Journal of the American Medical Informatics Association, 24(e1), 2017, e121–e128
doi: 10.1093/jamia/ocw123
Advance Access Publication Date: 11 September 2016
Research and Applications

Research and Applications

Assessing electronic health record phenotypes against gold-standard diagnostic criteria for diabetes mellitus

Susan E Spratt,[1] Katherine Pereira,[2] Bradi B Granger,[2] Bryan C Batch,[1] Matthew Phelan,[3] Michael Pencina,[3,4] Marie Lynn Miranda,[5] Ebony Boulware,[1] Joseph E Lucas,[6] Charlotte L Nelson,[3] Benjamin Neely,[3] Benjamin A Goldstein,[3,4] Pamela Barth,[7] Rachel L Richesson,[2] Isaretta L Riley,[1] Leonor Corsino,[1] Eugenia R McPeek Hinz,[1] Shelley Rusincovitch,[7] Jennifer Green,[1] Anna Beth Barton,[1] and the DDC Phenotype Group (Carly Kelley, Kristen Hyland, Monica Tang, Amanda Elliott, Ewa Ruel, Alexander Clark, Melanie Mabrey, Kay Lyn Morrissey, Jyothi Rao, Beatrice Hong, Marjorie Pierre-Louis, Katherine Kelly, and Nicole Jelesoff)

[1]Department of Medicine, Duke University School of Medicine, Durham, NC, USA, [2]Duke University School of Nursing, Durham, NC, USA, [3]Duke Clinical Research Institute, Durham, NC, USA, [4]Department of Biostatistics & Bioinformatics, Duke University School of Medicine, Durham, NC, USA, [5]Rice University and Baylor College of Medicine, Houston, TX, USA, [6]Department of Statistical Science, Duke University, Durham, NC, USA and [7]Duke Translational Medicine Institute, Durham, NC, USA

Corresponding Author: Susan E Spratt, Duke Box 3311, Duke University Medical Center, Durham, NC 27710, USA; susan.spratt@duke.edu; Tel: 919-668-1367

Received 23 March 2016; Revised 19 July 2016; Accepted 20 July 2016

ABSTRACT

Objective: We assessed the sensitivity and specificity of 8 electronic health record (EHR)-based phenotypes for diabetes mellitus against gold-standard American Diabetes Association (ADA) diagnostic criteria via chart review by clinical experts.

Materials and Methods: We identified EHR-based diabetes phenotype definitions that were developed for various purposes by a variety of users, including academic medical centers, Medicare, the New York City Health Department, and pharmacy benefit managers. We applied these definitions to a sample of 173 503 patients with records in the Duke Health System Enterprise Data Warehouse and at least 1 visit over a 5-year period (2007–2011). Of these patients, 22 679 (13%) met the criteria of 1 or more of the selected diabetes phenotype definitions. A statistically balanced sample of these patients was selected for chart review by clinical experts to determine the presence or absence of type 2 diabetes in the sample.

Results: The sensitivity (62–94%) and specificity (95–99%) of EHR-based type 2 diabetes phenotypes (compared with the gold standard ADA criteria via chart review) varied depending on the component criteria and timing of observations and measurements.

Discussion and Conclusions: Researchers using EHR-based phenotype definitions should clearly specify the characteristics that comprise the definition, variations of ADA criteria, and how different phenotype definitions and components impact the patient populations retrieved and the intended application. Careful attention to phenotype definitions is critical if the promise of leveraging EHR data to improve individual and population health is to be fulfilled.

Key words: EHR phenotypes, diabetes identification, diabetes registries

INTRODUCTION

Diabetes mellitus affects 29 million Americans. Costs attributable to diabetes are $284 billion annually.[1–3] The American Diabetes Association (ADA) 2016 guidelines give 4 options as the "gold standard" for accurate diagnosis of diabetes: (1) fasting plasma glucose of 126 mg/dL, (2) 2-hour plasma glucose >200 mg/dL during an oral glucose tolerance test, (3) hemoglobin A1c >6.5%, or (4) random plasma glucose over 200 mg/dL.[4] While the ADA has published standard criteria for diagnosis of types 1 and 2 diabetes, the identification of these criteria in electronic health record (EHR) data is often missing, unclear, or unreliable.[5] Historically, self-reported data[1,3,6] have been the norm in assessing population health; more recently, secondary data analysis of EHR data to create disease registries has ballooned, in part due to the passage of the Health Information Technology for Economic and Clinical Health Act.[7] The accurate identification of patients with diabetes using secondary data is challenging; however, use of standardized and reproducible EHR-based phenotype definitions will support research and quality improvement by enabling direct comparison of population characteristics, risk factors, and complications. In addition, development and use of phenotypic standards will allow stakeholders to identify evidence-based interventions and apply them to appropriate patient populations.[8–23] The purpose of this project was to compare 8 different diabetes phenotypes to gain clear insight into the relative components of each definition and to better understand, compare, and design population health projects

Previously, we showed that the prevalence of diabetes in Durham County, North Carolina, varies (from 7 to 13%) depending on the specific EHR-based diabetes phenotype definition implemented.[5] In this analysis, we used the same EHR-based diabetes phenotype definitions (International Classification of Diseases, Ninth Revision, code 250.xx [ICD_250]; expanded ICD-9 codes [CCW]; abnormal A1c, based on New York City Health Department A1c Registry [A1c]; diabetes medication, based on pharmacy benefit manager (PBM) data [Med]; Durham Diabetes Coalition [DDC]; Surveillance, Prevention, and Management of Diabetes Mellitus [SUPREME DM or Sup-DM]; electronic medical records and genomics [Northwest or eMERGE]; abnormal A1c + diabetes medication [A1c_Med]) and assessed their sensitivity and specificity based on comparison to the gold-standard ADA diagnostic criteria implemented by clinical experts via chart review of a statistically based sample from the Duke Health System Enterprise Data Warehouse.

The phenotype definitions we selected for our investigation were mature and used in active population health programs or research studies by various groups, such as academic medical centers, health departments, government agencies, or medication managers (Box 1). While these different phenotypes used the same components (ICD-9 codes, laboratory data, and/or medication use), our previous work showed that the multiple ways in which the components are assembled, included, or excluded in terms of frequency, clinical context, and time can drastically change the performance of one phenotype against another.[5]

As shown in Table 1, 2 of the 8 phenotypes are made up solely of ICD-9 codes. The first (ICD_250) is based on Healthcare Effectiveness Data and Information Set ICD-9 codes 250.xx for types 1 and 2 diabetes. The second (CCW) is an expanded list of codes that includes Healthcare Effectiveness Data and Information Set codes and diagnoses that indicate secondary diabetes and diabetes complications: ICD-9 codes 249.xx, 362, and 357.

The third phenotype (A1c) uses the presence of an abnormal hemoglobin A1c to identify patients with diabetes. A growing number of health systems and local health departments rely on laboratory data to help identify patients with diabetes. For example, the New York City Health Department partnered with all the health systems and clinics in 5 boroughs to mandate reporting of all hemoglobin A1c results. They standardized a process to notify patients of their results, explain A1c goals to patients, and alert both patients and providers when testing is overdue.[18]

We constructed a fourth phenotype (Med) that is based solely on documentation of diabetes-associated medication, as this would be the data that PBMs use to identify patients with diabetes. PBMs do not have access to laboratory data or visit diagnoses; they identify patients based on medications prescribed and dispensed. PBMs create disease registries to give employer groups data on gaps in care (e.g., lack of statin therapy in patients on diabetes medications). PBM data are limited and cannot identify patients who are undiagnosed, untreated, or fail to use benefits to fill prescriptions.

The next 3 phenotypes use inpatient and outpatient diagnosis codes, laboratory test results, and medication prescription data to identify patients with diabetes. The combinations of how the components are assembled changes the nature of the cohort the phenotype identified. These differences are described here as the fifth, sixth, and seventh phenotypes.

The DDC, the fifth phenotype, aims to identify patients in Durham County, North Carolina, with type 2 diabetes. In order to reduce death, disability, and cost in patients, the DDC developed a geographic health information system based on analysis of secondary data, including health, census, demographic, social, and environmental data; the project matches resources to individuals and communities based on risk with diagnosed and undiagnosed type 2 diabetes.[19]

The sixth phenotype, Sup-DM, was designed by a consortium of 11 integrated health systems (the Health Care Systems Research Network [previously the health management organization (HMO) Research Network]), where diabetes prevalence ranges from 4.6 to 12% (average 6.9%) across sites. The Sup-DM phenotype was developed for epidemiological study and public health intervention in types 1 and 2 diabetes patients.[20–22]

The seventh phenotype, electronic medical records and genomics (NW), was developed by investigators at Northwestern University to identify patients with type 2 diabetes for genotype–phenotype correlation studies, and was designed specifically to exclude type 1 diabetes patients, including those who have ever been coded with

Box 1: Purpose and benefits of disease registries

Quality improvement
- Identifying gaps in care (HbA1c not checked)
- Identifying care goals not obtained (HbA1c > 8%, BP > 140/90 mmHg, etc.)
- Identifying medications not used (statins, ACEi/ARB)

Understanding the burden of disease and complications
- Comprehending disease disparities
- Finding undiagnosed cases

Identifying patients for research projects
Comparing care quality across sites
Comparing complexity of patients across sites
Comparative effectiveness research
Epidemiologic surveillance, including longitudinal analyses
Population-based care management studies of people with diabetes

ACEi, angiotensin-converting enzyme inhibitor; ARB, angiotensin receptor blocker; BP, blood pressure; HbA1c, glycated hemoglobin.

Table 1. Components and algorithm for 7 EHR phenotypes for diabetes

	Abbreviation Full name	ICD_250 ICD-9 code 250.xx	CCW Expanded ICD-9 codes	A1c Abnormal A1c (≥ 6.5%)	Med DM Med	DDC DDC	Sup-DM SUPREME DM	NW eMERGE NW	A1c_Med Abnormal A1c + DM Med[a]
	Based on	ICD-9 billing codes	Other DM codes HEDIS misses	NYC A1c h ealth depart-ment registry	PBM	Population Health Type 2 DM	Population Health Any DM	Exclusively Type 2 DM	A combination of abnormal A1c and PBM
	Components								
1a	ICD-9 250.x0, 250.x2	IP or AMB x1	IP x1 or AMB x2	(–)	(–)	IP, AMB, or ED x 1	IP x1 or AMB x2	IP, AMB, or ED x1	(–)
1b	ICD-9 250.x1, 250.x3	IP or AMB x1	IP x1 or AMB x2	(–)	(–)	(–)	IP x1 or AMB x2	(–)	(–)
1c	ICD-9 249.xx	(–)	IP x1 or AMB x2	(–)	(–)	IP, AMB, or ED x1	(–)	(–)	(–)
1d	ICD-9 other codes	(–)	IP x1 or AMB x2	(–)	(–)	IP, AMB, or ED x1	IP x1 or AMB x2	(–)	(–)
2a	Diabetes Med Group 1: insulin	(–)	(–)	(–)	1 or more DM med on AMB med rec	1 + AMB med rec	1+ AMB med rec	Excludes any patient on insulin or amylin	1 or more DM med on AMB med rec
2b	Diabetes Med Group 2: insulin secreta-gogues and incretins	(–)	(–)	(–)		1+ AMB med rec	1+ AMB med rec	Must have type 2 code	
2c	Diabetes Med Group 3: thiazolidine-diones and metformin	(–)	(–)	(–)		1+ AMB med rec	Excludes if this is the only criteria met		
3	Abnormal glucose lab	(–)	(–)		(–)	2 abnormal labs in past 365 d	2 abnormal labs in past 720 d		(–)
3a	A1c ≥ 6.5% once	(–)	(–)	1+ abnormal	(–)				1+ abnormal
3b	A1c ≥ 6.5 % twice	(–)	(–)	1+ abnormal	(–)				1+ abnormal
3c	Fasting glucose ≥ 126 × 2	(–)	(–)	(–)	(–)				(–)
3d	Random glucose ≥ 200 × 2	(–)	(–)	(–)	(–)				(–)
3e	Abnormal OGTT	(–)	(–)	(–)	(–)				(–)
3f	Two of above	(–)	(–)	(–)	(–)				(–)
	Simplified algorithm	1a or 1b	1a or 1b or 1c or 1d	3a	2a or 2b or 2c	1a or 1c or 1d or 2a or 2b or 2c or 3b or 3c or 3d or 3e or 3f	1a or 1b or 1d or 2a or 2b or 3a or 3b or 3c or 3d or 3e or 3f	1a but never 1b and never 2a unless con-trolled on oral agents	3a or 2
	No. patients identified by phenotype	18 893	16 320	12 182	11 800	22 050	18 958	11 620	15 478
	Extrapolated no. patients with type 2 DM knowing sensi-tivity/specificity that phenotype would identify	13 906	12 804	10 507	9481	14 414	13 422	10 073	12 480
	Extrapolated no. patients with any DM knowing sensi-tivity/specificity that phenotype would identify	15 833	14 521	11 741	10 668	16 387	15 281	10 408	13 904
	PPV type 2 DM	0.74	0.78	0.87	0.80	0.66	0.71	0.86	0.81
	PPV any DM	0.84	0.89	0.97	0.91	0.75	0.81	0.89	0.90

[a]The eighth phenotype, A1c_Med, was a combination of phenotypes 3 (A1c) and 4 (Med) and was devised and studied after the sampling strategy.

AMB, ambulatory; DDC, Durham Diabetes Coalition; DM, diabetes mellitus; ED, emergency department; eMERGE NW, Electronic Medical Records and Genomics Network, Northwestern University; HEDIS, Healthcare Effectiveness Data and Information Set; ICD-9, International Classification of Diseases, Ninth Revision; IP, inpatient; med, medication; NYC, New York City; PBM, pharmacy benefit manager; OGTT, oral glucose tolerance test; PPV, positive predictive value; rec, reconciliation; SUPREME DM, Surveillance, Prevention, and Management of Diabetes Mellitus Project; T2, type 2.

any type 1 diabetes code or any patient ever prescribed insulin, unless the patient is currently well controlled on oral agents alone.[23,24]

The eighth phenotype (A1c_Med) was designed after the data were pulled and is a combination of the A1c phenotype and the Med phenotype. Thus, this phenotype included patients meeting either the A1c phenotype (A1c over 6.5% twice) or being prescribed a diabetes medication (including metformin).

Our aim for this study is to measure the sensitivity and specificity of these 8 different diabetes phenotypes by querying the data from our clinical data warehouse so that researchers, policy makers, and health advocates can better understand, compare, and design population health projects.

MATERIALS AND METHODS

Population of interest and endpoints

The population of interest was the subset of 173 503 adult patients (18 years old or older) living in Durham County, North Carolina, who had electronic health data in the Duke Enterprise Data Warehouse and who met 1 or more of the 7 phenotype definitions for diabetes. Two definitions were specifically designed to identify type 2 diabetes, while the others were designed to identify diabetes more broadly. Some included type 1 and some included secondary causes of diabetes. Clinical expert chart review was conducted using an algorithmic approach (supplemental figures 1 and 2) based on ADA diagnostic criteria (the established gold standard). This approach was conducted to determine whether the patient had diabetes, and, if so, what type. Reviewers were asked to specify whether the patient had type 1, type 2, unspecified/could not determine type, or steroid-induced diabetes.

Methods

This project required several sequential steps. Each of the phenotypes was translated into an algorithm that could be applied to the data in the Duke Enterprise Data Warehouse. These algorithms were developed by a Duke data analyst and verified by another as described in our previous work.[5] The algorithms for each phenotype were then applied to a 5-year extract (between January 1, 2007, and December 31, 2011) from the data warehouse representing 173 503 unique adult patients 18 years or older and residing in Durham County, North Carolina. Each patient was classified as meeting or not meeting the criteria for each phenotype; patients were grouped based on how many phenotype definitions were met: none, 1–4, 5–6, or all 7 phenotype algorithms. (The eighth phenotype was a combination of phenotypes 3 and 4 and was devised and studied after the sampling strategy.) Once a cohort of patients was identified for each different phenotype, a sampling strategy for chart review was designed and implemented based on the literature.[24,25] This strategy was designed to reduce the burden on clinical expert reviewers by targeting their review to a statistically and strategically selected set of patients records, which is described in detail in a later section. We established a protocol for chart review to validate whether phenotype-identified patients truly had type 2 diabetes, type 1 diabetes, unspecified diabetes, or had been falsely identified and did not actually have diabetes. These expert assessments were used as gold-standard diagnoses for diabetes for the calculation of the sensitivity and specificity of each phenotype.

In order to determine which charts should be reviewed for validation, we extracted data from the Duke Enterprise Data Warehouse, which integrates EHRs that contain clinical data (laboratory, diagnostic, clinical notes, tests, etc.) as well as administrative and financial data from clinical encounters across the health system. The 7 diabetes phenotypes were applied to the Duke Enterprise Data Warehouse records (Table 1). An eighth phenotype (A1c_Med), based on combining 2 existing phenotypes, was used in the sensitivity and specificity analysis. A detailed algorithm for how each phenotype was applied is outlined in the supplemental material and in our previous work.[13] Of note, diagnosis codes for gestational diabetes were not included in any of the phenotypes; diagnosis codes for secondary causes of diabetes were included in some of the phenotypes but not the DDC. Some implementers prefer to specifically remove patients with a code of gestational diabetes within 12 months in a cohort of patients with diabetes. There is no specific code for Maturity Onset Diabetes of Youth (MODY) except for type 2 diabetes or secondary diabetes codes; however, if a MODY diagnosis was found in chart review, the patient was coded as other diabetes for purposes of this study.

Sample size design, sampling strategy, and data analysis plan

The phenotype algorithms, once programmed, are applied to the EHR data at a low cost with minimal time. However, obtaining a gold standard based on diagnostic criteria applied by expert clinicians via chart review for assessment purposes is time-consuming and expensive. Due to the large number of records (173 503), we were unable to review all of them. Therefore, to reduce the burden of expert review, we used statistical sampling methods to identify a representative sample of charts for review, as outlined below. Sampling was performed with the goal of a precision of 0.05 around our sensitivity and specificity estimates. To produce robust estimates for all phenotypes, a stratified random sample was necessary; this allowed us to sample more heavily where we believed there to be a higher probability of finding positives or negatives. Sample size calculations were derived from Begg and Greenes estimates for verification bias, and the number of samples per strata was proportional to the strata's variance. Using this approach, we stratify observations based on having similar operating characteristics.

As previously mentioned based on an initial descriptive analysis of the full sample, we divided the population into 4 strata: those who were positive for all 7 phenotype definitions (All 7 Group or Stratum); those who were positive for 5 or 6 of the definitions (5–6 Group or Stratum); those who were positive for 1, 2, 3, or 4 definitions (1–4 Group or Stratum); and those who were negative for all 7 phenotypes (All Negative Group or Stratum). A patient identified as having type 2 diabetes by 5 or 6 different phenotypes should have a higher chance of accuracy than a patient who is positive for only 1 to 4 phenotypes. Additionally, we chose our strata with the goal of ensuring that they contained similar types and numbers of patients (with the exception of the All Negative Stratum).[23] By stratifying our population in this way, we improve our sample design. We expect a different variance around our estimates for each stratum; to minimize our sample size, we can sample differently among the strata. We expect that the sensitivity estimate in the All 7 Stratum would have a smaller variance than the estimate within the 1–4 Stratum. Thus, fewer charts were required for the All 7 Stratum as compared to the 1–4 Stratum. Our final sampling plan is provided in Table 2. As shown: (1) 50 charts in the All Negative stratum, (2) 30 charts in the All 7 Stratum, (3) 160 charts in the 1–4 Stratum, and (4) 160 charts in the 5–6 Stratum (Table 2 and Supplemental Table 1).

Table 2. Patients identified as having diabetes and sampling strategy

Patients identified as having diabetes and sampling strategy	Identified	Sampled
No. of adult patients in Durham County from 2007 to 2011	173 503	400
No. of patients not identified by any of the phenotypes (All Negative Stratum)	150 824	50
No. of patients identified by at least 1 phenotype	22 679	350
No. of patients identified by 1, 2, 3, or 4 phenotypes (1–4 Definitions Stratum)	8033	160
No. of patients identified by 5 or 6 phenotypes (5–6 Definitions Stratum)	9392	160
No. of patients identified by all phenotypes (All 7 Stratum)	5254	30

Chart review, validation, and the adjudication process

Chart review by clinical experts was used as the gold standard to determine whether the electronic algorithm using an EHR phenotype correctly identified patients as having diabetes. Informatics experts designed a data-collection system with feedback from 8 diabetes providers (physicians and nurse practitioners) to support the review of sampled EHRs.[24] Two independent clinical reviewers were assigned to review each sampled record and were instructed to review all available data, including encounter notes and scanned documents that contained handwritten information, from the EHR that was recorded during the selected 5-year observation period (January 1, 2007, to December 31, 2011). All electronic, legacy, and scanned paper documents were reviewed to answer 3 main questions: Is there a diagnosis of diabetes in the chart? Is the patient on a diabetes medication (for diabetes), and is there evidence for abnormal labs? The last question asked each reviewer to determine by clinical acumen whether the patient had diabetes, and what type.[25]

We used the adjudicated answer to the final question to calculate the sensitivity and specificity of each phenotype for identifying diabetes. A subset of the final question was what type of diabetes the patient had; the answer to this subset question was used to calculate the sensitivity and specificity of each phenotype in identifying type 2 diabetes. A schema (supplementary figures 1 and 2) was developed to aid the practitioner in applying the gold-standard ADA criteria to the EHR data for determining whether a patient had diabetes. Each expert chart reviewer used the gold standard to assess whether the patient did indeed have diabetes and what type—type 1 or type 2 diabetes mellitus (DM). In many cases, this involved relying on lab data. However, in some cases, this meant relying on the medical notes of providers caring for the patient who had already used the American Diabetes Association (ADA) gold standard to diagnose the patient in the past and there was mention of diabetes or diabetes medications in problem lists, impressions, diagnoses, or in the treatment plan to the extent that it was obvious the patient had diabetes. This schema was used by each reviewer and any discordant determinations between reviewers were ultimately decided by the senior endocrinologist.

Four hundred records were sampled and reviewed separately by 20 experts (15 physicians and 5 nurse practitioners) with an average of 9.6 years of postdegree endocrinology experience (range 1–15 years). There were 150 discordant records; these were subsequently reviewed and adjudicated by a senior endocrinologist (SS) with 16 years of experience in diabetes care.

Within the appointed time period, there were 173 503 adult patients living in Durham County with available data in the Duke Enterprise Data Warehouse (Table 1). Of those, 22 679 met at least 1 of the phenotype criteria for diabetes. The number of patients identified using each phenotype algorithm ranged from 11 620 to 22 050.

In an individual patient, diabetes is diagnosed based on laboratory criteria (fasting glucose ≥ 126 mg/dL, random glucose ≥ 200 mg/dL, or a HgbA1c $\geq 6.5\%$). However, secondary data analysis can also leverage diagnosis codes and prescriptions to identify patients with diabetes. Each chart reviewer was instructed to look for a diabetes diagnosis on problem lists or within notes, the presence of diabetes medication(s), and abnormal laboratory values (glucose and HgbA1c) and make a clinical decision about whether the identification of diabetes was correct; if so, reviewers were asked to further characterize diabetes by type (1, 2, or unspecified). They were also asked to identify whether the patient had any disease that could lead to a false association with diabetes (e.g., obesity, prediabetes, polycystic ovary syndrome, nonalcoholic fatty liver disease, or steroid-induced hyperglycemia). The final endpoint was whether the patient actually had diabetes, along with diabetes type (Supplemental Figures 1 and 2).

Reviewers' assessments of particular data points were recorded on a 15-item electronic review form. The Research Electronic Data Capture platform[26]—a secure, web-based application for building and managing online databases—was used to manage the random assignment of charts to reviewers and the collection of data for each review. Experts (physicians and nurse practitioners) were recruited from the Duke University Health System and were trained on chart review in Maestro Care (Duke Medicine's EHR, powered by Epic of Verona, WI, USA) as well as Research Electronic Data Capture in 1-hour training sessions. A manual of operations was developed as a reference to supplement training. The reviewers examined electronic charts for a defined time range (2007–2011) to match the time period of the phenotype queries. In the event that reviewing clinicians disagreed on a diagnosis, the chart was sent to a senior adjudicator for final determination. We also sampled 5% of records about which the clinicians agreed on diagnosis. Discordant records were defined by differences in the final diagnosis designation: diabetes type 1, diabetes type 2, unspecified diabetes, or no diabetes.

Statistical analysis

The sensitivity and specificity of each phenotype was calculated with chart review being treated as the gold standard. Our sampling approach—designed to optimize estimation of sensitivity and specificity—naturally induces verification bias,[27] which we accounted for when estimating the operating characteristics of each phenotype.

We used methods from Begg and Greenes[24] to generate point estimates and 95% confidence intervals around the sensitivity and specificity of each phenotype for type 2 diabetes and any diabetes (type 1, type 2, or unspecified). Since the size of the All Negative Stratum ($n = 150\,824$) is so much larger than the others, unnormalized estimates of sensitivity are somewhat unstable. A single false positive in the largest symptom has a small impact on the false positive rate for that class, but it dramatically changes the estimate of the total number of patients with disease. Therefore, we use a Bayesian prior to stabilize our estimates.[28] The Bayesian priors were chosen to be uniform in all strata except in the All Negative Stratum. In this stratum, we placed an informative Bayesian prior on the false

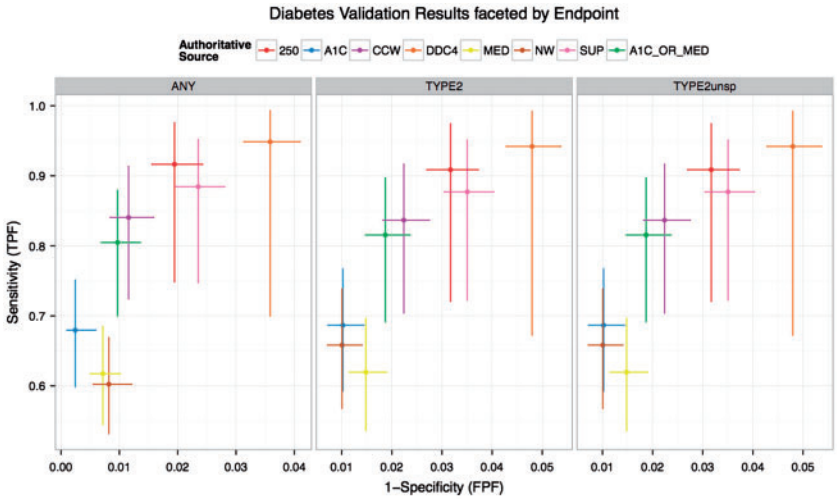

Figure 1. Sensitivity and specificity of diabetes phenotypes. ICD_250, International Classification of Diseases, Ninth Revision, code 250.xx; CCW, expanded ICD-9 codes; A1c, abnormal A1c, based on New York City Health Department A1c Registry; Med, DM medication, based on pharmacy benefit manager data; DDC, Durham Diabetes Coalition; SUPREME DM or Sup-DM, Surveillance, Prevention, and Management of Diabetes Mellitus; eMERGE NW or NW, electronic medical records and genomics; A1c_Med, abnormal A1c + DM medication.

negative cells that approximates the estimated prevalence of diabetes in Durham County. We note that this is an overestimate of the prevalence of type 2 diabetes (because many patients with disease will be identified by 1 of the computable phenotypes); this leads to the likelihood of underestimating sensitivity.[24] These Bayesian priors will normalize the false-negative rate to more closely reflect the prevalence of diabetes. Note that because all definitions share the same false-negative rate in the All Negative Stratum, comparison between the sensitivity of definitions is quite accurate even if confidence bands around sensitivity estimates are somewhat large.

RESULTS

Figure 1 shows the sensitivity and specificity of the 7 EHR phenotypes described plus the combined phenotype (A1c_Med) in identifying type 2 diabetes. Against the gold standard for identifying type 2 diabetes, the DDC phenotype had a sensitivity of 0.942 (0.672–0.992) and specificity of 0.952 (0.946–0.957). The A1c, NW, and Med phenotypes had sensitivities of 0.687 (0.593–0.767), 0.658 (0.567–0.739), and 0.62 (0.536–0.697), respectively, and specificities of 0.99 (0.986–0.993), 0.99 (0.986–0.993), and 0.985 (0.981–0.988).

Knowing the true rate of type 2 diabetes (derived from clinical expert chart review), we extrapolated the number of patients with true type 2 diabetes out of those identified by phenotype: in the range 9481 to 14 414 (Table 1).

Of those patients who were false positive for any diabetes, many had conditions or criteria that were associated with other forms of abnormal glucose metabolism or prediabetes (Supplemental Table 2). Forty-five percent of patients identified as having diabetes who did not have diabetes upon chart review actually had inpatient hyperglycemia as defined by random glucose over 200 mg/dL with no abnormal glucose outside the inpatient admission. In other cases it was difficult to determine if the patient had type 1 or type 2 diabetes, for instance in an obese patient on insulin therapy.

While Figure 1 showed the sensitivity and specificity of the algorithms for identifying type 2 diabetes, many of these phenotypes were not originally designed to identify only type 2 diabetes. Thus, analysis for the sensitivity of finding any diabetes was also determined: ICD_250 0.916 (0.748–0.976), CCW 0.84 (0.724–0.914), A1c 0.68 (0.598–0.751), DDC 0.949 (0.7–0.993), Sup-DM 0.884 (0.747–0.952), NW 0.602 (0.532–0.669), and A1c_Med 0.805 (0.7–0.879). Specificity was as follows: ICD_250 0.981 (0.976–0.984), CCW 0.988 (0.984–0.992), A1c 0.998 (0.994–0.999), DDC 0.964 (0.959–0.969), Sup-DM 0.976 (0.972–0.98), and NW 0.992 (0.988–0.994; Supplemental Tables 3 and 4).

DISCUSSION

We used a targeted sampling technique to measure the sensitivity and specificity of 7 different phenotype definitions. We conducted this analysis using only 400 sampled charts. By extrapolating our results, we found 15 303 of 22 679 patients (67.5%) were accurately identified as having type 2 diabetes. For any given phenotype definition, the number of patients that had to be extrapolated for accurate identification ranged from 9481 to 14 414 patients. Although our analysis shows variation in sensitivity and specificity across many different diabetes definitions, some phenotypes had comparable sensitivity estimates (DDC, ICD_250, and Sup-DM). As expected, these were the definitions with the broadest criteria, including the widest range of diagnoses codes and multiple classification criteria. The phenotype definitions that were least sensitive (A1c, Med, and NW) were more specific, with only 1 data source (e.g., HbA1c test result) or very specific inclusion and exclusion criteria. In an ideal world, a phenotype definition would be highly sensitive and highly specific, but one measurement often comes at the expense of the other.

There are situations where one might prefer sensitivity over specificity. Less specific but more sensitive phenotypes included those using inpatient glucose data or diabetes medications that can be used to treat conditions other than diabetes, like polycystic ovary syndrome, prediabetes, obesity, and nonalcoholic fatty liver disease. Of those patients identified as having diabetes who did not actually have diabetes, 10% were on metformin. Thiazolidinediones and alpha-glucosidase were not identified as posing a risk for falsely identifying patients, likely because they are not used often. Of those patients identified as having diabetes, 30% had stress hyperglycemia and 15% had steroid-induced hyperglycemia rather than diabetes. Broadening a phenotype definition to make it more sensitive can come at the expense of including patients with prediabetes or dis-

eases associated with insulin resistance. Thirty-seven percent of phenotype-identified patients our reviewers decided actually did not have type 2 diabetes did have a diagnosis of something related to abnormal glucose metabolism, either prediabetes, type 1 diabetes, or other types of diabetes (pancreatic disease), or were cases for which there were not enough data in the medical record to make a determination. In addition, allowing a diagnosis code to occur just once as an inclusion rule increased the risk of producing a false positive.

There are scenarios in which a more specific phenotype definition is warranted (e.g., identifying a cohort for a drug-effectiveness study). Qualities of more specific but less sensitive phenotypes include excluding patients with any type 1 diabetes code, excluding patients who have not been seen for a diabetes office visit within 1 year, requiring more than 1 diagnostic code, and excluding patients on certain diabetes medications, particularly insulin.

In the case of the NW phenotype, it is not surprising that it was more specific than it was sensitive because it was designed to specifically identify a cohort of patients who definitely had type 2 diabetes. The purpose of the NW phenotype was not to create an all-inclusive list of patients with diabetes. This particular phenotype has been used for a wide range of purposes: genetic studies, research projects, and to prompt providers to add diabetes to the problem list in EHRs.[23]

The availability of data is also critical to selecting a phenotype definition. PBMs only have access to medication data, health departments only have access to laboratory data, and insurance companies only have access to claims data. The Med phenotype (our attempt to mimic a PBM database) and A1c phenotype (our attempt to mimic the New York City Health Department registry) are examples of using the data that are available.

The characteristics of the data themselves, such as timing of data or the number of patient visits, can impact the performance of a phenotype definition. We identified nearly 4000 patients with diabetes who did not have a coded diagnosis of diabetes in the EHR and had not been prescribed diabetes medication. Concerned that we had an abundance of undiagnosed diabetes, we reviewed charts and found that 45% of those patients had inpatient hyperglycemia (as defined by random glucose over 200 mg/dL). On the other hand, patients with well-controlled diabetes may have normal glucose readings, or fasting glucose data could be labeled as random.[7] Patients who regularly attend medical appointments have more "opportunity" for diabetes codes (observation bias), and consequently are more likely to be identified using any phenotype definition that includes diagnosis codes.

Errors can also occur when using diagnosis codes in identifying diabetes cohorts. Patients can be accidentally coded as having the wrong type of diabetes (type 1 vs type 2) or patients can be coded as having chronic diabetes when they in fact have prediabetes or acute steroid-induced hyperglycemia.[29,30] Phenotype definitions based purely on automated searches by visit diagnosis codes miss those patients with undiagnosed diabetes or those seen for a different problem. Using ICD-9 criteria alone can fail to detect patients who are undiagnosed or who are admitted for problems unrelated to the disease in question.[15–17] Diagnosis codes can be inaccurate, identifying patients as type 2 when they have type 1. Strategies that use the ratio of type 1 vs type 2 codes or use laboratory data such as C-peptide levels can be employed to refine a phenotype.[16]

For the purposes of the DDC, creating a sensitive phenotype that identified the most patients with diabetes allowed us to implement a broad range of projects to reduce complications, death, and disability from diabetes in our community. With our goal of finding all patients with diabetes and our concern that many of the patients at the highest risk for diabetes might be less connected to health care, we designed a very broad phenotype. The DDC phenotype was the most sensitive, likely due to having the broadest inclusion criteria, such as use of metformin, diagnostic codes generated from emergency department visits, and allowing only 1 outpatient (ambulatory) code to count. However, these broad inclusion criteria affect specificity, which was lowest. Many patients identified by the phenotype did not have type 2 diabetes; many had only stress- or steroid-induced hyperglycemia or prediabetes states. The Sup-DM phenotype did not allow secondary diabetes codes, which limited its ability to find all patients with diabetes but avoided the pitfall of tracking those with steroid- or stress-induced hyperglycemia. As the DDC expands its goal to reduce death and disability in patients with all types of diabetes, we recommend editing our current phenotype to include type 1 diabetes, require 2 instances of a diagnosis code, and eliminate metformin alone as a criterion to identify diabetes.

CONCLUSIONS

Phenotypes used to identify patients with diabetes can be created from a variety of components. The way in which these components are assembled and applied affects the number of patients identified. Analysis of electronic health data presents many opportunities for generating hypotheses, validating exploratory and predictive models, and testing new types of statistical approaches; however, these opportunities can be adversely affected by the challenges of unstandardized cohort definitions and the complexity of defining diseases. One of the greatest challenges is that phenotypic definition encompasses all of the possible data elements. The most important step in mitigating this challenge is to understand how the chosen phenotype performs when comparing projects to use the same phenotype. Accurately identifying all patients with diabetes is essential to assessing risk, developing interventions, and preventing complications and disability attributable to diabetes. Choosing a phenotype depends on the intended use for the cohort and the availability of the components of each phenotype definition. This study establishes important groundwork for diabetes phenotype definitions, and the approach used in generating the framework may be generalizable to other conditions. Understanding how phenotype definitions differ informs planning, assessing, and comparing diabetes research studies, quality-improvement or community projects, and governmental policies.

FUNDING

The projects and the work described in this article are supported in part by (1) Grant Number 1C1CMS331018-01-00 from the Department of Health and Human Services, Centers for Medicare & Medicaid Services, and in part by (2) the Bristol Myers Squibb Foundation Together on Diabetes program, (3) NIH T32 grant Endocrinology and Metabolism Research Training Program of the National Institutes of Health under award number NIH 5T32DK007012, and (4) Grant UG1DA040317 from the National Institute on Drug Abuse. The contents of this article are solely the responsibility of the authors and have not been approved by the Department of Health and Human Services, Centers for Medicare & Medicaid Services, or the NIH.

DISCLOSURES

The authors have no disclosures to report.

CONTRIBUTIONS

SES: Guarantor. Concept design, literature review, chart reviewer education, chart review, chart adjudicator, analysis, manuscript

writing, and editing. KP: Literature review, chart review, analysis, manuscript writing, and editing. BBG: Analysis, manuscript writing, and editing. BCB: Literature review, chart review, analysis, manuscript writing, and editing. MP: Statistics, analysis. MLM: Concept design, analysis, manuscript writing, and editing. EB: Analysis, manuscript writing, and editing. JEL: Statistics. CLN: Statistics. BN: Statistics. BG: Conceptualization of the statistical approach as well as analysis, writing, and editorial review of the final paper. PB: Statistics, education of chart reviewers, analysis, manuscript writing, and editing. RLR: Concept design, chart reviewer education, literature review, analysis, manuscript writing, and editing. ILR: Chart review, literature review, analysis, manuscript writing, and editing. BH: Chart review, literature review, analysis, manuscript writing, and editing. LC: Chart review, literature review, analysis, manuscript writing, and editing. ERMcPH: Chart review, literature review, analysis, manuscript writing, and editing. MP-L: Chart review, literature review, analysis, manuscript writing, and editing. KK: Chart review, literature review, analysis, manuscript writing, and editing. SR: Chart review, literature review, analysis, manuscript writing, and editing. JG: Chart review, literature review, analysis, manuscript writing, and editing. ABB: Chart review, literature review, analysis, manuscript writing, and editing. NJ: Chart review, literature review, analysis, manuscript writing, and editing.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGMENTS

We would like to acknowledge Dr Robert M Califf for encouraging us to do this project, and Morgan deBlecourt for manuscript editing and preparation. We would like to thank and acknowledge Dr Christopher B Granger for manuscript editing.

REFERENCES

1. American Diabetes Association. Economic costs of diabetes in the U.S in 2012. *Diabetes Care*. 2013;36:1033–46.
2. The Diabetes Control and Complications Trial Research Group. The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. *N Engl J Med*. 1993;329:977–86.
3. Centers for Disease Control and Prevention. National Diabetes Statistics Report, 2014: *Estimates of Diabetes and Its Burden in the United States*. Atlanta, GA: US Department of Health and Human Services; 2014. http://www.cdc.gov/diabetes/pubs/statsreport14/national-diabetes-report-web.pdf. Accessed November 2, 2015.
4. American Diabetes Association. Classification and diagnosis of diabetes. *Diabetes Care*. 2016;39 (Suppl 1): S13–22.
5. Richesson RL, Rusincovitch SA, Wixted D, *et al*. A comparison of phenotype definitions for diabetes mellitus. *J Am Med Inform Assoc*. 2013;20:e319–26.
6. Centers for Disease Control and Prevention. *Behavioral Risk Factor Surveillance System*. http://www.cdc.gov/brfss/. Accessed May 8, 2015.
7. US Department of Health and Human Services. *HITECH Act Enforcement Interim Final Rule*. http://www.hhs.gov/hipaa/for-professionals/special-topics/HITECH-act-enforcement-interim-final-rule/index.html. Accessed February 16, 2016.
8. Nichols GA, Schroeder EB, Karter AJ, *et al*. Trends in diabetes incidence among 7 million insured adults, 2006-2011: the SUPREME-DM project. *Am J Epidemiol*. 2015;181:32–9.
9. Holt TA, Stables D, Hippisley-Cox J, *et al*. Identifying undiagnosed diabetes: cross-sectional survey of 3.6 million patients' electronic records. *Br J Gen Pract*. 2008;58:192–6.
10. Vinker S, Fogelman Y, Elhayany A, *et al*. Usefulness of electronic databases for the detection of unrecognized diabetic patients. *Cardiovasc Diabetol*. 2003;2:13.
11. Kudyakov R, Bowen J, Ewen E, *et al*. Electronic health record use to classify patients with newly diagnosed versus preexisting type 2 diabetes: infrastructure for comparative effectiveness research and population health management. *Popul Health Manag*. 2012;15:3–11.
12. Baus A, Wood G, Pollard C, *et al*. Registry-based diabetes risk detection schema for the systematic identification of patients at risk for diabetes in West Virginia primary care centers. *Perspect Health Inf Manag*. 2013;10:1f. eCollection 2013.
13. Ho ML, Lawrence N, van Walraven C, *et al*. The accuracy of using integrated electronic health care data to identify patients with undiagnosed diabetes mellitus. *J Eval Clin Pract*. 2012;18:606–11.
14. Mishra NK, Son RY, Arnzen JJ. Towards automatic diabetes case detection and ABCS protocol compliance assessment. *Clin Med Res*. 2012;10:106–21.
15. Schultz S, Seddig T, Hanser S, *et al*. Checking coding completeness by mining discharge summaries. *Stud Health Technol Inform*. 2011;169:594–8.
16. 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.
17. Morris AD, Boyle DI, MacAlpine R, *et al*. The diabetes audit and research in Tayside Scotland (DARTS) study: electronic record linkage to create a diabetes register. DARTS/MEMO Collaboration. *BMJ*. 1997;315:524–8.
18. Chamany S, Silver LD, Bassett MT, *et al*. Tracking diabetes: New York City's A1C Registry. *Milbank Q*. 2009;87:547–70.
19. Spratt SE, Batch BC, Davis LP, *et al*. Methods and initial findings from the Durham Diabetes Coalition: integrating geospatial health technology and community interventions to reduce death and disability. *J Clin Transl Endocrinol*. 2015;2:26–36.
20. 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.
21. Desai JR Wu P, Nichols GA, *et al*. Diabetes and asthma case identification, validation, and representativeness when using electronic health data to construct registries for comparative effectiveness and epidemiologic research. *Med Care*. 2012;50 (Suppl):S30–5.
22. Vogt TM, Elston-Lafata J, Tolsma D, *et al*. The role of research in integrated healthcare systems: the HMO Research Network. *Am J Manag Care*. 2004;10:643–648.
23. Pacheco JA, Thompson W. *Type 2 Diabetes Mellitus Electronic Medical Record Case and Control Selection Algorithms*. 2011. https://phekb.org/sites/phenotype/files/T2DM-algorithm.pdf. Accessed November 2, 2015.
24. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics*. 1983;39:207–15.
25. Richesson RL, Barth P, Nelson CL, *et al*. A Chart Review System for the Validation of Computable Phenotypes in Diabetes. *Poster presented at the American Medical Informatics Association (AMIA) Joint Summits for Translational Research*. San Francisco; March 2015.
26. REDCap website. https://redcap.vanderbilt.edu/. Accessed November 2, 2015.
27. Cronin AM, Vickers AJ. Statistical methods to correct for verification bias in diagnostic studies are inadequate when there are few false negatives: a simulation study. *BMC Med Res Methodol*. 2008;8:75.
28. Pepe, MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction (Oxford Statistical Science Series)*. New York: Oxford University Press; 2003.
29. Seidu S, Davies MJ, Mostafa S, *et al*. Prevalence and characteristics in coding, classification and diagnosis of diabetes in primary care. *Postgrad Med J*. 2014;90:13–17.
30. Liaw ST, Taggart J, Yu H, *et al*. Data extraction from electronic health records - existing tools may be unreliable and potentially unsafe. *Aust Fam Physician*. 2013;42:820–3.