Avens Publishing Group
Adv Diabetes Endocrinol
January 2023 Volume:7, Issue:1

Validated Models Using EHRs or Claims Data to Distinguish Diabetes Type among Adults

Keywords: Diabetes classification; EHR; Claims data; Model validation

Abstract

Purpose: Clinical data provides the opportunity for efficient and timely disease surveillance. We developed and validated advanced phenotyping models to classify adult patients with diabetes to type 1, type 2, or other/indeterminate using structured fields from EHR data. To simulate the use of claims data supplemented with medication information, we compared model performance before and after the removal of body mass index (BMI) and laboratory results.

Methods: We used 3 years of EHR data from a sample of 2,465 adult patients with diabetes from a health care system's clinical data warehouse. A weighted ratio of type 1 diabetes codes to all diabetes codes was created by down-weighting codes from care settings that do not treat diabetes. We developed two multinomial regression models and a machine learning conditional inference tree to classify patients to type 1, type 2, or other/indeterminate. The models were validated by calculating sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) relative to a gold standard.

Results: For all models, the weighted ratio of type 1 diabetes was the strongest predictive factor. The models had validation statistics ≥ 93% for sensitivity; ≥ 87% for specificity; ≥ 88% for PPV, and ≥ 93% for NPV. After removal of BMI and laboratory data from the regression model the largest decline in performance from the full model was in type 2 diabetes specificity (90.8% to 89.2%).

Conclusion: Prediction models and machine learning conditional inference trees using either structured fields from EHR data or claims data supplemented with medication data can be used to accurately distinguish diabetes type among adults. The inclusion of BMI and laboratory results improves model specificity for type 2 diabetes.

Open Access Research Article

Advances in Diabetes & Endocrinology

Campione JR[1*], Nooney JG[2], Kirkman MS[2], Pfaff E[3], Mardon R[1], Benoit SR[4], McKeever-Bullard K[4], Yang DH[1], Rivero G[1], Rolka D[4] and Saydah S[4]

[1]*Westat, Rockville, MD, USA*

[2]*Division of Endocrinology and Metabolism, Department of Medicine, University of North Carolina, Chapel Hill, NC, USA*

[3]*NC TraCS Institute, University of North Carolina, Chapel Hill, NC, USA*

[4]*Division of Diabetes Translation, Centers for Disease Control and Prevention, Atlanta, GA, USA*

***Address for Correspondence**
Campione JR, Westat, Rockville, MD, USA; Tel: 919-768-7325; E-mail: joannecampione@westat.com

Submission: 21 December, 2022
Accepted: 23 January, 2023
Published: 26 January, 2023

Introduction

The Centers for Disease Control and Prevention (CDC) estimates that 13% of adults in the United States have undiagnosed or diagnosed type 1 or type 2 diabetes mellitus (T1DM, T2DM) [1]. T1DM and T2DM are distinct conditions with unique epidemiology, treatment, and complications. As the prevalence of T2DM in adolescents and young adults continues to increase [2], there is a growing need for diabetes surveillance systems to distinguish between the types of diabetes to help in planning and budgeting for public health diabetes programs; to measure the cost and quality of care for the two types; and to trend type-specific pathophysiology, prevalence, morbidity and mortality especially for the more rare T1DM and youth onset T2DM [3-6].

To measure national diabetes prevalence among adults, CDC relies upon national surveys such as the National Health Interview Survey and the National Health and Nutrition Examination Survey (NHANES). However, it is difficult to use surveys for estimating prevalence by type because, when surveyed, individuals with diabetes may be uncertain about their diabetes type diagnosis and/or their prescribed diabetes-related medications.

Aside from surveys, another approach to diabetes type classification and surveillance is through the use of electronic health record (EHR) data. As more health systems in the U.S. have implemented EHR data ware houses and health information exchanges, EHR clinical data provides the opportunity for efficient and timely disease surveillance through modeling [7-9].

Within the youth population with diabetes, EHR-based methods to distinguish between T1DM and T2DM have been well-developed and validated [10-15]. However, among adults with diabetes, the studies on EHR-based models have either a) had access to diabetes onset information from a registry, b) not performed validation on both Type 1 and Type 2 classification, or c) validated on a small sample of cases [16-19].

One notable effort to develop an EHR-based algorithm for classifying diabetes type among adults has been the work of Klompas et al. and SUPREME DM. The SUPREME DM algorithm, includes a combination of criteria from diagnoses codes, drug use, and laboratory results for positive auto antibodies and c-peptide results to identify adults with T1DM or T2DM [18,20]. Schroeder et al. subsequently validated the SUPREME DM algorithm, however chart review for the study was only performed for T1DM patients and, therefore, the validation only reported T1DM positive predictive value (96.4%) [21].

While the field develops phenotyping models using EHR-data, in reality, there remains the interoperable challenge of combining EHR data across multiple providers' systems. For this reason, to perform these models at the regional or state level, researchers are more likely to use larger-scaled claims databases supplemented with medication data. Therefore, as part of our study, we removed from our full EHR model BMI and laboratory data (i.e., data often unavailable in claims data) to simulate a claims + medication-based ("simulated claims") model and compared the models' performance.

ISSN: 2475-5591

Material & Methods

Data source and population

We used EHR data from the University of North Carolina's (UNC) affiliated health system, called UNC Health, which consists of 12 hospitals and over 200 practices across the state of North Carolina. During the study time period, UNC Health used the Epic EHR system and stored the data in the Carolina Data Warehouse for Health (CDW-H). The CDW-H is refreshed daily and contains clinical, research, and administrative data sourced from UNC Health, covering over 2.7 million unique patients since 2004. The population for the study included adults ≥ 18 years of age who had at least one diagnosis code for non-gestational diabetes and two or more office visits at a UNC Health outpatient facility during the 18 months 4/1/2016 - 9/30/2017. This identified 100,743 recently active patients with at least one diagnosis code for diabetes.

Diabetes case identification and sample selection

Within this population, we aimed to identify a stratified random sample highly likely to have diabetes since patients who do not actually have diabetes are not useful for developing models to distinguish diabetes type. Our first step was to use diabetes case-finding criteria (see Appendix Table 1) based on diabetes diagnosis codes, diabetes-related laboratory results, and diabetes-related medications, similar to prior "straw man" criteria [18,22].

We then narrowed the population to include only patients who had at least one visit to the following clinic types that are likely to address diabetes: endocrinology, family medicine, general internal medicine, and obstetrics/gynecology. Women with a diagnosis code for abnormal glucose in pregnancy or gestational diabetes were excluded. These restrictions reduced the sample by 59%, resulting in a sample frame of 41,614 adult patients with presumed non-gestational diabetes.

We selected a stratified random sample of 2,500 adult patients with diabetes. To facilitate detection of the rarer T1DM, we oversampled probable T1DM by identifying patients with two or more T1DM codes on separate occasions OR one T1DM code on the patient's problem list AND no outpatient prescription for non-insulin hypoglycemic medications. The sample was further stratified by three age categories, sex, and race/ethnicity to allow for equal representation across these demographics. Once the sample was finalized, each patient in the sample was assigned a sample weight calculated as the inverse of the selection probability in their stratum for the purpose of validating the model on the health system's "real" patient population of adult patients with type 1 or type 2 diabetes. Thus, the sum of the sampling weights equaled 41,614.

From the CDW-H, relational data files of structured EHR data were pulled for the 2,500 patients during October 1, 2014 - September 30, 2017. Structured EHR data include patient demographics, health care service dates and settings, diagnoses codes, patient vital signs such as blood pressure and body mass index (BMI), laboratory results, and prescription medication information. Structured fields do not include physician's notes. The laboratory results file included hemoglobin A1c, lipids, c-peptide, auto antibodies, and triglycerides. Using EHR data for distinguishing diabetes can be challenging when information is not contained in structured data elements.

Gold standard classification

Currently, there can be considerable overlap in the diagnoses that physicians list in the patient records of adults with unclear diabetes type. Thus, there is no existing gold standard for diabetes type. For this study, to develop a gold standard diabetes type (T1DM, T2DM, and Other/Indeterminate type) we did chart review using REDCap electronic data capture tools on patients with any inconsistency in diagnosis codes [23]. For these patients, trained abstractors reviewed the patient's information to collect age at diagnosis, historical use of insulin and oral antidiabetic medications, and other elements not available in the EHR structured fields. We then applied two quantitative models independently to each case - a decision tree and a weighting equation. A decision tree used sequential rules to classify patients based on clinical factors and a weighting equation simultaneously considered twelve clinical factors using a scoring system in which clinical characteristics weighed towards or against Type 1 or Type 2. Both methods permitted a classification of "indeterminate." When the two methods did not agree, or when both models assigned the individual to Other/Indeterminate type (n=282), the study's endocrinologist reviewed and classified those cases. The other forty-one percent of the sample were straightforward cases that were already distinguishable – these patients had two or more of only one type-specific diabetes diagnosis code and consistent medication associated with that type (i.e., T1DMs with evidence of insulin only and T2DMs with no evidence of insulin not on insulin). The gold standard classifications were also used for the validation of new survey questions, in a separate study [24].

After chart review, we excluded thirty-five patients found to not have any diabetes or recently deceased. Among the 2,465 remaining, the sample consisted of 52% females. The race distribution was 33% Non-Hispanic white; 28% Non-Hispanic black; 23% Hispanic and 16% Non-Hispanic other. The gold standard classification was 663 T1DM, 1,738 T2DM, and 64 Other/Indeterminate types. After applying the sample weights, the gold standard prevalence was 4.8% T1DM, 94.6% T2DM, and 0.5% Other/Indeterminate type; similar to survey-based national estimates of T1DM and T2DM among adults diagnosed with diabetes [25]. Hence, a measure of internal validity, The Other/Indeterminate type includes secondary diabetes due to genetic defects of beta-cell function or insulin action, diabetes after a pancreatectomy or other surgery (i.e., post-procedural diabetes), secondary diabetes not elsewhere classified, and case types that were indeterminate.

Calculation

Development of Model Variables

Patient information found in the EHR data between October 1, 2014 - September 30, 2017 was used to develop the model variables. Prior algorithms to classify T1DM among patients with diabetes included the use of the ratio of T1DM codes to the sum of T1DM and T2DMdiagnosis codes [10,18,20]. Therefore, we created weighted diagnosis-based ratios for T1DM, T2DM, and Other/Indeterminate type. To do this, we categorized each diabetes code found into one of four subgroups: 1) High Value (HV): when the code was linked to a visit with one of UNC Health's diabetes / endocrinology clinics, primary care clinics, or when the visit type was "Return Diabetes"; 2) Problem List (PL): when the code was on the Patient Problem

ISSN: 2475-5591

List; 3) Primary Diagnosis (PD): when the code was not from a high value setting, but listed as the primary diagnosis; and 4) All Other (AO) diabetes diagnosis codes found. The AO subgroup was down-weighted to reduce the influence of diabetes diagnosis codes from care settings that do not treat diabetes because these codes may be less reliable than codes from health care settings that do treat diabetes. For each subgroup, type-specific ratios were created by dividing the type-specific count of codes (i.e., numerator) by the count of all diabetes codes. The all diabetes count excluded diagnosis codes for gestational diabetes, diabetes mellitus due to underlying condition (Version 10 of International Classification of Diseases Clinical Modification (ICD-10-CM) E08), and drug or chemical induced diabetes mellitus (ICD-10-CM E09). The final weighted ratio for each type was the weighted average of the subgroup ratios.

Using laboratory result data, we found patients' highest value for hemoglobin A1c and triglycerides and their lowest value for high-density lipoproteins (HDL). Indicator flags were created if a patient had positive auto-antibodies or a c-peptidevalue < 0.1ng/mL.

Prescribed medications were put into one of six categories: 1) Metform alone, 2) insulin, 3) Sulfonylurea, 4) other oral agents, 5) non-insulin injectables (i.e., liraglutide and exenatide) or 6) Glucagon. Insulin use is universal for patients with T1DM, with the exception of those newly diagnosed with T1DM, and oral hypoglycemic medications (alone or with insulin) are often used to treat patients with T2DM. We aggregated the counts of metformin alone, oral agent, and non-insulin injectables to create a patient's count of all oral agents. The count values of the six categories and of all oral agents were used to develop two indicator variables: "Oral Agent Use Only" and "Insulin Use Only".

Development of Regression Model

After variable development, we randomly cut the sample in half. One half was used for model development (n=1,233) and the other half for model validation. To choose the candidate variables for predicting diabetes type among adults we reviewed the correlations of each variable to each gold standard type, and considered the factors used in previously published models to distinguish diabetes type in light of our clinical knowledge [17,20]. This resulted in a list of 26 candidate variables, not including patient race, age category, or gender as these characteristics were used for stratification and weighting.

We then estimated and refined several multinomial regression models on the development sample. For each patient, multinomial model produced a probability of T1DM, T2DM, and Other/Indeterminate type. The highest of these probabilities is that model's predicted type for that case. We refined the models using Least Absolute Shrinkage and Selection Operator (LASSO) to assist in finding the subset of variables that best predicted diabetes type [26]. We prioritized the use of continuous values when possible, however, we also explored the use of cut points for continuous variables such as age, highest observed BMI, highest hemoglobin A1c results, and diagnosis ratios. We reviewed the area under the curve (i.e., c-statistic) to choose the best performing model within the development sample taking into account clarity, simplicity, and clinical plausibility, as well as statistical performance. The data preparation and the regression models were developed using SAS version 9.4 (SAS Institute Inc., Cary, NC).

Development of Inference Tree

As an alternative to regression modeling, we applied a supervised machine learning (ML) approach to develop a conditional inference tree that classified each patient into one of the three gold standard types. Supervised learning, in which the machine learning program is provided a set of input variables (e.g. the 26 study variables) and a known output variable (e.g., gold standard type), is a common approach used for disease prediction and diagnosis [27]. The conditional inference tree was built using the *ctree* function from **party** package in R version 3.6.0 [28]. As part of the process, first the ML program uses a significance test procedure in order to select key variables, and then, when necessary, determines implicit binary splits for continuous variables [29-30]. The overall criterion for the tree was optimal balanced sensitivity, specificity, PPV and NPV.

Validation

After model development, we validated both the multinomial regression model and the conditional inference tree in the validation sample (n=1,232). For each model, we calculated the weighted sensitivity, specificity, PPV, and NPV for both T1DM and T2DM relative to the gold standard. To assist with comparison of model performance, we also looked at the combined accuracy score defined as all correct predictions divided by total sample. Lastly, we tested and validated the SUPREME DM algorithm (Appendix Table 2) for classifying diabetes type [20].

Results

The frequencies or mean values of the twenty-six EHR candidate variables used in model development are reported in Table 1 by the full sample's gold standard classification and in total. Interestingly, among the patients classified with Other/Indeterminate type, 54.7% took insulin only and 15.6% had one or more conditions qualifying them for the T1DM flag.

Multinomial Regression Model Estimates

The final multinomial regression model included seven variables. Table 2 shows the regression model's odds ratios and maximum likelihood estimates (i.e., model coefficients) for predicting T1DM and Other/Indeterminate type, with T2DM as the reference outcome. Three factors significantly increased the likelihood for T1DM in comparison to T2DM: higher weighted T1DM ratio ($p < .001$), insulin use only ($p < .001$), and higher glucagon count ($p = .04$). Three variables significantly decreased the likelihood for T1DM in comparison to T2DM: older patient age ($p = .01$), oral agent use only ($p < .001$), and higher BMI ($p < .001$). For example, for each kg/m^2 unit increase in highest observed BMI, the likelihood or probability of an individual having T1DM in comparison to T2DM decreases by 0.207. For the dichotomous variables, it is easier to interpret the meaning of the odds ratio. For example, for patients that only used insulin, the probability of having T1DM rather than T2DM is 11 times that for patients that did not use insulin only. In predicting Other/Indeterminate type, the weighted T1DM ratio ($p < .001$) was the only factor that increased the likelihood for Other/Indeterminate type in comparison to T2DM. Two variables significantly decreased the likelihood for Other/Indeterminate type in comparison to T2DM: older patient age ($p = .01$) and highest observed BMI ($p = .002$).

ISSN: 2475-5591

Table 1: Characteristics of study sample based on the twenty-six EHR-based variables used in developing the models to predict diabetes type (N=2,465 adults; Data from 10/1/14 – 9/30/17).

	Frequency (%) for Dichotomous Variables or Mean for Continuous Variables			
	GS T1DM (n=663)	**GS T2DM (n=1,738)**	**GS Oth/Ind (n=64)**	**Total (N=2,465)**
Diagnoses-Related				
Total count of diabetes diagnosis codes*	30.9	20.9	32.0	23.9
Weighted Ratio of T1DM to all diabetes diagnosis codes*	83.0%	2.5%	39.1%	25.1%
Weighted ratio of T2DM to all diabetes diagnosis codes*	15.4%	96.0%	49.3%	73.1%
Ratio of Other DM to all diabetes diagnosis codes*	1.6%	1.7%	12.5%	2.0%
T1DM Flag of one or more of the six factors below	29.7%	2.2%	15.6%	9.9%
1. Insulin pump use	6.9%	0.2%	3.1%	2.1%
2. Celiac disease	1.5%	0.1%	0%	0.5%
3. Diabetic ketoacidosis without T2DM diagnosis	11.2%	0.5%	9.4%	3.2%
4. Hypoglycemia	10.6%	1.2%	1.6%	3.7%
5. C-peptide result < 0.1 (*lab result*)	7.1%	0.0%	1.6%	2.0%
6. Positive diabetes autoantibodies (*lab result*)	4.5%	0.4%	3.1%	1.6%
Polyneuropathy and under age 40 years	11.2%	7.8%	18.8%	9.0%
Retinopathy under age 40 years	1.6%	3.0%	1.0%	1.6%
Medication-related				
Glucagon prescription count	0.7	0.1	0.7	0.3
Insulin use only	90.5%	9.7%	54.7%	32.6%
Count of non-inpatient insulin prescriptions	7.0	2.0	6.2	3.4
Oral agent use only	0.0%	47.2%	3.1%	33.4%
Count of oral agent prescriptions	0.2	4.0	1.3	2.9
Sulfonylurea prescription count	0.00	1.08	0.27	0.77
Other Laboratory Results				
Patient's highest triglyceride (mg/dL)	136.2	213.5	263.1	194.0
Patient's highest Hemoglobin A1C (%)	9.2	8.7	10.0	8.9
Patient's lowest HDL (mg/dL)	55.7	43.5	48.0	46.9
Other				
Outpatient visit count	22.3	24.3	24.6	23.7
Patient's highest body mass index (kg/m^2)	28.2	35.4	30.5	33.3
Patient Age (years)	44.2	54.3	44.0	51.4
Family histories only included T1DM	3.5%	1.3%	0.0%	1.8%

EHR = Electronic Health Record; GS = Gold Standard; T1DM = type 1 diabetes; T2DM = type 2 diabetes. *The denominator for this measure excludes gestational diabetes, diabetes mellitus due to underlying condition (ICD-10-CM E08), and drug or chemical induced diabetes mellitus (ICD-10-CM E09).

Table 2: Results from multinomial regression model for 1,323 adult patients with diabetes with service dates from 10/1/14 – 9/30/17. *(Development sample with gold standard T2DM as the reference*)

Factor	Odds ratio for T1DM	T1DM Estimate	Estimate P-value	Odds ratio for Oth/Ind	Other DM Estimate	Estimate P-value
Full Model						
Intercept		1.762	0.4272		7.2339	0.0212
Patient age	0.960	-0.041	0.0129	0.962	-0.0383	0.0087
Weighted T1DM Ratio	1.149	0.139	< .0001	1.091	0.0868	< .0001
Oral agent use only	0.003	-5.927	0.0005	0.004	-3.3147	0.0126
Insulin use only	11.307	2.425	0.0007	0.330	-1.1101	0.3578
Glucagon count	2.466	0.903	0.0431	2.688	0.9887	0.0472
Highest observed BMI	0.813	-0.207	< .0001	0.732	-0.3116	0.0016
T1DM Flag*	1.659	0.506	0.5735	2.297	0.8317	0.3644
Akaike information criterion (AIC) = 2222.75						
Reduced Model without BMI and Laboratory Results						
Intercept		-6.317	< .0001		-4.564	< .0001
Patient age	0.990	-0.010	0.5841	1.006	-0.006	0.7929
Weighted T1DM Ratio	1.136	0.127	< .0001	1.079	0.076	< .0001
Oral agent use only	0.003	-5.881	< .0001	0.038	-3.276	< .0001
Insulin use only	14.818	2.696	< .0001	0.492	2.696	0 .0087
Glucagon count	2.005	0.695	0.0387	2.054	0.720	0.0436
T1DM Flag without lab results**	1.841	0.610	0.4059	3.745	1.321	0.0383
Akaike information criterion (AIC) = 2696.87						

T2DM = type 2 diabetes; T1DM = type 1 diabetes; Oth/Ind = Other/Indeterminate; BMI = body mass index
Note: Positive estimates indicate increase risk and negative estimates indicate decrease risk over T2DM (reference).Odds ratios are per one unit increase for age, weighted T1DM ratio, glucagon count, and highest BMI. The AIC statistic represents sample fit used for model selection; a lower AIC statistic is better.

ISSN: 2475-5591

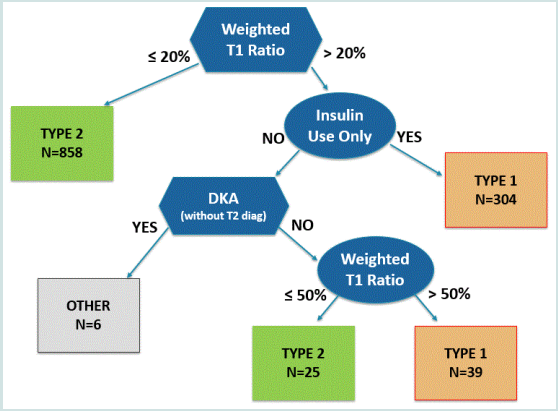

Figure 1: Conditional Inference Tree Unweighted Results using Test Sample (n=1,232 adult patients; 10/01/14- 9/30/17).

Table 2 also shows the results of the multinomial regression model after the removal of three risk factors not easily accessible in claims data: the two laboratory results included in the T1DM Flag (positive auto-antibodies and c-peptide results) and BMI. When the two laboratory results and BMI were removed, the model fit decreased only slightly in comparison to the full model.

Conditional Inference Tree Results

The machine learning approach chose only three variables for the conditional inference tree to classify diabetes type: weighted T1DM ratio, insulin use only, and DKA without T2DM codes. The tree made two binary splits of the weighted T1DM diagnosis ratio: 18.78% full sample and 42.99% among insulin users. Because these precise split values "over fit" the development sample, the split values were smoothed and rounded up to 20% and 50%, respectively. Figure 1 display how the inference tree classifies patients from the test sample, with each box at the end of a branch displaying the number of patients classified to the type of diabetes.

Table 3 shows the weighted classification results from both the regression model and the inference tree using the test sample, cross-tabulated with the gold standard classifications. The full regression model and the inference tree each had an accuracy of 98.9%. The reduced regression model without laboratory results and BMI had an accuracy of 98.8%.

Validation Results

As shown in Table 4, the regression models and the inference tree had similar T1DM sensitivity (93.5% and 93.4%) and were significantly higher than SUPREME DM's T1DM sensitivity (88.1%). The reduced regression model's T2DM specificity (89.2%) decreased by 1.6 percentage points in comparison to the full regression model (90.8%), yet was still higher than both the inference tree (87.2%) and SUPREME DM (76.1%).

The full regression model misclassified 58 cases (4.7% unweighted), the reduced regression model misclassified 65 cases (5.3% unweighted) and the inference tree misclassified 70 cases (5.7% unweighted). Forty-nine of the inaccurate cases were classified to the same type by the full regression model and the tree: 20 T1DM, 28 T2DM AND 1 Other/Indeterminate types. We investigated these 49 cases and found that 41 cases (84%) had required a review by the study's endocrinologist during the gold standard process. Twenty-six of the 49 cases (53%) had a gold standard type of other/Indeterminate, several with cystic fibrosis-related diabetes. All 49 patients took insulin. There were 6 patients that used oral medications, as found by the gold standard chart review, yet no evidence of oral medications in the EHR structured fields. Furthermore, among cases incorrectly classified as T1DM, the patients had valid T1DM factors such as a high weighted T1DM ratio and/or a prescription for glucagon.

Discussion

We developed an EHR-based regression model, a simulated claims-based regression model, and an EHR-based inference tree to distinguish diabetes type among adults. All three models yielded ≥ 89% accuracy on a test set comprising 1,232 adult patients with diabetes. The results offer enhanced models to classify diabetes type among adults using EHR or claims data for the purposes of surveillance, targeting interventions, evaluating treatment processes, and measuring type-specific patient outcomes [8,31]. The full EHR-based multinomial regression model had very strong performance, especially T1DM sensitivity, and may be ideal for analysts with access to diabetes-related medication data, BMI values, c-peptide results, and auto antibodies. Interestingly, the machine learning inference tree approach did not use BMI or laboratory results to distinguish type. Therefore, our tree model may be ideal for researchers using claims data supplemented with medication data.

The conditional inference tree displayed the ability of machine learning to successfully find optimal cut-points of the weighted T1DM ratio variable twice, resulting in high performance. For example, for a person with ≥ 20% T1DM ratio and insulin use only the tree's upper branches assigns that person to T1DM type, whereas the SUPREME DM algorithm would have not classified those patients as T1DM (unless they had positive autoantibodies or c-peptide result < 0.1 ng/mL) because the SUPREME DM algorithm uses a T1DM ratio cut point of > 50% for T1DM classification.

The weighted T1DM ratio was the strongest factor in the models. Because of the increased granularity of ICD-10-CM compared to ICD-9-CM, we suspect the ratio variable would gain precision in phenotyping models that are no longer using ICD-9-CM codes [32]. Clinicians now must choose among one of five categories when coding a diabetes-related condition or complications under ICD-10-CM: 1) E08: Diabetes mellitus (DM) due to underlying condition, 2) E09: Drug or chemical induced DM, 3) E10: Type 1 DM, 4) E11: Type 2 DM, and 5) E13: Other specified diabetes mellitus. Yet, having to make this choice can be difficult at diabetes onset, even for endocrinologists. Therefore, future research is needed to measure physician accuracy and consistency of their coding especially among specialty-type providers who do not usually diagnose diabetes.

Most often, adults with diabetes may be incorrectly identified as having T2DM. Notably, among the models validated in this study, we found that the full multinomial regression model, with use of c-peptide results, positive auto antibodies, and BMI values, was better than the reduced model and the inference tree at detecting T2DM true negatives. More specifically, this clinical information enhances the ability to distinguish when an adult should not be classified as

ISSN: 2475-5591

Table 3: Full Regression Model and Inference Tree Weighted Classification Results by Gold Standard on Test Sample (N= 1,232 adult patients; Data from 10/1/14 – 9/30/17)

	Gold Standard Diabetes Classification			
	Other/ Indeterminate	Type 1	Type 2	Total Weighted Test Sample
Regression Model Classification				
Other/Indeterminate type	13.8	5.0	12.2	31.0
Type 1	51.8	936.2	63.1	1,051.0
Type 2	42.0	59.7	19,506.0	19,608.0
Total	108	1,001	19,581	20,690
	Accuracy* = 98.9%			
Reduced Regression Model Classification without BMI and Laboratory Results				
Other/Indeterminate type	0	2.4	1.9	4.2
Type 1	51.1	935.1	69.6	1,055.8
Type 2	56.4	63.5	19,509.8	19,629.7
Total	108	1,001	19,581	20,690
	Accuracy* = 98.8%			
Inference Tree Classification				
Other/Indeterminate type	0.0	12.1	7.1	19.2
Type 1	19.8	935.0	43.5	998.3
Type 2	87.7	53.8	19,531.0	19,672.5
Total	108	1,001	19,581	20,690
	Accuracy* = 98.9%			

*Accuracy = (# of True Positive cases) divided by all patients. Example: Tree accuracy = (0 + 935 + 19531) / 20,690 = 98.9%.
Note: The regression model is a prediction model using multinomial logistic regression. The inference tree is a sequential decision true developed with machine learning.

Table 4: Performance of Regression Model, Inference Tree, and SUPREME DM among the test sample (N=1,232).

	Measure Percent and 95% Confidence Interval			
Validation Measure	**Full Regression Model[1]**	**Reduced Regression Model[1]**	**Inference Tree[1]**	**SUPREME DM[2]**
T1DM Sensitivity	93.5% (88.5%, 98.6%)	93.4% (88.4%, 98.5%)	93.4% (88.3%, 98.6%)	88.1% (84.5%, 91.8%)
T1DM Specificity	99.4% (98.9%, 99.9%)	99.4% (98.9%, 99.9%)	99.7% (99.6%, 99.8%)	99.8% (99.8%, 99.9%)
T1DM PPV	89.1% (80.3%, 97.8%)	88.6% (79.8%, 97.3%)	93.7% (91.1%, 96.2%)	95.9% (94.3%, 97.4%)
T1DM NPV	99.7% (99.4%, 99.9%)	99.7% (99.4%, 99.9%)	99.7% (99.4%, 99.9%)	99.4% (99.2%, 99.6%)
T2DM Sensitivity	96.6% (99.2%, 100%)	96.6% (99.2%, 100%)	99.7% (99.62%, 99.9%	99.9% (99.8%, 99.9%)
T2DM Specificity	90.8% (85.9%, 95.7%)	89.2% (84.2%, 94.1%)	87.2% (80.3%, 94.2%	76.1% (63.7%, 88.4%)
T2DM PPV	99.5% (99.2%, 99.8%)	99.4% (99.1%, 99.7%)	99.3% (98.8%, 99.7%)	98.6% (97.7%, 99.5%)
T2DM NPV	93.0% (86.1%, 100%)	93.3% (86.2%, 100%)	95.0% (92.7%, 97.4%)	97.3% (95.9%, 98.6%)

T1DM = type 1 diabetes; T2DM = type 2 diabetes; PPV = positive predictive value; NPV = negative predictive value; SUPREME-DM = Surveillance, prevention, and management of diabetes mellitus.
[1] The regression models and inference tree validations were performed on the weighted test sample (n=20,690).
[2]SUPREME-DM validation was performed on the weighted full sample (n=41,614). Details about SUPREME-DM are accessible at https://www.sentinelinitiative.org/sites/default/files/Methods/Mini-Sentinel_Methods_Validating-Diabetes-Mellitus_MSDD_Using-SUPREME-DM-DataLink.pdf

TD2M and, thus, considered to have T1DM or other rare type of diabetes. This is important to improve patient-centered care and the public health monitoring of diabetes trends. Thus, more research is needed to develop phenotyping models that include the processing of free text notes and other factors that will assist in classifying adult patients that have other types of diabetes.

We found that claims data with medication information is a sufficient data source for classifying diabetes type. The removal of BMI and laboratory data had little impact on the regression model's performance with the exception of a slight decline in detecting T2DM true negatives. Nevertheless, in comparison to claims data, EHR data can provide more comprehensive, timely, and longitudinal information for patients who change insurers [33]. Additionally, most health care providers' EHR databases are updated continually, and thus, automatic programs against EHR data can analyze the data on a routine basis to produce timely, granular, and detailed surveillance summaries and/or patient predicted type [34]. We have provided in the Appendix the definition of the 7 variables in the model (Table 3), the SAS Code® for the multinomial regression model (Table 4), and the code using the SAS/STAT® Proc PLM SCORE statement to apply the coefficients (Table 5) [35].

Lastly, the study has limitations. Because we developed and

ISSN: 2475-5591

tested our algorithm among patients from one health system's EHR database, it is possible that a patient received health care outside of UNC Health, and thus, have incomplete clinical data in our analysis [8,36]. Forty-one of the patients did not have the REDCap chart review performed offering the possibility of gold standard misclassification and also inflating the validation results. We randomly sampled 30 of these "clean cases" for chart review (blindly) and found 100% compliance to the classification. Although 41% of the patients did not undergo chart review, only 8 of those patients were misclassified by the full model and the tree model. This suggests that classification of the non-chart reviewed patients was accurate, remaining the same after adding in age, laboratory results, BMI, and the T1DM flag.

Conclusion

This study was the first to validate the classification of both type 1 and type 2 diabetes among adults from data fields commonly available in EHR data and claims data supplemented with medication information. Validation of the models against a gold standard classification found that a regression-based prediction model and a conditional inference tree using EHR data or claims plus medications data could be used to accurately distinguish diabetes type.

Acknowledgements

This work was supported by Grant Number DP006327-01, funded by the Centers for Disease Control and Prevention. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Centers for Disease Control and Prevention or the Department of Health and Human Services.

J.R.C. is the guarantor of this work and, as such, had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Author contributions: J.R.C. conceptualized the development of the models. J.R.C., J.N, M.S.K., E.P., D.Y., and G.R analyzed data, contributed to discussion, and reviewed/edited the manuscript. K.M.B., S.R.B, S.S, D.R., and R.M. contributed to discussion and reviewed/edited the manuscript.

References

1. Centers for Disease Control and Prevention. National Diabetes Statistics Report, 2020. Atlanta, GA: Centers for Disease Control and Prevention, U.S. Dept of Health and Human Services.
2. Nadeau KJ, Anderson BJ, Berg EG, Chiang JL, Chou H, et al. (2016) Youth-onset Type 2 Diabetes Consensus Report: Current status, challenges, and priorities. Diabetes Care 39: 1635-1642.
3. Genuth SM, Palmer JP, Nathan DM (2018) Chapter 1 in Diabetes in America, 3rd ed. Cowie CC, Casagrande SS, Menke A, Cissell MA, et al, Eds. Bethesda, MD, National Institutes of Health, NIH Pub No. 17-1468.
4. Centers for Disease Control and Prevention. Diabetes Report Card 2017 (2018) Atlanta, GA: Centers for Disease Control and Prevention, US Dept of Health and Human Services.
5. Spratt SE, Batch BC, Davis LP, Dunham AA, Easterling M, et al. (2015) Methods and initial findings from the Durham Diabetes Coalition: Integrating geospatial health technology and community interventions to reduce death and disability. J Clin Transl Endocrinol 2: 26-36.
6. Mayer-Davis EJ, Lawrence JM, Dabelea D, Divers J, Isom S, et al. (2017) Incidence trends of type 1 and type 2 diabetes among youths, 2002-2012. N Engl J Med 376: 1419-1429.
7. Horth RZ, Wagstaff S, Jeppson T, Patel V, McClellan J, et al. (2019) Use of electronic health records from a statewide health information exchange to support public health surveillance of diabetes and hypertension. BMC Public Health 19: 1106.
8. Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, et al. (2013) Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. J Am Med Inform Assoc. 20: e147-e154.
9. Klompas M, Cocoros NM, Menchaca JT, Erani D, Hafer E, et al. (2017) State and local chronic disease surveillance using electronic health record systems. Am J Public Health 9: 1406-1412.
10. Dabelea D, Pihoker C, Talton JW, D'Agostino RB Jr, Fujimoto W, et al. (2011) Etiological approach to characterization of diabetes type: the SEARCH for Diabetes in Youth Study. Diabetes Care. 34: 1628-1633.
11. Zhong V, Obeid J, Craig J, Pfaff E, Thomas J, et al. (2016). 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 Inform Assoc 23: 1060-1067.
12. Kosowan L, Wicklow B, Queenan J, Yeung R, Amed S, et al. (2019) Enhancing health surveillance: Validation of a novel electronic medical record-based pediatric type 1 and type 2 diabetes mellitus case definition. Can J Diabetes. 43: 392-398.
13. Zhong VW, Pfaff ER, Beavers DP, Thomas J, Jaacks LM, et al. (2014) Use of administrative and electronic health record data for development of automated algorithms for childhood diabetes case ascertainment and type classification: the SEARCH for Diabetes in Youth Study. Pediatr Diabetes. 15: 573-584.
14. Wells BJ, Lenoir KM, Wagenknecht LE, Mayer-Davis EJ, Lawrence JM, et al. (2020) Detection of diabetes status and type in youth using electronic health records: The SEARCH for Diabetes in Youth Study. Diabetes Care. 43: 2418-2425.
15. Lawrence JM, Black MH, Zhang JL, Slezak JM, Takhar HS, et al. (2014) Validation of pediatric diabetes case identification approaches for diagnosed cases by using information in the electronic health records of a large integrated managed health care organization. Am J Epidemiol 179: 27-38.
16. Sharma M, Petersen I, Nazareth I, Coton SJ (2016) An algorithm for identification and classification of individuals with type 1 and type 2 diabetes mellitus in a large primary care database. Clin Epidemiol 8: 373-380.
17. Lo-Ciganic W, Zgibor JC, Ruppert K, Arena VC, Stone RA (2011) Identifying Type 1 and Type 2 Diabetic Cases Using Administrative Data: A Tree-Structured Model. J Diabetes Sci Technol 5: 486-493.
18. Klompas M, Eggleston E, McVetta J, Lazarus R, Li L, et al. (2013) Automated detection and classification of type 1 versus type 2 diabetes using electronic health record data. Diabetes Care 36: 914-921.
19. Weisman A, Tu K, Young J, Kumar M, Austin PC, et al. (2020) 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 8: e001224.
20. Raebel MA, Schroeder EB, Goodrich G, Paolino AR, Donahoo WT, et al. (2016) Mini-sentinel methods validating type 1 and type 2 diabetes mellitus in the mini-sentinel distributed database using the surveillance, prevention, and management of diabetes mellitus (SUPREME-DM) datalink.
21. Schroeder EB, Donahoo WT, Goodrich GK, Raebel MA (2018) Validation of an algorithm for identifying type 1 diabetes in adults based on electronic health record data. Pharmacoepidemiol Drug Saf 27: 1053-1059.
22. Upadhyaya SG, Murphree DH Jr, Ngufor CG, Knight AM, Cronk DJ, et al. (2017) Automated diabetes case identification using electronic health record data at a tertiary care facility. Mayo Clin Proc Innov Qual Outcomes. 1: 100-110.
23. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, et al. Research electronic data capture (REDCap) - A metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 42: 377-381.

ISSN: 2475-5591

24. Nooney JG, Kirkman MS, Bullard KM, White Z, Meadows K, et al. (2020) Identifying optimal survey-based algorithms to distinguish diabetes type among adults with diabetes. J Clin Transl Endocrinol 21: 100231.

25. Bullard KM, Cowie CC, Lessem SE, Saydah SH, Menke A, et al. (2018) Prevalence of diagnosed diabetes in adults by diabetes type - United States, 2016. MMWR Morb Mortal Wkly Rep 67: 359-361.

26. Tibshirani R (1995) Regression shrinkage and selection via the lasso. J Royal Statistical Soc, Series B 58: 267-288.

27. Banda JM, Seneviratne M, Hernandez-Boussard T, Shah NH (2018) Advances in electronic phenotyping: from rule-based definitions to machine learning models. Annu Rev Biomed Data Sci 1: 53□68.

28. R Core Team (2014) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

29. Hothorn T, Hornik K, Zeileis A (2016) Unbiased Recursive Partitioning: A Conditional Inference Framework. J Computational Graphical Statistics 15: 651-674.

30. Goto T, Camargo CA Jr, Faridi MK, Yun BJ, Hasegawa K (2018) Machine learning approaches for predicting disposition of asthma and COPD exacerbations in the ED. Am J Emerg Med 36: 1650-1654.

31. Hripcsak G, Albers DJ (2013) Next-generation phenotyping of electronic health records. J Am Med Inform Assoc. 20: 117-121.

32. Chi GC, Li X, Tartof SY, Slezak JM, Koebnick C, Lawrence JM (2019) Validity of ICD-10-CM codes for determination of diabetes type for persons with youth-onset type 1 and type 2 diabetes. BMJ Open Diabetes Res Care 7: e000547.

33. Laws MB, Michaud J, Shield R, McQuade W, Wilson IB (2018) Comparison of electronic health record-based and claims-based diabetes care quality measures: Causes of discrepancies. Health Serv Res. Suppl 1: 2988-3006.

34. Birkhead GS, Klompas M, Shah NR (2015) Uses of electronic health records for public health surveillance to advance public health. Annu Rev Public Health 36: 345-359.

35. SAS Institute Inc. (2010) SAS/STAT 9.22 User's Guide, Cary, NC: SAS Institute Inc.

36. Wei WQ, Leibson CL, Ransom JE, Kho AN, Caraballo PJ, et al. (2012) Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus. J Am Med Inform Assoc 19: 219-224.