency of Canada. A Report on Mental Illnesses in Canada - Schizophrenia. 2012;

215. Offord DR, Boyle MH, Campbell D, Goering P, Lin E, Wong M, et al. One-Year Prevalence of Psychiatric Disorder in Ontarians 15 to 64 Years of Age. Can J Psychiatry. 1996 Nov 26;41(9):559–64.

216. Widdifield J, Paterson JM, Bernatsky S, Tu K, Tomlinson G, Kuriya B, et al. The Epidemiology of Rheumatoid Arthritis in Ontario, Canada. Arthritis Rheumatol. 2014 Apr;66(4):786–93.

217. Lujan ME, Chizen DR, Pierson RA. Diagnostic criteria for polycystic ovary syndrome: pitfalls and controversies. J Obstet Gynaecol Can. 2008 Aug;30(8):671–9.

218. Arora P, Vasa P, Brenner D, Iglar K, McFarlane P, Morrison H, et al. Prevalence estimates of chronic kidney disease in Canada: results of a nationally representative survey. Can Med Assoc J. 2013 Jun 11;185(9):E417–23.

219. Government of Canada PHA of C. What is the impact of osteoporosis in Canada and what are Canadians doing to maintain healthy bones? 2010;

220. Government of Canada Statistics Canada. Census Profile, 2011. 2012;

221. Government of Canada Statistics Canada. Population by age and sex, 2012. 2015;

222. Government of Canada Statistics Canada. Body composition of Canadian adults, 2009 to 2011. 2013;

223. Bang H, Edwards AM, Bomback AS, Ballantyne CM, Brillon D, Callahan MA, et al. Development and validation of a patient self-assessment score for diabetes risk. Ann Intern Med. 2009 Dec 1;151(11):775–83.

224. Kaczorowski J, Robinson C, Nerenberg K. Development of the CANRISK questionnaire to screen for prediabetes and undiagnosed type 2 diabetes. Can J Diabetes. 2009 Jan 1;33(4):381–5.

225. Gray LJ, Taub NA, Khunti K, Gardiner E, Hiles S, Webb DR, et al. The Leicester Risk Assessment score for detecting undiagnosed Type 2 diabetes and impaired glucose regulation for use in a multiethnic UK setting. Diabet Med. 2010 Aug 3;27(8):887–95.

226. Chen L, Magliano DJ, Balkau B, Colagiuri S, Zimmet PZ, Tonkin AM, et al. AUSDRISK: an Australian Type 2 Diabetes Risk Assessment Tool based on demographic, lifestyle and simple anthropometric measures. Med J Aust. 2010 Feb 15;192(4):197–202.

227. Parikh NI, Pencina MJ, Wang TJ, Benjamin EJ, Lanier KJ, Levy D, et al. A risk score for predicting near-term incidence of hypertension: the Framingham Heart Study. Ann Intern Med. 2008 Jan 15;148(2):102–10.

228. 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 May;122(5):464–71.

229. Kshirsagar A V, Chiu Y-L, Bomback AS, August PA, Viera AJ, Colindres RE, et al. A hypertension risk score for middle-aged and older adults. J Clin Hypertens (Greenwich). 2010 Oct 8;12(10):800–8.

230. Joseph GB, McCulloch CE, Nevitt MC, Neumann J, Gersing AS, Kretzschmar M, et al. Tool for osteoarthritis risk prediction (TOARP) over 8 years using baseline clinical data, X-ray, and MRI: Data from the osteoarthritis initiative. J Magn Reson Imaging. 2018 Jun;47(6):1517–26.

231. Zhang W, McWilliams DF, Ingham SL, Doherty SA, Muthuri S, Muir KR, et al. Nottingham knee osteoarthritis risk prediction models. Ann Rheum Dis. 2011 Sep 1;70(9):1599–604.

232. Kerkhof HJM, Bierma-Zeinstra SMA, Arden NK, Metrustry S, Castano-Betancourt M, Hart DJ, et al. Prediction model for knee osteoarthritis incidence, including clinical, genetic and biochemical risk factors. Ann Rheum Dis. 2014 Dec;73(12):2116–21.

233. Riddle DL, Stratford PW, Perera RA. The incident tibiofemoral osteoarthritis with rapid progression phenotype: development and validation of a prognostic prediction rule.

Osteoarthr Cartil. 2016 Dec;24(12):2100–7.

234. Ferrannini E, Cushman WC. Diabetes and hypertension: the bad companions. Lancet. 2012 Aug;380(9841):601–10.

235. Florey C V, Uppal S, Lowy C. Relation between blood pressure, weight, and plasma sugar and serum insulin levels in schoolchildren aged 9-12 years in Westland, Holland. Br Med J. 1976 Jun 5;1(6022):1368–71.

236. Stamler J, Stamler R, Rhomberg P, Dyer A, Berkson DM, Reedus W, et al. Multivariate analysis of the relationship of six variables to blood pressure: findings from Chicago community surveys, 1965--1971. J Chronic Dis. 1975 Nov;28(10):499–525.

237. Salomaa V V, Strandberg TE, Vanhanen H, Naukkarinen V, Sarna S, Miettinen TA. Glucose tolerance and blood pressure: long term follow up in middle aged men. BMJ. 1991 Mar 2;302(6775):493–6.

238. Medalie JH, Papier CM, Goldbourt U, Herman JB. Major Factors in the Development of Diabetes Mellitus in 10,000 Men. Arch Intern Med. 1975 Jun 1;135(6):811.

239. Louati K, Vidal C, Berenbaum F, Sellam J. Association between diabetes mellitus and osteoarthritis: systematic literature review and meta-analysis. RMD Open. 2015 Jun 2;1(1).

240. Schett G, Kiechl S, Bonora E, Zwerina J, Mayr A, Axmann R, et al. Vascular cell adhesion molecule 1 as a predictor of severe osteoarthritis of the hip and knee joints. Arthritis Rheum. 2009 Aug;60(8):2381–9.

241. Yoshimura N, Muraki S, Oka H, Kawaguchi H, Nakamura K, Akune T. Association of knee osteoarthritis with the accumulation of metabolic risk factors such as overweight, hypertension, dyslipidemia, and impaired glucose tolerance in Japanese men and women: the ROAD study. J Rheumatol. 2011 May 1;38(5):921–30.

242. Hart DJ, Doyle D V, Spector TD. Association between metabolic factors and knee osteoarthritis in women: the Chingford Study. J Rheumatol. 1995 Jun;22(6):1118–23.

243. Visser W, den Heijer M, Spoelman W, Rosendaal FR, Huizinga TW, Kloppenburg M. Glucose and Insulin Concentrations in Association with Hand Osteoarthritis: The Neo Study. Ann Rheum Dis. 2013 Jun 23;72(Suppl 3):A702.1-A702.

244. Nieves-Plaza M, Castro-Santana LE, Font YM, Mayor AM, Vilá LM. Association of Hand or Knee Osteoarthritis With Diabetes Mellitus in a Population of Hispanics From Puerto Rico. JCR J Clin Rheumatol. 2013 Jan;19(1):1.

245. Yoshimura N, Muraki S, Oka H, Tanaka S, Kawaguchi H, Nakamura K, et al. Accumulation of metabolic risk factors such as overweight, hypertension, dyslipidaemia, and impaired glucose tolerance raises the risk of occurrence and progression of knee osteoarthritis: a 3-year follow-up of the ROAD study. Osteoarthr Cartil. 2012 Nov;20(11):1217–26.

246. Dahaghin S, Bierma-Zeinstra SMA, Koes BW, Hazes JMW, Pols HAP. Do metabolic factors add to the effect of overweight on hand osteoarthritis? The Rotterdam Study. Ann Rheum Dis. 2007 Jul 2;66(7):916–20.

247. Puenpatom RA, Victor TW. Increased Prevalence of Metabolic Syndrome in Individuals with Osteoarthritis: An Analysis of NHANES III Data. Postgrad Med. 2009 Nov 13;121(6):9–20.

248. Engström G, Gerhardsson de Verdier M, Rollof J, Nilsson PM, Lohmander LS. C-reactive protein, metabolic syndrome and incidence of severe hip and knee osteoarthritis. A population-based cohort study. Osteoarthr Cartil. 2009 Feb;17(2):168–73.

249. Marks R, Allegrante JP. Comorbid disease profiles of adults with end-stage hip osteoarthritis. Med Sci Monit. 2002 Apr;8(4).

250. Conaghan PG, Vanharanta H, Dieppe PA. Is progressive osteoarthritis an atheromatous vascular disease? Ann Rheum Dis. 2005 Nov 26;64(11):1539–41.

251. Bae Y-H, Shin J-S, Lee J, Kim M, Park KB, Cho J-H, et al. Association between Hypertension and the Prevalence of Low Back Pain and Osteoarthritis in Koreans: A Cross-Sectional Study. Fuchs FD, editor. PLoS One. 2015 Sep 22;10(9):e0138790.

252. Kim HS, Shin J-S, Lee J, Lee YJ, Kim M-R, Bae Y-H, et al. Association between Knee Osteoarthritis, Cardiovascular Risk Factors, and the Framingham Risk Score in South Koreans: A Cross-Sectional Study. Böttcher Y, editor. PLoS One. 2016 Oct 20;11(10):e0165325.

253. Findlay DM. Vascular pathology and osteoarthritis. Rheumatology. 2007 Aug 5;46(12):1763–8.

254. Imhof H, Sulzbacher I, Grampp S, Czerny C, Youssefzadeh S, Kainberger F. Subchondral bone and cartilage disease: a rediscovered functional unit. Invest Radiol. 2000 Oct;35(10):581–8.

255. Berger CE, Kröner AH, Minai-Pour MB, Ogris E, Engel A. Biochemical markers of bone metabolism in bone marrow edema syndrome of the hip. Bone. 2003 Sep;33(3):346–51.

256. Wang X, Wang F, Hu J, Sorrentino R. Exploring joint disease risk prediction. AMIA . Annu Symp proceedings AMIA Symp. 2014;2014:1180–7.

257. Wu Y, Ding Y, Tanaka Y, Zhang W. Risk Factors Contributing to Type 2 Diabetes and Recent Advances in the Treatment and Prevention. Int J Med Sci. 2014;11(11):1185–200.

258. Delli AJ, Larsson HE, Ivarsson S-A, Lernmark A, Kong APS, Chan JCN. Type 1 Diabetes. In: Textbook of Diabetes. Oxford, UK: Wiley-Blackwell; 2010. p. 139–59.

259. Canadian Diabetes Association. An economic tsunami: the cost of diabetes in Canada. 2009.

260. Levey AS, Becker C, Inker LA, Author C. Glomerular Filtration Rate and Albuminuria for Detection and Staging of Acute and Chronic Kidney Disease in Adults: A Systematic Review HHS Public Access. JAMA Febr. 2015;24(3138):837–46.

261. Medical Council of Canada. Objectives for the Qualifying Examination: Clinical Laboratory Tests Normal Values. 2017.

# Appendices

## **Appendix A: Summary of Validation Results for CPCSSN Diseases** (193)

| **Condition** | **Sensitivity %<br>(95% CI)** | **Specificity %<br>(95% CI)** | **PPV %<br>(95% CI)** | **NPV %<br>(95% CI)** |
|---|---|---|---|---|
| Hypertension | 84.9<br>(82.6 to 87.1) | 93.5<br>(92.0 to 95.1) | 92.9<br>(91.2 to 94.6) | 86.0<br>(83.9 to 88.2) |
| Diabetes | 95.6 | 97.1 | 87.0 | 99.1 |

| | | | | |
|---|---|---|---|---|
| | (93.4 to 97.9) | (96.3 to 97.9) | (83.5 to 90.5) | (98.6 to 99.6) |
| Depression | 81.1<br>(77.2 to 85.0) | 94.8<br>(93.7 to 95.9) | 79.6<br>(75.7 to 83.6) | 95.2<br>(94.1 to 96.3) |
| COPD | 82.1<br>(76.0 to 88.2) | 97.3<br>(96.5 to 98.0) | 72.1<br>(65.4 to 78.8) | 98.4<br>(97.9 to 99.0) |
| Osteoarthritis | 77.8<br>(74.5 to 81.1) | 94.9<br>(93.8 to 96.1) | 87.7<br>(84.9 to 90.5) | 90.2<br>(88.7 to 91.8) |
| Dementia | 96.8<br>(93.3 to 100.0) | 98.1<br>(97.5 to 98.7) | 72.8<br>(65.0 to 80.6) | 99.8<br>(99.6 to 100.0) |
| Epilepsy | 98.6<br>(96.6 to 100.0) | 98.7<br>(98.2 to 99.2) | 85.6<br>(80.2 to 91.1) | 99.9<br>(99.7 to 100.0) |
| Parkinsonism | 98.8<br>(96.4 to 100.0) | 99.0<br>(98.6 to 99.5) | 82.0<br>(74.5 to 89.5) | 99.9<br>(99.8 to 100.0) |

COPD: chronic obstructive pulmonary disease; NPV: negative predictive value; PPV: positive predictive value

## Appendix B: Risk Factor Case Definitions

| Risk Factor | Table Name | Value |
|---|---|---|
| Alcohol | Billing<br>Health Condition<br>Encounter Diagnosis | ICD-9 Codes:<br>• 303: Alcohol dependence syndrome<br>• 305.0: Non-dependent alcohol abuse |
| | Health Condition | Inclusion:<br>• "alcohol"<br>Exclusion:<br>• "fam"<br>• "no"<br>• "FAS" |
| | Encounter Diagnosis | Inclusion:<br>• "alcohol dependence"<br>• "alcohol abuse"<br>• "alcoholism" |
| | Risk Factor | Inclusion:<br>• "alcohol"<br>Exclusion:<br>• "no"<br>• "alcohol n"<br>• "alcohol -" |
| Epilepsy | Disease Case* | Epilepsy |
| Stress | Billing<br>Health Condition<br>Encounter Diagnosis | ICD-9 Codes:<br>• 308: Acute reaction to stress<br>• 309: Adjustment reaction |
| | Risk Factor | Inclusion:<br>• "stress"<br>Exclusion: |

| | | |
|---|---|---|
| | | • “no” |
| Schizophrenia | Billing<br>Health Condition<br>Encounter Diagnosis | ICD-9 Code:<br>• 295: Schizophrenic disorders |
| | Health Condition<br>Encounter Diagnosis | Inclusion:<br>• “schizo”<br>Exclusion:<br>• “fam” |
| | Medication | Prescription of second-generation anti-psychotics:<br>• Aripiprazole (Abilify)<br>• Asenapine (Saphris)<br>• Brexpiprazole (Rexulti)<br>• Cariprazine (Vraylar)<br>• Clozapine (Clozaril)<br>• Iloperidone (Fanapt)<br>• Lurasidone (Latuda)<br>• Olanzapine (Zyprexa)<br>• Paliperidone (Invega)<br>• Quetiapine (Seroquel)<br>• Risperidone (Risperdal)<br>• Ziprasidone (Geodon)<br>Prescription of first-generation anti-psychotics:<br>• Chlorpromazine<br>• Fluphenazine<br>• Haloperidol<br>• Perphenazine |
| Anxiety | Billing<br>Health Condition<br>Encounter Diagnosis | ICD-9 Code:<br>• 300.0: anxiety related neurotic disorders |
| | Health Condition | Inclusion:<br>• “anxiety”<br>Exclusion:<br>• “fam” |
| | Encounter Diagnosis | Inclusion:<br>• “anxiety” |
| Cancer | Billing<br>Health Condition<br>Encounter Diagnosis | ICD-9 Codes:<br>• 140-149: malignant neoplasm of lip, oral cavity, and pharynx<br>• 150-159: malignant neoplasm of digestive organs and peritoneum<br>• 160-169: malignant neoplasm of respiratory and intrathoracic organs<br>• 170-175: malignant neoplasm of bone, |

| | | |
|---|---|---|
| | | connective tissue, skin, and breast<br>• 176: Kaposi's sarcoma<br>• 179-189: malignant neoplasm of genitourinary organs<br>• 190-199: malignant neoplasm of other and unspecified sites<br>• 200-208: malignant neoplasm of lymphatic and hematopoietic tissue<br>• 209: neuroendocrine tumours<br>• 239: neoplasms of unspecified nature |
| | Health Condition | Inclusion:<br>• "cancer"<br>• "neoplasm"<br>Exclusion:<br>• "fam" |
| | Medication | Prescription of chemotherapy drugs:<br>• Mechlorethamine (nitrogen mustard, Mustargen)<br>• Melphalan (Alkeran, L-PAM)<br>• Chlorambucil (Leukeran)<br>• Cyclophosphamide (Cytoxan, Procytox)<br>• Ifosfamide (Ifex)<br>• Estramustine (Emcyt)<br>busulfan (Myleran, Busulfex)<br>• Dacarbazine (DTIC)<br>• Temozolomide (Temodal)<br>• Carmustine (BiCNU, BCNU)<br>• Lomustine (CeeNU, CCNU)<br>• Streptozocin (Zanosar)<br>• Cisplatin (Platinol AQ, Platinol)<br>• Carboplatin (Paraplatin, Paraplatin AQ)<br>• Oxaliplatin (Eloxatin)<br>• Thiotepa (ThioTEPA)<br>• Methotrexate<br>• Raltitrexed (Tomudex)<br>• Pemetrexed (Alimta)<br>• Cladribine (Leustatin)<br>• Fludarabine (Fludara)<br>• Mercaptopurine (Purinethol, 6-MP)<br>• Thioguanine (Lanvis, 6-TG)<br>• Azactidine (Vidaza)<br>• Capecitabine (Xeloda)<br>• Cytarabine (Cytosar, Ara-C) |

| | | |
|---|---|---|
| | | • 5-fluorouracil (Adrucil, 5-FU, Efudex [topical])<br>• Gemcitabine (Gemzar)<br>• Bleomycin (Blenoxane)<br>• Dactinomycin (Cosmegen, actinomycin-D)<br>• Daunorubicin (Cerubidine, daunomycin)<br>• Doxorubicin (Adriamycin)<br>• Epirubicin (Pharmorubicin)<br>• Idarubicin (Idamycin)<br>• Mitomycin (Mutamycin)<br>• Mitoxantrone (Novantrone)<br>• Liposomal daunorubicin (DaunoXome)<br>• Liposomal doxorubicin (Myocet)<br>• Pegylated liposomal doxorubicin (Caelyx)<br>• Asparaginase (Kidrolase)<br>• Docetaxel (Taxotere)<br>• Paclitaxel (Taxol)<br>• Vinblastine (Velbe)<br>• Vincristine (Oncovin)<br>• Vinorelbine (Navelbine)<br>• Vindesine (Eldesine)<br>• Irinotecan (Camptosar)<br>• Topotecan (Hycamtin)<br>• Etoposide (Vepesid, VP-16)<br>• Teniposide (Vumon, VM-26)<br>• Hydroxyurea (Hydrea)<br>• Octreotide (Sandostatin, Sandostatin LAR)<br>• Mitotane (Lysodren)<br>• Procarbazine hydrochloride (Matulane)<br>• Arsenic trioxide<br>• Pofimer sodium (Photofrin)<br>• Altretamine (Hexalen, Hexastat) |
| Cardiovascular Disease | Billing<br>Health Condition<br>Encounter Diagnosis | ICD-9 Codes:<br>• 410-414: ischemic heart disease<br>• 415-417: diseases of pulmonary circulation<br>• 420-429: other forms of heart disease |
| | Health Condition<br>Encounter Diagnosis | Inclusion:<br>• “cardiovascular disease”<br>• “CVD” |

| | | |
|---|---|---|
| | | • "coronary artery disease"<br>• "CAD"<br>• "heart attack"<br>• "myocardial infarction"<br>• "heart disease"<br>Exclusion:<br>• "fam" |
| | Medication | Prescription of anticoagulant medications:<br>• Rivaroxaban (Xarelto)<br>• Dabigatran (Pradaxa)<br>• Apixaban (Eliquis)<br>• Heparin (various)<br>• Warfarin (Coumadin)<br>Prescription of antiplatelet agents:<br>• Clopidogrel (Plavix)<br>• Dipyridamole<br>• Prasugrel (Effient)<br>• Ticagrelor (Brilinta) |
| Diabetes | Disease Case* | Diabetes |
| COPD | Disease Case* | COPD |
| Rheumatoid Arthritis | Billing<br>Health Condition<br>Encounter Diagnosis | ICD-9 Code:<br>• 714: rheumatoid arthritis and other inflammatory polyarthropathies |
| | Health Condition<br>Encounter Diagnosis | Inclusion:<br>• "rheumatoid arthritis"<br>Exclusion:<br>• "fam" |
| Hypertension | Disease Case* | Hypertension |
| Lipid Disorder | Billing<br>Health Condition<br>Encounter Diagnosis | ICD-9 Code:<br>• 272: disorders of lipid metabolism |
| | Health Condition | Inclusion:<br>• "lipid"<br>• "cholesterol"<br>Exclusion:<br>• "fam" |
| | Lab | LDL measurement: 3.37-9 mmol/L |
| | Medications | Inclusion:<br>• "statin"<br>Exclusion:<br>• "nystatin" |
| Bipolar Affective Disorder | Billing<br>Health Condition<br>Encounter Diagnosis | ICD-9:<br>• 296.4: bipolar affective disorder, manic<br>• 296.5: bipolar affective disorder, |

| | | |
|---|---|---|
| | | depressed<br>• 296.6: bipolar affective disorder, mixed<br>• 296.7: bipolar affective disorder, unspecified |
| | Health Condition<br>Encounter Diagnosis | Inclusion:<br>• "bipolar"<br>Exclusion:<br>• "fam" |
| Chronic Kidney Disease | Billing<br>Health Condition<br>Encounter Diagnosis | ICD-9:<br>• 585: chronic renal failure |
| | Health Condition<br>Encounter Diagnosis | Inclusion:<br>• "chronic kidney disease"<br>• "CKD"<br>• "chronic renal failure"<br>Exclusion:<br>• "fam" |
| | Lab | Occurrence of the following laboratory results:<br>• Estimated glomerular filtration rate (eGFR) less than 60 mL/min/1.73 m$^2$ (260)<br>• Serum creatinine greater than 120 µmol/L for men or 90 µmol/L for women (261)<br>• Urine albumin/creatinine ratio greater than 20 mg/mmol for men or 28 mg/mmol for women (84)<br>• Serum albumin greater than 300 mg/L (260) |
| Tricyclic Antidepressant (TCA) use | Medication | Prescription of:<br>• Amitriptyline<br>• Amoxapine<br>• Desipramine (Norpramin)<br>• Doxepin<br>• Imipramine (Tofranil)<br>• Nortriptyline (Pamelor)<br>• Protriptyline (Vivactil)<br>• Trimipramine (Surmontil) |
| Osteoporosis | Billing<br>Health Condition<br>Encounter Diagnosis | ICD-9 Code:<br>• 733: Osteoporosis and other bone disorders |
| | Health Condition<br>Encounter Diagnosis | Inclusion:<br>• "osteoporosis"<br>Exclusion: |

| | | |
|---|---|---|
| | | • “fam” |
| | Medications | Prescription of:<br>• Alendronic acid<br>• Risedronic acid<br>• Ibandronic acid |
| Leg Injury | Billing<br>Health Condition<br>Encounter Diagnosis | ICD-9 Codes:<br>• 820-29: fracture of lower limb<br>• 843: sprain or strain of hip and thigh<br>• 844: sprain or strain of knee and leg<br>• 928: crushing injury to lower limb |
| BMI | Exam | Based on:<br>• BMI (kg/$m^2$) as recorded in EMR<br>• Height (m) and weight (kg) as recorded in the EMR on the same date |
| Family History of Osteoarthritis | Family History | Inclusion:<br>• “osteoarthritis”<br>Exclusion:<br>• “no” |
| Family History of Diabetes | Family History | Inclusion:<br>• “diabet”<br>Exclusion:<br>• “no” |
| Family History of Hypertension | Family History | Inclusion:<br>• “hypertens”<br>Exclusion:<br>“no” |
| Family History of Depression | Family History | Inclusion:<br>• “depress”<br>Exclusion:<br>“no” |
| Stroke | Billing<br>Health Condition<br>Encounter Diagnosis | ICD-9 Codes:<br>• 430: Subarachnoid hemorrhage<br>• 431: Intracerebral hemorrhage<br>• 432: Other and unspecified intracranial hemorrhage<br>• 434: Occlusion of cerebral arteries |
| | Health Condition<br>Encounter Diagnosis | Inclusion:<br>• “stroke”<br>Exclusion:<br>• “fam” |
| Asthma | Billing<br>Health Condition<br>Encounter Diagnosis | ICD-9 Code:<br>• 493: Asthma |

|  |  |  |
|---|---|---|
|  | Health Condition<br>Encounter Diagnosis | Inclusion:<br>• “asthma”<br>Exclusion:<br>• “fam” |

# Curriculum Vitae

|  |  |
|---|---|
| **Name:** | Jason Black |
| **Post-secondary Education and Degrees:** | Western University<br>London, Ontario, Canada<br>2012-2016 Bachelor of Medical Sciences |
| **Honours and Awards:** | NSERC Undergraduate Student Research Award<br>2012 |
| **Related Work Experience** | Research Analyst<br>Western University<br>2017-2018<br><br>Teaching Assistant |

Western University
2018

Research Assistant
Western University
2016

**Publications:**

Alexandria Ratzki-Leewing, Stewart Harris, Selam Mequanint, Natalie H. Au, **Jason E. Black**, Sonja Reichert, Judith B. Brown, Bridget L. Ryan. Severe Hypoglycemia Rates Highest Among Those with Suboptimal Reporting Behaviour: Results of the InHypo- DM Study. ADA, 2018. Accepted. Abstract.

Alexandria Ratzki-Leewing, Stewart Harris, Selam Mequanint, Natalie H. Au, **Jason E. Black**, Sonja Reichert, Judith B. Brown, Bridget L. Ryan. Real-world risk indicators of severe hypoglycemia in TD: Results of the InHypo-DM Study. ADA, 2018. Accepted. Abstract.

Alexandria Ratzki-Leewing, Stewart Harris, Selam Mequanint, Sonja M. Reichert, Judith Belle Brown, **Jason E. Black**, Bridget L. Ryan. The real-world crude incidence of hypo- glycemia in adults with diabetes: Results of the InHypo-DM study, Canada. BMJ Open Diabetes Research and Care, 2017 In Press.

**Jason E. Black**, Amanda L. Terry, Daniel J. Lizotte. FRAMR-EMR: Framework for Prognostic Predictive Model Development Using Electronic Medical Record Data with a Case Study in Osteoarthritis Risk. BMC Prognostic and Diagnostic Research, 2017. Under Revision.