Clinical Epidemiology

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Validation of Register-Based Diabetes Classifiers in Danish Data

Anders Aasted Isaksen[1,2], Annelli Sandbæk[1,2], Lasse Bjerg[1,2]

[1]Department of Public Health, Aarhus University, Aarhus, Denmark; [2]Steno Diabetes Center Aarhus, Aarhus University Hospital, Aarhus N, Denmark

Correspondence: Anders Aasted Isaksen, Institut for Folkesundhed, Aarhus Universitet, Bartholins Allé 2, Aarhus, 8000, Denmark, Email aai@ph.au.dk

Purpose: To validate two register-based algorithms classifying type 1 (T1D) and type 2 diabetes (T2D) in a general population using Danish register data.

Patients and Methods: After linking data on prescription drug usage, hospital diagnoses, laboratory results and diabetes-specific healthcare services from nationwide healthcare registers, diabetes type was defined for all individuals in Central Denmark Region age 18–74 years on 31 December 2018 according to two distinct register-based classifiers: 1) a novel register-based diabetes classifier incorporating diagnostic hemoglobin-A1C measurements, the *Open-Source Diabetes Classifier* (OSDC), and 2) an existing Danish diabetes classifier, the *Register for Selected Chronic Diseases* (RSCD). These classifications were validated against self-reported data from the *Health in Central Denmark* survey – overall and stratified by age at onset of diabetes. The source-code of both classifiers was made available in the open-source *R* package *osdc*.

Results: A total of 2633 (9.0%) of 29,391 respondents reported having any type of diabetes, divided across 410 (1.4%) self-reported cases of T1D and 2223 (7.6%) cases of T2D. Among all self-reported diabetes cases, 2421 (91.9%) were classified as diabetes cases by both classifiers. In T1D, sensitivity of OSDC-classification was 0.773 [95% CI 0.730–0.813] (RSCD: 0.700 [0.653–0.744]) and positive predictive value (PPV) 0.943 [0.913–0.966] (RSCD: 0.944 [0.912–0.967]). In T2D, sensitivity of OSDC-classification was 0.944 [0.933–0.953] (RSCD: 0.905 [0.892–0.917]) and PPV 0.875 [0.861–0.888] (RSCD: 0.898 [0.884–0.910]). In age at onset-stratified analyses of both classifiers, sensitivity and PPV were low in individuals with T1D onset after age 40 and T2D onset before age 40.

Conclusion: Both register-based classifiers identified valid populations of T1D and T2D in a general population, but sensitivity was substantially higher in OSDC compared to RSCD. Register-classified diabetes type in cases with atypical age at onset of diabetes should be interpreted with caution. The validated, open-source classifiers provide robust and transparent tools for researchers.

Keywords: type 1 diabetes, type 2 diabetes, classification, population-based, open-source

Plain Language Summary

Why was this study done?

General-purpose registers and other administrative databases often provide the basis of diabetes epidemiology, but rarely contain validated diabetes-specific data. If the diabetes-specific data are not accurate, bias may be induced into studies, and findings could differ with various diabetes definitions. Therefore, we set out to answer a key question for diabetes epidemiologists: Are register-based classifications of diabetes type accurate when applied to a general population?

What did the researchers do and find?

Using nationwide Danish healthcare register data, we implemented and validated two distinct register-based classifiers of diabetes type in a general survey population. We found that both classifiers accurately identified type 1 and type 2 diabetes in a general population, but performed poorly in the minority of cases with atypical age at onset of diabetes.

What do these results mean?

To provide robust findings, register-based diabetes studies should use diabetes classifiers validated on data and source populations similar to those under study. To support this, we made the source code of both classifiers available in the open-source *R* package *osdc*, empowering diabetes epidemiologists with a convenient tool for future studies.

Received: 15 February 2023
Accepted: 31 March 2023
Published: 5 May 2023

Introduction

Epidemiological studies are often based on administrative data from general-purpose registers and other databases, which may not be validated for a particular purpose. In diabetes epidemiology, it is important to have an accurate tool to identify individuals with diabetes in the registries, as findings may differ with various diabetes definitions due to misclassification bias.[1,2] During the last decades, considerable efforts have been made towards establishing such a tool for diabetes research in several countries.[3–6] However, a review of US databases found just two validated classifiers for identifying type 1 diabetes (T1D) and type 2 diabetes (T2D),[7] and these were only validated within cohorts limited to individuals with diabetes.[8–10] Later studies have validated additional type-specific diabetes classifiers in diabetes-only cohorts,[11–14] but none have been validated in a general population, to the best of our knowledge, and performance of type-specific diabetes classifiers in population-based settings remains unclear.

Denmark is well known for its numerous nationwide healthcare and administrative registries, which offer unique opportunities for population-based research in health care and epidemiology. Data on diagnoses in the primary care sector – where most diabetes cases are diagnosed and treated – are not collected in a nationwide register, and this presents a challenge to identification of diabetes cases, as classification algorithms must use data on indirect process indicators instead. In Denmark, the first nationwide resource readily available to researchers to identify diabetes cases was the National Diabetes Register, established in 2006, which defined diabetes (of any type) based on register data on hospital diagnoses, prescription drug purchases, diabetes-specific podiatrist services and frequency of blood-glucose measurements.[15,16] The National Diabetes Register was discontinued in 2012, and a later validation study questioned its validity and called for future registers to adopt inclusion based on elevated hemoglobin-A1c (HbA1c) levels.[17] A national diabetes database based on reports from outpatient diabetes clinics and primary care physicians was also established, but due to lack of reporting from general practices, only a minority of individuals with T2D were included.[18] In 2014, the Danish Health Data Authority launched the Register of Selected Chronic Diseases (RSCD),[19] which aimed to identify incidence and prevalence of a range of chronic diseases based register data on hospital diagnoses and prescription drug purchases, including T1D and T2D. At present, RSCD is the only publicly available resource to identify diabetes cases in Danish register data (by application to the Danish Health Data Authority), but it has not been publicly validated nor is the source code behind the algorithm publicly available. Notably, the algorithm lacks inclusion based on elevated HbA1c levels.

The aims of this study were 1) to develop a novel open-source diabetes-classification algorithm (the Open-Source Diabetes Classifier, *OSDC*) incorporating elevated HbA1c levels as an inclusion criterion, 2) validate this algorithm against independent self-reported data from health surveys, and 3) validate the diabetes-classification algorithms of RSCD as a point of reference against the performance of OSDC.

Materials and Methods

Setting

From nationwide healthcare registers (described below), data covering the Central Denmark Region were used to identify diabetes populations using the OSDC and RSCD classifiers. The Central Denmark Region is one of five administrative regions in Denmark, with a population of 1.3 million inhabitants (22% of the entire Danish population). Survey data from the *Health in Central Denmark* survey[20] was used to validate a subset of the diabetes populations generated by the OSDC and RSCD classifiers.

Data

Survey data

Health in Central Denmark is a digital and postal questionnaire survey conducted in 2020 on all inhabitants of Central Denmark Region aged 18 to 74 years identified as prevalent diabetes cases by OSDC on 31 December 2018, plus an equally-sized group of OSDC non-diabetes cases (matched to diabetes cases by sex, age, and municipality).[20] The survey collected self-reported data related to health in general, with an additional focus on items related to diabetes mellitus, such as current disease, diabetes type, and age at onset. On the index date, 942,572 individuals aged 18–74 years resided

in Central Denmark Region, of whom 44,659 OSDC diabetes cases and 46,195 matched OSDC non-diabetes cases were invited to the survey. In total, 51,854 (57%) responded.

Register Data

Due to the public healthcare system in Denmark, the registers cover the entire population and can be linked at the individual level.[21] Information on age, sex, and immigrant origin was obtained from the *Danish Civil Registration System*.[22] Information on hospital admissions and outpatient contacts was obtained from the *Danish National Patient Register* from 1994 through 2018.[23] Information on diabetes-specific podiatrist services in the primary healthcare sector was obtained from the *Danish National Health Service Register* from 1990 through 2018.[24] Information on purchases of glucose-lowering drugs (GLD) was obtained from the *Danish National Prescription Registry* from 1995 through 2018.[25] Information on routine clinical HbA1c samples was obtained from the *Clinical Laboratory Information System* of Central Denmark Region and the *Register of Laboratory Results for Research* from 2011 through 2018. Supplementary S1 provides detailed description of data codes used in the registers.

Diabetes Classification Algorithms

The Open-Source Diabetes Classifier

Diabetes is defined at the second occurrence of any event across four types of inclusion events: 1) HbA1c measurements of ≥48 mmol/mol (6.5%) (censoring events in pregnancies as potential gestational diabetes mellitus (GDM)), 2) hospital diagnoses of diabetes, 3) diabetes-specific services received at podiatrist, 4) purchases of GLD (excluding brand drugs for weight loss, eg *Saxenda*, censoring purchases during pregnancies as potential GDM, and metformin purchases in women below age 40 as potential polycystic ovary syndrome (PCOS)). All available data is used, except purchases of GLD, which are restricted to data from 1997 onwards.

Diabetes type is classified as either T1D or T2D based on patterns of purchases of insulins (including analogues) and hospital primary diagnoses of T1D and T2D. Classification as T1D requires an individual to have either 1) purchased only insulins and never any other type of GLD, and have at least one diagnosis of T1D, or 2) have a majority of T1D diagnoses from endocrinological departments (or from other medical departments, in the absence of contacts to endocrinological departments), and a purchase of insulin within 180 days after onset of diabetes, with insulin contributing at least two-thirds of all defined daily doses of GLD purchased.[26] In populations generated on a fixed index date (such as this cross-sectional study), individuals classified as T1D must have purchased insulins in the last year prior to the index date. Individuals not classified as T1D are classified as T2D.

The Register for Selected Chronic Diseases

We developed an implementation of the algorithms of the RSCD diabetes classifier (version 1.0, August 2016) according to official documentation.[27] Diabetes is defined by two inclusion events: i) type-specific diagnoses of T1D and T2D, and ii) purchases of GLD.

Individuals are classified as T2D if they have at least one purchase of non-insulin GLD, or have a hospital diagnosis of T2D as the most recent type-specific diabetes diagnosis. Exclusions from the T2D population include 1) women with only metformin purchases and any diagnoses of PCOS or purchases of clomifene or antiandrogens and estrogens, 2) individuals with only one recorded inclusion event, and 3) individuals with no recorded inclusion events in the last 10 years prior to the index date.

Individuals are classified as T1D if they have at least one purchase of insulins, or have a diagnosis of T1D as the most recent type-specific diabetes diagnosis, and fulfill no exclusion criteria. Exclusions from the T1D population include 1) women with any diagnoses of GDM, who have made purchases of GLD only in the period from 280 days prior to their first diagnosis of GDM to 280 days after their last diagnosis of GDM; 2) individuals classified as T2D; 3) individuals without any purchases of GLD, or have made only one purchase and have no hospital records of T1D; 4) individuals with no insulin purchases in the last 10 years prior to the index date.

Figure 1 shows the flow of diabetes classification in OSDC and RSCD (more details on OSDC are available in Supplementary S2).

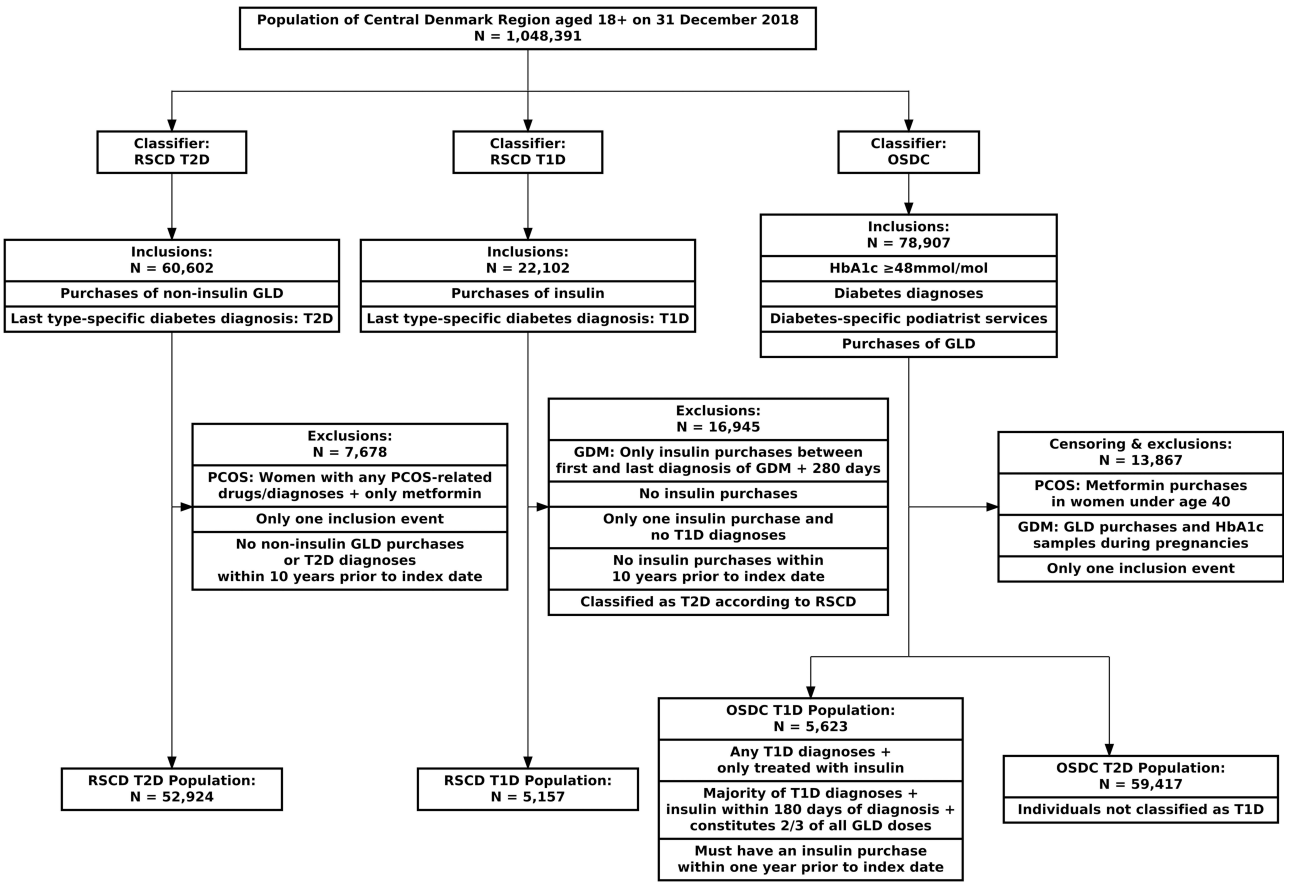

Figure 1 Algorithm flow of individuals in each diabetes classifier.
Abbreviations: T1D, type 1 diabetes; T2D, type 2 diabetes; OSDC, Open-Source Diabetes Classifier; RSCD, Register for Selected Chronic Diseases; GDM, gestational diabetes mellitus; PCOS, polycystic ovary syndrome; GLD, glucose-lowering drugs.

Self-Reported Diabetes Variable

In the survey data, self-reported diabetes type was categorized as either T1D (self-reported T1D), T2D (all other types of diabetes) or no diabetes, corresponding to the diabetes types discernable by the register-based classifiers.

Study Population

Among survey respondents, 2411 were excluded due to missing data on diabetes items. In addition, 2093 individuals with self-reported onset of diabetes after 31 December 2018 were excluded, as diabetes cases onset after the index date would erroneously evaluate to false-negative cases in the register-based classifiers, due to the delay from register-classification on the index date (31 December 2018) until questionnaire responses (November 2020). After these exclusions, 47,350 individuals remained, but due to survey invites being conditioned on OSDC diabetes status, the survey population was biased towards the OSDC classification and OSDC diabetes prevalence was 43.7% (20,692 individuals). To account for this, we first estimated the OSDC diabetes prevalence to be 9.3% (2483 individuals) in a random sample of 26,665 individuals from the background population with the same age, sex and municipality distributions as the OSDC non-diabetes cases of respondents. To offset the oversampling of OSDC diabetes cases in the survey, OSDC diabetes cases were randomly subsampled to 2733 individuals to achieve an unbiased OSDC diabetes prevalence of 9.3% in the final study population of 29,391 individuals. Figure 2 shows the flow of the survey study population.

Statistical Analysis

Characteristics of the study population were tabulated according to self-reported, OSDC-classified, and RSCD-classified diabetes type. Validation analyses were performed separately for T1D and T2D, where each register-based diabetes

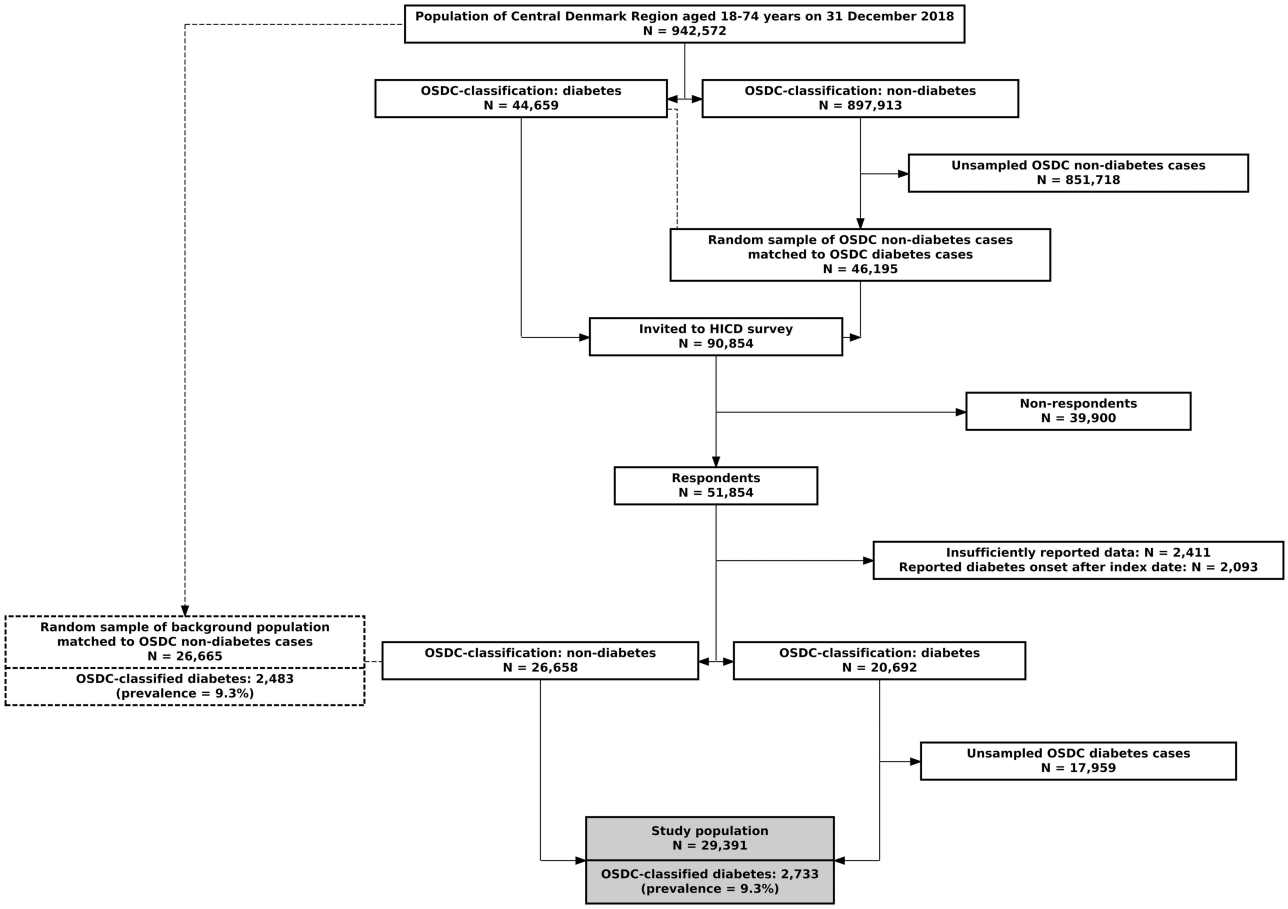

Figure 2 Flow of individuals into the study population.
Abbreviations: OSDC, Open-Source Diabetes Classifier; HICD, Health in Central Denmark survey.

classifier was validated against self-reported diabetes status, treating diabetes status as a distinct binary variable for each diabetes type (eg, in the analyses of T1D, diabetes status was treated as T1D vs no T1D, the latter category including both T2D and non-diabetes cases). Concordance tables and associated validation metrics were computed with 95% confidence intervals: sensitivity (true positives/(true positives + false negatives)), specificity (true negatives/(true negatives + false positives)), positive predictive value (PPV: true positives/(true positives + false positives)), and negative predictive value (NPV: true negatives/(true negatives + false negatives)). Finally, analyses stratified by self-reported age at diabetes onset (including all self-reported non-diabetes cases in both strata of age at onset) were performed to assess the influence of age at onset on sensitivity and PPV.

Main validation analyses were bootstrapped in 1000 random subsamples in order to assess robustness (available in Supplementary Material S3). Exploratory validation analyses stratified by sex are available in (Supplementary Material S5 and S6). Potential bias between self-reported and register-classified diabetes duration across calendar year of diabetes onset was examined in hexagon-plots with smoothed LOESS regression lines with 95% confidence intervals (Supplementary Material S7).

The algorithms behind the diabetes classifiers were implemented in *R*, and all statistical analyses were performed using *R*[28] and the *epiR* package.[29] The implemented OSDC and RSCD algorithms are fully open-source, publicly available as part of the *osdc R*-package.[30]

Results

Diabetes Populations

Among the 29,391 individuals in the study population, self-reported diabetes prevalence was 2633 (9.0%), as 410 (1.4%) reported having T1D, and 2223 (7.6%) reported having T2D. OSDC-classified diabetes prevalence was 2733 (9.3%), with 336 (1.1%) classified as T1D and 2397 (8.2%) as T2D, while RSCD-classified diabetes prevalence was 2544 (8.7%), with 304 (1.0%) classified as T1D and 2240 (7.6%) as T2D. For diabetes of any type vs no diabetes, 2873 individuals were defined as having diabetes according to any of the three definitions, and 2421 individuals (84.3%) were concordantly defined as diabetes cases by all three definitions. Figure 3 shows Venn diagrams of concordance between the different diabetes definitions.

In both T1D and T2D, the mean age at onset was lowest in the self-reported data (T1D: 26.2 years, T2D: 51.6 years) and higher in the RSCD (T1D: 28.6 years, T2D: 52.3 years) and OSDC-classifications (T1D: 30.1 years, T2D: 53.5 years). In both T1D and T2D, the proportion of women in the self-reported diabetes population was lower than in the OSDC and RSCD cohorts (self-reported: 44.4% and 39.8% vs OSDC: 48.2% and 40.8% vs RSCD: 45.7% and 40.6%),

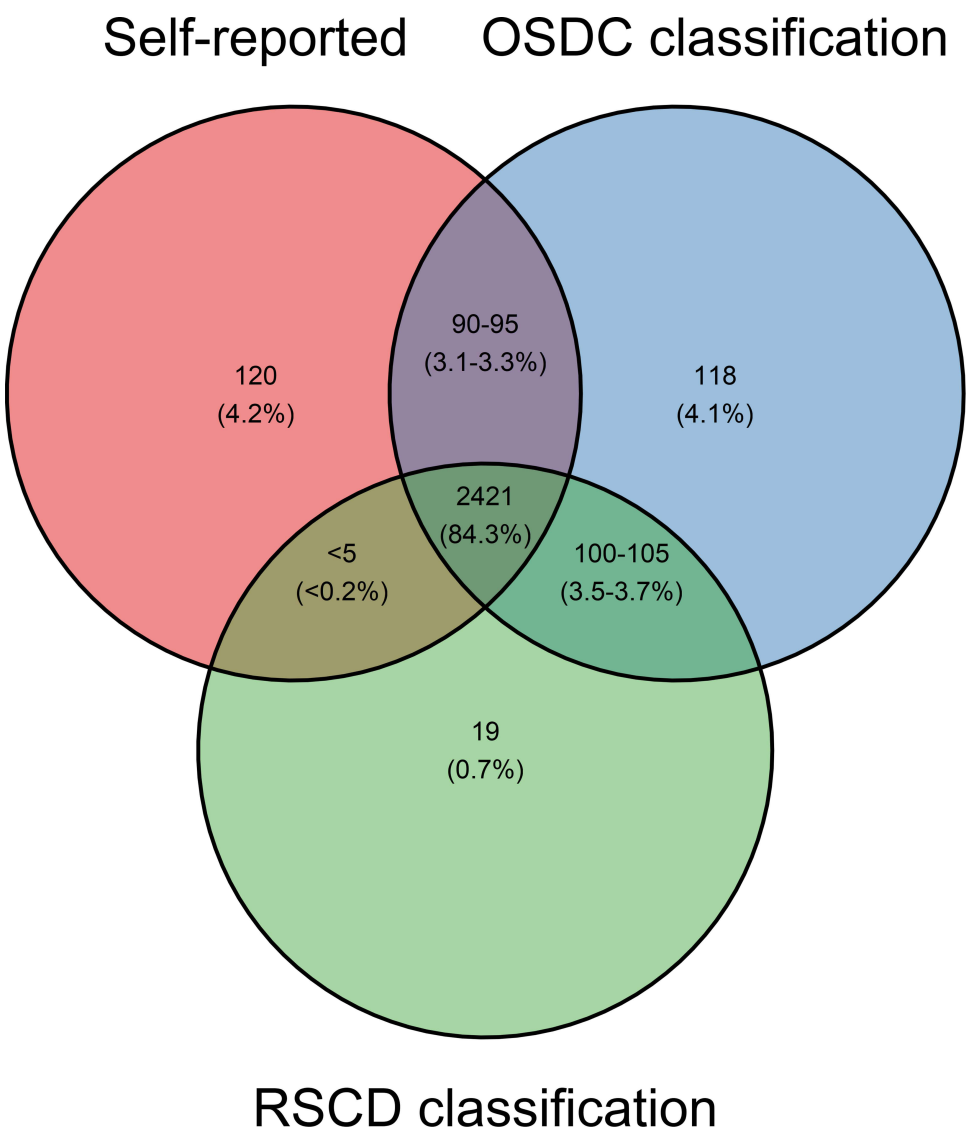

Figure 3 Concordance between individuals defined as having diabetes (any type). Three cells are censored in order to comply with statistical disclosure requirements of Statistics Denmark to protect individual-level data privacy.
Abbreviations: OSDC, Open-Source Diabetes Classifier; RSCD, Register for Selected Chronic Diseases.

and mean age was higher (self-reported: 49.9 and 62.2 vs OSDC: 48.4 and 61.7 vs RSCD: 48.0 and 61.4). In T1D, the proportion of migrants tended to be higher in the self-reported cohort than in the OSDC and RSCD cohorts (self-reported : 5.9% vs OSDC: 4.2% vs RSCD: 4.9%), while the proportion was lower in the self-reported T2D population (self-reported: 7.1% vs OSDC: 7.8% vs RSCD: 7.5%). Table 1 shows the characteristics of the study population according to self-reported, OSDC-classified, and RSCD-classified diabetes type (characteristics without subsampling in Supplementary Material S4).

Overall Validation Analyses

For T1D, sensitivity in OSDC was 0.773 (95% CI [0.730; 0.813]), which was higher than RSCD at 0.700 (95% CI [0.653; 0.744]). This difference persisted in NPV, which in OSDC was 0.997 (95% CI [0.996; 0.997]) vs 0.996 (95% CI [0.995; 0.996]) in RSCD. Specificity was practically identical in both classifiers at 0.999 (95% CI [0.999; 1.000]), and so was PPV, which in OSDC was 0.943 (95% CI [0.913; 0.966]) vs 0.944 (95% CI [0.912; 0.967]) in RSCD.

Table 1 Characteristics of Study Population by Diabetes Definition

Source of Diabetes Definition		T1D	T2D	No Diabetes
Characteristics				
Self-reported diabetes				
Diabetes type (self-reported)	N (%)	410 (1.4)	2223 (7.6)	26,758 (91.0)
Age at onset (self-reported)	Mean (SD)	26.2 (16.2)	51.6 (11.4)	
Sex: Female	N (%)	182 (44.4)	885 (39.8)	11,699 (43.7)
Age	Mean (SD)	49.9 (15.2)	62.2 (9.2)	59.9 (11.4)
Migrant origin	N (%)	24 (5.9)	157 (7.1)	1208 (4.5)
OSDC-classified diabetes				
Diabetes type (OSDC-classified)	N (%)	336 (1.1)	2397 (8.2)	26,658 (90.7)
Age at onset (OSDC-classified)	Mean (SD)	30.1 (13.2)	53.5 (10.0)	
Sex: Female	N (%)	162 (48.2)	977 (40.8)	11,627 (43.6)
Age	Mean (SD)	48.4 (15.2)	61.7 (9.6)	60.0 (11.4)
Migrant origin	N (%)	14 (4.2)	186 (7.8)	1189 (4.5)
RSCD-classified diabetes				
Diabetes type (RSCD-classified)	N (%)	304 (1.0)	2240 (7.6)	26,847 (91.3)
Age at onset (RSCD-classified)	Mean (SD)	28.6 (13.3)	52.3 (10.9)	
Sex: Female	N (%)	139 (45.7)	909 (40.6)	11,718 (43.6)
Age	Mean (SD)	48.0 (15.3)	61.4 (9.9)	60.0 (11.4)
Migrant origin	N (%)	15 (4.9)	169 (7.5)	1205 (4.5)

Abbreviations: T1D, type 1 diabetes; T2D, type 2 diabetes; OSDC, Open-Source Diabetes Classifier; RSCD, Register for Selected Chronic Diseases.

For T2D, sensitivity in OSDC was 0.944 (95% CI [0.933; 0.953]), which again outperformed RSCD at 0.905 (95% CI [0.892; 0.917]), and translated to a difference in NPV, where OSDC was 0.995 (95% CI [0.994; 0.996]) compared to RSCD's 0.992 (95% CI [0.991; 0.993]). Specificity was higher in RSCD at 0.992 (95% CI [0.990; 0.993]) vs OSDC's 0.989 (95% CI [0.988; 0.990]), with corresponding PPV in RSCD at 0.898 (95% CI [0.884; 0.910]) vs OSDC's 0.875 (95% CI [0.861; 0.888]).

All estimates were robust in Supplementary Bootstrapped Analyses (S3). Table 2 shows concordance tables and validation metrics of each register-based diabetes classifier for T1D and T2D.

Analyses Stratified by Age at Onset of Diabetes

In both classifiers, sensitivity of T1D-classification was much higher in individuals with diabetes onset before age 40 (OSDC: 0.884, 95% CI [0.844; 0.917], RSCD: 0.819, 95% CI [0.772; 0.859]) than in individuals with onset later in life

Table 2 Overall Validation Analyses of Type 1 Diabetes and Type 2 Diabetes

T1D				
OSDC	Survey: +T1D	Survey: -T1D	Total N	
OSDC: +T1D	317	19	336	PPV: 0.943 (0.913, 0.966)
OSDC: -T1D	93	28,962	29,055	NPV: 0.997 (0.996, 0.997)
Total N	410	28,981	29,391	
	Sensitivity: 0.773 (0.730, 0.813)	Specificity: 0.999 (0.999, 1.000)		
RSCD				
RSCD: +T1D	287	17	304	PPV: 0.944 (0.912, 0.967)
RSCD: -T1D	123	28,964	29,087	NPV: 0.996 (0.995, 0.996)
Total N	410	28,981	29,391	
	Sensitivity: 0.700 (0.653, 0.744)	Specificity: 0.999 (0.999, 1.000)		
T2D				
OSDC	Survey: +T2D	Survey: -T2D		
OSDC: +T2D	2098	299	2397	PPV: 0.875 (0.861, 0.888)
OSDC: -T2D	125	26,869	26,994	NPV: 0.995 (0.994, 0.996)
Total N	2223	27,168	29,391	
	Sensitivity: 0.944 (0.933, 0.953)	Specificity: 0.989 (0.988, 0.990)		
RSCD				
RSCD: +T2D	2011	229	2240	PPV: 0.898 (0.884, 0.910)
RSCD: -T2D	212	26,939	27,151	NPV: 0.992 (0.991, 0.993)
Total N	2223	27,168	29,391	
	Sensitivity: 0.905 (0.892, 0.917)	Specificity: 0.992 (0.990, 0.993)		

Notes: "-T1D" designates individuals with type 2 diabetes or no diabetes according to the source (classifier or survey), and "-T2D" designates individuals with type 1 diabetes or no diabetes.

Abbreviations: T1D, type 1 diabetes; T2D, type 2 diabetes; OSDC, Open-Source Diabetes Classifier; RSCD, Register for Selected Chronic Diseases; PPV, positive predictive value; NPV, negative predictive value.

(OSDC: 0.378, 95% CI [0.278; 0.486], RSCD: 0.278, 95% CI [0.189; 0.382]). PPV was also lower in T1D onset after age 40 (OSDC: 0.708, 95% CI [0.559; 0.830], RSCD: 0.658, 95% CI [0.486; 0.804]) than in cases with earlier onset (OSDC: 0.956, 95% CI [0.926; 0.976], RSCD: 0.960, 95% CI [0.929; 0.980]).

In T2D-classification, sensitivity was lower in both classifiers among those with onset before age 40 (OSDC: 0.863, 95% CI [0.814; 0.902], RSCD: 0.855, 95% CI [0.806; 0.896]) compared to those with onset later in life (OSDC: 0.954, 95% CI [0.944; 0.963], RSCD: 0.911, 95% CI [0.898; 0.923]). PPV was lower in those with onset before age 40 (OSDC: 0.471, 95% CI [0.425; 0.517], RSCD: 0.563, 95% CI [0.512; 0.613]) compared to those with onset later in life (OSDC:

Table 3 Sensitivity and Positive Predictive Value Stratified by Age at Onset of Diabetes

T1D			**T2D**		
Onset before age 40			**Onset before age 40**		
OSDC	Survey: +T1D	Survey: -T1D	OSDC	Survey: +T2D	Survey: -T2D
+T1D	283	13	+T2D	220	247
-T1D	37	27,000	-T2D	35	26,831
N	320	27,013	N	255	27,078
	Sensitivity: 0.884 (0.844, 0.917)	PPV: 0.956 (0.926, 0.976)		Sensitivity: 0.863 (0.814, 0.902)	PPV: 0.471 (0.425, 0.517)
RSCD	Survey: +T1D	Survey: -T1D	RSCD	Survey: +T2D	Survey: -T2D
+T1D	262	11	+T2D	218	169
-T1D	58	27,002	-T2D	37	26,909
N	320	27,013	N	255	27,078
	Sensitivity: 0.819 (0.772, 0.859)	PPV: 0.960 (0.929, 0.980)		Sensitivity: 0.855 (0.806, 0.896)	PPV: 0.563 (0.512, 0.613)
Onset after age 40			**Onset after age 40**		
OSDC	Survey: +T1D	Survey: -T1D	OSDC	Survey: +T2D	Survey: -T2D
+T1D	34	14	+T2D	1878	265
-T1D	56	28,712	-T2D	90	26,583
N	90	28,726	N	1968	26,848
	Sensitivity: 0.378 (0.278, 0.486)	PPV: 0.708 (0.559, 0.830)		Sensitivity: 0.954 (0.944, 0.963)	PPV: 0.876 (0.862, 0.890)
RSCD	Survey: +T1D	Survey: -T1D	RSCD	Survey: +T2D	Survey: -T2D
+T1D	25	13	+T2D	1793	175
-T1D	65	28,713	-T2D	175	26,673
N	90	28,726	N	1968	26,848
	Sensitivity: 0.278 (0.189, 0.382)	PPV: 0.658 (0.486, 0.804)		Sensitivity: 0.911 (0.898, 0.923)	PPV: 0.911 (0.898, 0.923)

Notes: "-T1D" designates individuals with type 2 diabetes or no diabetes according to the source (classifier or survey), and "-T2D" designates individuals with type 1 diabetes or no diabetes. Survey-reported non-diabetes cases were included in both strata of age at onset.

Abbreviations: T1D, type 1 diabetes; T2D, type 2 diabetes; OSDC, Open-Source Diabetes Classifier; RSCD, Register for Selected Chronic Diseases; PPV, positive predictive value.

0.876, 95% CI [0.862; 0.890], RSCD: 0.911, 95% CI [0.898; 0.923]). Table 3 shows concordance tables, sensitivity and PPV from analyses stratified by age at onset of diabetes.

Discussion

We validated two predefined algorithms classifying diabetes type in Danish register data against self-reported diabetes type in a Danish survey population. Overall, both classifiers performed excellent in terms of PPV in both T1D and T2D, as well as sensitivity in T2D classification (all estimates 0.875 and above), and had near-perfect accuracy in terms of specificity and NPV in both T1D and T2D (all estimates 0.989 and above). Both classifiers were unable to accurately classify diabetes type in individuals with T1D onset after age 40 and T2D onset before age 40.

The main strength of this real-world study is the large study population from a general population with data on diabetes type, which allows a complete validation including sensitivity and specificity. To limit the risk of circular bias,[31] we validated the register-based classifications against self-reported diabetes type, which is independent of the register data. Clinical audit of electronic patient records, an approach commonly used in validation studies,[32] was unsuited to capture a general population in our setting, as the vast majority of patients in Denmark – including T2D – are handled in the primary care sector, where electronic patient records are not available for research on the scale of this study.

Self-reported data on diabetes type and duration may contain inaccuracies, and an imperfect validation golden standard is a limitation of this study.[33] As the T1D population is much smaller than the T2D population, analyses of T1D would be particularly vulnerable to inaccuracies in the self-reported data. Since age at onset of diabetes is higher among individuals with T2D, we would expect the mean age at onset of diabetes in the T1D population to increase if the self-reported data were inaccurate. Bias from this drift would lower the sensitivity of the register-based classifiers, particularly among individuals with self-reported T1D onset later in life. While sensitivity was poorest in those with T1D onset after age 40, the distribution of self-reported age at onset of T1D was similar to previously reported distributions,[34] when accounting for differences in source populations (age at onset of T1D: 0–14 years: 28%, 15–39 years: 47%, 40–64 years: 23%, 65+ years: 2%). Thus, we found no evidence of a substantial proportion of misreported T2D cases in the self-reported T1D population. The poor performance of the register-based classifiers in subgroups with diabetes onset at atypical age is likely to reflect clinical uncertainty in these cases resulting in inaccurate type-specific diabetes diagnoses and uncharacteristic GLD purchase patterns,[35] which may lead the algorithms to misclassify. Sensitivity in T1D-classification has previously been reported to be highly age at onset-dependent,[11] and our findings indicate that this issue also extends to PPV and to T2D classification.

Although additional inaccuracies in the self-reported diabetes data cannot be ruled out, especially in subpopulations with lower health literacy, such as migrants,[36] we expect these inaccuracies to be minor and non-differential between the two register-based classifiers. Thus, any inaccuracies would result in *bias towards the mean*: an underestimation of performance in both classifiers, and attenuation of differences between them.[31] Indeed, a substantial proportion of self-reported T1D cases among migrants in our study may have been true T2D cases, as the proportion of self-reported T1D was higher in migrants than native Danes, which contradicts previous findings.[37,38]

Due to the survey-based nature of our study, selection bias cannot be ruled out, but several factors suggest that this bias was limited: In the *Health in Central Denmark* survey, response rates were high (>50%) in both the T1D, T2D and no-diabetes groups, and all groups shared similar non-response patterns.[20] In addition, the subsampling approach in our study compensated for the slightly lower response rate of the diabetes group compared to the non-diabetes group.

Sensitivity & NPV were higher in OSDC than in RSCD for both diabetes types. In T1D classification, these differences are attributable to differences in the algorithms, as both algorithms relied on the same data sources for identifying T1D (GLD use and type-specific diabetes diagnoses).[27] In T2D, this difference may be explained by the use of HbA1c data in OSDC, which enables inclusion of diabetes cases at the time of diagnosis, rather than requiring subsequent initiation of GLD treatment or hospitalization. Specificity and PPV in T2D classification were higher in RSCD than in OSCD, possibly explained by inclusion of milder cases in the latter, who may be less likely to correctly report having diabetes, eg, if an individual had never purchased GLD or been hospitalized in relation to the disease. Notably, the demographics of the register-classified diabetes populations differed, particularly in T1D, where OSDC's higher sensitivity in women (0.846 vs 0.725, Supplementary S5) resulted in a higher prevalence of women in the OSDC

population compared to RSCD. This is in line with a previous study comparing the Danish National Diabetes Register (which did not use HbA1c) against a local database containing HbA1c as an alternative inclusion criterion, which found higher prevalences of diabetes in the HbA1c-augmented definition, especially among women.[1] In terms of overall performance, both classifiers achieved excellent accuracy despite their different approaches to classifying diabetes type, handling potentially erroneous data, and censoring PCOS and GDM cases. This is perhaps a testament to the high quality of the underlying Danish register data, rather than the specific design of each algorithm – suggesting that other algorithms using similar data and censoring approaches[39] may yield comparable levels of overall performance.

To the best of our knowledge, this study is the first to validate the performance of a type-specific diabetes classifier in a general population setting of individuals with and without diabetes. Other studies have validated classifier performance in populations consisting only of individuals with diabetes, which fails to test the algorithm's ability to discern individuals with diabetes from those without, and overestimates the accuracy of the algorithm compared to its performance in a general population. In addition, both algorithms were pre-specified, as opposed to the data-driven, exploratory approaches used in other validation studies to design algorithms optimized for a particular dataset, which risks overfitting and overestimation of performance compared to a general population.[8,11,12] Despite these differences, the performance of both classifiers in our study was superior to the T1D-classification accuracy previously reported by studies in the United States and Hong Kong,[8,13] comparable to those reported in the United Kingdom,[14] but inferior to that reported by a study in Canada.[12]

While accurate classification of diabetes type is the most important task for a register-based classifier, the accuracy of the estimated age at onset of diabetes (or diabetes duration) is also important in diabetes epidemiology, but no prior studies reported the accuracy of diabetes onset classification. Our Supplementary Analyses (S7) showed that, compared to self-reported data, the register-based classifiers underestimated diabetes duration in many cases, eg T1D onset prior to prescription data becoming available in 1995 and T2D onset before 2010. However, as self-reported age at onset of diabetes may be susceptible to recall bias, validation studies with more robust measures of age at onset of diabetes are needed to clarify this issue.

Finally, the open-source implementation of the classifiers validated in this study allows researchers to easily utilize and modify the validated algorithms in future studies of type 1 and T2D,[40] particularly those based on Danish register data or countries with similar healthcare systems and register data infrastructure.

Conclusion

Using Danish health register data, the *Open-Source Diabetes Classifier* and the *Register for Selected Chronic Diseases* generate valid populations of T1D and T2D in a general population (PPVs from 87.5% to 94.4%, all NPVs above 98.9%), but sensitivity was substantially higher in OSDC compared to RSCD in T1D (77.3% vs 70.0%) and in T2D (94.4% vs 90.5%). Neither algorithm was able to accurately classify diabetes type in individuals with T1D onset after age 40 nor T2D onset before age 40, and results from register-based studies of these groups should be interpreted with caution. The validated classifiers provide a robust foundation for register-based studies, and their open-source nature makes them transparent and flexible tools for researchers.

Ethics Approval Statement

This study was approved by the Health in Central Denmark steering committee. The Health in Central Denmark project is registered in the Central Denmark Region internal register of research projects (reg. no. 1-16-02-165-20). Access to register data was provided and approved by the Danish Health Data Authority and Statistics Denmark. In Denmark, studies based entirely on survey and register data do not require specific ethical approval.

Acknowledgments

The authors are grateful to all participants of the *Health in Central Denmark* cohort for their contributions to the study. The authors are grateful to J. Støy and S.T. Knudsen from Steno Diabetes Center Aarhus for providing valuable clinical inputs to algorithm design during development of OSDC, and to the Clinical Epidemiology Research group at Steno Diabetes Center Copenhagen for insightful feedback on preliminary analyses.

Disclosure

Financial support was provided to AAI by a research training supplement grant from *Aarhus University*, Denmark, as well as unrestricted grants from the *Public Health in Central Denmark Region* foundation and *Steno Diabetes Center Aarhus*, Denmark. AS and LB are employed by *Steno Diabetes Center Aarhus*, Denmark, which is partially funded by an unrestricted donation from the *Novo Nordisk Foundation*. The authors report no other conflicts of interest in this work.

References

1. Nielsen AA, Christensen H, Lund ED, Christensen C, Brandslund I, Green A. Diabetes mortality differs between registers due to various disease definitions. *Dan Med J*. 2014;61(5):A4840.
2. Rawshani A, Landin-Olsson M, Svensson AM, et al. The incidence of diabetes among 0–34 year olds in Sweden: new data and better methods. *Diabetologia*. 2014;57(7):1375–1381. doi:10.1007/s00125-014-3225-9
3. Bak JCG, Serné EH, Kramer MHH, Nieuwdorp M, Verheugt CL. National diabetes registries: do they make a difference? *Acta Diabetol*. 2021;58 (3):267–278. doi:10.1007/s00592-020-01576-8
4. Hallgren Elfgren I-M, Grodzinsky E, Törnvall E. The Swedish National Diabetes Register in clinical practice and evaluation in primary health care. *Prim Health Care Res Dev*. 2016;17(6):549–558. doi:10.1017/S1463423616000098
5. Cooper JG, Thue G, Claudi T, Løvaas K, Carlsen S, Sandberg S. The Norwegian diabetes register for adults – an overview of the first years. *Norsk Epidemiol*. 2013;23(1). doi:10.5324/nje.v23i1.1599
6. Nishioka Y, Takeshita S, Kubo S, et al. Appropriate definition of diabetes using an administrative database: a cross-sectional cohort validation study. *J Diabetes Investig*. 2022;13(2):249–255. doi:10.1111/jdi.13641
7. Raebel MA, Schroeder EB, 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. Sentinel initiative; 2016.
8. Klompas M, Eggleston E, McVetta J, Lazarus R, Li L, Platt R. Automated detection and classification of type 1 versus type 2 diabetes using electronic health record data. *Diabetes Care*. 2013;36(4):914–921. doi:10.2337/dc12-0964
9. Schroeder EB, Donahoo WT, Goodrich GK, Raebel MA. Validation of an algorithm for identifying type 1 diabetes in adults based on electronic health record data. *Pharmacoepidemiol Drug Saf*. 2018;27(10):1053–1059. doi:10.1002/pds.4377
10. Lo-Ciganic W, Zgibor JC, Ruppert K, Arena VC, Stone RA. Identifying type 1 and type 2 diabetic cases using administrative data: a tree-structured model. *J Diabetes Sci Technol*. 2011;5(3):486–493. doi:10.1177/193229681100500303
11. Ke C, Stukel TA, Luk A, et al. Development and validation of algorithms to classify type 1 and 2 diabetes according to age at diagnosis using electronic health records. *BMC Med Res Methodol*. 2020;20(1):35. doi:10.1186/s12874-020-00921-3
12. Weisman A, Tu K, Young J, et al. Validation of a type 1 diabetes algorithm using electronic medical records and administrative healthcare data to study the population incidence and prevalence of type 1 diabetes in Ontario, Canada. *BMJ Open Diabetes Res Care*. 2020;8(1):e001224. doi:10.1136/bmjdrc-2020-001224
13. Lynam A, McDonald T, Hill A, et al. Development and validation of multivariable clinical diagnostic models to identify type 1 diabetes requiring rapid insulin therapy in adults aged 18–50 years. *BMJ Open*. 2019;9(9):e031586. doi:10.1136/bmjopen-2019-031586
14. Thomas NJ, McGovern A, Young KG, et al. Identifying type 1 and 2 diabetes in research datasets where classification biomarkers are unavailable: assessing the accuracy of published approaches. *J Clin Epidemiol*. 2022;153:34–44. doi:10.1016/j.jclinepi.2022.10.022
15. Kristensen JK, Drivsholm TB, Carstensen B, Steding-Jensen M, Green A. Validering af metoder til identifikation af erkendt diabetes på basis af administrative sundhedsregistre [Validation of methods to identify known diabetes on the basis of health registers]. *Ugeskr Laeger*. 2007;169 (18):1687–1692. Danish.
16. Carstensen B, Kristensen JK, Marcussen MM, Borch-Johnsen K. The national diabetes register. *Scand J Public Health*. 2011;39(7 Suppl):58–61. doi:10.1177/1403494811404278
17. Green A, Sortsø C, Jensen PB, Emneus M. Validation of the Danish national diabetes register. *Clin Epidemiol*. 2014;7:5–15. doi:10.2147/CLEP.S72768
18. Jørgensen ME, Kristensen JK, Reventlov Husted G, Cerqueira C, Rossing P. The Danish adult diabetes registry. *Clin Epidemiol*. 2016;8:429–434. doi:10.2147/CLEP.S99518
19. The Danish Health Data Authority. The register of selected chronic diseases; 2014. Available from: https://www.esundhed.dk/Dokumentation/DocumentationExtended?id=29. Accessed November 25, 2022.
20. Bjerg L, Dalsgaard E-M, Norman K, Isaksen AA, Sandbæk A. Cohort profile: health in Central Denmark (HICD) cohort - a register-based questionnaire survey on diabetes and related complications in the Central Denmark Region. *BMJ Open*. 2022;12(7):e060410. doi:10.1136/bmjopen-2021-060410
21. Schmidt M, Schmidt SAJ, Adelborg K, et al. The Danish health care system and epidemiological research: from health care contacts to database records. *Clin Epidemiol*. 2019;11:563–591. doi:10.2147/CLEP.S179083
22. Schmidt M, Pedersen L, Sørensen HT. The Danish Civil Registration System as a tool in epidemiology. *Eur J Epidemiol*. 2014;29(8):541–549. doi:10.1007/s10654-014-9930-3
23. Schmidt M, Schmidt SA, Sandegaard JL, Ehrenstein V, Pedersen L, Sørensen HT. The Danish National Patient Registry: a review of content, data quality, and research potential. *Clin Epidemiol*. 2015;7:449–490. doi:10.2147/CLEP.S91125
24. Andersen JS, Olivarius Nde F, Krasnik A. The Danish National Health Service Register. *Scand J Public Health*. 2011;39(7 Suppl):34–37. doi:10.1177/1403494810394718
25. Pottegård A, Schmidt SAJ, Wallach-Kildemoes H, Sørensen HT, Hallas J, Schmidt M. Data resource profile: the Danish National Prescription Registry. *Int J Epidemiol*. 2017;46(3):798–798f. doi:10.1093/ije/dyw213
26. Wertheimer AI. The defined daily dose system (DDD) for drug utilization review. *Hosp Pharm*. 1986;21(3):233–234, 239–241, 258.

27. The Danish Health Data Authority. Algoritmer for udvalgte kroniske sygdomme og svære psykiske lidelser; 2016. Available from: https://www.esundhed.dk/-/media/Files/Dokumentation/Register-for-Udvalgte-Kroniske-Sygdomme-og-Svaere-Psykiske-Lidelser/29_Algoritmer_for_RUKS.ashx. Accessed November 25, 2022.
28. R Core Team R. *R: A Language and Environment for Statistical Computing [Computer Program]. Version 4.1.3: R Foundation for Statistical Computing*. Vienna, Austria: R Core Team R; 2022.
29. epiR: tools for the analysis of epidemiological data [computer program]. Version 2.0.52; 2022. Available from: https://CRAN.R-project.org/package=epiR. Accessed April 19, 2023.
30. Isaksen AA. osdc - an R package for classifying diabetes in Danish registers. Available from: https://github.com/Aastedet/osdc. Accessed November 25, 2022.
31. Fox MP, Lash TL, Bodnar LM. Common misconceptions about validation studies. *Int J Epidemiol*. 2020;49(4):1392–1396. doi:10.1093/ije/dyaa090
32. Nissen F, Quint JK, Morales DR, Douglas IJ. How to validate a diagnosis recorded in electronic health records. *Breathe*. 2019;15(1):64–68. doi:10.1183/20734735.0344-2018
33. Wacholder S, Armstrong B, Hartge P. Validation studies using an alloyed gold standard. *Am J Epidemiol*. 1993;137(11):1251–1258. doi:10.1093/oxfordjournals.aje.a116627
34. Green A, Hede SM, Patterson CC, et al. Type 1 diabetes in 2017: global estimates of incident and prevalent cases in children and adults. *Diabetologia*. 2021;64(12):2741–2750. doi:10.1007/s00125-021-05571-8
35. Thomas NJ, Lynam AL, Hill AV, et al. Type 1 diabetes defined by severe insulin deficiency occurs after 30 years of age and is commonly treated as type 2 diabetes. *Diabetologia*. 2019;62(7):1167–1172. doi:10.1007/s00125-019-4863-8
36. Pettersson S, Hadziabdic E, Marklund H, Hjelm K. Lower knowledge about diabetes among foreign-born compared to Swedish-born persons with diabetes-a descriptive study. *Nurs Open*. 2019;6(2):367–376. doi:10.1002/nop2.217
37. Hussen HI, Yang D, Cnattingius S, Moradi T. Type I diabetes among children and young adults: the role of country of birth, socioeconomic position and sex. *Pediatr Diabetes*. 2013;14(2):138–148. doi:10.1111/j.1399-5448.2012.00904.x
38. Neu A, Willasch A, Ehehalt S, Kehrer M, Hub R, Ranke MB. Diabetes incidence in children of different nationalities: an epidemiological approach to the pathogenesis of diabetes. *Diabetologia*. 2001;44(Suppl 3):B21–26. doi:10.1007/PL00002948
39. Carstensen B, Rønn PF, Jørgensen ME. Prevalence, incidence and mortality of type 1 and type 2 diabetes in Denmark 1996–2016. *BMJ Open Diabetes Res Care*. 2020;8(1):e001071. doi:10.1136/bmjdrc-2019-001071
40. Marszalek RT, Flintoft L. Being open: our policy on source code. *Genome Biol*. 2016;17(1):172. doi:10.1186/s13059-016-1040-y