Detection of Diabetes Status and Type in Youth Using Electronic Health Records: The SEARCH for Diabetes in Youth Study

Diabetes Care 2020;43:2418–2425 | <https://doi.org/10.2337/dc20-0063>

Brian J. Wells,¹ Kristin M. Lenoir,¹
Lynne E. Wagenknecht,¹
Elizabeth J. Mayer-Davis,²
Jean M. Lawrence,³ Dana Dabelea,⁴
Catherine Pihoker,⁵ Sharon Saydah,⁶
Ramon Casanova,¹ Christine Turley,⁷
Angela D. Liese,⁸ Debra Standiford,⁹
Michael G. Kahn,¹⁰ Richard Hamman,⁴ and
Jasmin Divers¹¹

OBJECTIVE

Diabetes surveillance often requires manual medical chart reviews to confirm status and type. This project aimed to create an electronic health record (EHR)-based procedure for improving surveillance efficiency through automation of case identification.

RESEARCH DESIGN AND METHODS

Youth (<20 years old) with potential evidence of diabetes ($N = 8,682$) were identified from EHRs at three children's hospitals participating in the SEARCH for Diabetes in Youth Study. True diabetes status/type was determined by manual chart reviews. Multinomial regression was compared with an ICD-10 rule-based algorithm in the ability to correctly identify diabetes status and type. Subsequently, the investigators evaluated a scenario of combining the rule-based algorithm with targeted chart reviews where the algorithm performed poorly.

RESULTS

The sample included 5,308 true cases (89.2% type 1 diabetes). The rule-based algorithm outperformed regression for overall accuracy (0.955 vs. 0.936). Type 1 diabetes was classified well by both methods: sensitivity (Se) (>0.95), specificity (Sp) (>0.96), and positive predictive value (PPV) (>0.97). In contrast, the PPVs for type 2 diabetes were 0.642 and 0.778 for the rule-based algorithm and the multinomial regression, respectively. Combination of the rule-based method with chart reviews ($n = 695$, 7.9%) of persons predicted to have non-type 1 diabetes resulted in perfect PPV for the cases reviewed while increasing overall accuracy (0.983). The Se , Sp , and PPV for type 2 diabetes using the combined method were ≥ 0.91 .

CONCLUSIONS

An ICD-10 algorithm combined with targeted chart reviews accurately identified diabetes status/type and could be an attractive option for diabetes surveillance in youth.

The SEARCH observational study (SEARCH for Diabetes in Youth Study) has conducted population-based incidence and prevalence ascertainment of nongestational diabetes in youth since 2001 (1–6). The methods employed by SEARCH for the identification and typing of youth-onset diabetes are well established, but they rely on a time-consuming process of manually reviewing clinical records to verify diabetes presence, diabetes type, and date of diagnosis. The now widespread use

¹Division of Public Health Sciences, Department of Biostatistics and Data Science, Wake Forest School of Medicine, Winston-Salem, NC

²Departments of Nutrition and Medicine, The University of North Carolina at Chapel Hill, Chapel Hill, NC

³Department of Research and Evaluation, Kaiser Permanente Southern California, Pasadena, CA

⁴Department of Epidemiology, Colorado School of Public Health, University of Colorado Denver, Aurora, CO

⁵Department of Pediatrics, University of Washington, Seattle, WA

⁶Division of Diabetes Translation, National Center for Chronic Disease Prevention and Health Promotion, Centers for Disease Control and Prevention, Atlanta, GA

⁷Department of Pediatrics, Medical University of South Carolina, Charleston, SC

⁸Department of Epidemiology and Biostatistics, Arnold School of Public Health, University of South Carolina, Columbia, SC

⁹Cincinnati Children's Hospital Medical Center, Cincinnati, OH

¹⁰Department of Pediatrics, University of Colorado Anschutz Medical Campus, Aurora, CO

¹¹Division of Health Services Research, NYU Winthrop Research Institute, NYU Long Island School of Medicine, Mineola, NY

Corresponding author: Brian J. Wells, bjwells@wakehealth.edu

Received 9 January 2020 and accepted 20 June 2020

This article contains supplementary material online at <https://doi.org/10.2337/figshare.12558875>.

© 2020 by the American Diabetes Association. Readers may use this article as long as the work is properly cited, the use is educational and not for profit, and the work is not altered. More information is available at <https://www.diabetesjournals.org/content/license>.

of electronic health records (EHRs) in the U.S. raises the question of whether they can be deployed to increase the efficiency and sustainability of surveillance of youth-onset diabetes. EHR-based algorithms for identifying children and young adults with diabetes and the associated type of diabetes (type 1, 2, other) have long been fraught with difficulties. Clinicians may use diabetes codes before objective measurements have confirmed the diagnosis. The American Diabetes Association recognizes both hemoglobin A_{1c} (HbA_{1c}) and fasting glucose as laboratory methods to use in the diagnosis of diabetes. Although fasting blood glucose is a simple method for determining diabetes status (7), the fasting state of the patient at the time of the blood glucose test may not be evident in structured electronic data. In regards to determination of diabetes type, EHR-derived records can also be problematic. Prior versions of the ICD codes, which provide the foundation for medical billing, did not completely distinguish diabetes types. For example, ICD-9 (used until October 2015 in the U.S.) does not have an explicit class for type 2 diabetes. Instead, type 2 diabetes is aggregated into a class for “type 2 diabetes or unspecified type.” Medications are also not infallible in helping to distinguish specific diabetes types, since some of the medications may be used for treating other conditions (e.g., metformin may be used to treat polycystic ovarian syndrome), and at least one-third of youth with type 2 diabetes are treated with insulin, a medication historically associated with the treatment of type 1 diabetes (8).

Previous researchers have developed computerized algorithms for the automatic detection of diabetes presence and type among children and young adults using ICD-9 and ICD-10 codes (9–13). In these previous studies, the accuracies of the diabetes type classifiers were assessed using true diabetes cases, which would likely overestimate the performance of these tools for surveillance. If an electronic algorithm is to be used for the surveillance of all possible persons with diabetes without prior knowledge of their true status, it would be impossible to exclude persons without diabetes or persons with diabetes of a certain type (e.g., persons with drug-induced

diabetes). In addition, in two studies (12,13) adults were included; thus, the proportion of persons with type 2 diabetes was much higher than that typically seen in children with diabetes. This is important because estimates of positive predictive value (PPV) will increase as prevalence increases. It should also be noted that the previous works were evaluated in a limited number of health care systems that did not always use the same gold standard for true diabetes status and type.

For facilitation of diabetes surveillance, an ideal algorithm would be capable of identifying both diabetes status and type from information in the EHR of pediatric patients with unknown diabetes status. In addition, the algorithm would need to work well across different health systems and would not require local tuning, thus facilitating its implementation at disparate clinical sites. Furthermore, the algorithm would not require special software or expertise beyond the use of the programming language used for the extraction of EHR data. Finally, the algorithm would be validated against an adjudicated gold standard definition of diabetes and diabetes type in the era of ICD-10. The goal of this study was to use the existing SEARCH infrastructure to develop new procedures for improving the efficiency of simultaneously identifying existing cases of diabetes and diabetes type using routine structured data that are available in most EHRs.

RESEARCH DESIGN AND METHODS

SEARCH identifies persons with diabetes among health plan members at the following locations: seven counties in southern California, the entire state of Colorado, Native American reservations in Arizona and New Mexico, eight counties in Ohio, the entire state of South Carolina, and five counties in Washington state. In addition to clinical diagnoses, cases are also identified by a variety of other sources that include referrals from other health care providers, community health systems, and diabetes registries. In SEARCH, an identification of a case is considered valid when there is information sufficient to determine the person has been diagnosed with diabetes by a physician. This determination can be made by provider report, medical record review, or self-report.

The current study was limited to patients who had an encounter documented in the EHR during 2017 in one of the following children's hospitals that are part of the SEARCH case ascertainment network: Cincinnati Children's Hospital, Cincinnati, OH; Seattle Children's Hospital, Seattle, WA; and Children's Hospital Colorado, Denver, CO. The study design allowed the investigators to explore the feasibility of using EHR data exclusively to identify diabetes cases while harnessing the SEARCH infrastructure to provide the gold standard for diabetes status and type. The study was conducted only after approval by local institutional review boards with waivers of informed consent and Health Insurance Portability and Accountability Act (HIPAA) authorization. Two of these study sites use EHRs developed by Epic (Verona, WI), while the other site uses a Cerner (Kansas City, MO) EHR.

Initial Case Identification

Based on prior literature (10,12,14,15) with modifications for the change to ICD-10 coding, criteria used to identify potential cases included HbA_{1c} $\geq 6.5\%$ (≥ 48 mmol/mol), fasting plasma glucose ≥ 126 mg/dL (≥ 7.0 mmol/L), random plasma glucose ≥ 200 mg/dL (≥ 11.1 mmol/L), at least one diabetes-related ICD-10 code (E08–E13), and prescription for or administration of a diabetes-related medication. These criteria were considered our “Wide Net” and were intended to have maximum sensitivity to avoid missing any true cases. Diabetes cases were required to meet at least one of the criteria during the 2017 calendar year. Supplementary Table 1 defines the Wide Net criteria and includes the medication classes considered. Each of the three hospitals applied the Wide Net algorithm and provided a set of demographic characteristics for each qualifying patient to the local SEARCH site staff. Individuals in this study were <20 years of age on 31 December 2017, had at least one clinical visit (inpatient, outpatient, or emergency department) during 2017, and had an address located within the SEARCH site geographical area. The requirement of a visit in 2017 and the geographical catchment area requirements were necessary to mirror the current SEARCH methodology. There were a small number of cases excluded due to a specific

geographical address outside of a SEARCH-defined area even though they appeared eligible according to the Wide Net criteria, which was based on zip code. The number of ineligible cases removed from Ohio, Washington, and Colorado was 28, 23, and 16, respectively.

Medical Record Review

SEARCH staff reviewed the medical record of every individual identified by the Wide Net using personal identifiers (name, date of birth, sex, etc.) to match persons to known SEARCH cases. Many of the persons identified by the Wide Net were previously registered for the SEARCH study and thus had already undergone adjudication for diabetes status and type. Individuals unknown to the SEARCH study underwent full chart review with use of the same techniques for case ascertainment and determination of diabetes type used in SEARCH. Data sets at each site were then stripped of protected health information excepting variable dates according to month, meeting the “limited data set” definition according to HIPAA. The limited data sets were transferred to and analyzed by the data-coordinating center at Wake Forest School of Medicine with institutional review board approval.

Variable Selection

Data sets included a panel of structured data fields that were used to define the Wide Net or were of interest for the predictive modeling. Candidate variables ($n = 29$) were considered for predicting diabetes status and type based on the expertise of the research team, which included physicians and epidemiologists. The complete list of candidate variables can be found in Supplementary Table 2. Variables were included from the following data domains: diagnostic codes, laboratory measurements, vital signs, demographics, and medications. Multiple variables were created from the ICD codes such as “presence of any diabetes-related diagnostic codes.” In addition, the number of ICD-10 codes occurring in each of the type-specific diabetes classes was counted (number of type 1 codes, number of type 2 codes, number of other diabetes code types), which is similar to the methodology of previous work (9,10,14). The type-specific code counts were limited to ICD-10 (implemented in the U.S. in October 2015) due to the lack of an exclusive type 2 class in ICD-9. The counts of ICD-specific codes

were raw counts regardless of whether they came from the same encounter. Other data were included from 2009 onward when most health systems were using EHRs. Blood pressure values were not included due to the extensive number of missing values. BMI was not included due to a lack of percentile conversion for persons <2 years of age in growth tables created by the Centers for Disease Control and Prevention (16). Percentiles were necessary for this study given the age range of the population, as the normal range for BMI varies during childhood development as body proportions and compositions change. A sensitivity analysis of BMI as a candidate predictor in the subpopulation aged 2–19 years ($n = 8,051$) was performed in order to assess whether the omission of BMI as a candidate variable posed a limitation for classification of type 2 cases. Presence of obesity was included as a dichotomous proxy for BMI.

Statistical Methods

The methods were designed to determine whether multinomial regression could improve upon the previous research (mentioned above [9,10,14]) that demonstrated the utility of using diabetes ICD code counts to determine diabetes type. Multinomial logistic regression was used to build a predictive model for classifying individuals into one of four mutually exclusive classes: no diabetes (according to the SEARCH definition and eligibility criteria), type 1 diabetes, type 2 diabetes, and other diabetes type. Bayesian information criteria (BIC) was used in a forward stepwise variable selection process in which the variables were selected in a fashion that minimized model BIC at each step in the process (17). This process continued until the inclusion of any remaining variable did not result in a significant reduction in the BIC. The complete set of potential predictor variables ($p = 29$) was used to fit the full regression model. Performance metrics were generated in a fashion similar to the method of a k-fold cross validation (18). Instead of splitting the data into equal folds, data from two of the three sites included in this study were used for model training and data from the third site were used to evaluate the model and calculate performance metrics in order to gain insight into how a model generated from external

site data might perform at a different site. This process was performed three times so that each site served as a test data set once. Performance metrics were averaged across the three sites to approximate the generalization performance of the model trained on the full data set and to allow for evaluation of robustness of the algorithm to differences between sites.

Performance of a rule-based algorithm was evaluated using the counts of diabetes type-specific ICD-10 codes. The counts of diabetes type-specific codes were raw counts regardless of whether they came from the same encounter. The algorithm required at least two diabetes-related ICD-10 codes to identify presence of diabetes. The most frequently occurring diabetes type-specific code (type 1, type 2, or other) was used to classify type. Ties ($n = 32$) in the counts of diabetes type-specific codes were handled in the following fashion: ties between type 1 and type 2 were assigned to type 1, ties between type 1 and other were assigned to type 1, and ties between type 2 and other were assigned to type 2. The two-code rule was based on previous research indicating improved specificity (Sp) and PPV with two codes as opposed to a single diabetes-related code (10).

A third method consisted of assessing the impact of hypothetical targeted chart reviews on the accuracy of the ICD-10 rule-based algorithm. In this scenario, predicted classes with poor performance would undergo a hypothetical SEARCH chart review. That is, all cases predicted to not have type 1 diabetes would be reviewed by trained staff using the procedures established in SEARCH to generate the gold standard for diabetes status and diabetes type, and all misclassifications would be corrected. Finally, the number of diabetes cases identified with each method as well as the proportion of diabetes types predicted among cases was compared.

The three methods were compared in their ability to classify individuals correctly into each of the four mutually exclusive classes (no diabetes, type 1 diabetes, type 2 diabetes, and other diabetes types). Sensitivity (Se), Sp , PPV, and negative predictive value (NPV) were calculated for each of the classes separately. Next, predicted class assignment and actual class assignment were used to create a 4×4 table and calculate the

accuracy (number of correct classifications/total *N*) of the methods.

RESULTS

Figure 1 provides a flowchart that outlines the study population and the number of individuals classified according to diabetes status and type. The top of the figure displays the number of youth (ages 0–19 years) within each of the hospitals. There were 197,574, 335,342, and 259,356 unique individuals in Washington, Ohio, and Colorado hospitals, respectively. The number of potential diabetes cases identified by the Wide Net in Washington ($n = 2,260$), Ohio ($n = 3,742$), and Colorado ($n = 2,680$) totaled 8,682. According to the SEARCH gold standard, 3,374 (39%) of the Wide Net sample did not have diabetes. Among 5,308 confirmed cases, 4,732 were of type 1 diabetes (89.2%), 400 type 2 diabetes (7.5%), and 176 other diabetes type (3.3%). The left and right columns display the predicted status/type according to the rule-based algorithm and the multinomial regression, respectively. Table 1 provides descriptive statistics of the SEARCH-eligible persons captured by the Wide Net. The proportion of persons determined not to have diabetes according to the SEARCH gold standard

was higher at the Ohio site due to a number of youth with obesity who were taking metformin (which was one of the Wide Net criteria) as part of their treatment for obesity. The percentages of type 1 diabetes cases according to the SEARCH gold standard were comparable across sites. Individuals in Colorado were more likely to report their ethnicity as Hispanic. The mean age (12 years old) and percentage of females (49%) in the data were comparable across sites.

Table 2 shows that *Se*, *Sp*, PPV, and NPV for the presence of diabetes and type 1 diabetes were all >0.95 for the rule-based and regression methods. The multinomial regression model had difficulty capturing all type 2 diabetes cases (*Se* 0.573) compared with the rule-based method (*Se* 0.899). In contrast, the regression model showed better PPV for type 2 diabetes than the rule-based algorithm (0.778 and 0.642, respectively). The regression and rule-based methods worked poorly in terms of *Se* for identifying individuals with the other diabetes types. The overall accuracy for the multinomial regression and the ICD counts algorithms was 0.936 and 0.955, respectively. In total, the estimated number of cases according to each method was 5,211 (multinomial), 5,426 (rule based), and

5,290 (rule based plus chart review). The rule-based method overestimated the proportion of type 2 diabetes among cases (0.107) versus the SEARCH gold standard proportion of 0.075, while the multinomial regression model underestimated the proportion of type 2 diabetes (0.057). A categorical and continuous version of maximum BMI percentile did not substantially change performance metrics for type 2 diabetes cases compared with the model generated from the original variable set and underperformed compared with the rule-based ICD-10 method.

Supplementary Table 3 displays results by site for the multinomial regression test data sets and the rule-based method to explore differences between sites. There was some variability in the ability of the multinomial regression method to identify type 2 and “other” diabetes type cases by site. The rule-based method consistently outperformed the multinomial regression method for the κ statistic and accuracy at each site. The performance of the rule-based algorithm was also very similar across race and ethnicity categories. *Se* and *Sp* for both the presence of diabetes and type 1 diabetes were ≥ 0.94 for Whites, non-Whites, Hispanics, and non-Hispanics (data not shown).

Figure 2 shows an evaluation of the rule-based algorithm combined with chart reviews of individuals classified as patients with type 2 diabetes or other diabetes. This method would require the review of 695 charts (7.9% of Wide Net cases) and would increase the accuracy of the rule-based algorithm from 0.955 to 0.983. The PPV for type 2 diabetes and other diabetes type would be equal to 1.0. The *Se* for type 2 diabetes and other diabetes type increased from 0.899 to 0.910 and 0.496–0.734, respectively. The addition of chart reviews to the rule-based method also improved the estimated proportion of type 2 diabetes, 0.109–0.069, which was close to the true proportion of 0.075. The addition of chart reviews also improved the rule-based algorithm–predicted proportion of type 1 diabetes from 0.872 to 0.909, which was closer to the true proportion of 0.892.

Figure 1—Flowchart of the study methods. *Ties ($n = 32$) in the counts of diabetes type-specific codes were handled in the following fashion: ties between type 1 diabetes (T1) and type 2 diabetes (T2) were assigned to type 1, ties between type 1 and other were assigned to type 1, and ties between type 2 and other were assigned to type 2. DM, diabetes mellitus.

CONCLUSIONS

The results of this study demonstrate that with use of EHR-based algorithms, the presence of two or more diabetes-related

Table 1—Characteristics of persons (<20 years old) with possible diabetes identified by the Wide Net algorithm at the three SEARCH sites in 2017

	Ohio*	Washington†	Colorado‡	Total
N	3,742	2,260	2,680	8,682
Age, mean (SD)	12.3 (5.4)	12.7 (5.3)	11.4 (5.2)	12.1 (5.3)
Sex, n (%)				
Female	1,868 (49.9)	1,046 (46.3)	1,347 (50.3)	4,261 (49.1)
Male	1,874 (50.1)	1,214 (53.7)	1,333 (49.7)	4,421 (50.9)
Race, n (%)				
White	2,750 (73.5)	1,404 (62.1)	1,747 (65.2)	5,901 (68.0)
Black	677 (18.1)	177 (7.8)	186 (6.9)	1,040 (12.0)
Other or unknown	315 (8.4)	679 (30.0)	747 (27.9)	1,741 (20.1)
Ethnicity, n (%)				
Hispanic	156 (4.2)	224 (9.9)	611 (22.8)	991 (11.4)
Non-Hispanic or unknown	3,586 (95.8)	2,036 (90.1)	2,069 (77.2)	7,691 (88.6)
Diabetes status, n (%)				
No diabetes	2,141 (57.2)	578 (25.6)	655 (24.4)	3,374 (38.9)
Diabetes	1,601 (42.8)	1,682 (74.4)	2,025 (75.6)	5,308 (61.1)
Diabetes type, n	1,601	1,682	2,025	5,308
Type 1, n (%)	1,379 (86.1)	1,523 (90.5)	1,830 (90.4)	4,732 (89.1)
Type 2, n (%)	169 (10.6)	134 (8.0)	97 (4.8)	400 (7.5)
Other, n (%)	53 (3.3)	25 (1.5)	98 (4.8)	176 (3.3)

Diabetes status and diabetes type are based on the SEARCH gold standard verified by chart reviews. Diabetes cases were evaluated for validity and eligibility for the SEARCH registry. Individuals that did not meet the SEARCH eligibility criteria were excluded. The number of cases excluded in Ohio, Washington, and Colorado were 30, 24, and 16, respectively. *Cincinnati Children's Hospital. †Colorado Children's Hospital. ‡Seattle Children's Hospital.

ICD codes is adequate for identifying diabetes cases in youth. The rule-based and regression methods tested in this project worked well for identification of presence of diabetes and accurately classifying persons with type 1 diabetes. The ability to accurately identify individuals with diabetes and type 1-specific diabetes using the rule-based algorithm, which is based on counts of diabetes-specific diagnostic codes, is consistent with previous research (9–11).

The predicted proportions of type 2 and other diabetes by the rule-based and regression algorithms alone without chart reviews were mediocre. The PPVs for classifying type 2 diabetes, using the computerized algorithms without chart reviews, were poorer than the PPV of 0.89 obtained by Chi et al. (9). However, the study was restricted to individuals with type 1 or type 2 diabetes and therefore excluded individuals with other diabetes types. The PPV for type 2 diabetes using the rule-based algorithm (0.642) was slightly less than that obtained by a similar rule-based algorithm by Zhong et al. (11). At two independent hospital systems, Zhong et al. obtained PPVs for type 2 diabetes of 0.63 and 0.76 without chart review. The lower PPV for type 2 diabetes measured

in this study may be due to the fact that the calculation of PPV in the work by Zhong et al. was limited to true diabetes cases and/or due to hospital differences. The current performance metrics are likely to be a more realistic estimate for future surveillance of diabetes in youth in the U.S. Furthermore, the combination of the rule-based algorithm in tandem with a targeted chart review resulted in excellent accuracy, even for persons with type 2 diabetes, despite the relatively small number.

It is possible that other types of machine learning methods like artificial neural networks or random forests could improve classification further, but these methods were not explored, as these tools would be more difficult to implement in current EHR systems than an ICD code-based algorithm. Natural language-processing methods were also not explored for similar concerns about implementation complexity and generalizability for future surveillance efforts. It is possible that these other methods would improve accuracy; future work may consider exploring these methods for automated surveillance. It is also unknown how much BMI or blood pressure values might have improved the regression model.

The results could have a big impact on diabetes surveillance projects like SEARCH. Creating a procedure that combines counts of diabetes-specific ICD codes with targeted chart reviews could allow a semiautomated procedure for the adjudication of diabetes status and type without an excessive number of manual chart reviews. Implementing a procedure requiring chart reviews of all possible cases (Wide Net) would result in the manual review of 8,682 charts across the three sites participating in this study. In these data, the hybrid method reduced the number of manually reviewed charts by >90% (from 8,682 to 695). It should be noted that the chart reviews include all individuals predicted to have type 2 diabetes. If the number of youth-onset type 2 diabetes cases in the U.S. continues to rise, it may become easier to correctly classify type 2 diabetes cases among youth. On the other hand, if the ability to classify patients with type 2 diabetes in an automated fashion does not improve, then an increase in the number of type 2 diabetes cases would necessitate additional chart reviews.

Assuming that the accuracy obtained in this study is replicated at external sites, surveillance initiatives will have to decide whether the level of accuracy is sufficient. The acceptance of this proposed method will likely hinge on its performance among persons with type 2 diabetes, since the performance metrics for diabetes presence and type 1 diabetes were exceptional and the number of individuals with other diabetes type was very small. The current authors suggest that the *Se* (0.910), *Sp* (1.00), and PPV (1.00) attained by the hybrid method for type 2 diabetes would be adequate for population-level surveillance.

There are several limitations to this work. Although this tool can be easily applied to other environments without adaptation, this project was conducted exclusively among large tertiary referral hospitals. The comprehensiveness of the data available on specific individuals may differ when compared with other environments depending on the use of a single EHR system across locations and specialties as well as whether people seek primary care services in addition to specialty services within the same system. Consequently, further work is required in other systems to validate these results in the U.S. Similarly, external validation would

Table 2—Performance of multinomial regression and a rule-based algorithm using ICD-10 codes for determining diabetes status and type using SEARCH cohort status as gold standard

	Multinomial regression*	Rule-based algorithm
Diabetes (n = 5,308)		
Se	0.964	0.991
Sp	0.982	0.966
PPV	0.987	0.969
NPV	0.918	0.983
Type 1 diabetes (n = 4,732)		
Se	0.953	0.978
Sp	0.963	0.968
PPV	0.978	0.980
NPV	0.957	0.967
Type 2 diabetes (n = 400)		
Se	0.573	0.899
Sp	0.992	0.975
PPV	0.778	0.642
NPV	0.977	0.995
Other diabetes type, e.g., medication-induced, monogenic (n = 176)		
Se	0.381	0.496
Sp	0.981	0.996
PPV	0.512	0.698
NPV	0.986	0.988
κ statistic	0.870	0.910
Accuracy	0.936	0.955

Accuracy = number correctly classified / N. Positive (LR+) and negative (LR−) likelihood ratios may be calculated with the following formulas: $LR+ = Se / (1 - Sp)$, $LR- = (1 - Se) / Sp$. *Variables in the final multinomial regression model included the following: most common diabetes type-specific code, maximum HbA_{1c}, proportion of type 2 diabetes codes, any elevated outpatient glucose, any metformin, any antidiabetes medicine, age, proportion of type 1 diabetes codes, multiple elevations in outpatient random glucose, obesity, any diabetic ketoacidosis, ethnicity, any contraceptive medication, count of type 1 diabetes codes, proportion of other diabetes codes, and polycystic ovarian syndrome.

be recommended before attempting to apply this tool outside of the U.S. It is possible that differences in the prevalence of type 1 and type 2 diabetes in other

parts of the world could impact the performance of this tool. In particular, many parts of the world have much lower rates of type 2 diabetes among children

than the U.S (19–21). Also, many countries have instituted government-funded universal health insurance where much of the health care is provided by a single payer. These systems frequently have centralized databases that capture services provided across an entire country. Algorithms like the one presented in this article may perform differently in these areas with access to more data across different specialties and health care settings. Although ICD-10 is endorsed by the World Health Organization and has been widely adopted by countries throughout the world (22), some locales may rely primarily on the Systematized Nomenclature of Medicine—Clinical Terms (SNOMED CT) to define diagnostic codes. In these instances, diabetes-related SNOMED CT codes can be translated to ICD-10 using the Unified Medical Language System (23,24).

The current study only included diabetes status according to documentation in the EHR system. People may have moved out of the area or sought care locally in another health care system. The current study was limited to individuals with at least one encounter in 2017. The characteristics of the population and the performance of the tool may be different in other years. However, we have no basis to believe that the 2017 results would be atypical. The number of persons with types of diabetes other than type 1 or type 2 was small, and thus it will be difficult to classify these individuals using either a rule-based or regression-based method.

The Wide Net algorithm for identifying potential diabetes cases could be further refined. For example, regional practice patterns (e.g., more use of metformin at Cincinnati Children's Hospital in patients with obesity) resulted in significant increases in the Wide Net population with few additional cases of diabetes. The Wide Net differences did not appear to impact accuracy, since we observed similar performance across all sites (data not shown). Future applications of the Wide Net could consider exclusion of individuals whose only evidence of possible diabetes was a prescription for metformin. There were 1,254 people (14%) who met the Wide Net criteria solely based on a prescription for metformin; excluding these individuals, we would have only missed four cases of diabetes.

Without Chart Review		Actual			
Predicted	none	3,213	3	21	19
	other	25	82	5	4
	type 1	25	57	4,632	17
	type 2	111	34	74	360

Changes in the 4x4 table after chart reviews of patients determined to have T2 or Other diabetes according to the deterministic algorithm based on counts of diabetes-related ICD-10 Codes

After Targeted Chart Review		Actual			
Predicted	none	3,349	3	21	19
	other	0	116	0	0
	type 1	25	57	4,711	17
	type 2	0	0	0	364

		Before Chart Review	After Chart Review
Diabetes n = 5,308	Se	0.991	0.992
	Sp	0.966	0.992
	PPV	0.969	0.995
	NPV	0.983	0.983
Type 1 Diabetes (T1D) n = 4,732	Se	0.978	0.995
	Sp	0.968	0.968
	PPV	0.98	0.980
	NPV	0.967	0.992
Type 2 Diabetes (T2D) n = 400	Se	0.899	0.910
	Sp	0.975	1.000
	PPV	0.642	1.000
	NPV	0.995	0.996
Other Diabetes Type n = 176 (e.g., medication-induced, monogenic)	Se	0.496	0.734
	Sp	0.996	1.000
	PPV	0.698	1.000
	NPV	0.988	0.993
κ Statistic		0.91	0.965
Accuracy		0.955	0.983

Figure 2—Impact of targeted chart review on 695 charts (7.9%) of patients of the Wide Net predicted to have type 2 diabetes or other diabetes type.

This is one of the first studies to show that combining ICD-10 rule-based algorithms with targeted chart reviews can accurately classify individuals simultaneously according to diabetes status and type in youth. The current study has several strengths. First, the rule-based algorithm was developed and compared across multiple sites with different EHR systems and is simple to use. Notably, the rule-based algorithm does not require the application of natural language processing to unstructured text in clinical notes. Furthermore, every instance of a person meeting the Wide Net criteria was fully adjudicated using manual chart reviews. Development and validation of computerized phenotyping of diabetes using EHR data are expensive due to the extensive amount of chart reviews necessary to provide a “gold standard” outcome of diabetes status and type. The ongoing SEARCH study provided this unique infrastructure in which to conduct this type of research, providing a gold standard by which to measure performance. Finally, the sample size for individuals with type 1 and type 2 diabetes was adequate to support the conclusions.

Conclusion

Surveillance of childhood diabetes is vital for public health and research initiatives aimed at reducing the burden of diabetes on society. The use of a computerized algorithm combined with targeted chart reviews presents an attractive option for future surveillance that could be more efficient than current methods.

Acknowledgments. The authors thank the following individuals who performed the critical tasks of data extraction and chart reviews for their indispensable work: Debra Standiford and Alka Chandel, Cincinnati Children’s Hospital Medical Center, Cincinnati, OH; Anna C. Bellatore, Ryan Natividad, Angela Comer, Sadaf Samay, and Sara Deakynne Davies, University of Colorado Denver, Aurora, CO; Jennifer Phillips, Beth Loots, and Cordelia Franklin, University of Washington, Seattle, WA. The SEARCH for Diabetes in Youth Study is indebted to the many youth and their families, and their health care providers, whose participation made this study possible. This study includes data provided by the Ohio Department of Health.

The provision of data by the Ohio Department of Health should not be considered an endorsement of this study or its conclusions.

Funding. The authors acknowledge the involvement in SEARCH of the South Carolina Clinical & Translational Research Institute at the Medical University of South Carolina, National Institutes

of Health (NIH)/National Center for Advancing Translational Sciences (NCATS) grants UL1 TR000062 and UL1 TR001450; Seattle Children’s Hospital and the University of Washington, NIH/NCATS grant UL1 TR00423; University of Colorado Pediatric Clinical and Translational Research Center, NIH/NCATS grant UL1 TR000154; the Barbara Davis Center for Diabetes at the University of Colorado Denver (Diabetes Endocrinology Research Center NIH grant P30 DK57516); the University of Cincinnati, NIH/NCATS grants UL1 TR000077 and UL1 TR001425; and the Children with Medical Handicaps Program managed by the Ohio Department of Health. SEARCH 3 is funded by the Centers for Disease Control and Prevention (PA numbers 00097, DP-05-069, and DP-10-001) and supported by the National Institute of Diabetes and Digestive and Kidney Diseases, NIH. SEARCH 4 (1UC4DK108173) is funded by the National Institute of Diabetes and Digestive and Kidney Diseases, NIH, and supported by the Centers for Disease Control and Prevention. The Population Based Registry of Diabetes in Youth Study (U18DP006131, U18DP006133, U18DP006134, U18DP006136, U18DP006138, and U18DP006139) is funded by the Centers for Disease Control and Prevention and supported by the National Institute of Diabetes and Digestive and Kidney Diseases, NIH. Sites with grant numbers (SEARCH 1 through 4) are as follows: Kaiser Permanente Southern California (U18DP006133, U48/CCU919219, U01 DP000246, and U18DP002714), University of Colorado Denver (U18DP006139, U48/CCU819241-3, U01 DP000247, and U18DP000247-06A1), Cincinnati Children’s Hospital Medical Center (U18DP006134, U48/CCU519239, U01 DP000248, and U18DP002709), The University of North Carolina at Chapel Hill (U18DP006138, U48/CCU419249, U01 DP000254, and U18DP002708), Seattle Children’s Hospital (U18DP006136, U58/CCU019235-4, U01 DP000244, and U18DP002710-01), and Wake Forest University School of Medicine (U18DP006131, U18DP006131S1, U48/CCU919219, U01 DP000250, and 200-2010-35171).

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention or the National Institute of Diabetes and Digestive and Kidney Diseases.

Duality of Interest. The authors acknowledge the involvement in SEARCH of Kaiser Permanente Southern California’s Clinical Research Center (funded by Kaiser Foundation Health Plan and supported in part by the Southern California Permanente Medical Group). No other potential conflicts of interest relevant to this article were reported.

Author Contributions. B.J.W. researched data, wrote the manuscript, and contributed to the discussion. K.M.L., L.E.W., R.C., and J.D. researched data, reviewed and edited the manuscript, and contributed to the discussion. E.J.M.-D., J.M.L., D.D., C.P., S.S., C.T., A.D.L., D.S., M.G.K., and R.H. reviewed and edited the manuscript and contributed to the discussion. B.J.W. 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.

Prior Presentation. Parts of this study were presented in abstract form at the 55th Annual Meeting of the European Association for the Study of Diabetes, 16–20 September 2019, Barcelona, Spain, and at the 79th Scientific Sessions

of the American Diabetes Association, 7–11 June 2019, San Francisco, CA.

References

1. Dabelea D, Bell RA, D’Agostino RB Jr., et al.; Writing Group for the SEARCH for Diabetes in Youth Study Group. Incidence of diabetes in youth in the United States [published correction appears in JAMA 2007;298:627]. JAMA 2007; 297:2716–2724
2. Mayer-Davis EJ, Lawrence JM, Dabelea D, et al.; SEARCH for Diabetes in Youth Study. Incidence trends of type 1 and type 2 diabetes among youths, 2002–2012. N Engl J Med 2017;376:1419–1429
3. Lawrence JM, Imperatore G, Dabelea D, et al.; SEARCH for Diabetes in Youth Study Group. Trends in incidence of type 1 diabetes among non-Hispanic white youth in the U.S., 2002–2009. Diabetes 2014;63:3938–3945
4. Liese AD, D’Agostino RB Jr., Hamman RF, et al.; SEARCH for Diabetes in Youth Study Group. The burden of diabetes mellitus among US youth: prevalence estimates from the SEARCH for Diabetes in Youth Study. Pediatrics 2006;118:1510–1518
5. Pettitt DJ, Talton J, Dabelea D, et al.; SEARCH for Diabetes in Youth Study Group. Prevalence of diabetes in U.S. youth in 2009: the SEARCH for Diabetes in Youth Study. Diabetes Care 2014;37:402–408
6. Dabelea D, Mayer-Davis EJ, Saydah S, et al.; SEARCH for Diabetes in Youth Study. Prevalence of type 1 and type 2 diabetes among children and adolescents from 2001 to 2009. JAMA 2014;311:1778–1786
7. American Diabetes Association. 2. Classification and diagnosis of diabetes: *Standards of Medical Care in Diabetes—2019*. Diabetes Care 2019; 42(Suppl. 1):S13–S28
8. Pinto CA, Stafford JM, Wang T, et al. Changes in diabetes medication regimens and glycemic control in adolescents and young adults with youth-onset type 2 diabetes: the SEARCH for diabetes in youth study. Pediatr Diabetes 2018; 19:1065–1072
9. Chi GC, Li X, Tartof SY, Slezak JM, Koebernick C, Lawrence JM. 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 2019;7:e000547
10. Lawrence JM, Black MH, Zhang JL, et al. 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 2014;179:27–38
11. 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 Inform Assoc 2016;23:1060–1067
12. Klompas M, Eggleston E, McVetta J, Lazarus R, Li L, Platt R. Automated detection and classification of type 1 versus type 2 diabetes using electronic health record data. Diabetes Care 2013;36:914–921
13. Teltsch DY, Fazeli Farsani S, Swain RS, et al. Development and validation of algorithms to identify newly diagnosed type 1 and type 2 diabetes in pediatric population using electronic

medical records and claims data. *Pharmacoepidemiol Drug Saf* 2019;28:234–243

14. Zhong VW, Pfaff ER, Beavers DP, et al.; Search for Diabetes in Youth Study Group. 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* 2014;15:573–584

15. Nichols GA, Desai J, Elston Lafata J, et al.; SUPREME-DM Study Group. Construction of a multisite DataLink using electronic health records for the identification, surveillance, prevention, and management of diabetes mellitus: the SUPREME-DM project. *Prev Chronic Dis* 2012;9:E110

16. Kuczmarski RJ. CDC growth charts; United States [Internet], 2000. Hyattsville, MD, National Center for Health Statistics. Available from <https://stacks.cdc.gov/view/cdc/11267>

17. Neath AA, Cavanaugh JE. The Bayesian information criterion: background, derivation, and applications. *Wiley Interdiscip Rev Comput Stat* 2012;4:199–203

18. Hastie T, Tibshirani R, Friedman J. Model assessment and selection. In *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Hastie T, Tibshirani R, Friedman J, Eds. New York, NY, Springer, 2009, p. 219–259

19. Fazeli Farsani S, van der Aa MP, van der Vorst MMJ, Knibbe CAJ, de Boer A. Global trends in the incidence and prevalence of type 2 diabetes in children and adolescents: a systematic review and evaluation of methodological approaches. *Diabetologia* 2013;56:1471–1488

20. Delvecchio M, Mozzillo E, Salzano G, et al.; Diabetes Study Group of the Italian Society of Pediatric Endocrinology and Diabetes (ISPED). Monogenic diabetes accounts for 6.3% of cases referred to 15 Italian pediatric diabetes centers

during 2007 to 2012. *J Clin Endocrinol Metab* 2017;102:1826–1834

21. Schober E, Rami B, Grabert M, et al.; DPV-Wiss Initiative of the German Working Group for Paediatric Diabetology and. Phenotypical aspects of maturity-onset diabetes of the young (MODY diabetes) in comparison with Type 2 diabetes mellitus (T2DM) in children and adolescents: experience from a large multicentre database. *Diabet Med* 2009;26:466–473

22. World Health Organization. International Classification of Diseases, 2020. Accessed 21 April 2020. Available from <http://www.who.int/classifications/icd/en/>

23. SNOMED International. Accessed 22 April 2020. Available from <http://snomed.org>

24. National Library of Medicine. Unified Medical Language System (UMLS), 2020. Accessed 22 April 2020. Available from <https://www.nlm.nih.gov/research/umls/index.html>