npj | Digital Medicine www.nature.com/npjdigitalmed

ARTICLE OPEN

Predicting adverse outcomes due to diabetes complications with machine learning using administrative health data

Mathieu Ravaut[1,2], Hamed Sadeghi[1], Kin Kwan Leung[1], Maksims Volkovs[1], Kathy Kornas[3], Vinyas Harish[3,4], Tristan Watson[3,5], Gary F. Lewis[6,7], Alanna Weisman[8,9], Tomi Poutanen[1] and Laura Rosella[3,5,10,11,12]✉

Across jurisdictions, government and health insurance providers hold a large amount of data from patient interactions with the healthcare system. We aimed to develop a machine learning-based model for predicting adverse outcomes due to diabetes complications using administrative health data from the single-payer health system in Ontario, Canada. A Gradient Boosting Decision Tree model was trained on data from 1,029,366 patients, validated on 272,864 patients, and tested on 265,406 patients. Discrimination was assessed using the AUC statistic and calibration was assessed visually using calibration plots overall and across population subgroups. Our model predicting three-year risk of adverse outcomes due to diabetes complications (hyper/hypoglycemia, tissue infection, retinopathy, cardiovascular events, amputation) included 700 features from multiple diverse data sources and had strong discrimination (average test AUC = 77.7, range 77.7–77.9). Through the design and validation of a high-performance model to predict diabetes complications adverse outcomes at the population level, we demonstrate the potential of machine learning and administrative health data to inform health planning and healthcare resource allocation for diabetes management.

npj Digital Medicine (2021)4:24 ; https://doi.org/10.1038/s41746-021-00394-8

INTRODUCTION

The global diabetes burden is projected to increase from 380 million people in 2013 to 590 million by 2035[1]. Patients living with diabetes have a higher risk for acute and long-term complications, such as hyperglycemia, nervous system damage, kidney disease, eye damage, and cardiovascular events, than the general population[2,3]. Furthermore, treatments for diabetes complications are a major contributor to the healthcare costs attributable to diabetes, particularly due to hospitalizations and emergency department visits[4,5]. Thus, predicting adverse outcomes due to diabetes complications is important for health system planning.

There is substantial evidence around the prevention of diabetes complications, as landmark studies have demonstrated the importance of controlling hyperglycemia, hypertension, hypercholesterolemia, and smoking cessation[6–10]. However, there are systems-level barriers, which compromise the ability to act upon this evidence and care for populations at scale[11]. These include socioeconomic status (SES) disparities broadly, shown internationally[12–14], the high cost of medications[15,16], access to care and healthcare personnel[17,18], and the built environment[19,20]. Limitations in public health planning and healthcare resource allocation can contribute to "cascades in care" where those who are not receiving care will not meet the targets vital for complications prevention[21].

Many prognostic models have been developed for diabetes complications in the clinical setting[22–24], including more recent applications of machine learning approaches[25–34]. These models generally have made use of rich suites of features (e.g., body mass index, smoking status, biomarkers ranging from commonly ordered lipids to extensive genetic panels) extracted from electronic medical records (EMRs)[25,27,31–33] or clinical trials[28,30]. However, while these models are important for clinical level risk prediction, they are not easily deployed by governments or private health insurance providers at the population level—which is precisely what is needed for addressing the aforementioned systemic barriers to diabetes complications care[35,36]. In contrast, administrative health data (AHD) consists of records collected automatically on diagnoses, procedures, medications, and demographics generated through the provision of health services by governments or other payers[37]. They most commonly do not contain imaging data, doctor text notes, laboratory results, or clinical measures. AHD are high-dimensional, and impossible to explore by clinicians or health systems planners manually. AHD have been long proposed as a tool to assess the quality of a healthcare system[38], but they also represent an enabler for automated analytic approaches to drive the efficiency and effectiveness of primary and secondary health prevention efforts[39,40]. See Supplementary Table 6 for a more detailed comparison between EMRs and AHD. The purpose of this study is to develop a single, large-scale machine learning model for common adverse outcome prediction due to diabetes complications that can be applied on routinely collected AHD for the purposes of public health planning and healthcare resource allocation (Fig. 1). Adverse outcomes are the manifestation of complications in a manner that results in hospital or ambulatory care. It is not our goal for this model to be applied in the context of individual patient care. We base our study on the single-payer health insurance system in Ontario, Canada. Canada has

[1]Layer 6 AI, Toronto, ON, Canada. [2]Department of Computer Science, University of Toronto, Toronto, ON, Canada. [3]Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada. [4]MD/PhD Program, Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada. [5]ICES, Toronto, ON, Canada. [6]Department of Medicine, Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada. [7]Department of Physiology, Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada. [8]Lunenfeld-Tanenbaum Research Institute, Mt. Sinai Hospital, Toronto, ON, Canada. [9]Division of Endocrinology and Metabolism, Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada. [10]Vector Institute, Toronto, ON, Canada. [11]Institute for Better Health, Trillium Health Partners, Mississauga, ON, Canada. [12]Department of Laboratory Medicine & Pathology, Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada. ✉email: laura.rosella@utoronto.ca

Published in partnership with Seoul National University Bundang Hospital

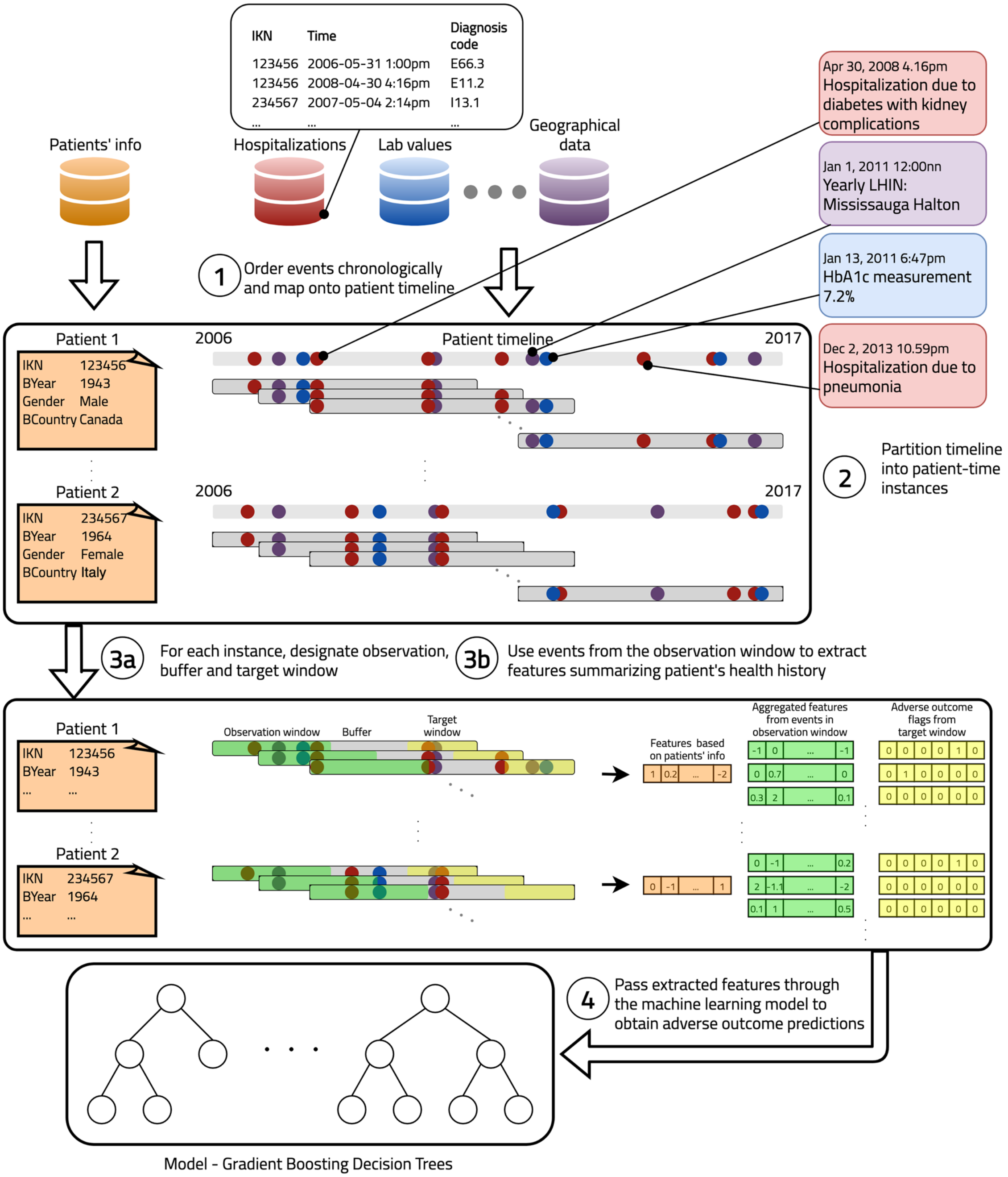

IKN Time Diagnosis code
123456 2006-05-31 1:00pm E66.3
123456 2008-04-30 4:16pm E11.2
234567 2007-05-04 2:14pm I13.1
... ...
Patients' info
Hospitalizations
Lab values
Geographical data
Apr 30, 2008 4.16pm Hospitalization due to diabetes with kidney complications
Jan 1, 2011 12:00nn Yearly LHIN: Mississauga Halton
Jan 13, 2011 6:47pm HbA1c measurement 7.2%
Dec 2, 2013 10.59pm Hospitalization due to pneumonia
1
Order events chronologically and map onto patient timeline
Patient 1
IKN 123456
BYear 1943
Gender Male
BCountry Canada
2006
Patient timeline
2017
Patient 2
IKN 234567
BYear 1964
Gender Female
BCountry Italy
2006
2017
2
Partition timeline into patient-time instances
3a
For each instance, designate observation, buffer and target window
3b
Use events from the observation window to extract features summarizing patient's health history
Patient 1
IKN 123456
BYear 1943
... ...
Observation window
Buffer
Target window
Features based on patients' info
Aggregated features from events in observation window
Adverse outcome flags from target window
1 0.2 ... -2
-1 0 ... -1
0 0.7 ... 0
0.3 2 ... 0.1
0 0 0 0 1 0
0 1 0 0 0 0
0 0 0 0 0 0
Patient 2
IKN 234567
BYear 1964
... ...
0 -1 ... 1
0 -1 ... 0.2
2 -1.1 ... -2
0.1 1 ... 0.5
0 0 0 0 1 0
0 0 0 0 0 0
0 0 0 0 0 0
4
Pass extracted features through the machine learning model to obtain adverse outcome predictions
Model - Gradient Boosting Decision Trees

Fig. 1 Overview of our end-to-end prediction pipeline. (Step 1) Events from multiple administrative health datasets are ordered chronologically and mapped onto patient timeline. (Steps 2 and 3a) Patient timeline is partitioned into patient-time instances and each instance is assigned an observation, buffer and target window. In this study, observation window is 2 years, buffer is 3 years, target window is 3 months and we use nonoverlapping target windows. To illustrate this partitioning, we consider the test period which runs from Jan 1, 2016 to Dec 31, 2016 as an example. In this period a given patient has four instances at Jan 1, 2016, Apr 1, 2016, Jul 1, 2016, and Oct 1, 2016. The first instance has observation window [Jan 1, 2011–Dec 31, 2012], buffer [Jan 1, 2013–Dec 31, 2015], and target window [Jan 1, 2016–Mar 31, 2016]. The second instance has observation window [Apr 1, 2011–Mar 31, 2013], buffer [Apr 1, 2013–Mar 31, 2016], and target window [Apr 1, 2016–Jun 30, 2016], and so on. (Step 3b) Events from the observation window are used to extract features that summarize patient's health history up to the end of the observation window. (Step 4) Extracted features are passed to the machine learning model to generate adverse outcome predictions. The goal of the model is to accurately predict which instances will have an adverse outcome from each of the six complications in the target window.

established some of the most comprehensive administrative health data holdings in the world, covering nearly the total population, in part owing to its universal healthcare system[41]. To ensure broad patient coverage we apply minimal selection criteria, and only require patients to be alive and diagnosed with diabetes at some point in their life. We used a "2-claim" algorithm to flag diabetes in administrative health data[42]. Since this algorithm does not differentiate type 1 and type 2, the resulting cohort is made of patients diagnosed with both types[42]. This results in a large and diverse cohort of over 1.5 million patients with a broad representation of different socio-demographic groups and patterns of interaction with the healthcare system (e.g., frequency of doctor visits, availability of laboratory results etc.). Our results indicate that machine learning can be successfully leveraged to draw insights from administrative health data with minimal restriction, opening up avenues for the deployment of advanced population health management systems to improve health management, promote health equity, and reduce barriers to diabetes care with low per-patient overhead and cost.

RESULTS

Cohort characteristics

We aim to predict three years in advance, whether a patient diagnosed with diabetes will experience a healthcare visit from an adverse outcome due to a diabetes complication within a target three month prediction window. For this study, an adverse outcome is defined as at least one hospitalization or ambulatory usage associated with any diabetes complication during the target prediction window. Note that our task differs from complication incidence prediction itself as multiple adverse outcomes can be associated with the same complication and can occur repeatedly as the complication evolves. Adverse outcomes generally indicate significant negative events during a patient's diabetes progression, and can be both detrimental to quality of life and place considerable cost burden on the healthcare system. Accurate advanced prediction can support preventative measures, and aid with how resources are deployed and managed within a health system. We target the following five complications: severe hyper/hypoglycemia, tissue infection, retinopathy, cardiovascular events (e.g., stroke), and amputation. Specific definitions for each complication's adverse outcomes are provided in the Supplementary Table 5. This set covers most major complications, with the exception of those that have very limited data in our cohort such as kidney failure or erectile dysfunction.

After applying the selection criteria, we obtain a cohort of 1,567,636 patients. The model is trained using the data from 1,029,366 patients, then validated and tested on the distinct sets of 272,864 and 265,406 patients, respectively. Patients in all sets are selected at random. Both validation and testing are done forward in time so all evaluation is out-of-time and out-of-sample. Given the large-scale size of our final cohort, we do not use k-fold cross-validation and stick to a single, fixed validation set, as is common practice with very large datasets[43,44]. We use patient-time instances for all modeling, where each instance represents a view of the patient at a specific point in time. All patients thus have multiple associated instances as we slide the model across time. This simulates the real-life application of advanced population risk assessment systems, that are typically used to continuously monitor patients at regular time intervals. Cohort statistics are summarized in Table 1. We observe that instance incidence rates within the narrow three month window varies significantly across complications between 0.04% (retinopathy) and 1.08% (cardiovascular events). In all cases, the binary classification is severely imbalanced.

For each patient, we consider 11 years of history from January 2006 to December 2016, and aggregate data from multiple administrative health sources such as demographics, outpatient doctor visits, hospitalizations, laboratory tests, etc. Figure 1 outlines our end-to-end pipeline. We first order all events from each source chronologically, and partition this data into patient-time instances. For each instance we then extract features that summarize a patient's health history at that point in time, and pass them to our machine learning model to get adverse outcome predictions.

Model performance

Figure 2 shows Area Under the Receiver Operating Curve (AUC) for predictive performance as well as calibration curves for each complication. We compute the AUC using all instances from the test cohort to measure model performance for the entire test time period from January to December 2016. Given that the incidence rates vary significantly across complications, the calibration curves have correspondingly different ranges. Furthermore, since all incidence rates are very low, we do not compute the Brier score as it was shown to be not well suited for rare events[45]. Calibration curves are computed using 20 bins of identical size, and we plot "observed" and "predicted" probabilities for each bin. We retrain the model five times with random restarts, and report AUC ranges across the restarts. Our model achieves an average AUC of 77.74 (77.7–77.9) over the test set. The best and worst AUCs of 84.4 (84.3–84.5) and 68.9 (68.9–69.2) are obtained for adverse outcomes due to hyper/hypoglycemia and amputation, respectively.

Feature contribution

Figure 3 displays features that contribute the most to model prediction as determined by the mean absolute Shapley values (see Methods for further details)[46]. For each complication we show the top eight most predictive features, and the corresponding administrative health dataset where each feature is derived from. We observe that across all complications the model is leveraging features from multiple datasets to generate predictions. This can be partially attributed to the fact that each individual dataset is sparse, and thus cannot be used exclusively even if it is highly predictive. We also observe that while there are some commonalities, such as the age that is used for all complications, most top features differ for each complication. For instance, diagnosis history is important for hyper/hypoglycemia and retinopathy, while it is absent from the top eight features for the other three complications.

Cost analysis

Due to the high incidence rate and significant number of complications, diabetes imposes one of the heaviest cost burdens on the healthcare system[47,48]. In Ontario, Canada with a population of 14.5 million people in 2019, ambulatory use and hospitalizations due to the diabetes complications considered here alone cost over $3.9 billion per year. This estimate is obtained using the validated costing methodology developed by Wodchis et al.[49] and we analyze cost further in Fig. 4. The costing algorithm gives us the annual patient expenditure for each patient and each healthcare channel (hospitalizations, ambulatory usage, drugs, etc). Since we also have access to each patient's health history, we computed estimations of the cost of each adverse outcome. See Methods for more details on the cost computation.

Figure 4a–c is built by ranking patients by decreasing likelihood of getting an adverse outcome as predicted by the model. Figure 4d shows the proportion of the total cost that is contributed by each of the five complications. Adverse outcomes due to cardiovascular events and tissue infection complications account for the largest proportion of the total cost, jointly contributing over 82%. This is expected since these are the most frequent complications in our cohort with 3.56% and 2.28% test patient incidence rate, respectively.

Figure 4a shows the total annual cost for most at risk patients predicted by our model. For each of the five complications, we sort patients according to the predicted likelihood of adverse outcome, then compute the annual cost for each percentage of patients in this sorted list from top 1–100%. As seen in the figure, the model captures around $440 M (11.1%), $850 M (21.8%), and $1.15B (28.1%) of the total annual cost in the top 1%, 3%, and 5% of predicted patients, respectively. Figure 4c further summarizes statistics for the top 1% of the most at-risk patients predicted by the model. We observe several interesting patterns. First, adverse events due to hyper/hypoglycemia are typically predicted for much younger patients with an average age of 21.8 years old

Table 1. Patient and instance counts across complications and population subgroups.

	Training		Validation		Test	
	(Jan. 2011–Dec. 2014)		(Jan. 2015–Dec. 2015)		(Jan. 2016–Dec. 2016)	
	Patients	Instances	Patients	Instances	Patients	Instances
Full cohort	1,029,366	15,862,818	272,864	1,077,964	265,406	1,046,122
Adverse outcomes[a]						
Hyper/Hypoglycemia	12,015 (1.17%)	16,462 (0.10%)	1116 (0.41%)	1279 (0.12%)	1191 (0.45%)	1396 (0.13%)
Tissue infection	67,200 (6.53%)	92,089 (0.58%)	5946 (2.18%)	6782 (0.63%)	6047 (2.28%)	6898 (0.66%)
Retinopathy	4875 (0.47%)	5902 (0.04%)	429 (0.16%)	482 (0.05%)	386 (0.15%)	418 (0.04%)
Cardiovascular events	105,310 (10.23%)	160,084 (1.01%)	10,030 (3.68%)	12,102 (1.12%)	9456 (3.56%)	11,254 (1.08%)
Amputation	44,560 (4.33%)	54,947 (0.35%)	3552 (1.30%)	3760 (0.35%)	3479 (1.31%)	3728 (0.36%)
Sex						
Male	535,903 (52.06%)	8,250,936 (52.01%)	142,037 (52.05%)	560,863 (52.03%)	137,315 (51.74%)	540,721 (51.69%)
Female	493,463 (47.94%)	7,611,882 (47.99%)	130,827 (47.95%)	517,101 (47.97%)	128,091 (48.26%)	505,401 (48.31%)
Age group[b]						
<20	17,197 (1.67%)	235,540 (1.48%)	3304 (1.21%)	13,160 (1.22%)	2873 (1.08%)	11,381 (1.09%)
20–44	187,359 (18.20%)	2,495,146 (15.73%)	35,783 (13.11%)	142,001 (13.17%)	32,683 (12.31%)	129,044 (12.34%)
45–64	533,188 (51.80%)	7,439,656 (46.90%)	123,393 (45.22%)	490,343 (45.49%)	117,955 (44.44%)	467,611 (44.70%)
65–79	345,604 (33.57%)	4,434,449 (27.96%)	84,375 (30.92%)	332,809 (30.87%)	85,225 (32.11%)	336,113 (32.13%)
80+	108,778 (10.57%)	1,258,027 (7.93%)	26,009 (9.53%)	99,651 (9.24%)	26,670 (10.05%)	101,973 (9.75%)
Immigration status[c]						
Immigrant	184,109 (17.89%)	2,770,495 (17.47%)	51,432 (18.85%)	202,624 (18.80%)	51,488 (19.40%)	202,458 (19.35%)
Long-term resident	845,257 (82.11%)	13,092,323 (82.53%)	221,432 (81.15%)	875,340 (81.20%)	213,918 (80.60%)	843,664 (80.65%)
Material deprivation marginalization score[d]						
1st quintile	180,587 (17.54%)	2,794,415 (17.62%)	48,768 (17.87%)	192,798 (17.89%)	47,502 (17.90%)	187,511 (17.92%)
2nd quintile	189,937 (18.45%)	2,935,301 (18.50%)	50,631 (18.56%)	200,175 (18.57%)	49,734 (18.74%)	196,204 (18.76%)
3rd quintile	199,671 (19.40%)	3,078,263 (19.41%)	52,727 (19.32%)	208,256 (19.31%)	51,794 (19.52%)	204,184 (19.52%)
4th quintile	207,273 (20.14%)	3,190,758 (20.11%)	54,447 (19.95%)	215,020 (19.95%)	52,419 (19.75%)	206,503 (19.74%)
5th quintile	229,932 (22.34%)	3,527,905 (22.24%)	60,433 (22.15%)	238,595 (22.13%)	58,477 (22.03%)	230,181 (22.00%)
Ethnicity marginalization score[d]						
1st quintile	170,001 (16.52%)	2,617,658 (16.50%)	43,868 (16.08%)	173,045 (16.05%)	41,958 (15.81%)	165,122 (15.78%)
2nd quintile	167,063 (16.23%)	2,579,779 (16.26%)	43,594 (15.98%)	172,093 (15.96%)	41,982 (15.82%)	165,391 (15.81%)
3rd quintile	168,147 (16.34%)	2,598,752 (16.38%)	44,389 (16.27%)	175,483 (16.28%)	42,785 (16.12%)	168,682 (16.12%)
4th quintile	192,984 (18.75%)	2,984,352 (18.81%)	51,214 (18.77%)	202,523 (18.79%)	50,409 (18.99%)	198,977 (19.02%)
5th quintile	309,205 (30.04%)	4,746,101 (29.92%)	83,941 (30.76%)	331,700 (30.77%)	82,792 (31.19%)	326,411 (31.20%)

We breakdown the cohort into age, sex, immigration status, material deprivation marginalization, and ethnicity marginalization. The observation window, buffer and target window are 2 years, 3 years, and 3 months, respectively. Target window date ranges are shown in brackets for training, validation, and test sets.

[a] A patient is considered to have an adverse outcome if there is a corresponding event anywhere in the training, validation, or test period. Similarly, an instance is considered to have an adverse outcome if there is a corresponding event in its target window.

[b] Age is computed at the start of training, validation, and test periods for each patient, and at the start of the observation window for each instance.

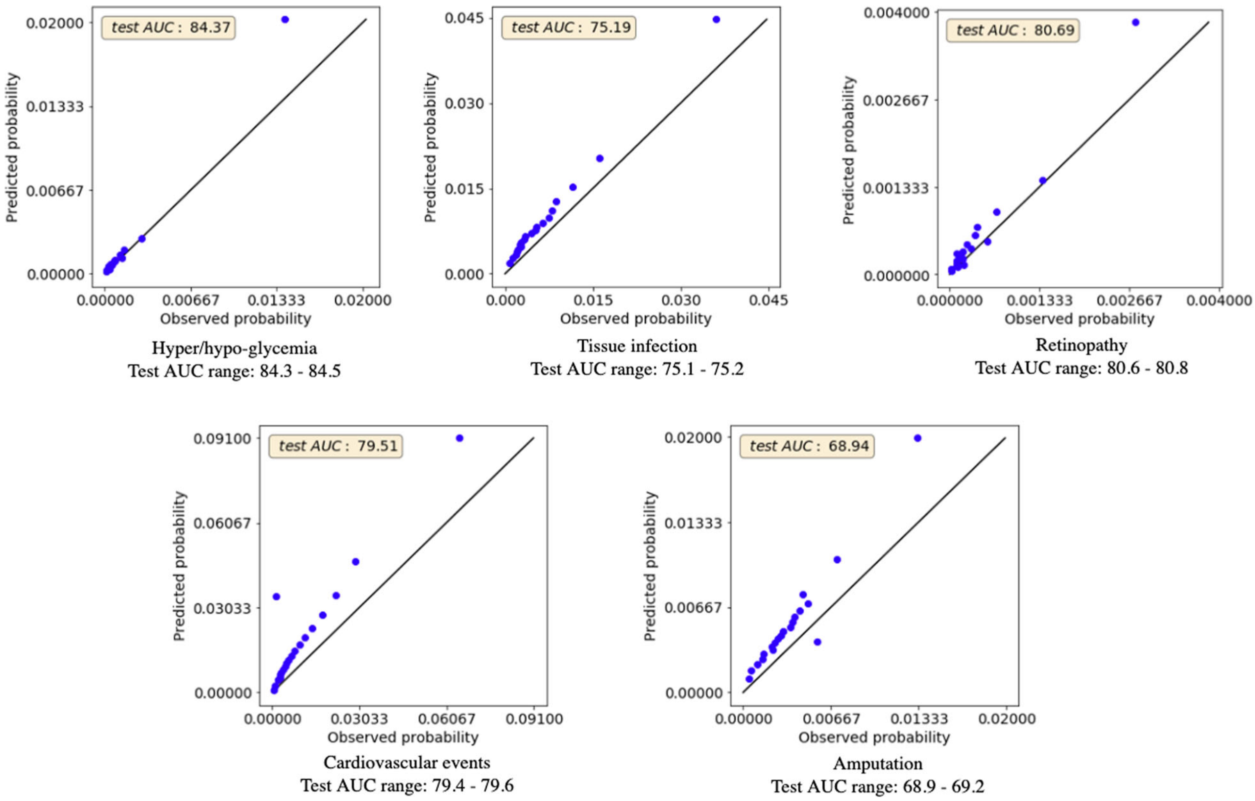
[c] Long-term residents correspond to patients born in Canada or who immigrated to Canada before 1985. Our immigrants cohort contains patients born in 19 different countries, from diverse regions such as South Asia, North Africa, and Eastern Europe. See Supplementary Material Table 1 for the details of the number of immigrants born in each country.

[d] Ethnicity and deprivation marginalization scores quantify the degree of marginalization within each District Administration (DA) according to ethnic concentration and material deprivation. A DA typically encompasses a few hundred inhabitants. These two scores are quintiles ranging from 1 to 5 based on each patient's history from the 2004–2008 period, where five represents a highest degree of marginalization.

compared to 60.9 for the full cohort. Second, the ratio of female versus male patients remains roughly constant across complications except for amputation where the fraction of female patients is higher. Third, the proportion of most at-risk patients that are immigrants is significantly lower for all complications relative to the full cohort. The top 1% most at-risk patients similarly have a lower ethnicity marginalization score for all complications. Finally, HbA1c measurements are higher for patients predicted for hyper/hypoglycemia than other complications. It has been shown that patients with diabetes with an out-of-control glycemia are at higher risk of severe hypoglycemia[50]. We emphasize that these findings in Fig. 4c may not fully be relevant in a clinical setting, but reflect attributes of the training data.

DISCUSSION

This research demonstrated the feasibility of applying machine learning methods to administrative health data for public health planning. Our model can predict the 3-year risk of adverse outcomes due to diabetes complications (hyper/hypoglycemia,



Hyper/hypo-glycemia
Test AUC range: 84.3 - 84.5

Tissue infection
Test AUC range: 75.1 - 75.2

Retinopathy
Test AUC range: 80.6 - 80.8

Cardiovascular events
Test AUC range: 79.4 - 79.6

Amputation
Test AUC range: 68.9 - 69.2

Fig. 2 Test set AUC and calibration curves for all six complications. The average test AUC is 77.74 (77.7–77.9). The model is retrained five times with random restarts. The reported AUC results are averaged across the restarts. Corresponding ranges are also shown, low variance signifies that the results are stable. Calibration curves are computed using 20 bins of identical size, well-calibrated models have curves close to the identity line. Incidence rates vary significantly across complications (from 0.04 to 1.01%) so calibration curves have correspondingly different ranges.

tissue infection, retinopathy, cardiovascular events, and amputation) with a test AUC of 77.7 (range 77.7–77.9, Fig. 2). It was not our goal for this model to be used for individual level patient care. Our model was trained on data from over 1.5 million patients from Ontario, which is among one of the most diverse populations in the world and, to our knowledge, one of the largest prediction modelling studies that takes into account multiple types of diabetes complications[22–34,51–53]. Our model was also well-calibrated and showed good discrimination.

While diabetes complications have been better managed in recent years, they remain a large burden because the incidence of diabetes continues to grow and even in the presence of interventions, not all cases can be prevented[54]. Thus, there is a need to effectively manage diabetes complications at both the individual patient and system levels. This is further emphasized as increasing age and years lived with diabetes have been found to independently predict diabetes morbidity and mortality[55]. Moreover, it has been well established that the complications of diabetes drive costs[4,5]. In Ontario alone, with a population of 14.5 million in 2019, adverse outcomes due to diabetes complications had an annual cost of over $3.9 billion, making diabetes a critical condition that warrants investment into analytic data-driven solutions for health system planning.

Health systems planning, for diabetes and other conditions, requires accurate assessments of population risk[35,36]. From the cost analysis in Fig. 4b, we observe that the top 1% of most at-risk patients predicted by our model account for over $440 M or 11% of the overall annual cost. This increases to over $850 M or nearly 22% of the top 3% of patients' total cost. In contrast, random selection would only capture 1% and 3% of the cost, respectively. Targeting policy interventions (e.g., subsidizing access to fruits and vegetables, community planning to facilitate active transportation) and resource allocation (e.g., incentivizing physicians to have more intensive diabetes follow-up care, either virtually or in-person) to communities projected at highest risk based on our model outputs could help maximize their effectiveness in changing the trajectories of diabetes complications[11,36].

The observed differences in model calibration across complication type may be impacted by the inclusion of both episodic and progressive types of diabetes complications, which by nature have a different epidemiology and trajectory[2]. More specifically, episodic complications, such as tissue infection, can be treated and recur multiple times, whereas progressive complications, such as cardiovascular disease, generally result over an elongated period of time due to chronic damage to the organ system[2]. The overprediction of high-risk individuals could also be due to the relationship between age and years lived with diabetes as key drivers of complications[55]. Finally, it is possible that our over-prediction of those at high risk could be due to the lack of valuable clinical features such as body mass index, smoking status, biomarkers in AHD. At the population level, applications that overpredict would still be appropriate for targeting resources and identifying individuals that would benefit from closer follow-up, including the use of other prediction models which include biomarkers and other individual risk factors.

The analysis of top features in Fig. 3 provides insight into the types of information used by our model to make predictions for each complication. Explainability is a major benefit of decision tree models, and is one of the main reasons why we focus on decision trees for this study. Administrative health databases typically have billions of records spread across multiple datasets making it highly challenging to work with. Moreover, predictive patterns inferred

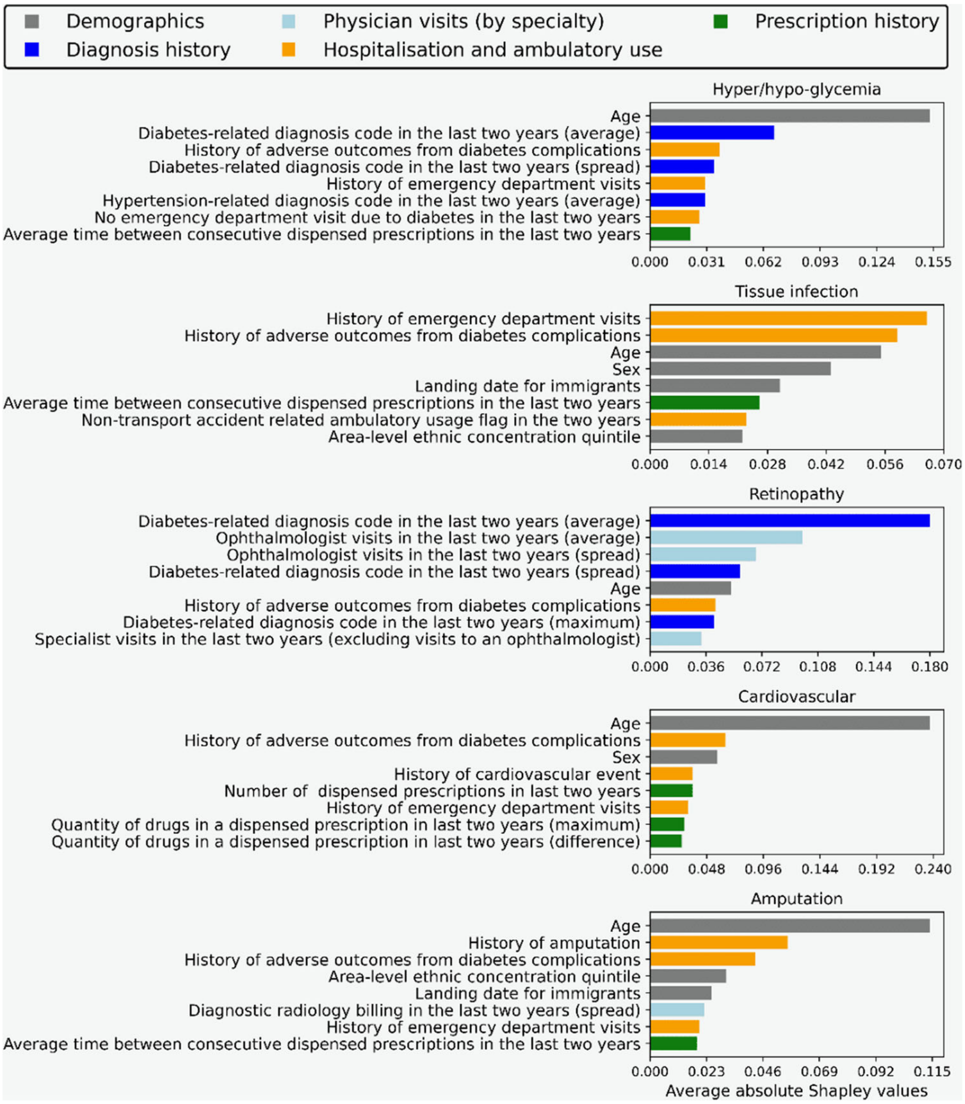
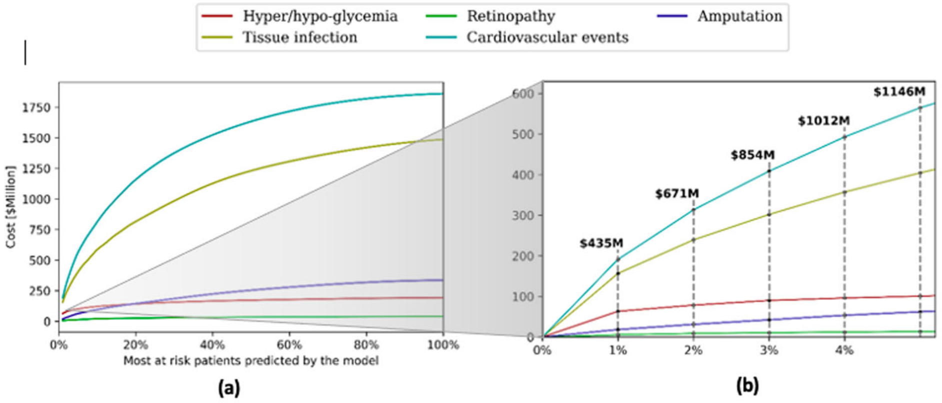


Fig. 3 Top eight features for each complication. For each feature we show the corresponding administrative health dataset where it is derived from, and the magnitude of the contribution to the model. The contribution is measured using the mean absolute Shapley values (see Methods) over a large random sample of 10,000 test instances. The feature contributions here do not represent causal effects, and only indicate correlation with the target predicted adverse outcomes as captured by the model.

by the model at this scale can identify new trends at the population level (or validate existing hypotheses)[56]. In Fig. 3, we observe that socio-demographic factors such as length of stay in Canada for immigrants and ethnic concentration in the area of residence, play an important role in model prediction. We consistently found that features based on immigration status, age/sex, area of residence (particularly census statistics such as neighbourhood-level income, unemployment, ethnic concentration etc.) and other related information appeared within the top 20 most predictive features. This is also observed from Fig. 4d, which shows that there are significantly fewer immigrants and lower ethnicity marginalization in the top 1% of the most at-risk patients predicted by the model as compared to the full cohort. Lower proportion of immigrants aligns with previous studies showing that immigrants have a different diabetes trajectory and are less at risk for these complications[57]. Clinical prediction models generally exclude such types of features and mainly focus on health data for each patient. Our results indicate that the social determinants of health, even at the census level, can be highly predictive for severe health outcomes. Thus the application of a model such as ours for population health planning, which leverages detailed information on the social determinants to allocate resources and plan policies to improve diabetes complications outcomes could offer a data-driven approach to addressing health disparities[58–60].

Our study features a number of important strengths and contributions. The proposed model was developed and tested on a large cohort of over 1.5 million patients with minimal exclusion criteria, capturing virtually all incidences of target adverse outcomes. The cohort is ethnically diverse, with wide representation from across world regions. We demonstrate the applicability of machine learning methods using population data available in multiple jurisdictions around the world. We conducted extensive feature engineering and selection to capture correlations between different AHD sources and target outcomes. The final model has over 700 features from a variety of datasets such as demographics, census information, laboratory results, diagnosis history, physician billing claims, hospitalization and ambulatory usage, prescription medication history and others. Given the nature of administrative data, we believe that our approach could be applied for the



	Age	% Females	% Immigrants	HbA1c	Ethnic marg. score	Material deprivation marg. score
All patients	60.9	48.3	16.6	7.1	3.46	3.22
Top 1% predicted at-risk patients						
Hyper/hypo-glycemia	21.8	47.8	2.7	9.3	3.26	3.24
Tissue infection	51.0	40.3	1.6	8.7	3.25	4.03
Retinopathy	59.8	43.0	7.9	8.5	3.20	3.48
Cardiovascular events	77.9	46.1	6.5	7.5	3.12	3.48
Amputation	59.3	63.5	5.1	7.1	2.92	3.09
Top 5% predicted at-risk patients						
Hyper/hypo-glycemia	37.5	47.5	6.0	8.9	3.46	3.51
Tissue infection	54.2	41.6	2.6	8.2	3.11	3.87
Retinopathy	62.2	45.5	9.9	8.1	3.30	3.47
Cardiovascular events	76.4	43.8	6.4	7.4	3.07	3.42
Amputation	62.2	63.5	5.7	7.1	2.95	3.13
Top 10% predicted at-risk patients						
Hyper/hypo-glycemia	46.2	46.2	6.6	8.7	3.44	3.55
Tissue infection	55.2	41.5	2.7	8.0	3.04	3.77
Retinopathy	62.8	45.9	10.7	7.9	3.35	3.49
Cardiovascular events	76.2	44.0	6.7	7.4	3.07	3.40
Amputation	62.6	68.4	4.7	7.2	2.87	3.12

(c)

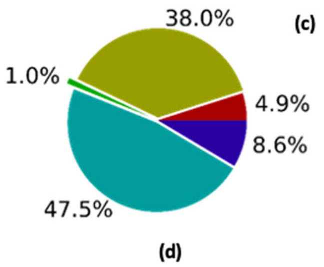


Fig. 4 Adverse outcome cost analysis and high-risk statistics across complications. The total annual cost for adverse outcomes across all five complications is estimated to be ~$3.9B. **a** Annual cost for most at-risk patients predicted by our model. For each complication, we sort all patients by the predicted likelihood of adverse outcome, then compute total cost for each percentage of patients in the sorted list from top 1–100%. **b** A detailed sub-view from panel **a** for top 5% of predicted patients with the total cost across all six complications. **c** Statistics for the top 1% of most at risk patients predicted by our model. We analyze age, sex, immigration status, and HbA1c values, and compare them to the full cohort. All statistics are computed at the end of the observation window for each patient when the model makes its prediction. **d** Breakdown of the contribution to the total annual cost by complication.

forecasting of other chronic diseases at the population level. This is especially important given rising rates of multimorbidity internationally. One study in 2009 found that 24.3% of Ontarians were diagnosed with multiple comorbidities[61]. Since AHD is thought to be the most basic level of information collected by a healthcare system, we believe that our approach to population-level risk prediction would be feasible in other jurisdictions with universal health coverage and databases suitable for linkage such as the Scandinavian countries, United Kingdom, Australia, and New Zealand or within large private insurers in the United States. Finally, modern machine learning approaches are often criticized for lack of interpretability, and are sometimes referred to as black-box models[62]. Using a model based on decision trees enabled us to determine which features are important for prediction, and how they are combined inside the model. This is important for transparency and practical deployment of such systems that clinical and health system specialists need to be audited.

Despite the mentioned strengths, our study also has several important limitations. First, we are limited by the algorithm that we used to flag diabetes and build our cohort[42]. This “2-claim” algorithm has a specificity of 97%, meaning that there are almost no healthy patients in our cohort. However, its 86% sensitivity means that we did not capture all patients diagnosed with diabetes. Moreover, we are working with a joint cohort of patients diagnosed with type 1 and patients diagnosed with type 2. While patients diagnosed with both types share the complications and adverse outcomes we explored in this paper, their diabetes trajectory differs, with type 1 patients typically being diagnosed at a much younger age[2]. We considered using a validated type 1 diabetes algorithm[63] to identify and remove the type 1 subcohort, but with a sensitivity of 80.6% on administrative health data, it would leave out hundreds of patients diagnosed with type 1 in our cohort. We argue that it is preferable for a system-level analysis to predict adverse outcomes of diabetes complications from both patients diagnosed with type 1 and patients diagnosed with type 2 since systems-level barriers are shared between the two populations[11]. We focused on hospitalization and ambulatory care service usage due to diabetes complications. Hence, we do not account for associated adverse events treated in the primary care settings as they could not be identified accurately in our data. In addition, we lacked prescription information for individuals under 65 years old (Ontario's health system provides age-based drug coverage for individuals 65 years and older and those receiving social assistance). If available, they may improve predictive performance even further. More generally, as AHD systems around the world are being increasingly integrated with other data sources such as EHRs, we can believe that our models could be retrained to leverage newly linked databases, with increased discriminative performance. However, as it has been shown with EHRs[64], one must always keep in mind that AHD reflect not only the health state of a patient but also the interactions they had with the healthcare system. Our temporal sliding window framework is robust to the bias of events in the administrative data reflecting past true health states (time of diagnosis is posterior to the time when symptoms started). Our model learns correlations between observed events and target adverse outcomes. Most of these correlations are not causal, and cannot be used to explain why a specific outcome has occurred. Inferring causal relationships would require a different conceptual and analytic framework, which is for future work[65]. Finally, as with other predictive models, external validation with recalibration and prospective validation as well as monitoring for distribution shifts over time would be important to conduct prior to widespread implementation and adoption.

In conclusion, we outline the development and validation of a machine learning model to predict adverse outcomes due to a range of diabetes complications three years ahead at the population level using routinely collected administrative data. We believe that after such models are externally and prospectively validated, public health officials will have a powerful tool for the ongoing risk assessment and cost-effective targeting of prevention efforts and resource allocation related to diabetes complications care at a population-scale.

METHODS

Data source

This study was undertaken using publicly funded administrative health service records linked with population and other data holdings for Ontario, Canada. In Ontario, all residents are eligible for universal health coverage, so AHD covers virtually every Ontarian. Moreover, Ontario is Canada's most populous province and among the most ethnically diverse populations in the world (Supplementary Table 4). In 2016, it had a population of 13.2 million (14.5 million in 2019), of which almost 30% were immigrants[66].

The data were accessed at ICES, which is an independent, nonprofit research institute, whose legal status under Ontario's health information privacy law allows it to collect and analyze healthcare and demographic data, without consent, for health system evaluation and improvement. We analyzed the data within the Health AI Data Analytics Platform (HAIDAP), a platform with high-performance computing resources required for advanced analytics.

The study used multiple diverse data sources including demographic information, census, physician claims, laboratory results, prescription medication history, hospital and ambulatory usage and others. These data sources were linked using the unique encoded identifiers from the Registered Persons Database (RPDB). The RPDB is a central population registry of all residents in Ontario who have ever received a health card number from the province's universal single-payer healthcare system. This registry enables linkage across datasets, and contains basic demographic information, including sex, age, and geographical residence information that we used in our model.

Patients with diabetes were identified using the Ontario Diabetes Database (ODD), a validated registry of Ontario residents diagnosed with diabetes[42]. For each identified patient we extracted data on healthcare utilization and services accessed from the following sources: physician and emergency claims from the Ontario Health Insurance Plan (OHIP), hospitalization history from the Discharge Abstract Database (DAD), emergency services from the National Ambulatory Care Reporting System (NACRS) and prescription medication claims for individuals aged 65 years or above and those receiving social assistance. Diabetes-related laboratory test results were obtained from the Ontario Laboratory Information System (OLIS). The Ontario portion of the Immigration Refugees and Citizenship Canada (IRCC) permanent resident database was used to identify immigration status and country of birth. Neighbourhood-level measures of socioeconomic status, such as ethnicity marginalization score and material deprivation marginalization score shown in Table 1, were obtained using data from the 2001, 2006, and 2011 Canadian censuses (ON-Marg)[67]. Finally, patient deaths that occurred during the observation period were identified from the Office of the Registrar General-Deaths (ORG-D) database. A detailed description of all the data sources can be found in the Supplementary Table 1.

Cohort and exclusion criteria

We used an eleven-year time period from Jan 1, 2006 to Dec 31, 2016 for this study. The Ontario Diabetes Database contained 1,645,089 patients that were flagged as being diagnosed with diabetes at some point in their life and alive on Jan 1, 2006. The algorithm used to identify these patients has demonstrated a sensitivity of 86% and a specificity of 97% compared to physician-assigned diagnoses identified in chart audits[68].

We excluded patients that were not alive as of January 1, 2012 ($n =$ 56,345), and immigrant patients who arrived in Canada later than the last test observation window ($n = 21{,}108$). This resulted in the final cohort of 1,567,636 patients, corresponding to more than 95% of the original cohort. Unlike previous studies that generally apply extensive selection criteria[69], we only excluded forced conditions where the patient is either deceased or not in the system at the time of prediction.

Study design

For each patient in the cohort we partitioned the 11-year time period into patient-time instances that represent a view of the patient at a specific point in time. Each instance was then assigned a 2-year observation window, 3-year buffer, and 3-month target window. We used nonoverlapping target windows so the first instance has an observation window [Jan 1, 2006–Dec 31, 2007], buffer [Jan 1, 2008–Dec 31, 2010], and target window [Jan 1, 2011–Mar 31, 2011]. Similarly, the last instance in our time period has an observation window [Oct 1, 2011–Sept 31, 2013], buffer [Oct 1, 2013–Sept 31, 2016], and target window [Oct 1, 2016–Dec 31, 2016]. Taking into account observation window and buffer time offsets, each patient can have up to 24 instances with nonoverlapping target windows in our 11-year time period. Following the exclusion criteria we removed all instances where the patient is not alive at the end of the target window (n $= 1{,}611{,}222$). We also excluded instances for immigrant patients where the landing date was after the start of the observation window ($n = 259{,}911$). Statistics on the resulting mean number of instances per year can be found in Supplementary Table 8. In addition to this setup, we experimented with the buffer sizes of one and five years. Performance results for these settings are shown in the Supplementary Figs. 1–4, 6, and 7, and the associated feature contributions are shown in Supplementary Figs. 8 and 9.

The health event data from the observation window is used to extract features that summarize a patient's health history at that point in time. We found that the 2-year window was sufficient to obtain the necessary information. There is a sweet spot to find between having an observation window long enough to extract meaningful information, and generating enough instances to train the model with. Indeed, as the observation window grows wider, sliding it through our eleven years time period with three months gaps generates less and less instances, thus reducing the model input size and decreasing performance. Two years was found to be an ideal compromise in our early experiments, and was used thereafter.

The extracted features are then fed to the model that generates instance-level adverse outcome predictions. An instance is considered to have an adverse outcome if there is a corresponding event in its target window. This means that there is at least one hospitalization or ambulatory episode flagged with an ICD-10 code related to one of the diabetes complications during the target window. Adverse outcomes are used here since from AHD, we cannot necessarily ascertain when a complication first became apparent, but rather when an individual sought care for that complication. See Supplementary Table 5 for the list of ICD-10 codes used for adverse outcomes from each complication. Refer to Supplementary Tables 9–11 for statistics on the resulting adverse outcomes, including mean duration and incidence.

An overview of our approach is displayed in the Supplementary Methods, while we delve into details of our end-to-end pipeline in the diagram of Fig. 1. This justifies our choice of using a similar machine learning approach. The multi-instance approach simulates continuous population screening in a practical application. Specifically, we simulate a system where the entire cohort with diabetes is screened every 3 months, and most at-risk patients identified by the model are selected for further analysis and action. The main task is thus to accurately capture all instances that have an adverse outcome in the target prediction window. To achieve this the model must perform well across patients in the cohort and across time for each patient.

Cohort partitioning

We partitioned the cohort into nonoverlapping sets of patients with 1,029,366 patients for model training then 272,864 and 265,406 patients for validation and testing, respectively. Patients in each set are selected at random. All model developments and parameter selections were performed on the training and validation sets, and we report the final model performance on the test set. To reduce the time bias we further partitioned the data in time. For patients in the training set we used instances that have target windows in [Jan 1, 2011–Dec 31, 2014]. Similarly, for validation and test sets we only kept instances with target windows in [Jan 1, 2015–Dec 31, 2015] and [Jan 1, 2016–Dec 31, 2016], respectively. The detailed statistics for each set are summarized in Table 1. Partitioning in time ensures that there is no overlap between the sets so all testing is performed both out-of-sample and out-of-time. This provides a more accurate estimate of performance since in practice, the model would be applied to patients who are newly diagnosed with diabetes (out-of-sample), and all predictions would be done forward in time compared to the training data (out-of-time).

Feature extraction

The main features that we examined were derived from demographics (not changing over time), geographical information, chronic conditions and healthcare utilization history. Stationary features included sex, birth year, immigrant status, and country of origin. Geographical information comprised residence statistics and measures of area-level socioeconomic status from recent census surveys at the level of the first three digits of the postal code. healthcare utilization included information on physician/specialist visits, hospitalization and ambulatory usage and prescription history as seen in Fig. 3. It comprised emergency department visits and laboratory results during the observation window.

We did not perform any preprocessing of continuous variables, except for laboratory results. Laboratory results can be reported in different units, such as mg/L and g/L, and we standardized the unit before doing feature extraction. One-hot encoding was used for all categorical variables, and we discarded categories that appeared with a frequency of less than 1%. Removing infrequent categories significantly reduced the feature size and improved model generalization.

As reported in previous studies, we also found that events in the observation window occur in highly irregular patterns[70,71]. Patients would typically have clusters of activity (multiple doctor/ER visits, laboratory tests, etc.) followed by quiet periods with few events. To summarize these patterns we performed various aggregations over different time intervals within the two year observation window. For time aggregation we counted events in the last month, quarter, 6 months, year, etc. For event aggregation we combined events of the same type such as doctor visits by physician specialty and prescription medication by drug type. This double aggregation resulted in features such as "number of ophthalmologist visits in the last month" and "total quantity of drug X prescribed in the last year". We found such features to be highly informative for adverse outcomes prediction, and many of them appear in the top features as seen in Fig. 3. During feature selection we adopted a greedy approach, and computed multiple combinations of time and event aggregation. These features were then incrementally added into the model and retained only if the validation set performance improved. Note that throughout this process, to prevent any model bias, the test set remained untouched and was only used for the test performance computation of the final model.

In addition to event aggregation, we included other features that summarize a patient's recent medical history. To estimate the recurrence frequency we computed time between consecutive events, as well as time since the most recent event. The goal was to estimate whether certain events are becoming more frequent or occur with a specific time pattern. We particularly focused on the past diabetes and diabetes-related complications as these are generally indicative of future complications. Moreover, we compared each patient's event history with histories from patients in the same sex, age, and immigration status groups. Within-group comparisons can identify "outlier" patients whose progression of condition trajectory deviates significantly from other patients[72,73]. All feature selection here was performed in a similarly greedy fashion by incrementally adding subsets of features to the model. After multiple rounds of feature selection we obtained a set of ~700 features that maximized the validation AUC, and used this set for all further experiments. Supplementary Table 2 provides additional details on feature engineering while Table 3 provides a guide for reading the feature names.

Model development

We trained the Gradient Boosting Decision Trees (GBDT) model implemented in Python in the XGBoost open source library[74,75]. GBDT was chosen due to its ability to handle different feature types (categorical, ordinal, numerical, missing values, etc.) as well as good support for explainability. Besides, XGBoost coupled with extensive feature engineering has consistently shown extremely competitive performance on tabular data. It was used in numerous winning solutions to Kaggle competitions[76] and ACM Recommender Systems Challenges[77,78], and was also proven successful on longitudinal healthcare data[79]. We also experimented with leading deep learning models such as recurrent neural networks including GRU-D[80], multilayer perceptron and Transformers with self-attention[80,81]. However, XGBoost consistently outperformed these models by a relatively large margin. This aligns with previous findings on similar heterogeneous tabular datasets[82,83]. Details on the comparison with logistic regression can be found in the Supplementary Methods and the logistic regression model's performance in Supplementary Table 7.

To handle the multi-class problem of predicting adverse outcomes for multiple complications, we adopted the Cross-Class Relevance Learning (CCRL) method, where class index is appended to the input features and the task is transformed into binary classification[84]. This significantly accelerated training since otherwise we require to optimize a separate XGBoost model for each class, i.e., one for each complication. Our model outputs five risk scores (one per complication) for each instance that is fed to it.

To find good settings of hyperparameters we ran grid search by first specifying ranges for each hyperparameter, and then exhaustively evaluating on points selected from those ranges. After grid search we selected the following settings: a tree depth of 10, learning rate of 0.05, minimum child weight of 50, alpha = 0.3, gamma = 0.1, lambda = 0.0, column sample by tree of 0.6 and column sample by level of 0.6 (relevant XGBoost parameter documentation can be found at: https://xgboost.readthedocs.io/en/latest/parameter). Since incidence rates for adverse outcomes are typically lower than 1%, we undersampled negative instances by a factor of up to 10× to balance the training data[85,86]. After training, the output probabilities from the model were recalibrated using the approach proposed by Pozzolo, et al.[87].

Model evaluation

Given the large-scale size of our final cohort (millions of patients, tens of millions of instances), we do not use k-fold cross-validation and stick to a single, fixed validation set, as is common practice with very large datasets[43,44]. Our validation set is large enough to capture the whole population distribution. We evaluated the test performance of our model on a distinct held out test set using the Area Under the Receiver Operating Curve (AUC) metric. AUC is robust to significant label imbalance[88], and is commonly used for such prediction tasks[89]. Calibration of the model was assessed by plotting calibration curves of the observed versus the predicted probabilities across 20 evenly partitioned bins. The calibration curves and AUC results for each complication are shown in Fig. 2. For practical application, we are particularly interested in the most at-risk patients predicted by the model. As we discussed earlier, such patients can be further analysed for possible preventative measures and resource management. To evaluate performance for the top predicted patients, we also computed the precision (positive predictive value) and recall (sensitivity) shown in the Supplementary Figs. 1 and 2.

We used the Shapley values to find top features that contribute the most to model prediction[46]. To estimate the contribution for each feature we averaged absolute Shapley values over a sample of 100,000 instances selected at random from the test cohort. Different samples were used for each complication to avoid biasing the estimates.

Costing methodology

To evaluate whether the model can capture a significant portion of the cost associated with treating adverse outcomes[90], we computed annual cost for most at-risk patients predicted by the model. Diabetes and its related complications place a significant cost burden on the healthcare system. Continuous population screening and early detection can lead to significant cost savings through preventative measures and resource planning. However, this would only be possible if the model can accurately predict the costly outcomes, meaning that it can make higher predictions on instances with costly adverse outcomes due to diabetes complications than on negative instances. The cost is computed by first applying the costing algorithm[49] to estimate total annual healthcare expenditure by category (hospitalizations, prescriptions, etc.) for each patient. The costing algorithm follows a bottom-up approach for ambulatory care to get person-level healthcare expenditure per year and per category of healthcare utilization by mapping the utilization data with cost information. For inpatient hospitalizations, emergency department visits and same day surgery costs, the algorithm estimates costs based on average provincial costs for these procedures weighted by the resource intensity in a given care setting. Utilization data is directly available through the administrative databases leveraged in this study. Cost information is estimated in the algorithm based on amounts billed to the Ministry of Health and Long Term Care (MOHLTC). We used this costing algorithm off-the-shelf (without any tweaking) on our cohort. This algorithm has been previously validated and is further described elsewhere[47,91]. From the category estimates, we then isolated the portion of the cost attributed to adverse outcomes by isolating cost from the relevant hospitalizations and ambulatory usage. Finally, we sorted all patients according to the model predicted probability of adverse outcome, and computed cumulative cost for each percentages of this sorted list. Cost in one percentile is just the sum of costs of all patients in this percentile. Results for this analysis are shown in Fig. 4.

Ethics

This study obtained ethics approval from the Research Ethics Board at the University of Toronto (Protocol # 37650).

Reporting summary

Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

DATA AVAILABILITY

The dataset for this study is held securely in coded from at ICES. While data sharing agreements prohibit ICES from making the dataset publicly available, access may be granted to those who meet prespecified criteria for confidential access, available at www.ices.on.ca/DAS. The full dataset creation plan is available from the authors upon request.

CODE AVAILABILITY

The data for this study was prepared with custom code from ICES using the SAS Enterprise v6.1 software. This 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.

Received: 18 August 2020; Accepted: 11 January 2021;
Published online: 12 February 2021

REFERENCES

1. Guariguata, L. et al. Global estimates of diabetes prevalence for 2013 and projections for 2035. *Diabetes Res. Clin. Pract.* **103**, 137–149 (2014).
2. Deshpande, A. D., Harris-Hayes, M. & Schootman, M. Epidemiology of diabetes and diabetes-related complications. *Phys. Ther.* **88**, 1254–1264 (2008).
3. Harding, J. L., Pavkov, M. E., Magliano, D. J., Shaw, J. E. & Gregg, E. W. Global trends in diabetes complications: a review of current evidence. *Diabetologia* **62**, 3–16 (2019).
4. Caro, J. J., Ward, A. J. & O'Brien, J. A. Lifetime costs of complications resulting from type 2 diabetes in the U.S. *Diabetes Care* **25**, 476–481 (2002).
5. Hazel-Fernandez, L. et al. Relationship of diabetes complications severity to healthcare utilization and costs among Medicare Advantage beneficiaries. *Am. J. Manag. Care* **21**, e62–e70 (2015).
6. Diabetes Control and Complications Trial Research Group. et al. The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. *N. Engl. J. Med.* **329**, 977–986 (1993).
7. Turner, R. et al. Tight blood pressure control and risk of macrovascular and microvascular complications in type 2 diabetes: UKPDS 38. UK Prospective Diabetes Study Group. *BMJ* **317**, 703–713 (1998).
8. Colhoun, H. M. et al. Primary prevention of cardiovascular disease with atorvastatin in type 2 diabetes in the Collaborative Atorvastatin Diabetes Study (CARDS): multicentre randomised placebo-controlled trial. *Lancet* **364**, 685–696 (2004).
9. Gaede, P. et al. Multifactorial intervention and cardiovascular disease in patients with type 2 diabetes. *N. Engl. J. Med.* **348**, 383–393 (2003).
10. An, Pan, Yeli, Wang, Mohammad, Talaei & Hu Frank, B. Relation of smoking with total mortality and cardiovascular events among patients with diabetes mellitus. *Circulation* **132**, 1795–1804 (2015).
11. Zgibor, J. C. & Songer, T. J. External barriers to diabetes care: addressing personal and health systems issues. *Diabetes Spectr.* **14**, 23–28 (2001).
12. Secrest, A. M. et al. Associations between socioeconomic status and major complications in type 1 diabetes: the Pittsburgh epidemiology of diabetes complication (EDC) Study. *Ann. Epidemiol.* **21**, 374–381 (2011).
13. Funakoshi, M. et al. Socioeconomic status and type 2 diabetes complications among young adult patients in Japan. *PLoS ONE* **12**, e0176087 (2017).
14. Rabi, D. M. et al. Association of socio-economic status with diabetes prevalence and utilization of diabetes care services. *BMC Health Serv. Res.* **6**, 124 (2006).
15. Egede, L. E. et al. Medication nonadherence in diabetes: longitudinal effects on costs and potential cost savings from improvement. *Diabetes Care* **35**, 2533–2539 (2012).
16. Booth, G. L. & Zinman, B. Diabetes: progress in reducing vascular complications of diabetes. *Nat. Rev. Endocrinol.* **10**, 451–453 (2014).
17. Mainous, A. G. 3rd, King, D. E., Garr, D. R. & Pearson, W. S. Race, rural residence, and control of diabetes and hypertension. *Ann. Fam. Med.* **2**, 563–568 (2004).
18. Booth, G. L. et al. Early specialist care for diabetes: who benefits most? A propensity score-matched cohort study. *Diabet. Med.* **33**, 111–118 (2016).
19. Creatore, M. I. et al. Association of neighborhood walkability with change in overweight, obesity, and diabetes. *JAMA* **315**, 2211–2220 (2016).
20. Shah, R., Luo, J., Gerstein, H. C. & Booth, G. Neighborhood walkability and diabetes-related complications. *Diabetes* **67**, Supplement 1 (2018).
21. Ali, M. K., Bullard, K. M., Gregg, E. W. & Del Rio, C. A cascade of care for diabetes in the United States: visualizing the gaps. *Ann. Intern. Med.* **161**, 681–689 (2014).
22. Selby, J. V., Karter, A. J., Ackerson, L. M., Ferrara, A. & Liu, J. Developing a prediction rule from automated clinical databases to identify high-risk patients in a large population with diabetes. *Diabetes Care* **24**, 1547–1555 (2001).
23. Pagano, E. et al. Prediction of mortality and macrovascular complications in type 2 diabetes: validation of the UKPDS Outcomes Model in the Casale Monferrato Survey, Italy. *Diabetologia* **56**, 1726–1734 (2013).
24. Parrinello, C. M. et al. Risk prediction of major complications in individuals with diabetes: the Atherosclerosis Risk in Communities Study. *Diabetes Obes. Metab.* **18**, 899–906 (2016).
25. Aminian, A. et al. Predicting 10-year risk of end-organ complications of type 2 diabetes with and without metabolic surgery: a machine learning approach. *Diabetes Care* **43**, 852–859 (2020).
26. Dworzynski, P. et al. Nationwide prediction of type 2 diabetes comorbidities. *Sci. Rep.* **10**, 1776 (2020).
27. Song, X. et al. Longitudinal risk prediction of chronic kidney disease in diabetic patients using a temporal-enhanced gradient boosting machine: retrospective cohort study. *JMIR Med. Inf.* **8**, e15510 (2020).
28. Segar, M. W. et al. Machine learning to predict the risk of incident heart failure hospitalization among patients with diabetes: the WATCH-DM risk score. *Diabetes Care* **42**, 2298–2306 (2019).
29. Alaa, A. M., Bolton, T., Di Angelantonio, E., Rudd, J. H. F. & van der Schaar, M. Cardiovascular disease risk prediction using automated machine learning: a

prospective study of 423,604 UK Biobank participants. *PLoS ONE* **14**, e0213653 (2019).

30. Rodriguez-Romero, V. et al. Prediction of nephropathy in type 2 diabetes: an analysis of the ACCORD trial applying machine learning techniques. *Clin. Transl. Sci.* **12**, 519–528 (2019).
31. Makino, M. et al. Artificial intelligence predicts the progression of diabetic kidney disease using big data machine learning. *Sci. Rep.* **9**, 11862 (2019).
32. Brisimi, T. S., Xu, T., Wang, T., Dai, W. & Paschalidis, I. C. Predicting diabetes-related hospitalizations based on electronic health records. *Stat. Methods Med. Res.* **28**, 3667–3682 (2019).
33. Dagliati, A. et al. Machine learning methods to predict diabetes complications. *J. Diabetes Sci. Technol.* **12**, 295–302 (2018).
34. Kazemi, M., Moghimbeigi, A., Kiani, J., Mahjub, H. & Faradmal, J. Diabetic peripheral neuropathy class prediction by multicategory support vector machine model: a cross-sectional study. *Epidemiol. Health* **38**, e2016011 (2016).
35. Manuel, D. G. & Rosella, L. C. Commentary: assessing population (baseline) risk is a cornerstone of population health planning-looking forward to address new challenges. *Int. J. Epidemiol.* **39**, 380–382 (2010).
36. Gruss, S. M. et al. Public health approaches to type 2 diabetes prevention: the US National Diabetes Prevention Program and Beyond. *Curr. Diab. Rep.* **19**, 78 (2019).
37. Virnig, B. A. & McBean, M. Administrative data for public health surveillance and planning. *Annu. Rev. Public Health* **22**, 213–230 (2001).
38. Iezzoni, L. I. Assessing quality using administrative data. *Ann. Intern. Med.* **127**, 666–674 (1997).
39. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56 (2019).
40. Panch, T., Pearson-Stuttard, J., Greaves, F. & Atun, R. Artificial intelligence: opportunities and risks for public health. *Lancet Digital Health* **1**, e13–e14 (2019).
41. Quan, H. et al. Mining administrative health databases to advance medical science: geographical considerations and untapped potential in Canada. *Can. J. Cardiol.* **28**, 152–154 (2012).
42. Hux, J. E., Ivis, F., Flintoft, V. & Bica, A. Diabetes in Ontario: determination of prevalence and incidence using a validated administrative data algorithm. *Diabetes Care* **25**, 512–516 (2002).
43. Deng, J. et al. ImageNet: a large-scale hierarchical image database. in *Proc. 2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255 https://arxiv.org/pdf/1804.07461.pdf (2009).
44. Wang, A. et al. GLUE: a multi-task benchmark and analysis platform for natural language understanding. *arXiv* (2018).
45. Assel, M., Sjoberg, D. D. & Vickers, A. J. The Brier score does not evaluate the clinical utility of diagnostic tests or prediction models. *Diagn. Progn. Res.* **1**, 19 (2017).
46. Lundberg, S. M. et al. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, 56–67 (2020).
47. Rosella, L. C. et al. Impact of diabetes on healthcare costs in a population-based cohort: a cost analysis. *Diabet. Med.* **33**, 395–403 (2016).
48. Williams, R., Van Gaal, L. & Lucioni, C., CODE-2 Advisory Board. Assessing the impact of complications on the costs of Type II diabetes. *Diabetologia* **45**, S13–S17 (2002).
49. Wodchis, W. P., Bushmeneva, K., Nikitovic, M. & McKillop, I. *Guidelines on Person-Level Costing Using Administrative Databases in Ontario*. https://tspace.library.utoronto.ca/bitstream/1807/87373/1/Wodchis%20et%20al_2013_Guidelines%20on%20Person-Level%20Costing.pdf (2013).
50. Lipska, K. J. et al. HbA1c and risk of severe hypoglycemia in type 2 diabetes: the Diabetes and Aging Study. *Diabetes Care* **36**, 3535–3542 (2013).
51. Perveen, S., Shahbaz, M., Ansari, M. S., Keshavjee, K. & Guergachi, A. A hybrid approach for modeling type 2 diabetes mellitus progression. *Front. Genet.* **10**, 1076 (2019).
52. Chen, T. et al. Prediction of cardiovascular outcomes with machine learning techniques: application to the Cardiovascular Outcomes in Renal Atherosclerotic Lesions (CORAL) study. *Int. J. Nephrol. Renovasc. Dis.* **12**, 49–58 (2019).
53. 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.* **44**, 16 (2019).
54. Gregg, E. W., Sattar, N. & Ali, M. K. The changing face of diabetes complications. *Lancet Diabetes Endocrinol.* **4**, 537–547 (2016).
55. Huang, E. S. et al. Rates of complications and mortality in older patients with diabetes mellitus: the diabetes and aging study. *JAMA Intern. Med.* **174**, 251–258 (2014).
56. Mehta, S. et al. 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.* **47**, 1571–1584 (2018).
57. Shah, B. R. Diabetes in visible minority populations in Ontario. *Healthc. Q* **16**, 14–17 (2013).
58. Chen, I. Y., Joshi, S. & Ghassemi, M. Treating health disparities with artificial intelligence. *Nat. Med.* **26**, 16–17 (2020).
59. Jack, L., Jack, N. H. & Hayes, S. C. Social determinants of health in minority populations: a call for multidisciplinary approaches to eliminate diabetes-related health disparities. *Diabetes Spectr.* **25**, 9–13 (2012).
60. Rivera, L. A., Lebenbaum, M. & Rosella, L. C. 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* **14**, 101 (2015).
61. Rosella, L. et al. Accumulation of chronic conditions at the time of death increased in Ontario from 1994 to 2013. *Health Aff.* **37**, 464–472 (2018).
62. Gilpin, L. H. et al. Explaining explanations: an overview of interpretability of machine learning. *arXiv*, https://arxiv.org/pdf/1806.00069.pdf (2018).
63. Weisman, A. 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. *Can. BMJ Open Diabetes Res. Care* **8**, e001224 (2020).
64. Agniel, D., Kohane, I. S. & Weber, G. M. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *BMJ* **361**, k1479 (2018).
65. Hernán, M. A., Hsu, J. & Healy, B. A second chance to get causal inference right: a classification of data science tasks. *Chance* **32**, 42–49 (2019).
66. Chui, T., Flanders, J. & Anderson, T. *Immigration and Ethnocultural Diversity in Canada—National Household Survey*. https://www12.statcan.gc.ca/nhs-enm/2011/as-sa/99-010-x/99-010-x2011001-eng.pdf (2011).
67. Matheson, F. I., Dunn, J. R., Smith, K. L. W., Moineddin, R. & Glazier, R. H. Building the Canadian marginalization index: a new tool for studying inequalities. *Can. J. Public Health* **103**, S12–S16 (2012).
68. Lipscombe, L. L. et al. Identifying diabetes cases from administrative data: a population-based validation study. *BMC Health Serv. Res.* **18**, 316 (2018).
69. Perveen, S., Shahbaz, M., Keshavjee, K. & Guergachi, A. Prognostic modeling and prevention of diabetes using machine learning technique. *Sci. Rep.* **9**, 13805 (2019).
70. Rajkomar, A. et al. Scalable and accurate deep learning with electronic health records. *npj Digital Med.* **1**, 18 (2018).
71. Razavian, N. et al. Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. *Big Data* **3**, 277–287 (2015).
72. Tomašev, N. et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature* **572**, 116–119 (2019).
73. Parikh, R. B. et al. Machine learning approaches to predict 6-month mortality among patients with cancer. *JAMA Netw. Open* **2**, e1915997 (2019).
74. Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **29**, 1189–1232 (2001).
75. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. *arXiv*, https://arxiv.org/pdf/1603.02754.pdf (2016).
76. Bojer, C. & Meldgaard, J. Learnings from Kaggle's Forecasting Competitions. *arXiv*, https://arxiv.org/ftp/arxiv/papers/2009/2009.07701.pdf (2020).
77. Volkovs, M. et al. Two-stage model for automatic playlist continuation at scale. in *Proc. ACM Recommender Systems Challenge* 1–6 (Association for Computing Machinery, 2018).
78. Volkovs, M., Yu, G. W. & Poutanen, T. Content-based neighbor models for cold start in recommender systems. in *Proc. Recommender Systems Challenge* 1–6 (Association for Computing Machinery, 2017).
79. Yelin, I. et al. Personal clinical history predicts antibiotic resistance of urinary tract infections. *Nat. Med.* **25**, 1143–1152 (2019).
80. Che, Z., Purushotham, S., Cho, K., Sontag, D. & Liu, Y. Recurrent neural networks for multivariate time series with missing values. *Sci. Rep.* **8**, 6085 (2018).
81. Vaswani, A. et al. in *Advances in Neural Information Processing Systems* (eds. Guyon, I. et al.) 5998–6008 (Curran Associates, Inc., 2017).
82. Shavitt, I. & Segal, E. Regularization learning networks: deep learning for tabular datasets. *arXiv*, https://papers.nips.cc/paper/2018/file/500e75a036dc2d7d2fec5da1b71d36cc-Paper.pdf (2018).
83. Brown, I. & Mues, C. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Syst. Appl.* **39**, 3446–3453 (2012).
84. Ma, J., Gorti, S. K., Volkovs, M., Stanevich, I. & Yu, G. Cross-class relevance learning for temporal concept localization. *arXiv*, https://arxiv.org/pdf/1911.08548.pdf (2019).
85. Ling, C. X. & Li, C. Data mining for direct marketing: problems and solutions. in *Proc. Fourth International Conference on Knowledge Discovery and Data Mining* 73–79 (AAAI Press, 1998).
86. Akbani, R., Kwek, S. & Japkowicz, N. in *Machine Learning: ECML* 39–50 (Springer Berlin Heidelberg, 2004).
87. Pozzolo, A. D., Caelen, O., Johnson, R. A. & Bontempi, G. Calibrating probability with undersampling for unbalanced classification. in *2015 IEEE Symposium Series on Computational Intelligence* 159–166 (2015).

88. Bradley, A. P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recogn.* **30**, 1145–1159 (1997).
89. Swets, J. A. Measuring the accuracy of diagnostic systems. *Science* **240**, 1285–1293 (1988).
90. Doucet, G. & Beatty, M. The cost of diabetes in Canada: the economic Tsunami. *Can. J. Diabetes* **34**, 27–29 (2010).
91. Wodchis, W. P., Austin, P. C. & Henry, D. A. A 3-year study of high-cost users of health care. *CMAJ* **188**, 182–188 (2016).

ACKNOWLEDGEMENTS

This study was supported by ICES, which is funded by an annual grant from the Ontario Ministry of Health and Long-Term Care (MOHLTC). Parts of this material are based on data and information compiled and provided by MOHLTC, Canadian Institutes for Health Information (CIHI), Immigration, Refugees, and Citizenship Canada (IRCC) and the Office of the Registrar General (ORG). The analyses, conclusions, opinions, and statements reported 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. We thank IMS Borgan Inc. for use of their Drug Information Database. For modeling infrastructure, we thank ICES IT staff. This study was funded by the Connaught Global Challenge Award 2018/19. L.R. is supported by a Canada Research Chair in Population Health Analytics. V.H. is supported by the Ontario Graduate Scholarship and Canadian Institutes of Health Research Banting and Best Canada Graduate Scholarship-Master's awards. M.R. was supported by a MITACS Accelerate Scholarship during his internship at Layer 6 AI prior to joining as a full-time employee.

AUTHOR CONTRIBUTIONS

M.R. and L.R. planned the study. T.W. prepared the cohort. M.R. analyzed the data with contributions from H.S., K.L., M.V., and V.H. M.R., L.R., H.S., K.L., M.V., and K.K. wrote the first draft of the manuscript. V.H., G.L., A.W., and T.P. contributed important revisions to the manuscript. All authors contributed to data interpretation, critically reviewed, and approved the final manuscript.

COMPETING INTERESTS

M.R., H.S., K.L., M.V., and T.P. are full-time employees of Layer 6 AI, co-founded by M.V. and T.P., owned by Toronto-Dominion Bank. V.H., K.K., and T.W. are employed at the Dalla Lana School of Public Health. The employers of the authors had no role in the design or funding of this research.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-021-00394-8.

Correspondence and requests for materials should be addressed to L.R.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.