cmajOPEN

Research

Developing a case definition for type 1 diabetes mellitus in a primary care electronic medical record database: an exploratory study

B. Cord Lethebe MSc, Tyler Williamson PhD, Stephanie Garies MPH, Kerry McBrien MD MPH, Charles Leduc MD MSc, Sonia Butalia MD MSc, Boglarka Soos MMath, Marta Shaw PhD, Neil Drummond PhD

Abstract

Background: Identifying cases of disease in primary care electronic medical records (EMRs) is important for surveillance, research, quality improvement and clinical care. We aimed to develop and validate a case definition for type 1 diabetes mellitus using EMRs.

Methods: For this exploratory study, we used EMR data from the Southern Alberta Primary Care Network within the Canadian Primary Care Sentinel Surveillance Network (CPCSSN), for the period 2008 to 2016. For patients identified as having diabetes mellitus according to the existing CPCSSN case definition, we asked family physicians to confirm the diabetes subtype, to create the reference standard. We used 3 decision-tree classification algorithms and least absolute shrinkage and selection operator logistic regression to identify variables that correctly distinguished between type 1 and type 2 diabetes cases.

Results: We identified a total of 1309 people with type 1 or type 2 diabetes, 110 of whom were confirmed by their physicians as having type 1 diabetes. Two machine learning algorithms were useful in identifying these cases in the EMRs. The first algorithm used "type 1" text words or age less than 22 years at time of initial diabetes diagnosis; this algorithm had sensitivity 42.7% (95% confidence interval [CI] 33.5%–52.5%), specificity 99.3% (95% CI 98.6%–99.7%), positive predictive value 85.5% (95% CI 72.8%–93.1%) and negative predictive value 94.9% (95% CI 93.5%–96.1%). The second algorithm used a combination of free-text terms, insulin prescriptions and age; it had sensitivity 87.3% (95% CI 79.2%–92.6%), specificity 85.4% (95% CI 83.2%–87.3%), positive predictive value 35.6% (95% CI 29.9%–41.6%) and negative predictive value 98.6% (95% CI 97.7%–99.2%).

Interpretation: We used machine learning to develop and validate 2 case definitions that achieve different goals in distinguishing between type 1 and type 2 diabetes in CPCSSN data. Further validation and testing with a larger and more diverse sample are recommended.

In Ontario over the period 1995 to 2005, diabetes had a prevalence of 8.8% and an annual incidence of 8.2 per 1000.[1] More recently, the Public Health Agency of Canada reported an age-adjusted overall population prevalence of 7.8% in 2013/14,[2] whereas Diabetes Canada reported an estimated 9.3% prevalence in 2015.[3] However, these rates aggregate all forms of diabetes and, importantly, do not differentiate between type 1 and type 2 diabetes mellitus. The management of diabetes in Canada, including management of the type 1 form, has shifted to a more interdisciplinary, team-based, integrated approach, based on implementation of the Chronic Care Model.[4] The ability to accurately distinguish between type 1 and type 2 diabetes is important for clinical quality improvement, given the difference in management approaches between the 2 conditions, as well as for health outcomes research and pragmatic trials.

Existing validated case definitions for diabetes include those developed by Clottey and colleagues,[5] Hux and colleagues,[6] Amed and colleagues[7] and Guttmann and colleagues[8] Few studies have validated case definitions distinguishing between type 1 and type 2 diabetes.[9,10] A recent systematic review[11] identified 16 studies that used administrative data coded with the *International Classification of Diseases* (ICD) to derive validated case definitions for diabetes in adults. None of these studies were able to differentiate between type 1 and type 2

Competing interests: None declared.

This article has been peer reviewed.

Correspondence to: B. Cord Lethebe, bcletheb@ucalgary.ca

CMAJ Open **2019. DOI:10.9778/cmajo.20180142**

diabetes. The Canadian Primary Care Sentinel Surveillance Network (CPCSSN) extracts, transforms, cleans and codes primary care data from electronic medical records (EMRs) into a standardized data model and makes the processed data available for research, surveillance, quality improvement and panel management.[12] Previously, the CPCSSN developed a definition for undifferentiated diabetes that had excellent accuracy (95.3% sensitivity and 97.1% specificity).[13] The current study builds on that CPCSSN definition by identifying patients by diabetes type using artificial intelligence machine learning. This data-oriented method is designed to find patterns or generate predictive models using large and complex data sets to identify the most accurate case definition within the data.[14] Our objective was to assess whether machine learning methods can be used to create human-readable case definitions that distinguish between type 1 and type 2 diabetes.

Methods

Data source and reference standard

The CPCSSN extracts de-identified clinical data from the EMR systems of about 1500 sentinel family physicians, nurse practitioners and community pediatricians, who contribute data for about 1.8 million patients in 7 provinces and 1 territory across Canada. These data include patient demographic characteristics, diagnoses, prescribed medications, laboratory results, physical measurements (e.g., weight, blood pressure), medical procedures, behavioural risk factors, physician billing, allergies, vaccinations and referrals. Previous work has shown that the CPCSSN database includes patients who are generally representative of the Canadian population.[15] CPCSSN data are available to researchers upon approval; further details about the CPCSSN have been published previously[16] and can be found on the CPCSSN website (www.cpcssn.ca).

For this study, we used data from one of the CPCSSN's participating practice-based research networks, the Southern Alberta Primary Care Research Network, extracted on Dec. 31, 2016, and derived for the period 2008–2016 inclusive. This network generally had more than 200 000 patients (e.g., 237 734 in the fourth quarter of 2016, of whom 17 003 had diabetes). We identified and sampled a cohort of patients of all ages, who were believed to have diabetes, using the current validated CPCSSN case definition. Family physicians who agreed to participate in this study were able to re-identify the patients with CPCSSN-defined diabetes in the EMR systems of their respective practices. We sent an e-mail invitation to family physicians who were part of the Southern Alberta Primary Care Research Network and who belonged to 1 of 4 clinics in southern Alberta that we had identified as having a large population of patients with diabetes. We first asked the family physicians to confirm that the specified patients had diabetes and then to determine whether each patient had type 1, type 2 or another diabetes subtype, according to their clinical expertise and any supporting evidence they chose to use. The physicians performed these tasks before any machine learning classification was performed. The list of physician-confirmed diabetes cases, along with the diabetes subtype for each case (type 1, type 2 or other) constituted the reference standard for the analysis. By using this method of sampling patients, we attempted to preserve the true distribution of type 1 and type 2 cases within the undifferentiated diabetes population. This approach allowed us to comfortably report positive predictive value (PPV) and negative predictive value (NPV) metrics.

Machine learning

We applied supervised machine learning methods to the large, complex, multivariable CPCSSN data set and the reference standard (physician-confirmed diabetes subtype) to "learn" the clinical characteristics (called "features") that differentiated people with type 1 diabetes mellitus from those with other subtypes of the disease.

Selection of features

Before the machine learning processing began, we selected and defined as binary outcomes all plausibly relevant variables within the CPCSSN data. More specifically, we selected features using information from various parts of the patient chart: age, physician billing, current and historical diagnoses, referrals and prescribed medications. Diagnoses in the Canadian primary care setting are generally coded using the *International Classification of Diseases, 9th Revision* (ICD-9). Therefore, we considered as a feature every unique ICD-9 code present in the EMR database. In addition, we generated features on the basis of 2 instances of a code in the year leading up to the chart review, as well as 2 instances within 2 years. We used a similar approach for the medication and referral codes. We included diagnoses, referrals and medications that were recorded as free text using a simple "bag-of-words" approach, which creates a binary indicator for each unique word that appears in any free-text field within the CPCSSN database. Similarly, we added non–case-sensitive wildcard searches for the following keywords and phrases related to diabetes status: "type 1," "type 2," "type i," "type ii," "insulin dependent," "insulin dep," "tidm," "tiidm," "non insulin dependent," "type 1 insulin dependent," "iddm," "niddm," "dm1," "dm2," "dmi," "dmii," "t2dm" and "t1dm." We also included the following combinations as features: "type 1+ insulin dependent+ insulin dep+ type 1 insulin dependent" and "type i+ tidm."

For each prescribed medication recorded in the EMR, CPCSSN assigns codes from the Anatomical Therapeutic Chemical (ATC) Classification system.[17] We included as a feature each unique ATC code appearing in the medication table, using truncated codes to identify families of drugs rather than specific examples. We also assessed the frequency of ATC codes, particularly whether 2 instances of the same code were used within 1 year, and 2 instances of the same code within 2 years.

We also included laboratory values. The diabetes-related tests available in the CPCSSN data are hemoglobin A1c and fasting plasma glucose measures. We created binary indicators for whether or not a patient had certain laboratory values over ranges of thresholds (e.g., HbA1c > 6.3%, > 6.4%, > 6.5%, > 6.6%).

We included various age cut-offs as features because recent evidence suggests that 48% of type 1 cases are diagnosed before the patient reaches 30 years of age, and that the remaining 52% of cases are diagnosed between ages 30 and 60 years.[18] The peak incidence period for diagnosis of type 1 diabetes is age 10 to 14 years.[19] We included each age-year between 18 and 50 inclusive as candidate features, with age being calculated at the date of intitially meeting the general case definition for diabetes.

Algorithms

We required that the generated case definitions be entirely human-readable and easily translated into a set of logic statements. These criteria ruled out many of the "black-box" machine learning algorithms like the random forest algorithm, support-vector machines and artificial neural networks. We used the following machine learning algorithms for feature selection in this analysis: the C5.0 decision tree,[14] the classification and regression tree decision tree,[20] the chi-square automated interaction detection (CHAID) decision tree[21,22] and the least absolute shrinkage and selection operator (LASSO) logistic regression.[23,24] These algorithms are commonly used in machine learning settings and were selected for their ability to generate human-readable rule sets that can be used as case definitions.[14]

Statistical analysis

Each of the machine learning algorithms has tuning parameters that can be manipulated to control the complexity and size of the final definition. These complexity parameters include maximum depth of the tree, a confidence factor or complexity parameter, a minimum number of cases required to make a split and a loss matrix. We selected the tuning parameters using a bootstrap method. We took a random sample with replacement of the study population for a range of possible values for the tuning parameters. We repeated the process 30 times for each tuning parameter value, until we could determine which tuning parameter values optimized the accuracy metrics. Specifically, we investigated the misclassification rate, the F1 score, the PPV and the Youden *J* statistic. The F1 score is defined as (sensitivity × PPV)/(sensitivity + PPV). The Youden *J* statistic is defined as sensitivity + specificity – 1.

Once we had selected the tuning parameters, we used 10-fold cross-validation to determine the validity estimates.[25] We accomplished this by splitting the study population into 10 segments or "folds." We conducted training of the model using 9 of these folds, and performed testing on the remaining fold. We repeated the cross-validation 10 times, such that each fold was used once for testing. After determining the validity estimates, we fitted the model with the entire study population to get the final case definition. We used R statistical software version 3.3.1 for all statistical analyses.

Ethics approval

The study received ethics approval from the University of Calgary's Conjoint Health Research Ethics Board (Ethics ID REB17–0091).

Results

A total of 189 physicians in the southern Alberta CPCSSN were approached about the study, and 23 agreed to participate. From the patient rosters of these physicians, we initially identified a sample of 1501 patients who were thought to have diabetes, of whom 102 were subsequently found not to have the disease (i.e., had been misclassified by the physician). Of the remaining 1399 patients, an additional 90 patients (6.0% of the original sample) were excluded for various reasons: 14 patients had died, 68 were no longer active in the physician's panel, and 8 had gestational diabetes or a relatively rare diabetes subtype (e.g., latent autoimmune diabetes of adults, mature onset diabetes of the young). Therefore, 1309 patients were confirmed to have the disease and were included in the analysis. Of this sample, 1199 people (91.6%) were classified by their family physician as having type 2 diabetes and 110 people (8.4%) were classified by their family physician as having type 1 diabetes; the cohort thus created included substantially more cases of type 2 diabetes than type 1 diabetes.

Patients with type 1 diabetes were younger, and this group included more females (Table 1). Also, a substantially greater proportion of patients with type 1 diabetes had insulin prescriptions, both issued in the past year (30.0% v. 6.6%) and at any time (76.4% v. 13.0%).

The 10-fold cross-validation results are presented in Table 2. Prevalence of the disease in the sample was relatively low, so the algorithms naturally favoured high-specificity models, except when the validity metrics that favour sensitivity were maximized (e.g., the Youden *J* statistic). The sensitivities for misclassification rate, F1 score and PPV ranged from 40.0% to 61.8%, whereas the specificities for these tuning parameters ranged from 96.3% to 99.3%. When the Youden *J* statistic was maximized, however, sensitivities ranged from 52.7% to 87.3%, and specificities from 85.4% to 97.9%.

Table 3 shows the final case definitions for 2 notable models from the 10-fold cross-validation results. The first is the CHAID method maximizing PPV. The cross-fold estimate here had a PPV of 85.5% but lacked sensitivity (42.7%). The case definition interpreted for this model used the free-text term "type 1" appearing anywhere in the text fields for problem list, encounter diagnosis, billing or medication reason. Also, all those who were under the age of 22 years at the time of first meeting the general case definition or diabetes were considered to represent type 1 cases.

The second reported case definition is the LASSO implementation maximizing the Youden *J* statistic. This method yielded sensitivity of 87.3% and specificity of 85.4%. The added sensitivity came at the cost of PPV, which was estimated as 35.6%. The features selected for this case definition were the term "type 1" appearing in any text field, a prescription for insulin or age less than 30 years at the time of first meeting the general case definition for diabetes. The importance of each feature, as ranked by the random forest model, is presented in Appendix 1 (available at www.cmajopen.ca/content/7/2/E246/suppl/DC1).

Table 1: Demographic and relevant clinical features comparing patients with type 1 and type 2 diabetes

	Group; % of patients (95% CI)*†		
Characteristic	Type 2 diabetes *n* = 1199	Type 1 diabetes *n* = 110	Total *n* = 1309
Sex, male	53.5 (50.6–56.3)	47.3 (37.7–57.0)	52.9 (50.2–55.7)
Age, yr, mean (95% CI)	64.6 (63.9–65.3)	46.0 (42.8–49.2)	63.0 (62.3–63.8)
No. of encounters in past year, mean (95% CI)	5.1 (4.8–5.3)	4.0 (3.2–4.8)	5.0 (4.8–5.2)
Prescription for insulin (A10AB - -)‡			
In past year	6.6 (5.3–8.2)	30.0 (21.8–39.6)	8.6 (7.1–10.2)
In past 2 years	8.8 (7.2–10.5)	47.3 (37.8–57.0)	12.0 (10.3–13.9)
At any time	13.0 (11.2–15.1)	76.4 (67.1–83.7)	18.3 (16.3–20.6)
Prescription for blood glucose–lowering drugs excluding insulin (A10B - - -)‡			
In past year	45.5 (42.6–48.3)	12.7 (7.4–20.8)	42.7 (40.0–45.4)
In past 2 years	54.6 (51.8–57.5)	20.9 (14.0–29.9)	51.8 (49.0–54.5)
At any time	71.9 (69.2–74.4)	26.4 (18.6–35.8)	68.1 (65.5–70.6)
Occurrence of “type 1” in any text field	0.7 (0.3–1.4)	40.0 (30.9–49.8)	4.0 (3.0–5.2)
Billing code 250.01 in past year	0	10.0 (5.3–17.6)	0.8 (0.4–1.5)
Occurrence of “type 2” in any text field	26.3 (23.8–28.9)	7.3 (3.4–14.3)	24.7 (22.4–27.1)
Occurrence of “diabetes” in any text field	95.3 (93.9–96.4)	99.1 (94.3–100)	95.6 (94.4–96.7)

Note: CI = confidence interval.
*Except where indicated otherwise.
†CIs for proportions are exact.
‡The parenthetical notation represents relevant codes in the Anatomical Therapeutic Chemical Classification system, where each code is 7 characters long and dashes represent “wild card” characters. Specifically, insulin is represented by various codes in which the first 5 characters are A10AB, and blood glucose–lowering drugs other than insulin are represented by various codes in which the first 4 characters are A10B.

Interpretation

We have shown that machine learning methods can be used to create interpretable case definitions that distinguish between type 1 and type 2 diabetes in CPCSSN-processed EMR data. Although we found no single case definition with high sensitivity, specificity and predictive values, we judge that at least 2 useful case definitions were identified. The first adopted the CHAID implementation maximizing PPV. This simple case definition has high PPV and NPV. High predictive values are ideal for creating cohorts in observational studies and for other screening purposes, because patients for whom there is a strong probability of having the condition of interest are identified with high accuracy. The second case definition adopted the LASSO approach maximizing the Youden *J* statistic and had good sensitivity and specificity (87.3% and 85.4%, respectively). This case definition would be useful for epidemiologic and surveillance purposes, such as examining population-level temporal trends of incidence and prevalence.

Clottey and colleagues[5] developed a case definition for undifferentiated diabetes which consisted of at least one of the following criteria: at least 2 physician billing claims within a 2-year period or 1 hospital admission with an ICD code for diabetes. Hux and colleagues[6] generated 2 definitions, the first involving either 1 claim or 1 hospital admission, and the second involving either 2 claims or 1 hospital admission. In British Columbia, Amed and colleagues[7] developed 2 additional definitions intended for use in children and adolescents. The first was based on 1 hospitalization, 2 physician billing claims in a single year, and combinations of insulin and oral antidiabetic medications. The second consisted of 4 billing codes over 2 years. Guttmann and colleagues[8] developed a definition for pediatric diabetes using claims data exclusively, concluding that 4 physician billing claims using ICD-9 code 250.X in a 2-year period provided optimal sensitivity and specificity. Each study included in the recent systematic review by Khokhar and colleagues[11] used physician claims either alone or in combination with hospital discharge data. Physician billing is not necessarily an accurate reflection of the content of a given encounter. For example, Wyse and colleagues[26] identified 15% under-reporting of polypectomy when validated against clinical records. Muhajarine and colleagues[27] identified similar misclassification rates for hypertension. Hux and colleagues[6] reported PPVs for their case definitions ranging from 0.61 to 0.80, which indicated substantial misclassification of diabetes relative to chart review. Hence, our study represents 2 important achievements: it exploits data other than hospital admissions and physician claims in determining cases and creating case definitions that maximize sensitivity, specificity, PPV and NPV, and it also presents validation metrics for the case definitions supporting differentiation between type 1 and type 2 diabetes.

Table 2: Ten-fold cross-validation results for each of 4 machine learning algorithms, minimizing or maximizing various metrics*

Metric and algorithm	Sensitivity, %	Specificity, %	PPV, %	NPV, %	Accuracy, %†
Misclassification rate					
C5.0	40.9 (31.8–50.7)	99.3 (98.6–99.7)	84.9 (71.9–92.8)	94.8 (93.4–95.9)	94.4 (93.0–95.5)
CaRT	40.9 (31.8–50.7)	99.3 (98.6–99.7)	84.9 (71.9–92.8)	94.8 (93.4–95.9)	94.4 (93.0–95.5)
CHAID	40.0 (30.9–49.8)	99.3 (98.6–99.7)	84.6 (71.4–92.7)	94.7 (93.3–95.9)	94.3 (92.9–95.5)
LASSO	40.9 (31.8–50.7)	99.3 (98.6–99.7)	84.9 (71.9–92.8)	94.8 (93.4–95.9)	94.4 (93.0–95.5)
F1 score					
C5.0	61.8 (52.0–70.8)	96.5 (95.2–97.4)	61.8 (52.0–70.8)	96.5 (95.2–97.4)	93.5 (92.0–94.8)
CaRT	60.9 (51.1–69.9)	96.3 (95.0–97.3)	60.4 (50.6–69.4)	96.4 (95.1–97.3)	93.3 (91.8–94.6)
CHAID	51.8 (42.1–61.4)	98.6 (97.7–99.1)	77.0 (65.5–85.7)	95.7 (94.3–96.7)	94.6 (93.2–95.8)
LASSO	40.9 (31.8–50.7)	99.3 (98.6–99.7)	84.9 (71.9–92.8)	94.8 (93.4–95.9)	94.4 (93.0–95.5)
PPV					
C5.0	43.6 (34.3–53.4)	99.1 (98.3–99.5)	81.4 (68.7–89.9)	95.0 (93.6–96.1)	94.4 (93.0–95.5)
CaRT	40.9 (31.8–50.7)	99.3 (98.6–99.7)	84.9 (71.9–92.8)	94.8 (93.4–95.9)	94.4 (93.0–95.5)
CHAID‡	42.7 (33.5–52.5)	99.3 (98.6–99.7)	85.5 (72.8–93.1)	94.9 (93.5–96.1)	94.5 (93.1–95.7)
LASSO	40.9 (31.8–50.7)	99.3 (98.6–99.7)	84.9 (71.9–92.8)	94.8 (93.4–95.9)	94.4 (93.0–95.5)
Youden *J* statistic					
C5.0	85.5 (77.2–91.2)	85.5 (83.4–87.5)	35.3 (29.7–41.5)	98.5 (97.4–99.1)	85.5 (83.5–87.4)
CaRT	80.9 (72.1–87.5)	89.2 (87.2–90.8)	40.8 (34.3–47.7)	98.1 (97.0–98.8)	88.5 (86.6–90.1)
CHAID	52.7 (43.0–62.2)	97.9 (96.9–98.6)	69.9 (58.7–79.2)	95.7 (94.4–96.8)	94.1 (92.6–95.3)
LASSO‡	87.3 (79.2–92.6)	85.4 (83.2–87.3)	35.6 (29.9–41.6)	98.6 (97.7–99.2)	85.5 (83.5–87.4)

Note: CaRT = classification and regression tree, CHAID = chi-square automated interaction detection, LASSO = least absolute shrinkage and selection operator, NPV = negative predictive value, PPV = positive predictive value.
*The misclassification rate metric was minimized, whereas the F1 score, PPV and Youden *J* statistic metrics were maximized.
†A dummy classifier that assumes all cases were type 2 diabetes would achieve an accuracy of 91.6%.
‡Instances reported as final case definitions.

Table 3: Final case definitions for 2 notable instances of cross-validation results*

Type of analysis	Case definition
CHAID with maximized PPV	Any of the following 2 criteria: • Anywhere text "type 1" • Age < 22 yr at time of original diabetes diagnosis
LASSO with maximized Youden *J* statistic	Any of the following criteria: • Anywhere text "type 1" • Any occurrence of A10AB- - in the medication table (insulin and analogues for injection, fast acting)† • Age < 30 yr at time of original diabetes diagnosis

Note: CHAID = chi-square automated interaction detection, LASSO = least absolute shrinkage and selection operator, PPV = positive predictive value.
*Disease status assumed to be type 2 diabetes or a diabetes subtype, unless the patient meets criteria for type 1 diabetes.
†The specified notation represents relevant codes in the Anatomical Therapeutic Chemical Classification system, where each code is 7 characters long and dashes represent "wild card" characters. Specifically, insulin is represented by various codes in which the first 5 characters are A10AB.

Limitations

This study had a small number of confirmed cases of type 1 diabetes. We believe that the under-recording of insulin prescriptions for patients confirmed as having type 1 diabetes derives from their receiving most of their diabetes-specific care from an endocrinologist or other diabetes specialist in an outpatient clinic setting, transactions that are usually not subsequently recorded in primary care EMRs (which was our data source). Future research on a larger sample would result in more stable validity results and feature selection. The validity measures should be interpreted with caution, given that the diabetes cohort was selected from patients meeting the previously validated case definition for diabetes, and their inclusion was conditional upon CPCSSN-processed data and criteria and the validity of that definition. The sample that defined the reference set may not have been representative of the Canadian population, despite the fact that CPCSSN is generally representative of patients seen in Canadian primary care. Misclassification as to the type of diabetes is also possible; however, we are confident that the rate of misdiagnosis was reasonably low, because type 1 diabetes has little diagnostic uncertainty (based on extreme thirst, urination and weight loss), requires immediate medical attention and immediate insulin use, and often presents with diabetic ketoacidosis (and hospital admission). Further study is required to determine the validation metrics of the case definitions for type 1 diabetes in non-CPCSSN EMR data. Furthermore, an external validation study would provide better evidence of the generalizability of the new case definitions.

Conclusion

We used machine learning to develop and validate 2 case definitions that achieve different goals in distinguishing between type 1 and type 2 diabetes in CPCSSN data. One of these case definitions is suited for screening and cohort development, with high PPV and NPV. The other case definition is suited for epidemiologic purposes, having a reasonable balance between sensitivity and specificity. Further validation and testing with a larger and more diverse sample are recommended.

References

1. Lipscombe LL, Hux JE. Trends in diabetes prevalence, incidence and mortality in Ontario, Canada 1995-2005: a population-based study. *Lancet* 2007;369:750-6.
2. Toews J, Pelletier C, McRae L. Diabetes trends, 2003/04–2013/14: data from the Canadian Chronic Disease Surveillance System. *Can J Diabetes* 2017;41:S77.
3. Diabetes statistics in Canada. Diabetes Canada. Available: http://www.diabetes.ca/how-you-can-help/advocate/why-federal-leadership-is-essential/diabetes-statistics-in-canada (accessed 2018 Feb. 5).
4. Clement M, Harvey B, Rabi DM, et al. Organization of diabetes care. *Can J Diabetes* 2013;37(Suppl 1):S20-5.
5. Clottey C, Mo F, LeBrun B, et al. The development of the National Diabetes Surveillance System (NDSS) in Canada. *Chronic Dis Can* 2001;22:67-9.
6. Hux JE, Ivis F, Flintoft V, et al. Diabetes in Ontario: determination of prevalence and incidence using a validated administrative data algorithm. *Diabetes Care* 2002;25:512-6.
7. Amed S, Vanderloo SE, Metzger D, et al. Validation of diabetes case definitions using administrative claims data. *Diabet Med* 2011;28:424-7.
8. Guttmann A, Nakhla M, Henderson M, et al. Validation of a health administrative data algorithm for assessing the epidemiology of diabetes in Canadian children. *Pediatr Diabetes* 2010;11:122-8.
9. Vanderloo SE, Johnson JA, Reimer K, et al. Validation of classification algorithms for childhood diabetes identified from administrative data. *Pediatr Diabetes* 2012;13:229-34.
10. Ng E, Dasgupta K, Johnson JA. An algorithm to differentiate diabetic respondents in the Canadian Community Health Survey. *Health Rep* 2008;19:71-9.
11. Khokhar B, Jette N, Metcalfe A, et al. Systematic review of validated case definitions for diabetes in ICD-9-coded and ICD-10-coded data in adult populations. *BMJ Open* 2016;6:e009952.
12. Greiver M, Williamson T, Barber D, et al. Prevalence and epidemiology of diabetes in Canadian primary care practices: a report from the Canadian Primary Care Sentinel Surveillance Network. *Can J Diabetes* 2014;38:179-85.
13. Williamson T, Green ME, Birtwhistle R. Validating the 8 CPCSSN case definitions for chronic disease surveillance in a primary care database of electronic health records. *Ann Fam Med* 2014;12:367-72.
14. Quinlan JR. *C4.5 programs for machine learning.* Burlington (MA): Morgan Kaufmann; 1992.
15. Queenan JA, Williamson T, Khan S, et al. Representativeness of patients and providers in the Canadian Primary Care Sentinel Surveillance Network: a cross-sectional study. *CMAJ Open* 2016;4:E28-32.
16. Garies S, Birtwhistle R, Drummond N, et al. Data resource profile: national electronic medical record data from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN). *Int J Epidemiol* 2017;46:1091-1092f.
17. *International language for drug utilization research*. Oslo: WHO Collaborating Centre for Drug Statistics Methodology; updated 2019 Jan. 7. Available: https://whocc.no (accessed 2018 Feb. 27).
18. Thomas NJM, Jones S, Weedon M, et al. Classifying diabetes by type 1 genetic risk shows autoimmune diabetes cases are evenly distributed above and below 30 years of age [abstract 264]. *Diabetologia* 2016;59(Suppl 1):S135.
19. Diaz-Valencia PA, Bougnères P, Valleron AJ. Global epidemiology of type 1 diabetes in young adults and adults: a systematic review. *BMC Public Health* 2015;15:255.
20. Therneau TM, Atkinson EJ. *An introduction to recursive partitioning using the RPART routines*. Rochester (MN): Mayo Foundation; 2018. Available: https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf (accessed 2018 Mar. 22).
21. Kass GV. An exploratory technique for investigating large quantities of categorical data. *Appl Stat* 1980;29:119-27.
22. The FoRt Student Project Team. *CHAID: chi-squared automated interaction detection, 2015. R package version 0.1-2*. Vienna (Austria): R Foundation for Statistical Computing; 2015
23. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* 2010;33:1-22.
24. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc B* 1996;58:267-88.
25. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence* [Montréal 1995 Aug. 20–25]. Vol. 2:1137-45. Stanford (CA): Morgen Kauman; 1995.
26. Wyse JM, Joseph L, Barkun AN, et al. Accuracy of administrative claims data for polypectomy. *CMAJ* 2011;183:E743-7.
27. Muhajarine N, Mustard C, Roos LL, et al. Comparison of survey and physicians claims data for detecting hypertension. *J Clin Epidemiol* 1997;50:711-8.

Affiliations: Department of Community Health Sciences (Lethebe, Williamson, Garies, McBrien, Soos, Shaw), Clinical Research Unit (Lethebe), Department of Family Medicine (Garies, McBrien, Leduc, Drummond) and Department of Medicine (Butalia), University of Calgary, Calgary, Alta.; Department of Family Medicine (Drummond), University of Alberta, Edmonton, Alta.

Contributors: B. Cord Lethebe contributed to the study design, to analysis of the data and to writing the manuscript. Tyler Williamson contributed to the study design, to analysis and interpretation of the data, and to writing the manuscript. Stephanie Garies, Kerry McBrien, Charles Leduc, Sonia Butalia, Boglarka Soos, Marta Shaw and Neil Drummond contributed to the study design, to interpretation of the data and to writing the manuscript. All of the authors revised the paper critically for important intellectual content, gave final approval of the version to be published and agreed to be accountable for all aspects of the work.

Funding: Alberta Innovates contributed to funding of the Secondary Use of Data Project, of which this research formed a component.

Acknowledgements: The authors thank Tim Murphy for contributions to development of the Secondary Use of Data Project (of which this research formed a component) and Rebecca Miyagishima, who assisted with preparing the manuscript for submission.

Supplemental information: For reviewer comments and the original submission of this manuscript, please see www.cmajopen.ca/content/7/2/E246/suppl/DC1.