View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Carolina Digital Repository

Performance of a computable phenotype for identification of patients with diabetes within PCORnet: The Patient-Centered Clinical Research Network

Andrew D. Wiese[1] | Christianne L. Roumie[2,3,4] | John B. Buse[5] | Herodes Guzman[5] | Robert Bradford[5] | Emily Zalimeni[5] | Patricia Knoepp[5] | Heather L. Morris[6] | William T. Donahoo[7] | Nada Fanous[7] | Britany F. Epstein[7] | Bonnie L. Katalenich[8] | Sujata G. Ayala[9] | Megan M. Cook[9] | Katherine J. Worley[10] | Katherine N. Bachmann[11,12,13] | Carlos G. Grijalva[1,4] | Russell L. Rothman[1,2,3] | Rosette J. Chakkalakal[2]

[1] Department of Health Policy, Vanderbilt University Medical Center, Nashville, TN, USA

[2] Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA

[3] Department of Pediatrics, Vanderbilt University Medical Center, Nashville, TN, USA

[4] Veterans Health Administration—Tennessee Valley Healthcare System, Geriatric Research Education Clinical Center (GRECC), Nashville, TN, USA

[5] Department of Medicine, University of North Carolina, Chapel Hill, NC, USA

[6] Department of Health Outcomes and Biomedical Informatics, University of Florida, Gainesville, FL, USA

[7] Department of Medicine, University of Florida, Gainesville, FL, USA

[8] LA CaTS Clinical Translational Unit, Tulane University School of Medicine, Tulane, LA, USA

[9] Institute for Medicine and Public Health, Vanderbilt University Medical Center, Nashville, TN, USA

[10] Vanderbilt Institute for Clinical and Translational Research, Vanderbilt University Medical Center, Nashville, TN, USA

[11] Veterans Health Administration—Tennessee Valley Healthcare System, CSR&D, Nashville, TN, USA

[12] Vanderbilt Translational and Clinical Cardiovascular Research Center, Vanderbilt University Medical Center, Nashville, TN, USA

[13] Division of Diabetes, Endocrinology, and Metabolism, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA

Abstract

Purpose: PCORnet, the National Patient-Centered Clinical Research Network, represents an innovative system for the conduct of observational and pragmatic studies. We describe the identification and validation of a retrospective cohort of patients with type 2 diabetes (T2DM) from four PCORnet sites.

Methods: We adapted existing computable phenotypes (CP) for the identification of patients with T2DM and evaluated their performance across four PCORnet sites (2012-2016). Patients entered the cohort on the earliest date they met one of three CP categories: (CP1) coded T2DM diagnosis (ICD-9/ICD-10) and an antidiabetic prescription, (CP2) diagnosis and glycosylated hemoglobin (HbA1c) ≥6.5%, or (CP3) an antidiabetic prescription and HbA1c ≥6.5%. We required evidence of health care utilization in each of the 2 prior years for each patient, as we also developed an incident T2DM CP to identify the subset of patients without documentation of T2DM in the 365 days before t_0. Among a systematic sample of patients, we calculated the positive predictive value (PPV) for the T2DM CP and incident-T2DM CP using electronic health record (EHR) review as reference.

Results: The CP identified 50 657 patients with T2DM. The PPV of patients randomly selected for validation was 96.2% (n = 1572; CI:95.1-97.0) and was consistently high across sites. The PPV for the incident-T2DM CP was 5.8% (CI:4.5-7.5).

Conclusions: The T2DM CP accurately and efficiently identified patients with T2DM across multiple sites that participate in PCORnet, although the incident T2DM CP requires further study. PCORnet is a valuable data source for future epidemiological and comparative effectiveness research among patients with T2DM.

Funding and prior posting statement: This manuscript has not been posted or presented elsewhere and is was funded through a Patient-Centered Outcomes Research Institute (PCORI) Award (CDRN-1306-04869).

Correspondence
A. D. Wiese, Department of Health Policy, Vanderbilt University Medical Center, Suite 2600, Village at Vanderbilt, 1500 21 Avenue South, Nashville, TN 37212, USA.
Email: andrew.d.wiese.1@vumc.org

Funding information
National Institutes of Health, Grant/Award Number: UL1TR002489; US Department of Veterans Affairs Clinical Sciences Research and Development (CSR&D) Program, Grant/Award Number: IK2 CX001678; Patient-Centered Outcomes Research Institute, Grant/Award Number: CDRN-1306-04869; PhRMA Foundation

KEYWORDS

distributed research network, electronic health records, PCORI, pharmacoepidemiology, type 2 diabetes

1 | INTRODUCTION

In 2014, the Patient-Centered Outcomes Research Institute (PCORI) established a national, distributed research network of interconnected health care data systems integrated under a standardized, common data model (CDM) known as PCORnet: The National Patient-Centered Clinical Research Network.[1,2] Similar to the Sentinel CDM, the PCORnet CDM facilitates the rapid and efficient conduct of research across approximately 80 sites while allowing sites to maintain control over their data, thereby reducing patient privacy and site autonomy concerns.[3] PCORnet represents a potentially transformative data resource, because it remains one of the few that allows for the inclusion of data derived from the electronic health record (EHR), including laboratory and clinical data. Nevertheless, few studies have characterized the use and validity of PCORnet data to conduct observational studies.

Type 2 diabetes mellitus (T2DM) is a common chronic condition that represents a major public health concern in the United States and abroad.[4-6] In addition to significantly increasing the risk of cardiovascular disease, renal disease, and mortality, T2DM contributes to rising health care costs.[4] There are multiple medications available for management of T2DM and comparative effectiveness research to determine the most effective and safe T2DM treatment regimens for specific groups of patients is a major research priority.[7] Large well-designed observational studies can help identify the benefits and harms of specific treatments among patients with T2DM, particularly among those who may be underrepresented in clinical trials (ie, the elderly and those with certain comorbidities such as heart and renal disease).[7-9] The identification of patients with T2DM from EHR and administrative databases has been conducted extensively in the prior literature.[10-23]

Most previous studies have identified and studied patients with T2DM using EHR and administrative databases within well-defined but single data systems. However, less is known about the implementation of those strategies across diverse systems and at a large scale.[10-23] Therefore, T2DM represented an optimal condition for us to construct, characterize and validate a cohort of patients with T2DM across four diverse PCORnet sites to determine the utility of using these integrated data to conduct epidemiological and comparative effectiveness research.

2 | METHODS

2.1 | Data sources

PCORnet is a distributed research network of 13 clinical data research networks (CDRNs) and 20 patient-powered research networks (PPRNs). Each CDRN is composed of one or more sites with an integrated health care data system that captures the information on the patient population receiving care within the system (representing approximately 80 individual sites). Each site is responsible for individual data standardization according to the PCORnet CDM, governance, security, and patient privacy policies. PPRNs represent individual networks of patients with a shared clinical condition with an established agreement to collect and share data relevant to the condition of interest.[1,2] The PCORnet CDM and data harmonization strategies have been previously reported.[1,2,24] In brief, individual PCORnet sites construct a standardized research dataset using the same CDM, allowing the same query program to be distributed and applied to each dataset across the network to attain a larger number of patients and facilitate multicenter collaborations. The PCORnet CDM, adapted from the Mini-Sentinel CDM, outlines the exact specifications for data organization and representation in the analytical research dataset created at each site from their available EHR information.[2,3,24]

The CDM includes data from the EHR, clinical, billing, pharmacy prescription, and laboratory information. Data types can include inpatient and outpatient encounters, medication data, laboratory data, and vital signs. Data may be structured (ICD9 CM/ICD10 codes) or semi-structured (laboratory tests and results).[25] The CDM uses a unique study patient identifier to allow integration of different data types for the generation of analytical datasets while maintaining all personal identifiers and dates protected within the secured CDM data structure.

2.2 | Constructing a retrospective cohort of patients with T2DM

Initial development and refinement of the pilot computable phenotype (CP) occurred at Vanderbilt University Health System (VUHS) (additional details in Supplement). We used the final iteration of the CP from VUHS to identify patients with T2DM across four diverse PCORnet sites in the southern United States using data from 1 January 2012 through 31 December 2016. The four sites included VUHS, the University of North Carolina (UNC), the OneFlorida CDRN, and the Tulane CDRN. The VUHS and UNC sites represent the Mid-South CDRN, and each include a large academic medical center with affiliated community clinics in middle Tennessee and North Carolina, respectively. The OneFlorida CDRN includes three large university systems and nine unique clinical systems providing care to 9.7 million patients. The Tulane CDRN represents a subset of the REACHnet CDRN that encompasses over 4 million patients across multiple academic medical centers and health systems in Southeastern Louisiana. The T2DM CP consisted of three CP categories and were based on previously validated T2DM CP definitions to identify patients with T2DM. These CPs were pilot tested at VUHS using EHR review of a random sample of 60 charts by two physicians using an iterative process to achieve consensus on the presence of T2DM.

We identified adult patients at each site on the earliest date (t_0) they met one of the three CP categories for T2DM: (1) A coded inpatient or outpatient T2DM diagnosis (ICD9/ICD10) [eTable 1] and an antidiabetic medication prescription (eTable 2) within the 90 days following the diagnosis date (CP1); (2) a coded T2DM diagnosis and an outpatient glycolated hemoglobin (HbA1C) value ≥6.5% within 90 days before or after the diagnosis date (CP2) (laboratory values queried using Logical Observation Identifiers, Names and Codes (LOINC) codes (eTable 3); or (3) any antidiabetic medication prescription within 90 days before or after an outpatient HbA1C value ≥6.5% (CP3). For each CP category, t_0 for each patient was the earliest date of the two criteria. If patients met more than one CP on t_0, we classified individuals hierarchically as CP1, CP2, and CP3, respectively. Eligible patients were required to have ≥1 health care encounter in each of the 2 years before t_0 to assure active use of their respective health care system. Finally, we excluded patients with coded diagnoses of gestational diabetes, prediabetes, and type 1 diabetes or evidence of a positive beta human chorionic gonadotropin test as a marker for pregnancy during the 90 days before or after t_0. Patients were identified with T2DM on the earliest of all eligible dates during the study period, such that the identification of an exclusion criterion within 90 days of an eligible t_0 might exclude patient entry on that date but did not preclude the patient from entering the cohort on a subsequent date when all eligibility criteria were met. In addition, we classified identified patients as having incident or new onset diabetes if they fulfilled the incident T2DM CP definition that required no evidence of T2DM (ie, diagnosis, medication prescribed, or elevated HbA1c) in the 365 days prior to t_0 within each CDRN. The IRB for each participating site approved the study protocol.

2.3 | Chart review process

To determine the accuracy of the T2DM CP to identify patients with T2DM within PCORnet sites, we conducted structured chart reviews in a systematic sample of patients identified using the T2DM CP. We verified diagnoses, medication use, and HbA1c values identified from the query of the site-specific CDM against the gold standard of information in the EHR. To attain a sample with CP distribution proportional to that observed at the site where the CPs were developed and pilot tested, each site randomly selected 400 patients identified in the following proportions, 71.5% CP1; 27% CP2; and 1.5% CP3. At least two trained study personnel conducted the EHR review at each site after completing a mandatory training session at each site (details included in Supplement).

KEY POINTS

- PCORnet (the National Patient-Centered Clinical Research Network) is a distributed research network composed of nearly 80 individual health care sites integrated under a standardized, common data model.
- Among 1572 patients randomly sampled from a cohort of 50 657 patients identified with type 2 diabetes from four PCORnet sites, the positive predictive value for the presence of type 2 diabetes was 96.2% using manual electronic health record review as the reference.
- PCORnet represents an innovative data source for future epidemiological and comparative effectiveness research among patients with T2DM.

2.3.1 | Chart reviews

After the training process, for each identified patient in the sample, we reviewed the EHR for each patient within 180 days (90 days before and after) of the index date (t_0). The review of each individual record in the EHR was only conducted by a single reviewer at each site, although each reviewer could flag a record for review and data abstraction by a designated physician expert at each site. Reviewers documented evidence within the ascertainment period of T2DM diagnoses (as well as type 1 diabetes, gestational diabetes, prediabetes, and diabetes controlled with lifestyle modification) from patient summaries, admission/discharge summaries, and clinic/outpatient visit summaries. Evidence of antidiabetic medication prescriptions was identified from prescription information in the EHR, and evidence of elevated HbA1c values was recorded only from site-specific HbA1c laboratory results in the EHR. Evidence of each criteria was abstracted using a standardized and customized REDCap electronic data capture instrument.[26] Additionally, the clinical reviewers documented any diagnosis of T2DM in all documentation available in the EHR during the time period before t_0. We assigned study specific identifiers to each patient at the individual sites so that no identifiable patient information was recorded in the data capture instrument. The information in the patients' EHR was considered the gold standard for all elements assessed. Thus, if there was no information in the EHR regarding an element of interest, we considered it evidence the patient did not have the condition, medication prescription, or laboratory value.

2.4 | Statistical analysis

We calculated the positive predictive value (PPV) of the T2DM CP for identifying adult patients with T2DM in the PCORnet data. Among all patients identified by the CP, confirmed cases were those patients with evidence in the medical record within 90 days of t_0 of a documented T2DM diagnosis (including diabetes controlled with lifestyle modification) and documented antidiabetic medication prescription (CP1), documented T2DM diagnosis, and documented elevated HbA1c (CP2) or documented antidiabetic medication prescription and elevated HbA1c (CP3), and without documented evidence in the medical record of type 1 diabetes, gestational diabetes, pregnancy, or prediabetes without evidence of type 2 diabetes within 90 days of t_0. We calculated 95% confidence intervals (CI) for the PPV using Wilson's formula.[27] Secondary analyses assessed the PPV for each CP category and each site. As a sensitivity analysis, we assessed the PPV for each CP and each site based on the ICD coding era in which each patient was identified (all four sites transitioned from ICD9 to ICD10 in October 2015, although Tulane allowed ICD9 codes through May 2016). We further assessed the PPV for the incident or new onset T2DM CP. All analyses were performed in Stata-IC, version 15.1 (College Station TX), and manuscript preparation was completed in part using StatTag (Northwestern University).[28]

3 | RESULTS

3.1 | Cohort assembly and sample characteristics

Using the T2DM CP, we initially identified 205 004 patients ≥18 years of age with possible T2DM among 4 696 145 patients across the four sites. The final retrospective cohort from which the validation sample was identified comprised 51 226 patients with T2DM (Figure 1). After the validation sample was identified and data collected in January 2018, a regularly scheduled process to refresh the PCORnet data structure incorporated additional health care encounter information at each site, resulting in minor changes to the underlying T2DM cohorts at each site from which descriptive statistics were derived (*n* = 50 657) (Table 1). Patients with T2DM were primarily ≥60 years

TABLE 1 Characteristics of patients in the retrospective cohort identified with T2DM, PCORnet, 2012-2016; April 2018 PCORnet Data Query

	VUHS[a]	OneFlorida[b]	UNC[c]	Tulane
N	**13 782**	**14 881**	**19 731**	**2263**
Age in years (%)				
18-29	79 (0.6)	103 (0.7)	88 (0.4)	20 (0.9)
30-39	253 (1.8)	425 (2.9)	392 (2.0)	67 (3.0)
40-49	1065 (7.7)	1357 (9.1)	1620 (8.2)	202 (8.9)
50-59	2743 (19.9)	3261 (21.9)	3822 (19.4)	425 (18.8)
60-69	4272 (31.0)	4762 (32.0)	5728 (29.0)	738 (32.6)
70-79	3586 (26.0)	3388 (22.8)	5249 (26.6)	548 (24.2)
80-89	1464 (10.6)	1325 (8.9)	2342 (11.9)	214 (9.5)
90+	319 (2.3)	243 (1.6)	473 (2.4)	49 (2.2)
Unavailable[a]	T[d]	17 (0.1)	17 (0.1)	0 (0.0)
Female (%)	6486 (47.1)	8468 (56.9)	10 214 (51.8)	1224 (54.1)
Race (%)				
Black	2573 (18.7)	5439 (36.5)	5454 (27.6)	1515 (66.9)
Missing	101 (0.7)	96 (0.6)	620 (3.1)	52 (2.3)
Other	351 (2.5)	1114 (7.5)	1201 (6.1)	78 (3.4)
White	10 756 (78.0)	8232 (55.3)	12 446 (63.1)	618 (27.3)
CP type (%)				
CP1	9870 (71.6)	12 302 (82.7)	12 803 (64.9)	1470 (65.0)
CP2	3698 (26.8)	2460 (16.5)	6900 (35.0)	783 (34.6)
CP3	214 (1.6)	119 (0.8)	28 (0.1)	10 (0.4)

[a] Vanderbilt University Health System.

[b] OneFlorida CDRN.

[c] University of North Carolina.

[d] PCORnet low threshold masking.

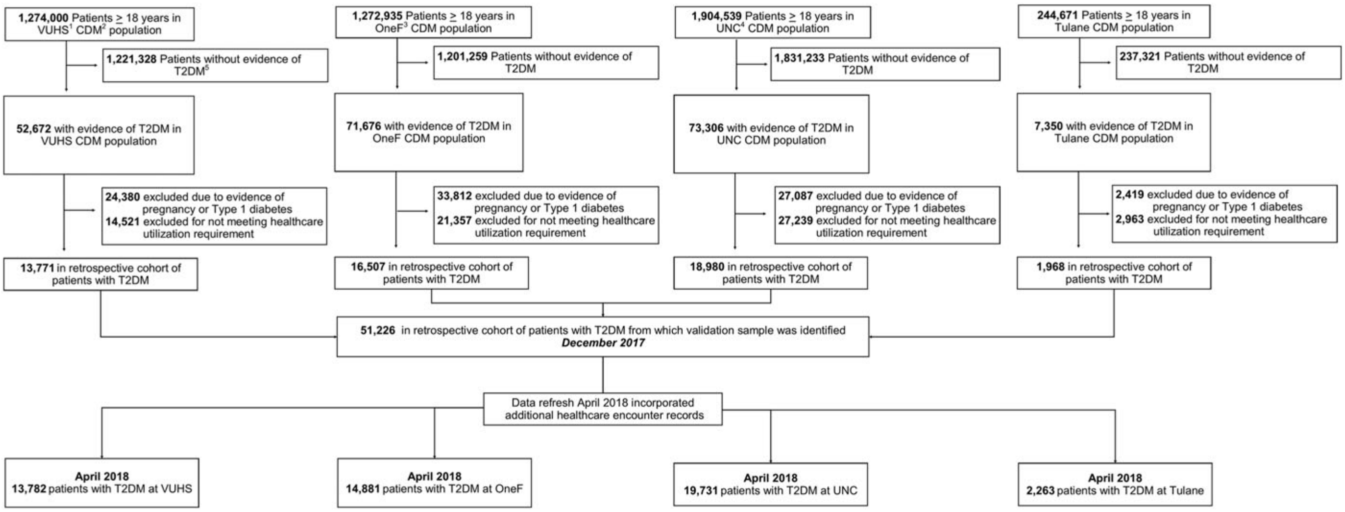
1,274,000 Patients ≥ 18 years in VUHS¹ CDM² population
1,221,328 Patients without evidence of T2DM⁵
52,672 with evidence of T2DM in VUHS CDM population
24,380 excluded due to evidence of pregnancy or Type 1 diabetes
14,521 excluded for not meeting healthcare utilization requirement
13,771 in retrospective cohort of patients with T2DM
1,272,935 Patients ≥ 18 years in OneF³ CDM population
1,201,259 Patients without evidence of T2DM
71,676 with evidence of T2DM in OneF CDM population
33,812 excluded due to evidence of pregnancy or Type 1 diabetes
21,357 excluded for not meeting healthcare utilization requirement
16,507 in retrospective cohort of patients with T2DM
1,904,539 Patients ≥ 18 years in UNC⁴ CDM population
1,831,233 Patients without evidence of T2DM
73,306 with evidence of T2DM in UNC CDM population
27,087 excluded due to evidence of pregnancy or Type 1 diabetes
27,239 excluded for not meeting healthcare utilization requirement
18,980 in retrospective cohort of patients with T2DM
244,671 Patients ≥ 18 years in Tulane CDM population
237,321 Patients without evidence of T2DM
7,350 with evidence of T2DM in Tulane CDM population
2,419 excluded due to evidence of pregnancy or Type 1 diabetes
2,963 excluded for not meeting healthcare utilization requirement
1,968 in retrospective cohort of patients with T2DM
51,226 in retrospective cohort of patients with T2DM from which validation sample was identified
December 2017
Data refresh April 2018 incorporated additional healthcare encounter records
April 2018
13,782 patients with T2DM at VUHS
April 2018
14,881 patients with T2DM at OneF
April 2018
19,731 patients with T2DM at UNC
April 2018
2,263 patients with T2DM at Tulane

FIGURE 1 Patients with T2DM identified from four PCORnet sites (2012-2016). [1]VUHS: Vanderbilt University Health System; [2]CDM: Common data model; [3]OneF: One Florida; [4]UNC: University of North Carolina; [5]T2DM: Type 2 diabetes mellitus

of age (68.5%), female (52.1%), and White (63.3%), although population characteristics varied by site (Table 1).

3.2 | Positive predictive value of CP

Each site systematically selected a random sample of 400 patients identified using the T2DM CP for review. We completed EHR abstraction for 98.3% (*n* = 1572) of the selected sample (>96% for each site) as 26 records across all sites were unavailable for review and two records were incomplete in data abstraction.

Of the 1572 patients identified and reviewed, 1512 [PPV: 96.2% (CI: 95.1-97.0)] had confirmed T2DM (Table 2). Interestingly, a small percentage of patients identified using the T2DM CP had evidence of type 1 diabetes, gestational diabetes, or prediabetes (0.5%), even though the T2DM CP excluded individuals with evidence of these conditions. The PPV was highest for CP1 [PPV: 96.4% (CI: 95.2-97.4)], high for CP2 [PPV: 96.1% (CI: 93.8-97.5)], and moderate for CP3 [PPV: 81.3% (CI: 57.0-93.4)], albeit there were a small number of patients identified using CP3. The overall PPV was >95.0% across all four sites (Table 2).

Of patients identified with T2DM using CP1, 63.1% also had an HbA1c value ≥6.5% within 180 days of t_0, which was not a required element of CP1. Similarly, 49.8% of patients identified using CP2 had evidence of an antidiabetic medication prescription and 75.0% of patients identified using CP3 had evidence of a T2DM diagnosis within 180 days of t_0 (eTable 4). Among the 60 patients (3.8% of total) that were not confirmed by EHR review, most patients identified using CP1 (*n* = 40) lacked evidence of an antidiabetic medication prescription only (85.0%), while those identified using CP2 primarily lacked evidence of an elevated HbA1c (47.1%). Only three patients identified using CP3 did not have confirmed T2DM, either due to the lack of an antidiabetic medication prescription (*n* = 2) or evidence of type 1 diabetes (*n* = 1).

Of the total 1572 charts reviewed, there were 916 (58.2%) patients identified as incident or new-onset T2DM. Among them, 53 [PPV: 5.8% (CI: 4.5-7.5)] were true new onset or incident T2DM after EHR review (Table 3). The incident T2DM CP performed poorly for all T2DM CP categories (PPV range: 3.9-18.2) and across all four sites (PPV range: 3.0-9.4) (Table 3). In the sensitivity analysis by ICD coding era, the PPV was similar among patients identified in the ICD9 era versus the ICD10 era [PPV: 96.5 (CI: 95.2-97.5) and PPV: 95.6 (CI: 93.4-97.1), respectively] [eTable 5]. The results were similar across ICD coding eras for CP1 and CP2, as well as for each site [eTable 5].

TABLE 3 Random sample of patients with possible incident T2DM by computable phenotype (CP) and site among selected health care systems in PCORnet, 2012-2016

	Total Reviewed	Total Confirmed	PPV[a] (95% CI[b])	
Overall	916	53	5.8	(4.5, 7.5)
By case type				
CP1—diabetes diagnosis and medication use	573	38	6.6	(4.9, 9.0)
CP2—diabetes diagnosis and HbA1c ≥6.5%	332	13	3.9	(2.3, 6.6)
CP3—HbA1c ≥6.5% and medication use	11	2	18.2	(5.1, 47.7)
By site				
Vanderbilt University Health System	199	6	3.0	(1.4, 6.4)
OneFlorida	213	20	9.4	(6.2, 14.1)
University of North Carolina	307	20	6.5	(4.3, 9.8)
Tulane	197	7	3.6	(1.7, 7.2)

[a] PPV: positive predictive value.

[b] Confidence interval using Wilson's formula.

TABLE 2 Random sample of patients with possible T2DM by computable phenotype (CP) and site at selected health care systems in PCORnet, 2012-2016

	Total Reviewed	Patients with Confirmed T2DM (*n* = 1512)			Patients without Confirmed T2DM (*n* = 60)	
		N	PPV[a] (95% CI[b])		Did not meet ≥2 diabetes criteria[c]	Type 1 diabetes, gestational diabetes, or prediabetes
Overall	1572	1512	96.2	(95.1, 97.0)	52	8
By computable phenotype						
CP1—diabetes diagnosis and medication use	1124	1084	96.4	(95.2, 97.4)	37	3
CP2—diabetes diagnosis and HbA1c ≥6.5%	432	415	96.1	(93.8, 97.5)	13	4
CP3—HbA1c ≥6.5% and medication use	16	13	81.3	(57.0, 93.4)	2	1
By site						
Vanderbilt University Health System	398	379	95.2	(92.7, 96.9)	18	1
OneFlorida	391	382	97.7	(95.7, 98.8)	8	1
University of North Carolina	396	380	96.0	(93.5, 97.5)	15	1
Tulane	387	371	95.9	(93.4, 97.4)	11	5

[a] PPV: positive predictive value.

[b] Confidence interval using Wilson's formula.

[c] Criteria included evidence of a diabetes diagnosis, diabetes medication use, or elevated HbA1c ≥6.5%.

4 | DISCUSSION

PCORnet is a large integrated data resource with potential for the conduct of clinical trial, epidemiological and comparative effectiveness research, yet the performance of systematic strategies to identify patients with common conditions such as T2DM in PCORnet has not been previously characterized. We identified a large number of patients with T2DM CP (n = 50 657) across four PCORnet sites (representing only around 5% of PCORnet sites). The vast majority of identified cohort members were confirmed to have T2DM, with an overall PPV of 96.2%. The PPVs were consistently high for all CP categories that were defined by combinations of T2DM diagnoses, antidiabetic medication prescriptions and elevated HbA1c laboratory values and across all four sites representing different EHR systems, patient populations, and geographical regions with a high prevalence of T2DM. In contrast, the incident T2DM CP incorrectly identified many prevalent T2DM cases as incident or new onset, an important consideration for future observational studies using PCORnet.

Strategies for the identification of patients with T2DM from a single hospital or within a single network-based EHR system have been reported and applied extensively, although prior studies used different inclusion and exclusion criteria to identify patients with T2DM.[10-23] Of those that applied specific algorithms requiring multiple sources of evidence to identify patients with T2DM (ie, >1 of either a coded diagnosis, prescribed antidiabetic medications or elevated laboratory values; in addition to evidence of baseline enrollment and health care utilization), the study populations ranged in size from 10 000 to 500 000 patients.[10-23] Distinct from our study, several of these prior studies centered only on a single health care system[13-16] or used coded diagnoses, antidiabetic medication use, and laboratory values to identify patients with T2DM from de-identified, commercially available databases, which could not be validated against patients' EHR information.[17-21] In contrast, we identified 50 657 patients with T2DM in only four of about 80 PCORnet sites using CP definitions of coded diagnoses, prescribed antidiabetic medications, and laboratory values similar to the Mini-Sentinel recommendations for identifying patients with T2DM.[22,23] Our study demonstrates the successful implementation of those strategies in PCORnet and highlights the scalability of the process to enable the identification of patients with T2DM for the conduct of retrospective and prospective cohort studies as well as pragmatic clinical trials.

The PPV of our T2DM CP was similar or higher than other prior studies that sought to identify patients with T2DM. In a prior study of 18 131 patients with clinically confirmed T2DM in two California counties, using an HbA1c ≥6.5% to identify patients with T2DM at any time during the 5-year study period had a high PPV (83.2%).[29] In another prior study among 101 278 children <20 years of age from two health care systems in North and South Carolina, multiple algorithms requiring evidence of ≥2 diagnosis codes (or the ratio of T2DM codes to type 1 diabetes codes) had low to moderate PPV (highest 75.5%) compared with manual chart review.[30] In a separate study conducted at a single ambulatory practice, a two-step definition that first identified all patients with evidence of diabetes based on diagnosis codes, laboratory values, and antidiabetic medications and further differentiated patients with T2DM from type 1 diabetes based on oral hypoglycemic use had a PPV around 90.0% for identifying patients with T2DM relative to manual EHR review.[31] Furthermore, a comprehensive study involving the application of 8 EHR-T2DM phenotypes to a large sample of 173 503 patients reported a varied sensitivity across definitions, but a consistently high specificity for each definition (95%-99%).[12] In comparison, our approach had a PPV of 96.2% among a large sample of cases in PCORnet (likely approximating a high specificity for identifying T2DM cases), providing evidence that patients with T2DM can be successfully and accurately identified in PCORnet.

We purposefully used a T2DM CP that required baseline health care utilization among identified members of our retrospective cohort to attempt to identify patients with incident T2DM. Although identifying the true date of T2DM disease onset is challenging even in purely clinical settings, it would be advantageous to be able to do so in the design of comparative effectiveness studies and to allow the characterization of the evolution of disease and management over time.[32] Nevertheless, upon EHR review, many of the patients identified using the incident T2DM CP were prevalent T2DM patients receiving diabetes treatment outside the health care system who had transitioned to receive their diabetes care within the network. In addition, the incident T2DM CP incorrectly identified patients with prevalent T2DM receiving regular non-diabetes care, imaging, or surgical procedures at each site as having incident or new-onset T2DM. Consequently, the performance of our CP for identification of incident T2DM was very poor. As noted in prior research, identification of patients with T2DM within large health care systems does not always indicate that such systems are the primary or initial site of diabetes management.[33] As outlined in a recent study, requiring at least one encounter coded as primary care or endocrinological/diabetes in each of the prior 2 years might better differentiate incident or new onset T2DM from patients with prevalent T2DM.[13,34] The identification of patients with incident T2DM in EHR data remains challenging, and so additional studies are needed to determine the performance of incident T2DM definitions in PCORnet.

An important limitation of our study was the inability to determine the sensitivity and specificity of the T2DM CP in PCORnet. To calculate sensitivity and specificity, we would have needed to review records for patients that were not identified using the T2DM CP. Although we did not calculate specificity directly, the estimated prevalence of diagnosed diabetes in the adult population nationally and in PCORnet is approximately 10%, and therefore our high PPV likely approximates a high specificity.[35] The high PPV of the T2DM CP suggests these definitions can be used to identify patients with T2DM for future observational or pragmatic clinical studies using PCORnet data.[35-37]

Similar to other observational studies which utilize EHR or claims data, missing or incomplete diagnosis, laboratory, or medication prescribing information could have limited our ability to identify or confirm patients with T2DM in the study (although the impact on the observed PPV would be minimal). Another important consideration for the use of PCORnet to study patients with T2DM is that we did not assess whether each site was the sole source of treatment and care, which could have implications for the conduct of longitudinal studies and for the demonstration of history of T2DM. Although most observational studies involving patients with T2DM use pharmacy information on filled medication prescriptions, we instead used prescribing information from the EHR to identify and characterize medication use. Future work is needed to determine the accuracy of the prescribing information in relation to actual medication dispensing and use in PCORnet.

In conclusion, our study demonstrates the successful implementation of a strategy to accurately identify a large cohort of patients with T2DM within PCORnet, although additional study is needed to identify patients with new-onset or incident T2DM. PCORnet represents a novel data platform for the identification of patients with chronic disease and the efficient conduct of clinical and epidemiological research.

ETHICS STATEMENT

The study protocol was approved by the IRB of each participating site.

ACKNOWLEDGEMENTS

The statements presented in this manuscript are solely the responsibility of the authors and do not necessarily represent the views of the PCORI, its Board of Governors, or Methodology Committee or other participants in PCORnet.

CONFLICT OF INTEREST

C.L.R. reports support from PCORI during the conduct of the study. J.B.B. has performed contracted consulting work with fees paid to his institution for Adocia, AstraZeneca, Dexcom, Elcelyx Therapeutics, Eli Lilly, Intarcia Therapeutics, Lexicon, MannKind, Metavention, NovaTarg, Novo Nordisk, Sanofi, Senseonics, and vTv Therapeutics; he has received grant support from AstraZeneca, Boehringer Ingelheim, Johnson & Johnson, Lexicon, Novo Nordisk, Sanofi, Theracos, and vTv Therapeutics; he is a consultant to Neurimmune AG; he holds stock options in Mellitus Health, PhaseBio, and Stability Health. R.L.R. reports personal fees from edlogics and Abbots Diabetes Care outside the submitted work. C.G.G. has received consulting fees from Pfizer, Sanofi-Pasteur, and Merck, and received research support from Sanofi-Pasteur, Campbell Alliance, the Centers for Disease Control and Prevention, National Institutes of Health, The Food and Drug Administration, and the Agency for Health Care Research and Quality. All other authors have no conflicts of interest to disclose.

FUNDING

This manuscript was funded through a Patient-Centered Outcomes Research Institute (PCORI) Award (CDRN-1306-04869) for development of the National Patient-Centered Clinical Research Network, known as PCORnet. A.D.W. is supported by a PhRMA Foundation Postdoctoral Fellowship Award in Health Outcomes. K.N.B. is supported by a Vanderbilt University Medical Center Faculty Research Scholars award and a Career Development Award (#IK2 CX001678) from the US Department of Veterans Affairs Clinical Sciences Research and Development (CSR&D) Program. J.B.B. and P.K. are supported by a grant from the National Institutes of Health (UL1TR002489) and by the PCORnet Diabetes and Obesity Collaborative Research Group.

ORCID

Andrew D. Wiese https://orcid.org/0000-0002-0699-4224

REFERENCES

1. Pletcher MJ. PCORnet's Collaborative Research Groups. *Patient Relat Outcome Meas*. 2018;9:91-95.
2. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching PCORnet, a national patient-centered clinical research network. *J Am Med Inform Assoc*. 2014;21(4):578-582.
3. Ball R, Robb M, Anderson SA, Dal Pan G. The FDA's sentinel initiative—a comprehensive approach to medical product surveillance. *Clin Pharmacol Ther*. 2016;99(3):265-268.
4. Bommer C, Sagalova V, Heesemann E, et al. Global economic burden of diabetes in adults: projections from 2015 to 2030. *Diabetes Care*. 2018;41(5):963-970.
5. Sarwar N, Gao P, Seshasai SR, et al. Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. *Lancet*. 2010;375(9733):2215-2222.
6. Beckman JA, Creager MA, Libby P. Diabetes and atherosclerosis: epidemiology, pathophysiology, and management. *JAMA*. 2002;287(19):2570-2581.
7. Greenfield S. Comparative effectiveness and the future of clinical research in diabetes. *Diabetes Care*. 2013;36(8):2146-2147.
8. Ridda I, Lindley R, MacIntyre RC. The challenges of clinical trials in the exclusion zone: the case of the frail elderly. *Australas J Ageing*. 2008;27(2):61-66.
9. Witham MD, McMurdo ME. How to get older people included in clinical studies. *Drugs Aging*. 2007;24(3):187-196.
10. Wei WQ, Leibson CL, Ransom JE, et al. Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus. *J Am Med Inform Assoc*. 2012;19(2):219-224.
11. Richesson RL, Rusincovitch SA, Wixted D, et al. A comparison of phenotype definitions for diabetes mellitus. *J Am Med Inform Assoc*. 2013;20(e2):e319-e326.
12. Spratt SE, Pereira K, Granger BB, et al. Assessing electronic health record phenotypes against gold-standard diagnostic criteria for diabetes mellitus. *J Am Med Inform Assoc*. 2017;24(e1):e121-e128.
13. Pantalone KM, Misra-Hebert AD, Hobbs TM, et al. Effect of glycemic control on the diabetes complications severity index score and development of complications in people with newly diagnosed type 2 diabetes. *J Diabetes*. 2018;10(3):192-199.
14. Roumie CL, Min JY, D'Agostino McGowan L, et al. Comparative safety of sulfonylurea and metformin monotherapy on the risk of heart failure: a cohort study. *J Am Heart Assoc*. 2017;6(4):e005379.
15. Rodgers LR, Weedon MN, Henley WE, Hattersley AT, Shields BM. Cohort profile for the MASTERMIND study: using the Clinical Practice Research Datalink (CPRD) to investigate stratification of response to treatment in patients with type 2 diabetes. *BMJ Open*. 2017;7(10):e017989.
16. Brennan MB, Hess TM, Bartle B, et al. Diabetic foot ulcer severity predicts mortality among veterans with type 2 diabetes. *J Diabetes Complications*. 2017;31(3):556-561.
17. Boye KS, Botros FT, Haupt A, Woodward B, Lage MJ. Glucagon-like peptide-1 receptor agonist use and renal impairment: a retrospective analysis of an electronic health records database in the U.S. population. *Diabetes Ther*. 2018;9(2):637-650.
18. Blonde L, Meneghini L, Peng XV, et al. Probability of achieving glycemic control with basal insulin in patients with type 2 diabetes in real-world practice in the USA. *Diabetes Ther*. 2018;9(3):1347-1358.
19. Sharma M, Nazareth I, Petersen I. Trends in incidence, prevalence and prescribing in type 2 diabetes mellitus between 2000 and 2013 in primary care: a retrospective cohort study. *BMJ Open*. 2016;6(1):e010210.

20. Berkowitz SA, Krumme AA, Avorn J, et al. Initial choice of oral glucose-lowering medication for diabetes mellitus: a patient-centered comparative effectiveness study. *JAMA Intern Med*. 2014;174(12):1955-1962.

21. Fu AZ, Qiu Y, Davies MJ, Radican L, Engel SS. Treatment intensification in patients with type 2 diabetes who failed metformin monotherapy. *Diabetes Obes Metab*. 2011;13(8):765-769.

22. 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.

23. Raebel M, Schroeder E, 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. In: Mini-Sentinel, ed 2016.

24. Garza M, Del Fiol G, Tenenbaum J, Walden A, Zozus MN. Evaluating common data models for use with a longitudinal community registry. *J Biomed Inform*. 2016;64:333-341.

25. *International Classification of Diseases, Ninth Revision, Clinical Modification*. Washington, DC: Public Health Service, US Dept of Health and Human Services; 1988.

26. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform*. 2009;42(2):377-381.

27. Niesner K, Murff HJ, Griffin MR, et al. Validation of VA administrative data algorithms for identifying cardiovascular disease hospitalization. *Epidemiology*. 2013;24(2):334-335.

28. *StatTag [Computer Program]*. Chicago, Illinois, United States: Galter Health Sciences LIbrary; 2016.

29. Richardson MJ, Van Den Eeden SK, Roberts E, Ferrara A, Paulukonis S, English P. Evaluating the use of electronic health records for type 2 diabetes surveillance in 2 California counties, 2010–2014. *Public Health Rep*. 2017;132(4):463-470.

30. 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.

31. 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.

32. Ray WA. Evaluating medication effects outside of clinical trials: new-user designs. *Am J Epidemiol*. 2003;158(9):915-920.

33. Pantalone KM, Misra-Hebert AD, Hobbs TM, et al. Antidiabetic treatment patterns and specialty care utilization among patients with type 2 diabetes and cardiovascular disease. *Cardiovasc Diabetol*. 2018;17(1):54.

34. Pantalone KM, Hobbs TM, Wells BJ, et al. Changes in characteristics and treatment patterns of patients with newly diagnosed type 2 diabetes in a large United States integrated health system between 2008 and 2013. *Clin Med Insights Endocrinol Diabetes*. 2016;9:23-30.

35. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J Clin Epidemiol*. 2005;58(4):323-337.

36. Brenner H, Gefeller O. Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. *Stat Med*. 1997;16(9):981-991.

37. Brenner H, Gefeller O. Use of the positive predictive value to correct for disease misclassification in epidemiologic studies. *Am J Epidemiol*. 1993;138(11):1007-1015.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.