




BMJ Open Using electronic health records to enhance surveillance of diabetes in children, adolescents and young adults: a study protocol for the DiCAYA Network

Annemarie G Hirsch ¹, Sarah Conderino,² Tessa L Crume,³ Angela D Liese,⁴ Anna Bellatorre,³ Stefanie Bendik,² Jasmin Divers,⁵ Rebecca Anthopoulos,² Brian E Dixon,^{6,7} Yi Guo ⁸, Giuseppina Imperatore,⁹ David C Lee ², Kristi Reynolds,¹⁰ Marc Rosenman,^{11,12} Hui Shao,¹³ Levon Utidjian,¹⁴ Lorna E Thorpe,² The DiCAYA Study Group

To cite: Hirsch AG, Conderino S, Crume TL, *et al.* Using electronic health records to enhance surveillance of diabetes in children, adolescents and young adults: a study protocol for the DiCAYA Network. *BMJ Open* 2024;**14**:e073791. doi:10.1136/bmjopen-2023-073791

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (<http://dx.doi.org/10.1136/bmjopen-2023-073791>).

Received 16 March 2023
Accepted 20 December 2023



© Author(s) (or their employer(s)) 2024. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

For numbered affiliations see end of article.

Correspondence to

Dr Annemarie G Hirsch;
aghirsch@geisinger.edu

ABSTRACT

Introduction Traditional survey-based surveillance is costly, limited in its ability to distinguish diabetes types and time-consuming, resulting in reporting delays. The Diabetes in Children, Adolescents and Young Adults (DiCAYA) Network seeks to advance diabetes surveillance efforts in youth and young adults through the use of large-volume electronic health record (EHR) data. The network has two primary aims, namely: (1) to refine and validate EHR-based computable phenotype algorithms for accurate identification of type 1 and type 2 diabetes among youth and young adults and (2) to estimate the incidence and prevalence of type 1 and type 2 diabetes among youth and young adults and trends therein. The network aims to augment diabetes surveillance capacity in the USA and assess performance of EHR-based surveillance. This paper describes the DiCAYA Network and how these aims will be achieved.

Methods and analysis The DiCAYA Network is spread across eight geographically diverse US-based centres and a coordinating centre. Three centres conduct diabetes surveillance in youth aged 0–17 years only (component A), three centres conduct surveillance in young adults aged 18–44 years only (component B) and two centres conduct surveillance in components A and B. The network will assess the validity of computable phenotype definitions to determine diabetes status and type based on sensitivity, specificity, positive predictive value and negative predictive value of the phenotypes against the gold standard of manually abstracted medical charts. Prevalence and incidence rates will be presented as unadjusted estimates and as race/ethnicity, sex and age-adjusted estimates using Poisson regression.

Ethics and dissemination The DiCAYA Network is well positioned to advance diabetes surveillance methods. The network will disseminate EHR-based surveillance methodology that can be broadly adopted and will report diabetes prevalence and incidence for key demographic subgroups of youth and young adults in a large set of regions across the USA.

STRENGTHS AND LIMITATIONS OF THIS STUDY

- ⇒ Electronic health record-based surveillance systems offer a potential opportunity to obtain more efficient and timely information on disease prevalence and incidence than is obtained from traditional disease surveillance.
- ⇒ The Diabetes in Children, Adolescents and Young Adults (DiCAYA) Network's large and diverse population will facilitate the estimation of diabetes prevalence and incidence by diabetes subtype for key demographic subgroups of youth and young adults in a large set of regions across the USA.
- ⇒ The diversity of clinical centres in the DiCAYA Network allows for the development and dissemination of surveillance methodology that is generalisable to a variety of settings with access to electronic health record data.
- ⇒ Because electronic health record data are limited to individuals affiliated with the reporting health systems and these populations may differ from the general population, the DiCAYA Network is testing bias adjustment and denominator selection approaches.
- ⇒ A limitation of using electronic health record data for surveillance is that the data were collected for a different purpose (ie, clinical care and billing) and thus lack the rigour and standardisation of traditional research data.

INTRODUCTION

More than 529 million people worldwide and 37 million people in the USA have diabetes.^{1–4} Among children and adolescents in the USA, diabetes is now the third most common chronic disease.⁵ Although prevalence and incidence have recently stabilised in the adult population, diabetes continues to increase among youth, with patterns varying by race

and ethnicity.^{6–9} People with early-onset diabetes face higher risk of chronic kidney disease, myocardial infarction and stroke at younger ages than those who develop diabetes later in life.^{10–13} Surveillance of diabetes is thus a critical function for public health authorities, in understanding the changing epidemiology of diabetes, guiding prevention strategies, allocating resources to at-risk communities and informing health policies for different age groups.

Achieving timely and valid estimates of diabetes is challenging. National estimates of diabetes in adults in the USA are based on surveys, but survey-based methods have been challenged by declining response rates and growing concerns regarding non-response bias.¹⁴ These methods can also be limited in their ability to identify diabetes type and to produce reliable estimates in children and adolescents.¹⁵ For example, the prevalence of diabetes by type using the National Health and Nutrition Examination Survey data is based on age of diagnosis and insulin use, an approach that is susceptible to type misclassification as patterns and treatment of type 1 and type 2 diabetes (T1D and T2D) change.¹⁵ Therefore, data on trends in incidence and prevalence of diabetes by type among young adults are poorly understood. Among youth, survey-based approaches also generate less accurate estimates given the lower disease burden in this age group.¹⁵

To address the limitations of traditional disease surveillance approaches, the Centers for Disease Control and Prevention (CDC) developed specialised surveillance efforts, including the SEARCH for Diabetes in Youth (SEARCH) study in 2000 and the Diabetes in Young Adults study in 2017 to establish diabetes registries using active case-finding surveillance efforts from networks of health systems.^{16 17} These initiatives have provided critical findings on the epidemiology of diabetes by type in children and young adults. SEARCH teams also piloted and validated new methods for improving the timeliness, efficiency and sustainability of surveillance of youth-onset diabetes using electronic health records (EHRs).¹⁸ Other federally funded consortiums, including the Surveillance, Prevention and Management of Diabetes Mellitus Study¹⁹ and the Veterans Affairs Diabetes Epidemiology Cohort,²⁰ have developed EHR-based approaches for identifying adults with diabetes, though the methods did not differentiate by diabetes subtype. Findings of these and other studies²¹ suggested that EHR-based surveillance had promise, but further refinement of methods across broader geographical areas was needed.

In 2020, CDC and the National Institutes of Diabetes and Digestive and Kidney Diseases jointly funded the Diabetes in Children, Adolescents and Young Adults (DiCAYA) Network through 2025. The DiCAYA Network aims to advance the efficiency, flexibility, sustainability and transportability of diabetes surveillance efforts in youth and young adults through the use of large-volume EHR data. The DiCAYA Network was competed through an open request for proposals process involving scientific review from CDC. Geographically diverse sites around

the USA were selected to work together as a network to jointly develop and evaluate innovative approaches to surveillance of diabetes in the target populations. The premise behind DiCAYA was that EHR-based surveillance holds promise for being relatively low cost, as no additional efforts for prospective data collection are required. Importantly, EHR systems can provide timely surveillance data, as data are collected in real time as people interact with the healthcare system, and case identification can be automated. EHR systems also offer large population sizes that overcome the sample size challenges of monitoring relatively rare diseases. The DiCAYA Network will conduct network-wide diabetes surveillance and test bias-adjustment methods, with the goal of informing future EHR-based surveillance strategies at the national level. The network will disseminate EHR-based surveillance methodology that can be broadly adopted and will report diabetes prevalence and incidence in youth and young adults by subtype, race/ethnicity, sex and age. This paper describes the DiCAYA Network, its structure and the methods that will be used to conduct EHR-based diabetes surveillance in youth and young adults in the USA. The protocol represents the work of multiple public health researchers and practitioners, all of whom aim to collectively advance surveillance of diabetes using EHR systems, applying methods that can be replicated by other institutions.

METHODS/DESIGN

Network overview

The DiCAYA Network is spread across eight US-based centres and a Coordinating Centre (CoC), with three centres conducting surveillance in youth aged 0–17 years only (component A), three centres conducting surveillance in young adults aged 18–44 years only (component B) and two centres conducting surveillance in both youth and young adults (components A and B) (figure 1). Component A sites include OneFlorida+ (OFL), PEDSnet and University of South Carolina (UofSC). Component B sites include Geisinger, Indiana University along with the Regenstrief Institute (IU/Regenstrief) and Kaiser Permanente Southern California (KPSC). Two centres are both component A and B sites, Lurie Children's Hospital (Lurie Children's) and University of Colorado Denver (CO). The CoC is housed at New York University (NYU) Langone Health, with researchers at NYU Long Island School of Medicine and NYU Grossman School of Medicine. The Network is composed of three types of centres—geographical-based centres (CO and UofSC), membership-based centres (KPSC) and health system centres (PEDSnet, Geisinger, IU/Regenstrief, Lurie Children's and OFL). Geographical-based centres represent well-delineated geographical and administrative areas and are designed to cover the entire states of Colorado and South Carolina. The membership-based centre, KPSC, is an integrated healthcare delivery system that combines health coverage and care delivery. Members of

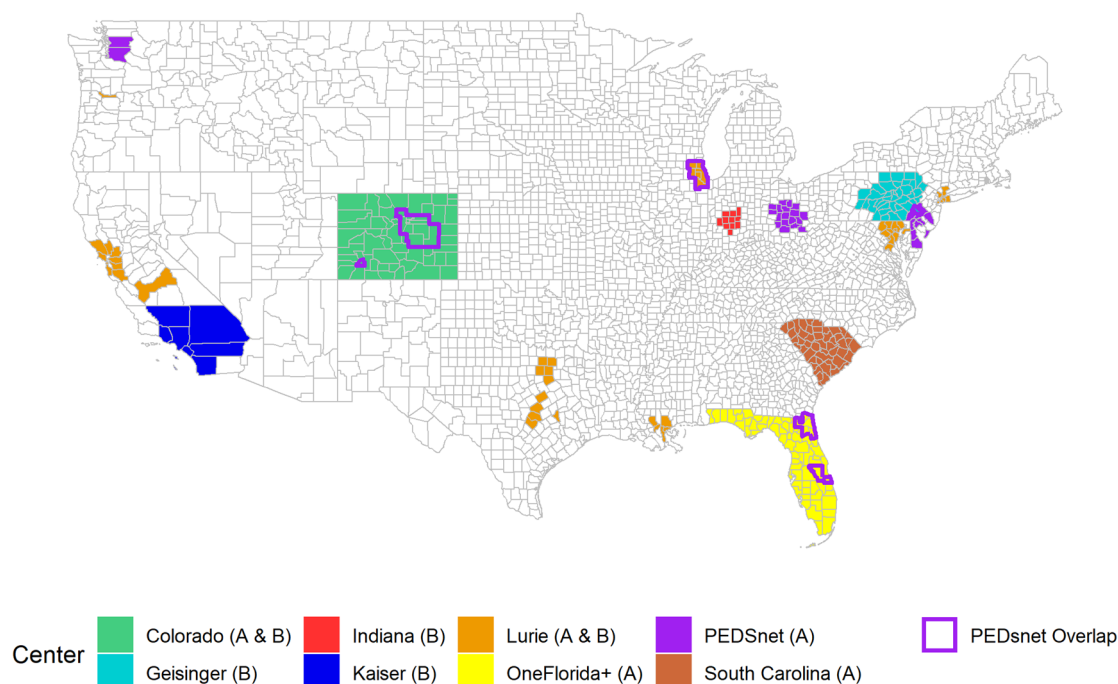


Figure 1 Map of counties included in the DiCAYA network by clinical centre. DiCAYA, Diabetes in Children, Adolescents and Young Adults.

the health plan prepay for and access all aspects of health-care from the same system, while health system centres represent healthcare delivery systems that deliver care to patients, with a range of payers (including the uninsured), who may not receive all aspects of their care from a single healthcare delivery system. Membership-based and health system centres access data from their given EHR data repository or from existing National Patient-Centered Outcomes Research Network (PCORnet) clinical research networks (CRNs). Geographical-based centres receive and integrate independent EHR data streams from all major health systems, with augmentation of records from medical claims data within their respective states (see [table 1](#)). Collectively, these centres cover approximately 36million patients, although the exact patient numbers will only be available when overlap in patient population across centres is determined.

The DiCAYA Network has two primary aims, including (1) to refine and validate EHR-based computable phenotype algorithms for accurate identification of incident and prevalent T1D and T2D among two age groups, youth (<18 years of age) and young adults (18 to <45 years of age), according to age, sex, race/ethnicity and geography and (2) to estimate the incidence and prevalence of T1D and T2D among youth and young adults and trends therein between 2018 and 2024, according to age, sex, race/ethnicity and geography. The protocol for achieving aims 1 and 2 is described in detail below. Ultimately, the network aims to augment diabetes surveillance capacity in the USA and assess performance of

EHR-based surveillance with respect to appropriate surveillance performance metrics (eg, simplicity, data quality, completeness, acceptability, accuracy, representativeness and timeliness).²²

Development of computable phenotype definitions

Aim 1 encapsulates the foundational research needed for the DiCAYA Network to assess the validity of a set of computable phenotype definitions, derived from data that can be processed from EHR systems, to determine diabetes status and type among youth and young adults. The performance of the computable phenotype definitions will be assessed by measuring sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) of the phenotypes against the gold standard of a sample of manually abstracted medical charts. Performance will be assessed for a number of computable phenotype definitions from the literature that can be implemented with the available EHR data, leading to refinement and ultimately the identification of one or more valid computable phenotypes. This process will be harmonised across centres, facilitated by a standardised REDCap (Research Electronic Data Capture) abstraction form and manual of procedures.

Applying computable phenotype definitions proceeds through a sequence of steps, based on previously published methods¹⁸ ([figure 2](#)). All individuals with any indication in the EHR of possible diabetes in each centre's source populations (termed 'wide net') are identified by applying the following criteria during the

Table 1 Centre descriptions

Centre name	Component*	Centre source Population†	Data streams	Geographical coverage (State)	No of counties covered
University of South Carolina	A	1.1 million	Two large health systems (Medical University of South Carolina; Prisma Health) and network for providers/institutions including three independent hospitals, one private paediatric endocrinology practice, three federally qualified health centres.	SC	46/46
University of Colorado Denver (CO)	A	1.3 million	<p>Seven health systems:</p> <ul style="list-style-type: none"> ► The Children's Hospital of Colorado ► Valley Wide Health Systems ► Intermountain Health ► Denver Health and Hospital Authority ► The Barbara Davis Centre ► University of Colorado Hospital ► Centura Health <p>One data warehouse: Health Data Compass</p> <p>Colorado All Payers Claims database</p>	CO	64/64
PEDSnet	A	6.5 million	<p>Subset of PEDSnet⁴⁹ institutions, a PCORnet Clinical Research Network⁴⁷</p> <ul style="list-style-type: none"> ► Children's Hospital of Philadelphia ► Cincinnati Children's Hospital Medical Centre ► The Children's Hospital of Colorado‡ ► Nationwide Children's Hospital ► Nemours Children's Health System (both Delaware and Florida‡ systems) ► Seattle Children's Hospital ► Ann & Robert H. Lurie Children's Hospital of Chicago‡ 	DE FL NJ OH PA WA CO IL	3/3 7/67 4/21 18/88 5/67 2/39 11/64 6/102
University of Florida (OFL)	A	12.0 million	OneFlorida+Clinical Research Consortium, ⁵⁰ a PCORnet Clinical Research Network ⁴⁷	FL	67/67
Lurie Children's Hospital (Lurie Children's)	A	2.0 million	<p>Several Clinical Research Networks from PCORnet⁴⁷: Lurie Children's Hospital (of the PEDSnet CRN), and sites in Illinois from the CAPriCORN CRN⁵¹</p> <p>REACHnet CRN⁵²</p> <p>INSIGHT CRN^{53 54}</p> <p>ADVANCE CRN⁵⁵</p> <p>Johns Hopkins University (of the PaTH CRN)⁵⁶</p>	IL LA TX CA NJ NY OR CA DC MD	6/102 5/64 10/254 9/58 1/21 7/62 1/36 1/58 1/1 11/23
University of Colorado Denver (CO)	B	2.2 million	See component A for details	CO	64/64
Kaiser Permanente of Southern California	B	4.6 million	EHR system (Epic), medical claims and administrative data	CA	7/58
Geisinger Clinic (Geisinger)	B	351 628	EHR system (Epic) Local PCORnet ⁴⁷ CDM PaTH CRN	PA	38/67

Continued

Table 1 Continued

Centre name	Component*	Centre source Population†	Data streams	Geographical coverage (State)	No of counties covered
Indiana University–Purdue University at Indianapolis (IU/Regenstrief)	B	2.1 million	Indiana Network for Patient Care clinical data repository	IN	11/92
Lurie Children's Hospital (Lurie Children's)	B	4.0 million	See component A for details.	Multiple states	52/679

*A: ages 0–17 years, B: ages 18–44 years.

†Population from which the wide net was drawn for the 2018 index year.

‡In the primary analysis, data from these institutions will not be included in the PEDSnet, but they may be included in PEDSnet in subsequent secondary DiCAYA analyses in which PEDSnet is its entirety as a bloc.

CA, California; CAPriCORN CRN, Chicago Area Patient-Centered Outcomes Research Clinical Research Network; DC, District of Columbia; DE, Delaware; DiCAYA, Diabetes in Children, Adolescents and Young Adults; EHR, electronic health record; FL, Florida; LA, Louisiana; MD, Maryland; NJ, New Jersey; NY, New York; OFL, OneFlorida+; OH, Ohio; OR, Oregon; PA, Pennsylvania; PaTH CRN, A Path Towards a Learning Health System Clinical Research Network; PCORnet, The National Patient-Centered Clinical Research Network; REACHnet, Research Action for Health Network; SC, South Carolina; TX, Texas; WA, Washington.

time window (the index surveillance year for state-based and membership-based centres, index surveillance year and prior 2 years for health system-based centres) in the respective age group (0–17, 18–44 years): (1) ≥ 1 haemoglobin A1c $\geq 6.5\%$; (2) ≥ 1 fasting glucose ≥ 126 mg/dL; (3) ≥ 1 random plasma glucose ≥ 200 mg/dL; (4) ≥ 1 diabetes-related diagnosis code from an inpatient or outpatient encounter (online supplemental file 1) or (5) ≥ 1 prescribed, administered or dispensed medication that is typically indicated for the treatment of diabetes (online supplemental file 2). The wide net was designed to have maximum sensitivity, to avoid missing any true diabetes cases.¹⁸

Next, the primary computable phenotype for presumed diabetes will be applied to the wide net population. The computable phenotype is defined as those with at least one diabetes diagnosis code (International Classification of Disease (ICD)-10-CM: E08–E11, E13) (online supplemental file 1) within the given time windows, based on a method that has been previously used in a cohort of individuals with youth-onset diabetes.²³ While an individual could have a code for gestational diabetes in the EHR, a code for gestational diabetes would not be sufficient to be classified as presumed diabetes. Consistent with prior literature, diabetes type (type 1, type 2, other) will be defined based on the proportion of diabetes type-specific diagnosis codes (type 1, type 2 or other) among total diabetes codes, using plurality to assign type.²³ In ties, type 1 is given preference over type 2, and type 2 is given preference over others.

To calculate diabetes incidence rates, computable phenotypes to distinguish newly diagnosed diabetes cases from existing cases will also be defined and validated. New diabetes cases will be defined as those who met the presumed diabetes case definition (ie, those who had at least one diagnosis code for diabetes) for the first time in the index surveillance year (ie, no prior diabetes

diagnosis). We will assess whether also to require a record of an earlier healthcare encounter, from before the date of diabetes diagnosis, when determining incidence in the healthcare systems or geographical surveillance areas. We will make this assessment using sensitivity analyses, chart reviews and data available from the EHR (eg, each individual's first-ever encounter date; each individual's first diabetes diagnosis code date; and each individual's last healthcare encounter date before the first diabetes diagnosis date).

All working computable phenotype definitions will be iteratively refined over the course of the project based on results of the validation and refinement study analyses. We will conduct a manual chart review on a subset of patients who meet the wide net criteria. Sensitivity, specificity, PPV and NPV for the working computable phenotypes will be calculated among the wide net patients for whom true diabetes status has been determined through the chart review. Given that the wide net includes individuals with any evidence of diabetes (ie, medication, laboratory measures and diagnoses), we will have the ability to compare multiple phenotype algorithms (including those that use medication or laboratory-based criteria) to wide net patients with completed chart reviews. Sensitivity will be calculated as the true positives (classified as having diabetes by chart review and by computable phenotype) divided by all those classified as having diabetes by manual review (including true positives and false negatives or those classified as having diabetes by manual review but not by computable phenotype). Specificity will be calculated as the true negatives (classified as not having diabetes by chart review and by computable phenotype) divided by all those classified as not having diabetes by manual review (including true negatives and false positives, or those classified as having diabetes by computable phenotype but not by manual review). PPV will be calculated as the true positives divided by all those

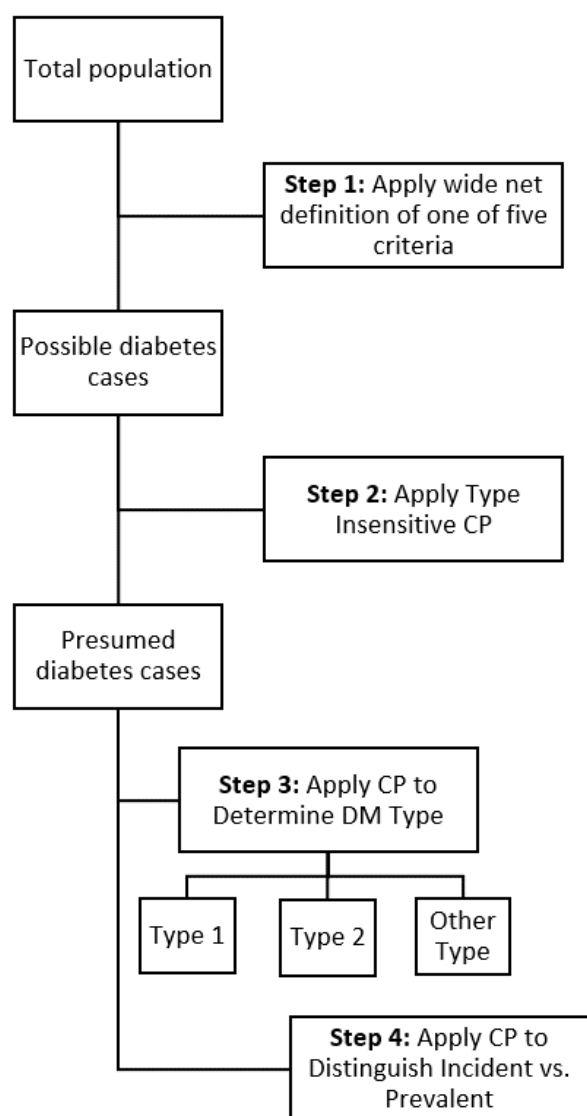


Figure 2 Computable phenotype (CP) for diabetes (DM) flow chart.

classified as having diabetes by computable phenotype (including true positives and false positives). Finally, NPV will be calculated as the true negatives divided by all those classified as not having diabetes by computable phenotype (including true negative and false negatives). These performance measures will be calculated overall and by subgroup (age, sex and race/ethnicity).

In preparation for the validation and refinement study, the network calculated the necessary sample sizes for the chart reviews using separate component A and component B detectable effect sizes for differences in sensitivity estimates across computable phenotypes, defining the overall detectable effect as the maximum of the two detectable effect sizes. The use of the wide net for validation reduces the number of charts that we would have to manually review to identify a sufficient number of true diabetes cases for measurement of sensitivity. In order to compare computable phenotypes from the literature, a

correlation of 0.707 ($R^2=0.5$) between phenotypes was assumed.^{24–26} To achieve an overall 80% power to detect small differences in sensitivity and specificity, the network will perform manual chart review on approximately 2600 wide net patients per component. This sample will be allocated across centres using a minimum of 400 individuals per centre, with the remaining sample proportional to the size of the centre's wide net population, up to a maximum of 750 individuals. Samples will be selected using stratified random sampling, with oversampling by race/ethnicity to achieve 10% Asian and Pacific Islander and 20% black individuals in the total sample (table 2) to facilitate evaluation of validity by race. Validity will also be assessed by age, sex and ethnicity.

To accomplish aim 2, estimation of incidence and prevalence of T1D and T2D among youth and young adults, the computable phenotypes developed in aim 1 will first be applied to each centre's EHR sample to define the numerators. Details on the EHR data used at each centre are outlined in table 1. Next, denominators for prevalence and incidence estimates will be defined using different methods based on type of centre. For geographical-based centres that draw from all major health systems in their respective states, the state civilian, non-institutionalised population will serve as the denominator, as determined using the 2020 US census data and the CDC National Center for Health Statistics' race-bridged post-census estimates of resident US population. Denominators for health system centres will be generated in two ways: (1) using utilisation data and (2) using US census population-based data. To capture the health system utilisation denominators, these centres will first define the number of unique patients with at least one health system encounter during a 3-year window that includes the index year and the two previous years, a time window selected based on national estimates of frequency of healthcare utilisation in the target age groups.²⁷ Based on patients' latest addresses at the start of the index year, health system centres will generate the coverage for each county represented in their utilisation data, defined as the number of unique patients divided by the appropriate population size for the county (including in subpopulations by sex, race/ethnicity and age). The network will evaluate different inclusion criteria for counties under surveillance, including coverage level and geographical contiguity/proximity to the health system. Population-based denominators for the selected counties will be defined using an average of 3 years (index year and two prior years). Finally, for the membership-based centre, the number of members of the health plan as of 1 January of the index year, determined through administrative databases, will serve as the denominator.

For all centres, relevant inclusion and exclusion criteria will be applied to ensure that the denominators represent the at-risk population defined in the

Table 2 Chart review sampling assignments for the computable phenotype refinement analysis

	Total	Asian/PI	Black	Hispanic	Other	Unknown	White
Component A							
PEDSnet	705 (27.1%)	92 (36.4%)	141 (25.7%)	162 (27%)	35 (27.1%)	35 (27.1%)	240 (25.5%)
OFL	588 (22.6%)	59 (23.3%)	118 (21.5%)	135 (22.5%)	29 (22.5%)	29 (22.5%)	218 (23.1%)
Lurie Children's	426 (16.4%)	43 (17%)	85 (15.5%)	98 (16.4%)	21 (16.3%)	21 (16.3%)	158 (16.8%)
UofSC	421 (16.2%)	13 (5.1%)	113 (20.6%)	97 (16.2%)	21 (16.3%)	21 (16.3%)	156 (16.5%)
Colorado	462 (17.8%)	46 (18.2%)	92 (16.8%)	107 (17.9%)	23 (17.8%)	23 (17.8%)	171 (18.1%)
Total	2602 (100%)	253 (100%)	549 (100%)	599 (100%)	129 (100%)	129 (100%)	943 (100%)
Component B							
Geisinger	469 (18%)	47 (18.1%)	94 (18.1%)	150 (18%)	5 (19.2%)	19 (18.1%)	154 (17.9%)
Lurie Children's	543 (20.9%)	54 (20.8%)	109 (21%)	174 (20.9%)	5 (19.2%)	22 (21%)	179 (20.9%)
Indiana	499 (19.2%)	50 (19.2%)	99 (19%)	160 (19.2%)	5 (19.2%)	20 (19%)	165 (19.2%)
Kaiser Permanente Southern California	652 (25%)	65 (25%)	130 (25%)	209 (25.1%)	7 (26.9%)	26 (24.8%)	215 (25.1%)
Colorado	440 (16.9%)	44 (16.9%)	88 (16.9%)	141 (16.9%)	4 (15.4%)	18 (17.1%)	145 (16.9%)
Total	2603 (100%)	260 (100%)	520 (100%)	834 (100%)	26 (100%)	105 (100%)	858 (100%)
OFL, OneFlorida+; UofSC, University of South Carolina.							

numerator. Preliminary analyses compared the population living in the DiCAYA counties to the rest of the US by age, sex, race, ethnicity and socioeconomic status (SES). Initial results show that distributions by age and sex are comparable, but non-Hispanic black and Hispanic people were over-represented, and non-Hispanic white persons were under-represented (table 3).

Unadjusted prevalence and incidence rates overall and for demographic or geographical subgroups will be estimated as the ratio of the numerator divided by the appropriate denominator. The prevalence will be expressed as the number of diabetes cases per 1000 individuals in a defined period. Incidence rates will be expressed as the number of newly diagnosed cases in a calendar year per 100 000 individuals. Incidence estimates will be provided for calendar years 2018 through 2024, and prevalence estimates will be provided for select calendar years from 2018 to 2024. Skew-corrected inverted score tests for binomial distribution will be used to compare two rates and compute 95% CIs. The prevalence and incidence rates will be presented as unadjusted estimates and as race/ethnicity, sex and age-adjusted estimates using Poisson regression. Analyses will be run separately for components A and B.

Bias correction and estimation

A key limitation to using EHR data for population health surveillance is the potential for patient populations to be non-representative of the general target population of inference. For example, EHRs have greater coverage among women and children, and those who frequent health systems tend to be more ill than the general population.¹⁹ In the analysis of non-probability samples such as EHRs, two main methodological frameworks may be

used to estimate population quantities.^{28–30} In the quasi-randomisation framework, pseudoinclusion probabilities are estimated based on covariates available for all population units and used to correct for selection bias. In contrast, in the superpopulation modelling approach, a statistical model is assumed for the outcome of interest in the non-probability sample and applied to the target population. Both quasi-randomisation and superpopulation modelling rely to varying degrees on auxiliary data from external surveys or administrative sources. Multilevel regression with poststratification (MLRP) is a variation on superpopulation modelling that has often been used in political science. Using MLRP with a highly non-representative survey sample from the Xbox gaming platform, Wang *et al*³¹ were able to predict the 2012 US presidential election.³¹ Although MLRP may be conducted in a frequentist or in a Bayesian setting, the latter may be well suited to handle issues of data sparsity. The DiCAYA Network will implement MLRP to produce estimates of incidence and prevalence. In sensitivity analysis, we will explore existing Bayesian hierarchical models that have been used in survey samples^{32–34} and will compare them to common survey methods such as raking and poststratification³⁵ that help correct for non-representativeness.

The DiCAYA Network will implement MLRP to produce estimates of incidence and prevalence. As sensitivity analyses, the network will also apply additional bias-adjustment methods, including propensity score weighting, poststratification, empirical Bayesian hierarchical modelling and geospatial small-area estimation.^{32–36} Through the bias correction methods, the network will generate estimates of prevalence and incidence rates of diabetes (type insensitive (type 1, type 2 or other), type 1 and type 2) in youth

Table 3 Comparison of population demographics within counties included in the DiCAYA Network versus all US counties

	Component A: ages 0–17 years				Component B: ages 18–44 years			
	US total population, 2020 census		Population, DiCAYA counties, US 2020 census		US total population, 2020 census		Population, DiCAYA counties, US 2020 census	
	N	%	N	%	N	%	N	%
Sex								
Female	35627710	48.9	9433357	49.0	58406670	49.4	15526923	49.5
Male	37194403	51.1	9824535	51.0	59826165	50.6	15832605	50.5
Age group (years)								
0–4	19301292	26.5	5183565	26.9				
5–9	20237711	27.8	5361445	27.8				
10–14	20754423	28.5	5443891	28.3				
15–17	12528687	17.2	3268991	17.0				
18–19					8432242	7.1	2040929	6.5
20–24					21594755	18.3	5303536	16.9
25–29					23231243	19.7	6348490	20.2
30–34					22838403	19.3	6372194	20.3
35–39					21828304	18.5	5910928	18.8
40–44					20307888	17.2	5383451	17.2
Race/ethnicity								
AI	1445621	2.0	248509	1.3	1931016	1.6	478258	1.5
API	5011051	6.9	1847103	9.6	9219241	7.8	4125945	13.2
His	17463322	24.0	5155471	26.8	23975505	20.3	9308455	29.7
NHB	11123707	15.3	3733021	19.4	17001026	14.4	4172525	13.3
NHW	37778412	51.9	8273788	43.0	66106047	55.9	13274345	42.3
Total	72822113		19257892		118232835		31359528	

AI, American Indian; API, Asian Pacific Islander; DiCAYA, Diabetes in Children, Adolescents and Young Adults; NHB, non-Hispanic Black; NHW, non-Hispanic White.

and young adults by various demographic characteristics, including race/ethnicity, age and sex.

Patient and public involvement

Patients and the public were not involved in the development of this protocol.

ETHICS AND DISSEMINATION

The DiCAYA Network is well positioned to lead a critical advancement in surveillance of diabetes and diabetes types in youth and young adults. While validation studies are needed, EHR-based surveillance systems offer an opportunity to mount more efficient systems than methods used for traditional disease surveillance. The use of existing data and potentially automated case ascertainment (ie, computable phenotypes) make EHR-based surveillance timely and flexible, critical features of a public health surveillance system.²¹ DiCAYA's large and diverse population will facilitate the estimation of diabetes prevalence and incidence by type for key demographic subgroups of youth and young adults in a large set of geographical regions across the USA. Through dissemination of the methods and results, we will inform

future strategies for conducting nationwide EHR-based surveillance of diabetes.

Each DiCAYA centre and the CoC received approval from their local institutional review boards for this protocol. To facilitate network-wide analyses at the CoC, each centre executed a data use agreement with the CoC that permits the sharing of EHR data elements with the CoC (online supplemental file 3). Data transfers between centres and the CoC are conducted via a secure file transfer protocol. The CoC manages these data centrally on a secure central platform. Centres have access to their own individual-level data and aggregate data from the network.

There are a number of potential limitations to using EHR data for population health surveillance. First, EHR data are limited to individuals affiliated with the reporting health systems, and these populations may differ from the general population for a variety of reasons, including services received, health insurance coverage and health insurance types.³⁷ EHRs include data on the subset of the population that seeks care, potentially biasing EHRs toward greater coverage of women, children and individuals who are more ill.³⁷ Moreover, the fragmentation

of healthcare in the USA implies that not all health conditions of a given individual are reflected in the EHR under study. DiCAYA's geographical-based centres are likely less vulnerable to this limitation, given their state-wide coverage and use of multiple data sources. The membership-based centre is also less vulnerable, given that all aspects of each member's healthcare and services are captured in the EHR, and members have a unique medical record number that does not change if members leave and rejoin the health plan. DiCAYA will deploy MLRP and Bayesian hierarchical modelling^{28 33} to minimise some of these biases. For a target population of interest, MLRP can combine data in EHRs with rich information from auxiliary sources like the census. MLRP may be especially useful when it is reasonable to hypothesise that selection into the EHR sample is not associated with missed outcomes from excluded individuals after accounting for observed information (ie, missing at random²⁹). However, the performance of MLRP may be sensitive to issues of model misspecification. A Bayesian framework may be conducive to the more complex scenario when the selection process is missing not at random²⁹ and the underlying health status is plausibly associated. Second, while the use of computable phenotypes based on discrete and easily extractable EHR data (eg, diagnoses codes, laboratory values and medications) is essential for large-scale and efficient surveillance, this approach does not leverage potentially informative free text clinician notes. Therefore, in developing our computable phenotypes, we will assess their performance as compared with manual chart review, and some sites will explore the use of natural language processing of free text data from the EHR to automate identification of key variables for diagnosing diabetes cases. Third, our proposed primary computable phenotype may result in some misclassification of disease, particularly among individuals who were initially misdiagnosed (eg, diagnosed as T2D before T1D was ultimately diagnosed or vice versa). Sensitivity of computable phenotypes using diagnoses alone (ie, not laboratory or medication evidence) may also be limited. Prior work has demonstrated, for example, that including laboratory results increases the sensitivity of diabetes computable phenotypes.³⁸ We will examine the extent of misclassification as well as the sensitivity, specificity and PPV and NPVs of all working computable phenotype definitions and anticipate that the computable phenotypes will be iteratively refined over time based on results of the validation analyses. Lastly, an inherent limitation of using EHR data is that data were collected for a different purpose (ie, clinical care, billing and operations) and thus lack the rigour and standardisation of traditional research data.

There are also some limitations inherent to the DiCAYA Network, as constructed. First, the DiCAYA Network is designed to provide prevalence and incidence estimates on populations in care who have been screened, treated and/or diagnosed with diabetes, potentially representing populations with higher SES than the general

population.³⁹ This could result in prevalence estimates that are lower than in the general population, given the higher prevalence of diabetes⁴⁰ among individuals with lower SES. Second, while DiCAYA has more geographical coverage than prior specialised surveillance or pilot efforts, there remains limited coverage in the Northeast, Northwest, North Central and Midwest regions of the USA. Conversely, in some parts of the USA, DiCAYA centres serve overlapping geographies, potentially leading to overlapping patient populations. For centres with overlapping geographies, we will conduct sensitivity analyses to determine the impact of removing data from overlapping centres. In addition, while SEARCH provided important insights into the great burden of diabetes among American Indian youth,⁴¹ DiCAYA will be limited in its ability to conduct surveillance in the American Indian population. Finally, completeness of case ascertainment is a key characteristic of surveillance that can be assessed when there is a second, independent source of cases.⁴² In the DiCAYA Network, only the state-based sites in South Carolina (component A) and Colorado (components A/B), have data sources required to assess completeness for a select number of years, capitalising on their prior SEARCH infrastructure and within-state network design to allow for complete coverage of complementary healthcare utilisation across hospital systems within the state over time. A subset of DiCAYA's health system centres gather data from multiple health systems, through participation in a health information exchange network or CRNs (eg, INSIGHT CRN) that may facilitate more complete case ascertainment than a single health centre serving part of the population in a geographical area.⁴³

Despite these limitations, the DiCAYA Network offers several important strengths. The diversity of clinical centres that comprise the DiCAYA Network allows for the development of surveillance methodology that is generalisable to a variety of settings with access to EHR data. As of 2019, about three-quarters of office-based physicians and nearly all non-federal acute care hospitals in the USA had adopted a certified EHR system.⁴⁴ The WHO reported that 47% of countries had national EHR systems, as of 2016.⁴⁵ By developing and validating computable phenotypes across a range of settings in distinct geographies, DiCAYA will publish automated approaches to case ascertainment of T1D and T2D that can be replicated broadly, in the USA and abroad. These EHR-based approaches can be adapted by other healthcare systems and applied to common data models⁴⁶ (eg, PCORnet,⁴⁷ Observational Medical Outcomes Partnership⁴⁸), which would facilitate rapid dissemination of the DiCAYA surveillance protocol. Similarly, the methods for identifying surveillance denominators from patient populations will be applicable to a range of healthcare delivery settings and chronic disease outcomes.

In addition to delivering surveillance estimates and advancing surveillance methodology, the work of the DiCAYA Network will identify large cohorts of youth

and young adults with incident and prevalent diagnosed T1D and T2D on whom the network has access to longitudinal EHR data. The deidentified data, code and other materials used in this study are available for use from the DiCAYA Network for ancillary studies or in collaboration with DiCAYA Network sites, pending review and approval by the network's Publications and Presentations Committee. The data available on these cohorts can inform future research on risk factors for diabetes onset and complications in these age groups. Importantly, these demographically and geographically diverse cohorts can be used to conduct future research on racial, ethnic and geographical disparities of diabetes in youth and young adults.

Author affiliations

¹Department of Population Health Sciences, Geisinger, Danville, Pennsylvania, USA

²Department of Population Health, New York University Grossman School of Medicine, New York, New York, USA

³Department of Epidemiology, Lifecourse Epidemiology of Adiposity and Diabetes (LEAD), University of Colorado - Anschutz Medical Campus, Aurora, Colorado, USA

⁴Department of Epidemiology and Biostatistics, University of South Carolina, Columbia, South Carolina, USA

⁵Department of Foundations of Medicine, New York University Long Island School of Medicine, Mineola, New York, USA

⁶Department of Epidemiology, Fairbanks School of Public Health, Indiana University, Indianapolis, Indiana, USA

⁷Center for Biomedical Informatics, Regenstrief Institute Inc, Indianapolis, Indiana, USA

⁸Department of Health Outcomes and Biomedical Informatics, University of Florida, Gainesville, Florida, USA

⁹Division of Diabetes Translation, Centers for Disease Control and Prevention, Atlanta, Georgia, USA

¹⁰Department of Research & Evaluation, Kaiser Permanente Southern California, Pasadena, California, USA

¹¹Department of Pediatrics, Ann & Robert H. Lurie Children's Hospital of Chicago, and Northwestern University Feinberg School of Medicine, Chicago, Illinois, USA

¹²Northwestern University Feinberg School of Medicine, Chicago, Illinois, USA

¹³Department of Pharmaceutical Outcomes and Policy, University of Florida College of Pharmacy, Gainesville, Florida, USA

¹⁴Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA

Twitter Brian E Dixon @dpugrad01

Acknowledgements The research reported in this publication was conducted using data from the DiCAYA (Assessing the Burden of Diabetes by Type in Children, Adolescents, and Young Adults) Network. DiCAYA is funded by the Centers for Disease Control and Prevention (CDC) (DP20-001) and the National Institute of Diabetes and Digestive and Kidney Diseases to modernize diabetes surveillance efforts using electronic health record data and advanced statistical analysis. This study includes data from the following institutions: University of South Carolina, University of Colorado Denver, Children's Hospital of Philadelphia, University of Florida, Lurie Children's Hospital, Kaiser Permanente Southern California, Geisinger, and Indiana University-Purdue University Indianapolis. The research reported in this study was also conducted using PEDSnet and the OneFlorida+ Clinical Research Network and several other clinical research networks in the project led by Lurie Children's Hospital. PEDSnet, A Pediatric Learning Health System, includes data from the following PEDSnet institutions: Children's Hospital Colorado, Children's Hospital of Philadelphia, Cincinnati Children's Hospital Medical Center, Lurie Children's Hospital, Nationwide Children's Hospital, Nemours Children's Health, and Seattle Children's Hospital. PEDSnet is a Partner Network Clinical Data Research Network in PCORnet, the National Patient-Centered Clinical Research Network, an initiative funded by the Patient-Centered Outcomes Research Institute (PCORI). OneFlorida+ is a collaboration among researchers, clinicians and patients in Florida, Georgia, and Alabama to create an enduring infrastructure for a wide range of health research, including pragmatic clinical trials, comparative effectiveness research, implementation science studies, observational research,

and cohort discovery. Network partners include the University of Florida, Florida State University, the University of Miami, the University of South Florida, Emory University in Atlanta, and the University of Alabama at Birmingham, along with the six universities' affiliated health systems and practices. Other partners include AdventHealth (Orlando), Tallahassee Memorial HealthCare, Tampa General Hospital, Bond Community Health (Tallahassee), Community Health IT (Kennedy Space Center), Nicklaus Children's Hospital (Miami), Capital Health Plan (Tallahassee), Bendcare (Boca Raton, Florida) and the Florida Agency for Health Care Administration, which oversees the Florida Medicaid Program. OneFlorida+ is also a network partner of PCORnet. The Lurie Children's Hospital DiCAYA project is composed of health care institutions from multiple clinical research networks: The INSIGHT Clinical Research Network (CRN) is a collaborative initiative which integrates clinical and social determinants of health data from over 15 million patients within New York City's leading health systems. As a member of PCORI's PCORnet, INSIGHT operates as one of the largest and most diverse CRNs. The INSIGHT network infrastructure is built to convene and meaningfully engage with patients, caregivers, families, researchers, health system leaders, clinicians, and funders to ensure a universal focus on patient-centered research and health equity. This work is supported in-part by the PCORI PCORnet grant to the INSIGHT Clinical Research Network (grant # RI-CORNELL-01-MC). The approach described in this manuscript was developed in partnership with Research Action for Health Network (REACHnet), funded by PCORI (PCORI Award RI-LPHI-01-MC). REACHnet is a partner network in PCORnet, which was developed with funding from PCORI. The Accelerating Data Value Across a National Community Health Center Network (ADVANCE) Clinical Research Network is a member of PCORnet. ADVANCE is a multicenter collaborative led by OCHIN (not an acronym) in partnership with Fenway Health, Health Choice Network, and Oregon Health & Science University. The Chicago Area Patient-Centered Outcomes Research Network (CAPriCORN) is a partnership between healthcare and research institutions, patients, patient advocates, clinicians, community-based organisations (CBOs), and non-profits committed to enabling and delivering patient-centered clinical research. A Patient Community Advisory Committee (PCAC) has worked with CAPriCORN since its inception to elevate the patient voice in research. CAPriCORN's mission is to develop, test, and implement clinical research in order to improve health care quality, health outcomes, and health equity for the diverse populations of Chicagoland and the surrounding states. Johns Hopkins is a member site of the PaTH CRN. PaTH is a Partner Network in PCORnet, the National Patient-Centered Clinical Research Network. PCORnet has been developed with funding from the Patient-Centered Outcomes Research Institute (PCORI). PaTH's participation in PCORnet was funded through PCORI Award (RI-PITT-01-PS1).

Collaborators The DiCAYA Study Group includes: Children's Hospital of Philadelphia/PEDSnet: Charles Bailey, MD, PhD (MPI); Christopher Forrest, MD, PhD (MPI); Levon Utidjian, MD; Mitch Maltenfort, PhD; Amy Shah, MD (for CCHMC/PEDSnet); Eneida A. Mendonca, MD, PhD (for CCHMC/PEDSnet); G. Todd Alonso, MD (for Colorado/PEDSnet); Sara Deakynne-Davies, MPH (for Colorado/PEDSnet); Tim Bunnell, PhD (for Nemours/PEDSnet); Anne Kazak, PhD (for Nemours/PEDSnet); Melody Kitzmiller, BS (for NCH/PEDSnet); Manmohan Kamboj, MD (for NCH/PEDSnet); Dimitri Christakis, MD, MPH (for SCH/PEDSnet); Daksha Ranade, MPH, MBA (for SCH/PEDSnet). Geisinger: Annemarie G. Hirsch, PhD, MPH (PI); Joseph J. DeWalle, BS; H. Lester Kirchner, PhD; Meredith Lewis, MS; Dione G. Mercer, BS, BA; Cara M. Nordberg, MPH; Amy Poissant, BS; Brian S. Schwartz MD, MS. Indiana University/Regenstrief: Brian E. Dixon, PhD, MPA (PI); Shaun Grannis, MD, MS; Seho Park, MA, PhD, MS; Katie Allen (for Regenstrief Institute); Anna Roberts, MS (for Regenstrief Institute); Nimish Valvi, PhD, MPH (for Regenstrief Institute); Jeff Warvel (for Regenstrief Institute); Ashley Wiensch, MPH (for Regenstrief Institute); Tamara Hannon, MD (for Indiana University). Kaiser Permanente of Southern California: Kristi Reynolds, PhD, MPH (PI); John Chang, MPH; Eva Lustigova, MPH; Don McCarthy, MA; Matthew T. Mefford, PhD; Rong Wei, MA; Hui Zhou, PhD. Lurie Children's Hospital: Marc Rosenman, MD (PI); Lu Zhang, PhD; George Lales, MS; Anthony Wong, PhD; Allison Zelinski, MS (for Children's Hospital of Philadelphia); Yuan Luo, PhD (for Northwestern University); Mark Weiner, MD (for Weill Cornell); Pedro Rivera, MS (for OCHIN); Thomas Carton, PhD, MS (for Louisiana Public Health Institute); Elizabeth Nauman, MPH, PhD (for Louisiana Public Health Institute); Harold P. Lehmann, MD, PhD (for Johns Hopkins University); Victor W. Zhong, PhD (for Shanghai Jiao Tong University). NYU Langone Health: Jasmin Divers, MS, PhD (MPI); Lorna E. Thorpe, MPH, PhD (MPI); Meredith Akerman, MS; Rebecca Anthopolos, DrPH; Stefanie Bendik, MPH; Sarah Conderino, MPH; Andrew Fair, ScM, MS; Jessica Guillaume, MPH; Shahidul Islam, DrPH, MPH; Alan Jacobson, MD; David C. Lee, MD; Chinyere Okpara, MS; Anand Rajan, MPH; Andrea Titus, PhD. University of Colorado Denver: Tessa Crume, PhD, MPH (MPI); Dana Dabelea, MD, PhD (MPI); Theresa Anderson, MS; Anna Bellatorre, PhD, MA; Rebecca Conway, PhD, MPH; Toan Ong, PhD; Jack Pattee, PhD; Shawna Burgett, PhD; Bethelhem (Betty) Shiferaw, MPH.

University of Florida: Jiang Bian, PhD (MPI); Yi Guo, PhD (MPI); Hui Shao, MD, PhD (MPI); Elizabeth A. Shenkman, PhD (MPI); Sarah J. Bost, MSLS; William T. Donahoe, MD; William R. Hogan, MD, PhD; Piaopiao Li; Tianchen Lyu, MS; Mattia Prosperi, PhD; Yonghui Wu, PhD. University of South Carolina: Angela D. Liese, PhD, MPH (MPI); Bo Cai, PhD, MSc; Lisa Knight, MD, MBA; Caroline Rudisill, MSc, PhD; Jessica Stucker, MSW; Deborah Bowlby, MD (for Medical University of South Carolina); Jihad S. Obeid, MD (for Medical University of South Carolina); Elaine Apperson MD (for Prisma Health); Alex Ewing, PhD, MPH (for Prisma Health). Centers for Disease Control and Prevention, Division of Diabetes Translation: Giuseppina Imperatore, Meda Pavkov, Deborah B. Rolka.

Contributors All members of The DiCAYA Study Group contributed to the conception and design of this protocol; AGH, SC, SB and LET drafted the manuscript; AGH, SC, TLC, ADL, AB, SB, JD, RA, BED, YG, GI, DCL, KR, MR, HS, LU and LET substantively revised the manuscript. AGE and SC contributed equally and are co-first authors.

Funding This work was supported by the Centers for Disease Control and Prevention and the National Institute for Diabetes and Digestive and Kidney Diseases. U18DP006521 Children's Hospital of Pennsylvania, U18DP006512 University of Florida, U18DP006509 Geisinger, U18DP006500 Indiana University–Purdue University at Indianapolis, U18DP006513 University of South Carolina, U18DP006506 Kaiser Foundation Hospitals, U18DP006693 Lurie Children's, U18DP006694 Lurie Children's, U18DP006517 University of Colorado Component-A, U18DP006518 University of Colorado Component-B, U18DP006510 NYU Long Island School of Medicine.

Disclaimer 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. The content of this publication is solely the responsibility of the authors and does not necessarily represent the views of PCORI or of other organisations participating in, collaborating with, or funding REACHnet or PCORnet.

Map disclaimer The inclusion of any map (including the depiction of any boundaries therein), or of any geographic or locational reference, does not imply the expression of any opinion whatsoever on the part of BMJ concerning the legal status of any country, territory, jurisdiction or area or of its authorities. Any such expression remains solely that of the relevant source and is not endorsed by BMJ. Maps are provided without any warranty of any kind, either express or implied.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

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

ORCID iDs

Annemarie G Hirsch <http://orcid.org/0000-0001-7699-2171>

Yi Guo <http://orcid.org/0000-0003-0587-4105>

David C Lee <http://orcid.org/0000-0002-7202-1893>

REFERENCES

- Ong KL, Stafford LK, McLaughlin SA. Global, regional, and national burden of diabetes from 19190 to 2021, with projections of prevalence to 2050: a systematic analysis for the global burden of disease study 2021. *Lancet* 2023;402:203–34.
- Saran R, Li Y, Robinson B, et al. US renal data system 2014 annual data report: epidemiology of kidney disease in the United States. *Am J Kidney Dis* 2015;66:S0272–6386(15)00744–1.
- Emerging Risk Factors Collaboration, Sarwar N, Gao P, et al. Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. *Lancet* 2010;375:2215–22.
- Steinmetz JD, Bourne RRA, Briant PS. Blindness and vision impairment collaborators, study Vlegotgbod. causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of Avoidable blindness in relation to VISION 2020: the right to sight: an analysis for the global burden of disease study. *Lancet Glob Health* 2021;9:e144–60.
- Centers for Disease Control and Prevention. *National Diabetes Statistics Report; 2020 Secondary National Diabetes Statistics Report 2020*. 2022. Available: <https://www.cdc.gov/diabetes/library/features/diabetes-stat-report.html>
- Dabelea D, Mayer-Davis EJ, Saydah S, et al. Prevalence of type 1 and type 2 diabetes among children and adolescents from 2001 to 2009. *JAMA* 2014;311:1778–86.
- Lawrence JM, Imperatore G, Dabelea D, et al. Trends in incidence of type 1 diabetes among non-Hispanic white youth in the U.S., 2002–2009. *Diabetes* 2014;63:3938–45.
- Benoit SR, Hora I, Albright AL, et al. New directions in incidence and prevalence of diagnosed diabetes in the USA. *BMJ Open Diab Res Care* 2019;7:e000657.
- Geiss LS, Wang J, Cheng YJ, et al. Prevalence and incidence trends for diagnosed diabetes among adults aged 20 to 79 years, United States, 1980–2012. *JAMA* 2014;312:1218–26.
- Bryden KS, Dunger DB, Mayou RA, et al. Poor prognosis of young adults with type 1 diabetes: a longitudinal study. *Diabetes Care* 2003;26:1052–7.
- Hillier TA, Pedula KL. Complications in young adults with early-onset type 2 diabetes: losing the relative protection of youth. *Diabetes Care* 2003;26:2999–3005.
- Jaiswal M, Divers J, Dabelea D, et al. Prevalence of and risk factors for diabetic peripheral neuropathy in youth with type 1 and type 2 diabetes: SEARCH for diabetes in youth study. *Diabetes Care* 2017;40:1226–32.
- Dabelea D, Stafford JM, Mayer-Davis EJ, et al. Association of type 1 diabetes vs type 2 diabetes diagnosed during childhood and adolescence with complications during teenage years and young adulthood. *JAMA* 2017;317:825–35.
- Czajka JL, Beyler A. *Background Paper: Declining Response Rates in Federal Surveys: Trends and Implications*. Washington, DC: Mathematica Policy Research, 2016.
- Saydah S, Imperatore G. Emerging approaches in surveillance of type 1 diabetes. *Curr Diab Rep* 2018;18:61.
- Dabelea D, Sauder KA, Jensen ET, et al. Twenty years of pediatric diabetes surveillance: what do we know and why it matters. *Ann N Y Acad Sci* 2021;1495:99–120.
- Lawrence JM, Slezak JM, Quesenberry C, et al. Incidence and predictors of type 1 diabetes among younger adults aged 20–45 years: the diabetes in young adults (Diya) study. *Diabetes Res Clin Pract* 2021;171:S0168–8227(20)30881–0.
- Wells BJ, Lenoir KM, Wagenknecht LE, et al. Detection of diabetes status and type in youth using electronic health records: the SEARCH for diabetes in youth study. *Diabetes Care* 2020;43:2418–25.
- Nichols GA, Desai J, Elston Lafata J, et al. 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.
- Miller DR, Safford MM, Pogach LM. Who has diabetes? best estimates of diabetes prevalence in the Department of veterans affairs based on computerized patient data. *Diabetes Care* 2004;27 Suppl 2:B10–21.
- Klompas M, Eggleston E, McVetta J, et al. Automated detection and classification of type 1 versus type 2 diabetes using electronic health record data. *Diabetes Care* 2013;36:914–21.
- Centers for Disease Control and Prevention (U.S.): Guidelines Working Group. Updated guidelines for evaluation public health surveillance systems; recommendations from the guidelines working group. *Morbidity and Mortality Weekly Report* 2001;50:1–35.
- Chi GC, Li X, Tartof SY, et al. Validity of ICD-10-CM codes for determination of diabetes type for persons with youth-onset type 1 and type 2 diabetes. *BMJ Open Diab Res Care* 2019;7:e000547.
- Chu H, Cole SR. Sample size calculation using exact methods in diagnostic test studies. *J Clin Epidemiol* 2007;60:1201–2.
- Li J, Fine J. On sample size for sensitivity and specificity in prospective diagnostic accuracy studies. *Stat Med* 2004;23:2537–50.

- 26 Steinberg DM, Fine J, Chappell R. Sample size for positive and negative predictive value in diagnostic research using case-control designs. *Biostatistics* 2009;10:94–105.
- 27 Villarroel M, Blackwell D, Jen A. Tables of Summary Health Statistics for U.S. Adults: 2018 National Health Interview Survey: National Center for Health Statistics. 2019.
- 28 Valliant R. Comparing alternatives for estimation from Nonprobability samples. *J Surv Stat Methodol* 2020;8:231–63.
- 29 Elliott MR, Valliant R. Inference for Nonprobability samples. *Statist Sci* 2017;32:16.
- 30 Zhang L-C. On valid descriptive inference from non-probability sample. *Statistical Theory and Related Fields* 2019;3:103–13.
- 31 Wang W, Rothschild D, Goel S, et al. Forecasting elections with non-representative polls. *Int J Forecast* 2015;31:980–91.
- 32 Malec D, Sedransk J, Moriarity CL, et al. Small area inference for binary variables in the national health interview survey. *Journal of the American Statistical Association* 1997;92:815–26.
- 33 Barker LE, Thompson TJ, Kirtland KA, et al. Bayesian small area estimates of diabetes incidence by United States County, 2009. *J Data Sci* 2013;11:269–80.
- 34 Zhang JL, Bryant J. Combining multiple imperfect data sources for small area estimation: a Bayesian model of provincial fertility rates in Cambodia. *Statistical Theory and Related Fields* 2019;3:178–85.
- 35 Smith TMF. On the validity of inferences from Non-Random samples. *J Royal Stat Soc Series A (General)* 1983;146:394.
- 36 Dunson DB. Commentary: practical advantages of Bayesian analysis of epidemiologic data. *Am J Epidemiol* 2001;153:1222–6.
- 37 Klompas M, Cocoros NM, Menchaca JT, et al. State and local chronic disease surveillance using electronic health record systems. *Am J Public Health* 2017;107:1406–12.
- 38 Raebel MA, Schroeder EM, Goodrich G, et al. 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. 2016. Available: https://www.sentinelinitiative.org/sites/default/files/Methods/Mini-Sentinel_Methods_Validating-Diabetes-Mellitus_MSDD_Using-SUPREME-DM-DataLink.pdf
- 39 Villarroel MA, Blackwell DL, Jen A. Tables of summary health statistics for U.S. adults: 2018 national health interview survey. secondary tables of summary health statistics for U.S. adults: 2018 national health interview survey. 2018. Available: <https://www.cdc.gov/nchs/nhis/shs/tables.htm>
- 40 Hill-Briggs F, Adler NE, Berkowitz SA, et al. Social determinants of health and diabetes: A scientific review. *Diabetes Care* 2020;44:258–79.
- 41 Dabelea D, DeGroat J, Sorrelman C, et al. Diabetes in Navajo youth: prevalence, incidence, and clinical characteristics: the SEARCH for diabetes in youth study. *Diabetes Care* 2009;32:S141–7.
- 42 Verlato G, Muggeo M. Capture-recapture method in the epidemiology of type 2 diabetes: a contribution from the Verona diabetes study. *Diabetes Care* 2000;23:759–64.
- 43 Dixon BE, Holmgren AJ, Adler-Milstein J, et al. Health information exchange and Interoperability. In: *Clinical Informatics Study Guide: Text and Review*. 2nd ed. Cham: Springer International Publishing, 2022: 203–19.
- 44 Office of the National Coordinator for Health Information Technology. Quick Stats. Secondary Quick Stats, 2022. Available: <https://www.healthit.gov/data/quickstats/adoption-electronic-health-records-hospital-service-type-2019-2021>
- 45 World Health Organization. *Global diffusion of eHealth: Making universal health coverage achievable. Report of the third global survey on eHealth*. Geneva: World Health Organization, 2016.
- 46 Weeks J, Pardee R. Learning to share health care data: A brief Timeline of influential common data models and distributed health data networks in U.S. *EGEMS (Wash DC)* 2019;7:4.
- 47 Forrest CB, McTigue KM, Hernandez AF, et al. Pcornet® 2020: Current state, accomplishments, and future directions. *J Clin Epidemiol* 2021;129:60–7.
- 48 Observational Health Data Sciences and Informatics (OHDSI). OMOP common data model. secondary OMOP common data model. 2022. Available: <https://ohdsi.org/data-standardization/the-common-data-model>
- 49 Forrest CB, Margolis PA, Bailey LC, et al. Pedsnet: a national pediatric learning health system. *J Am Med Inform Assoc* 2014;21:602–6.
- 50 Hogan WR, Shenkman EA, Robinson T, et al. The OneFlorida data trust: a centralized, Translational research data infrastructure of statewide scope. *J Am Med Inform Assoc* 2022;29:686–93.
- 51 Kho AN, Hynes DM, Goel S, et al. Capricorn: Chicago area patient-centered outcomes research network. *J Am Med Inform Assoc* 2014;21:607–11.
- 52 Haynes SC, Rudov L, Nauman E, et al. Engaging Stakeholders to develop a patient-centered research agenda: lessons learned from the research action for health network (Reachnet). *Med Care* 2018;56:S27–32.
- 53 Weill Cornell Medicine. INSIGHT clinical research network. *Secondary INSIGHT Clinical Research Network* 2022. Available: <https://phs.weill.cornell.edu/research-collaboration/research-programs/insight-clinical-research-network>
- 54 Kaushal R, Hripcsak G, Ascheim DD, et al. Changing the research landscape: the New York City clinical data research network. *J Am Med Inform Assoc* 2014;21:587–90.
- 55 DeVoe JE, Gold R, Cottrell E, et al. The ADVANCE network: accelerating data value across a national community health center network. *J Am Med Inform Assoc* 2014;21:591–5.
- 56 Amin W, Tsui FR, Borromeo C, et al. Path network team. path: towards a learning health system in the mid-Atlantic region. *J Am Med Inform Assoc* 2014;21:633–6.