International Journal of Public Health (2019) 64:441–450
https://doi.org/10.1007/s00038-018-1186-3

ORIGINAL ARTICLE

Identifying diabetes cases in health administrative databases: a validation study based on a large French cohort

Sonsoles Fuentes[1] · Emmanuel Cosson[2,3] · Laurence Mandereau-Bruno[1] · Anne Fagot-Campagna[4] · Pascale Bernillon[1] · Marcel Goldberg[5] · Sandrine Fosse-Edorh[1] · CONSTANCES-Diab Group

Received: 18 June 2018 / Revised: 3 September 2018 / Accepted: 26 November 2018 / Published online: 4 December 2018

Abstract

Objectives In the French national health insurance information system (SNDS) three diabetes case definition algorithms are applied to identify diabetic patients. The objective of this study was to validate those using data from a large cohort.

Methods The CONSTANCES cohort (Cohorte des consultants des Centres d'examens de santé) comprises a randomly selected sample of adults living in France. Between 2012 and 2014, data from 45,739 participants recorded in a self-administrated questionnaire and in a medical examination were linked to the SNDS. Two gold standards were defined: known diabetes and pharmacologically treated diabetes. Sensitivity, specificity, positive and negative predictive values (PPV, NPV) and kappa coefficients (k) were estimated.

Results All three algorithms had specificities and NPV over 99%. Their sensitivities ranged from 73 to 77% in algorithm A, to 86 and 97% in algorithm B and to 93 and 99% in algorithm C, when identifying known and pharmacologically treated diabetes, respectively. Algorithm C had the highest k when using known diabetes as the gold standard (0.95). Algorithm B had the highest k (0.98) when testing for pharmacologically treated diabetes.

Conclusions The SNDS is an excellent source for diabetes surveillance and studies on diabetes since the case definition algorithms applied have very good test performances.

Keywords Information systems · Diabetes · Algorithms · Validation studies · CONSTANCES

Introduction

Diabetes is one of the leading causes of morbidity and mortality worldwide. The growing diabetes epidemic represents a major challenge to public health (Cho et al. 2018). In this context, surveillance is fundamental in the development and evaluation of public health programmes to reduce the burden of diabetes, by improving the knowledge of the disease, by assessing the prevalence and incidence of diabetes and its complications and by defining target populations (Geiss et al. 2017, 2018; Kirtland et al. 2014; Schmittdiel et al. 2014). Data for diabetes surveillance are accessible through different sources, including national health surveys (Dwyer-Lindgren et al. 2016) and patient registries (Richesson 2011). Recently, health administrative databases have emerged as an efficient source of data for diabetes surveillance (Saydah et al. 2004). In addition, they can be used for other purposes such as studies on pharmacoepidemiology or cost-effectiveness evaluation of

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00038-018-1186-3) contains supplementary material, which is available to authorized users.

✉ Sonsoles Fuentes
sonsoles.fuentes@santepubliquefrance.fr

[1] Santé publique France (SpF), F-94415 Saint-Maurice, France

[2] Department of Endocrinology-Diabetology-Nutrition, AP-HP, Jean Verdier Hospital, Paris 13 University, Sorbonne Paris Cité, CRNH-IdF, CINFO, Bondy, France

[3] Sorbonne Paris Cité, UMR U1153 Inserm/U1125 Inra/Cnam/Université Paris 13, Bobigny, France

[4] Strategy and Research Department, Caisse nationale de l'assurance maladie, Paris, France

[5] Population-Based Epidemiological Cohorts Unit, Inserm UMS 011, Villejuif, France

public health programmes (Bezin et al. 2017; Goldberg 2006).

Health administrative databases can be accessed easily and quickly, associated costs are low and they are quite exhaustive. However, using these databases for surveillance purposes is not a simple matter, because of the large volumes of data stored and because these data have not been necessarily recorded for epidemiological purposes(Walraven 2017). They also have other limitations since many of them are regional, not national, databases (Dart et al. 2011; Lipscombe and Hux 2007; Monesi et al. 2012) or, like Medicare, they concern only certain groups of population (Day and Parker 2013; Sakshaug et al. 2014).

Created in 1999, the French national health insurance information system (Système national inter-régime de l'Assurance maladie—SNIIRAM-, recently renamed *Système National des Données de Santé* –SNDS-) is one of the largest health administrative databases in the world (Maura et al. 2015; Tubiana et al. 2017; Tuppin et al. 2017; Weill et al. 2016). Today, it covers more than 99% of the French population (approximately 65 million people) including people living in French overseas territories (Tuppin et al. 2017). In the absence of a registry of diabetic patients in France, the Diabetes National Surveillance System was developed through this health administrative database which is used to estimate the national prevalence of pharmacologically treated diabetes and the incidence of diabetes-related complications, as well as their temporal trends and their territorial variations (Fosse-Edorh et al. 2017). To estimate these indicators, a diabetes case definition algorithm only based on antidiabetic drug consumption was applied. Two other diabetes case definitions have been proposed in France (de Lagasnerie et al. 2018; Fosse-Edorh et al. 2017). One uses information on individuals with diabetes who benefit from full insurance coverage for this chronic illness (*affection de longue durée-diabète,* hereafter ALD-diabetes). The French national health insurance scheme offers full coverage of healthcare costs for people presenting certain chronic diseases, including diabetes, based on the medical doctor request and an insurance physician validation. The other, which is the latest algorithm to be introduced, adds hospital diagnoses codes to the combination of information on ALD-diabetes and antidiabetic drug consumption.

Certain diabetes case definitions applied in other health administrative databases like Medicare (Hebert et al. 1999) or regional Canadian information systems (Leong et al. 2013) have already been validated. In those validation studies, the gold standard references were based on various sources such as linkage with registries of diabetic patients, laboratory data or primary care medical chart reviews. Recently, the setting up of the CONSTANCES cohort in France, which links self-reported data and data from medical examinations with health administrative databases, opens new perspectives for the validation of the three diabetes case definition algorithms developed to date for the French national health insurance information system.

The main objective of this study was to assess the test performance for different characteristics of the three diabetes case definition algorithms introduced above, in identifying both "known diabetes" and "pharmacologically-treated diabetes", using data from a large sample of adults living in Metropolitan France.

Methods

The French national health insurance information system

Two main databases comprise the French national health insurance information system: the Inter-Scheme Consumption Data (*Données de consommation inter-régimes,* DCIR) and the Medical Information System Program (*Progamme de médicalisation des systèmes d'information,* PMSI) (Tuppin et al. 2017). In the DCIR, out-of-hospital reimbursement information on dispensed health care and full insurance coverage due to chronic disease diagnosis codes are complemented with demographic data (age, gender and residence). However, diagnoses that apply to outpatient visits as well as the results of biological exams are not recorded in this database. In the PMSI, inpatient data from public and private hospitals are recorded, including admission and discharge dates, primary, related and associated diagnoses and certain medical procedures—but not the results of biological examinations. Both databases are linked through anonymized identification for each beneficiary.

Diabetes case definition algorithms

Three diabetes case definition algorithms used to date in the French national health insurance information system are outlined below (Fosse-Edorh et al. 2017).

Algorithm A Positive if the individual benefits during the given year from full health insurance coverage due to a chronic disease with a ICD-10 code of diabetes (E10 or E14), i.e. ALD-Diabetes.

Algorithm B Positive if the individual has a reimbursement of an antidiabetic drug (class A10 from Anatomic Treatment Classification (ATC)—except Benfluorex-) on at least three different dates in a given year or on two dates if at least one large package of antidiabetic drugs was dispensed.

Algorithm C Positive if the individual meets at least one of following conditions: (a) is registered as having ALD-Diabetes during the given year; (b) is reimbursed for an antidiabetic drug on at least three different dates in the previous 2 years, or on two dates if at least one large package of antidiabetic drugs was dispensed; (c) was hospitalized with a principal or related diagnosis of diabetes (E10–E14) or with a principal or related diagnosis of a diabetes-related complication (G59.0*,G63.2*, G73.0*, G99.0*, H28.0*, H36.0*, I79.2*, L97, M14.2*, M14.6*, N08.3) and an associated diagnosis of diabetes (E10-E14) in the previous 2 years.

The CONSTANCES cohort

CONSTANCES is a prospective population-based general-purpose cohort designed to serve as an open epidemiological research infrastructure (Zins et al. 2010). A five-year recruitment process started in 2012. The CONSTANCES cohort aims to constitute a representative sample of the French adult population aged 18–69 at cohort inception. People in the cohort were randomly selected within the National Health Insurance Fund beneficiaries. In France, all salaried workers—whether active or retired—and their families, are affiliated to the National Health Insurance Fund ("*Caisse Nationale d'Asssurance Maladie des travailleurs salaries*", CNAMTS) which covers approximately 86% of the French population.

A self-administered questionnaire with items on lifestyle factors, socio-economic status, occupational exposures and health status is completed by the participants at home. They also attend one of CONSTANCES's 22 dedicated recruitment sites, distributed throughout Metropolitan France for a medical examination. These sites are Health Screening Centers (HSC) managed by the CNAMTS which provide a free medical check-up every 5 years to salaried workers and their families. As part of the medical examination, an exhaustive questionnaire on personal and family disease history and health conditions is completed by HSC's physicians. The medical questionnaire is followed by a physical examination, anthropometric measurements, blood sampling and other tests.

Once a year, the CNAMTS transfers data on healthcare reimbursements and hospitalization, as well as other data regarding the cohort's participants from the French national health insurance information system to the central CONSTANCES database. Data collected in the self-administered questionnaire and in the medical examination at cohort inclusion are then linked with the data provided from the French national health insurance information system from three years prior to the inclusion of the participant in the study. Further information on the CONSTANCES cohort can be found elsewhere (Goldberg et al. 2017; Ruiz et al. 2016; Zins et al. 2010).

Study population

The study population was selected among CONSTANCES participants recruited between 2012 and 2014. Women who declared in the self-administered questionnaire that they had gestational diabetes mellitus or were pregnant were not included in the study population. Individuals for whom data from the French national health insurance information system were not available were secondarily excluded from the resulting validation population, as were those who neither filled out the self-questionnaire nor the medical questionnaire. A descriptive analysis on socio-economic, sociodemographic and anthropometric characteristics was performed in the validation population and in the population excluded due to unavailable data (health insurance data or self-reported questionnaire and medical questionnaire).

Gold standard: "known diabetes"

Data from the self-administered questionnaire and the medical questionnaire were used to define the gold standard for known diabetes cases. In the self-administered questionnaire, participants reported to have diabetes through the item: "*Have you ever been told by a doctor or other health care professional that you had diabetes?*". In the medical questionnaire, completed during the medical examination, the physician asked each participant if they had diabetes. Based on both items a gold standard variable "known diabetes" was constructed with two categories "positive" and "negative".

Gold standard: "pharmacologically-treated diabetes"

Two questions in the self-administered questionnaire were related to diabetes treatment: *"Are you currently being treated for diabetes with oral medication?"* and "*Are you currently being treated for diabetes with one or more insulin injections?*". Among the participants already categorized under "positive" for known diabetes, those who reported diabetes treatment (insulin, oral medication or both) constituted the "positive" category of the second gold standard, entitled "pharmacologically-treated diabetes".

Statistical analysis

The three diabetes case definition algorithms A, B and C outlined above were applied and their test characteristics compared with the two gold standards. Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy and Cohen's kappa coefficient (κ-coefficient) together with their 95% CI were all estimated to evaluate the performance of each algorithm in identifying "known" diabetes cases and "pharmacologically-treated" diabetes cases. The level of agreement was assessed as follows: poor (κ-coefficient < 0.20); fair ($0.20 \leq \kappa$-coefficient < 0.40); moderate ($0.40 \leq \kappa$-coefficient < 0.60); good ($0.60 \leq \kappa$-coefficient < 0.80); and very good (κ-coefficient ≥ 0.80) (Leong et al. 2013). A supplementary analysis was done stratifying the previous analysis by sex and age groups (18–29 years, 30–54 years and 55 years or more). The analyses were performed using SAS 9.4 and STATA 14 software packages.

Results

Validation study population

A total of 50,954 participants were recruited between 2012 and 2014 (see Fig. 1). Women who reported a previous diagnosis of gestational diabetes mellitus ($n = 545$) and those who reported being pregnant in the self-administered questionnaire ($n = 179$) were excluded. Participants for whom full national health insurance information system data ($n = 4477$, 8.7%), or both self-administered questionnaire and medical questionnaire ($n = 14$) were not available, were secondarily excluded from the validation population (see Fig. 1).

The characteristics of the validation population ($n = 45{,}739$) were compared with the population excluded due to absence of either full health insurance data or self-administered questionnaire and medical questionnaire data ($n = 4491$).The individuals in the validation population were more likely to be men, to be obese, to have been treated less frequently for hypertension and to be smokers. They also had a higher socio-economic status, were more likely to have been born in France (including overseas territories) and to have a professional activity (**see** Table 1).

Among the individuals who constituted the validation population, 1157 were classified as having known diabetes and 1018 pharmacologically treated diabetes (see Fig. 1).

Gold standard: "known diabetes"

Test performances to identify known diabetes cases of the three algorithms, previously developed, are described in Table 2. Irrespective of the algorithm used, the proportion of true negatives among those not having diabetes (specificity) was above 99.9%. Sensitivity varied between 73.7% (95% CI 71.1, 76.2) for algorithm A, 85.8% (95% CI 83.7, 87.8) for algorithm B and 93.8% (95% CI 92.2, 95.1) for algorithm C. No algorithm had a NPV below 99% or a PPV below 96%. The level of agreement with the gold standard for all three algorithms was very good (κ-coefficient 0.85, 0.91 and 0.95 for the algorithms A, B and C, respectively) without overlapping of the 95% CI of the values. In the results of the supplementary analysis, stratified by sex and age groups, no relevant differences in the validation tests were observed (see Electronic supplementary material ESM table A)".

Gold standard: "pharmacologically-treated diabetes"

Algorithm C's sensitivity in identifying pharmacologically treated diabetes cases (99.3%, 95% CI 98.6, 99.7) was higher than both algorithm B's (97.3%, 95% CI 96.2, 98.2) and algorithm A's (77.2%, 95% CI 74.5, 79.7) sensitivity (see Table 3). A value close to 100% was observed for all three algorithms' specificities. Seven percentage points separated the highest PPV (algorithm B: 97.9%) from the lowest PPV (algorithm C = 90.6%); the 95% CI of each algorithm did not overlap. All NPV were over 99%. Concerning the level of agreement, algorithm B had the highest κ-coefficient (0.98, 95% CI 0.97, 0.98), followed by algorithm C (0.95, 95% CI 0.94, 0.96) and algorithm A the lowest one (0.84, 95% CI 0.82, 0.86). In the supplementary analysis by sex and age groups, no significant differences between categories were observed (see ESM table B).

Discussion

Test characteristics of three diabetes case definition algorithms used in the French national health insurance information system were assessed using two gold standards (entitled "known diabetes" and "pharmacologically-treated diabetes") in a large cohort of more than 45,000 individuals which combined self-reported data, data from medical examination and data from the French national health insurance information system. All three diabetes case definition algorithms had very good test performances. The most exhaustive, algorithm C—which combined ALD-Diabetes, treatment reimbursement and hospitalizations—showed the best test characteristics for identifying known diabetes cases, by definition; it also had the highest sensitivity when using pharmacologically treated diabetes as gold standard. The algorithm B, based only on treatment's reimbursement, exhibited by definition the best test characteristics when using pharmacologically treated diabetes as a gold standard. Algorithm A, which used only ALD-

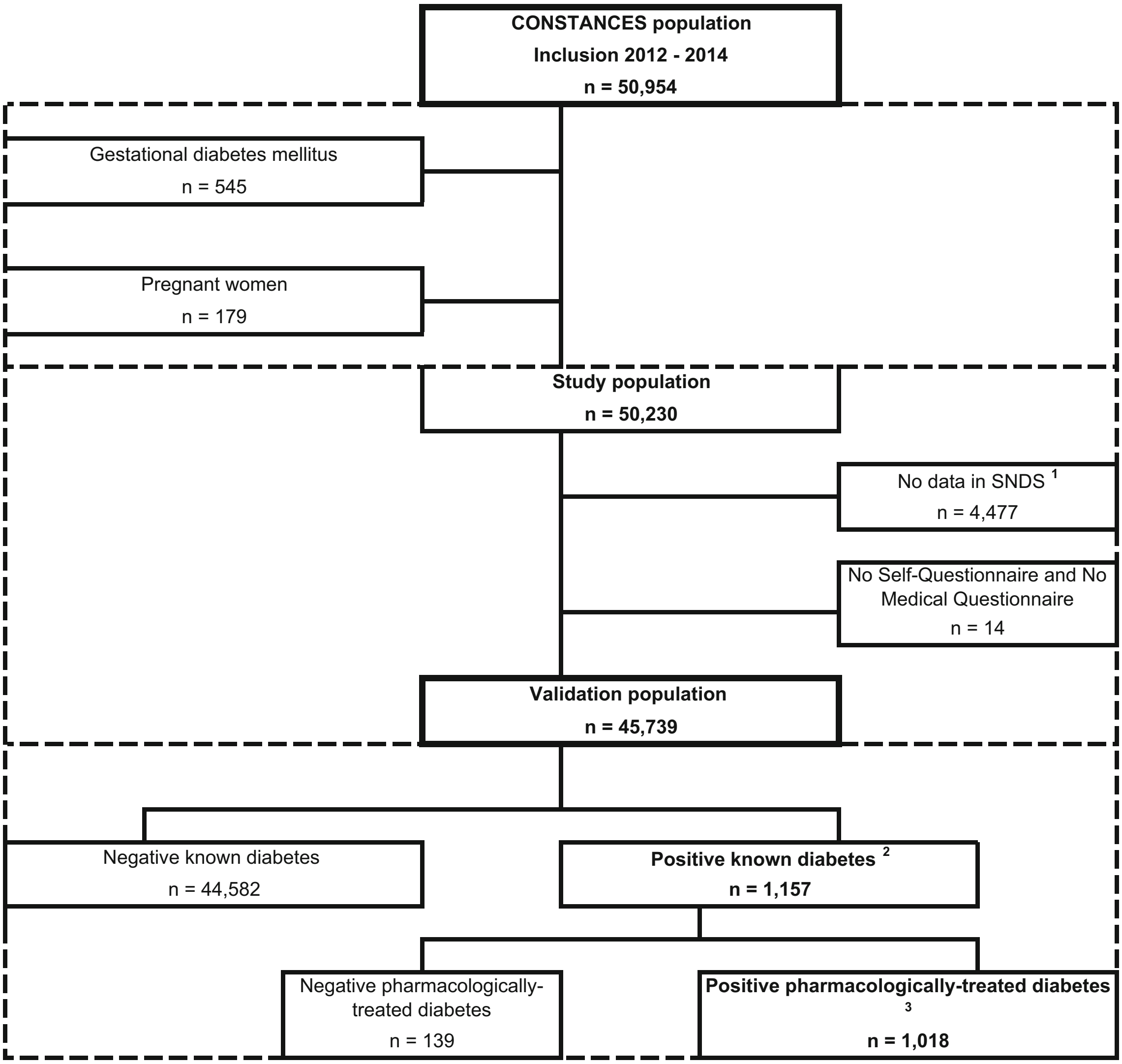

Fig. 1 Flow chart and number of people with known diabetes and with pharmacologically treated diabetes in the validation population. 1 *Système National des Données de Santé* (SNDS) French national health insurance information system, 2 Gold standard "Known Diabetes" and 3 gold standard "pharmacologically-treated diabetes"

Diabetes-full health insurance coverage for diabetes-, had the weakest test results for both gold standards.

In the absence of a diabetes registry of patients, health administrative databases are a valuable tool for diabetes surveillance and medical research. Canada's diabetes surveillance system, entitled the National Diabetes Surveillance System (NDSS), is based on its regional health administrative database. The NDSS diabetes case definition is as follows: two physician claims within a two-year period or one hospitalization with a ICD-code for diabetes (Clottey et al. 2001). An extensive meta-analysis recently estimated the pooled sensitivity of the NDSS case definition at 82.3% and the pooled specificity at 97.9% (Leong et al. 2013). Frequently, diabetes case definition algorithms have higher values for specificity than for sensitivity because when chronic diseases are ascertained in administrative databases, the proportion of false positives is usually lower than the proportion of false negatives (Muggah et al. 2013). In an interesting study performed in the USA, six algorithms for identifying Medicare beneficiaries with diabetes were validated using self-reported diabetes status from the Medicare Current Beneficiary Survey as the gold standard (Hebert et al. 1999). While all six algorithms had high specificity, the maximum value for sensitivity and the kappa coefficient were 71% and 0.7, respectively.

Validation studies of diabetes case definition algorithms published prior to this study relied on small samples or samples which were not representative of the general population (Leong et al. 2013). Instead, the CONSTANCES cohort enabled us to reach a larger and general population representative sample. Other strength of this

Table 1 Descriptive table of validation and excluded populations' characteristics from 50,230 participants in CONSTANCES[a] cohort recruited between 2012 and 2014 in France

n	Validation population 45,739	Excluded population 4491	*p* value[b]
Age (mean, ± sd)	49.1 ± 13.2	49.5 ± 15.2	0.058
Gender, men (%)	47.41	40.26	< 0.001
Current smoking status (%)			
Never smoked	45.35	50.64	< 0.001
Former smoker	19.50	17.27	
Current smoker	35.15	32.09	
Height and weight (%)			
Body mass index, kg/m^2	25.1 ± 4.5	24.9 ± 4.6	0.004
Self-reported disease (%)			
Treated hypertension	13.17	14.71	0.004
Treated dyslipidemia	10.57	11.62	0.220
Family medical history (%)			
Mother or father diagnosed with diabetes	15.75	15.92	0.768
Socio-economic status (%)			
Education (ISCED 2011[c]) (%)			
No education–primary education	3.25	3.69	< 0.001
Lower secondary education	7.12	12.55	
Upper secondary education	34.35	40.74	
Lower tertiary education	33.34	28.79	
Upper tertiary education	21.94	14.22	
Geographical origin			
France	89.00	84.08	< 0.001
DOM-TOM[d]	0.89	1.76	
Europe	4.25	4.34	
North Africa	2.93	5.49	
Sub-Saharan Africa	1.19	1.81	
Asia	0.74	1.03	
Others	0.99	1.49	
Professional activity			
Employed	65.10	44.10	< 0.001
Unemployed	6.22	5.21	
Retired	23.40	30.17	
Student	1.79	9.98	
Unemployed due to disability	1.58	7.68	
No professional activity	1.89	2.86	

[a] Constances "*Cohorte des consultants des Centres d'examens de santé*", [b] Student *T* Test (continuous variables) and Chi square test (categorical variables), [c] ISCED: International Standard Classification of Education, [d] DOM-TOM: French overseas territory

study is that the gold standard relies not only on self-reported data but also on data collected by a physician during a medical examination. Moreover, the exhaustive data collection in the CONSTANCES cohort ensured two validation analyses by using two gold standards "known diabetes" and "pharmacologically treated diabetes.

Despite having the weakest performance, Algorithm A has been frequently used in the French literature (Fromont et al. 2013; Perlbarg et al. 2013; Ricci et al. 2013) because it is simple to implement. One factor to consider with respect to algorithm A is that information on ALD-Diabetes before 2014 was either not available or not exhaustively recorded for some specific French health insurance funds, for example those for farmers or self-employed people (Fosse-Edorh et al. 2017). Moreover, in 2014, 21% of people pharmacologically treated for diabetes did not have ALD-diabetes status. This rate varied depending on geographic area and socio-economic level. No false

Table 2 Test characteristics of three diabetes case definition algorithms applied in the French national health insurance information system using known diabetes as the gold standard (based on data from participants of CONSTANCES[a] cohort recruited between 2012 and 2014 in France)

	TP	FP	TN	FN	Se (%) (95% CI)	Sp (%) (95% CI)	PPV (%) (95% CI)	NPV (%) (95% CI)	Acc (%) (95% CI)	*K* (95% CI)
Algorithm A[b]	853	0	44,582	304	73.73 (71.09, 76.24)	100.0 (99.99, 100.0)	100.0 (99.57, 100.0)	99.32 (99.24, 99.40)	99.34 (99.26,99.41)	0.85 (0.83, 0.86)
Algorithm B[c]	993	19	44,563	164	85.83 (83.68, 87.79)	99.96 (99.93, 99.97)	98.12 (97.08, 98.87)	99.63 (99.57, 99.69)	99.60 (99.54,99.66)	0.91 (0.90, 0.93)
Algorithm C[d]	1085	31	44,551	72	93.78 (92.23, 95.10)	99.93 (99.90, 99.95)	97.22 (96.08, 98.11)	99.84 (99.80, 99.87)	99.77 (99.73,99.82)	0.95 (0.94, 0.96)

TP true positives, *FP* false positives, *TN* true negatives, *FN* false negatives, *Se* sensitivity, *Sp* specificity, *PPV* positive predictive value, *NPV* negative predictive value, *Acc* accuracy, *K* kappa coefficient, 95% *CI* confidence interval

[a]CONSTANCES "*Cohorte des consultants des Centres d'examens de santé*". [b]Algorithm A: Benefiting from ALD-Diabetes ["Affection de longue durée -diabète" (chronic disease -diabetes)] status or 100% reimbursement of care due to a previous diagnosis of diabetes by a physician that was validated by an insurance doctor. [c]Algorithm B: Having at least 3 antidiabetic drug reimbursements recorded in the previous year (or 2 if one of them was a large package), [d]Algorithm C: At least one of the three following conditions: (a) benefiting from ALD-Diabetes status; (b) having at least 3 antidiabetic drug reimbursements recorded in the previous 2 years (or 2 if one of them was a large package); (c) having had at least one hospitalization related to diabetes in the previous 2 years

Table 3 Test characteristics of three diabetes case definition algorithms applied in the French national health insurance information system using pharmacologically treated diabetes as the gold standard (based on data from participants of CONSTANCES[a] cohort recruited between 2012 and 2014 in France)

	TP	FP	TN	FN	Se (%) (95% CI)	Sp (%) (95% CI)	PPV (%) (95% CI)	NPV (%) (95% CI)	Acc (%) (95% CI)	*K* (95 %CI)
Algorithm A[b]	786	67	44,654	232	77.21 (74.51, 79.75)	99.85 (99.81, 99.88)	92.15 (90.13, 93.86)	99.48 (99.41, 99.55)	99.35 (99.27,99.42)	0.84 (0.82, 0.86)
Algorithm B[c]	991	21	44,700	27	97.35 (96.16, 98.25)	99.95 (99.93, 99.97)	97.92 (96.85, 98.71)	99.94 (99.91, 99.96)	99.90 (99.86,99.92)	0.98 (0.97, 0.98)
Algorithm C[d]	1011	105	44,616	7	99.31 (98.59, 99.72)	99.77 (99.72, 99.81)	90.59 (88.73, 92.24)	99.98 (99.97, 99.99)	99.76 (99.71,99.80)	0.95 (0.94, 0.96)

TP true positives, *FP* false positives, *TN* true negatives, *FN* false negatives, *Se* sensitivity, *Sp* specificity, *PPV* positive predictive value, *NPV* negative predictive value, *Acc* accuracy, *K* kappa coefficient, 95% *CI* confidence Interval

[a]CONSTANCES "*Cohorte des consultants des Centres d'examens de santé*". [b]Algorithm A: Benefiting from ALD-Diabetes ("Affection de longue durée-diabète" (chronic disease -diabetes)) status or 100% reimbursement of care due to a previous diagnosis of diabetes by a physician that was validated by an insurance doctor. [c]Algorithm B: Having at least 3 antidiabetic drug reimbursements recorded in the previous year (or 2 if one of them was a large package), [d]Algorithm C: At least one of the three following conditions: (a) benefiting from ALD-Diabetes status; (b) having at least 3 antidiabetic drug reimbursements recorded in the previous 2 years (or 2 if one of them was a large package); (c) having had at least one hospitalization related to diabetes in the previous 2 years

positives were found in the assessment of algorithm A with known diabetes as gold standard. This could be due to the administrative procedure involved for people who wish to benefit from ALD-Diabetes status. An individual's general practitioner should first sign an application for long-term illness recognition; this application is then sent to a health insurance physician for approval.

The sensitivity of algorithm B (based on antidiabetic drug reimbursements) for pharmacologically treated diabetes cases was 12 percentage points higher than its sensitivity for known diabetes. The higher number of false negatives when using "known diabetes" compared with using "pharmacologically-treated diabetes" as a gold standard can be explained by the fact that diabetic patients controlling their glycaemia through diet and physical activity are not classified as positive "known diabetes cases" by algorithm B.

The highest sensitivity in both validation analyses was observed in algorithm C, which combined data on ALD-diabetes, drug reimbursements and hospitalization diagnoses. Since its case definition was broader, both a lower number of false negatives and a higher number of false positives were expected. Because part of this algorithm is based on ALD-diabetes information, the practical considerations of algorithm A must be acknowledged (Fosse-Edorh et al. 2017). Beyond the objectives of the present study, we have validated the different components of algorithm C and their combinations; the results of these supplementary analyses are described in the tables C and D in the EMS. When removing the component related to ALD-diabetes from algorithm C, the test characteristics are similar to those of algorithm B. In addition, algorithm C is more complex, requiring the combination of a large number of variables from two databases (DCIR and PMSI) in the French national health insurance information system. This makes this algorithm more computationally expensive.

This study has some limitations. The prevalence of diabetes in our study population is lower than the estimated prevalence in all France (Carrere et al. 2018; Kusnik-Joinville et al. 2008), due to selection biases of the cohort (people with chronic diseases are less likely to participate) and the exclusion of certain groups with a high prevalence of diabetes (people aged over 70 years and those living in overseas territories) (Santin et al. 2016). However, we believe that these differences are not large enough to have an impact on results of the analyses (Wong and Lim 2011). Almost 10% of the selected study population in CONSTANCES had no linked data with the French national health insurance information system. As previously described, they had substantial differences from the validation population. The absence of data on reimbursement and hospitalization is partly the result of participants not giving their permission to link data, but mostly due to recent changes in health insurance affiliation (e.g. young adults affiliated to the student health insurance system change to the National Health Insurance Fund when they start working).

Conclusion

The French national health insurance information system is an excellent source for the study of diabetes, since the three diabetes case definition algorithms currently applied had very good test performances. Besides the performance of test characteristics, the objectives of the study, together with the accessibility of data and the workload expected (in terms of time and computational skills required), should be considered in the selection of the algorithm.

Algorithm C was found to be the most suitable to identify known diabetes because it also captures patients who were hospitalized or who died before having 3 drug deliveries. This algorithm is thus better suited for studies on complications and cost of care. This algorithm also had somewhat highest costs in terms of time and computational skills needed. Furthermore, its sensitivity may not be stable when studying temporal trends before 2014 or territorial variations. We found that algorithm B is preferable when the objective is to study temporal trends, territorial or socio-economic variations. Moreover, unlike algorithms A and C, algorithm B can be used in other countries since information on antidiabetic drug consumption is commonly available in national health administrative databases. However, one shortcoming common to any diabetes case definitions based on health administrative databases is their inability to identify undiagnosed diabetes.

Acknowledgements The CONSTANCES cohort is supported by the Caisse Nationale d'Assurance Maladie des travailleurs salariés-CNAMTS. CONSTANCES is accredited as a "National Infrastructure for Biology and health" by the governmental Investissements d'avenir programme and was funded by the Agence nationale de la recherche (ANR-11-INBS-0002 Grant). CONSTANCES also receives funding from MSD, AstraZeneca and Lundbeck managed by INSERM-Transfert. This study has received a funding from the Interministerial Mission for Combating Drugs and Addictive Behaviors ("Mission Interministérielle de Lutte contre les Drogues et les Conduites Addictives", MILDECA). None of the authors are salaried by the funders of the CONSTANCES cohort. The funders did not have any role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. All authors declared no potential conflict of interest relevant to this article.

Compliance with ethical standards

Conflict of interest All authors declared no potential conflict of interest relevant to this article.

Research involving human participants and/or animals This article does not contain any studies with human participants performed by any of the authors.

Informed consent The CONSTANCES study was approved by authorities regulating ethical data collection in France (CCTIRS: Comité Consultatif pour le Traitement des Informations Relatives à la Santé; CNIL-Commission Nationale Informatique et Liberté) and all participants signed an informed consent.

References

Bezin J, Duong M, Lassalle R, Droz C, Pariente A, Blin P, Moore N (2017) The national healthcare system claims databases in France, Sniiram And Egb: powerful tools for pharmacoepidemiology. Pharmacoepidemiol Drug Saf 26:954–962. https://doi.org/10.1002/Pds.4233

Carrere P, Fagour C, Sportouch D, Gane-Troplent F, Helene-Pelage J, Lang T, Inamo J (2018) Diabetes mellitus and obesity in the French Caribbean: a special vulnerability for women? Women Health 58:145–159. https://doi.org/10.1080/03630242.2017.1282396

Cho NH, Shaw JE, Karuranga S, Huang Y, Da Rocha Fernandes JD, Ohlrogge AW, Malanda B (2018) IDF Diabetes Atlas: global estimates of diabetes prevalence for 2017 and projections for

2045. Diabet Res Clin Pract 138:271–281. https://doi.org/10.1016/j.diabres.2018.02.023

Clottey C, Mo F, Lebrun B, Mickelson P, Niles J, Robbins G (2001) The development of the national diabetes surveillance system (NDSS) in Canada. Chronic Dis Can 22:67–69

Dart D, Martens PJ, Sellers EA, Brownell MD, Rigatto C, Dean HJ (2011) Validation of a pediatric diabetes case definition using administrative health data in Manitoba. Canada Diabetes Care 34:898–903. https://doi.org/10.2337/Dc10-1572

Day HR, Parker JD (2013) Self-report of diabetes and claims-based identification of diabetes among medicare beneficiaries. National Health Statistics Report 1–14

De Lagasnerie G, Aguade AS, Denis P, Fagot-Campagna A, Gastaldi-Menager C (2018) The economic burden of diabetes to French national health insurance: a new cost-of-illness method based on a combined medicalized and incremental approach. Eur J Health Econ 19:189–201. https://doi.org/10.1007/s10198-017-0873-y

Dwyer-Lindgren L, Mackenbach JP, Van Lenthe FJ, Flaxman AD, Mokdad AH (2016) Diagnosed and undiagnosed diabetes prevalence by county in the US 1999–2012. Diabetes Care 39:1556–1562. https://doi.org/10.2337/dc16-0678

Fosse-Edorh S, Rigou A, Morin S, Fezeu L, Mandereau-Bruno L, Fagot-Campagna A (2017) Algorithms based on medico-administrative data in the field of endocrine, nutritional and metabolic diseases, especially diabetes. Rev Epidemiol Sante Publique 65(Suppl 4):S168–S173. https://doi.org/10.1016/J.Respe.2017.05.001

Fromont A et al (2013) Comorbidities at multiple sclerosis diagnosis. J Neurol 260:2629–2637. https://doi.org/10.1007/S00415-013-7041-9

Geiss LS, Kirtland K, Lin J, Shrestha S, Thompson T, Albright A, Gregg EW (2017) Changes in diagnosed diabetes, obesity, and physical inactivity prevalence in US counties, 2004–2012. PLoS ONE 12:E0173428. https://doi.org/10.1371/journal.pone.0173428

Geiss LS, Bullard KM, Brinks R, Hoyer A, Gregg EW (2018) Trends in type 2 diabetes detection among adults in the USA, 1999–2014. BMJ Open Diabetes Res Care 6:E000487. https://doi.org/10.1136/bmjdrc-2017-000487

Goldberg M (2006) Administrative data bases: could they be useful for epidemiology? Rev Epidemiol Sante Publique 54:297–303

Goldberg M et al (2017) Constances: a general prospective population-based cohort for occupational and environmental epidemiology: cohort profile. Occup Environ Med 74:66–71. https://doi.org/10.1136/oemed-2016-103678

Hebert PL, Geiss LS, Tierney EF, Engelgau MM, Yawn BP, Am M (1999) Identifying persons with diabetes using medicare claims data. Am J Med Qual 14:270–277. https://doi.org/10.1177/106286069901400607

Kirtland KA, Burrows NR, Geiss LS (2014) Diabetes interactive atlas. Prev Chronic Dis 11:130300. https://doi.org/10.5888/Pcd11.130300

Kusnik-Joinville O, Weill A, Salanave B, Ricordeau P, Allemand H (2008) Prevalence and treatment of diabetes in France: trends between 2000 and 2005. Diabetes Metab 34:266–272. https://doi.org/10.1016/J.Diabet.2008.01.005

Leong A, Dasgupta K, Bernatsky S, Lacaille D, Avina-Zubieta A, Rahme E (2013) Systematic review and meta-analysis of validation studies on a diabetes case definition from health administrative records. PLoS ONE 8:E75256. https://doi.org/10.1371/Journal.Pone.0075256

Lipscombe LL, Hux JE (2007) Trends in diabetes prevalence, incidence, and mortality in Ontario, Canada 1995–2005: a population-based study. Lancet 369:750–756. https://doi.org/10.1016/s0140-6736(07)60361-4

Maura G, Blotiere PO, Bouillon K, Billionnet C, Ricordeau P, Alla F, Zureik M (2015) Comparison of the short-term risk of bleeding and arterial thromboembolic events in nonvalvular atrial fibrillation patients newly treated with dabigatran or rivaroxaban versus vitamin K antagonists: a French nationwide propensity-matched cohort study. Circulation 132:1252–1260. https://doi.org/10.1161/circulationaha.115.015710

Monesi L et al (2012) Prevalence, incidence and mortality of diagnosed diabetes: evidence from an Italian population-based study. Diabet Med 29:385–392. https://doi.org/10.1111/J.1464-5491.2011.03446.X

Muggah E, Graves E, Bennett C, Manuel DG (2013) Ascertainment of chronic diseases using population health data: a comparison of health administrative data and patient self-report. BMC Public Health 13:16. https://doi.org/10.1186/1471-2458-13-16

Perlbarg J, Allonier C, Boisnault P, Daniel F, Le Fur P, Szidon P, Bourgueil Y (2013) Feasibility and practical value of statistical matching of a general practice database and a health insurance database applied to diabetes and hypertension. Sante Publique (Bucur) 26:355–363

Ricci P, Mezzarobba M, Blotière P, Polton D (2013) Reimbursed health expenditures during the last year of life, in France, in the year 2008. Rev Epidemiol Sante Publique 61:29–36

Richesson RL (2011) Data standards in diabetes patient registries. J Diabetes Sci Technol 5:476–485. https://doi.org/10.1177/193229681100500302

Ruiz F et al (2016) High quality standards for a large-scale prospective population-based observational cohort: Constances. BMC Public Health 16:877. https://doi.org/10.1186/S12889-016-3439-5

Sakshaug S, Weir DR, Nicholas LH (2014) Identifying diabetics in medicare claims and survey data: implications for health services research. BMC Health Serv Res 14:150. https://doi.org/10.1186/1472-6963-14-150

Santin G et al (2016) Estimation De Prévalencesdans Constances: Premières explorations. Bull Épidémiologique Hebd 35–36:622–628

Saydah S, LS G, Tierney E, SM B, Engelgau M, Brancati F (2004) Review of the performance of methods to identify diabetes cases among vital statistics, administrative, and survey data. Ann Epidemiol 14:507–516. https://doi.org/10.1016/J.Annepidem.2003.09.016

Schmittdiel JA et al (2014) Prescription medication burden in patients with newly diagnosed diabetes: a surveillance, prevention, and management of diabetes mellitus (Supreme-Dm) study. J Am Pharm Assoc 54:374–382. https://doi.org/10.1331/japha.2014.13195

Tubiana S et al (2017) Dental procedures, antibiotic prophylaxis, and endocarditis among people with prosthetic heart valves: nationwide population based cohort and a case crossover. Study BMJ 358:J3776. https://doi.org/10.1136/Bmj.J3776

Tuppin P et al (2017) Value of a national administrative database to guide public decisions: from the Systeme National D'information Interregimes De L'assurance Maladie (SNIIRAM) to the Systeme National Des Donnees De Sante (SNDS) in France. Rev Epidemiol Sante Publique 65(Suppl 4):S149–S167. https://doi.org/10.1016/j.respe.2017.05.004

Walraven CV (2017) A comparison of methods to correct for misclassification bias from administrative database diagnostic codes. Int J Epidemiol 0:1–12. https://doi.org/10.1093/ije/dyx253

Weill A et al (2016) Low dose oestrogen combined oral contraception and risk of pulmonary embolism, embolism, stroke, and myocardial infarction in five million french women: cohort. Study BMJ 353:I2002. https://doi.org/10.1136/Bmj.I2002

Wong HB, Lim GH (2011) Measures of diagnostic accuracy: sensitivity, specificity, PPV and NPV. Proc Singap Healthc 20:316–318. https://doi.org/10.1177/201010581102000411

Zins M et al (2010) The constances cohort: an open epidemiological laboratory. BMC Public Health 10:1