Identifying type 1 / type 2 diabetes in medico-administrative database to improve health surveillance, medical research and prevention in diabetes: Algorithm development and application
Sonsoles Fuentes, Rok Hrzic, Romana Haneef, Sofiane Kab, Emmanuel Cosson, Sandrine Fosse-Edorh

To cite this version:
Sonsoles Fuentes, Rok Hrzic, Romana Haneef, Sofiane Kab, Emmanuel Cosson, et al.. Identifying type 1 / type 2 diabetes in medico-administrative database to improve health surveillance, medical research and prevention in diabetes: Algorithm development and application. Diabetes Epidemiology and Management, 2023, 10, pp.1-7. 10.1016/j.deman.2023.100137. hal-04165493

HAL Id: hal-04165493
https://cnam.hal.science/hal-04165493
Submitted on 19 Jul 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives| 4.0 International License
Original article

Identifying type 1 / type 2 diabetes in medico-administrative database to improve health surveillance, medical research and prevention in diabetes: Algorithm development and application

Sonsoles Fuentes*a, Rok Hrzićb, Romana Haneeff, Sofiane Kabc, Emmanuel Cossond,e, Sandrine Fosse-Edorha,*,1

* Santé Publique France (SpF), 12 rue du Val d’Osne, Saint Maurice Cedex 94415, France
b Faculty of Health, Medicine and Life Sciences, International Health, School for Public Health and Prim Care, Maastricht University, Maastricht, the Netherlands
c Population-Based Epidemiological Cohorts Unit, Inserm UMS 011, Villejuif, France
d Department of Endocrinology-Diabetology-Nutrition, AP-HP, Avicenne Hospital, Paris 13 University, Sorbonne Paris Cité, CRNH-IDF, CNFOD, Bobigny, France
e Sorbonne Paris Cité, UMR U1153 Inserm/U1125 Inra/Cnam/Université Paris 13, Bobigny, France

ARTICLE INFO

Article History:
Received 11 January 2023
Revised 1 March 2023
Accepted 1 March 2023
Available online 4 March 2023

Keywords:
Machine learning
Algorithm
Type 1 diabetes
Type 2 diabetes and prevalence

ABSTRACT

Introduction: Big data sources represent an opportunity for diabetes research. One example is the French national health data system (SNDS), gathering information on medical claims of out-of-hospital health care and hospitalizations for the entire French population (66 million). Currently, a validated algorithm based on antidiabetic drug reimbursement is able to identify people with pharmacologically-treated diabetes in the SNDS. But it cannot distinguish type 1 from type 2 diabetes. Differentiating type 1 and type 2 diabetes is crucial in diabetes surveillance, because they carry differences in their prevention, populations at risk, disease natural history, pathophysiology, management and risk of complications.

This article investigates the development of a type 1/type 2 diabetes classification algorithm using artificial intelligence and its application to estimate the prevalence of type 1 and type 2 diabetes in France.

Methods: The final data set comprised all diabetes cases from the CONSTANCES cohort (n = 951). A supervised machine learning method based on eight steps was used: final data set selection, target definition (type 1), coding features, final data set splitting into training and testing data sets, feature selection and training and validation and selection of algorithms. The selected algorithm was applied to SNDS data to estimate the type 1 and type 2 diabetes prevalence among adults 18–70 years of age.

Results: Among the 3481 SNDS features, 14 were selected to train the different algorithms. The final algorithm was a linear discriminant analysis model based on the number of reimbursements for fast-acting insulin, long-acting insulin and biguanides over the previous year (specificity 97% and sensitivity 100%). In 2016, after adjusting for algorithm performance, type 1 and type 2 diabetes prevalence in France was estimated to be 0.3% and 4.4%, respectively.

Conclusion: Our type 1/type 2 classification algorithm was found to perform well and to be applicable to any prescription or medical claims database from other countries. Artificial intelligence opens new possibilities for research and diabetes prevention.

© 2023 Published by Elsevier Masson SAS. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

Introduction

Diabetes is a leading cause of morbidity and mortality worldwide [1,2], and public health surveillance is fundamental in decreasing the global burden of diabetes [3]. In recent decades, big data have emerged and offered new opportunities for surveillance [4,5]. Big data refer to massive volumes of information collected from different sources, as characterized by the “three Vs”: volume, velocity and variety [6].
One example of a big data source for public health surveillance is the French national health data system, the SNDS [7,8]. In the SNDS, individual, updated and exhaustive health information from the entire French population (66 million people) is electronically collected, including information on claims from out-of-hospital health care consumption and on hospital stays in public and private hospitals. Currently, a validated algorithm based on antidiabetic drug reimbursement is able to identify people with pharmacologically treated diabetes [9,10]. But it cannot distinguish type 1 from type 2 diabetes. Differentiating type 1 and type 2 diabetes is crucial in diabetes surveillance, because they carry differences in their prevention, populations at risk, disease natural history, pathophysiology, management and risk of complications [11]. This limitation is commonly encountered in studies based on medico-administrative data, in which clinical diagnoses are not accessible or not reliably reported, e.g., studies based on Medicare or Medicaid. In these studies, conclusions are drawn on the basis of studying type 2 diabetes. Otherwise, to investigate type 1 diabetes, studies must be restricted to young individuals. Artificial intelligence, especially supervised machine learning, might be able to overcome this limitation by enabling the development of an innovative algorithm to classify type 1 and type 2 diabetes cases.

Supervised machine learning includes different methods in which classification or predictive algorithms are developed through linking known features in the assessment of targets by using a training data set in which these targets are characterized. The algorithm is then applied to additional data sources in which the targets are unknown [6].

The objectives of our study were (i) to develop an algorithm to distinguish type 1 and type 2 diabetes cases on the basis of information available in the SNDS through a supervised machine learning method and (ii) to apply this algorithm to the SNDS, to estimate the prevalence of type 1 and type 2 diabetes in France among adults in 2016.

Methods

Dataset

Since 2012, the CONSTANCES population-based general-purpose cohort has recruited 200,000 participants comprising a representative sample of the French population between 18 and 69 years of age (at inclusion) [12]. Individuals are randomly selected from among all beneficiaries of the National Health Insurance Fund. First, participants complete a self-administered questionnaire on health status, health-related behaviors, socioeconomic and demographic characteristics. They then attend a health screening center and receive a medical examination including medical questionnaires, physical examination and blood sampling for further biological tests. Finally, the SNDS information from the participants who provided consent is extracted and linked with the information collected in previous phases.

Training and evaluation methodology

A supervised machine learning method based on the following eight steps was applied [13] (Fig. 1): (i) selection of the final data set, (ii) target definition, (iii) coding features for a given window of time, (iv) splitting the final data set into a training data set and a testing data set, (v) feature selection, (vi) training algorithms, algorithm validation and (viii) final algorithm selection.

Step 1: selection of the final data set

All diabetes cases were selected among the participants recruited by CONSTANCES between 2012 and 2014, after exclusion of women who reported gestational diabetes mellitus, women who were pregnant during the study and participants without accessible data in the SNDS [9]. Subsequently, only individuals pharmacologically treated for diabetes for whom complete data on their diabetes diagnosis and treatment were available were retained in the final data set.

Step 2: Target definition

Type 1 and type 2 diabetes cases were identified with a decision tree developed in the ENTRED study and based on three items: age at diabetes diagnosis, current insulin treatment, and the delay between diabetes diagnosis and first insulin treatment [14]. Type 1 cases were defined as target positive, and type 2 cases were defined as target negative. A descriptive analysis of socioeconomic, sociodemographic and lifestyle factors, as well as anthropometric characteristics, was performed to assess the differences between the two groups.

Step 3: Coding features for a given window of time

A total of 3481 continuous features from SNDS data were coded regarding out-of-hospital health care reimbursement over the 12 months before the date of the self-administered questionnaire (numbers of medical consultations, dispensed drugs coded with the fifth level of the Anatomical Therapeutic Chemical code (ATC 05), biological tests, medical procedures, treatments and medical devices) and information on hospitalizations in the 24 months before the same date. Sex, age and the characteristics of the city/town of residence were also considered [15–17].

Step 4: Training data set and testing data set

The final data set was divided into training (80%) and testing (20%) data sets. Due to a substantial imbalance in the number of target positives and target negatives, a random down-sampling was performed in the target negatives.

Step 5: Feature selection

After removal of all features with a variance equal to zero, the ReliefExp score was estimated, on the basis of the relevance of each feature, to differentiate between the target positive and target negative groups in the target. The ReliefExp method is noise tolerant and is not affected by feature interactions [18,19]. The remaining features were ranked according to ReliefExp score.

Steps 6–8: Algorithm training, validation and selection

The following types of models were applied to the training data set: Linear discriminant analysis (LDA), logistic regression, flexible discriminant analysis (FDA) and C5 decision tree (C5) [20]. For each model, the features were selected with three different thresholds of ReliefExp scores: 0.35, 0.1 and 0.05 (Fig. 2). After an initial validation of the algorithms using the training data set (k-fold cross-validation), the algorithm performance was assessed with the testing data set. The estimated performance metrics for each algorithm were sensitivity, specificity, Kappa, F1 score and area under the receiver operating characteristic curve. Finally, we retained a single model on the basis of three criteria: performance, computational parsimony and applicability to additional databases.

Application

The French health care system has universal coverage, and all beneficiaries have a unique identification number and a personal smartcard allowing information on health care utilization to be electronically recorded [7,21]. This information is collected and anonymized by the SNDS, which comprises two main databases: inter-scheme consumption data (Données de consommation inter-régimes, DCIR) and the French national hospital discharge database.
The DCIR contains exhaustive information from medical claims on reimbursement for out-of-hospital dispensed health care together with demographic information (sex, age, and town or village of residence). The PMSI includes information from public and private hospitals.

All pharmacologically treated diabetes cases in France in 2016 were ascertained in the SNDS using a validated algorithm that...

Fig. 1. Supervised machine learning for developing a classification/prediction algorithm.

Fig. 2. Feature selection based on ReliefExp score using three different thresholds (0.35, 0.1 and 0.05)

No: number of; -12m-: over the prior 12 months; -24m-: over the prior 24 months; [+T2D]: the mean in the type 2 diabetes group is higher than that in the type 1 diabetes group.
identifies a diabetes case if an individual had a reimbursement for an antidiabetic drug (class ATC A10) 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 [9]. To exclude gestational diabetes mellitus cases, women identified with pharmacologically treated diabetes who gave birth in 2016 were excluded. In addition, all individuals with ages below 18 or above 70 years were not included in the study population. The algorithm selected in the previous section was applied in the study population to characterize each case as type 1 or type 2 diabetes.

To study the prevalence of type 1 and type 2 diabetes in France in 2016, we used the mean French population in 2016, estimated by the National Institute for Statistics (Institut national de la statistique et des études économiques, INSEE), as the denominator. The results were declined by sex and age (1-year class). Finally, the prevalence of type 1 and type 2 diabetes for the entire study population (adults 18–70 years of age) was adjusted to the performance of the algorithm (taking into account the positive and the negative predicted value) [22]. Analysis of the SNDS were performed with SAS 7.1, and supervised machine learning was performed with the R packages CORElearn and caret. All confidentiality, safety and security procedures were approved by the French legal authorities.

Results

Final data set

Among the 50,954 participants recruited by the CONSTANCES cohort between 2012 and 2014, a total of 1161 diabetes cases were identified. The final data set for developing the algorithm was composed of 951 pharmacologically treated diabetes cases. The number of type 1 diabetes cases (target 1) was 49, and the number of type 2 diabetes cases (target 2) was 902 (Fig. a ESM). In Table a ESM, the main characteristics of type 1 and type 2 diabetes cases are presented.

Feature selection for the type 1/type 2 diabetes classification algorithm

All 3481 features were ranked on the basis of their ReliefExp Score or their ability to differentiate between type 1 and type 2 diabetes (Fig. 2). The first feature was the number of reimbursements for fast-acting insulin/insulin analogues (ATC – A10AB-), which was followed by the number of reimbursements for long-acting insulins/insulin analogues (ATC – A10AE) and the number of reimbursements for biguanides (ATC – A10BA-).

The other features with a ReliefExp Score above 0.05 included those associated with the number of reimbursements for medical devices for self-monitoring (test strips for blood glucose tests, test strips for blood prothrombin, devices for glucose testing, or test strips for urine glucose and ketone bodies), the number of reimbursements for screening tests performed in out-of-hospital laboratories (glucose, microalbuminuria and prostate-specific antigen), information on hospitalizations associated with diabetes (total number of hospitalizations and number of hospitalizations with a duration between 1 and 7 days) and age at inception. When comparing the distribution of the selected features in the type 1 and type 2 groups, we found that only four features had a higher mean in the type 2 group: the number of reimbursements for biguanides, age at inception, and number of reimbursements for prostate-specific antigen screening and out-of-hospital glucose tests.

Type 1 / type 2 diabetes classification algorithm

After selection of the features, four different types of models (LDA, logistic regression, FDA and C5) with 3, 9 or 14 features (on the basis of three ReliefExp score thresholds of 0.35, 0.1 and 0.05, respectively) were trained with the training data set and subsequently validated. The results of k-fold cross validation within the training data set are shown in Fig. b ESM; all algorithms had an area under the receiver operating characteristic curve above 0.94, and the upper limit of the 95% confidence interval exceeded 0.99. In performance testing within the testing data set, the three algorithms based on LDA had the highest specificity and accuracy (above 97%) as well as the highest F1 score (0.8) (Table b ESM). Among them, the algorithm with the highest parsimony and the best applicability to further databases was the one with three features (number of reimbursements for fast-acting insulin, long-acting insulin and biguanides). Figs. 3 and 4 provide a graphic representation of the selected type 1/type 2 classification algorithm. After the algorithm was applied in the testing data set, only five type 2 diabetes cases were misclassified as type 1 cases, and no type 1 cases were misclassified as type 2 cases.

Fig. 3. Validation using data from the testing data set (n = 189) of the type 1/type 2 diabetes classification algorithm (linear discriminant model with ReliefExp score threshold for feature selection of 0.35–3 features).

No*: number of reimbursements over the prior 12 months.
In 2016, a total of 1,844,329 diabetes cases ranging in age from 18 to 69 years (after excluding 7248 pregnant women) were identified. In Fig. 4, type 1 and type 2 diabetes prevalence is presented by 1-year age group and sex. Before the age of 32–34 years, the prevalence of type 1 diabetes was higher than that of type 2 diabetes, but the prevalence of type 2 diabetes prevalence increased sharply with age above 34 years, reaching rates of 18% and 12% among men and women, respectively, in the 70-year age group. Regarding sex, the prevalence rates of type 1 diabetes were higher among men than women across all age groups, whereas those of type 2 diabetes were higher among women until the age of 32, at which point the prevalence became higher in men. After adjusting for algorithm performance, the percentage of type 1 cases among all diabetes cases was 6.9%, the prevalence of type 1 diabetes was 0.32% (0.36% in men and 0.29% in women), and the prevalence of type 2 diabetes was 4.36% (5.03% in men and 3.72% in women).

Feature selection: from 3481 to 14 features

From the 3481 features coded, 14 were selected for developing the algorithms, because of their ability to differentiate type 1 and type 2 diabetes. Most of the selected features were expected. The first three features with the highest ReliefExp scores were associated with diabetes treatment. Long-acting combined with fast-acting insulins are the most common treatment for type 1 diabetes, whereas type 2 cases are more commonly treated with biguanides [23]. Other groups of features selected were the number of reimbursements for self-monitoring devices for measuring glucose, such as test strips for blood glucose, or for urine glucose and ketone bodies; these devices are more frequently used by individuals with type 1 than type 2 diabetes, who have higher risks of hypoglycemia and ketosis. Features of hospitalizations with a diabetes diagnosis (total number and number of hospitalizations from 1 to 7 days over the prior 2 years) were also highly discriminant, because people with type 1 diabetes usually experience more acute complications, such as diabetic ketoacidosis, than those with type 2 diabetes [24]. Two selected features were associated with reimbursement for screening tests for follow-up performed in out-of-hospital laboratories. The first feature was the number of glucose tests, which was more frequent in the type 2 group, because those individuals are less likely to monitor blood glucose themselves [25]. The second feature was the greater number of tests for the urinary albumin excretion rate in type 1 than type 2 cases [26].

However, some unexpected features were highly discriminant between type 1 and type 2 diabetes. One such feature was the age of 32–34 years, at which point the prevalence of type 2 diabetes began to exceed that of type 1 diabetes. The prevalence of both types of diabetes was higher in men, except for the prevalence of type 2 diabetes in the population between 18 and 32 years of age, which was higher among women.

Feature selection: from 3481 to 14 features

From the 3481 features coded, 14 were selected for developing the algorithms, because of their ability to differentiate type 1 and type 2 diabetes. Most of the selected features were expected. The first three features with the highest ReliefExp scores were associated with diabetes treatment. Long-acting combined with fast-acting insulins are the most common treatment for type 1 diabetes, whereas type 2 cases are more commonly treated with biguanides [23]. Other groups of features selected were the number of reimbursements for self-monitoring devices for measuring glucose, such as test strips for blood glucose, or for urine glucose and ketone bodies; these devices are more frequently used by individuals with type 1 than type 2 diabetes, who have higher risks of hypoglycemia and ketosis. Features of hospitalizations with a diabetes diagnosis (total number and number of hospitalizations from 1 to 7 days over the prior 2 years) were also highly discriminant, because people with type 1 diabetes usually experience more acute complications, such as diabetic ketoacidosis, than those with type 2 diabetes [24]. Two selected features were associated with reimbursement for screening tests for follow-up performed in out-of-hospital laboratories. The first feature was the number of glucose tests, which was more frequent in the type 2 group, because those individuals are less likely to monitor blood glucose themselves [25]. The second feature was the greater number of tests for the urinary albumin excretion rate in type 1 than type 2 cases [26].
number of reimbursements for prostate-specific antigen, whose discriminant ability may relate to this type of screening usually being recommended for older men—a group relatively more likely to have type 2 diabetes [27]. The other feature was the number of reimbursements for test strips for self-monitoring of blood prothrombin, which was higher among individuals with type 1 than type 2 diabetes. We hypothesize that individuals with type 1 diabetes, who also may tend to be highly concerned about heart disease [28], are more likely to self-monitor blood characteristics related to heart disease than individuals with type 2 diabetes.

The type 1 / type 2 diabetes classification algorithm

The final algorithm was a LDA model based on three features: the number of reimbursements for long-acting insulin, for fast-acting insulin and for biguanides over the previous 12 months. Most of the algorithms applied to health administrative databases to characterize type 1 and type 2 diabetes cases are based on the International Classification of diseases (ICD) 9/10 diagnostic codes [29]. Unfortunately, in the SNDS, as in other medico-administrative databases, out-of-hospital diagnostic codes are either not available or not reliable. Diagnostic codes are usually recorded manually by health care professionals for financial purposes. Therefore, they are at risk for error and bias [30]. For example, in countries (such as France) where hospitals are paid through a diagnosis-related group’s system, diseases with lower reimbursement for hospitals may be under-recorded [8,31]. Because our algorithm is based on drug reimbursements electronically recorded at the point of sale, it is not exposed to this limitation. In addition, the sensitivity and specificity of this classification algorithm are better than those of previous algorithms [29]. The classification algorithm had very good performance, with a sensitivity of 100%; this is an exceptional characteristic for algorithms applied in health administrative databases, which usually have moderate sensitivity [22].

The combination of therapeutic features constituting the algorithm was consistent with treatment guidelines in France, other European countries and the US [23,32]. Metformin monotherapy is the recommended starting pharmacological treatment for type 2 diabetes cases, whereas type 1 diabetes cases should be treated with multiple daily injections of rapid-acting insulin with meals combined with daily basal insulin. This characteristic of the algorithm enhances its applicability to prescription or medical claims databases from other countries where these guidelines are followed. This aspect is important, because some countries or regions use this type of database only for diabetes surveillance, whereas other countries, such as Norway or the US, use these databases to complete information from other sources, such as national diabetes registers or national surveys [33–35].

The estimations of type 1 diabetes prevalence among adult population are scarce [27,36]. By applying the classification algorithm to the SNDS, we were able to estimate for the first time in France the prevalence of type 1 and type 2 diabetes in adults aged between 18 and 70 years. The observed prevalence of type 1 diabetes among adults in the UK and the US are close to the one described in our study, while type 2 prevalence was higher especially in the US (8.5% in 2016) [27,36].

Strengths and limitations

Our study has several strengths. The algorithm was developed using data from a large sample representative of the population living in metropolitan France [12]. Moreover, to estimate the prevalence of type 1 and type 2 diabetes among adults, we used the SNDS, a nationwide population based database including all residents in France, thereby overcoming the limitations of other studies based on national population such as selection bias or recall bias [8].

Our study also has some limitations. Because the CONSTATES cohort includes only adults between 18 and 70 years of age, the performance of this algorithm for other age groups may differ. In addition, other types of diabetes, such as latent autoimmune diabetes in adulthood or maturity onset diabetes, were not assessed in the phase of target definition. The algorithm will be adapted over the years, because care may change over time. For example, we tested our algorithm generated on the basis of 2012–2014 data in the SNDS in 2016 but not later, because in 2017 in France, continuous interstitial glucose monitoring devices (e.g., the FreeStyle Libre® flash glucose monitoring device) began to be fully reimbursed by the Public Health System for patients receiving intensified insulin therapy [37]. This likely modified the ranking of features on the basis of their ability to discriminate between the two types of diabetes in developing these algorithms in data sets after 2017. Another limitation of the algorithm is related to CONSTATES’ population since as generalist cohort, the type 2 diabetes cases suffering severe complications for whom the insulin treatment has been intensified are less likely to be recruited.

Conclusion

Through supervised machine learning methods, we developed a type 1/type 2 diabetes classification algorithm based on the number of reimbursements for fast-acting insulin, long-acting insulin and biguanides over the prior year. This algorithm has very good performance, as well as high applicability to prescription or medical claims databases from other countries. It also allowed us to produce the first estimate of the prevalence of type 1 and type 2 diabetes in France, in individuals 18 to 70 years of age. Artificial intelligence is a useful tool in developing methods to exploit big data sources, which may open up new areas in diabetes research and prevention.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This research was supported by Santé Publique France, the French National Public Health Agency.

The CONSTATES cohort is supported by the Caisse Nationale d’Assurance Maladie des travailleurs salariés- CNAMTS. CONSTATES is accredited as a “National Infrastructure for Biology and Health” by the governmental Investissements d’avenir program and was funded by the Agence Nationale de la Recherche (ANR-11-INBS-0002 Grant). CONSTATES 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 CONSTATES cohort. The funders did not have any role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgments

The authors want to thank the Endocrinology ReDSiam Working Group where the algorithms were discussed.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.deman.2023.100137.
References


