International Journal of Population Data Science

Journal Website: www.ijpds.org

Development of a prognostic prediction model to estimate the risk of multiple chronic diseases: constructing a copula-based model using Canadian primary care electronic medical record data

Jason E. Black[1,*], Jacqueline K. Kueper[2], Amanda L. Terry[3], and Daniel J. Lizotte[2]

Submission History

Submitted:	28/08/2020
Accepted:	30/10/2020
Published:	19/01/2021

[1]Department of Epidemiology and Biostatistics, Western University, 1151 Richmond Street London, Ontario, Canada, N6A 3K7
[2]Department of Computer Science, Department of Epidemiology and Biostatistics, Western University, 1151 Richmond Street London, Ontario, Canada, N6A 3K7
[3]Department of Family Medicine, Department of Epidemiology and Biostatistics, Western University, 1151 Richmond Street London, Ontario, Canada, N6A 3K7

Abstract

Introduction

The ability to estimate risk of multimorbidity will provide valuable information to patients and primary care practitioners in their preventative efforts. Current methods for prognostic prediction modelling are insufficient for the estimation of risk for multiple outcomes, as they do not properly capture the dependence that exists between outcomes.

Objectives

We developed a multivariate prognostic prediction model for the 5-year risk of diabetes, hypertension, and osteoarthritis that quantifies and accounts for the dependence between each disease using a copula-based model.

Methods

We used data from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) from 2009 onwards, a collection of electronic medical records submitted by participating primary care practitioners across Canada. We identified patients 18 years and older without all three outcome diseases and observed any incident diabetes, osteoarthritis, or hypertension within 5-years, resulting in a large retrospective cohort for model development and internal validation (n=425,228). First, we quantified the dependence between outcomes using unadjusted and adjusted ϕ coefficients. We then estimated a copula-based model to quantify the non-linear dependence between outcomes that can be used to derive risk estimates for each outcome, accounting for the observed dependence. Copula-based models are defined by univariate models for each outcome and a dependence function, specified by the parameter θ. Logistic regression was used for the univariate models and the Frank copula was selected as the dependence function.

Results

All outcome pairs demonstrated statistically significant dependence that was reduced after adjusting for covariates. The copula-based model yielded statistically significant θ parameters in agreement with the adjusted and unadjusted ϕ coefficients. Our copula-based model can effectively be used to estimate trivariate probabilities.

Discussion

Quantitative estimates of multimorbidity risk inform discussions between patients and their primary care practitioners around prevention in an effort to reduce the incidence of multimorbidity.

Keywords

prognostic prediction model; risk estimation; multimorbidity; electronic medical records; CPCSSN; copula; multivariate; diabetes; osteoarthritis; hypertension; primary care

*Corresponding Author:
Email Address: jblack85@uwo.ca (Jason E. Black)

https://doi.org/10.23889/ijpds.v6i1.1395

Introduction

Harnessing observational health data to improve patient care, such as through decision support tools embedded into electronic medical records (EMRs), is a topic of great interest [1, 2]. Prognostic prediction models can provide decision support through quantitative estimates of disease risk based on a patient's individual predictors (e.g., age, sex, physical activity level) [3–5]. Understanding a patient's risk of disease empowers prevention efforts, a hallmark of population health, by guiding decision-making processes and identifying patients at increased risk [6]. Research related to decision support at the point of care requires both methodological and clinical considerations. Methodological considerations span from data source selection and pre-processing to model development and evaluation; clinical considerations include identifying what disease(s) or aspects of clinical care could benefit from decision support and the types of information or tools that will accomplish this.

There is a gap between one of the most prominent clinical challenges faced by primary care practitioners and their patients and the development of prognostic prediction models thus far. Multimorbidity, where a patient has two or more chronic diseases, is increasing in prevalence and presents several challenges in terms of identification and treatment [7, 8]. The ability to estimate a patient's risk of multimorbidity is needed [7, 9]. Research into multimorbidity has predominantly focused on establishing patterns or clusters of multimorbidity or establishing risk factors by investigating the associations between multimorbidity and potential risk factors [10, 11]. While related to multimorbidity risk, the latter does not allow for risk estimation. Recently, there has been a focus on developing strategies to prevent multimorbidity as health policy makers and health care practitioners recognize its importance [12, 13]. There are many existing prognostic prediction models for individual diseases but few for multimorbidity [14, 15]. Using a series of single-disease models in a clinical setting to estimate risk of multiple diseases is not only burdensome but also may give inaccurate perceptions of risk.

Methodological complexity may be a barrier to developing tools for multimorbidity risk prediction; standard off-the-shelf packages for developing prediction models are not expected to perform correctly. Prognostic prediction models are commonly developed to estimate the risk of a single disease. To estimate the risk of multimorbidity, one might combine the risks of multiple single disease models. For example, if one were interested in estimating a patient's risk of diabetes and hypertension co-occurring, they might multiply the patient's risk of diabetes by their risk of hypertension, giving the risk of both diseases occurring. However, this method assumes independence between the incidence of diseases, which rarely occurs. Instead, this dependence must be accounted for when estimating the risk of multiple diseases. We hypothesize that a lack of clear methodology for how to account for dependence between disease incidence is a barrier to the development of prognostic prediction models for multimorbidity and targeting this methodological gap is a necessary first step towards proper estimation of multimorbidity risk.

The objective of this study is to present a methodology for prognostic prediction models that accounts for dependence between disease incidence. This is achieved in the context of Canadian primary health care, whereby we developed a prognostic prediction model that estimates the 5-year risk of diabetes, hypertension, and osteoarthritis. These diseases were selected as a case study based on their prevalence, availability of validated case-detecting algorithms [16], and clinical importance. To accomplish this, we first developed univariate multivariable models for each disease. We then explored the dependence between disease incidence, which led to the development of a model capable of predicting each disease and their co-occurrence while accounting for the dependence between disease incidence.

Methods

Data source

Primary care is typically the first contact for patients within the Canadian healthcare system. A patient is managed in primary care by their primary care practitioner or referred to secondary or tertiary care, depending on the level of care required [17]. Primary care is an ideal setting for the deployment of interventions aimed at reducing multimorbidity risk given the broad population it serves who typically are in earlier stages of disease compared to patients of secondary or tertiary care.

All data used were derived from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) database [18]: a database containing patient information from the EMRs of primary care practices across Canada starting in 2008 [19]. Nearly 1,200 primary care practitioners voluntarily contribute deidentified records of more than 1.5 million patients. Patients provide consent via an opt-out system, where patients who do not wish to contribute their data may choose to opt-out, except in Quebec, where an opt-in process is mandated by provincial law. In 2013, CPCSSN patients were older and more likely to be female compared to the overall Canadian population as reported in census data [20], which is typical of primary care [21–23].

All structured data from the EMR are available in CPCSSN, including patient demographics, diagnoses, laboratory results, prescriptions, referrals, risk factors, medical procedures, vaccinations, and allergies. For privacy reasons, the free-text narrative where primary care practitioners record their notes is not available in CPCSSN.

Measures

Outcome

CPCSSN researchers developed and validated case-detecting algorithms for several chronic diseases to identify cases of disease within the database [16]. These case detecting algorithms were developed using published evidence and input from primary care and specialist physicians and validated by a comprehensive chart review. Validation demonstrated high sensitivity and specificity.

We used CPCSSN case-detecting algorithms to identify cases of diabetes, osteoarthritis, and hypertension. Sensitivity and specificity for these case-detecting algorithms were high; see Appendix Table 1. Our use of validated disease case-detecting algorithms helps ensure that the identification of outcomes is accurate. Inaccurate outcome identification (a form of measurement error) will decrease the accuracy of risk estimates due to biased relationships between the predictors and true disease development. This poor performance would not be revealed by internal validation as the data used for validation would be subject to the same issue of inaccuracy in outcome identification as the data used to construct the model. Often only internal validation is feasible, reinforcing the importance of using a validated case-detecting algorithm for the identification of outcomes.

Blinding of predictor information during outcome assessment was not possible. Predictor assessment and subsequent outcome assessment were both conducted by the primary care practitioner; thus, it is likely that primary care practitioners had some knowledge of the patient's predictors while assessing the outcomes, which may have introduced measurement bias. However, each outcome has clearly defined diagnostic criteria; thus, the impact of this is likely minimal.

Predictors

We identified predictors for each outcome through review of relevant literature. These predictors are presented in Appendix Table 2. We then attempted to identify predictors in the CPCSSN database. We identified 5 predictors of osteoarthritis, 8 of diabetes, and 6 of hypertension; see Table 1. Where possible, we used CPCSSN validated case-detecting algorithms. Otherwise, we developed an algorithm to identify each predictor using CPCSSN data: some combination of diagnostic terms and codes; medications used for specific indications; and laboratory results. These algorithms were reviewed by a primary care practitioner to ensure accuracy. See Supplemental Appendix 1 for predictor case-detecting algorithms.

We estimated each patient's income by linking their Forward Sortation Area (FSA) to area-level income data collected by the National Household Survey conducted in 2011 [24]. Rurality was assessed based on the second digit of the FSA.

As suggested in TRIPOD [25], we included all continuous risk factors in their original form. We did not transform or categorize continuous variables.

Participants

We included all patients aged 18 or older who did not have diabetes, osteoarthritis, and hypertension at baseline (i.e., we excluded patients with all 3 outcomes) and had some interaction with their primary care practitioner in 2009 or 2010 (i.e., an interaction that resulted in a billing occurrence, encounter recording or diagnosis, exam, or health condition diagnosis in the EMR). For each patient, we considered the first interaction with their primary care practitioner between 1 January 2009 and 31 December 2010 the patient's unique start-date. We assessed the patient's predictors at this point (including diabetes, hypertension, and osteoarthritis as one may predict another). We then noted any diagnosis of diabetes, osteoarthritis, or hypertension over the following 5 years. We included all eligible patients to maximize predictive performance.

Missing data

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. Multiple imputation was used to address missing data, which produced 5 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 [26] to create a single statistic whose variance has been adjusted to account for the uncertainty of deriving an estimate from multiple datasets.

Statistical analysis

To construct a prognostic prediction model for diabetes, hypertension, and osteoarthritis, we analyzed the dependence between these diseases. We selected copulas [27, 28] to model the dependence between outcomes because they account for more than two diseases, adjust for both continuous and discrete variables, and can be used to construct a prognostic prediction model. First, we constructed univariate models for each outcome then we used a copula to describe the dependence between outcomes.

Univariate multivariable logistic regression

We constructed univariate multivariable logistic regression models for each outcome. We included patients without the outcome at baseline when estimating the univariate model. For example, we used a subgroup of patients who did not have diabetes at baseline to construct the diabetes univariate model. We internally validated each univariate model by measuring its discrimination and calibration. We assessed

Table 1: Predictors available in CPCSSN database

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

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

$$\dot{C}_\theta\left(u,v\right)=\frac{\dfrac{e^\theta\theta\left(\left(u-1\right)\left(-e^{\theta v}\right)-e^{\theta\left(u+v\right)}+ue^{\theta v+\theta}-\left(v-1\right)e^{\theta u}+ve^{\theta u+\theta}-e^\theta\left(u+v\right)+u+v-1\right)}{\left(e^\theta-1\right)\left(-e^{\theta\left(u+v\right)}+e^{\theta u+\theta}+e^{\theta v+\theta}-e^\theta\right)}+\log\left(\dfrac{\left(e^{-\theta u}-1\right)\left(e^{-\theta v}-1\right)}{e^{-\theta}-1}+1\right)}{\theta^2} \tag{2}$$

$$\overline{\pi}_{ik}=1-\pi_{ik} \tag{3}$$

$$\overline{\pi}_{il}=1-\pi_{il} \tag{4}$$

where C_θ is the copula function; $\dot{C}_\theta$ is the derivative of the copula function; π_{ik} and π_{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.

discrimination (the ability to assign higher risk to true positive cases) by determining the area under the receiver operator characteristic curve (AUC). We assessed calibration (how well the model fits the data) by examining calibration plots. To investigate the potential impact of censoring (e.g., a patient changing providers), we conducted a sensitivity analysis where we required that each patient have at least one interaction with their primary care practitioner after the end of their follow-up period. We compared parameter estimates from this restricted cohort to those of the overall cohort.

Analysis of dependence

We explored the dependence between outcomes in a pairwise fashion. For each pairwise analysis, we included patients who did not have either outcome at baseline. For example, in the analysis of diabetes and hypertension, we included patients who did not have diabetes or hypertension.

We estimated the unadjusted pairwise correlation between outcomes using the ϕ coefficient (also known as the mean square contingency coefficient). The ϕ coefficient is a measure of association between two binary variables, analogous to the Pearson correlation coefficient for continuous variables [29]. In fact, estimating a Pearson correlation coefficient for two binary variables gives the ϕ coefficient [29].

We then estimated the adjusted pairwise correlation (also known as partial correlation) between outcomes using the ϕ coefficient adjusted for the predictors of both outcomes.

To enable predictions that account for the dependence between outcomes, we estimated a copula-based model that captures the dependence between each outcome pair. Copula models are able to capture dependence among variables without imposing any requirements on marginal distributions of the variables; for example, the marginal distributions do not need to be Gaussian. Many parametric copula forms exist that are characterized by the structure of the dependence they can best describe. We selected the Frank copula [30] based on its ability to describe weak dependence based on the weak correlations we observed between outcome pairs.

$$C_\theta\left(u,v\right)=-\frac{1}{\theta}\ln\left(1+\frac{\left(e^{-\theta u}-1\right)\left(e^{-\theta v}-1\right)}{e^{-\theta}-1}\right) \tag{5}$$

When modelling the dependence between binary variables, the copula is defined by both the parameter θ and the marginal distributions [27]. As such, we used the two-stage estimation procedure based on the composite likelihood suggested by Zhao and Joe [31] for the estimation of θ. First, we determined the marginal models using the maximum likelihood estimation procedure, yielding β estimates that we used in the second step. From these univariate models, we estimated the probabilities for the independent occurrence of each outcome (π_j), by:

$$\pi_j\left(\mathbf{x}\right)=\frac{\exp\left(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_j\right)}{1+\exp\left(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_j\right)} \tag{6}$$

where $\boldsymbol{\beta}_j$ is a vector containing the β estimates for each outcome j and $\mathbf{x}$ is a matrix of covariate data. Second, we obtained estimates of θ, 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_θ [Equation 1]) equal to zero, we estimated θ.

A dependence structure using copulas is completely specified by its univariate models and copula, which is specified by its θ estimate. For each disease pair, we estimated the parameter θ and bootstrapped confidence intervals using the percentile method [32] and 1,000 replicates. Additionally, we tested the null hypothesis that the observed outcome frequencies are no different than what would be expected under independence [27] using the following hypothesis test based on the score test. We rejected the null hypothesis if z_{obs} is larger in absolute value than a critical value derived from the standard Normal distribution, denoted $N(0,\,1)$.

$$z_{obs}=\sum_{i=1}^{n}\frac{\dot{C}_{\theta_{\mathbf{o}}}\left(\widehat{\overline{\pi}}_{ik},\widehat{\overline{\pi}}_{il}\right)\left(Y_{ik}-\hat{\pi}_{ik}\right)\left(Y_{il}-\hat{\pi}_{il}\right)}{\hat{\pi}_{ik}\hat{\pi}_{il}\widehat{\overline{\pi}}_{ik}\widehat{\overline{\pi}}_{il}}\Bigg/\sqrt{\sum_{i=1}^{n}\frac{\dot{C}^2_{\theta_{\mathbf{o}}}\left(\widehat{\overline{\pi}}_{ik}\widehat{\overline{\pi}}_{il}\right)}{\hat{\pi}_{ik}\hat{\pi}_{il}\widehat{\overline{\pi}}_{ik}\widehat{\overline{\pi}}_{il}}} \tag{7}$$

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.

$$p\left(x_1, x_2, x_3\right) = p\left(X_1 = x_1, X_2 = x_2, X_3 = x_3\right) \qquad (8)$$

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

$$p\left(x_1, x_2\right) = \sum_{x_3 \in \{0,1\}} p\left(x_1, x_2, x_3\right) \qquad (9)$$

Similar expressions are true for $p\left(x_1, x_3\right)$ and $p\left(x_2, x_3\right)$. Based on $\hat{p}\left(x_1, x_2\right)$, $\hat{p}\left(x_1, x_3\right)$, and $\hat{p}\left(x_2, x_3\right)$ as estimated by the copula model, trivariate probability mass functions $\left(\hat{p}\left(x_1, x_2, x_3\right)\right)$ can be found such that their bivariate distributions match the specified bivariate marginal distributions. In fact, there may be many trivariate probability mass functions whose bivariate marginals match the specified bivariate distributions. We chose the trivariate with the highest entropy (highest uncertainty), as this gives the most conservative estimate in terms of the model's predictions.

To find the trivariate distribution with maximum entropy, we first note that it is possible to define the space of all trivariate probability mass functions that satisfy the bivariate constraints using a single parameter, α (see Supplementary Appendix 2). Therefore, to find the distribution with maximum entropy, we first determined the permitted bounds of α, such that all estimated probabilities fall in the range 0 to 1, and we then defined the entropy of a potential solution distribution as a function of α [33]. By searching over possible α, we found the distribution that maximizes entropy. Given the resulting trivariate distribution, all joint probabilities of disease incidence can be estimated. I.e., the risk of developing any combination of diabetes, hypertension, and osteoarthritis all within a 5-year window can be estimated.

We have included code to estimate the copula-based model, perform hypothesis testing, and estimate trivariate probabilities using the copula-based model, see Supplementary Appendix 3.

Results

Descriptive statistics

We followed a cohort of 425,228 adult patients who did not have multimorbid diabetes, hypertension, and osteoarthritis (i.e., they had at most two of these three conditions) who had received care between 1 January 2009 and 31 December 2010 for 5 years. Figure 1 details the flow of patients into the cohort.

At baseline, the majority of patients were female (58%) and had a body mass index (BMI) greater than 25 kg/m^2 (64%) with a median age of 49 years old (interquartile range: 34 to 59). For a detailed description of all patient characteristics, see Appendix Table 3.

After 5 years, hypertension was the most commonly acquired outcome (n=39,882; incidence proportion of 9.4%), followed by diabetes (n=18,769; 4.4%), then osteoarthritis (n=12,803; 3.0%).

Predictors

For BMI, the most recent value before baseline was used. For each predictor found in the CPCSSN database, we assessed its face validity by comparing its prevalence in CPCSSN during 2009 and 2010 with national averages from 2010 (data not shown). Polycystic ovarian syndrome and alcohol use disorder were much lower than national averages; we did not include these predictors in our analysis. Additionally, family history data was not collected in several networks; thus, we did not include family history in our analysis.

The following predictors were missing to some degree: smoking information, sex, BMI, age, and income (Table 2). We used multiple imputation by chained equations to account for missing data in sex, BMI, age, and income. We did not impute smoking information due to its high degree of missingness.

Many patients were missing BMI values. We could not determine the reason why patients were missing BMI values; however, the distribution among patients with BMI values was approximately similar to that of the Canadian population (Appendix Table 4). We examined the kernel density distribution of imputed BMI values compared to known BMI values: all imputed BMI values were within a reasonable range of values (Appendix Figure 1).

Figure 1: Cohort based on CPCSSN database

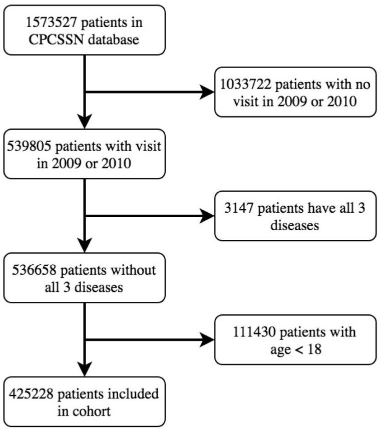

1573527 patients in CPCSSN database
1033722 patients with no visit in 2009 or 2010
539805 patients with visit in 2009 or 2010
3147 patients have all 3 diseases
536658 patients without all 3 diseases
111430 patients with age < 18
425228 patients included in cohort

Table 2: Predictors with missing data

	Development set ($n = 265,228$)		**Validation set** ($n = 160,000$)	
	n missing	%	*n* missing	%
Smoking	247,918	93%	149,401	93%
Sex	44	0.02%	25	0.02%
BMI	175,632	66%	105,768	66%
Age	167	0.06%	92	0.06%
Income	13,824	5.2%	8,579	5.4%

BMI: body mass index

Table 3: Univariate logistic regression models

	Reference category/units	β estimate	95% CI	Odds ratio	95% CI
Diabetes univariate model (AUC = 0.85)					
Hypertension	No	Reference		Reference	
	Yes	0.3	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	($\mathrm{kg/m^2}$)	0.07	0.07 to 0.08	1.07	1.07 to 1.08
Sex	Male	Reference		Reference	
	Female	-0.3	-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	($10,000)	-0.89	-1.15 to -0.64	0.41	0.32 to 0.53
Hypertension univariate model (AUC = 0.84)					
Diabetes	No	Reference		Reference	
	Yes	0.18	0.12 to 0.23	1.19	1.13 to 1.26
Age	(Years)	0.07	0.06 to 0.07	1.07	1.06 to 1.07
BMI	($\mathrm{kg/m^2}$)	0.06	0.06 to 0.07	1.06	1.06 to 1.07
Chronic Kidney Disease	No	Reference		Reference	
	Yes	0.8	0.74 to 0.85	2.22	2.09 to 2.35
Tricyclic Antidepressant Use	No	Reference		Reference	
	Yes	0.55	0.49 to 0.62	1.74	1.63 to 1.86
Osteoarthritis univariate model (AUC = 0.83)					
Age	(Years)	0.06	0.05 to 0.06	1.06	1.05 to 1.06
Sex	Male	Reference		Reference	
	Female	0.22	0.17 to 0.27	1.25	1.19 to 1.31
BMI	($\mathrm{kg/m^2}$)	0.04	0.03 to 0.04	1.04	1.04 to 1.05
Previous Leg Injury	No	Reference		Reference	
	Yes	1.6	1.52 to 1.68	4.94	4.57 to 5.35
Osteoporosis	No	Reference		Reference	
	Yes	0.9	0.83 to 0.98	2.47	2.29 to 2.66

AUC: area under the receiver operator characteristic curve; BMI: body mass index;
CI: confidence interval.

Univariate results

Univariate results are displayed in Table 3. We found that all predictors were associated with the corresponding outcome. Each univariate model displayed strong discrimination and moderate calibration (see Appendix Figures 2a-c for calibration plots). Sensitivity analyses revealed that censoring was not a concern: model estimates based on a cohort restricted to patients with at least one interaction with their primary care practitioner after the follow-up period (n=315,859) were similar to those of the overall cohort (Appendix Table 5).

Dependence analysis

We estimate the unadjusted and adjusted correlation between each outcome pair (Table 4a). All pairs were positively correlated. Diabetes and hypertension displayed the highest correlation, followed by hypertension and osteoarthritis, then

Table 4a: Unadjusted correlation (ϕ coefficients)

	Diabetes	Hypertension	Osteoarthritis
Diabetes	1		
Hypertension	0.240 (0.238 to 0.246, p < 0.0001)	1	
Osteoarthritis	0.098 (0.093 to 0.102, p < 0.0001)	0.209 (0.205 to 0.213, p < 0.0001)	1

Table 4b: Adjusted correlation (partial correlation)

	Diabetes	Hypertension	Osteoarthritis
Diabetes	1		
Hypertension	0.132 (0.128 to 0.137, p < 0.0001)	1 1	
Osteoarthritis	0.038 (0.034 to 0.042, p < 0.0001)	0.123 (0.118 to 0.127, p < 0.0001)	1

Table 4c: Adjusted dependence (θ estimates)

	Diabetes	Hypertension	Osteoarthritis
Diabetes			
Hypertension	1.677 (1.566 to 1.788, p < 0.0001)		
Osteoarthritis	0.683 (0.526 to 0.841, p < 0.0001)	1.949 (1.822 to 2.076, p < 0.0001)	

Table 5: Trivariate probabilities for simulated patient

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

Simulated patient: 79 year-old woman whose BMI is 34 kg/m^2 with an income of roughly $35,000 and free of any other risk factors.

diabetes and osteoarthritis. This was consistent after adjusting for predictors, though smaller in magnitude (Table 4b).

We estimated copulas for each outcome pair (Table 4c). Hypothesis testing demonstrated a significant positive dependence between all outcome pairs after adjusting for risk factors.

To demonstrate the use of our model, we estimated the trivariate probabilities for a simulated patient accounting for the dependence between outcomes using the copula model and without accounting for the dependence between outcomes by multiplying the probability from each univariate model (Table 5). Risk estimates differed between these approaches, demonstrating the need to account for the dependence between outcomes.

Discussion

We developed and internally validated univariate models for diabetes, hypertension, and osteoarthritis based on EMR records. All models were highly discriminative and moderately calibrated. We then explored the dependence between each outcome by estimating the unadjusted and adjusted correlation in a pairwise fashion. All outcome pairs were positively correlated. After adjusting for predictors, outcome pairs remained positively correlated with reduced magnitudes. Finally, we estimated a copula-based model that describes the dependence between outcomes while enabling risk predictions.

Existing research for multimorbidity risk prediction includes four areas. First, establishing risk factors for multimorbidity that can be used to identify high-risk patients. Many studies have found that older age, female gender, and lower socioeconomic status are associated with multimorbidity [34]. While our model was constructed for prediction purposes, rather than to derive causal inferences, these factors were all included in our model and found to be predictive of diabetes, hypertension, or osteoarthritis.

Second, two prognostic prediction models have been developed for the onset of the first of several possible chronic disease outcomes. Ng et al. (2020) developed a model from national survey data linked with provincial health administrative data in Canada that is primarily intended for population-level predictions to aid health policy makers [35]; May et al. (2019) developed a model with primary care clinical data in the United States that is intended for implementation with EMRs for individual patient-level predictions [36]. These models require the absence of all possible outcome diseases at baseline (6 and 10, respectively) and predict the first instance of any of the diseases, which may signal the beginning of progression towards multimorbidity. In contrast, our methods account for dependence between diseases and allow for the presence of some of the outcome conditions at baseline such that predictions may be made for individuals who are further along in the natural history of diseases.

A third line of research relevant to multimorbidity prediction includes Bayesian networks rather than regression-based models [37–39]. Lappenshaar et al. developed multilevel Bayesian network methodology and applied it to explore cardiovascular multimorbidity from primary care data in the Netherlands [38]. Their methodology allows for predicting multiple outcomes, explicit modelling of interactions and dependence between variables, formally incorporating domain knowledge, and accounting for practice-level variation which is commonly present in large health databases. In contrast to our regression-based methods that estimate conditional probability distributions with predictions based on all variables in a parametric model, multilevel Bayesian networks model a joint probability distribution and make predictions based on variables in the Markov blanket of the outcome(s) of interest. Lappenshaar et al. extended their models to include changes over time through multilevel temporal Bayesian networks [37]. While this methodology has the potential for individual risk prediction, it has not been evaluated in that setting; the main focus of the work thus far was to understand interactions between diseases and the progression of multimorbidity over time and to predict of future rates of multimorbidity at a group level.

Finally, Wang et al. (2014) developed a multitask machine learning framework for EMR-based multiple disease prediction [40]. Their framework includes learning groupings of common risk factors across the outcome diseases, which serve as high-level latent predictors to use instead of raw EMR features and learning regression coefficients to weight these groupings. The resulting model can be used both for risk prediction and to explore the groupings to identify potential shared or unique risk factors across outcome diseases. A case study with chronic obstructive pulmonary disease and congestive heart failure found the multitask learning framework had better AUCs than a single-outcome dimensionality reduction approach (Principal Component Analysis) and similar performance to a logistic regression-based approach.

Strengths and limitations

Our analysis was limited by the availability of risk factor information within the EMRs. Data such as behavioral or environmental factors are not typically collected during a clinical encounter, thus not stored in the EMR. As such, the univariate models likely underestimated risk among patients who possess the unavailable risk factor. For the dependence analysis, the observed dependence might have been influenced by an unavailable risk factor that could not be adjusted for. Such a factor could act in either direction; a risk factor could increase or decrease the observed dependence between the outcomes, thus the true dependence could be less than or greater than what we observed.

Our analysis may be subject to bias introduced by patterns in physicians' diagnosis of diabetes, hypertension, and osteoarthritis. For example, when a physician diagnoses a patient with diabetes, they will likely assess for related conditions that may have otherwise gone undetected, such as hypertension. This may explain some of the dependence that we observed between disease pairs.

The CPCSSN case definition we used to identify patients with diabetes identifies both type 1 and type 2 diabetes but does not distinguish between the two. However, type 1 cases typically constitute the minority of diabetes cases (10%, [41]) and are more commonly diagnosed in children [42]; thus, incident cases of diabetes that we observed in adults were more likely type 2. Risk factor information for type 1 diabetes (i.e., genetic factors [43]) were not available; however, risk factors for type 2 diabetes (e.g., age, sex, obesity, income [44]) were available and included in the model. Indeed, our model estimates risk of diabetes (type 1 or type 2) based on risk factors for type 2 diabetes. The same issue discerning type 1 from type 2 diabetes exists when treating diabetes as a predictor for hypertension. Because the majority of patients with diabetes at baseline likely have type 2 diabetes, the association between diabetes and incident hypertension will largely be determined by these patients. Any patients with type 1 diabetes at baseline will essentially be assigned the risk of a patient with type 2 diabetes at baseline. If the true risk of hypertension differs between patients with type 1 and type 2 diabetes, there may be some misspecification of risk.

The CPCSSN validated case-detecting algorithm for osteoarthritis has lower sensitivity than the algorithms for the other conditions. Assuming misclassified osteoarthritis cases are similar to correctly classified cases among truly positive cases, this may reduce the strength of associations that are observed between the predictors and osteoarthritis and result in underestimated risk. However, the difference in sensitivity is small, thus any underestimation in risk would be expected to be small.

Ideally, a prognostic prediction model should be deployed in the same setting that it was developed [14]. Our use of CPCSSN data strongly positions our model for deployment in the Canadian primary care setting, especially among physicians who submit data to CPCSSN. Use in new settings requires model ‘updating’ using data from the new setting [45]. This is also ideal operationally, as no additional measures beyond those already collected in the physician’s EMR were used in development; thus, no additional measures are required when applying our model to a patient in practice. A future direction could be to pilot test implementation of our model in CPCSSN-contributing settings to passively operate in the background of a physician’s EMR, flagging patients whose estimated risk is above some specified risk threshold.

Conclusion

Prevention efforts are needed to mitigate the increasing population health burden of multimorbidity. Quantitative estimates of risk can play a valuable role by providing a means to better understand potential future health trajectories and to foster discussions between patients and their primary care practitioners about appropriate preventative measures. Our research presents a model that can be used to provide such risk estimates while understanding and accounting for the dependence that exists between outcomes. The methods described above should be considered whenever predicting multiple outcomes where there may be some dependence between diseases. Further research will determine how best to incorporate this model into primary care practitioners’ clinical workflow and assess its real-world performance.

Acknowledgements

We would like to acknowledge Dr. Sonny Cejic for his contributions to this work. Dr. Cejic helped establish the clinical relevance of this work and assessed the accuracy of all case-detecting algorithms constructed for this study.

Statement on conflicts of interest

The authors have no conflicts of interest.

Ethics statement

Ethics approval was obtained from the Western University Research Ethics Board #107572.

Supplemental appendices

Supplemental Appendix 1: Predictor case-detecting algorithms. Case-detecting algorithms for all predictors used in our study.

Supplemental Appendix 2: Maximum entropy trivariate distribution. A brief description of the trivariate probability mass functions that satisfy the bivariate constraints that are specified by a single parameter, α.

Supplemental Appendix 3: R code for copula-based model. R code we developed to estimate the copula-based model, perform hypothesis testing, and estimate trivariate probabilities using the copula-based model.

References

1. Beam AL, Kohane IS. Big data and machine learning in health care. Vol. 319, JAMA - Journal of the American Medical Association. American Medical Association; 2018. p. 1317–8. https://jamanetwork.com/journals/jama/fullarticle/2675024. https://doi.org/10.1001/jama.2017.18391

2. Anthony Celi L, Fine B, Stone DJ. An awakening in medicine: the partnership of humanity and intelligent machines. Lancet Digit Heal. 2019;1:e255–7. https://www.thelancet.com/digital-health. https://doi.org/10.1038/s41436-019-0566-2

3. 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. http://www.ajconline.org/article/S0002914904004370/fulltext. https://doi.org/10.1016/j.amjcard.2004.03.023

4. Hendriksen JMT, Geersing GJ, Moons KGM, de Groot JAH. Diagnostic and prognostic prediction models. J Thromb Haemost. 2013 Jun;11 Suppl 1:129–41. http://www.ncbi.nlm.nih.gov/pubmed/23809117. https://doi.org/10.1111/jth.12262

5. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J. 2014 Aug 1;35(29):1925–31. http://www.ncbi.nlm.nih.gov/pubmed/24898551. https://doi.org/10.1093/eurheartj/ehu207

6. Steyerberg EW, Moons KGM, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. PLoS Med. 2013;10(2). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3564751/. https://doi.org/10.1371/journal.pmed.1001381

7. Muth C, Blom JW, Smith SM, Johnell K, Gonzalez-Gonzalez AI, Nguyen TS, et al. Evidence supporting the best clinical management of patients with multimorbidity and polypharmacy: a systematic guideline review and expert consensus. Vol. 285, Journal of Internal Medicine. Blackwell Publishing Ltd; 2019. p. 272–88. https://pubmed.ncbi.nlm.nih.gov/30357955/. https://doi.org/10.1111/joim.12842

8. Mercer S, Salisbury C, Fortin M. ABC of Multimorbidity. 2014. 20–64 p. https://www.wiley.com/en-ca/ABC+of+Multimorbidity-p-9781118383889.

9. Geersing GJ, De Groot JA, Reitsma JB, Hoes AW, Rutten FH. The impending epidemic of chronic cardiopulmonary disease and multimorbidity: The need for new research approaches to guide daily practice. Vol. 148, Chest. American College of Chest Physicians; 2015. p. 865–9. https://pubmed.ncbi.nlm.nih.gov/25856418/. https://doi.org/10.1378/chest.14-3172

10. Seyed Parham Khalili; Marianna LaNoue. Population Health At Home: Building A Data Warehouse And Applying Cluster Analysis To Identify Patterns In Healthcare Utilization And Multimorbidity Among Primary Care Patients. J Gen Intern Med. 2017;32:S786. https://link.springer.com/article/10.1007/s11606-017-4028-8.

11. Vos R, Aarts S, van Mulligen E, Metsemakers J, van Boxtel MP, Verhey F, et al. Finding potentially new multimorbidity patterns of psychiatric and somatic diseases: Exploring the use of literature-based discovery in primary care research. J Am Med Informatics Assoc. 2014;21(1):139–45. https://pubmed.ncbi.nlm.nih.gov/23775174/. https://doi.org/10.1136/amiajnl-2012-001448

12. 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. http://www.jcomorbidity.com/index.php/test/article/view/72/255. https://doi.org/10.15256/joc.2016.6.72

13. Rijken M, Struckmann V, Van Der Heide I, Hujala A, Barbabella F, Van Ginneken E, et al. How to improve care for people with multimorbidity in Europe? 2016. https://www.euro.who.int/__data/assets/pdf_file/0004/337585/PB_23.pdf?ua=1.

14. 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(7709):1487–90. http://www.ncbi.nlm.nih.gov/pubmed/19502216. https://doi.org/10.1136/bmj.b606

15. O'Caoimh R, Cornally N, Weathers E, O'Sullivan R, Fitzgerald C, Orfila F, et al. Risk prediction in the community: A systematic review of case-finding instruments that predict adverse healthcare outcomes in community-dwelling older adults. Vol. 82, Maturitas. Elsevier Ireland Ltd; 2015. p. 3–21. https://pubmed.ncbi.nlm.nih.gov/25866212/. https://doi.org/10.1016/j.maturitas.2015.03.009

16. 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. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4096475&tool=pmcentrez&rendertype=abstract. https://doi.org/10.1370/afm.1644

17. Tarlier D. TUTOR-PHC 2003/2004 RESEARCH TRAINEES "NO COOKIE-CUTTER RESPONSE" CONCEPTUALIZING PRIMARY HEALTH CARE 1 R. Thomas-MacLean, D. Tarlier, S. Ackroyd-Stolarz, M. Fortin, M. Stewart. 2007;(2001). https://www.uwo.ca/fammed/csfm/tutor-phc/documentation/trainingpapers/TUTOR_Definitio_ of_primar_health_care.pdf.

18. Birtwhistle R V. Canadian Primary Care Sentinel Surveillance Network. Canadian Family Physician 2011. http://cpcssn.ca/.

19. 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. http://www.ncbi.nlm.nih.gov/pubmed/19587256. https://doi.org/10.3122/jabfm.2009.04.090081

20. 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. http://www.ncbi.nlm.nih.gov/pubmed/27331051. https://doi.org/10.9778/cmajo.20140128

21. 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. http://doi.wiley.com/10.1111/j.1365-2753.2008.00952.x. https://doi.org/10.1111/j.1365-2753.2008.00952.x

22. 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. http://www.ncbi.nlm.nih.gov/pubmed/10718692.

23. 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. http://www.ncbi.nlm.nih.gov/pubmed/9614260. https://doi.org/10.1056/NEJM199806043382307

24. Statistics Canada. National Household Survey Profile, 2011. 2013. https://www12.statcan.gc.ca/nhs-enm/2011/dp-pd/prof/index.cfm?Lang=E.

25. 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. http://www.ncbi.nlm.nih.gov/pubmed/25560730. https://doi.org/10.7326/M14-0698

26. Rubin DB, Wiley J, New York Chichester Brisbane Toronto Singapore S. Multiple Imputation for Nonresponse in Surveys. 1987. http://doi.wiley.com/10.1002/9780470316696. https://doi.org/10.1002/9780470316696

27. 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. http://projecteuclid.org/euclid.bjps/1369746494. https://doi.org/10.1214/11-BJPS165

28. Joe H. Multivariate models and dependence concepts. Chapman & Hall; 1997. 399 p. (Monographs on statistics and applied probability). https://books.google.ca/books/about/Multivariate_Models_and_Multivariate_Dep.html?id=iJbRZL2QzMAC&redir_esc=y.

29. Cramer H. Mathematical Methods of Statistics. Princet Univ Press. 1946;282. https://www.jstor.org/stable/j.ctt1bpm9r4.

30. Frank MJ. On the simultaneous associativity of F(x, y) and x+y −F(x, y). Aequationes Math. 1979;21:194–226. https://link.springer.com/article/10.1007/BF02189866.

31. 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. http://www.jstor.org/stable/25046184. https://doi.org/10.2307/25046184

32. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. Chapman & Hall. 1994. https://books.google.ca/books?hl=en&lr=&id=gLlpIUxRntoC&oi=fnd&pg=PR14&dq=bootstrap&ots=A9yvW9ObE0&sig=7rTIwhSbwqrsnJGajHIFWh10_Rw#v=onepage&q=bootstrap&f=false.

33. Gray RM. Entropy and Information Theory. New York. Springer; 2009. 409 p. https://www.springer.com/gp/book/9781441979698.

34. 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 Sep;10(4):430–9. http://linkinghub.elsevier.com/retrieve/pii/S1568163711000249. https://doi.org/10.1016/j.arr.2011.03.003

35. Ng R, Sutradhar R, Kornas K, Wodchis WP, Sarkar J, Fransoo R, et al. Development and Validation of the Chronic Disease Population Risk Tool (CDPoRT) to Predict Incidence of Adult Chronic Disease. JAMA Netw Open. 2020 Jun 4;3(6):e204669. https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2766780. https://doi.org/10.1001/jamanetworkopen.2020.4669

36. May HT, Lappé DL, Knowlton KU, Muhlestein JB, Anderson JL, Horne BD. Prediction of Long-Term Incidence of Chronic Cardiovascular and Cardiopulmonary Diseases in Primary Care Patients for Population Health Monitoring: The Intermountain Chronic Disease Model (ICHRON). Mayo Clin Proc. 2019 Jul 1;94(7):1221–30. http://www.ncbi.nlm.nih.gov/pubmed/30577973. https://doi.org/10.1016/j.mayocp.2018.06.029

37. Lappenschaar M, Hommersom A, Lucas PJF, Lagro J, Visscher S, Korevaar JC, et al. Multilevel temporal Bayesian networks can model longitudinal change in multimorbidity. J Clin Epidemiol. 2013 Dec 1;66(12):1405–16. http://www.ncbi.nlm.nih.gov/pubmed/24035172. https://doi.org/10.1016/j.jclinepi.2013.06.018

38. Lappenschaar M, Hommersom A, Lucas PJF, Lagro J, Visscher S. Multilevel Bayesian networks for the analysis of hierarchical health care data. Artif Intell Med. 2013 Mar 1;57(3):171–83. https://www.sciencedirect.com/science/article/pii/S093336571300002X?via%3Dihub. https://doi.org/10.1016/J.ARTMED.2012.12.007

39. Oniśko A, Druzdzel MJ, Wasyluk H. Extension of the Hepar II model to multiple-disorder diagnosis. In: Advances in Soft Computing. Physica-Verlag (A Springer-Verlag Company; 2000. p. 303–13. https://www.pitt.edu/~druzdzel/psfiles/springer00.pdf. https://doi.org/10.1007/978-3-7908-1846-8_27

40. Wang X, Wang F, Hu J, Sorrentino R. Exploring joint disease risk prediction. AMIA . Annu Symp proceedings AMIA Symp. 2014;2014:1180–7. http://www.ncbi.nlm.nih.gov/pubmed/25954429.

41. Canadian Diabetes Association. An economic tsunami: the cost of diabetes in Canada. 2009. http://www.diabetes.ca/CDA/media/documents/publications-and-newsletters/advocacy-reports/economic-tsunami-cost-of-diabetes-in-canada-english.pdf.

42. DeFronzo RA, Ferrannini E, Zimmet P, Alberti KGMM. International Textbook of Diabetes Mellitus. DeFronzo RA, Ferrannini E, Zimmet P, Alberti KGMM, editors. International Textbook of Diabetes Mellitus. Chichester, UK: John Wiley & Sons, Ltd; 2015. http://doi.wiley.com/10.1002/9781118387658. https://doi.org/10.1002/9781118387658

43. Holt RIG, Cockram CS, Flyvbjerg A, Goldstein BJ. Textbook of diabetes. 5th ed. Wiley-Blackwell; 2017. https://www.wiley.com/en-us/Textbook+of+Diabetes%2C+5th+Edition-p-9781118912027.

44. 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. http://www.ncbi.nlm.nih.gov/pubmed/25249787. https://doi.org/10.7150/ijms.10001

45. Janssen KJM, Vergouwe Y, Kalkman CJ, Grobbee DE, Moons KGM. A simple method to adjust clinical prediction models to local circumstances. Can J Anesth. 2009 Mar;56(3):194–201. https://pubmed.ncbi.nlm.nih.gov/19247740/. https://doi.org/10.1007/s12630-009-9041-x

46. 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. http://www.ncbi.nlm.nih.gov/pubmed/25737033. https://doi.org/10.1016/j.diabres.2015.02.002

47. 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. http://www.ncbi.nlm.nih.gov/pubmed/12610029. https://doi.org/10.2337/diacare.26.3.725

48. 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. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4295827&tool=pmcentrez&rendertype=abstract. https://doi.org/10.1007/s13668-014-0103-5

49. 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. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4205573&tool=pmcentrez&rendertype=abstract. https://doi.org/10.1177/2040622314548679

50. Bates GW, Legro RS. Longterm management of Polycystic Ovarian Syndrome (PCOS). Mol Cell Endocrinol. 2013 Jul 5;373(1–2):91–7. http://www.ncbi.nlm.nih.gov/pubmed/23261983. https://doi.org/10.1016/j.mce.2012.10.029

51. 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. http://www.ncbi.nlm.nih.gov/pubmed/11087022. https://doi.org/10.1093/oxfordjournals.schbul.a033504

52. 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. http://www.ncbi.nlm.nih.gov/pubmed/12113916. https://doi.org/10.1016/S0165-0327(01)00456-6

53. 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. http://link.springer.com/10.1007/s40265-015-0347-4. https://doi.org/10.1007/s40265-015-0347-4

54. 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. http://www.ncbi.nlm.nih.gov/pubmed/12971015. https://doi.org/10.1177/070674370304800705

55. 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):1–11. https://pubmed.ncbi.nlm.nih.gov/25894943/. https://doi.org/10.1007/s11892-015-0603-8

56. 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. http://www.ncbi.nlm.nih.gov/pubmed/22971751. https://doi.org/10.1038/jhh.2012.33

57. 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; http://www.ncbi.nlm.nih.gov/pubmed/25833704. https://doi.org/10.1038/jhh.2015.30

58. Cuffee Y, Ogedegbe C, Williams NJ, Ogedegbe G, Schoenthaler A. Psychosocial Risk Factors for Hypertension: an Update of the Literature. Curr Hypertens Rep. 2014;16(10). http://link.springer.com/article/10.1007/s11906-014-0483-3. https://doi.org/10.1007/s11906-014-0483-3

59. Gargiulo R, Suhail F, Lerma E V. Hypertension and chronic kidney disease. Disease-a-Month. 2015 Sep;61(9):387–95. http://www.ncbi.nlm.nih.gov/pubmed/26328515. https://doi.org/10.1016/j.disamonth.2015.07.003

60. 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. http://hyper.ahajournals.org/cgi/doi/10.1161/HYPERTENSIONAHA.108.126698. https://doi.org/10.1161/HYPERTENSIONAHA.108.126698

61. 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. http://www.ncbi.nlm.nih.gov/pubmed/22987869. https://doi.org/10.1542/peds.2011-3870

62. Floras JS. Hypertension and sleep apnea. Can J Cardiol. 2015 May;31(7):889–97. http://www.ncbi.nlm.nih.gov/pubmed/26112299. https://doi.org/10.1016/j.cjca.2015.05.003

63. 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. https://pubmed.ncbi.nlm.nih.gov/25510755/. https://doi.org/10.3349/ymj.2015.56.1.124

64. 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. https://pubmed.ncbi.nlm.nih.gov/9521177/. https://doi.org/10.1093/oxfordjournals.aje.a009482

65. Neogi T, Zhang Y. Osteoarthritis prevention. Curr Opin Rheumatol. 2011 Mar;23(2):185–91. https://doi.org/10.1097/BOR.0b013e32834307eb

66. 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. http://www.ncbi.nlm.nih.gov/pubmed/25447976. https://doi.org/10.1016/j.joca.2014.11.019

67. 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. http://www.ncbi.nlm.nih.gov/pubmed/16777458. https://doi.org/10.1016/j.jbspin.2006.03.001

68. Harvey WF, Yang M, Cooke TDV, 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):287–95. https://pubmed.ncbi.nlm.nih.gov/20194234/. https://doi.org/10.7326/0003-4819-152-5-201003020-00006

69. 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. https://onlinelibrary.wiley.com/doi/10.1111/jphp.12196. https://doi.org/10.1111/jphp.12196

70. 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. https://pubmed.ncbi.nlm.nih.gov/19921240/. https://doi.org/10.1007/s00420-009-0486-6

71. Centre for Chronic Disease Prevention Public Health Agency of Canada. Chronic Disease Indicator Framework, 2013 Edition. 2013; https://www.phac-aspc.gc.ca/publicat/hpcdp-pspmc/34-1-supp/assets/pdf/34-S1_E_v6.pdf.

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

Abbreviations

AUC:	area under the receiver operator characteristic curve
BMI:	body mass index
CPCSSN:	Canadian Primary Care Sentinel Surveillance Network
EMR:	electronic medical record
FSA:	forward sortation area

Appendix Table 1: Validation of CPCSSN case-detecting algorithms [16]

Outcome	Sensitivity % (95% CI)	Specificity % (95% CI)
Hypertension	84.9 (82.6 to 87.1)	93.5 (92.0 to 95.1)
Diabetes	95.6 (93.4 to 97.9)	97.1 (96.3 to 97.9)
Osteoarthritis	77.8 (74.5 to 81.1)	94.9 (93.8 to 96.1)

Appendix Table 2: Outcome predictors

Diabetes	Hypertension	Osteoarthritis
Hypertension [46, 47]	Older age [56]	Osteoporosis [63]
Older age [46, 47]	Diabetes [56, 57]	Previous leg injury [64–67]
Lipid disorders [46]	Obesity [56, 57]	Leg length inequality [68]
Obesity [46–49]	Smoking [56]	Older age [63, 65, 66, 69, 70]
Waist circumference	Stress [58]	Obesity [63–67, 69, 70]
Smoking [46, 47]	Kidney disease [59]	Female sex [63, 65, 66, 69]
Stress [46]	Tricyclic antidepressant (TCA) use [60]	Family history of osteoarthritis [69]
Male sex [46]		
Polycystic ovarian syndrome (PCOS) [50]	High salt intake [56, 61]	Physically intensive occupations [69]
Schizophrenia [51, 52]	Sleep apnea [62]	
Depression [53]		
Bipolar disorder [52, 54]		
Low physical activity [49]		
Family history of type 2 diabetes [46]		
Air pollution [55]		
Low socioeconomic status [49]		

Appendix Table 3: Descriptive statistics

	Development set		**Validation set**	
	(n = 265,228)		(n = 160,000)	
	n cases	%	n cases	%
Osteoarthritis	26,013	9.8%	15,840	9.9%
Diabetes	18,140	6.8%	10,839	6.8%
Hypertension	41,185	15.5%	24,845	15.5%
Depression	38,629	14.6%	23,348	14.6%
Smoking	11,037	63.8%	6,807	64.2%
Female Sex	153,664	57.9%	93,202	58.3%
Alcohol	4,038	1.5%	2,429	1.5%
Stress	7,907	3.0%	4,729	3.0%
Epilepsy	1,842	0.7%	1,137	0.7%
Schizophrenia	3,955	1.5%	2,424	1.5%
Anxiety	18,894	7.1%	11,432	7.1%
Cancer	11,139	4.2%	6,514	4.1%
Cardiovascular Disease	14,730	5.6%	8,772	5.5%
COPD	4,515	1.7%	2,750	1.7%
Rheumatoid Arthritis	2,039	0.8%	1,224	0.8%
Lipid Disorder	47,619	18.0%	28,634	17.9%
Polycystic Ovarian Syndrome	706	0.5%	448	0.5%
Chronic Kidney Disease	9,283	3.5%	5,484	3.4%
Tricyclic Antidepressant Use	8,114	3.1%	4,921	3.1%
Osteoporosis	8,971	3.4%	5,413	3.4%
Leg Injury	7,808	2.9%	4,603	2.9%
Family History of Osteoarthritis	168	0.1%	114	0.1%
Family History of Diabetes	2,851	1.1%	1,727	1.1%
Family History of Hypertension	1,817	0.7%	1,087	0.7%
Lives in a rural location	55,527	20.9%	33,371	20.9%
Morbidity				
−1 disease*	79,671	30.0%	48,110	30.1%
Multimorbidity				
−2 disease*	23,565	8.9%	14,114	8.8%
−3 disease*	6,286	2.4%	3,777	2.4%
Age				
−18 to 24	36,962	13.9%	21,985	13.7%
−25 to 44	89,608	33.8%	54,052	33.8%
−45 to 64	95,161	35.9%	57,763	36.1%
−65 and older	43,330	16.3%	26,108	16.3%
BMI				
−Underweight (< 18.5 kg/m^2)	1,680	1.9%	1,014	1.9%
−Normal (18.5 to 24.9 kg/m^2)	30,541	34.1%	18,379	33.9%
−Overweight (25 to 29.9 kg/m^2)	30,736	34.3%	18,644	34.4%
−Obese (≥ 30 kg/m^2)	26,639	29.7%	16,195	29.9%
Personal Income				
−Less than $30000	1,401	0.6%	**	**
−$30000 to $49999	185,981	74.0%	111,810	73.8%
−$50000 to $74999	64,016	25.5%	38,768	25.6%
−Greater than $75000	6	0.0%	**	**

*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 [71].

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

Percentages are based on patients with complete data for the characteristic.

BMI: body mass index.

COPD: chronic obstructive pulmonary disease.

Appendix Table 4: BMI distribution compared to Canadian population

	Development set (n = 265228)		**Validation set** (n = 160000)		Canadian Population
	n cases	%	*n* cases	%	%
BMI					
-Underweight (< 18.5 kg/m^2)	1,680	1.9%	1,014	1.9%	
-Normal (18.5 to 24.9 kg/m^2)	30,541	34.1%	18,379	33.9%	32%*
-Overweight (25 to 29.9 kg/m^2)	30,736	34.3%	18,644	34.4%	40%*
-Obese (> 30 kg/m^2)	26,639	29.7%	16,195	29.9%	27%*

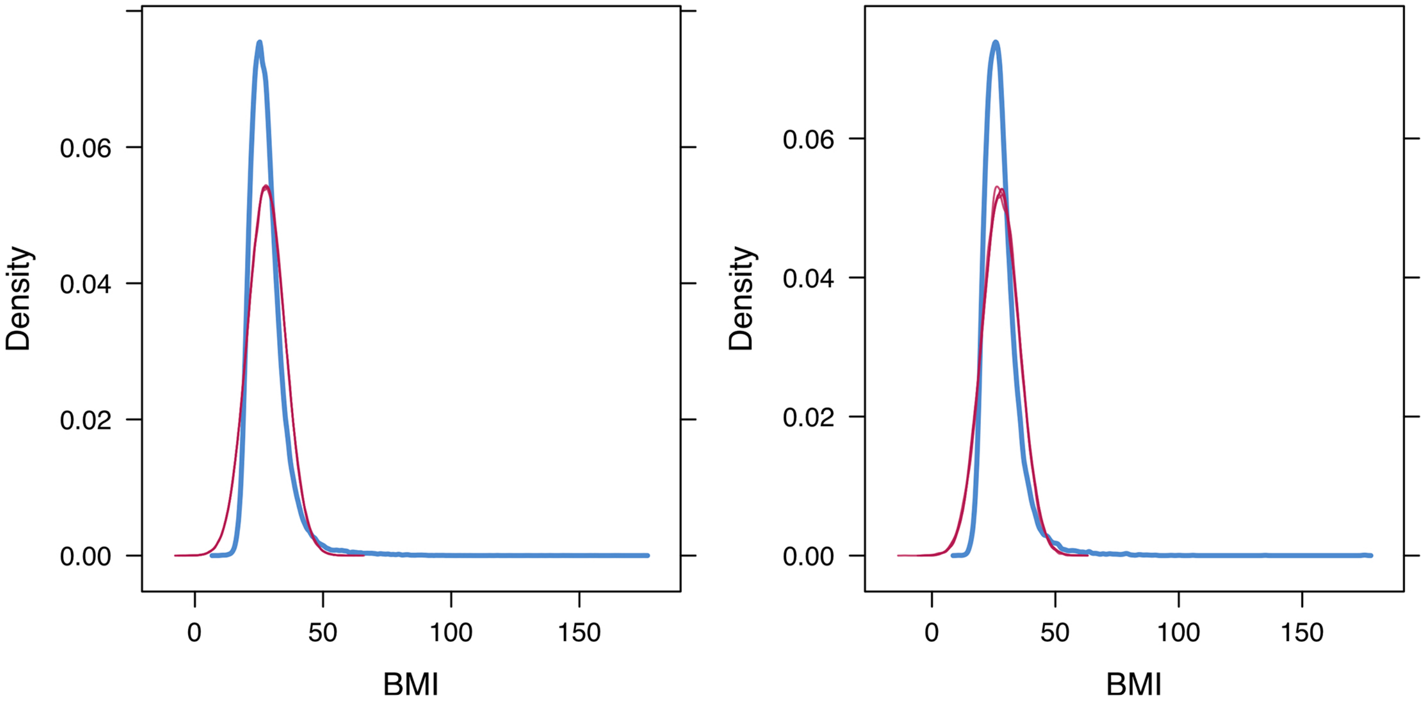
*Canadian Health Measures Survey (CHMS) 2009/10 [72].
BMI: body mass index.

Appendix Table 5: Sensitivity analysis excluding patients without encounter after follow-up

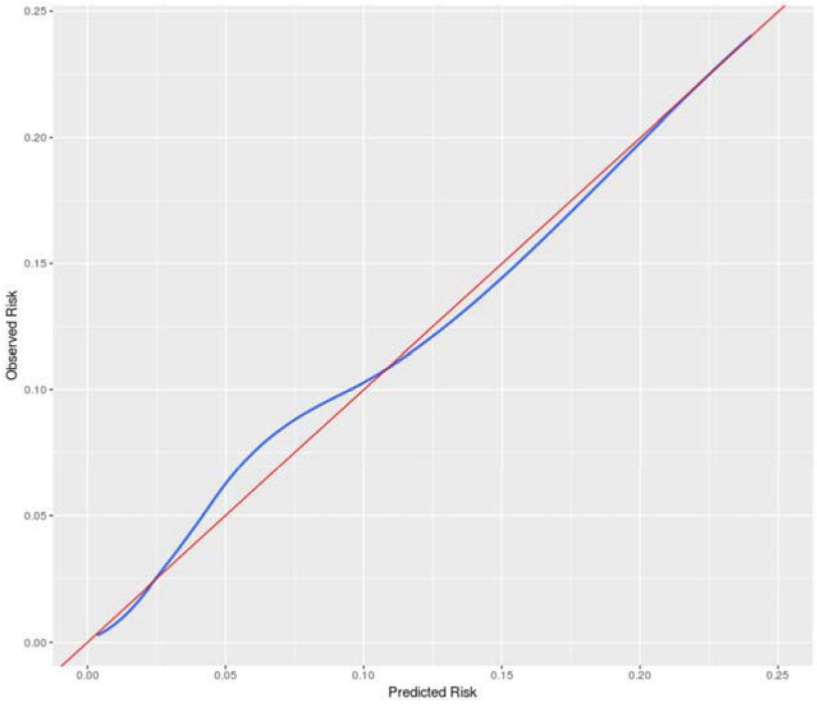
	Reference category/units	**β estimate**	**95% CI**	**Odds ratio**	**95% CI**
Diabetes univariate model					
Hypertension	No	Reference		Reference	
	Yes	0.23	0.18 to 0.28	1.26	1.20 to 1.32
Age	(Years)	0.03	0.03 to 0.04	1.03	1.03 to 1.04
Lipid disorders	No	Reference		Reference	
	Yes	1.55	1.50 to 1.60	4.71	4.50 to 4.93
BMI	(kg/m^2)	0.06	0.06 to 0.07	1.06	1.06 to 1.07
Sex	Male	Reference		Reference	
	Female	-0.28	-0.32 to -0.23	0.76	0.72 to 0.79
Schizophrenia	No	Reference		Reference	
	Yes	0.56	0.42 to 0.70	1.75	1.52 to 2.01
Depression	No	Reference		Reference	
	Yes	0.07	0.01 to 0.14	1.08	1.01 to 1.15
Income	($10,000)	-0.06	-0.09 to -0.03	0.94	0.91 to 0.97
Hypertension univariate model					
Diabetes	No	Reference		Reference	
	Yes	0.15	0.10 to 0.20	1.16	1.10 to 1.22
Age	(Years)	0.06	0.06 to 0.07	1.06	1.06 to 1.07
BMI	(kg/m^2)	0.04	0.04 to 0.05	1.05	1.04 to 1.05
Chronic Kidney Disease	No	Reference		Reference	
	Yes	0.64	0.58 to 0.70	1.9	1.79 to 2.02
Tricyclic Antidepressant Use	No	Reference		Reference	
	Yes	0.57	0.51 to 0.64	1.77	1.66 to 1.89
Osteoarthritis univariate model					
Age	(Years)	0.06	0.05 to 0.06	1.06	1.06 to 1.07
Sex	Male	Reference		Reference	
	Female	0.19	0.14 to 0.25	1.21	1.15 to 1.28
BMI	(kg/m^2)	0.04	0.03 to 0.04	1.04	1.03 to 1.04
Previous Leg Injury	No	Reference		Reference	
	Yes	1.52	1.43 to 1.60	4.56	4.18 to 4.97
Osteoporosis	No	Reference		Reference	
	Yes	0.78	0.70 to 0.86	2.18	2.01 to 2.37

AUC: area under the receiver operator characteristic curve; BMI: body mass index; CI: confidence interval.

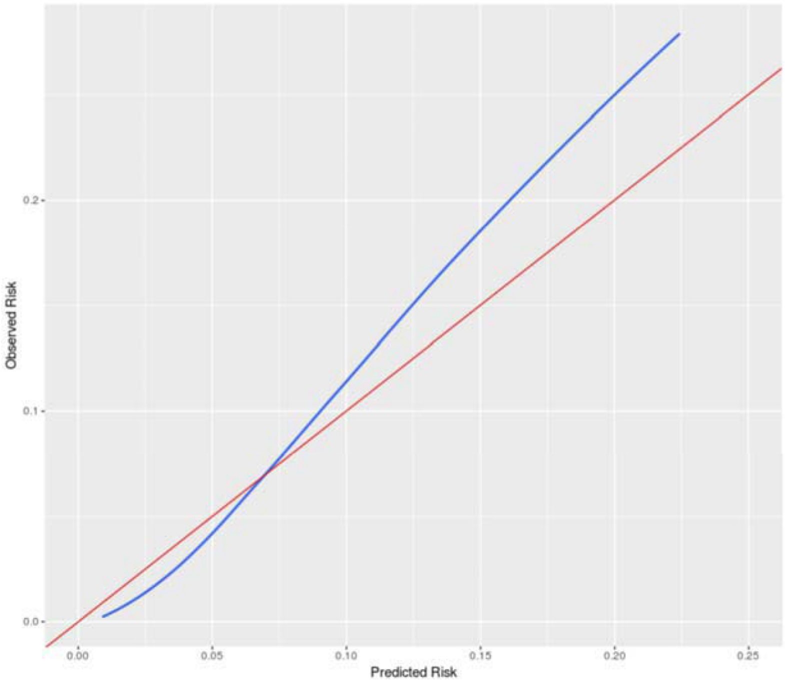
Appendix Figure 1: Kernel density estimates for the marginal distribution of the five imputed datasets (red) and the original data (blue) (left: development sets; right: validation sets)



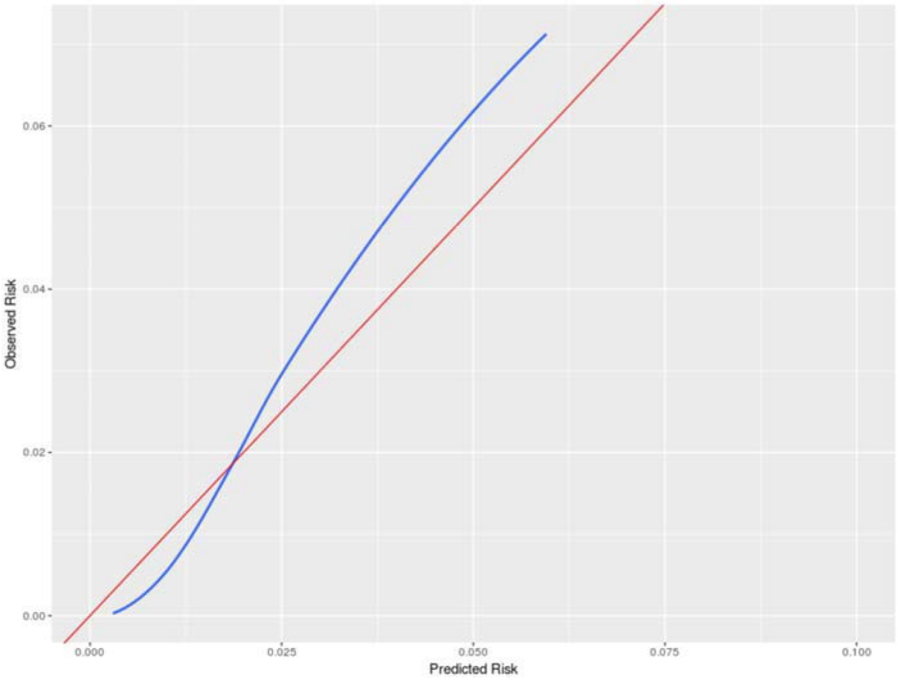
Appendix Figure 2a: Calibration plot for diabetes



Appendix Figure 2b: Calibration plot for hypertension



Appendix Figure 2c: Calibration plot for osteoarthritis


Observed Risk
Predicted Risk