HHS Public Access

Author manuscript

Diabetes Obes Metab. Author manuscript; available in PMC 2026 January 01.

Published in final edited form as:

Diabetes Obes Metab. 2025 January ; 27(1): 102–110. doi:10.1111/dom.15987.

Developing an Automated Algorithm for Identification of Children and Adolescents with Diabetes using Electronic Health Records from OneFlorida+ Clinical Research Network

Piaopiao Li, MS[1], **Eliot Spector, MS**[2], **Khalid Alkhuzam, MS**[1], **Rahul Patel, PharmD**[1], **William T Donahoo, MD**[3], **Sarah Bost, MS**[2], **Tianchen Lyu, MS**[2], **Yonghui Wu, PhD**[2], **William Hogan, MD**[2], **Mattia Prosperi, PhD**[2], **Brian E Dixon, PhD**[4], **Dana Dabelea, PhD, MD**[5], **Levon H Utidjian, MD**[6], **Tessa L Crume, PhD**[7], **Lorna Thorpe, PhD**[8], **Angela D. Liese, PhD**[9], **Desmond A Schatz, MD**[10], **Mark A Atkinson, PhD**[11], **Michael J. Haller, MD**[10], **Elizabeth A Shenkman, PhD**[2], **Yi Guo, PhD**[2], **Jiang Bian, PhD**[2], **Hui Shao, PhD, MD**[1,12,13]

[1]Department of Pharmaceutical Outcomes and Policy, College of Pharmacy, University of Florida, Gainesville, FL

[2]Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, Gainesville, FL

[3]Division of Endocrinology, Diabetes & Metabolism, College of Medicine, University of Florida, Gainesville, FL

[4]Department of Epidemiology, Indiana University (IU) Richard M. Fairbanks School of Public Health, IN

[5]Lifecourse Epidemiology of Adiposity & Diabetes Center, University of Colorado Anschutz Medical Campus, Aurora, CO

[6]Division of General Pediatrics & Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia, Philadelphia, PA

[7]Department of Epidemiology, LEAD Center, Colorado School of Public Health, University of Colorado Anschutz Medical Campus, Aurora, CO

[8]Department of Population Health, NYU Langone Health, New York, NY

[9]Department of Epidemiology and Biostatistics, Arnold School of Public Health, University of South Carolina, SC

[10]Department of Pediatrics, University of Florida College of Medicine, Gainesville, FL,

[11]Diabetes Institute, University of Florida, Gainesville, FL

Corresponding author: Hui Shao, MD, PhD, Associate Professor, Hubert Department of Global Health, Emory Rollins School of Public Health, 1518 Clifton Road, NE, CNR Room 7041, Atlanta, GA 30322, hui.shao@emory.edu.

Author Contributions and Guarantor Statement: P.L., E.S., K.A., R.P. reviewed clinical charts. P.L. analyzed data and prepared the results. P.L., H.S. drafted the manuscript. Y.G., J.B. provided critical revision of the draft. All authors contributed critically to the discussion and participated in the manuscript development.

Conflicts of Interest: The authors have disclosed no conflicts of interest.

[12]Hubert Department of Global Health, Rollin School of Public Health, Emory University, Atlanta, GA

[13]Department of Family and Preventive Medicine, School of Medicine, Emory University, Atlanta, GA

Abstract

Objective—The rapid growth of electronic health records (EHRs) nationwide presents a unique opportunity for conducting automated diabetes surveillance in the United States. However, the validity of such a surveillance system relies on the accuracy of algorithms used to identify diabetes cases, which are currently lacking. This study aimed to develop an automated computable phenotype (CP) algorithm for identifying diabetes cases in children and adolescents within the EHR.

Materials and Methods—The CP algorithm was iteratively derived based on structured data from EHRs (UF Health system 2012–2020). We randomly selected 536 presumed cases among individuals < 18 years old who has (1) HbA1c ≥ 6.5%; or (2) fasting glucose ≥ 126 mg/dL; or (3) random plasma glucose ≥ 200 mg/dL; (4) diabetes-related diagnosis code from an inpatient or outpatient encounter; or (5) prescribed, administered, or dispensed diabetes-related medication. Four reviewers independently reviewed the patient charts to determine diabetes status and type.

Results—Presumed cases without type 1 (T1D) or type 2 (T2D) diabetes diagnosis codes were categorized as non-diabetes/other types of diabetes. The rest were categorized as T1D if the most recent diagnosis was T1D, or otherwise categorized as T2D if the most recent diagnosis was T2D. Next, we applied a list of diagnoses and procedures that can determine diabetes type (e.g., steroid use suggests induced diabetes) to correct misclassifications from step 1. Among the 536 reviewed cases, 159 and 64 had T1D and T2D. The sensitivity, specificity, and positive predictive values of the CP algorithm were 94%, 98%, and 96% for T1D; 95%, 95%, and 73% for T2D.

Conclusion—We developed a highly accurate EHR-based CP for diabetes in youth based on EHR data from UF Health. Consistent with prior studies, T2D was more difficult to identify using these methods.

Background

As a common chronic disease, diabetes imposes significant health and economic challenges.[1] In recent decades, the burden of diabetes in children and adolescents has been increasing in the United States. The estimated prevalence increased from 1.48 to 2.15 per 1000 individuals between 2001 and 2017 for type 1 diabetes (T1D) and from 0.34 to 0.67 per 1000 individuals between 2001 and 2017 for type 2 diabetes (T2D).[1] To identify factors associated with the increasing rate of diabetes onset and provide timely surveillance of diabetes in children and adolescents, building a timely and accurate surveillance system is not only necessary but critical.

Currently, diabetes surveillance relies heavily on national surveys such as the National Health and Nutrition Examination Survey (NHANES) and the National Health Interview Survey (NHIS).[3] However, the NHIS does not collect any health information from children

and adolescents, and the sample size of children and adolescents is limited in NHANES.[3] To overcome the limitations of national surveys on the surveillance of diabetes in children and adolescents, the Centers for Disease and Prevention (CDC) and the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) funded the SEARCH for Diabetes in Youth Study (SEARCH)[5]. SEARCH provided critical information on the incidence and prevalence of diabetes in children and adolescents.[6–9] However, it only included patients from a small number of states, limiting its generalizability to certain high populous states, such as Florida, which was not covered in the SEARCH study[10] Besides, cases were ascertained primarily through networks of pediatric endocrinologists, where the remainder of the cases were identified through local pediatric diabetes databases and electronic health records (EHRs).[11] Such a traditional method is logistically complex and requires intensive human labor, which is both costly and time-consuming, leading to delays in important surveillance metrics.

The rich real-world data, including EHRs and claims data, such as those in the OneFlorida+ Clinical Research Network (CRN)[12] offer an unique opportunity to developing a diabetes surveillance system. The OneFlorida+ CRN is one of the eight CRNs funded by PCORI that contributes to the National Patient-Centered Clinical Research Network (PCORnet).[13] The OneFlorida+ CRN has a large statewide repository of EHRs from its clinical partners linked with a number of other data sources (e.g., Medicaid claims) that contains patients' demographics, diagnosis, procedures, laboratory test results, and more, which are necessary for developing an accurate and timely statewide diabetes surveillance system.

Developing an accurate automated algorithm to identify diabetes cases from EHRs is the critical step in developing a diabetes surveillance system. Many efforts have been made in identifying people with T1D and T2D through EHRs in recent years, leading to the development of several EHR-based diabetes CP algorithms.[14–21] For example, Sharma and colleagues developed a two-step algorithm to identify T1D or T2D based on diagnostic records, treatment, and clinical test results.[14] Daniel et al. used electronic health records in primary care in the UK to develop and validate a machine-learning algorithm to predict T1D children.[22] Lo-ciganic et al. developed a recursive partitioning and regression tree model to identify T1D and T2D using administrative data.[23] However, previous studies found that using only diagnostic codes (i.e., ICD-9/10-CM) to identify cases in EHRs has poor sensitivity and specificity.[24] Besides, it's difficult to ascertain diabetes and subtypes when the diagnoses recorded in the EHR system for the same patient across different encounters were conflicting. Instead, algorithms combining diagnoses, medication, and lab results from EHRs were reported to perform better with higher sensitivity and specificity.[25,26] Most current algorithms used a combination of diabetes diagnosis codes, blood glucose level, insulin use, cumulative days of non-insulin prescriptions, and age at first diagnosis to classify the diabetes sub-type.[23,27–29] Although the reported sensitivity and positive predictive value (PPV) for T1D are generally acceptable, most studies failed to correctly identify T2D cases and often required additional manual review to ascertain the diabetes type.[23,27–29] In addition, there were data quality issues in the data sources used to develop these algorithms. For example, patient records may be incomplete due to patients seeking care from multiple healthcare systems with unconnected EHR systems. As one of the largest statewide clinical data repositories, OneFlorida+ CRN includes 13 unique

healthcare organizations providing care for about half of all Floridians (~17.2 million), which could provide more comprehensive clinical profiles for patients and reduce the data quality issues in the data source. Morris and colleagues developed an automated algorithm to identify diabetes in children and adolescents using data from the UF Health system, one of the data partners of OneFlorida+ CRN.[30] The algorithm performed well in identifying T1D. However, the PPV for identifying T2D was low. This algorithm only used diabetes diagnoses, blood glucose level, and glucose-lowering medication information. There was still abundant clinical information in EHR, such as comorbidity, steroid use, and lab results of T1D-related autoantibody, which can be leveraged to improve the overall performance of the automated algorithm for diabetes.

This study aimed to develop a new automatic algorithm for identifying diabetes cases in children and adolescents by integrating diagnoses, patient characteristics, and sets of clinical features in EHRs from the UF Health system, one of the main data contributors to the OneFlorida+ CRN.

Method

Data sources

OneFlorida+ CRN includes 13 unique healthcare systems providing care for about half of all Floridians (~17.2 million).[31] We used EHRs from the UF Health System Integrated Data Repository (IDR), one of the main data contributors to the OneFlorida+ CRN, because we have access to patients' complete clinical charts to validate diabetes cases. The UF IDR is a large-scale medical network that collects data from UF Health System, including both clinical and research enterprises. UF Health System is a medical network associated with the University of Florida. The IDR serves as a secure, clinical data warehouse that aggregates data from the university's clinical and administrative information systems, including the electronic health record system. UF IDR contains demographics, clinical encounter data, diagnoses, procedures, lab results, medications, comorbidity information, etc. As of January 2020, the IDR contains records of 1.2 million patients with over 1 billion observation facts. EHR data is composed of two main parts: structured data and unstructured data. Structured data includes demographic information, laboratory results, patient diagnosis lists, medication lists, procedure lists, etc. The unstructured data mainly includes physicians' clinical notes.

Identification of presumptive cases

The study population included individuals under 18 years of age, who had at least one encounter in OneFlorida+ CRN between January 1, 2012, and December 31, 2020. Age was determined using the age at the end of the index year (i.e., the year first met any of the five presumptive criteria, details see below).

We applied the following rules to identify a group of individuals with presumed diabetes (i.e., presumptive cases): (1) at least one HbA1c ≥ 6.5%; or (2) at least one fasting glucose ≥ 126 mg/dL; or (3) at least one random plasma glucose ≥ 200 mg/dL; or (4) at least one diagnosis code for diabetes from an inpatient or outpatient encounter; or (5) at least one

prescribed, administered, or dispensed diabetes-related medication, including metformin, sulfonylurea, glucagon-like peptide-1 receptor agonists, thiazolidinediones, SGLT2 inhibitor insulin, and other hypoglycemic agents. The goal of using such criteria is to capture as many potential diabetes cases as possible (i.e., high sensitivity) so that those who were not selected could be safely assumed to have no diabetes.

Diabetes diagnoses were identified using ICD-9-CM (International Classification of Diseases, Ninth Revision, Clinical Modification) code 250.xx and ICD-10-CM (the International Classification of Diseases, Tenth Revision, Clinical Modification after Oct 2015) code E08-E13. Laboratory records including hemoglobin A1c (HbA1c) level, fasting blood glucose level, and random blood glucose level were identified using the Logical Observation Identifiers Names and Codes (LOINC®). All the lab tests were standardized across the different sites within the UF health system. RxNorm was used to identify the prescription refill records of diabetes-related medication.

Chart review

A stratified random sampling algorithm was applied to select 536 individuals from the overall presumptive cases (Appendix sTable 1). A standardized data extraction form (see Supplement) was developed to extract information from the EHR, including the lab test results (hemoglobin A1c, random glucose level, fasting glucose level, and T1D-related autoantibody), and diagnosis information (including encounter type and provider specialty), and medication use. We also extracted text written by clinicians in unstructured clinical notes.

Four reviewers (PL, ES, RP, KA) independently reviewed the patient charts to determine diabetes status and subtypes. For each selected patient chart, at least two reviewers were involved in the review process. Discrepancies between the two reviewers were resolved through discussions with the entire study team. If disagreement persisted after the group discussion, the case was sent to an endocrinology clinician (TD) to resolve the conflicts by reviewing the patient charts. The diabetes status and type for each individual from the chart review were considered the gold standard when evaluating the performance of the developed automated algorithm.

Algorithm development

We developed two algorithms following distinctive principles. The first one mimicked the reviewers' detailed thought process in determining diabetes status. We drew a preliminary empirical decision tree to summarize how to determine diabetes status when we reviewed the first 100 charts. We used 2022 Standards of Care in Diabetes guideline as a starting point and instructed our chart reviewers to specifically adjudicate diabetes cases and types during the chart review.[32] We refined the decision tree iteratively and got the optimal algorithm as we reviewed the remaining 536 charts. The algorithm developed under this principle is a decision-tree-based algorithm. This tree captured the flow of thought for how the reviewers determined the diabetes type by using the available information from the EHR. The second algorithm was a rule-based algorithm. During the process of chart reviewing, we tried to explore and summarize clinical information which plays an important role in

identifying cases and subtypes, e.g., T1D antibodies such as glutamic acid decarboxylase autoantibodies, islet cell autoantibodies, insulin autoantibodies, insulinoma-associated-2 autoantibodies, or zinc transporter 8 autoantibodies as the gold standard to confirm T1D. This rule-based algorithm was developed by summarizing and simplifying the reviewer's decision-making process and utilizing direct determinative factors to determine diabetes status.

Assess the performance of the algorithms

We used the two newly developed algorithms to predict the diabetes status of the 536 reviewed individuals. Using the results from the manual chart review as the gold standard, we evaluated the performance of the two algorithms. We used sensitivity, specificity, and positive predictive value (PPV) to assess the performance of the developed algorithms, which was the recommended matrix by the Centers for Disease Control and Prevention.[33] Sensitivity measures the ability of the algorithm to identify patients with diabetes correctly. Specificity measures the ability of the algorithm to identify people without diabetes correctly. The positive predictive value measures the proportion of individuals who truly have that diabetes after being categorized as diabetes by the algorithm. The algorithm developed by the SEARCH research team was used as a reference.[34]

Data analysis was conducted using R, version 4.0.3. This study has been approved by the UF Institutional Review Board (IRB).

Results

The demographic characteristics of the reviewed cases are presented in table 1. We identified 135,764 presumptive cases <18 years of age from January 2012 to December 2020. Among the 135,764 presumptive cases, 536 cases were randomly selected for manual chart review, among whom 169 cases (31.5%) had T1D, 70 cases had T2D (13.1%), 297 (55.4%) cases were either other types of diabetes (including secondary diabetes mellitus, diabetes mellitus due to underlying condition, drug or chemical induced diabetes mellitus, other specified diabetes mellitus) or non-diabetes.

Figure 1 and Figure 2 present two separate automated algorithms to determine diabetes status. Figure 1 is a decision-tree-based algorithm, which mimics the reviewer's reasoning steps to determine the diabetes status of the selected cases based on information obtained from the EHR, such as lab tests (HbA1c, fasting glucose, T1D antibody), medication (long-term usage of insulin, metformin, steroid), diagnosis (T1D, T2D, Other), comorbidity (heart transplant, cystic fibrosis, etc.), etc. Figure 2 presents a rule-based algorithm: among individuals who met the presumptive criteria, we first used the most recent diagnosis to determine the presumptive diabetes status (e.g., if the most recent diagnosis was T1D, then the case was classified as T1D). In the second step, we used a list of diagnosis and procedure records to correct misclassifications from the first step: (1) if T1D-related autoantibody (islet cell antibodies [ICA], glutamic acid decarboxylase 65 [GAD65], insulin autoantibodies [IAA]) is positive, then the case was corrected as T1D; (2) If having any of the comorbidities (cystic fibrosis, Cushing's disease, heart transplantation, liver transplantation, lung transplantation, AIDS, and drug or chemical-induced diabetes), then

the case was re-classified as non-diabetes/other types of diabetes; (3) If having diabetes insipidus, then the case was re-classified as non-diabetes/other types of diabetes; (4) If using any of the following steroid (Cortisone Acetate, Dexamethasone, Hydrocortisone, Methylprednisolone, Prednisolone, and Prednisone), then the case was re-classified as non-diabetes/other types of diabetes.

The performance matrix of the two developed algorithms is presented in table 2. The sensitivity, specificity, and PPVs of the decision-tree-based CP algorithm were 93%, 99%, and 97% for T1D, 93%, 95%, and 71% for T2D, and 93%, 96%, and 97% for non-diabetes/other types of diabetes, respectively. The rule-based CP algorithm showed similar performance as the decision-tree-based CP algorithm. The sensitivity, specificity, and positive predictive values of the rule-based CP algorithm were 94%, 98%, and 96% for type 1 diabetes, 95%, 95%, and 73% for T2D, and 93%, 98%, and 98% for non-diabetes/other types of diabetes. The sensitivity, specificity, and positive predictive values of the SEARCH algorithm were 95%, 97%, and 94% for T1D, 96%, 92%, and 62% for T2D, and 87%, 99%, and 99% for non-diabetes/other types of diabetes when implemented in the selected sample. Both newly developed algorithms achieved comparable performance for identifying T1D cases and improved performance for T2D.

Discussion

Conducting diabetes surveillance in children and adolescents is challenging because the available data to support the surveillance is scarce.[35,36] The current diabetes surveillance systems are mainly built upon existing data sources, such as the NHANES and the NHIS, which contain limited information for children and adolescents. Using real-world large-volume data such as EHR offers a potential solution for diabetes surveillance in these populations. In the two newly developed algorithms, we integrated diagnosis, medication, and lab information in EHR to identify diabetes cases and determine diabetes subtypes. Overall, the two algorithms can both accurately identify individuals with T1D with high sensitivity, specificity, and PPV (all above 90%) using EHRs from UF Health – one clinical site in the OneFlorida+ CRN. Our study demonstrated the feasibility of using EHRs to build a diabetes surveillance system for children and adolescents.

In a previous study, Morris and colleagues found that developing an automated algorithm to identify diabetes in children and adolescents using data from the UF Health system is challenging with a low PPV of 0.52 for T2D,[30] which was also consistent with studies conducted in the other health systems.[37,38] One potential explanation for the low PPV of the T2D algorithm was the low prevalence of T2D in children and adolescent, because PPV tend to be closely associated with the disease prevalence.[39] The estimated T1D prevalence in the young population (15–19 years) was 3.23 per 1000 youths in 2009 and 3.91 per 1000 youths in 2017. However, the estimated T2D prevalence in the young population (15–19 years) was only 0.68 per 1000 youths in 2009 and 1.04 per 1000 youths in 2017.[40] The low prevalence of T2D increased the difficulty of achieving a high PPV for the algorithm even though both sensitivity and specificity are higher than 0.9. The algorithm developed by Morris and colleagues was mainly based on diabetes diagnosis, lab results, and pharmacy data. Our new algorithms, on the other hand, used additional information such as diagnoses

of comorbidities, steroid use, and lab results of T1D-related autoantibody. For example, it is common that children and adolescents who are undergoing a heart/liver transplantation show continuous abnormal glucose levels during hospitalization due to steroid use during surgery.[41] In such a scenario, the diabetes conditions shown in the patient problem list are temporary. Identifying cases using only diagnosing codes may increase the false positive rate. Our algorithms decreased the false positive rate by excluding those misclassified cases using those additional specific clinical diagnoses and procedures. Both new algorithms have improved PPV compared to the algorithm developed by Morris and colleagues for T2D identification (PPV: 0.52).[30] And the rule-based algorithm showed a slightly higher PPV than the decision-tree-based algorithm (0.73 vs 0.71).

The SEARCH algorithm was developed based on a voting mechanism: T1D was determined by the ratio of type 1 codes to the sum of type 1 and type 2 codes.[34] This algorithm achieved PPV of 0.76 and 0.63 at the Medical University of South Carolina (MUSC) and the University of North Carolina Health Care System (UNC), two SEARCH sites where the algorithm was initially developed. However, the algorithm showed a lower PPV (i.e., 0.62) in OneFlorida+CRN data, indicating the necessity to build our health system-tailored algorithm. While the PPV of the two new algorithms was still lower than 0.8 for T2D identification, the PPV of the rule-based algorithm increased by 0.11 (0.73 vs. 0.62) when compared with the performance of the SEARCH algorithm in the OneFlorida+ data repository.

The decision-tree-based algorithm relies on complex steps to categorize a potential case, which is harder to implement compared to the rule-based algorithm. Besides, the chance of data overfitting is increased with the decision-tree-based algorithm because it uses highly specific information such as “≥ one diagnosis of diabetes from the provider whose specialty is Endocrinology, Diabetes and Metabolism” or “≥ two HbA1C or glucose on separate days”. On the contrary, the rule-based algorithm is easier to implement due to less complex steps and more concise and structured rules. Considering the similar performance between the decision-tree-based and the rule-based algorithms, we believe the rule-based algorithm is superior to the decision-tree-based algorithm and should be used for building the diabetes surveillance system in OneFlorida+CRN. We will also recommend this approach be externally validated and used in other EHR systems to develop their localized diabetes surveillance system.

This study has several strengths. First, we used data from EHR to build the algorithms, which is less costly and time-consuming than a diabetes registry approach.[42] The development of the algorithms lay a solid foundation for timely and accurate surveillance of T1D and T2D prevalence for children and adolescents using EHR, which is crucial for monitoring the burden of the disease in this population. It can enable the detection and tracking of temporal trends in T1D and T2D prevalence. This is essential for understanding the changing patterns of T1D and T2D over time, identifying high-risk populations or regions. These information can assist policymakers in making informed decisions about resource allocation, screening, and prevention programs based on the prevalence estimates in the EHR surveillance system. Second, our algorithms are capable of differentiating between

T1D and T2D, providing the possibility of conducting subtype-specific surveillance and inferring subtype-specific prevention programs or policies.

Our study is also of a few limitations. First, we assumed that individuals who did not meet the five presumptive criteria were negative cases. We may miss a small number of diabetes cases. However, the five presumptive criteria were designed to be inclusive and highly sensitive, thus the chance of omitting diabetes cases will be low.[43] Second, the algorithms were developed using data from the OneFlorida+ CRN. They might not be generalizable to other healthcare systems. External validation is necessary for widespread utilization outside of OneFlorida+ CRN. Researchers should first check the validity of the algorithm in other health systems before implementing them. Third, there is potential for misclassification between latent autoimmune diabetes in adults (LADA) and T1D due to their similar autoimmune markers. LADA cases in youth are likely to be diagnosed as T1D. However, since the target population of the algorithm is children and adolescents, LADA cases are expected to be relatively rare, and the potential misclassification is unlikely to be a significant issue. Fourth, the decision-tree-based algorithm, which mimicked the reviewers' detailed thought process in determining diabetes status, could be updated to enhance performance as new clinical guidelines and diagnostic criteria become available. Fifth, if resources allow, we can review additional charts to further refine the algorithm development. Sixth, some high-risk population such as those with obesity may remain undiagnosed if they do not undergo routine blood glucose or hemoglobin testing during medical visits, which was greatly influenced by clinical training of physicians and the institution's practice guidelines. Lastly, some information, such as antibody positivity, may not be fully documented in structured tables. Natural language processing can be used to capture additional information from unstructured clinical notes, thereby improving the accuracy of the algorithm. In conclusion, the newly developed algorithms can accurately identify diabetes cases in children and adolescents, especially for T1D. This new algorithm can be used to develop an EHR-based diabetes surveillance system for children and adolescents.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding Source:

This study was funded in part by Centers for Disease Control and Prevention and Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health through the award to University of Florida (award number U18DP006512). The study is not related to the DiCAYA Network. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Centers for Disease Control and Prevention, the National Institutes of Health, or the DiCAYA Network.

Data Sharing and Data Accessibility

This research employs the OneFlorida dataset. The OneFlorida dataset contains individual-level information derived from electronic health records from providers of the OneFlorida Research Consortium. The dataset can be accessed upon request, and subject to Institutional

Review Board approval and fulfillment of institutional requirement. To request access to the OneFlorida dataset, interested parties should contact the OneFlorida coordination office (https://onefloridaconsortium.org/contact-2/), who will facilitate the process in line with ethical and institutional guidelines.

Reference

1. Shao Y, Wang Y, Bigman E, Imperatore G, Holliday C, Zhang P. Lifetime Medical Spending Attributed to Incident Type 2 Diabetes in Medicare Beneficiaries: A Longitudinal Study Using 1999–2019 National Medicare Claims. Diabetes Care. 2024;47(8):1311–1318. doi:10.2337/dc24-0466 [PubMed: 38913956]
2. Lawrence JM, Divers J, Isom S, et al. Trends in Prevalence of Type 1 and Type 2 Diabetes in Children and Adolescents in the US, 2001–2017. JAMA. 2021;326(8):717. doi:10.1001/jama.2021.11165 [PubMed: 34427600]
3. Zhong VW, Pfaff ER, Beavers DP, et al. 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(8):573–584. doi:10.1111/pedi.12152 [PubMed: 24913103]
4. Duncan GE. Prevalence of diabetes and impaired fasting glucose levels among US adolescents: National Health and Nutrition Examination Survey, 1999–2002. Arch Pediatr Adolesc Med. 2006;160(5):523–528. doi:10.1001/archpedi.160.5.523 [PubMed: 16651496]
5. SEARCH for Diabetes in Youth. https://www.searchfordiabetes.org/dspHome.cfm
6. Mayer-Davis EJ, Lawrence JM, Dabelea D, et al. Incidence trends of type 1 and type 2 diabetes among youths, 2002–2012. New England Journal of Medicine. 2017;376(15):1419–1429. [PubMed: 28402773]
7. Powell J, Isom S, Divers J, et al. Increasing burden of type 2 diabetes in Navajo youth: The SEARCH for diabetes in youth study. Pediatric Diabetes. 2019;20(7):815–820. doi:10.1111/pedi.12885 [PubMed: 31260152]
8. Snyder LL, Stafford JM, Dabelea D, et al. Socio-economic, demographic, and clinical correlates of poor glycaemic control within insulin regimens among children with Type 1 diabetes: the SEARCH for Diabetes in Youth Study. Diabetic Medicine. 2019;36(8):1028–1036. doi:10.1111/dme.13983 [PubMed: 31050009]
9. Kim G, Divers J, Fino NF, et al. Trends in prevalence of cardiovascular risk factors from 2002 to 2012 among youth early in the course of type 1 and type 2 diabetes. The SEARCH for Diabetes in Youth Study. Pediatric Diabetes. 2019;20(6):693–701. doi:10.1111/pedi.12846 [PubMed: 30903717]
10. https://www.searchfordiabetes.org/dspHome.cfm.
11. SEARCH for Diabetes in Youth. https://www.searchfordiabetes.org/docs/SEARCH_Phase_4_Protocol.pdf
12. Shenkman E, Hurt M, Hogan W, et al. OneFlorida Clinical Research Consortium: Linking a Clinical and Translational Science Institute With a Community-Based Distributive Medical Education Model. Acad Med. 2018;93(3):451–455. doi:10.1097/ACM.0000000000002029 [PubMed: 29045273]
13. Collins FS, Hudson KL, Briggs JP, Lauer MS. PCORnet: turning a dream into reality. J Am Med Inform Assoc. 2014;21(4):576–577. doi:10.1136/amiajnl-2014-002864 [PubMed: 24821744]
14. Sharma M, Petersen I, Nazareth I, Coton SJ. An algorithm for identification and classification of individuals with type 1 and type 2 diabetes mellitus in a large primary care database. Clin Epidemiol. 2016;8:373–380. doi:10.2147/CLEP.S113415 [PubMed: 27785102]
15. Lo-Ciganic W, Zgibor JC, Ruppert K, Arena VC, Stone RA. Identifying Type 1 and Type 2 Diabetic Cases Using Administrative Data: A Tree-Structured Model. J Diabetes Sci Technol. 2011;5(3):486–493. doi:10.1177/193229681100500303 [PubMed: 21722564]
16. Bobo WV, Cooper WO, Stein CM, et al. Positive predictive value of a case definition for diabetes mellitus using automated administrative health data in children and youth exposed to antipsychotic

drugs or control medications: a Tennessee Medicaid study. BMC Medical Research Methodology. 2012;12(1):128. doi:10.1186/1471-2288-12-128 [PubMed: 22920280]

17. 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(4):914–921. doi:10.2337/dc12-0964 [PubMed: 23193215]
18. Vanderloo SE, Johnson JA, Reimer K, et al. Validation of classification algorithms for childhood diabetes identified from administrative data. Pediatric Diabetes. 2012;13(3):229–234. doi:10.1111/j.1399-5448.2011.00795.x [PubMed: 21771232]
19. 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(6):1060–1067. doi:10.1093/jamia/ocv207 [PubMed: 27107449]
20. 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(1):27–38. doi:10.1093/aje/kwt230 [PubMed: 24100956]
21. Zhong VW, Pfaff ER, Beavers DP, et al. 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. Pediatric Diabetes. 2014;15(8):573–584. doi:10.1111/pedi.12152 [PubMed: 24913103]
22. Daniel R, Jones H, Gregory JW, et al. Predicting type 1 diabetes in children using electronic health records in primary care in the UK: development and validation of a machine-learning algorithm. The Lancet Digital Health. 2024;6(6):e386–e395. doi:10.1016/S2589-7500(24)00050-5 [PubMed: 38789139]
23. Lo-Ciganic W, Zgibor JC, Ruppert K, Arena VC, Stone RA. Identifying type 1 and type 2 diabetic cases using administrative data: a tree-structured model. J Diabetes Sci Technol. 2011;5(3):486–493. doi:10.1177/193229681100500303 [PubMed: 21722564]
24. Guo Y, He X, Lyu T, et al. Developing and Validating a Computable Phenotype for the Identification of Transgender and Gender Nonconforming Individuals and Subgroups. AMIA Annu Symp Proc. 2020;2020:514–523. [PubMed: 33936425]
25. Mo H, Thompson WK, Rasmussen LV, et al. Desiderata for computable representations of electronic health records-driven phenotype algorithms. Journal of the American Medical Informatics Association. 2015;22(6):1220–1230. doi:10.1093/jamia/ocv112 [PubMed: 26342218]
26. Wei WQ, Denny JC. Extracting research-quality phenotypes from electronic health records to support precision medicine. Genome Med. 2015;7(1):41. doi:10.1186/s13073-015-0166-y [PubMed: 25937834]
27. Vanderloo SE, Johnson JA, Reimer K, et al. Validation of classification algorithms for childhood diabetes identified from administrative data. Pediatr Diabetes. 2012;13(3):229–234. doi:10.1111/j.1399-5448.2011.00795.x [PubMed: 21771232]
28. 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(4):914–921. doi:10.2337/dc12-0964 [PubMed: 23193215]
29. Bobo WV, Cooper WO, Stein CM, et al. Positive predictive value of a case definition for diabetes mellitus using automated administrative health data in children and youth exposed to antipsychotic drugs or control medications: a Tennessee Medicaid study. BMC Med Res Methodol. 2012;12:128. doi:10.1186/1471-2288-12-128 [PubMed: 22920280]
30. Morris HL, Donahoo WT, Bruggeman B, et al. An Iterative Process for Identifying Pediatric Patients With Type 1 Diabetes: Retrospective Observational Study. JMIR Med Inform. 2020;8(9):e18874. doi:10.2196/18874 [PubMed: 32886067]
31. Hogan WR, Shenkman EA, Robinson T, et al. The OneFlorida Data Trust: a centralized, translational research data infrastructure of statewide scope. Journal of the American Medical Informatics Association. 2022;29(4):686–693. doi:10.1093/jamia/ocab221 [PubMed: 34664656]
32. American Diabetes Association Professional Practice Committee. 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes-2022. Diabetes Care. 2022;45(Suppl 1):S17–S38. doi:10.2337/dc22-S002 [PubMed: 34964875]

33. https://www.cdc.gov/mmwr/preview/mmwrhtml/rr5013a1.htm.
34. Zhong VW, Obeid JS, Craig JB, et al. An efficient approach for surveillance of childhood diabetes by type derived from electronic health record data: the SEARCH for Diabetes in Youth Study. Journal of the American Medical Informatics Association. 2016;23(6):1060–1067. doi:10.1093/jamia/ocv207 [PubMed: 27107449]
35. Desai J, Geiss L, Mukhtar Q, et al. Public Health Surveillance of Diabetes in the United States: Journal of Public Health Management and Practice. 2003;9(Supplement):S44–S51. doi:10.1097/00124784-200311001-00008
36. Li P, Lyu T, Alkhuzam K. The Role of Health System Penetration Rate in Estimating Prevalence of Type 1 Diabetes in Children and Adolescents Using Electronic Health Records. Journal of the American Medical Informatics Association. doi:10.1093/jamia/ocad194
37. Rhodes ET, Laffel LMB, Gonzalez TV, Ludwig DS. Accuracy of Administrative Coding for Type 2 Diabetes in Children, Adolescents, and Young Adults. Diabetes Care. 2007;30(1):141–143. doi:10.2337/dc06-1142 [PubMed: 17192348]
38. 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. American Journal of Epidemiology. 2014;179(1):27–38. doi:10.1093/aje/kwt230 [PubMed: 24100956]
39. Akobeng AK. Understanding diagnostic tests 1: sensitivity, specificity and predictive values. Acta Paediatrica. 2007;96(3):338–341. doi:10.1111/j.1651-2227.2006.00180.x [PubMed: 17407452]
40. Lawrence JM, Divers J, Isom S, et al. Trends in Prevalence of Type 1 and Type 2 Diabetes in Children and Adolescents in the US, 2001–2017. JAMA. 2021;326(8):717. doi:10.1001/jama.2021.11165 [PubMed: 34427600]
41. Kaplan NM, Palmer BF, Mora PF. Post-Transplantation Diabetes Mellitus. The American Journal of the Medical Sciences. 2005;329(2):86–94. doi:10.1097/00000441-200502000-00006 [PubMed: 15711425]
42. Eggleston EM, Klompas M. Rational Use of Electronic Health Records for Diabetes Population Management. Curr Diab Rep. 2014;14(4):479. doi:10.1007/s11892-014-0479-z [PubMed: 24615333]
43. Trevethan R. Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice. Front Public Health. 2017;5:307. doi:10.3389/fpubh.2017.00307 [PubMed: 29209603]

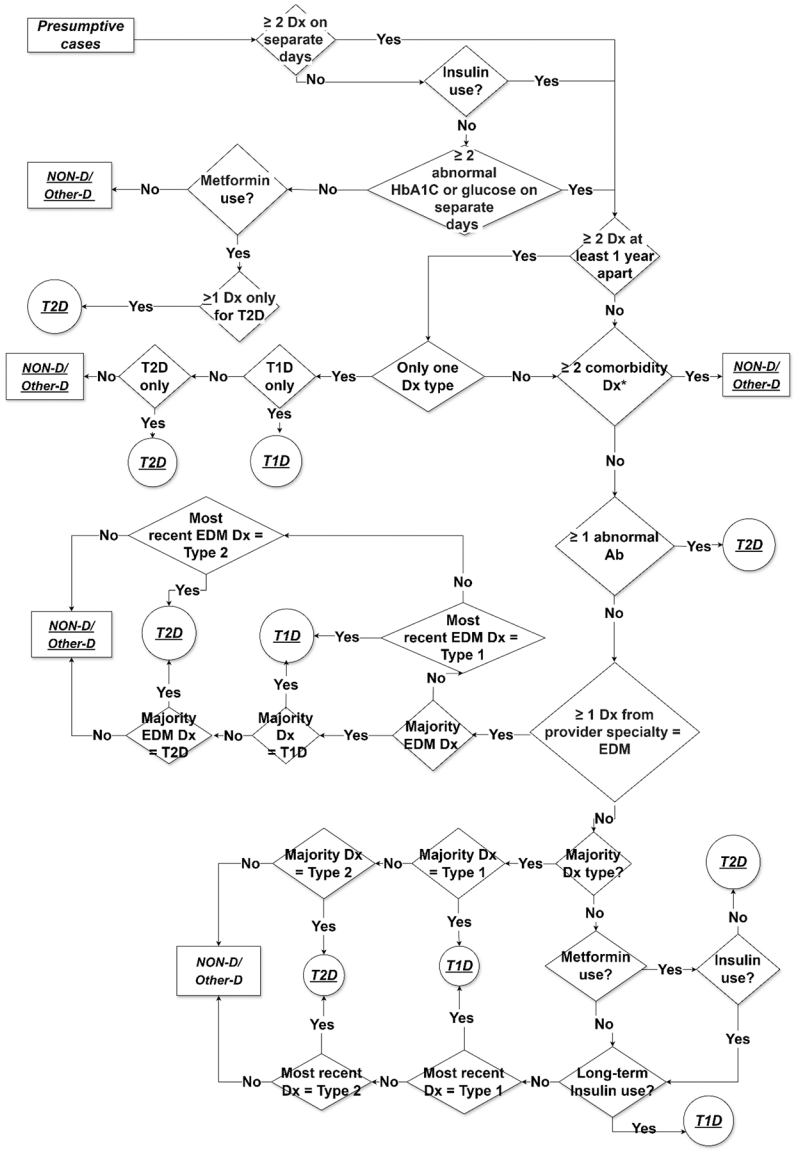

Figure 1.

Illustration of decision-tree-based algorithm

Note: Dx refers to diagnosis; T1D refers to type 1 diabetes; T2D refers to type 2 diabetes; NON-D refers to non-diabetes; Other-D refers to other type of diabetes; EDM refers that the provider's specialty is "Endocrinology, Diabetes and Metabolism"; Ab refers to autoantibodies that are markers of type 1 diabetes, including include glutamic acid decarboxylase autoantibodies (GADA or Anti-GAD), islet cell autoantibodies (ICA), insulin autoantibodies (IAA), insulinoma-associated-2 autoantibodies (IA-2A or Anti-IA-2), or zinc transporter 8 autoantibodies (ZnT8A or Anti-ZnT8). Other types of diabetes include secondary diabetes mellitus, diabetes mellitus due to underlying condition, drug or chemical induced diabetes mellitus, other specified diabetes mellitus.

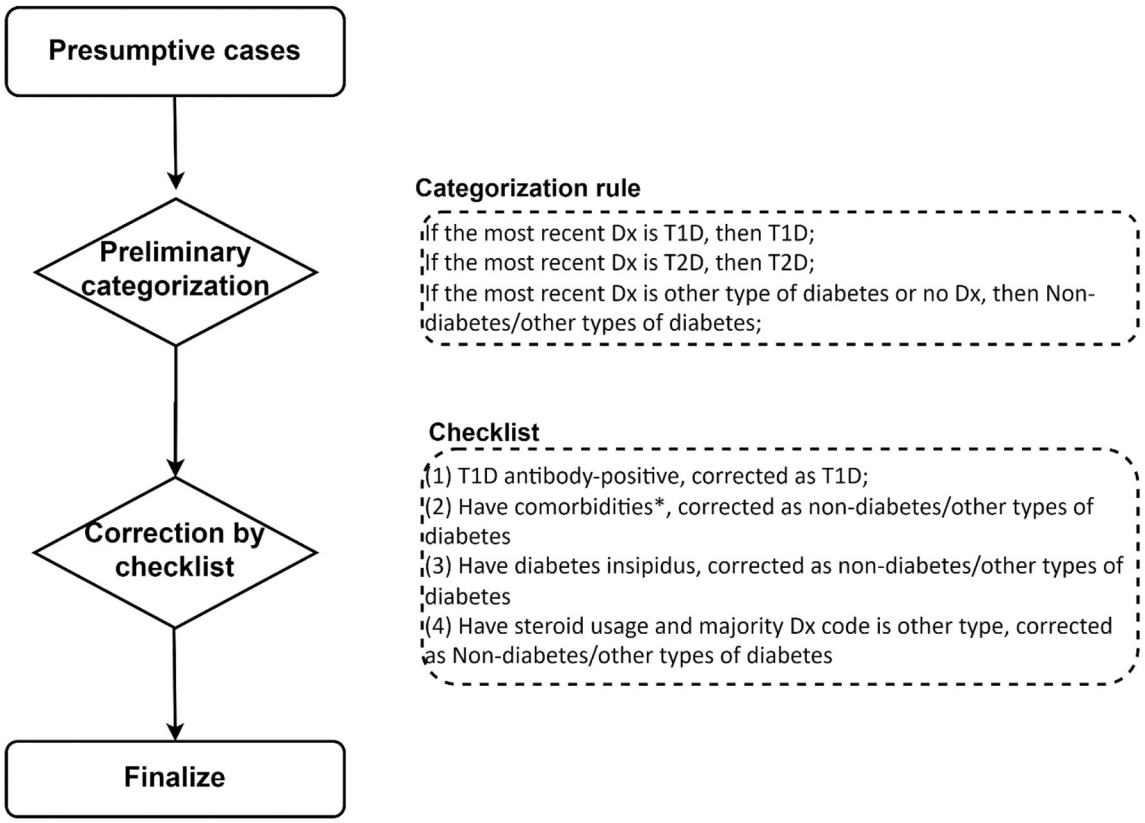

Figure 2.

Illustration of rule-based algorithm

Note: Dx refers to diagnosis; T1D refers to type 1 diabetes; T2D refers to type 2 diabetes. Other types of diabetes include secondary diabetes mellitus, diabetes mellitus due to underlying condition, drug or chemical induced diabetes mellitus, other specified diabetes mellitus.

T1D antibodies include glutamic acid decarboxylase autoantibodies (GADA or Anti-GAD), islet cell autoantibodies (ICA), insulin autoantibodies (IAA), insulinoma-associated-2 autoantibodies (IA-2A or Anti-IA-2), or zinc transporter 8 autoantibodies (ZnT8A or Anti-ZnT8).

Table 1.

Demographic and clinical characteristics of the included individuals.

	Total		T1D		T2D	
	Count	%	Count	%	Count	%
Total	536	100	169	100	70	100
Gender						
Male	267	49.8	81	47.9	47	67.1
Female	269	50.2	88	52.1	23	32.9
Age (years)						
0–4.9	180	33.6	22	13.0	2	2.9
5.0–9.9	88	16.4	39	23.1	10	14.3
10.0–14.9	140	26.1	58	34.3	33	47.1
15.0–17.9	128	23.9	50	29.6	25	35.7
Race/ethnicity						
Non-Hispanic White	252	47.0	105	62.1	19	27.1
Non-Hispanic Black	174	32.5	28	16.6	36	51.4
Hispanic	54	10.1	21	12.4	7	10.0
Other	36	6.7	9	5.3	4	5.7
Unknown	20	3.7	6	3.6	4	5.7
Meeting presumptive criteria 1	214	39.9	139	82.2	44	62.9
Meeting presumptive criteria 2	6	1.1	1	0.6	4	5.7
Meeting presumptive criteria 3	402	75.0	141	83.4	28	40.0
Meeting presumptive criteria 4	288	53.7	167	98.8	69	98.6
Meeting presumptive criteria 5	410	76.5	163	96.4	59	84.3

Note: Presumptive criteria 1: at least one HbA1c ≥ 6.5%; presumptive criteria 2: at least one fasting glucose ≥ 126 mg/dL; presumptive criteria 3: at least one random plasma glucose ≥ 200 mg/dL; presumptive criteria 4: at least one diagnosis code for diabetes from an inpatient or outpatient encounter; presumptive criteria 5: at least one prescribed, administered, or dispensed diabetes-related medication, including metformin, sulfonylureas, alpha-glucosidase inhibitors, thiazolidinediones, meglitinides, dipeptidyl peptidase 4 inhibitors, amylin analogs, GLP-1 receptor agonists, SGLT-2 inhibitors. Mg refers to milligrams; dL refers to deciliter.

Table 2.

Performance of decision-tree-based and rule-based algorithms in the presumptive cases.

	Performance	Non-diabetes/Other types of diabetes	T1D	T2D
Decision-tree-based algorithm	Sensitivity	0.93	0.93	0.93
	Specificity	0.96	0.99	0.95
	PPV	0.97	0.97	0.71
Rule-based algorithm	Sensitivity	0.93	0.94	0.95
	Specificity	0.98	0.98	0.95
	PPV	0.98	0.96	0.73
SEARCH algorithm	Sensitivity	0.87	0.95	0.96
	Specificity	0.99	0.97	0.92
	PPV	0.99	0.94	0.62

Note: PPV refers to positive predictive value; SEARCH refers to SEARCH for Diabetes in Youth Study, a national multi-center study aimed to aimed at understanding diabetes among children and young adults in the United States. Other types of diabetes include secondary diabetes mellitus, diabetes mellitus due to underlying condition, drug or chemical induced diabetes mellitus, other specified diabetes mellitus.