Developing a case definition for type I diabetes mellitus in a primary care electronic medical record database: A validation study using a machine learning approach

Journal:	*CMAJ*
Manuscript ID	CMAJ-18-0579
Manuscript Type:	Research - Descriptive study
Date Submitted by the Author:	18-Jul-2018
Complete List of Authors:	Lethebe, Brendan; University of Calgary, Community Health Sciences Williamson, Tyler; University of Calgary Department of Community Health Sciences Garies, Stephanie; University of Calgary, Family Medicine McBrien, Kerry; University of Calgary, Family Medicine and Community Health Sciences Leduc, Charles; University of Calgary Department of Family Medicine Butalia, Sonia; University of Calgary, Medicine Soos, Boglarka; University of Calgary Department of Family Medicine Shaw, Marta; University of Calgary, Community Health Sciences Drummond, Neil; University of Alberta, Family Medicine; University of Calgary, Family Medicine
Keywords:	Community Medicine, Diabetes, Family Medicine, General Practice, Primary Care, Medical Informatics, Medical Education
More Detailed Keywords:	
Abstract:	Background: Identifying caseness in primary care electronic medical records (EMR) is important for surveillance, research, quality improvement and clinical care. We aimed to develop and validate a case definition for type 1 diabetes using artificial intelligence machine learning methods. Methods: EMR data from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) in Alberta were used from 2008-2016. Patients who were identified as having diabetes mellitus according to the existing CPCSSN case definition were selected and their family physician was asked to confirm diabetes subtype, forming the reference standard. We then applied machine learning to identify variables that correctly distinguish between type 1 and 2 cases. Results: After excluding non-diabetes patients, a total of 1309 people with diabetes were used; 110 of these were confirmed as type 1 by physicians. Two machine learning algorithms were found to be useful: 1) definition consisting of "type 1" text words or age <30 years with sensitivity 52.3% (95% CI 42.6, 61.7), specificity 99.3% (95% CI 98.7, 99.7), PPV 87.9% (95% CI 77.0, 94.3), NPV 95.8% (95% CI 94.5, 96.8); 2) definition consisting of combinations of medications, endocrinology-specific referrals, and age criteria with sensitivity 79.3% (95% CI 70.3, 86.2), specificity

	89.4% (95% CI 87.5, 91.1), PPV 40.7% (95% CI 34.2, 47.6), NPV 88.6% (95% CI 86.7, 90.2). Interpretation: The first algorithm may be useful for cohort creation and disease registry development given its high PPV. The second algorithm may be more appropriate for public health surveillance and epidemiology, having better sensitivity and reasonably high specificity.

Developing a case definition for type I diabetes mellitus in a primary care electronic medical record database: A validation study using a machine learning approach

Brendan Cord Lethebe, MSc, Department of Community Health Science, University of Calgary
Tyler Williamson, PhD, Department of Community Health Science, University of Calgary
Stephanie Garies, MSc, Department of Family Medicine, University of Calgary
Kerry McBrien, PhD, Department of Community Health Science, University of Calgary
Charles Leduc, PhD, Department of Family Medicine, University of Calgary
Sonia Butalia, PhD, Department of Medicine, University of Calgary
Boglarka Soos, MSc, Department of Family Medicine, University of Calgary
Marta Shaw, MSc, Department of Community Health Science, University of Calgary
Neil Drummond, PhD, Department of Family Medicine, University of Alberta, Department of Family Medicine, University of Calgary

Corresponding author: Neil Drummond, neil.drummond@ualberta.ca

Abstract

Background: Identifying caseness in primary care electronic medical records (EMR) is important for surveillance, research, quality improvement and clinical care. We aimed to develop and validate a case definition for type 1 diabetes using artificial intelligence machine learning methods.

Methods: EMR data from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) in Alberta were used from 2008-2016. Patients who were identified as having diabetes mellitus according to the existing CPCSSN case definition were selected and their family physician was asked to confirm diabetes subtype, forming the reference standard. We then applied machine learning to identify variables that correctly distinguish between type 1 and type 2 diabetes cases.

Results: After excluding those without diabetes from the study, a total of 1309 people with diabetes were used; 110 of these were confirmed as type 1 by physicians. Two machine learning algorithms were found to be useful: 1) definition consisting of "type 1" text words *or* age <30 years with sensitivity 52.3% (95% CI 42.6, 61.7), specificity 99.3% (95% CI 98.7, 99.7), PPV 87.9% (95% CI 77.0, 94.3), NPV 95.8% (95% CI 94.5, 96.8); 2) definition consisting of combinations of medications, endocrinology-specific referrals, and age criteria with sensitivity 79.3% (95% CI 70.3, 86.2), specificity 89.4% (95% CI 87.5, 91.1), PPV 40.7% (95% CI 34.2, 47.6), NPV 88.6% (95% CI 86.7, 90.2).

Interpretation: The first algorithm may be useful for cohort creation and disease registry development given its high PPV. The second algorithm may be more appropriate for public health surveillance and epidemiology, having better sensitivity and reasonably high specificity.

Introduction

The use of large clinical datasets coupled with high performance computing and advanced analytics offers significant potential for understanding disease distribution in the community, effective management, risk and prevention, at levels of statistical precision that smaller datasets cannot achieve. Diabetes had an Ontario prevalence of 8.8% and an annual incidence of 8.2 per 1000 in 2007 [1]. More recently the Public Health Agency of Canada reported an age adjusted, overall population prevalence of 7.8% in 2013/4 [2], while Diabetes Canada [3] report an estimated 9.3% in 2015. But these rates aggregate all forms of diabetes, and importantly do not differentiate between type 1 and type 2 diabetes. The management of diabetes in Canada, including of type 1, has shifted to a more interdisciplinary, team-based, integrated approach based on implementation of the Chronic Care Model [4]. Therefore, developing a valid case definition for type 1 diabetes using primary care data is important and timely.

Existing validated case definitions for diabetes include those developed by Clottey et al [5], Hux et al [6], Amed et al [7] and Guttmann et al [8]. A recent systematic review [9] identified 16 studies which utilized International Classification of Diseases (ICD) - coded data to derive validated case definitions for diabetes in adults using administrative data sources. None are able to differentiate between type 1 and type 2. We sought to generate and validate a case definition for type 1 diabetes using routinely collected clinical and demographic data derived from primary care EMRs. This approach eliminates the requirement for costly and delayed linkage to administrative data or other sources. The Canadian Primary Care Sentinel Surveillance Network (CPCSSN) extracts, transforms, cleans and codes the data into a standardized data model and makes the processed data available for research, surveillance, quality improvement and panel management [10]. Previously, CPCSSN developed a definition for undifferentiated diabetes that demonstrates excellent accuracy (95.3% sensitivity and 97.1% specificity) [11]. This study builds on that CPCSSN definition by identifying patients by type using artificial intelligence machine learning. This is a data-oriented method designed to find patterns or generate predictive models using large and complex datasets to identify the most accurate case definition within the data. [12]

Methods

Data source and reference standard

CPCSSN extracts de-identified clinical data from the EMR systems of around 1500 sentinel family physicians, nurse practitioners and community pediatricians, who contribute data for approximately 1.8 million patients in seven provinces and one territory across Canada. CPCSSN extracts, transforms, cleans and codes the data into a standardized data model and makes the processed data available for research, surveillance, quality improvement and panel management. These data include patient demographics, diagnoses, prescribed medications, laboratory results, physical measurement (i.e. weight, blood pressure), medical procedures, behavioural risk factors, physician billing, allergies, vaccinations, and referrals.

Data from one of CPCSSN's participating practice-based research networks, the Southern Alberta Primary Care Research Network, extracted on December 31, 2016 and derived from the period 2008-2016 inclusive, were used. A cohort of 1399 patients of all ages, believed to have diabetes, was identified using the current CPCSSN case definition. Family physicians who agreed to participate in this study were able to re-identify the CPCSSN-defined diabetes patients in their own practice EMR system and were asked first to confirm that the patient had diabetes and then to determine whether the patient had type 1, type 2, another diabetes subtype or no diabetes, based on their clinical expertise and any supporting evidence they chose to make use of. This list of physician-confirmed diabetes cases along with their diabetes subtype (type 1, 2 or other) constituted the reference standard for the analysis.

Machine Learning

This study employed “supervised machine learning” to combine the large, complex, multi-variable CPCSSN dataset with the reference standard (physician confirmed diabetes subtype) to “learn” the clinical characteristics (called ‘features’) that differentiated those with type 1 diabetes mellitus from those with other subtypes of the disease.

Feature Selection

All plausibly relevant variables within the CPCSSN data were selected and defined as binary outcomes before the machine learning processing occurred. In our study, features were selected using information from various parts of the patient chart: age, sex, physician billing, current and historical diagnoses, referrals, and prescribed medication. Diagnoses in Canadian primary care are generally coded using the International Classification of Diseases version 9 (ICD-9). Therefore, every unique ICD-9 code present in the EMR database was considered as a feature. Special consideration was also given to two instances of a code within one year, and two within two years. A similar approach was used for all coded information in the EMR database. Any diagnoses, referrals, and medications that were recorded as free text were included using a simple bag-of-words approach. This creates a binary indicator for each unique word that appears in any free text field within the CPCSSN database. Similarly, non-case sensitive wildcard searches for keywords and phrases related to diabetes status were added. These were the following phrases: “type 1”, “insulin-dependent”, “t1dm”, "type 2","type I ","type II", "insulin dependent", "insulin dep","tIdm","tIIdm", "non-insulin dependent" and "type 1 insulin dependent". The following combinations were also included 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 (WHO, 2018))[13]. Each unique ATC code appearing in the medication table was included as a feature, including truncated codes to identify families of drugs rather than individual ones. The frequency of ATC codes was also assessed, particularly whether two instances of the same code were used within one year, and two within two years.

Laboratory values were also included. The diabetes-related tests available in the CPCSSN data are hemoglobin A1c and fasting plasma glucose measures. Binary indicators were created 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%, etc.).

Age was included as a variable because 53% of type 1 cases are diagnosed before the age of 30 years, with a peak around 14 years. Recent evidence suggests that the remaining 47% of type 1 cases are diagnosed between ages 30 and 60y [14], [15]. We included each age-year between 18 and 50 inclusive as candidate features.

Algorithms

Machine learning feature selection algorithms used in this analysis include the C5.0 decision tree [16], the Classification and Regression Tree (CaRT) decision tree [17,18], the Chi-Squared Automated Interaction Detection (CHAID) decision tree [19, 20], and Least Absolute Shrinkage and Selection Operator (LASSO) logistic regression [21,22]. 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 [12].

Statistical Analysis

Each of the machine learning algorithms have tuning parameters that can be manipulated to control the complexity and size of the final case definition. Tuning parameters were selected using a bootstrap method. A random sample with replacement of the study population was taken for a range of possible tuning parameter values. These were repeated 30 times per tuning parameter value, until it could be determined which tuning parameter values optimized the accuracy metrics. Specifically, the misclassification rate, the F1 Score, G-Mean, and the unweighted mean of sensitivity and specificity (known as the 'naïve mean') were all investigated. The F1 score is defined as

$$F1 = \frac{Sensitivity\ \times\ PPV}{Sensitivity + PPV}$$

And the G-mean is defined as

$$G\text{-}Mean = \sqrt{Sensitivity \times PPV}$$

Once the tuning parameters were selected, 10-fold cross validation was used to determine the validity estimates [23]. This was done by splitting the study population into 10 segments or “folds”. The training of the model is conducted on 9 of them, and testing is performed on the remaining fold. This was repeated 10 times, such that each fold was used once for testing. After the validity estimates were determined, the model was fitted on the entire study population to get the final case definition. All statistical analysis was conducted using R Statistical Software version 3.3.1 [24].

Results

Of the 1399 total sample of CPCSSN diabetes patients, 1309 individuals were confirmed with the disease and were included in the analysis. Ninety (6.4%) were excluded for a variety of reasons, including being (on investigation) misclassified as having type 1 by their family physician, identified as not having diabetes at all, being deceased or no longer active in a physician's panel, having gestational diabetes or a relatively rare diabetes subtype (e.g. latent autoimmune diabetes of adults (LADA), mature onset diabetes of the young (MODY), etc.). This resulted in confirmation of 1199 people with type 2 diabetes (91.6%) and 110 people with type 1 (8.4%).

Table 1 describes the demographic and clinical characteristics of the 1309 individuals with type 1 or type 2 diabetes. Type 1 patients were younger and included more females. People with type 1 diabetes had substantially more insulin prescriptions, both issued in the past year (27.3% vs 6.3%) and at any time (75.7% vs 12.9%).

The 10-fold cross validation results are presented in table 2. Generally, sensitivities were found to range from 40% to 55%, and specificities from 97-99%. Due to the low prevalence of type 1 diabetes, the natural inclination of the machine learning algorithms is to achieve high specificity at the expense of sensitivity. However, the set of algorithms minimizing the naïve mean of sensitivity and specificity show much higher sensitivities for the C5.0 and CaRT decision trees. Here the sensitivities are 81.1% and 79.3%, and the specificities are 87.9% and 89.4%. These approaches have a tuning parameter that allows a user to weight false negatives higher than false positives, thereby producing a case definition with a higher sensitivity.

Table 3 shows the final case definitions for two notable models from the 10-fold cross validation results. The first is the CaRT case definition which minimizes the misclassification rate. This is a simple case definition with high specificity (99.3%), very modest sensitivity (52.3%), but good positive and negative predictive values (87.9%, 95.8% respectively).

The second case definition in table 3 is the CaRT implementation maximizing the naïve mean. This has good sensitivity (79.3%), specificity (89.4%) and negative predictive value (97.9%) but poor positive predictive value (40.7%). This case definition includes as a feature the presence of the text “metabolism” in the referral table. This feature appears when a patient is referred to an “Endocrinology & Metabolism” specialist.

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 primary care EMR data. Although we found no single case definition that boasts high sensitivity, specificity *and* predictive values, we judge that two useful case definitions have been demonstrated here. The first (table 3) adopts the CaRT implementation minimizing misclassification. This is a simple case definition that has high positive and negative predictive values. High predictive values are ideal for cohort creation 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 adopts the CaRT approach maximizing the naive mean and has good sensitivity and specificity (79.3% and 89.4% respectively). This case definition is useful for epidemiologic and surveillance purposes, such as examining population level temporal trends of incidence and prevalence.

Clottey et al [5] developed a case definition for undifferentiated diabetes consisting of at least one of the following criteria: at least two physician billing claims within a two-year period or one hospitalization with an ICD code for diabetes. Hux et al [6] generated two definitions including one or two claims and a hospital admission. In British Columbia, Amed and colleagues [7] developed two additional definitions intended for use in children and adolescents. The first consists of one hospitalization, two physician billing claims in a single year, and combinations of insulin or oral anti-diabetic medications. The second consists of four billing codes over two years. Guttmann and colleagues [8] developed a definition for pediatric diabetes using claims data exclusively, concluding that four physician billing claims using ICD-9 250.X in a two-year period provided optimal sensitivity and specificity. Each study included in the recent systematic review by Khokhar et al [9] 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. Wyse [25] identified a 15% under-reporting of polypectomy validated against clinical records. Mujaharine et al [26] identified similar misclassification rates for hypertension. Hux et al [6] reported positive predictive values for their case definitions ranging from 0.61 to 0.80 indicating substantial misclassification of diabetes compared to chart review. Hence the ability of our study to exploit the usefulness of data other than hospital admission and physician claims in determining caseness, to create case definitions which, together, maximize sensitivity and specificity as well as positive and negative predictive values, and to present case definition validation metrics in support of the differentiation between type 1 and type 2 diabetes are significant achievements.

Limitations to our study include a fairly small number of confirmed type 1 diabetes cases (n = 110). We believe the under-recording of insulin prescription for patients confirmed as Type 1 derives from their receiving most of their diabetes-specific care from their endocrinologist or other diabetes specialist in an outpatient clinic setting. These transactions are usually not subsequently recorded in primary care EMRs as well. Future research on a larger sample would result in more stable validity results and feature selection. The validity measures should also be interpreted with caution, as our diabetes cohort was selected from patients meeting the previously validated case definition for diabetes and are conditional upon CPCSSN-processed data, criteria and the validity of that definition. Further study is required to determine the validation metrics of case definitions for type 1 diabetes mellitus in non-CPCSSN EMR data. Despite these limitations, this study has resulted for the first time in the development and validation of usable case definitions about patients with type 1 diabetes in primary care settings using routinely collected primary care EMR data which *has* undergone CPCSSN processing.

In conclusion, we have developed and validated two case definitions using machine learning that achieve different goals in distinguishing between type 1 and type 2 diabetes in CPCSSN data. One case definition is suited for screening and cohort development, with high positive and negative predictive values. The other is suited for epidemiological purposes, having a reasonable balance between sensitivity and specificity. Further validation and testing using a larger and more diverse sample are recommended.

Funding

The study was funded by Alberta Innovates.

Conflicts of Interest

We declare no conflict of interest.

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 CanadianChronic Disease Surveillance System. Can J Diabetes 41 (2017) S22–S83

3. Diabetes Canada. Available at: http://www.diabetes.ca/how-you-can-help/advocate/why-federal-leadership-is-essential/diabetes-statistics-in-canada. Accessed on: 5th February 2018.

4. Clement M, Harvey B, Rabi DM, Roscoe RS, Sherifali D. Organization of Diabetes Care. Canadian Journal of Diabetes 2013; 37 (supple 1): S20-S25.

5. Clottey C, Mo F, LeBrun B, Mickelson P, Niles J, Robbins G. The development of the National Diabetes Surveillance System (NDSS) in Canada. Chronic Diseases in Canada 2001; 22: 67-9.

6. Hux JE, Fintoft V, Ivis F, Bica A. Diabetes in Ontario: determination of prevalence and incidence using a validated adminsitrative data algorithm. Diabetes care 2002; 25: 512-516.

7. Amed S,Vanderloo SE, Metzger D, Collet JP, McCrae P, Johnson JA. Validation of diabetes case definitions using administrative claims data. Diabetes medicine 2011; 28: 424-427.

8. Guttmann A, Nakhla M, Henderson M, To T, Daneman D, Cauch-Dudek K, Wang X, Lam K, Hux J. Validation of a health administrative data algorithm for assessing the epidemiology of diabetes in Canadaian children Pediatric Diabetes 2010; 11: 122-8.

9. Khokhar B, Jette N, Metcalfe A, Cunningham CT, Quan H, Kaplan GG, Butalia S, Rabi D. Systematic review of validated case defintions for diabetes in ICD-9-coded and ICD-10-coded data in adult populations.BMJ Open 2016; 6: e009952. doi:10.1136/bmjopen-2015-009952.

10. Greiver M, Williamson T, Barber D, Birtwhistle R, Aliarzadeh B, Khan S, Morkem R, Halas G, Harris S, katz A. Prevalence and epidemiology of diabetes in Canadian primary care prctices: a report from the Canadian Primary Care Sentinel Surveillance Network. Canadian Journal of Diabetes 2014: 38: 179-85. doi: 10.1016/j.jcjd.2014.02.030.

11. T Williamson, ME Green, and R Birtwhistle. Validating the 8 CPCSSN case definitions for chronic disease surveillance in a primary care database of electronic health records. Annals of Family Medicine, 12(4):367-372, July/August 2014.

12. JR Quinlan. C4.5 Programs for Machine Learning. Morgan Kaufmann, 1992.

13. WHO Collaborating Centre for Drug Statistics Methodology. International language for drug utilization research. Available at: https://whocc.no. Accessed on: 27th February 2018.

14. Thomas NJM, Jones S, Weedon M, Hattersley A, Oram R. Classifying diabetes bt type 1 genetic risk shows autoimmune diabetes cases are evenly distributed above and below 340 years of age. Diabetologia 2016: 59 (Suppl 1): S1-S581 (Abstract 264).

15. Diaz-Valencia PA, Bougneres P, Valleron A-J. Global epidemiology of type 1 diabetes in younfd adults and adults: a systematic review. BMC Public Health 2015; 15:255. DOI: 10.1186/s12889-015-1591-y

16. Kuhn M, Weston S, Coulter N, and Culp M. C code for C5.0 by Quinlan R. C50: C5.0 Decision Trees and Rule-Based Models, 2015. R package version 0.1.0-24.

17. Therneau TM, Atkinson EJ. An introduction to recursive partitioning using the RPART routines. Mayo Foundation 2018Available at: https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf. Accessed on: 22nd March 2018..

18. Therneau T, Atkinson EJ, and Ripley B. RPART: Recursive Partitioning and Regression Trees, 2015. R Package Version 4.1-10.

19. Kass GV. An exploratory technique for investigating large quantities of categorical data. Applied Statistics, pages 119-127, 1980.

20. The FoRt Student Project Team. CHAID: CHi-squared Automated Interaction Detection, 2015. R Package Version 0.1-2.

21. Friedman J, Hastie T, and Tibshirani R. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software, 33(1):1-22, 2010.

22. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), pages 267-288, 1996.

23. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th International Joint Conference on Atificial Intelligence, volume 2, pp 1137-1145. Morgen Kauman, Stanford, CA, 1995.

24. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2013.

25. Wyse JM, Joseph L, Barkun AN, Sewitch MJ. Accuracy of administrative claims data for polypectomy. CMAJ 2011; DOI:10.1503/cmaj.100897.

26. Muhajarine N, Mustard C, Roos LL, Young K, Gelskey DE. Comparison of survey and physicians claims data for detecting hypertension. J Clin Epidemiol 1997; 50: 711-718.

Table 1: Demographic and relevant clinical features comparing type 1 and type 2 patients from this sample. Confidence intervals for proportions are exact.

Feature	**Type 2 Diabetes (n=1,199)**	**Type 1 Diabetes (n=110)**	**Total (n=1309)**
Sex [% male (95% CI)]	53.5 (50.6, 56.3)	47.3 (37.7, 57)	52.9 (50.2, 55.7)
Age [mean (95% CI)]	64.6 (63.9, 65.3)	46 (42.8, 49.2)	63 (62.3, 63.8)
# Encounters in past year [Median (Q1, Q3)]	5 (2, 8)	4 (1, 6)	5 (2, 8)
# HbA1c tests in past year [Median (Q1, Q3)]	1 (1, 1)	1 (0, 1)	1 (1, 1)
Insulin Prescription in past year (A10AB) [% (95% CI)]	6.3 (5, 7.8)	27.3 (19.2, 36.6)	8 (6.6, 9.6)
Insulin at any time	12.9 (11.1, 14.9)	75.7 (66.6, 83.3)	18.2 (16.1, 20.4)
Anti-Hyperglycemic Drugs (excluding Insulin) in past year (A10B) [% (95% CI)]	44.5 (41.6, 47.3)	12.7 (7.1, 20.4)	41.8 (39.1, 44.5)
Anti-Hyperglycemic Drugs (excluding Insulin) at any time	71.2 (68.6, 73.8)	26.1 (18.2, 35.3)	67.4 (64.8, 70)
Occurrence of “type 1” in any text field	0.7 (0.3, 1.3)	40.5 (31.3, 50.3)	4 (3, 5.2)

Table 2: 10-fold cross validation results for each of the four machine learning algorithms, minimizing/maximizing different metrics. The misclassification rate is minimized, while the G-Mean, F1 Score, and Naïve mean were maximized.

Metric	Method	Sensitivity	Specificity	PPV	NPV	Accuracy
Misclassification Rate	C5.0	49.6 (40.0, 59.1)	99.3 (98.5, 99.6)	85.9 (74.5, 93.0)	95.5 (94.2, 96.6)	95.1 (93.7, 96.2)
	CaRT	52.3 (42.6, 61.7)	99.3 (98.7, 99.7)	87.9 (77.0, 94.3)	95.8 (94.5, 96.8)	95.4 (94.1, 96.4)
	CHAID	51.4 (41.7, 60.9)	99.3 (98.7, 99.7)	87.7 (76.6, 94.2)	95.7 (94.4, 96.7)	95.3 (94.0, 96.4)
	LASSO	40.5 (31.5, 50.3)	99.3 (98.7, 99.7)	84.9 (71.9, 92.8)	94.8 (93.4, 95.9)	94.4 (93.0, 95.6)
G-Mean	C5.0	46.9 (37.4, 56.5)	99.1 (98.3, 99.5)	82.5 (70.5, 90.6)	95.3 (94.0, 96.4)	94.7 (93.3, 95.8)
	CaRT	52.3 (42.6, 61.7)	99.3 (98.7, 99.7)	87.9 (77.0, 94.3)	95.8 (94.5, 96.8)	95.4 (94.1, 96.4)
	CHAID	50.5 (40.9, 60.0)	99.3 (98.7, 99.7)	87.5 (76.3, 94.1)	95.6 (94.3, 96.7)	95.2 (93.9, 96.3)
	LASSO	40.5 (31.5, 50.3)	99.3 (98.7, 99.7)	84.9 (71.9, 92.8)	94.8 (93.4, 95.9)	94.4 (93.0, 95.6)
F1-Score	C5.0	54.1 (44.4, 63.5)	97.4 (96.3, 98.2)	65.9 (55.2, 75.3)	95.9 (94.5, 96.9)	93.8 (92.3, 95.0)
	CaRT	52.3 (42.6,61.7)	99.3 (98.7, 99.7)	87.9 (77.0, 94.3)	95,8 (94.5, 96.8)	95.4 (94.1, 96.4)
	CHAID	53.2 (43.5,62.6)	99.0 (98.2, 99.5)	83.1 (71.9, 90.6)	95.8 (94.5, 96.9)	95.2 (93.8, 96.2)
	LASSO	40.5 (31.5, 50.3)	99.3 (98.7, 99.7)	84.9 (71.9, 92.8)	94.8 (93.4, 95.9)	94.4 (93.0, 95.6)
Naïve Mean Sens + Spec	C5.0	81.1 (72.3, 87.7)	87.9 (85.9, 89.7)	38.1 (32.0, 44.7)	98.1 (97.0, 98.8)	87.4 (85.4, 89.1)
	CaRT	79.3 (70.3, 86.2)	89.4 (87.5, 91.1)	40.7 (34.2, 47.6)	97.9 (96.8, 98.7)	88.6 (86.7, 90.2)
	CHAID	56.8 (47.0, 66.0)	98.0 (97.0, 98.7)	72.4 (61.6, 81.2)	96.1 (94.8, 97.1)	94.6 (93.2, 95.7)
	LASSO	40.5 (31.5, 50.3)	99.3 (98.7, 99.7)	84.9 (71.9, 92.8)	94.8 (93.4, 95.9)	94.4 (93.0, 95.6)

Table 3: The final case definitions for 3 notable instances of the cross-validation results. Disease status assumed to be type 2 diabetes or a diabetes subtype, unless a patient meets the T1 criteria.

Type	Case Definition
CaRT with Minimized Misclassification	Anywhere text "Type 1" **OR** Age less than 30
CaRT with Maximized Naïve Mean	Medication ATC code A10AB (Insulin) **and** not Medication ATC code A10B twice in one year (Glucose Lowering Drugs, excluding Insulin) **OR** Referral text "Metabolism" **and** Medication ATC code A10B twice in one year **and** Medication ATC code A10AB **OR** Age less than 34 **and** ATC code A10AB