JAMA Network Open

Original Investigation | Public Health

Development and Validation of a Machine Learning Model Using Administrative Health Data to Predict Onset of Type 2 Diabetes

Mathieu Ravaut, MSc; Vinyas Harish, BCompH; Hamed Sadeghi, PhD; Kin Kwan Leung, PhD; Maksims Volkovs, PhD; Kathy Kornas, MPH; Tristan Watson, MPH; Tomi Poutanen, MSc; Laura C. Rosella, PhD

Abstract

IMPORTANCE Systems-level barriers to diabetes care could be improved with population health planning tools that accurately discriminate between high- and low-risk groups to guide investments and targeted interventions.

OBJECTIVE To develop and validate a population-level machine learning model for predicting type 2 diabetes 5 years before diabetes onset using administrative health data.

DESIGN, SETTING, AND PARTICIPANTS This decision analytical model study used linked administrative health data from the diverse, single-payer health system in Ontario, Canada, between January 1, 2006, and December 31, 2016. A gradient boosting decision tree model was trained on data from 1 657 395 patients, validated on 243 442 patients, and tested on 236 506 patients. Costs associated with each patient were estimated using a validated costing algorithm. Data were analyzed from January 1, 2006, to December 31, 2016.

EXPOSURES A random sample of 2 137 343 residents of Ontario without type 2 diabetes was obtained at study start time. More than 300 features from data sets capturing demographic information, laboratory measurements, drug benefits, health care system interactions, social determinants of health, and ambulatory care and hospitalization records were compiled over 2-year patient medical histories to generate quarterly predictions.

MAIN OUTCOMES AND MEASURES Discrimination was assessed using the area under the receiver operating characteristic curve statistic, and calibration was assessed visually using calibration plots. Feature contribution was assessed with Shapley values. Costs were estimated in 2020 US dollars.

RESULTS This study trained a gradient boosting decision tree model on data from 1 657 395 patients (12 900 257 instances; 6 666 662 women [51.7%]). The developed model achieved a test area under the curve of 80.26 (range, 80.21-80.29), demonstrated good calibration, and was robust to sex, immigration status, area-level marginalization with regard to material deprivation and race/ethnicity, and low contact with the health care system. The top 5% of patients predicted as high risk by the model represented 26% of the total annual diabetes cost in Ontario.

CONCLUSIONS AND RELEVANCE In this decision analytical model study, a machine learning model approach accurately predicted the incidence of diabetes in the population using routinely collected health administrative data. These results suggest that the model could be used to inform decision-making for population health planning and diabetes prevention.

JAMA Network Open. 2021;4(5):e2111315. doi:10.1001/jamanetworkopen.2021.11315

Key Points

Question Can a machine learning model trained on routinely collected administrative health data be used to accurately predict the onset of type 2 diabetes at the population level?

Findings In this decision analytical model study of 2.1 million residents in Ontario, Canada, a machine learning model was developed with high discrimination, population-level calibration, and calibration across population subgroups.

Meaning Study results suggest that machine learning and administrative health data can be used to create population health planning tools that accurately discriminate between high- and low-risk groups to guide investments and targeted interventions for diabetes prevention.

+ Supplemental content

Author affiliations and article information are listed at the end of this article.

Introduction

The global incidence and prevalence of diabetes is rising steadily, imposing considerable burden on health care systems. Between 2010 and 2030, it is projected that the prevalence of all forms of diabetes in adults will increase by 69% in developing countries and by 20% in developed countries.[1] In 2030, it is projected that the prevalence of diabetes will reach 55 million people in the US, 62 million in China, and 87 million in India.[1,2] Finally, in 2015, the global cost of diabetes was estimated to be $1.31 trillion US dollars.[3]

Serious efforts and investments into the prevention of type 2 diabetes are vital, and it has been well established that prevention programs are effective not only in clinical trials but in pragmatic, real-world settings.[4,5] However, it has proved difficult to scale diabetes prevention from the individual patient to the population due to systems-level barriers.[6] These barriers include disparities in socioeconomic status,[7-9] lack of access to healthy foods and medications,[10-12] lack of access to health care,[13,14] and the built environments in which people at risk of diabetes live.[15,16] These barriers, many of which are also known as the social determinants of health, contribute to "cascades in care" in which large segments of the population do not meet prevention targets.[17]

Identifying those most in need of interventions (eg, communities that could benefit from subsidies to access healthy foods or diabetes screening and prevention clinics) at the system level by governments, health insurance providers, and public health planners may be hampered by the lack of efficient systems to identify the distribution of risk in the population accurately.[5,18] Extensive research exists on building diabetes risk prediction models with traditional statistical approaches and machine learning[19-34]; however, the vast majority of these models are created for direct patient care and not for application at the level of the entire population for public health planning. A systematic review conducted on traditional, statistical diabetes risk scores in 2011 concluded that most risk scores are rarely used because they rely on uncommon tests or were not developed with end users in mind.[34] The review also concluded that using risk scores on population data sets to identify targets for public health interventions is a promising direction for continued work.[34] These population-level data sets, also known as administrative health data, are high-dimensional, are impossible to fully explore by clinicians or health system administrators using traditional methods, and represent opportunities for automated, machine learning–based approaches.

We aimed to develop and validate a population-level machine learning model to predict the incidence of type 2 diabetes 5 years before the actual onset of diabetes with high performance using routinely collected administrative health data. The main purpose of our model was to inform population health planning and management for the prevention of diabetes that incorporates health equity. It was not our goal for this model to be applied in the context of individual patient care. We developed and validated our model using a large, contemporary cohort from Ontario, Canada's single-payer health insurance system that covers all residents. We created our model with the intention that it could be used on data that are routinely collected by governments or health insurance systems, thereby offering efficient, population-level applicability while maintaining robust performance. Our model was assessed for discrimination and calibration, as well as calibration in key demographic subgroups. We also estimated the costs associated with the incident cases predicted by our model each year to demonstrate the financial incentives of using such an approach to target preventive efforts at the health system level.

Methods

Study Design and Participants

This decision analytical model study used administrative health services records linked with population and other data holdings covered under a single-payer health system in Ontario, Canada. We used an 11-year period from January 1, 2006, to December 31, 2016. This study obtained ethics approval from the Research Ethics Board at the University of Toronto (protocol No. 37650). The need for informed consent was waived owing to the use of deidentified patient data. In Ontario, all residents are eligible for universal health coverage; therefore, administrative health data cover virtually every resident. Moreover, Ontario is Canada's most populous province and among the most ethnically diverse populations in the world.[35] In 2016, it had a population of 13.2 million, of whom almost 30% were immigrants.[35]

The study linked multiple diverse data sources including demographic information, census, physician claims, laboratory results, prescription medication history, hospital and ambulatory usage, and others. Our administrative health data are significantly distinct from electronic medical records. Details on the specific administrative health data that we selected from the Institute of Clinical Evaluative Sciences (ICES) can be found in eTables 1 and 2 in the Supplement.

We randomly sampled 3 000 000 patients linked with Ontario's Registered Persons Database, with no initial exclusion criteria, which decreased to 2 137 343 after excluding patients not alive as of January 1, 2013, the earliest date of prediction of the model in this study's design. From this cohort, we also removed patients not residing in Ontario, patients already diagnosed with diabetes, and patients not in contact with the health care system. This last criterion designates patients having a last registered interaction with the health care system before the end of the target window. The proportion of patients in our final cohort with incident diabetes and those without diabetes reflects incidence rates reported in studies at the population level.[36] In designing and reporting this study, we adhered to the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) and Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guidelines.[37,38]

Model Development

For each patient in the cohort, we partitioned the entire time period into sliding (over the time dimension) patient-time instances that represent a view of the patient at a specific point in time. The detailed diagram of our end-to-end pipeline is shown in eFigure 1 in the Supplement, and further explanations on the instance creation procedure can be found in eMethods 1 in the Supplement. Following a recently proposed approach also training a model at the instance level,[39] each instance corresponded to pairs made of a 2-year block of the patient's history and its associated binary diabetes onset label 5 years later. Instances were separated by 3-month gaps, which allowed us to make quarterly predictions.

This sliding-window, multi-instance approach simulated continuous population screening in a practical application and was also conceptually similar to discrete-time survival analysis methods in which covariates are processed in sequential chunks.[40-42] We simulated a system in which the entire cohort was screened every 3 months, and the risk of developing diabetes was computed for each patient. The system's task was to accurately capture all instances of developed diabetes in the target prediction window (answering the question: Will the patient develop diabetes at any time during the target window?), which required the model to perform well across patients and across time. Three months is a typical update frequency in our administrative health databases; thus, running the model to make a new patient's future state prediction every 3 months allowed us to constantly refresh the predictions as new data became available.[43]

We partitioned the cohort into 3 nonoverlapping sets of patients, with 1 953 494 patients for model training, 300 000 for validation, and 300 000 for testing. Patients in each set were selected at random. All model developments and parameter selections were performed on the training and validation sets, and the test set was kept untouched for final performance reports. To reduce the time bias, we further partitioned the data in time. For patients in the training set, we used instances that had target windows in the period of January 1, 2013, to December 31, 2014 (2 years, or 8 instances). Similarly, for validation and test sets, only instances with target windows within the periods of January 1, 2015, to December 31, 2015, and January 1, 2016, to December 31, 2016, were used (1 year in each case, or 4 instances). The detailed statistics for each set are summarized in **Table 1**. Partitioning the training, validation, and test sets in time as well as patients ensured zero overlap between the sets. This process provided a more accurate estimate of performance because, in practice, the model would also be applied to patients who are newly added to the system (ie, unseen during model training and internal validation), and all predictions would be done forward in time compared with the training data.

We examined more than 300 features derived from demographic details, geographic information, chronic conditions, and health care use history. Stationary demographic features included sex, birth year, immigrant status, and country of origin. Geographic information comprised residence statistics and measures of area-level socioeconomic status from recent census surveys at the Dissemination Area (400-700 individuals) level. Race/ethnicity and material deprivation

Table 1. Cohort Description[a]

	No. (%)					
	Training (January 2013 to December 2014)		Validation (January to December 2015)		Test (January to December 2016)	
Variable	Total	Positives	Total	Positives	Total	Positives
Full cohort						
Unique patients, No.	1 657 395	23 979	243 442	1874	236 506	1967
Instances, No.	12 900 257	23 979	959 276	1874	927 230	1967
Sex						
Male	6 233 595 (48.3)	12 249 (51.1)	459 715 (47.9)	971 (51.8)	440 433 (47.5)	999 (50.8)
Female	6 666 662 (51.7)	11 730 (48.9)	499 561 (52.1)	903 (48.2)	486 797 (52.5)	968 (49.2)
Age group, y						
<10	1 616 100 (12.5)	205 (0.9)	102 462 (10.7)	14 (0.7)	88 668 (9.6)	8 (0.4)
10-19	1 954 979 (15.2)	358 (1.5)	142 442 (14.8)	32 (1.7)	136 183 (14.7)	32 (1.6)
20-29	1 939 960 (15.0)	696 (4.0)	148 168 (15.4)	75 (4.0)	144 396 (15.6)	79 (4.0)
30-39	1 882 470 (14.6)	2624 (10.9)	140 953 (14.7)	220 (11.7)	135 758 (14.6)	203 (10.3)
40-49	2 108 830 (16.3)	5374 (22.4)	155 409 (16.2)	423 (22.6)	149 244 (16.1)	437 (22.2)
50-59	1 657 299 (12.8)	6353 (26.5)	130 529 (13.6)	486 (25.9)	130 880 (14.1)	524 (26.7)
60-69	987 254 (7.7)	4701 (19.6)	80 069 (8.3)	364 (19.4)	82 448 (8.9)	423 (21.5)
70-79	510 517 (4.0)	2438 (10.2)	39 803 (4.1)	182 (9.7)	40 475 (4.4)	181 (9.2)
80-89	222 638 (1.7)	902 (3.8)	17 637 (1.8)	72 (3.8)	17 239 (1.9)	74 (3.8)
90-100	19 840 (0.2)	53 (0.2)	1761 (0.2)	6 (0.3)	1924 (0.2)	6 (0.3)
Immigration status						
Immigrant	1 537 571 (11.9)	4293 (17.9)	122 532 (12.8)	338 (18.0)	122 607 (13.2)	384 (19.5)
Long-term resident	11 362 686 (88.1)	19 686 (82.1)	836 744 (87.2)	1536 (82.0)	804 623 (86.8)	1583 (80.5)
Race/ethnicity marginalization score, quintile[b]						
1st	19 588 853 (15.2)	3690 (15.4)	144 694 (15.1)	275 (14.7)	136 943 (14.8)	303 (15.4)
2nd	2 083 902 (16.2)	3604 (15.0)	153 306 (16.0)	274 (14.6)	147 340 (15.9)	250 (12.7)
3rd	2 279 478 (17.7)	3711 (15.5)	167 552 (17.5)	304 (16.2)	162 545 (17.5)	318 (16.2)
4th	2 698 267 (20.9)	4441 (18.5)	201 623 (21.0)	355 (18.9)	194 554 (21.0)	366 (18.6)
5th	3 710 695 (28.8)	8126 (33.9)	279 566 (29.1)	642 (34.3)	273 841 (29.5)	703 (35.8)
Deprivation marginalization score, quintile[b]						
1st	3 041 507 (23.6)	4339 (18.1)	227 873 (23.8)	366 (19.5)	220 439 (23.8)	358 (18.2)
2nd	2 566 726 (19.9)	4569 (19.1)	190 232 (19.8)	333 (17.8)	185 106 (20.0)	383 (19.5)
3rd	2 442 622 (18.9)	4572 (19.1)	182 185 (19.0)	359 (19.2)	173 694 (18.7)	372 (18.9)
4th	2 288 370 (17.7)	4714 (19.7)	170 096 (17.7)	394 (21.0)	164 405 (17.7)	420 (21.4)
5th	2 391 970 (18.5)	5378 (22.4)	176 355 (18.4)	398 (21.2)	171 579 (18.5)	407 (20.7)

[a] We give the number of patients, number of instances, and associated number of positive data points for the training, validation, and test sets. Note that the number of positive patients and instances match exactly as a patient can only be diagnosed once with diabetes; we also give the distribution of each set in terms of sex, age, and immigration status.

[b] Race/ethnicity and deprivation marginalization scores quantify the degree of marginalization within each dissemination area according to ethnic concentration and material deprivation. A dissemination area typically encompasses a few hundred inhabitants. These 2 scores are quintiles ranging from 1 to 5 based on each patient's history from the 2004-2008 period, where 5 represents a highest degree of marginalization.

marginalization scores were built with the Ontario Marginalization Index and reflected neighborhood-level socioeconomic information.[44] Health care use included information on physician or specialist visits, emergency department visits, laboratory results, hospitalizations and ambulatory usage, and prescription history during the observation window (eFigure 2 and 3 in the Supplement). Extensive details on feature engineering can be found in eMethods 2 in the Supplement.

We trained the gradient boosting decision tree model implemented in Python in the XGBoost (The XGBoost Contributors) open source library.[45] The gradient boosting decision tree model was chosen owing to its ability to handle different feature types and missing values and good support for explainability. Not all patients have values for all features, given variation in health care use and laboratory tests. We did not remove patients with missing values, because XGBoost can still produce predictions without complete case data. Details on the XGBoost model parameters can be found in eMethods 3 in the Supplement. Results for different buffer sizes can be found in eTable 3 in the Supplement. XGBoost was compared with logistic regression in eTable 4 in the Supplement.

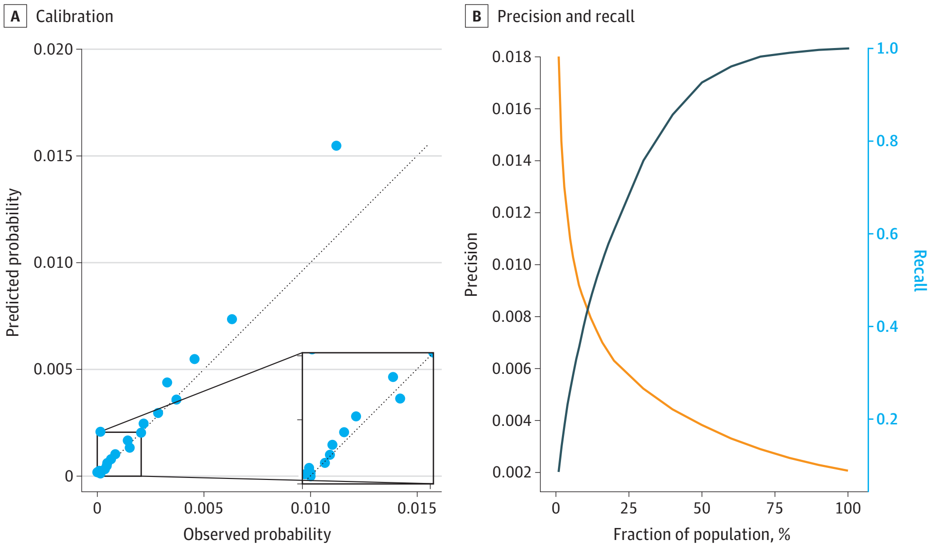
Statistical Analysis

To assess model performance, given the extremely unbalanced class ratio, we tracked the area under the receiver operating characteristic curve (AUC). The AUC is commonly used for such prediction tasks and is robust to class imbalances.[46] We reported the model's calibration curve in **Figure 1** for a visual verification of calibration. For practical application, it was relevant to focus on high-risk patients (ie, those with the highest predicted probability of developing type 2 diabetes) given that our cohort is at the population level. To evaluate the model's performance on the highest-risk patients, we display the precision and recall curves in eFigure 4 in the Supplement.

As shown in **Figure 2**, we evaluated the model on several subsets of the data, separating patients by sex, age, immigration status, marginalization (in terms of both race/ethnicity and material deprivation), and number of events. The number of events was defined as the total number of times that a patient interacted with the health care system in any way during the observation window. It was possible for the patient to have zero events during 1 or several observation windows, in which case the only nonzero variables in the patient's instance features would be never-missing stationary variables, such as the country of birth or sex. We reported the feature contribution using the Shapley values, further described in eFigure 3 in the Supplement.[47,48]

To assess the financial burden of the cohort of patients with diabetes, we used a costing algorithm developed by ICES.[49] This algorithm provides the total public health care expenditure per

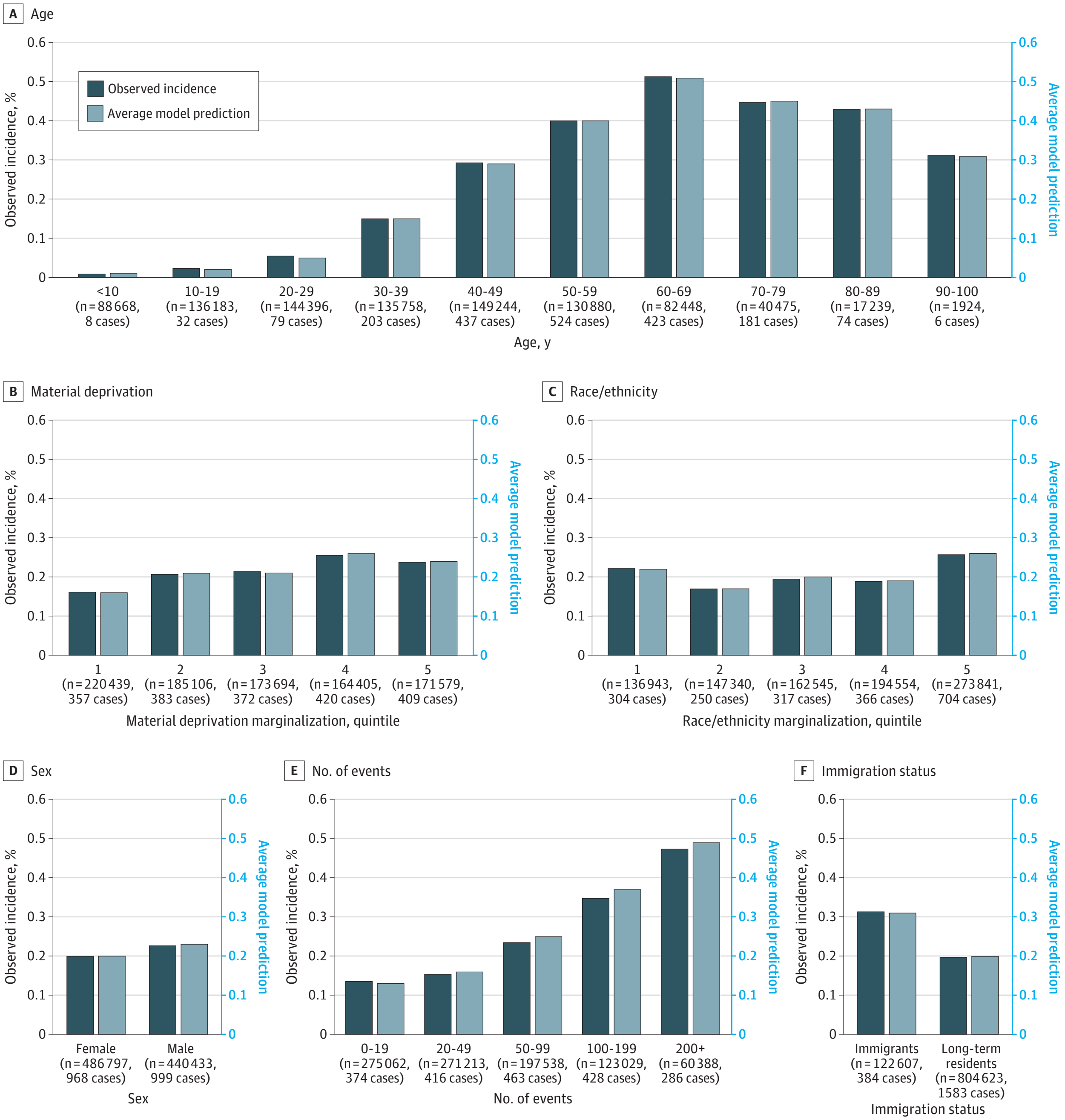
Figure 1. Diabetes Onset Prediction Performance



A, Calibration is assessed visually with a calibration curve composed of 20 population bins of equal size. B, Precision and recall curves are displayed. The left y-axis corresponds to precision, and the right y-axis to recall. The test area under the receiver operating curve is 80.26.

year for each patient, based on the patient's billing information across health care services. With this costing algorithm, we derived the annual cost of the cohort of patients with diabetes in Ontario, as well as the annual change in this cost. This cohort grows over time as the number of patients newly diagnosed with diabetes is greater than the number of patients with diabetes who die each year. We used this algorithm in combination with our model's predictions to estimate how cost-effective the policies implemented with the model could be. This process was done by sorting patients in the test

Figure 2. Diabetes Onset Calibration Across Population Groups



The model is evaluated on specific subsets of the population: sex (2 categories), age (10 bins of 10 years), immigration status (2 categories), race/ethnicity marginalization score (5 quintiles), material deprivation marginalization score (5 quintiles), and number of events in the observation window (5 categories). We display the incidence rate (left y-axis, dark blue bars), average model prediction (right y-axis, light blue bars), and number of positive cases within each subset. The size of each subset can be read on the x-axis. Note that incidence rates can vary dramatically between subsets, especially for age, making comparisons between subsets challenging.

set by decreasing model prediction (from the highest likelihood of getting diabetes as predicted by the model to the lowest) and computed the cumulated cost of these patients. All costs were reported in 2020 US dollars. Data were analyzed using SAS Enterprise software, version 6.1 (SAS Institute Inc) from January 1, 2006, to December 31, 2016.

Results

After applying the exclusion criteria, the resulting cohort sizes were 1 657 395 patients for training (12 900 257 instances; 6 666 662 women [51.7%]), 243 442 for validation, and 236 506 for testing (Table 1). That is, we used 83.7% of the patients in our analytic cohort, substantially more than similar studies.[23] A total of 416 151 patients were excluded: 191 999 patients owing to date of last contact being before the earliest possible target window of their set (training, validation, or test), 103 613 because they were immigrants who arrived in Canada after the end of their observation window, and 120 539 because they were already diagnosed with diabetes. The training, validation, and test sets contained 12 900 257, 959 276, and 927 230 patient instances, respectively. All of the results reported in this section are referring to the test set unless mentioned otherwise.

Figure 1 displays the performance of the model in different evaluation setups. We computed the AUC for all instances in the test set, spanning all 4 quarters of 2016 for each test patient. As seen in Figure 1, the model achieved a test AUC of 80.26 (range, 80.21-80.29) on this held-out set. The calibration curve contained 20 bins with an equal number of patients and was well aligned with the identity line, which showed a good calibration overall. The only exception was the last bin, which showed model overprediction for high-risk patients.

As shown in Figure 2, we evaluated the model on several partitions of the test population, and for each subset, we reported the size, incidence rate, and average model prediction. Incidence rates varied significantly across subsets: it was less than 0.1% for patients aged 20 to 29 years, whereas it was greater than 0.5% among those aged 60 to 69 years. We used the following partitions: sex, age, immigration status, material deprivation marginalization, race/ethnicity marginalization, and number of events in the observation window. We observed that, for all partitions, the model was well calibrated across all subsets except for the number of events; for a higher number of events, the model was slightly overpredicting.

We included an analysis that demonstrates how such a prediction model could be informative at the population level by examining predicted risk across the population into groups that can be segmented for health system planning, such as targeted interventions or resources. **Table 2** depicts an analysis of the model prediction bins and the same analysis within subgroups of the population.

Table 2. Model Prediction Risk Levels[a]

		Individuals, %			Marginalization scores		
Bin	Age, mean	Women	Immigrants	Time in Canada, y	Ethnicity	Deprivation	HbA_{1c}, mean
Model prediction							
Top 1%	58.3	59.6	38.8	17.3	4.22	3.63	5.84
Next 5%	59.4	42.3	26.5	18.4	3.85	3.45	5.81
Next 15%	58.3	40.8	16.5	19.4	3.44	3.15	5.73
Bottom 79%	31.8	55.3	11.4	19.7	3.38	2.87	5.53
Label							
Positive	53.7	49.2	19.5	19.1	3.54	3.15	5.92
Negative	37.4	52.5	13.2	19.6	3.42	2.95	5.63

Abbreviation: HbA_{1c}, hemoglobin A_{1c}.

[a] In the first setup, we rank patients by their model's output in decreasing order, then bin them into 4 categories: top 1%, next 5% (between top 1% and top 6%), next 15% (between top 6% and top 21%), and the remaining 79%. For each bin, we display statistics pertaining to general demographic factors (mean age, fraction of women, fraction of immigrants and time in Canada for immigrants) and socioeconomic factors (race/ethnicity and deprivation marginalization scores of the neighborhood), as well as the mean HbA_{1c}. Means are computed across nonmissing values from patients within each bin. For instance, time in Canada is computed only for immigrants of each model output bin as the value is missing for long-term residents. The second setup evaluates the same variables but when splitting patients according to their label (positive or negative).

Given the incidence rate of 0.2%, the top 1% constituted the high-risk patients, the next 5% were moderate-risk patients, the next 15% were low-risk patients, and the remaining 79% were negligible-risk patients. Analysis of these risk bins reflected the variables and thresholds used by the model to make predictions.

Patients developing diabetes were typically much older (mean age, 53.7 years) than patients who did not develop diabetes (mean age, 37.4 years) within the time frame of our study. Similarly, the very high-risk patients selected by our model had a mean age of 58.3 years. The model evaluated a greater proportion of immigrants considered to be at high risk. Patients at higher risk were more likely to live in neighborhoods with a high concentration of ethnic minority groups and material deprivation, as 4.22 and 3.63 were the mean scores, respectively, for high-risk patients compared with 3.38 and 2.87 for low-risk patients.

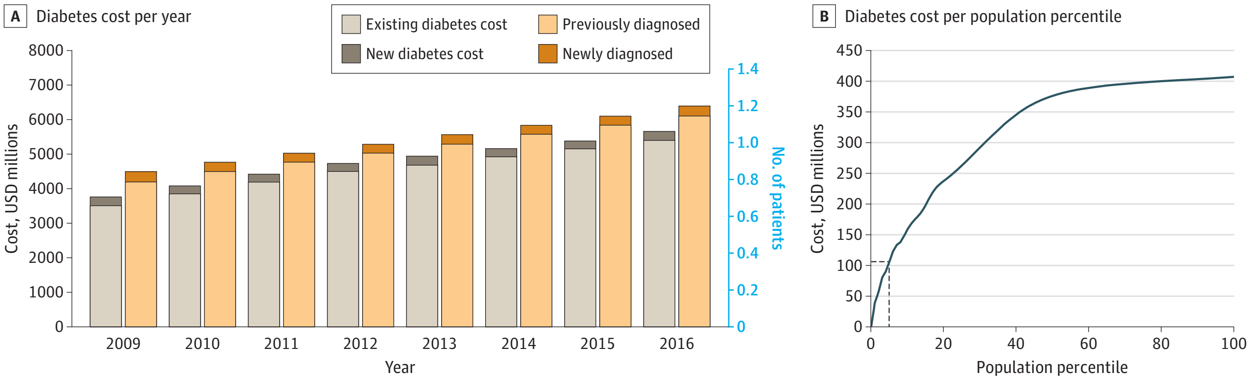
In **Figure 3**, we display an estimation of the total cost of the cohort predicted to develop diabetes in Ontario from 2009 to 2016. Figure 3A represents an estimation of this cohort and the associated cost after scaling our cohort to the entire population of Ontario. Although the number of patients with diabetes is estimated to be 785 000 with an associated cost of $3.5 billion in 2009, these figures increased to 1 144 000 and $5.4 billion, respectively, only 7 years later. The cohort with diabetes grew at an average of 51 800 new patients per year between 2009 and 2016, which added, on average, $242 million per year to the financial burden of diabetes. Moreover, in Figure 3B, the patients who were predicted to be at the highest risk by our model composed a large fraction of the cost: moderate-risk and high-risk patients were 5% of the population but represented 26% of the total diabetes cost.

Further results on the performance of our model are displayed in eFigures 2, 3, and 4 and eTable 3 in the Supplement. We conducted an ablation study over data sources, analyzed feature contribution from each data set, and reported precision and recall curves and AUC results for buffers of 1 year and 3 years, respectively.

Discussion

This decision analytical model study found that accurate prediction of type 2 diabetes onset at the population level 5 years in advance was possible solely from routinely collected administrative health data for the purposes of public health planning and health resource allocation. It was not our goal for this model to be applied in the context of individual patient care. Our model was trained and

Figure 3. Estimation of the Total Cost of the Cohort Predicted to Develop Diabetes in Ontario from 2009-2016



Diabetes cost per year (A) and per population percentile (B) are displayed. The 5% most at-risk patients concentrate 26% of the total cost. USD indicates US dollars.

validated on more than 2 million patients, which, to our knowledge, is one of the largest cohorts for predicting diabetes incidence. Our model showed consistent calibration across sex, immigration status, racial/ethnic and material deprivation, and a low to moderate number of events in the health care history of the patient. The cohort was representative of the whole population of Ontario, which is itself among the most diverse in the world.[50] The model was well calibrated, and its discrimination, although with a slightly different end goal, was competitive with results reported in the literature for other machine learning–based studies that used more granular clinical data from electronic medical records without any modifications to the original test set distribution.[23-25]

Assessing risk in populations is the basis of health system planning and a critical element of diabetes prevention.[51,52] When managing risk in populations, there are critical questions regarding the most efficient usage of resources, and without a comprehensive estimate of risk in populations, strategies can be costly and ineffective. Furthermore, it is widely recognized that the prevention of diabetes is not only influenced by factors at the individual level but must be complemented by whole population approaches, such as food policies and environmental changes.[6] The use of machine learning methods for predicting risk in populations offers an important opportunity to inform resource and policy-level decisions that can change diabetes risk trajectories as well as allow for more efficient targeting of resources within a health system.

The growing burden of diabetes is a challenge faced by other jurisdictions across the globe.[1-3] Continuous risk assessment using the multi-instance approach we proposed could reduce this cost through the targeting of preventive health measures, even more so given the fact that our model did not require additional data collection. Such an approach could be feasible in countries such as the UK, Australia, New Zealand, and the Scandinavian countries, which have large, administrative databases suitable for linkage.[53-57] Furthermore, this approach could also be deployed in populations covered under a singular health insurance system, such as Medicare or private insurers.[58]

Our features not only captured each patient's medical history but also included the social and demographic determinants of health, which are important predictors of a patient's overall risk of developing diabetes and are often missing in clinical data sources.[59-61] Moreover, the calibration of our machine learning model across demographic subgroups suggests that it may be possible to apply it to target-specific population segments with preventive measures (Table 2 and Figure 3). Diabetes prevention strategies can be targeted toward those above a certain risk threshold.[62] Our model results suggest that older patients from the most marginalized neighborhoods in terms of race/ethnicity and material deprivation were at the highest risk and may therefore benefit the most from preventive measures. Given the growing costs associated with the diabetes cohort, our work suggests a quantitative financial incentive toward the direction of preventive measures that consider those at greatest risk, including from a socioeconomic perspective.[59] Because our machine learning model included social determinants of health that are known to contribute to diabetes risk, our population-wide approach to risk assessment may represent a tool for addressing health disparities.[59,63,64]

Strengths and Limitations

Our study approach had several strengths. Owing to the nature of administrative data, such an approach could be applied to other chronic diseases. In 2009, 24.3% of Ontarians were found to be affected by multiple comorbidities.[65] Accurately forecasting other prominent chronic conditions, such as hypertension, could lead to potential considerable reductions in health care costs while also improving the health and well-being of the population. Similar work to create risk prediction models has been done in a primary prevention cohort from New Zealand to determine 5-year cardiovascular disease risk, and research from the UK reinforces that reducing cardiovascular event risk by even 1% would result in both large cost savings and improved population health.[54,66] Moreover, we included a detailed calibration assessment, both overall and within key population subgroups, which suggests that our model did not only have strong discrimination but was well calibrated in a diverse population.[67] Finally, the choice of using a gradient boosting machine model permitted the usage of

Shapley values to enhance explainability.[68] Our proposed approach also had some important limitations. First, there was the potential for misclassification of patients with type 1 diabetes given limitations with the algorithm used in label construction of type 1 and type 2 diabetes.[69,70] Of the roughly 2% to 3% of individuals aged 20 years or younger who tested positive for diabetes, we are uncertain how many were actually diagnosed with type 1 diabetes. However, we chose not to exclude younger patients in our cohort owing to the rising incidence of type 2 diabetes in youths and young adults.[71,72] Second, the input administrative health data were highly heterogeneous: only 23.4% of patients had at least 1 laboratory value, and only the patients older than 65 years had a prescription history. We believe that more consistency and fewer missing values in the input data would improve the model's discrimination. Third, administrative data often does not capture certain features known to be highly predictive of diabetes onset, such as body mass index; however, we achieved competitive performance when our machine learning model was compared to those trained on richer sources of data while allowing for applicability at the population level. Fourth, although we can interpret the model's decisions and the way it splits variables to separate patients into risk score categories, the model strictly captured correlations in the data and not causal pathways. Finally, our model would need to be further validated through prospective studies before deployment.

Conclusions

In this decision analytical model study, we developed and validated a population-level machine learning model to predict the incidence of type 2 diabetes 5 years ahead in a large, contemporary cohort from Ontario, Canada's single-payer health system. Study results suggest that our model had strong discrimination and was robust in calibration across several subgroups, including sex, immigration status, race/ethnicity marginalization, and material deprivation marginalization. Following external and prospective validation, our findings suggest that administrative health data and machine learning may be leveraged for the continuous risk assessment and cost-effective targeting of prevention efforts of type 2 diabetes at the population level with a focus on health equity.

ARTICLE INFORMATION

Accepted for Publication: April 1, 2021.

Published: May 25, 2021. doi:10.1001/jamanetworkopen.2021.11315

Corresponding Author: Laura C. Rosella, PhD, Dalla Lana School of Public Health, University of Toronto, 155 College St, Ste 672, Toronto, ON M5T 3M7, Canada (laura.rosella@utoronto.ca).

Author Affiliations: Layer 6 AI, Toronto, Ontario, Canada (Ravaut, Sadeghi, Leung, Volkovs, Poutanen); Department of Computer Science, University of Toronto, Toronto, Ontario, Canada (Ravaut); Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada (Harish, Kornas, Watson, Rosella); Temerty Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada (Harish, Rosella); Temerty Centre for Artificial Intelligence Research and Education in Medicine, University of Toronto, Toronto, Ontario, Canada (Harish, Rosella); Vector Institute, Toronto, Ontario, Canada (Harish, Rosella); Institute of Clinical Evaluative Sciences (ICES), Toronto, Ontario, Canada (Watson, Rosella); Institute for Better Health, Trillium Health Partners, Mississauga, Ontario, Canada (Rosella).

Author Contributions: Mr Ravaut had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Concept and design: Ravaut, Sadeghi, Leung, Volkovs, Poutanen, Rosella.

Acquisition, analysis, or interpretation of data: Ravaut, Harish, Sadeghi, Leung, Volkovs, Kornas, Watson, Rosella.

Drafting of the manuscript: Ravaut, Harish, Sadeghi, Leung, Rosella.

Critical revision of the manuscript for important intellectual content: All authors.

Statistical analysis: Ravaut, Sadeghi, Leung, Volkovs, Rosella.

Obtained funding: Poutanen, Rosella.

Administrative, technical, or material support: Ravaut, Sadeghi, Kornas, Watson, Poutanen.

Supervision: Sadeghi, Volkovs, Poutanen, Rosella.

Conflict of Interest Disclosures: Mr Harish reported receiving grants from the Canadian Institutes of Health Research (scholarship), the University of Toronto and Province of Ontario (scholarship), and the Vector Institute (scholarship) during the conduct of the study. Dr Rosella reported receiving grants from Canada Research Chairs, the Connaught Global Challenge Award, and the New Frontiers in Research Fund during the conduct of the study. No other disclosures were reported.

Funding/Support: This study was funded by grant FRN 72055702 from the Connaught Global Challenge Award 2018/2019 (Dr Rosella). Dr Rosella was supported by grant FRN 72060091 from the Canada Research Chair in Population Health Analytics. Mr Harish was supported by the Ontario Graduate Scholarship and Canadian Institutes of Health Research Banting and Best Canada Graduate Scholarship-Master's awards and the Vector Institute Postgraduate Affiliate program award.

Role of the Funder/Sponsor: The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Disclaimer: The analyses, conclusions, opinions, and statements expressed herein are solely those of the authors and do not reflect those of the funding or data sources; no endorsement is intended or should be inferred.

Additional Contributions: We thank IMS Borgan Inc for use of their Drug Information Database. For modeling infrastructure, we thank the ICES information technology staff.

Additional Information: This study was supported by ICES, which is funded by an annual grant from the Ontario Ministry of Health and Long-Term Care. Parts of this material are based on data and information compiled and provided by the Ministry of Health and Long-Term Care; Canadian Institutes for Health Information, Immigration, Refugees, and Citizenship Canada; and the Office of the Registrar General. The data set for this study is held securely in coded form at ICES. Although data-sharing agreements prohibit ICES from making the data set publicly available, access may be granted to those who meet prespecified criteria for confidential access, available at https://www.ices.on.ca/DAS. The full data set creation plan is available from the authors upon request. The data for this study was prepared with custom code from ICES using SAS Enterprise software, version 6.1 (SAS Institute Inc). These data were later analyzed with custom code from Layer 6 AI in the Java 8 and Python 3.6 programming languages. The analytic code is available from the authors upon request, understanding that the computer programs may rely upon coding templates or macros that are unique to ICES and this data and thus may require modification.

REFERENCES

1. Shaw JE, Sicree RA, Zimmet PZ. Global estimates of the prevalence of diabetes for 2010 and 2030. *Diabetes Res Clin Pract*. 2010;87(1):4-14. doi:10.1016/j.diabres.2009.10.007

2. Rowley WR, Bezold C, Arikan Y, Byrne E, Krohe S. Diabetes 2030: insights from yesterday, today, and future trends. *Popul Health Manag*. 2017;20(1):6-12. doi:10.1089/pop.2015.0181

3. Bommer C, Heesemann E, Sagalova V, et al. The global economic burden of diabetes in adults aged 20-79 years: a cost-of-illness study. *Lancet Diabetes Endocrinol*. 2017;5(6):423-430. doi:10.1016/S2213-8587(17)30097-9

4. Ali MK, Echouffo-Tcheugui J, Williamson DF. How effective were lifestyle interventions in real-world settings that were modeled on the Diabetes Prevention Program? *Health Aff (Millwood)*. 2012;31(1):67-75. doi:10.1377/hlthaff.2011.1009

5. Dunkley AJ, Bodicoat DH, Greaves CJ, et al. Diabetes prevention in the real world: effectiveness of pragmatic lifestyle interventions for the prevention of type 2 diabetes and of the impact of adherence to guideline recommendations: a systematic review and meta-analysis. *Diabetes Care*. 2014;37(4):922-933. doi:10.2337/dc13-2195

6. Zgibor JC, Songer TJ. External barriers to diabetes care: addressing personal and health systems issues. *Diabetes Spectr*. 2001;14(1):23-28. doi:10.2337/diaspect.14.1.23

7. Secrest AM, Costacou T, Gutelius B, Miller RG, Songer TJ, Orchard TJ. Associations between socioeconomic status and major complications in type 1 diabetes: the Pittsburgh Epidemiology of Diabetes Complication (EDC) Study. *Ann Epidemiol*. 2011;21(5):374-381. doi:10.1016/j.annepidem.2011.02.007

8. Rabi DM, Edwards AL, Southern DA, et al. Association of socio-economic status with diabetes prevalence and utilization of diabetes care services. *BMC Health Serv Res*. 2006;6:124. doi:10.1186/1472-6963-6-124

9. Funakoshi M, Azami Y, Matsumoto H, et al. Socioeconomic status and type 2 diabetes complications among young adult patients in Japan. *PLoS One*. 2017;12(4):e0176087. doi:10.1371/journal.pone.0176087

10. Egede LE, Gebregziabher M, Dismuke CE, et al. Medication nonadherence in diabetes: longitudinal effects on costs and potential cost savings from improvement. *Diabetes Care*. 2012;35(12):2533-2539. doi:10.2337/dc12-0572

11. Booth GL, Zinman B. Diabetes: progress in reducing vascular complications of diabetes. *Nat Rev Endocrinol*. 2014;10(8):451-453. doi:10.1038/nrendo.2014.90

12. Breland JY, McAndrew LM, Gross RL, Leventhal H, Horowitz CR. Challenges to healthy eating for people with diabetes in a low-income, minority neighborhood. *Diabetes Care*. 2013;36(10):2895-2901. doi:10.2337/dc12-1632

13. Mainous AG III, King DE, Garr DR, Pearson WS. Race, rural residence, and control of diabetes and hypertension. *Ann Fam Med*. 2004;2(6):563-568. doi:10.1370/afm.119

14. Booth GL, Shah BR, Austin PC, Hux JE, Luo J, Lok CE. Early specialist care for diabetes: who benefits most? a propensity score-matched cohort study. *Diabet Med*. 2016;33(1):111-118. doi:10.1111/dme.12801

15. Creatore MI, Glazier RH, Moineddin R, et al. Association of neighborhood walkability with change in overweight, obesity, and diabetes. *JAMA*. 2016;315(20):2211-2220. doi:10.1001/jama.2016.5898

16. Shah R, Luo J, Gerstein HC, Booth G. Neighborhood walkability and diabetes-related complications. *Diabetes*. 2018;67(suppl 1). doi:10.2337/db18-309-OR

17. Ali MK, Bullard KM, Gregg EW, Del Rio C. A cascade of care for diabetes in the United States: visualizing the gaps. *Ann Intern Med*. 2014;161(10):681-689. doi:10.7326/M14-0019

18. Polonsky KS. The past 200 years in diabetes. *N Engl J Med*. 2012;367(14):1332-1340. doi:10.1056/NEJMra1110560

19. Cahn A, Shoshan A, Sagiv T, et al. Prediction of progression from pre-diabetes to diabetes: development and validation of a machine learning model. *Diabetes Metab Res Rev*. 2020;36(2):e3252. doi:10.1002/dmrr.3252

20. Nusinovici S, Tham YC, Chak Yan MY, et al. Logistic regression was as good as machine learning for predicting major chronic diseases. *J Clin Epidemiol*. 2020;122:56-69. doi:10.1016/j.jclinepi.2020.03.002

21. Garcia-Carretero R, Vigil-Medina L, Barquero-Perez O, Ramos-Lopez J. Pulse wave velocity and machine learning to predict cardiovascular outcomes in prediabetic and diabetic populations. *J Med Syst*. 2019;44(1):16. doi:10.1007/s10916-019-1479-y

22. Choi BG, Rha SW, Kim SW, Kang JH, Park JY, Noh YK. Machine learning for the prediction of new-onset diabetes mellitus during 5-year follow-up in non-diabetic patients with cardiovascular risks. *Yonsei Med J*. 2019;60(2):191-199. doi:10.3349/ymj.2019.60.2.191

23. Perveen S, Shahbaz M, Keshavjee K, Guergachi A. Prognostic modeling and prevention of diabetes using machine learning technique. *Sci Rep*. 2019;9(1):13805. doi:10.1038/s41598-019-49563-6

24. Nguyen BP, Pham HN, Tran H, et al. Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records. *Comput Methods Programs Biomed*. 2019;182:105055. doi:10.1016/j.cmpb.2019.105055

25. Farran B, AlWotayan R, Alkandari H, Al-Abdulrazzaq D, Channanath A, Thanaraj TA. Use of non-invasive parameters and machine-learning algorithms for predicting future risk of type 2 diabetes: a retrospective cohort study of health data from Kuwait. *Front Endocrinol (Lausanne)*. 2019;10:624. doi:10.3389/fendo.2019.00624

26. Abbas HT, Alic L, Erraguntla M, et al. Predicting long-term type 2 diabetes with support vector machine using oral glucose tolerance test. *PLoS One*. 2019;14(12):e0219636. doi:10.1371/journal.pone.0219636

27. Talaei-Khoei A, Wilson JM. Identifying people at risk of developing type 2 diabetes: a comparison of predictive analytics techniques and predictor variables. *Int J Med Inform*. 2018;119:22-38. doi:10.1016/j.ijmedinf.2018.08.008

28. Pimentel A, Carreiro AV, Ribeiro RT, Gamboa H. Screening diabetes mellitus 2 based on electronic health records using temporal features. *Health Informatics J*. 2018;24(2):194-205. doi:10.1177/1460458216663023

29. Alghamdi M, Al-Mallah M, Keteyian S, Brawner C, Ehrman J, Sakr S. Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: the Henry Ford ExercIse Testing (FIT) project. *PLoS One*. 2017;12(7):e0179805. doi:10.1371/journal.pone.0179805

30. Casanova R, Saldana S, Simpson SL, et al. Prediction of incident diabetes in the Jackson Heart Study using high-dimensional machine learning. *PLoS One*. 2016;11(10):e0163942. doi:10.1371/journal.pone.0163942

31. Anderson JP, Parikh JR, Shenfeld DK, et al. Reverse engineering and evaluation of prediction models for progression to type 2 diabetes: an application of machine learning using electronic health records. *J Diabetes Sci Technol*. 2015;10(1):6-18. doi:10.1177/1932296815620200

32. Ozery-Flato M, Parush N, El-Hay T, et al. Predictive models for type 2 diabetes onset in middle-aged subjects with the metabolic syndrome. *Diabetol Metab Syndr*. 2013;5(1):36. doi:10.1186/1758-5996-5-36

33. Mani S, Chen Y, Elasy T, Clayton W, Denny J. Type 2 diabetes risk forecasting from EMR data using machine learning. *AMIA Annu Symp Proc*. 2012;2012:606-615.

34. Noble D, Mathur R, Dent T, Meads C, Greenhalgh T. Risk models and scores for type 2 diabetes: systematic review. *BMJ*. 2011;343:d7163. doi:10.1136/bmj.d7163

35. Chui T, Flanders J, Anderson T. *Immigration and Ethnocultural Diversity in Canada*. Statistics Canada; 2011.

36. Lipscombe LL, Hux JE. Trends in diabetes prevalence, incidence, and mortality in Ontario, Canada 1995-2005: a population-based study. *Lancet*. 2007;369(9563):750-756. doi:10.1016/S0140-6736(07)60361-4

37. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD statement. *Br J Surg*. 2015;102(3):148-158. doi:10.1002/bjs.9736

38. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *BMJ*. 2007;335(7624):806-808. doi:10.1136/bmj.39335.541782.AD

39. Ravaut M, Sadeghi H, Leung KK, et al. Predicting adverse outcomes due to diabetes complications with machine learning using administrative health data. *NPJ Digit Med*. 2021;4(1):24. doi:10.1038/s41746-021-00394-8

40. Graham SE, Willett JB, Singer JD. Using discrete-time survival analysis to study event occurrence. In: Newsom JT, Jones RN, Hofer SM, eds. *Longitudinal Data Analysis: A Practical Guide for Researchers in Aging, Health, and Social Sciences.* Routledge Taylor & Francis Group; 2012:329-373.

41. Singer JD, Willett JB. It's about time: using discrete-time survival analysis to study duration and the timing of events. *J Educ Behav Stat*. 1993;18(2):155-195. doi:10.3102/10769986018002155

42. Xie H, McHugo G, Drake R, Sengupta A. Using discrete-time survival analysis to examine patterns of remission from substance use disorder among persons with severe mental illness. *Ment Health Serv Res*. 2003;5(1):55-64. doi:10.1023/A:1021759509176

43. Hirdes JP, Poss JW, Caldarelli H, et al. An evaluation of data quality in Canada's Continuing Care Reporting System (CCRS): secondary analyses of Ontario data submitted between 1996 and 2011. *BMC Med Inform Decis Mak*. 2013;13:27. doi:10.1186/1472-6947-13-27

44. Matheson FI, Dunn JR, Smith KLW, Moineddin R, Glazier RH. Élaboration de l'indice de marginalisation canadien: un nouvel outil d'étude des inégalités. *Can J Public Health*. 2012;103(8)(suppl 2):S12-S16. doi:10.1007/BF03403823

45. Chen T, He T. XGBoost: extreme gradient boosting. Published January 15, 2021. Accessed April 28, 2021. https://cran.r-project.org/web/packages/xgboost/vignettes/xgboost.pdf

46. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit*. 1997;30(7):1145-1159. doi:10.1016/S0031-3203(96)00142-2

47. Lundberg SM, Erion GG, Lee S-I. Consistent individualized feature attribution for tree ensembles. Updated March 7, 2019. Accessed April 28, 2021. https://arxiv.org/abs/1802.03888

48. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*. 2020;2(1):56-67. doi:10.1038/s42256-019-0138-9

49. Wodchis WP, Bushmeneva K, Nikitovic M, McKillop I. Guidelines on person-level costing using administrative databases in Ontario. Volume 1. Published May 2013. Accessed April 28, 2021. http://www.sky9games.com/hsprn/uploads/files/Guidelines_on_PersonLevel_Costing_May_2013.pdf

50. Quan H, Smith M, Bartlett-Esquilant G, Johansen H, Tu K, Lix L; Hypertension Outcome and Surveillance Team. Mining administrative health databases to advance medical science: geographical considerations and untapped potential in Canada. *Can J Cardiol*. 2012;28(2):152-154. doi:10.1016/j.cjca.2012.01.005

51. Manuel DG, Rosella LC. Commentary: assessing population (baseline) risk is a cornerstone of population health planning–looking forward to address new challenges. *Int J Epidemiol*. 2010;39(2):380-382. doi:10.1093/ije/dyp373

52. Gruss SM, Nhim K, Gregg E, Bell M, Luman E, Albright A. Public health approaches to type 2 diabetes prevention: the US National Diabetes Prevention Program and beyond. *Curr Diab Rep*. 2019;19(9):78. doi:10.1007/s11892-019-1200-z

53. Herrett E, Gallagher AM, Bhaskaran K, et al. Data resource profile: clinical practice research datalink (CPRD). *Int J Epidemiol*. 2015;44(3):827-836. doi:10.1093/ije/dyv098

54. Mehta S, Jackson R, Pylypchuk R, Poppe K, Wells S, Kerr AJ. Development and validation of alternative cardiovascular risk prediction equations for population health planning: a routine health data linkage study of 1.7 million New Zealanders. *Int J Epidemiol*. 2018;47(5):1571-1584. doi:10.1093/ije/dyy137

55. Clarke P, Leal J, Kelman C, Smith M, Colagiuri S. Estimating the cost of complications of diabetes in Australia using administrative health-care data. *Value Health*. 2008;11(2):199-206. doi:10.1111/j.1524-4733.2007.00228.x

56. Dworzynski P, Aasbrenn M, Rostgaard K, et al. Nationwide prediction of type 2 diabetes comorbidities. *Sci Rep*. 2020;10(1):1776. doi:10.1038/s41598-020-58601-7

57. Ruiz PLD, Stene LC, Bakken IJ, Håberg SE, Birkeland KI, Gulseth HL. Decreasing incidence of pharmacologically and non-pharmacologically treated type 2 diabetes in Norway: a nationwide study. *Diabetologia*. 2018;61(11):2310-2318. doi:10.1007/s00125-018-4681-4

58. Razavian N, Blecker S, Schmidt AM, Smith-McLallen A, Nigam S, Sontag D. Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. *Big Data*. 2015;3(4):277-287. doi:10.1089/big.2015.0020

59. Jack L, Jack NH, Hayes SC. Social determinants of health in minority populations: a call for multidisciplinary approaches to eliminate diabetes-related health disparities. *Diabetes Spectr*. 2012;25(1):9-13. doi:10.2337/diaspect.25.1.9

60. Ludwig J, Sanbonmatsu L, Gennetian L, et al. Neighborhoods, obesity, and diabetes–a randomized social experiment. *N Engl J Med*. 2011;365(16):1509-1519. doi:10.1056/NEJMsa1103216

61. Walker RJ, Gebregziabher M, Martin-Harris B, Egede LE. Relationship between social determinants of health and processes and outcomes in adults with type 2 diabetes: validation of a conceptual framework. *BMC Endocr Disord*. 2014;14:82. doi:10.1186/1472-6823-14-82

62. Saaristo T, Moilanen L, Korpi-Hyövälti E, et al. Lifestyle intervention for prevention of type 2 diabetes in primary health care: one-year follow-up of the Finnish National Diabetes Prevention Program (FIN-D2D). *Diabetes Care*. 2010;33(10):2146-2151. doi:10.2337/dc10-0410

63. Chen IY, Joshi S, Ghassemi M. Treating health disparities with artificial intelligence. *Nat Med*. 2020;26(1):16-17. doi:10.1038/s41591-019-0649-2

64. Rivera LA, Lebenbaum M, Rosella LC. The influence of socioeconomic status on future risk for developing type 2 diabetes in the Canadian population between 2011 and 2022: differential associations by sex. *Int J Equity Health*. 2015;14:101. doi:10.1186/s12939-015-0245-0

65. Rosella L, Kornas K, Huang A, Bornbaum C, Henry D, Wodchis WP. Accumulation of chronic conditions at the time of death increased in Ontario from 1994 to 2013. *Health Aff (Millwood)*. 2018;37(3):464-472. doi:10.1377/hlthaff.2017.1150

66. Barton P, Andronis L, Briggs A, McPherson K, Capewell S. Effectiveness and cost effectiveness of cardiovascular disease prevention in whole populations: modelling study. *BMJ*. 2011;343:d4044. doi:10.1136/bmj.d4044

67. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW; Topic Group 'Evaluating Diagnostic Tests and Prediction Models' of the STRATOS Initiative. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17(1):230. doi:10.1186/s12916-019-1466-7

68. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*. 2019;1(5):206-215. doi:10.1038/s42256-019-0048-x

69. Weisman A, Tu K, Young J, et al. Validation of a type 1 diabetes algorithm using electronic medical records and administrative healthcare data to study the population incidence and prevalence of type 1 diabetes in Ontario, Canada. *BMJ Open Diabetes Res Care*. 2020;8(1):e001224. doi:10.1136/bmjdrc-2020-001224

70. Hux JE, Ivis F, Flintoft V, Bica A. Diabetes in Ontario: determination of prevalence and incidence using a validated administrative data algorithm. *Diabetes Care*. 2002;25(3):512-516. doi:10.2337/diacare.25.3.512

71. Lascar N, Brown J, Pattison H, Barnett AH, Bailey CJ, Bellary S. Type 2 diabetes in adolescents and young adults. *Lancet Diabetes Endocrinol*. 2018;6(1):69-80. doi:10.1016/S2213-8587(17)30186-9

72. Wilmot EG, Davies MJ, Yates T, Benhalima K, Lawrence IG, Khunti K. Type 2 diabetes in younger adults: the emerging UK epidemic. *Postgrad Med J*. 2010;86(1022):711-718. doi:10.1136/pgmj.2010.100917

SUPPLEMENT.
eFigure 1. Overview of Our Approach
eMethods 1. Instance Creation Process
eTable 1. Comparing Electronic Medical Records vs Administrative Health Data
eTable 2. Descriptions of Administrative Health Datasets Used
eMethods 2. Feature Engineering

eMethods 3. Model Development and Evaluation
eFigure 2. Ablation Study
eFigure 3. Feature Contribution
eFigure 4. Model Precision and Recall Curves
eTable 3. Results With Other Buffer Sizes
eTable 4. Comparison Against Logistic Regression Baseline
eReferences