

Validation of a type 1 diabetes algorithm using electronic medical records and administrative healthcare data to study the population incidence and prevalence of type 1 diabetes in Ontario, Canada

Alanna Weisman ^{1,2,3}, Karen Tu,^{3,4,5,6} Jacqueline Young,¹ Matthew Kumar,¹ Peter C Austin,^{1,3} Liisa Jaakkimainen,^{1,3,6} Lorraine Lipscombe,^{1,2,3,7} Ronnie Aronson,⁸ Gillian L Booth^{1,2,3,9}

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

► Additional material is published online only. To view please visit the journal online (<http://dx.doi.org/10.1136/bmjdr-2020-001224>).

Received 24 January 2020
Revised 12 May 2020
Accepted 19 May 2020



© Author(s) (or their employer(s)) 2020. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

For numbered affiliations see end of article.

Correspondence to

Dr Alanna Weisman;
alanna.weisman@mail.utoronto.ca

ABSTRACT

Introduction We aimed to develop algorithms distinguishing type 1 diabetes (T1D) from type 2 diabetes in adults ≥ 18 years old using primary care electronic medical record (EMRPC) and administrative healthcare data from Ontario, Canada, and to estimate T1D prevalence and incidence.

Research design and methods The reference population was a random sample of patients with diabetes in EMRPC whose charts were manually abstracted (n=5402). Algorithms were developed using classification trees, random forests, and rule-based methods, using electronic medical record (EMR) data, administrative data, or both. Algorithm performance was assessed in EMRPC. Administrative data algorithms were additionally evaluated using a diabetes clinic registry with endocrinologist-assigned diabetes type (n=29 371). Three algorithms were applied to the Ontario population to evaluate the minimum, moderate and maximum estimates of T1D prevalence and incidence rates between 2010 and 2017, and trends were analyzed using negative binomial regressions.

Results Of 5402 individuals with diabetes in EMRPC, 195 had T1D. Sensitivity, specificity, positive predictive value and negative predictive value for the best performing algorithms were 80.6% (75.9–87.2), 99.8% (99.7–100), 94.9% (92.3–98.7), and 99.3% (99.1–99.5) for EMR, 51.3% (44.0–58.5), 99.5% (99.3–99.7), 79.4% (71.2–86.1), and 98.2% (97.8–98.5) for administrative data, and 87.2% (81.7–91.5), 99.9% (99.7–100), 96.6% (92.7–98.7) and 99.5% (99.3–99.7) for combined EMR and administrative data. Administrative data algorithms had similar sensitivity and specificity in the diabetes clinic registry. Of 11 499 711 adults in Ontario in 2017, there were 24 789 (0.22%, minimum estimate) to 102 140 (0.89%, maximum estimate) with T1D. Between 2010 and 2017, the age-standardized and sex-standardized prevalence rates per 1000 person-years increased (minimum estimate 1.7 to 2.56, maximum estimate 7.48 to 9.86, $p < 0.0001$). In contrast, incidence rates decreased (minimum estimate 0.1 to 0.04, maximum estimate 0.47 to 0.09, $p < 0.0001$).

Conclusions Primary care EMR and administrative data algorithms performed well in identifying T1D and demonstrated increasing T1D prevalence in Ontario. These algorithms may permit the development of large, population-based cohort studies of T1D.

Significance of this study

What is already known about this subject?

- Validated algorithms identifying diabetes in large electronic medical record and administrative healthcare databases permit population-based research in diabetes but generally do not distinguish type 1 and type 2 diabetes.
- The inability to identify type 1 diabetes in large databases is a barrier to understanding type 1 diabetes epidemiology and outcomes.

What are the new findings?

- Algorithms identifying type 1 diabetes in a primary care electronic medical record database had excellent performance, and algorithms identifying type 1 diabetes in administrative healthcare data had good performance.
- Application of the administrative data algorithms demonstrated that type 1 diabetes prevalence rates among adults in Ontario, Canada have been increasing since 2010, while incidence rates have been stable.

How might these results change the focus of research or clinical practice?

- Application of these validated algorithms for identifying type 1 diabetes will permit population-based studies on the epidemiology, healthcare utilization, and outcomes of type 1 diabetes.

INTRODUCTION

Type 1 diabetes (T1D) results from the autoimmune destruction of the pancreatic beta cells, leading to lifelong insulin deficiency and hyperglycemia. Although T1D accounts for only an estimated 5%–10% of all diabetes cases, it continues to be associated with excess morbidity, premature mortality, and

an economic burden to healthcare systems.^{1–3} The relative rarity of T1D compared with the burden of type 2 diabetes (T2D) and difficulties in distinguishing T1D from T2D in large databases have been major barriers to understanding T1D epidemiology and outcomes at a population level.⁴ For example, while the epidemiology of T1D in children has been well described in a number of countries,^{5–7} there is a paucity of data concerning the epidemiology of T1D in adults, which is important for public health and resource planning.⁸

Population-based data sources such as electronic medical records (EMRs) or administrative healthcare databases are efficient and cost-effective methods for studying diabetes epidemiology and outcomes.⁹ An important strategy for conducting population-based studies is ‘electronic phenotyping’, in which algorithms are developed and used to identify patients with a particular condition.¹⁰ A number of algorithms distinguishing T1D from T2D have been developed using EMR databases. The majority of these algorithms were developed using rule-based methods or decision trees, although Lethebe and others applied various other machine learning approaches.¹¹ These algorithms commonly rely heavily on physician billing codes that are specific for T1D or T2D.^{11–15} However, physician billing codes differentiating T1D and T2D are not available in all regions and, even when they are available, result in substantial misclassification due to errors in coding.¹⁶ Furthermore, there has been limited comparison in the performance of algorithms derived from different methods (ie, rule-based vs decision trees vs machine learning).

Algorithms distinguishing T1D and T2D using administrative healthcare data have been developed for use in the pediatric population in countries such as the USA and Canada, but are more challenging to apply to adult populations in whom the proportion of T1D to T2D is much lower.^{17,18} To our knowledge, no algorithms distinguishing T1D from T2D in adults have been developed using administrative healthcare data alone, which lack the rich clinical details available in EMR databases. A major advantage of administrative healthcare data algorithms is their potential to be applied to an entire population.

Our objectives were to (1) derive and validate algorithms distinguishing T1D from T2D in diabetes populations using a primary care EMR database, population-based administrative healthcare data, and combined EMR and administrative data from Ontario, Canada; and (2) evaluate temporal trends in prevalence and incidence of T1D in this population by applying the best performing administrative healthcare data algorithms.

RESEARCH DESIGN AND METHODS

This study used EMR and administrative healthcare data from the province of Ontario, Canada, which has a single-payer healthcare system and a population of approximately 14 million. Records for a given individual

were linked across multiple databases using encoded identifiers based on health card numbers and analyzed at ICES (formerly known as the Institute for Clinical Evaluative Sciences). The study was designed and reported according to guidelines for validation studies using administrative data.¹⁹

Creation of the reference population

The reference population was derived from a primary care EMR database, known as the Electronic Medical Records Primary Care (EMRPC). EMRPC (formerly referred to as EMRALD) contains all clinical information from the medical chart of ~450 000 patients, contributed by over 400 primary care physicians using the Practice Solutions EMR in Ontario, Canada. Available data include the Cumulative Patient Profile (CPP) (medical history, problem list, medications, risk factors and allergies), progress notes, laboratory tests, specialist consultation letters, diabetes education center notes, and hospital discharge summaries. The population of EMRPC is generally representative of the entire Ontario population, with slight under-representation of young adult men and slight over-representation of young adult women.²⁰

All patients in EMRPC with diabetes were identified using a previously validated algorithm that has a sensitivity of 83.1% and specificity of 98.2%.^{21,22} Inclusion criteria were age ≥ 18 years and use of EMR by the primary care physician for at least 1 year. Exclusion criteria included individuals registered to the primary care physician who had not yet had a clinic visit prior to 30 September 2015 and missing date of birth.

A random 25% subsample of EMRPC patients with diabetes were selected to create the reference population. The number of individuals required in validation studies of a disease state depends on the prevalence of the disease. A larger number of individuals is required for diseases with low prevalence to ensure there are a sufficient number of cases in the validation sample.²³ To increase the efficiency of chart abstraction, we used an approach similar to the ‘Strawman’ algorithm employed by Klompas *et al*,¹² by selecting patients more likely to have T1D to undergo chart abstraction, thereby increasing the number of cases for the same number of charts abstracted. A search of medications in the CPP categorized patients into those with possible T1D (prescribed insulin with or without metformin, but no other antihyperglycemic agent) who were selected for chart abstraction, and those unlikely to have T1D (not prescribed insulin or prescribed a non-insulin, non-metformin antihyperglycemic agent) whose charts were not abstracted but were included in the reference population as they were assumed to not have T1D.

Patients classified as having possible T1D based on these criteria were manually reviewed by an endocrinologist (AW) using a standardized abstraction platform that was piloted on the first 100 charts. Fifty charts were randomly selected to be reabstracted by the same

abstractor for intrarater reliability assessment, and an additional 50 charts were randomly selected to be abstracted by a second abstractor (a primary care physician, LJ) for inter-rater reliability assessment. During the abstraction, diabetes type was classified as T1D or T2D using the entire EMR record and specific classification criteria (online supplementary appendix 1).

Validation of algorithms in an independent external sample

Administrative healthcare data algorithms that were derived using EMRPC as the reference population were additionally validated in an independent sample, the LMC Diabetes Registry. LMC is a network of community-based endocrinology practices with a central database that includes diabetes type as determined by an endocrinologist.²⁴ The database used for this study included records for patients from seven clinic sites.

Variables for EMR data algorithms

Automated searches of free text in the CPP were performed, which is the active health history section of the EMR. CPP descriptions of diabetes were classified as 'Definite T1D', 'Possible T1D' or 'T2D' (see table 1 footnote). Automated searches of structured fields were performed for medications (categorized as any insulin, bolus insulin, metformin, or antihyperglycemic medication other than insulin or metformin), body mass index and age.

Variables for administrative healthcare data algorithms

Hospitalization and emergency department diagnoses were obtained from the CIHI (Canadian Institute for Health Information)-Discharge Abstract Database and the National Ambulatory Care Reporting Services held at ICES. Physician service claims were determined from the Ontario Health Insurance Plan. Prescription medication information was obtained from the Ontario Drug Benefit Database, which includes claims for prescription medications for individuals age 65 or older, or those individuals under age 65 with disability or social assistance. Laboratory data were obtained from the Ontario Laboratories Information System. Insulin pump use was obtained from the Ontario Ministry of Health Assistive Devices Program, which has provided funding for insulin pumps to children with T1D since 2006 and adults with T1D since 2008. Date of diabetes diagnosis (incidence) was determined from the Ontario Diabetes Database, which requires two physician service claims for diabetes within 2 years or one hospitalization record with diabetes.⁹ Pediatric diabetes was identified by a previously validated algorithm, which requires four physician service codes with a diagnosis of diabetes within 1 year prior to the age of 19 years.²⁵ Hospitalization and emergency records transitioned from the International Classification of Diseases-9 (ICD-9) to ICD-10 in 2002. Prior to 2002, hospitalization records did not distinguish diabetic ketoacidosis (DKA) or diabetes as being specific to T1D or T2D. Therefore, three variables for DKA were considered: any DKA

consisted of a hospitalization code for DKA using either ICD-9 or ICD-10 criteria from any date, T1D-specific DKA consisted of a hospitalization code for DKA using ICD-10 criteria for 2002 or later, and T2D-specific DKA consisted of a hospitalization code for DKA using ICD-10 criteria for 2002 or later. A full list of variables and the corresponding codes is available in online supplementary appendix 2.

Derivation of algorithms

Algorithms were derived using the EMRPC reference population. Three approaches to algorithm derivation (all considering the same potential set of variables derived from the data sources above) were used: (1) classification trees; (2) random forests; and (3) rule-based methods. Classification trees categorize individuals as having T1D or T2D based on a sequence of binary splits or partitions that, at each step, consider all possible binary splits of the data (recursive partitioning). First, the single variable which best splits the data into two groups is selected. For each resulting subgroup, another variable (which could be the same variable) is selected that best splits the subgroup into two further subgroups. The process is repeated recursively on each of the two groups until no further partitioning can be performed.²⁶ Variables selected in the classification tree analysis are thus chosen without investigator input or selection. Random forests grow a sequence of classification trees, each classification tree grown in a different bootstrap sample. At each node, only a random sample of the predictor variables are considered when deciding how to partition the given node. Results are combined across the classification trees using a majority-vote approach. Random forests are less susceptible to overfitting compared with conventional classification trees.²⁷ Rule-based methods were initially based on the descriptive characteristics of the study sample and a priori clinical knowledge, and were subsequently refined based on evaluation of the performances of all algorithms.

Evaluation of algorithms

Algorithms were examined with respect to sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) and corresponding 95% CIs. Three algorithms were selected to give minimum, moderate and maximum estimates for prevalence and incidence rates based on their performance characteristics: a low sensitivity, moderate PPV algorithm (minimum estimate); a moderate sensitivity, moderate PPV algorithm (moderate estimate); and a high sensitivity, low PPV algorithm (maximum estimate).

Prevalence and incidence of T1D

The three selected algorithms were then applied to the entire Ontario population to determine yearly crude and age-standardized and sex-standardized prevalence and incidence rates for T1D per 1000 person-years of follow-up time from 1 April 2010 to 31 March

Table 1 Descriptive characteristics of the EMRPC reference population by diabetes type

Characteristics	Abstracted T1D (n=195)	Abstracted T2D (n=240)	Not abstracted (n=4967)
EMR variables			
Age at index, median (IQR)	42 (31–54)	67 (56–76)	66 (57–75)
Male	95 (48.7%)	131 (54.6%)	2667 (53.7%)
Cumulative Patient Profile terms*			
Definite T1D	152 (77.9%)	<6	37 (0.7%)
Possible T1D	21 (10.6%)	11 (4.6%)	17 (0.3%)
T2D	8 (4.0%)	142 (59.2%)	3017 (60.7%)
Renal insufficiency†	13 (6.5%)	26 (10.8%)	103 (2.1%)
Any insulin use	195 (100.0%)	240 (100.0%)	831 (16.7%)
Bolus insulin use	172 (88.2%)	160 (66.7%)	359 (7.2%)
Metformin use	14 (7.2%)	142 (59.2%)	3814 (76.8%)
Non-insulin, non-metformin antihyperglycemic agent use	0 (0%)	0 (0.0%)	2378 (47.9%)
BMI (mean±SD)	27±7	32±7	32±7
Administrative data variables			
Age at diabetes incidence, median (IQR)	25 (15–34)	50 (42–58)	56 (47–64)
Pediatric diabetes	58 (29.7%)	0 (0%)	14 (0.3%)
Any insulin use‡	48 (92.3%)	133 (90.5%)	457 (16.3%)
Bolus insulin use‡	47 (97.9%)	116 (78.9%)	242 (53.0%)
Metformin use‡	<6	79 (53.7%)	1722 (61.6%)
Non-insulin, non-metformin glycemic agent use‡	<6	<6	1045 (37.4%)
Insulin pump	63 (32.3%)	<6	17 (0.3%)
Hospital codes			
T1D DKA (ICD-10)	38 (19.5%)	<6	11 (0.2%)
T2D DKA (ICD-10)	8 (4.1%)	13 (5.4%)	27 (0.5%)
Any DKA (ICD-9+ICD-10)	58 (29.7%)	15 (6.3%)	43 (0.9%)
T1D	46 (23.6%)	13 (5.4%)	49 (1.0%)
T2D	42 (21.5%)	172 (71.7%)	2266 (45.6%)
eGFR (mean±SD)	101.4±24.9	75.1±28.0	80.3±21.7

*Definite T1D: T1D, T1DM, latent autoimmune diabetes, LADA, juvenile diabetes, juvenile-onset diabetes, juvenile-onset diabetes, diabète de type 1, diabète juvenile, diabète type 1, DB1. Possible T1D: type 1, IDDM, insulin-dependent diabetes, insulin-dependent diabetes, Schmidt, polyglandular. T2D: type 2 diabetes, T2D, T2DM, NIDDM, mature onset, diabète du type 2, non-insulin-dependent diabetes, non-insulin-dependent diabetes, non-insulin-dependent diabetes, non-insulin-dependent diabetes, diet-controlled diabetes, diet-controlled diabetes, diabète type 2, DB2.

†Renal insufficiency defined as eGFR ≤30 (or if end-stage renal disease documented for abstracted charts).

‡The denominator for medication use using administrative data was individuals who had any medication claim within ±365 days of index date, since not all individuals have medication data available in the administrative data.

BMI, body mass index; DB1, diabète type 1; DB2, diabète type 2; DKA, diabetic ketoacidosis; eGFR, estimated glomerular filtration rate; EMR, electronic medical record; EMRPC, Electronic Medical Records Primary Care; ICD, International Classification of Diseases; IDDM, Insulin-dependent diabetes mellitus; LADA, Latent Autoimmune Diabetes in Adults; NIDDM, non-insulin dependent diabetes mellitus; T1D, type 1 diabetes; T2D, type 2 diabetes; T1DM, type 1 diabetes mellitus; T2DM, type 2 diabetes mellitus.

2017. Denominators for these rates were individuals who were alive and eligible for healthcare at the start of each fiscal year (which begins on 1 April each year and concludes on 31 March the following year), and numerators were the number of individuals who met the criteria for the specified algorithm. Rates were standardized to the 1991 Canadian census population.

Sample size calculation

A sample size calculation was performed a priori and abstraction of 385 charts permitted estimation of the

95% CI for sensitivity and specificity with a maximum width of ±4%.²⁸ The sample size for manual chart abstraction could not be estimated precisely a priori since medication criteria were applied to a random subset of the entire diabetes population, and the actual sample size was larger than our sample size requirement calculated.

Statistical analysis

Intrarater and inter-rater reliability were assessed by unweighted Cohen's kappa. For the classification tree analyses, the R package rpart (V.4.1–8) was used to grow

and prune a classification tree with the outcome being T1D.^{26,29} A stopping rule was that a node did not undergo subsequent splits once the number of subjects in the node was less than 10. The minimum cross-validated mean error was used as a criterion in pruning the final tree. The bootstrapping approach of Harrell *et al*³⁰ was applied to provide optimism-corrected estimates of performance. Random forests were performed using the R package randomForest (V.4.5–36), and missing data were handled in multiple ways: complete case analysis, simple imputation with mean/median, or imputation using rfimpute.³¹

Trends in prevalence and incidence rates per 1000 person-years of follow-up were analyzed by negative binomial regression with follow-up time as an offset, comparing the rate in each year with 2010 as a reference. An Omnibus likelihood ratio was calculated for each negative binomial regression and model fit was verified. This process was repeated for each set of estimates (minimum, moderate and maximum). R V.3.1.2 (31 October 2014) was used for performing classification tree and random forest analyses.³² SAS V.9.4 was used for all other analyses.

Sensitivity analyses

As a post-hoc sensitivity analysis, the performances of selected algorithms in the LMC Diabetes Registry were evaluated when stratified by age as ≤ 40 or >40 years old.

Administrative healthcare data in Ontario date back to 1991; thus, in order to identify an individual as having incident diabetes younger than 19 years old, age at index date would have had to be 42 years or younger.

RESULTS

Reference population characteristics

There were 26 238 individuals with diabetes in EMRPC as of 30 September 2015, and 21 518 remained after applying the inclusion and exclusion criteria (figure 1). The reference population included 5402 individuals with diabetes (a random 25% subsample), of whom 435 had charts abstracted for meeting the criteria for possible T1D and 4967 were included as having T2D without chart abstraction based on medication use alone. Of the 435 patients with possible T1D, 195 (44.8%) had definite T1D based on chart abstraction. Thus, the overall proportion of T1D in this sample of primary care patients with diabetes was 3.6%. Unweighted Cohen's kappa for intrarater and inter-rater agreement was 0.96 and 0.82, respectively. Descriptive characteristics based on EMR and administrative data are reported in table 1.

The LMC Diabetes Registry was used for validation of administrative healthcare data algorithms in an independent sample population. There were 29 371 individuals in the registry, of whom 2379 had T1D and 26 992 had T2D

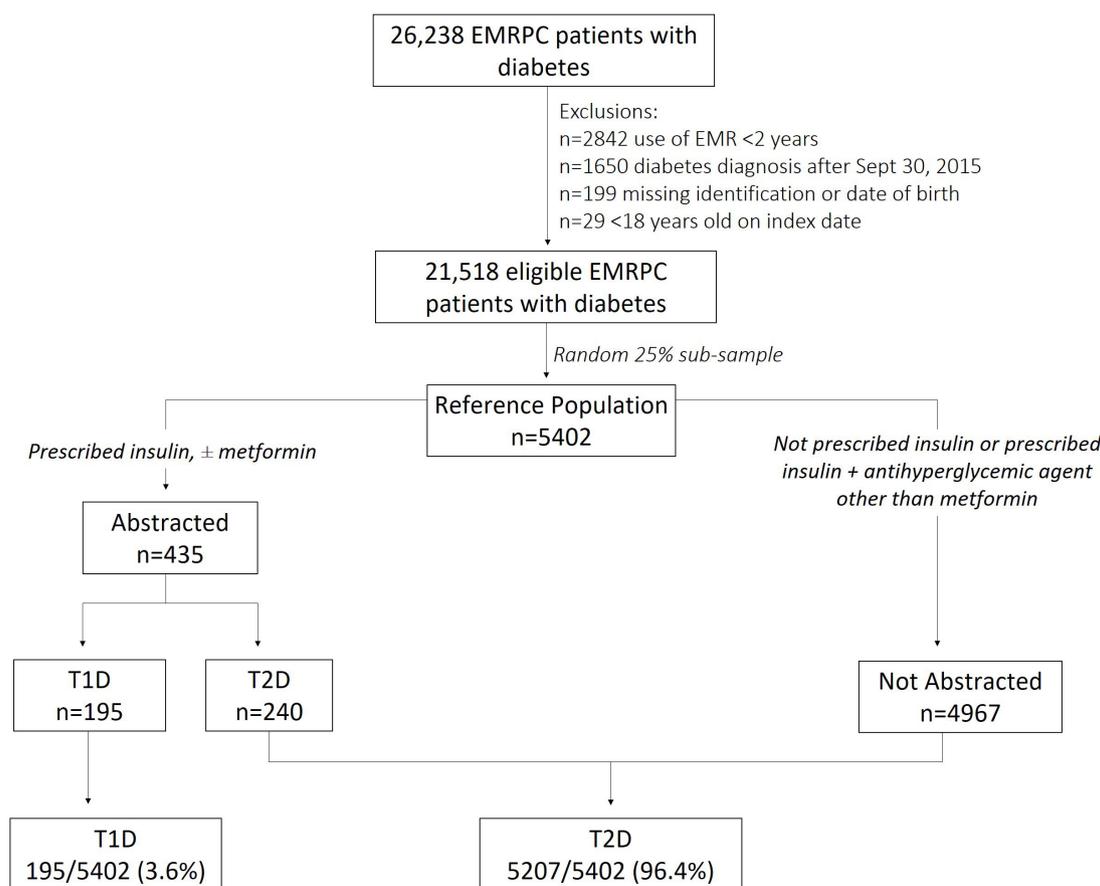


Figure 1 Selection of subjects for the reference population. EMR, electronic medical record; EMRPC, Electronic Medical Records Primary Care; T1D, type 1 diabetes; T2D, type 2 diabetes.

Table 2 Performances of selected algorithms*

Algorithm description	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)
EMR algorithms				
EMR _{Tree} †	80.6 (75.9 to 87.2)	99.8 (99.7 to 100)	94.9 (92.3 to 98.7)	99.3 (99.1 to 99.5)
EMR _{Forest} ‡	83.1 (77.1 to 88.1)	99.8 (99.6 to 99.9)	93.1 (88.3 to 96.)	99.4 (99.1 to 99.6)
EMR _{Rule 1} : definite T1D CPP + no non-insulin, non-metformin glycaemic agent	77.9 (71.5 to 83.6)	99.5 (99.3 to 99.7)	86.4 (80.4 to 91.1)	99.2 (98.9 to 99.4)
EMR _{Rule 2} : definite T1D CPP or pump	81.5 (75.4 to 86.7)	99.2 (98.9 to 99.4)	79.1 (72.8 to 84.5)	99.3 (99 to 99.5)
EMR _{Rule 3} : definite T1D CPP + no T2D CPP	76.9 (70.4 to 82.6)	99.4 (99.1 to 99.6)	82 (75.6 to 87.2)	99.1 (98.8 to 99.4)
EMR _{Rule 4} : (definite T1D CPP + insulin + no non-insulin, non-metformin glycaemic agent) or (possible T1D CPP + age at index <45)	81 (74.8 to 86.3)	97.5 (94.6 to 99.1)	96.3 (92.2 to 98.6)	86.3 (81.7 to 90.2)
EMR _{Rule 5} : definite T1D CPP	77.9 (71.5 to 83.6)	98.8 (96.4 to 99.7)	98.1 (94.4 to 99.6)	84.6 (79.9 to 88.7)
EMR _{Rule 6} : definite T1D CPP or BMI <25	85.6 (79.9 to 90.2)	89.6 (85 to 93.1)	87 (81.4 to 91.4)	88.5 (83.8 to 92.2)
Administrative data algorithms				
Admin _{Tree} †	61.4 (57.5 to 71.3)	99.2 (99.0 to 99.5)	73.9 (70.1 to 83.5)	98.6 (98.3 to 99.0)
Admin _{Forest} ‡	58.0 (50.7 to 65.0)	99.3 (99.0 to 99.5)	74.8 (67.1 to 81.5)	98.4 (98.1 to 98.8)
Admin _{Rule 1} : pediatric diabetes or pump or T1D DKA	58.5 (51.2 to 65.5)	99.3 (99.1 to 99.5)	76.5 (68.9 to 83.1)	98.5 (98.1 to 98.8)
Admin _{Rule 2} : pediatric diabetes or pump	51.3 (44.0 to 58.5)	99.5 (99.3 to 99.7)	79.4 (71.2 to 86.1)	98.2 (97.8 to 98.5)
Admin _{Rule 3} : (diabetes incidence <30 + (pediatric diabetes or any DKA)) or pump or T1D DKA	61.5 (54.3 to 68.4)	99.3 (99.0 to 100)	75.9 (68.5 to 100)	98.6 (98.2 to 98.9)
Admin _{Rule 4} : diabetes incidence <30 or pump or any DKA	80.0 (73.7 to 85.4)	96.7 (96.2 to 97.2)	47.6 (42.0 to 53.1)	99.2 (99.0 to 99.5)
Admin _{Rule 5} : diabetes incidence <30 or (diabetes incidence <40 and no T2D) or pump or any DKA	91.3 (86.4 to 94.8)	92.6 (91.8 to 93.3)	31.5 (27.7 to 35.5)	99.6 (99.4 to 99.8)
Combined EMR and administrative data algorithms				
EMR+Admin _{Tree} †	80.9 (76.5 to 87.6)	99.9 (99.8 to 100)	95.7 (93.1 to 99.0)	99.3 (99.0 to 99.5)
EMR+Admin _{Forest} ‡	88.2 (82.8 to 92.4)	99.8 (99.7 to 100)	95.0 (90.8 to 97.7)	99.6 (99.3 to 99.7)
EMR+Admin _{Rule 1} : insulin + no non-insulin, non-metformin glycaemic agent + (definite T1D CPP or pump or pediatric diabetes)	87.2 (81.7 to 91.5)	99.9 (99.7 to 100)	96.6 (92.7 to 98.7)	99.5 (99.3 to 99.7)
EMR+Admin _{Rule 2} : insulin + no non-insulin, non-metformin glycaemic agent + (pediatric diabetes or pump)	51.3 (44.0 to 58.5)	100 (99.9 to 100)	98 (93.1 to 99.8)	98.2 (97.8 to 98.5)
EMR+Admin _{Rule 3} : bolus insulin + no non-insulin, non-metformin glycaemic agent + (definite T1D CPP or pump or pediatric diabetes)	79.5 (73.1 to 84.9)	99.9 (99.7 to 100)	96.3 (92.1 to 98.6)	99.2 (99 to 99.5)
EMR+Admin _{Rule 4} : definite T1D CPP + insulin + no non-insulin, non-metformin glycaemic agent	77.4 (70.9 to 83.1)	99.9 (99.8 to 100)	96.8 (92.7 to 99)	99.2 (98.9 to 99.4)

*EMRPC reference standard included 5402 individuals (195 with T1D, 5207 with T2D).

†Classification tree estimates were optimism-adjusted.

‡Random forest used rImpute for missing data (age at diabetes incidence, eGFR and BMI).

BMI, body mass index; CPP, Cumulative Patient Profile; DKA, diabetic ketoacidosis; eGFR, estimated glomerular filtration rate; EMR, electronic medical record; EMRPC, Electronic Medical Records Primary Care; NPV, negative predictive value; PPV, positive predictive value; T1D, type 1 diabetes; T2D, type 2 diabetes.

(proportion of T1D out of all diabetes cases: 8.1%; the characteristics are reported in online supplementary table 3).

EMR data algorithms

The best performing EMR algorithm was derived using a classification tree (EMR_{Tree} in table 2 and figure 2A), which included the following variables: definite or possible T1D terms in the CPP, insulin, absence of non-insulin, non-metformin antihyperglycaemic agent, and age at index date. This algorithm had a sensitivity of 80.6% (95% CI 75.9 to 87.2), specificity of 99.8% (99.7 to 100), PPV of 94.9% (92.3 to 98.7) and NPV of 99.3% (99.1 to 99.5). The random forest (EMR_{Forest}) had similar performance to the classification tree, although with a slightly lower PPV. Variable importance factors for the random forest are reported in online supplementary table 4. Rule-based algorithms did not perform as well as the classification tree and random forest. The best performing algorithm using rule-based

methods was EMR_{Rule 3}, which included definite T1D terms in the CPP and absence of insulin and antihyperglycaemic agents other than metformin (sensitivity 77.9% (71.5 to 83.6), specificity 99.5% (99.3 to 99.7), PPV 86.4% (80.4 to 91.1), and NPV 99.2% (98.9 to 99.4)).

Performances of administrative healthcare data algorithms

The best performing algorithm (ie, highest PPV) was Admin_{Rule 2}, which was a combination including pediatric diabetes or insulin pump (sensitivity 51.3% (44 to 58.5), specificity 99.5% (99.3 to 99.7), PPV 79.4% (71.2 to 86.1), and NPV 98.2% (97.8 to 98.5)). Admin_{Rule 3}, which was a simple combination including age at diabetes incidence younger than 30 years old, pediatric diabetes, DKA and insulin pump use, had a higher sensitivity of 61.5% (54.3 to 68.4) but lower PPV of 75.9% (68.5 to 100). The classification tree (Admin_{Tree}; figure 2B) and random forest (Admin_{Forest}; online

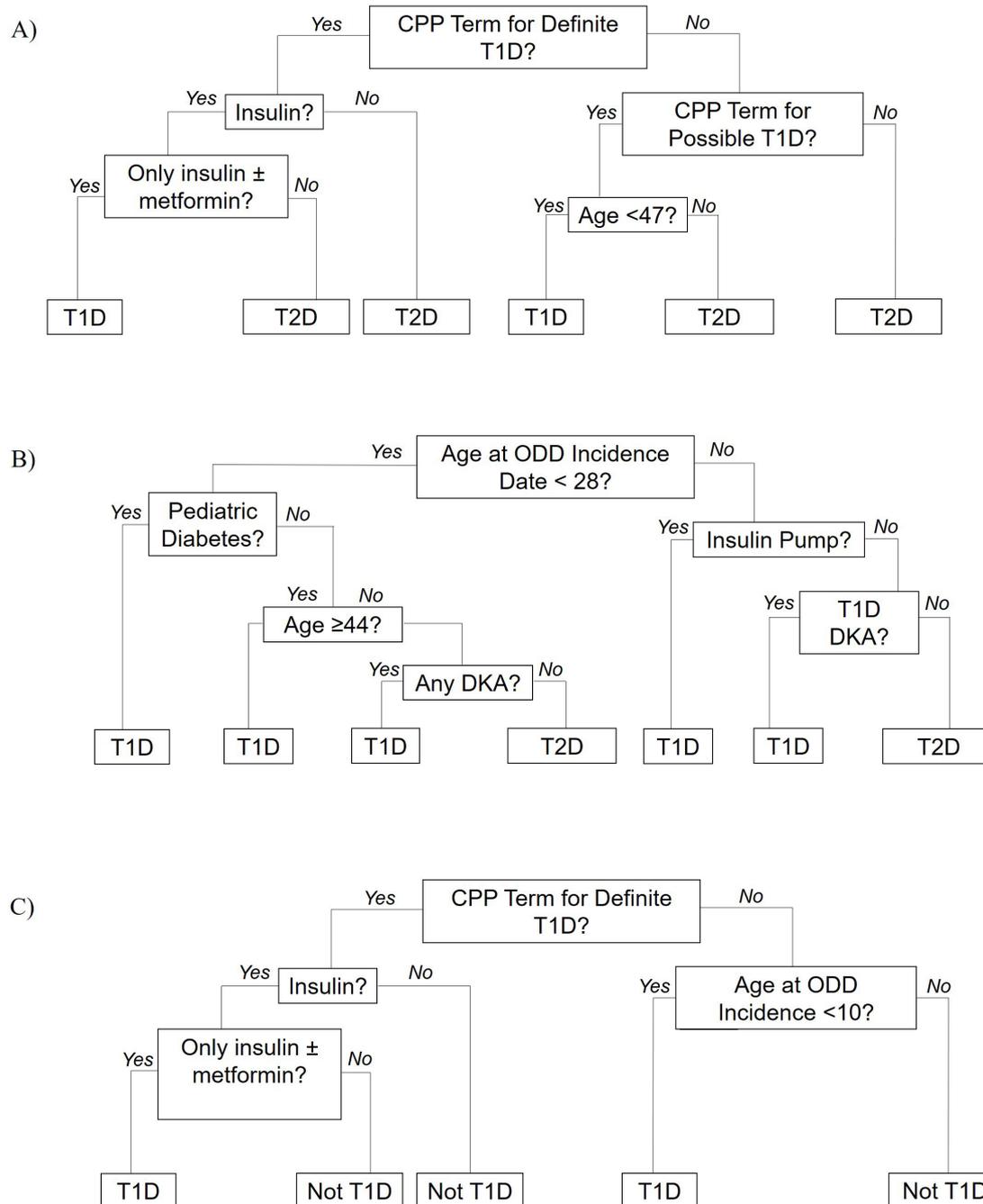


Figure 2 Classification tree algorithms for EMR data (A), administrative data (B), and combined EMR and administrative data (C). CPP, Cumulative Patient Profile; DKA, diabetic ketoacidosis; EMR, electronic medical record; ODD, Ontario Diabetes Database; T1D, type 1 diabetes; T2D, type 2 diabetes.

supplementary table 4) had similar performances, but both had lower PPV than the optimally performing algorithms using rule-based algorithms.

Sensitivity and specificity for all algorithms described above were similar in the LMC Diabetes Registry compared with EMRPC. However, PPVs were consistently higher and NPVs slightly lower in the LMC sample, as expected based on the higher proportion of patients with T1D in this sample (online supplementary table 5). For example, Admin_{Rule 2} had a sensitivity of 58.9% (56.9 to 60.9), specificity of 99.6% (99.5 to 99.7), PPV of 93.2% (91.8 to 94.4) and NPV of 98.5%

(96.3 to 96.7), and Admin_{Rule 3} had a sensitivity of 65.4% (63.4 to 67.3), specificity of 99.3% (99.2 to 99.4), PPV of 89.5% (87.9 to 90.9) and NPV of 97% (96.8 to 97.2).

Combined EMR and administrative data algorithms

A rule-based algorithm including insulin, absence of antihyperglycemic medication other than metformin, definite T1D terms in the CPP, insulin pump or pediatric diabetes (EMR+Admin_{Rule 1}) was the best performing algorithm, with a sensitivity of 87.2% (81.7 to 91.5), specificity of 99.9% (99.7 to 100), PPV of 96.6% (92.7 to 98.7), and NPV of 99.5% (99.3 to 99.7).

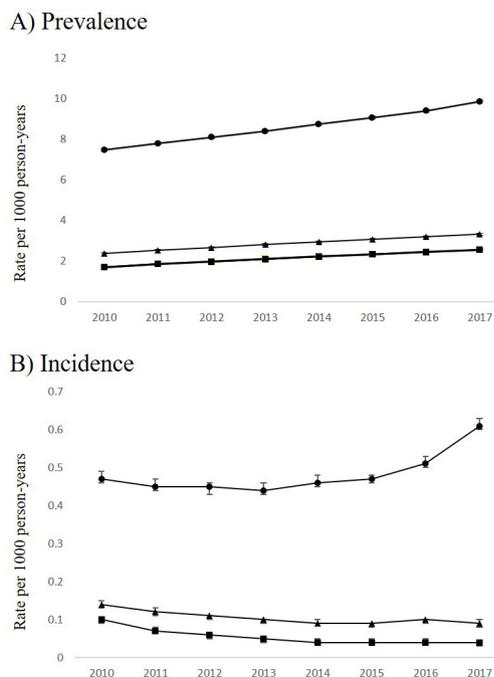


Figure 3 Age-standardized and sex-standardized (A) prevalence and (B) incidence trends of T1D in Ontario, Canada per 1000 person-years*. Legend: closed circles: high sensitivity, low PPV algorithm (Admin_{Rule 4}); closed triangles: moderate sensitivity, moderate PPV algorithm (Admin_{Rule 3}); closed squares: low sensitivity, moderate PPV algorithm (Admin_{Rule 2}). *The denominator for determination of prevalence rates was person-years of follow-up for all eligible subjects in each fiscal year. The denominator for determination of incidence rates was person-years of follow-up for eligible subjects excluding those previously identified with T1D prior to the start of each fiscal year. PPV, positive predictive value; T1D, type 1 diabetes.

Prevalence and incidence of T1D in Ontario

Three algorithms were applied to the Ontario population to estimate the minimum, moderate, and maximum prevalence and incidence rates of T1D from 2010 to 2017 (online supplementary tables 6–8). The algorithms selected were (1) Admin_{Rule 2}: a low sensitivity, moderate PPV algorithm; (2) Admin_{Rule 3}: a moderate sensitivity, moderate PPV algorithm; and (3) Admin_{Rule 4}: a high sensitivity, low PPV algorithm. The number of individuals with prevalent T1D in 2017 ranged from 24 789 (minimum estimate using Admin_{Rule 2}) to 102 140 (maximum estimate using algorithm 14). Age-standardized and sex-standardized prevalence rates increased between 2010 and 2017, with the minimum estimate being an increase from 1.7 (95% CI 1.68 to 1.73) to 2.56 (2.53 to 3.6) per 1000 person-years using Admin_{Rule 2}, and the maximum estimate being an increase from 7.48 (7.43 to 7.54) to 9.86 (9.79 to 9.92) per 1000 person-years using Admin_{Rule 4}. Overall, this represented a significant increase in T1D prevalence of 32%–50% ($p < 0.0001$ for Omnibus likelihood ratio). Using Admin_{Rule 2} and Admin_{Rule 3}, age-standardized and sex-standardized incidence decreased between 2010 and 2014 and subsequently remained

stable (figure 3), with the incidence rate per 1000 person-years in 2017 being between 0.04 (0.03–0.04) using the minimum estimate and 0.09 (0.08–0.1) using the moderate estimate (Omnibus likelihood ratio $p < 0.0001$). However, using Admin_{Rule 4} (maximum estimate), incidence rates for T1D decreased from 0.47 (0.46–0.49) per 1000 person-years in 2010 to a nadir of 0.44 in 2013, then increased to 0.61 in 2017 (Omnibus likelihood ratio $p < 0.0001$). Using Admin_{Rule 2} and Admin_{Rule 3}, the estimated number of individuals newly diagnosed with T1D in 2017 in Ontario was between 397 and 1013.

Sensitivity analyses

The following EMR variables had missing data: body mass index (12.9% missing) and estimated glomerular filtration rate (11.6% missing). The following administrative data variables had missing data: age at diabetes incidence (3.7%) and estimated glomerular filtration rate (22%). For all random forests, multiple approaches for missing data were evaluated (complete case analysis, simple imputation using means/medians, and imputation using rfimpute), and all results were consistent with the primary analysis. Sensitivities of algorithms were higher in individuals ≤ 40 years old than in those > 40 years old, whereas specificities were similar (online supplementary table 9; for example, sensitivity 73.0% (70.4 to 75.5) vs 44.1% (41.2 to 47) for Admin_{Rule 2}, and 77.7% (75.3 to 80) vs 52.3% (49.4 to 55.2) for Admin_{Rule 3}).

CONCLUSIONS

We derived and validated algorithms identifying T1D, among those with diabetes, using primary care EMR data, administrative healthcare data, and EMR combined with administrative healthcare data in Ontario, Canada. Algorithms using EMR data alone or EMR combined with administrative data were able to identify T1D with excellent performance. However, algorithms using administrative data alone had important limitations, namely low sensitivity and only moderate PPV.

In our study, rule-based algorithms outperformed or performed similarly to classification trees and random forests. It is important to note that the best performing rule-based algorithms mimicked classification trees, after the classification tree had been developed. Thus, it is possible to use data-driven approaches such as classification trees to determine optimal variables for inclusion in an algorithm and subsequently develop a rule-based approach that may be simpler for implementation. Rule-based approaches also permitted us to select an algorithm prioritizing specificity and PPV, even at the expense of sensitivity. In contrast, classification trees and random forests generated a single algorithm that best balanced all parameters (sensitivity, specificity, PPV and NPV).

The EMR data algorithm had superior performance compared with previously published algorithms. One algorithm based on the ratio of billing codes for T1D versus T2D, prescriptions and blood test results had a

sensitivity of 65% and a PPV of 88% but did not report specificity or NPV.^{12 13} Lethebe and others used machine learning models to develop an algorithm that included text terms for T1D diagnosis and age at meeting criteria for diabetes diagnosis, and had a sensitivity of 43%, specificity of 99%, PPV of 85% and NPV of 95%.¹¹ Another algorithm that considered diagnostic codes, medications, and age at diabetes incidence had perfect agreement with a gold standard for diagnosis of T1D versus T2D, but likely included a small number of subjects with T1D.¹⁴ Finally, a classification tree that included medications, DKA, billing codes, and age had a sensitivity of 92.8%, specificity of 99.3%, PPV of 89.5%, and NPV of 99.5%.¹⁵

We used three candidate administrative healthcare data algorithms to evaluate the minimum, moderate, and maximum estimates of the number of prevalent and incident T1D cases. All three algorithms demonstrated increasing prevalence of T1D from 2010 to 2017. We observed decreasing incidence of T1D between 2010 and 2017 using the minimum and moderate estimate algorithms. However, uptake of insulin pumps would be expected to have been highest shortly after the funding program was initiated in 2008, and therefore early ‘incident’ cases may actually reflect individuals with prevalent T1D who were newly starting insulin pump therapy. Interestingly, the moderate sensitivity, low PPV (maximum estimate) algorithm demonstrated a divergent trend in T1D incidence, with rates decreasing until 2014 but subsequently increasing. Since this algorithm includes DKA that was not specified to be associated with T1D or T2D, we hypothesize that the increasing incidence may reflect rising rates of DKA in patients with T2D taking sodium-glucose transport protein 2 inhibitors, who were misclassified as having T1D by this algorithm.³³ Accounting for these factors, incidence rates of T1D in Ontario appear to be stable.

Worldwide, the incidence of T1D in children and youth is reported as $\leq 0.5\%$ of the population.^{5–7} Incidence in children has generally been found to be increasing, although in some geographical regions there have been reports of a plateau in incidence rates.^{5 7 34–39} Our estimates for T1D incidence rates in adults were lower than those reported for children, which is not surprising since the majority of T1D cases are diagnosed in childhood or adolescence.⁴⁰ Less information on T1D incidence trends in adults is available, but our results are consistent with decreasing incidence rates in individuals over the age of 15 and 25 in Belgium, Lithuania and Sweden.^{6 8 41 42} In these countries, a concurrent increase in T1D incidence in younger ages has been noted, suggesting a shift in diagnosis to younger ages rather than a true increase in overall incidence of T1D. Although T1D prevalence rates in adults are not explicitly reported in the literature to our knowledge, our findings of increasing prevalence are consistent with reports of decreasing mortality in individuals with T1D and perhaps a shift in age at diagnosis to younger ages.⁴³

There are some important considerations in the application of the administrative healthcare data algorithms derived in this study. First, the proportion of patients with T1D in the reference population may have been lower than the true population prevalence since some individuals with T1D only see specialists and not primary care physicians. Indeed, the proportion of diabetes cases in EMRPC assigned as being T1D (3.6%) was lower than observed in other populations, which is generally quoted as 5%–10% of all diabetes cases.^{4 13 14} If the true population proportion of T1D is higher, then the PPV of applied algorithms would also be expected to be higher. Second, administrative healthcare data in Ontario date back to 1991, which means we could not determine if diabetes incidence criteria were met during childhood for individuals who were 42 years or older at the study end-date. This explains why individuals older than age 44 with incident diabetes younger than 28 years old but who did not meet the criteria for pediatric diabetes were classified as having T1D in the classification tree using administrative healthcare data (*figure 2B*), since it could not have been determined if these individuals met the criteria for pediatric diabetes. Thus, the algorithm performs better in younger individuals, and we expect that as the retrospective availability of data lengthens, algorithm performance will improve for older ages. Third, the algorithm relies on a database of insulin pump users that is exclusive for T1D, which may not be available in all settings. Finally, we did not assess alternative methods for building predictive models, such as logistic regression or machine learning approaches.

Our study has a number of strengths. The sample size for the reference population was large because automated search criteria based on medications were used to first identify individuals with possible T1D. In addition, we validated the algorithms in two different sample populations. We also evaluated multiple methods for deriving algorithms (classification trees, random forests, and rule-based methods). Limitations of our study include possible misclassification of T1D as T2D in the charts that were not manually abstracted, although the number of false-negatives is likely to be low since antihyperglycemic medications other than metformin were rarely used in T1D care prior to 2015. In addition, internal validation of the algorithms (eg, splitting the sample into derivation and validation sets) was not performed due to sample size concerns; however, classification tree estimates were adjusted for optimism, which corrects for the tendency of predictive models to perform better in training data sets than external data. Finally, algorithms with the highest PPV had only modest sensitivity, which led us to evaluate ranges of plausible estimates for prevalence and incidence rates.

In summary, we have derived and validated algorithms identifying T1D that have excellent performance using primary care EMR data and combined EMR and administrative data. Algorithms using administrative data alone have modest performance, but the benefit of being able

to be applied to an entire population. Application of these algorithms demonstrated increasing prevalence of T1D in Ontario since 2010 and stable incidence. These algorithms will permit further study of the epidemiology, healthcare utilization, and outcomes of T1D in large populations.

Author affiliations

¹ICES, Toronto, Ontario, Canada

²Division of Endocrinology & Metabolism, Department of Medicine, University of Toronto, Toronto, Ontario, Canada

³Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, Ontario, Canada

⁴Department of Family and Community Medicine, University of Toronto, Toronto, Ontario, Canada

⁵Toronto Western Hospital Family Health Team, University Health Network, Toronto, Ontario, Canada

⁶North York General Hospital, Toronto, Ontario, Canada

⁷Women's College Research Institute, Women's College Hospital, Toronto, Ontario, Canada

⁸LMC Diabetes & Endocrinology, Toronto, Ontario, Canada

⁹Li Ka Shing Knowledge Institute of St. Michael's Hospital, Toronto, Ontario, Canada

Twitter Alanna Weisman @AlannaWeisman

Acknowledgements The authors would like to acknowledge Esha Homenauth and Vicki Ling from ICES, and Ruth Brown from LMC Diabetes & Endocrinology, for preparing the data. The authors thank IMS Brogan for use of their Drug Information Database.

Contributors AW, JY and MK researched and analyzed the data. AW, GLB, LJ, KT, PCA and LL contributed to study design. AW conducted the literature search and wrote the first draft of the manuscript. All authors interpreted the data, provided critical revisions, and approved the final version. AW and GLB are the guarantors for this study and take full responsibility for the work as a whole.

Funding This project was funded by Diabetes Action Canada and the Ontario SPOR Support Unit. This study was supported by ICES, a research institute affiliated with the University of Toronto which is funded by an annual grant from the Ontario Ministry of Health and Long-Term Care (MOHLTC). The opinions, results and conclusions reported in this paper are those of the authors and are independent from the funding sources. No endorsement by ICES or the Ontario MOHLTC is intended or should be inferred. Parts of this material are based on data and/or information compiled and provided by CIHI. However, the analyses, conclusions, opinions and statements expressed in the material are those of the author(s), and not necessarily those of CIHI. AW is supported by a Canadian Institutes of Health Research Fellowship and a Diabetes Canada fellowship, and the Clinician-Scientist Training Program in the Department of Medicine at the University of Toronto. GLB is supported by a Clinician-Scientist Merit Award from the Department of Medicine at the University of Toronto. PCA is supported by a Mid-Career Investigator Award from the Heart and Stroke Foundation. KT and LJ are supported by Research Scholar Awards from the Department of Family and Community Medicine at the University of Toronto.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available. Data are held at ICES and are not publicly available.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iD

Alanna Weisman <http://orcid.org/0000-0003-3778-653X>

REFERENCES

- Foster NC, Beck RW, Miller KM, *et al.* State of type 1 diabetes management and outcomes from the T1D exchange in 2016-2018. *Diabetes Technol Ther* 2019;21:66-72.
- Johnson JA, Pohar SL, Majumdar SR. Health care use and costs in the decade after identification of type 1 and type 2 diabetes: a population-based study. *Diabetes Care* 2006;29:2403-8.
- Diabetes Control and Complications Trial Research Group, Nathan DM, Genuth S, *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* 1993;329:977-86.
- Saydah S, Imperator G. Emerging approaches in surveillance of type 1 diabetes. *Curr Diab Rep* 2018;18:61.
- Patterson CC, Harjutsalo V, Rosenbauer J, *et al.* Trends and cyclical variation in the incidence of childhood type 1 diabetes in 26 European centres in the 25 year period 1989-2013: a multicentre prospective registration study. *Diabetologia* 2019;62:408-17.
- Dahlquist GG, Nyström L, Patterson CC, *et al.* Incidence of type 1 diabetes in Sweden among individuals aged 0-34 years, 1983-2007: an analysis of time trends. *Diabetes Care* 2011;34:1754-9.
- Haynes A, Bulsara MK, Jones TW, *et al.* Incidence of childhood onset type 1 diabetes in Western Australia from 1985 to 2016: evidence for a plateau. *Pediatr Diabetes* 2018;19:690-2.
- Diaz-Valencia PA, Bougnères P, Valleron A-J. Global epidemiology of type 1 diabetes in young adults and adults: a systematic review. *BMC Public Health* 2015;15:255.
- 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.
- Banda JM, Seneviratne M, Hernandez-Boussard T, *et al.* Advances in electronic phenotyping: from rule-based definitions to machine learning models. *Annu Rev Biomed Data Sci* 2018;1:53-68.
- Lethebe BC, Williamson T, Garies S, *et al.* Developing a case definition for type 1 diabetes mellitus in a primary care electronic medical record database: an exploratory study. *CMAJ Open* 2019;7:E246-51.
- Klompas M, Eggleston E, McVetta J, *et al.* Automated detection and classification of type 1 versus type 2 diabetes using electronic health record data. *Diabetes Care* 2013;36:914-21.
- Schroeder EB, Donahoo WT, Goodrich GK, *et al.* Validation of an algorithm for identifying type 1 diabetes in adults based on electronic health record data. *Pharmacoepidemiol Drug Saf* 2018;27:1053-9.
- Sharma M, Petersen I, Nazareth I, *et al.* An algorithm for identification and classification of individuals with type 1 and type 2 diabetes mellitus in a large primary care database. *Clin Epidemiol* 2016;8:373-80.
- Lo-Ciganic W, Zgibor JC, Ruppert K, *et al.* Identifying type 1 and type 2 diabetic cases using administrative data: a tree-structured model. *J Diabetes Sci Technol* 2011;5:486-93.
- de Lusignan S, Khunti K, Belsey J, *et al.* A method of identifying and correcting miscoding, misclassification and misdiagnosis in diabetes: a pilot and validation study of routinely collected data. *Diabet Med* 2010;27:203-9.
- 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.
- Zhong VW, Obeid JS, Craig JB, *et al.* An efficient approach for surveillance of childhood diabetes by type derived from electronic health record data: the search for diabetes in youth study. *J Am Med Assoc* 2016;23:1060-7.
- Benchimol EI, Manuel DG, To T, *et al.* Development and use of reporting guidelines for assessing the quality of validation studies of health administrative data. *J Clin Epidemiol* 2011;64:821-9.
- Ivers N, Pylypenko B, Tu K. Identifying patients with ischemic heart disease in an electronic medical record. *J Prim Care Community Health* 2011;2:49-53.
- Tu K, Manuel D, Lam K, *et al.* Diabetics can be identified in an electronic medical record using laboratory tests and prescriptions. *J Clin Epidemiol* 2011;64:431-5.
- Ivers NM, Tu K, Young J, *et al.* Feedback gap: pragmatic, cluster-randomized trial of goal setting and action plans to increase the effectiveness of audit and feedback interventions in primary care. *Implement Sci* 2013;8:142.
- Jones SR, Carley S, Harrison M. An introduction to power and sample size estimation. *Emerg Med J* 2003;20:453-8.
- Aronson R, Orzech N, Ye C, *et al.* Specialist-Led diabetes registries and predictors of poor glycemic control in type 2 diabetes: insights into the functionally refractory patient from the LMC diabetes registry database. *J Diabetes* 2016;8:76-85.

- 25 Guttman 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.
- 26 Terry Therneau BAaBR. rpart: recursive partitioning and regression trees. R package version 4.1-8 2014.
- 27 Henrard S, Speybroeck N, Hermans C. Classification and regression tree analysis vs. multivariable linear and logistic regression methods as statistical tools for studying haemophilia. *Haemophilia* 2015;21:715–22.
- 28 Hahn GJ, Meeker WQ. *Statistical intervals*. John Wiley & Sons, Inc, 1991.
- 29 MKe al. caret: classification regression training. R package version 6.0-41 2015.
- 30 Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996;15:361–87.
- 31 ALaM W. Classification and regression by randomForest. *R News* 2002;2:18–22.
- 32 Team RC. *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, 2014.
- 33 Fralick M, Schneeweiss S, Paterno E. Risk of diabetic ketoacidosis after initiation of an SGLT2 inhibitor. *N Engl J Med* 2017;376:2300–2.
- 34 Mayer-Davis EJ, Lawrence JM, Dabelea D, *et al.* Incidence trends of type 1 and type 2 diabetes among Youths, 2002-2012. *N Engl J Med* 2017;376:1419–29.
- 35 Newhook LA, Grant M, Sloka S, *et al.* Very high and increasing incidence of type 1 diabetes mellitus in Newfoundland and Labrador, Canada. *Pediatr Diabetes* 2008;9:62–8.
- 36 Toth EL, Lee KC, Couch RM, *et al.* High incidence of IDDM over 6 years in Edmonton, Alberta, Canada. *Diabetes Care* 1997;20:311–3.
- 37 Blanchard JF, Dean H, Anderson K, *et al.* Incidence and prevalence of diabetes in children aged 0-14 years in Manitoba, Canada, 1985-1993. *Diabetes Care* 1997;20:512–5.
- 38 Legault L, Polychronakos C. Annual incidence of type 1 diabetes in Québec between 1989-2000 in children. *Clin Invest Med* 2006;29:10–13.
- 39 Fox DA, Islam N, Sutherland J, *et al.* Type 1 diabetes incidence and prevalence trends in a cohort of Canadian children and youth. *Pediatr Diabetes* 2018;19:501–5.
- 40 Tuomilehto J. The emerging global epidemic of type 1 diabetes. *Curr Diab Rep* 2013;13:795–804.
- 41 Weets I, De Leeuw IH, Du Caju MVL, *et al.* The incidence of type 1 diabetes in the age group 0-39 years has not increased in Antwerp (Belgium) between 1989 and 2000: evidence for earlier disease manifestation. *Diabetes Care* 2002;25:840–6.
- 42 Ostrauskas R, Žalinkevičius R, Jurgevičienė N, *et al.* The incidence of type 1 diabetes mellitus among 15-34 years aged Lithuanian population: 18-year incidence study based on prospective databases. *BMC Public Health* 2011;11:813.
- 43 Rawshani A, Rawshani A, Franzén S, *et al.* Mortality and cardiovascular disease in type 1 and type 2 diabetes. *N Engl J Med* 2017;376:1407–18.