Primary Care Diabetes xxx (2020) xxx–xxx

Contents lists available at ScienceDirect

Primary Care Diabetes

journal homepage: http://www.elsevier.com/locate/pcd

Original research

Coded diagnoses from general practice electronic health records are a feasible and valid alternative to self-report to define diabetes cases in research

A.W. de Boer[a,b,*], J.W. Blom[b], M.W.M. de Waal[b], R.C.A. Rippe[c], E.J.P. de Koning[d], I.M. Jazet[d], F.R. Rosendaal[a], M. den Heijer[a,e], M.E. Numans[b], R. de Mutsert[a]

[a] *Department of Clinical Epidemiology, Leiden University Medical Center (LUMC), Leiden, The Netherlands*
[b] *Department of Public Health and Primary Care, LUMC, Leiden, The Netherlands*
[c] *Centre for Child and Family Science, Leiden University, Leiden, The Netherlands*
[d] *Department of Internal Medicine, LUMC, Leiden, The Netherlands*
[e] *Department of Internal Medicine, VU Medical Center, Amsterdam, The Netherlands*

ARTICLE INFO

Article history:
Received 2 January 2020
Received in revised form 17 August 2020
Accepted 18 August 2020
Available online xxx

Keywords:
Diabetes mellitus
General practice electronic health records
Epidemiology
Self-report
Measures of agreement
General practice

ABSTRACT

Aims: To examine the feasibility and validity of obtaining International Classification of Primary Care (ICPC)-coded diagnoses of diabetes mellitus (DM) from general practice electronic health records for case definition in epidemiological studies, as alternatives to self-reported DM.
Methods: The Netherlands Epidemiology of Obesity study is a population-based cohort study of 6671 persons aged 45–65 years at baseline, included between 2008–2012. Data from electronic health records were collected between 2012–2014. We defined a reference standard using diagnoses, prescriptions and consultation notes and investigated its agreement with ICPC-coded diagnoses of DM and self-reported DM.
Results: After a median follow-up of 1.8 years, data from 6442 (97%) participants were collected. With the reference standard, 506 participants (79/1000 person-years) were classified with prevalent DM at baseline and 131 participants (11/1000 person-years) were classified with incident DM during follow-up. The agreement of prevalent DM between self-report and the reference standard was 98% (kappa 0.86), the agreement between ICPC-coded diagnoses and the reference standard was 99% (kappa 0.95). The agreement of incident DM between ICPC-coded diagnoses and the reference standard was >99% (kappa 0.92).
Conclusions: ICPC-coded diagnoses of DM from general practice electronic health records are a feasible and valid alternative to self-reported diagnoses of DM.

1. Introduction

Accurate information on the exposure, confounding factors and participants' health outcomes is crucial in epidemiological and clinical research. Methods such as questionnaires, medical registries and interviews can be used to define the case status. Self-report is a method that is often used for the definition of type 2 diabetes mellitus (DM) cases in research [1–3]. Previous studies have reported a high agreement between self-reported diagnoses of DM and medical records (kappa ranging between 0.71 and 0.92) [4–8]. However, the suboptimal response rate and the absence of the exact date of diagnosis are disadvantages of self-reporting [9].

In the Netherlands, type 2 DM is mainly diagnosed and treated in general practice. General practitioners (GPs) code health problems in a patient's electronic health record using the International Classification of Primary Care (ICPC) [10]. Therefore, an alternative source to obtain information about diagnoses of DM is by asking the GP or by linking with the general practice electronic health records. In previous cohort studies, questionnaires were sent to GPs to obtain information about the diagnoses of DM [3,11,12]. For example, in a Dutch cohort study, two out of three ascertained DM cases (via self-report, hospital discharge diagnoses or urinary glucose strip) were confirmed to have been diagnosed with DM, as reported by their GP or pharmacist on a questionnaire [3]. Case definition using ICPC-coded diagnoses from the general practice electronic health records

* Corresponding author at: Department of Clinical Epidemiology, LUMC, PO Box 9600, 2300 RC Leiden, The Netherlands.
E-mail address: a.w.de_boer@lumc.nl (A.W. de Boer).

https://doi.org/10.1016/j.pcd.2020.08.011
1751-9918/

may be even more accurate than a questionnaire among GPs. In addition, compared with self-reported diagnoses, using ICPC-coded diagnoses from electronic health records easily provides the health information of more participants, and more detailed information, like the exact date of diagnosis. Whereas there has been an increasing number of studies using electronic medical records to collect patient information [13], for information on diagnoses of disease commonly medical records from hospitals are used, using the International Classification of Diseases, Ninth Revision (ICD-9) codes. However, the feasibility and validity of the definition of DM cases using ICPC-coded diagnoses from electronic health records in primary care has not been investigated previously. It is unknown to what extent GPs code health problems accurately.

In this study, we examined the feasibility of obtaining ICPC-coded diagnoses from the general practice electronic health records. Next, we aimed to examine the validity of ICPC-coded diagnoses of DM from electronic health records as an alternative to self-reported DM in the Netherlands Epidemiology of Obesity (NEO) study, a prospective cohort study of 6671 persons aged 45–65 years at baseline. To that extent, we compared the ICPC-coded diagnoses of DM with the self-reported DM and with a reference standard based on all available information in the electronic health records.

2. Methods

2.1. Study design and study population

The NEO study is a population-based prospective cohort study of 6671 persons aged 45–65 years with an oversampling of participants with a body mass index (BMI) ≥ 27 kg/m^2. Detailed information about the study design and data collection has been described elsewhere [14]. Participants with a self-reported BMI ≥ 27 kg/m^2 were recruited between 2008 and 2012 from the greater area of Leiden in the Netherlands through GPs, municipal registers and advertisements. In one municipality (Leiderdorp), all inhabitants aged 45–65 years were invited irrespective of their BMI, allowing for a reference distribution of BMI.

Prior to the baseline visit, participants completed a questionnaire at home including questions about demography, lifestyle and clinical information. The participants were asked to bring all medication they were using in the month preceding the study visit to the NEO study site. Names and dosages of all medication were recorded by trained staff. During the baseline visit at the NEO study centre of the Leiden University Medical Center (LUMC), participants underwent an extensive physical examination, including fasting blood sampling. Plasma concentrations of glucose were determined in the central clinical chemistry laboratory of the LUMC. Within two weeks after the NEO study visit, the participants received a letter with several test results, including their fasting plasma glucose concentration and the upper limit of a normal fasting glucose concentration of 7 mmol/L. Body weight and height were measured during the study visit with a calibrated scale and a vertically fixed, calibrated tape measure during the study visit. The trained staff reported the height in cm, body weight was rounded to 100 g and one kilogram was subtracted to correct for the weight of clothing. BMI was calculated by dividing the weight (in kg) by the square of the height (in metres).

Between April 2012 and November 2014, the GPs of the participants were contacted and visited to extract health information from the electronic health records of the participants. The data was extracted from 2008 until the date of extraction including the ICPC-coded lifetime medical history and merged into one database (GP database). Detailed information of the methods of data extraction from electronic health records is provided in Appendix A.

The study was approved by the medical ethics committee of the LUMC and all participants gave written informed consent for participation in the study and for obtaining medical information from their GP or medical specialists during follow-up of the study. All researchers with access to GP information signed a confidentiality agreement.

2.2. DM at baseline by self-report

Prevalent DM by self-report was defined as a self-reported medical history of DM, type 1 or type 2, on the questionnaire or the use of glucose lowering medication (oral or insulin) in the month preceding the baseline visit.

2.3. ICPC-coded DM diagnoses from general practice electronic health records

ICPC-coded diagnoses of DM and the corresponding dates were extracted from the medical history in the GP database. Diagnoses are coded by GPs in the primary care ICT system according to ICPC version 1 [10]. The DM diagnosis is coded with code T90, T90.1 or T90.2. The index date was defined as the date of diagnosis. Coded diagnoses with an index date before the baseline visit were defined as prevalent ICPC-coded diagnoses of DM, while coded diagnoses with an index date at or after the baseline visit were defined as incident ICPC-coded DM diagnoses.

2.4. Reference standard of diagnosed DM

To evaluate the validity of self-reported DM and ICPC-coded diagnoses for the definition of DM cases, a reference standard was developed by the Diabetes adjudication committee of the NEO study, using all data from the GP database. The Diabetes adjudication committee was composed of GPs, endocrinologists, epidemiologists and data-managers, complemented with clinicians from the diabetes work package within the NEO study (Appendix B). The committee defined the reference standard of the diagnosis DM as having one of the following: (1) a correctly ICPC-coded diagnosis of DM or ICPC-coded consultation note for DM; or (2) a prescription of glucose-lowering medication; or (3) a strong indication for the diagnosis of DM by screening keywords in the GP database.

First, the GP database was searched for coded diagnoses and coded consultations with ICPC-code T90 and the corresponding date. To verify whether the diagnoses were correctly coded, the corresponding descriptions were screened for conflicting descriptions. In the description of an ICPC-code, the general practitioners describe the diagnosis in their own words. The extracted data of all participants with a conflicting description were read and recoded when there was no diagnosis of DM. For example, a participant with the diagnosis 'Uterus extirpation 2002' coded with ICPC-code T90 (DM) rather than an ICPC-code in chapter X (Female genital system and breast), without a prescription of glucose-lowering medication or DM consultations was recoded as no DM diagnosis. In addition, temporary steroid-induced DM in the medical history was classified as no diagnosis of DM.

Second, the GP database was searched for prescriptions of glucose-lowering medication and the corresponding date, often registered according to the Anatomical Therapeutic Chemical (ATC) codes listed under A10 (Drugs used in diabetes) [15]. Prescriptions without an ATC code were screened by keywords of the various types of glucose-lowering medication (Appendix C).

Third, of the remaining participants without an ICPC-coded diagnosis or consultation and without a prescription for glucose-lowering medication, the medical history and the consultation notes were screened by the keywords 'DM', 'diab', 'gluc', 'suiker'.

Table 1
Baseline characteristics of the participants of the Netherlands Epidemiology of Obesity study, aged 45 to 65 years with an oversampling of body mass index ≥27 kg/m^2 (n = 6671).

	N	Characteristics
Age (years)	6671	56 (6)
Sex (men)	6671	3156 (47)
Body mass index (kg/m^2)	6671	30.1 (4.9)
Fasting plasma glucose (mmol/L)	6617	5.7 (1.1)
Self-reported diabetes	6654	459 (7)
Glucose-lowering therapy	6671	356 (5)

Data expressed as the mean (SD) or number (percentage).

When a keyword was detected, the extracted data were read and coded using a decision rule developed by the Diabetes adjudication committee to code the data consistently. When there was a strong indication for the diagnosis of DM, the data was coded as DM with the corresponding date. A strong indication was defined as a written diagnosis of DM or DM consultations.

The index date was defined as the first date of an ICPC-coded diagnosis, ICPC-coded consultation notes, prescription or strong indication for the diagnosis DM. A diagnosis of DM according to the reference standard with an index date before the date of the baseline visit was defined as prevalent DM. A diagnosis of DM according to the reference standard with an index date at or after the date of the baseline visit was defined as incident DM.

2.5. Statistical analysis

Baseline characteristics of the total population were expressed as the mean (SD), or number (percentage). To examine the feasibility of data extraction from electronic health records, we calculated the proportion of participants who gave informed consent to extract data and the proportion of participants for whom data was obtained. We also determined the time and costs related to the extraction of data from electronic health records.

We estimated the prevalence and incidence of DM for each case definition, as described above, while excluding those with missing data on self-reported DM or without health information from the electronic health records. To evaluate the validity of the case definitions, we estimated sensitivity, specificity, positive predictive value, negative predictive value, likelihood ratios and kappa statistics. The strength of the agreement was classified as almost perfect for kappa values of more than 0.80 [16]. For prevalent diagnosed DM, we compared two case definitions with the reference standard, namely, self-report and ICPC-coded diagnoses. For incident diagnoses of DM, we compared ICPC-coded diagnoses with the reference standard.

For all analyses, STATA statistical software (Statacorp, College Station, TX), version 14 was used.

3. Results

In total, 6671 persons have been included in the NEO study between September 2008 and October 2012. The baseline characteristics of the participants are shown in Table 1.

3.1. Feasibility of data extraction from general practice electronic health records

Of all participants, 6652 (99.7%) participants gave informed consent for the collection of medical information. Of these, 6622 participants reported the contact details of their current GP when they made the appointment for the baseline visit. Between April 2012 and November 2014, 352 general practices received a letter with a request to extract data from electronic health records. In 264 general practices, data were extracted from twelve different ICT systems. In addition, 52 general practices gave written information about diagnoses and prescriptions. Of 180 participants, no information was obtained due to non-response of the GP (11 participants of 11 general practices), a lack of permission from the GP (41 participants of 25 general practices), the participant was not registered in the reported general practice (120 participants), or death (8 participants).

Table 2
Feasibility of data extraction from general practice electronic health records in the Netherlands Epidemiology of Obesity study (n = 6671).

Informed consent to collect data, n (%)	6652 (>99)
Participants of whom data was obtained, n (%)	6442 (97)
Obtained unique ICPC-codes, n	1230
Time to extract data[a], hours	2850
Time to process data[b], hours	2000
Costs to collect data, euro's	148,800

Abbreviations: ICPC, International Classification of Primary Care.

[a] Contact with general practitioners, preparation of data extraction, travel to the general practices, data extraction.

[b] Building a database, case definition.

In total, health information was obtained from 6442 (97%) participants, after a median follow-up of 1.8 years (interquartile range 1.2–3.1) after baseline. The total costs of data extraction and processing was estimated at €148,800 and the total time was estimated at 4850 h (Table 2). When only coded diagnoses were extracted and processed, the time and costs were estimated to be 25% lower. A detailed overview of the costs and time is shown in Appendix D. In total, 1230 different ICPC-codes *(of a total of ∼1300 existing ICPC-codes)* were registered in the obtained information from the electronic health records; 742 diseases/disorders, 404 complaints/symptoms and 84 process codes (e.g., preventive procedures, administration, referrals).

3.2. Prevalence of DM at baseline by self-report

Of the 6654 (>99%) participants who answered the questions at baseline about a medical history of DM, 459 (7%) participants reported to be diagnosed with DM type 1 or type 2. All participants brought their medication to the study visit and 356 (5%) participants were using glucose-lowering medication. Of the participants with self-reported DM, 106 participants were not using glucose lowering medication. One participant was using glucose-lowering medication, but did not report being diagnosed with DM. Thus, according to the definition, in total 460 participants had DM by self-report at baseline.

3.3. Contribution of the separate criteria of the reference standard for diagnosis of DM

According to the reference standard in the total population, combining prevalent and incident DM, 586 participants had a ICPC-coded diagnosis of DM. Of these participants, one participant had an incorrect ICPC-coded diagnosis of DM. In addition to these 585 diagnoses, 39 participants had an ICPC-coded consultation note for DM, 10 participants had a prescription of glucose-lowering medication and 3 participants a strong indication for the diagnosis of DM by screening keywords. In total, 637 participants were defined as having a diagnosis DM with the reference standard, of whom 506 prevalent and 131 incident DM.

3.4. Prevalence and incidence of DM according to the case definitions

Table 3 shows the case status definitions of prevalent diagnosed DM according to each case definition, including the overlap

Table 3
Prevalent diabetes diagnosed according to the case definitions in 6671 participants of the Netherlands Epidemiology of Obesity study.

		Self-report			ICPC-coded diagnosis		
		No diabetes	Diabetes	Missing	No diabetes	Diabetes	No data extraction
Reference standard	No diabetes	5893	31	12	5935	1	0
	Diabetes	89	413	4	46	460	0
	No data extraction	212	16	1	0	0	229
ICPC-coded diagnosis	No diabetes	5919	50	12			
	Diabetes	63	394	4			
	No data extraction	212	16	1			

Abbreviations: ICPC, International Classification of Primary Care.

Table 4
Prevalent and incident diagnosed diabetes cases according to the case definitions in participants of the Netherlands Epidemiology of Obesity study.

Definition of diabetes cases	Study population (N)	Prevalent diabetes cases (N)	Prevalence (per 1000 persons)	Incident diabetes cases (N)	Incidence rate (per 1000 person-years)
Self-report	6654	460	69		
ICPC-coded diagnoses	6442	461	72	125	11
Reference standard	6442	506	79	131	11

Abbreviations: ICPC, International Classification of Primary Care.

between the definitions. With self-report, 460 (69 per 1000) participants were defined as having prevalent diagnosed DM at baseline. With ICPC-coded diagnoses, 461 (72 per 1000) participants were defined as having prevalent DM at baseline. With the reference standard, 506 (79 per 1000) participants were defined as having prevalent DM at baseline.

After excluding participants with prevalent ICPC-coded DM at baseline, 5981 participants were at risk of DM. During a total of 11,880 person-years of follow-up, 125 (21 per 1000) participants were defined as having incident DM, giving an overall incidence of 11 per 1000 person-years. With the reference standard, 5936 participants were at risk of DM after excluding participants with prevalent DM at baseline. During a total of 11,777 person-years of follow-up, 131 (22 per 1000) participants were defined as having incident DM, giving an overall incidence of DM of 11 per 1000 person-years. Table 4 shows prevalent and incident DM cases according to each case definition.

3.5. *Measures of agreement between the case definitions*

All definitions of diagnosed DM showed an almost perfect agreement (kappa > 0.8) with the reference standard (Table 5). For prevalent diagnosed DM, the agreement between self-report and the reference standard was 98% (kappa 0.86) and the agreement between ICPC-coded diagnoses and the reference standard was 99% (kappa 0.95). The agreement between self-report and ICPC-coded diagnoses for prevalent diagnosed DM was 98% (kappa 0.87). Of the 103 participants with self-reported DM who did not use glucose lowering medication, 70 participants (68%) had an ICPC-coded diagnosis for prevalent diagnosed DM. For incident diagnosed DM, the agreement between ICPC-coded diagnoses of incident DM and the reference standard of incident DM was 99% (kappa 0.92).

4. Discussion

In this study, we aimed to examine ICPC-coded diagnoses of DM from general practice electronic health records as an alternative to self-reported DM for the definition of case status in epidemiological studies. Both self-report and ICPC-coded diagnoses from electronic health records had an excellent agreement with the reference standard for the definition of DM diagnosis, using all information from general practice electronic health records. Case definition with information from general practice electronic health records provided a high follow-up rate of 97% and detailed health information was obtained.

In our study, we found a kappa value of 0.86 for self-reported DM compared with the reference standard. This is in line with previous studies that reported kappa values of 0.71–0.92 for self-reported DM compared with diagnoses in hospital medical records [4–8]. We observed that the agreement between ICPC-coded diagnosis and the reference standard was highest, with a kappa value of 0.95 for prevalent DM and 0.92 for incident DM. This means that GPs code the diagnosis DM accurately in the electronic health records and this finding supports our hypothesis that extracting ICPC-coded diagnoses from electronic health records in the Netherlands is a valid and better alternative to self-reported DM.

A strength of this study is the availability of information from general practice electronic health records of the majority of the participants of the NEO study to make a reference standard of DM. In addition to comparing the feasibility and validity of ICPC-coded diagnosis from electronic health records with a reference standard, we also compared the results with DM by self-report, one of the most commonly used methods for DM diagnosis definitions in epidemiological studies.

According to the World Health Organization DM is diagnosed in case of a fasting plasma glucose ≥7.0 mmol/L on two different occasions, a history of diabetes diagnosis, or use of insulin or oral glucose-lowering medication [17,18]. However, how to obtain the medical history of DM or the use of medication are not specified. We used a reference standard developed by the Diabetes adjudication committee of the NEO study using all information from the GP medical records, including diagnoses, consultation notes, and prescriptions. We assumed that the GP based the DM diagnosis on the guidelines including laboratory tests, but we were not able to include laboratory test results in the reference standard because this information was not completely available in the electronic health records, or could not be extracted from certain GP systems. Any extra detected cases based on laboratory results only would have reduced the agreement between the case definitions.

Because DM type 2 is mainly diagnosed and treated in general practice in the Netherlands, the reference standard reflects the known diagnoses of DM in the general population. In contrast to the known diagnoses from the GP data, we were not able to detect unknown and undiagnosed DM in our population with the data collected in the NEO study, because the fasting plasma glucose concentrations were only measured once at baseline, whereas

Table 5
Measures of agreement and 95% confidence interval between the reference standard and the case definitions for the definition of diagnosed diabetes in participants of the Netherlands Epidemiology of Obesity study.

	Prevalent diagnosed diabetes		Incident diagnosed diabetes
	Self-report (n = 6426)	ICPC-coded diagnoses (n = 6442)	ICPC-coded diagnoses (n = 5935)
Kappa	0.86 (0.84–0.89)	0.95 (0.93–0.96)	0.92 (0.88–0.96)
Sensitivity	82 (79–86)	91 (88–93)	86 (79–92)
Specificity	99.5 (99.3–99.6)	100 (99.9–100)	100 (99.9–100)
Positive predictive value	93 (90–95)	99.8 (98.8–100)	99.1 (95.2–100)
Negative predictive value	98.5 (98.2–98.8)	99.2 (99.0–99.4)	99.7 (99.5–99.8)
Likelihood ratio of positive test	157 (110–224)	5396 (760–3.8 * 10^4)	5001 (704–3.6 * 10^4)
Likelihood ratio of negative test	0.18 (0.15–0.22)	0.09 (0.07–0.12)	0.14 (0.09–0.21)

Abbreviations: ICPC, International Classification of Primary Care.

a new diagnosis of DM requires confirmatory symptoms or laboratory tests on another day [18]. Nevertheless, we showed that using ICPC-coded diagnoses is a valid method for the case definition of known DM in the population.

We acknowledge that including ICPC-codes in both the reference standard and the ICPC-coded method, inherently results in agreement between the two. Nevertheless, our goal was to investigate if ICPC-codes only would be sufficient for use in research. Because of the excellent agreement between the two we are confident to conclude that it is not necessary to use of all information from the electronic health records and that ICPC-codes suffice to identify DM cases for outcome research.

Our study population was middle-aged at baseline (45–65 years at baseline) with an oversampling of patients with a BMI > 27 kg/m^2. Patients of middle age, and those with a high BMI may consult their GP more frequently than younger and leaner patients. This could lead to a more timely detection of DM and a more accurately registration of the diagnosis in the electronic medical records. This may have led to an improved agreement between ICPC-coded diagnoses and the reference standard in our population. Nevertheless, because type 2 DM typically develops at middle-age and at high BMI, our results pertain the majority of the DM patients detected in primary care.

After the baseline study visit, all participants received a letter informing them about their fasting plasma glucose concentration and the upper limit of a normal fasting plasma glucose concentration of 7.0 mmol/L. We did not inform the GPs of the participants and do not know if participants with a fasting plasma glucose concentration >7.0 mmol/L consulted their GP with this test result. In a previous study, the onset of DM was estimated to occur 4–7 years prior to its clinical diagnosis [19]. Some of the incident DM cases in our study may therefore have been detected early because of the test results. When we considered participants without an ICPC-coded diagnosis of DM, but with a fasting plasma glucose concentration >7.0 mmol/L not at risk of developing DM during follow-up, the incidence of DM using the reference standard was 6 per 1000 person-years (data not shown), compared with 11 per 1000 person-years when these participants were considered at risk of developing DM. The true incidence of DM in a median follow-up of 1.8 years most likely lies between these two estimates. In- or excluding these cases had no influence on the agreement between the ICPC-coded diagnoses and the reference standard. In the future, prospective analyses of the NEO study with DM as an outcome variable and a longer follow-up time, the effect of this potential early detection because of participation in the study will become negligible.

In this study, both participants with DM type 1 and participants with DM type 2 are included in the case definitions of DM diagnoses. We did not make separate subgroups in the case definitions of prevalent DM because both a medical history of DM type 1 and DM type 2 at baseline will be excluded in future prospective analyses on incident DM. Seven percent of all patients with DM aged between 50 and 59 years in the Netherlands have been diagnosed with DM type 1 [20]. With regard to incident DM, we assume that all participants with incident DM are diagnosed with DM type 2, because DM type 1 is usually diagnosed before the age of 30 years [21].

An advantage of using electronic health records for the follow-up of cohort studies is the high follow-up rate, 97% in our study. The response rate to a survey, or a follow-up questionnaire may be substantially lower. For example, the response rate of a follow-up questionnaire that we sent to the NEO participants in 2013 was 78%. A higher follow-up rate increases the statistical power of the study and selection bias is less likely to occur [22]. Moreover, the reasons why we were unable to obtain health information from electronic health records are likely to be unrelated to the diagnosis DM. For future data extractions, we aim to trace the current GP of the participants with missing GP information. The high follow-up rate in combination with the excellent agreement with the reference standard supports that the ICPC-coded diagnoses of DM are a valid and better alternative than self-reported DM.

Another advantage of using general practice electronic health records is that these records are a rich source of information. In the NEO study, 1230 different ICPC-codes are registered in the obtained information from the electronic health records. In addition, detailed information about diseases is available in the consultation notes. In cohort studies, the number of obtained diseases via questionnaires is much lower (10–56 diseases) and often limited to the presence of a diagnosis and the year of diagnosis [23]. However, for each diagnosis to be extracted from general practice medical records, it is important to investigate the validity of ICPC-coded diagnoses and to take the specific primary care setting into account.

Worldwide, electronic general practice data is increasingly used to conduct research [24]. ICPC-coded diagnoses of DM from general practice data may also be a valid source for case-definitions in other countries with a primary care system that is comparable with that in the Netherlands, like in the United Kingdom. In addition, in many countries, an infrastructure for GP data-sharing is implemented in general practices to develop a research data warehouse [25]. In the NEO study, a Dutch research data warehouse will be used in future follow-up. A research data warehouse will reduce the time and costs related to the collection of diagnoses and other medical information and will make the use of GP data for researchers even more feasible.

In conclusion, the excellent agreement with the reference standard in combination with the high follow-up rate supports that ICPC-coded diagnoses of DM are a feasible, valid and a better alternative to self-reported diagnoses of DM for ascertainment of DM cases in large cohort studies.

Funding

The NEO study is supported by the participating departments, the Division and the Board of Directors of the LUMC, and by the

Leiden University, Research Profile Area 'Vascular and Regenerative Medicine'

Conflict of interest

None.

Acknowledgements

The authors express their gratitude to all individuals who participate in the NEO study. We are grateful to all participating general practitioners for inviting eligible participants and for extracting data from the electronic health records. Furthermore, we thank P.R. van Beelen and all research nurses for collecting the data, P.J. Noordijk and her team for laboratory management, H.J. de Jong for management of the GP database and I. de Jonge, MSc for data management of the NEO study.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.pcd.2020.08.011.

References

[1] J.E. Manson, G.A. Colditz, M.J. Stampfer, W.C. Willett, A.S. Krolewski, B. Rosner, et al., A prospective study of maturity-onset diabetes mellitus and risk of coronary heart disease and stroke in women, Arch. Intern. Med. 151 (6) (1991) 1141–1147.

[2] K.L. Margolis, Q. Lihong, R. Brzyski, D.E. Bonds, B.V. Howard, S. Kempainen, et al., Validity of diabetes self-reports in the Women's Health Initiative: comparison with medication inventories and fasting glucose measurements, Clin. Trials 5 (3) (2008) 240–247, http://dx.doi.org/10.1177/1740774508091749.

[3] I. Sluijs, A.D. van der, J.W. Beulens, A.M. Spijkerman, M.M. Ros, D.E. Grobbee, et al., Ascertainment and verification of diabetes in the EPIC-NL study, Neth. J. Med. 68 (1) (2010) 333–339.

[4] D.L. Ngo, L.M. Marshall, R.N. Howard, J.A. Woodward, K. Southwick, K. Hedberg, Agreement between self-reported information and medical claims data on diagnosed diabetes in Oregon's Medicaid population, J. Public Health Manag. Pract. 9 (6) (2003) 542–544.

[5] Y. Okura, L.H. Urban, D.W. Mahoney, S.J. Jacobsen, R.J. Rodeheffer, Agreement between self-report questionnaires and medical record data was substantial for diabetes, hypertension, myocardial infarction and stroke but not for heart failure, J. Clin. Epidemiol. 57 (10) (2004) 1096–1103, http://dx.doi.org/10.1016/j.jclinepi.2004.04.005, doi:S0895-4356(04)00113-1 [pii].

[6] J.R. Robinson, T.K. Young, L.L. Roos, D.E. Gelskey, Estimating the burden of disease. Comparing administrative data and self-reports, Med. Care. 35 (9) (1997) 932–947.

[7] C.F. Simpson, C.M. Boyd, M.C. Carlson, M.E. Griswold, J.M. Guralnik, L.P. Fried, Agreement between self-report of disease diagnoses and medical record validation in disabled older women: factors that modify agreement, J. Am. Geriatr. Soc. 52 (1) (2004) 123–127, doi:52021 [pii].

[8] K.M. Skinner, D.R. Miller, E. Lincoln, A. Lee, L.E. Kazis, Concordance between respondent self-reports and medical records for chronic conditions: experience from the Veterans Health Study, J. Ambul. Care Manage. 28 (2) (2005) 102–110, doi:00004479-200504000-00002 [pii].

[9] K. Kelley, B. Clark, V. Brown, J. Sitzia, Good practice in the conduct and reporting of survey research, Int. J. Qual. Health Care 15 (3) (2003) 261–266.

[10] D.M. Kriegsman, B.W. Penninx, J.T. van Eijk, A.J. Boeke, D.J. Deeg, Self-reports and general practitioner information on the presence of chronic diseases in community dwelling elderly. A study on the accuracy of patients' self-reports and on determinants of inaccuracy, J. Clin. Epidemiol. 49 (12) (1996) 1407–1417.

[11] H. Galenkamp, M. Huisman, A.W. Braam, F.G. Schellevis, D.J. Deeg, Disease prevalence based on older people's self-reports increased, but patient-general practitioner agreement remained stable, 1992-2009, J. Clin. Epidemiol. 67 (7) (2014) 773–780, http://dx.doi.org/10.1016/j.jclinepi.2014.02.002.

[12] Wonca International Classification Committee, The International Classification of Primary Care, 2013 http://wwwph3corg/4daction/w3_CatVisu/en/icpchtml?wCatIDAdmin=1106.

[13] C. Shivade, P. Raghavan, E. Fosler-Lussier, Pj Embi, N. Elhadad, Sb Johnson, Am Lai, A review of approaches to identifying patient phenotype cohorts using electronic health records, J. Am. Med. Inform. Assoc. 21 (2014) 221–230.

[14] R. de Mutsert, M. den Heijer, T.J. Rabelink, J.W. Smit, J.A. Romijn, J.W. Jukema, et al., The Netherlands Epidemiology of Obesity (NEO) study: study design and data collection, Eur. J. Epidemiol. 28 (6) (2013) 513–523, http://dx.doi.org/10.1007/s10654-013-9801-3.

[15] M.C. Poortvliet, M. Lamkaddem, W. Devillé, Niet op naam ingeschreven (NONI) bij de huisarts, Inventarisatie en gevolgen voor de ziekenfondsverzekerden, Utrecht (2005).

[16] R. Dijkstra, Mijn HIS is het best!, 2015 https://wwwnhgorg/actueel/columns/mijn-his-het-best.

[17] WHO Collaborating Centre for Drug Statistics Methodology, Anatomical Therapeutic Chemical (ATC) Classification System, 2013 http://wwwwhoccno/atc/structure_and_principles/.

[18] J.R. Landis, G.G. Koch, The measurement of observer agreement for categorical data, Biometrics 33 (1) (1977) 159–174.

[19] NCD Risk Factor Collaboration (NCD-RisC), Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4.4 million participants, Lancet 387 (10027) (2016) 1513–1530, http://dx.doi.org/10.1016/s0140-6736(16)00618-8.

[20] World Health Organization, Global Report on Diabetes, Geneva, 2016.

[21] M.I. Harris, R. Klein, T.A. Welborn, M.W. Knuiman, Onset of NIDDM occurs at least 4-7 yr before clinical diagnosis, Diabetes Care 15 (7) (1992) 815–819.

[22] D.M. Maahs, N.A. West, J.M. Lawrence, E.J. Mayer-Davis, Epidemiology of type 1 diabetes, Endocrinol. Metab. Clin. North Am. 39 (3) (2010) 481–497, http://dx.doi.org/10.1016/j.ecl.2010.05.011.

[23] V. Kristman, M. Manno, P. Cote, Loss to follow-up in cohort studies: how much is too much? Eur. J. Epidemiol. 19 (8) (2004) 751–760.

[24] S. de Lusignan, C. van Weel, The use of routinely collected computer data for research in primary care: opportunities and challenges, Fam. Pract. 23 (2) (2006) 253–263, http://dx.doi.org/10.1093/fampra/cmi106.

[25] A.M. Cole, K.A. Stephens, G.A. Keppel, C.P. Lin, L.M. Baldwin, Implementation of a health data-sharing infrastructure across diverse primary care organizations, J. Ambul. Care Manage. 37 (2) (2014) 164–170, http://dx.doi.org/10.1097/jac.0000000000000029.