# Use of Artificial Intelligence for Public Health Surveillance: A case study to develop a Machine Learning-algorithm to estimate the incidence of Diabetes Mellitus in France

**Romana Haneef** ( Romana.HANEEF@santepubliquefrance.fr )
Sante publique France https://orcid.org/0000-0001-7741-0268

**Sonsoles Fuentes**
Sante publique France

**Sandrine Fosse-Edorh**
Sante publique France

**Rok Hrzic**
Maastricht University Faculty of Health, Medicine and Life Sciences: Maastricht Universitair Medisch Centrum+

**Sofiane Kab**
INSERM

**Emmanuel Cosson**
Paris 13 University: Universite Sorbonne Paris Nord

**Anne Gallay**
Sante publique France

**Methodology**

**Keywords:** Artificial intelligence, Machine learning technique, Supervise learning, Health indicator, Incidence, Diabetes Mellitus, Electronic health records and Public health surveillance

**Posted Date:** January 16th, 2021

**DOI:** https://doi.org/10.21203/rs.3.rs-139421/v1

# Abstract

**Background** The use of machine learning techniques is increasing in healthcare which allows to estimate and predict health outcomes from large administrative data sets more efficiently. The main objective of this study was to develop a generic machine learning (ML) algorithm to estimate the incidence of diabetes based on the number of reimbursements over the last 2 years.

**Methods** We selected a training data set from a population-based epidemiological cohort (i.e., CONSTANCES) linked with French National Health Database (i.e., SNDS) to develop a ML-algorithm for estimating the incidence of diabetes. To develop this algorithm, we adopted a supervised ML approach. Following steps were performed: i. selection of final data set, ii. target definition, iii. coding variables for a given window of time, iv. split final data into training and test data sets, v. variables selection, vi. training model, vii. validation of model with test data set and viii. selection of the model.

**Results** The final data set used to develop the algorithm included 44,659 participants from CONSTANCES. Out of 3,468 variables, which were similar in SNDS and CONSTANCES cohort were coded, 23 variables were selected to train different algorithms. The final algorithm to estimate the incidence of diabetes was a Linear Discriminant Analysis model based on number of reimbursements of selected variables related to biological tests, drugs, medical acts and hospitalization without a procedure over the last two years. This algorithm has a sensitivity of 62%, a specificity of 67% and an accuracy of 67% [95% CI: 0.66 – 0.68].

**Conclusions** Supervised ML is an innovative tool for the development of new methods to exploit large health administrative databases. In context of InfAct project, we have developed and applied the first time a generic ML-algorithm to estimate the incidence of diabetes for public health surveillance. The ML-algorithm we have developed, has a moderate performance. The next step is to apply this algorithm on SNDS to estimate the incidence of type 2 diabetes cases. More research is needed to apply various MLTs to estimate the incidence of various health conditions and to calculate the contribution of various risk factors on developing type 2 diabetes.


# Background

The availability of administrative data generated from different sources is increasing and the possibility to link these data sources with other databases offers unique opportunity to answer those research questions, which require a large sample size or detailed data on hard-to-reach population [1]. French National Health Data System (i.e., SNDS [*Système National de Données Santé*]) is an example of a big data/large administrative linked data set, which is used for public health surveillance in France [2]. It includes most updated, individual level health information about health insurance claims, hospital discharge and mortality of whole French population (i.e., 66 million people) [2]. However, the estimation of health indicators from linked administrative data is challenging due to several reasons such as variability in data sources and data collection methods, availability of a large number of variables, lack of

skills and capacity to analyze big data [3]. More efficient ways of analyzing health information using big data across European countries are required. In that context, the use of artificial intelligence (AI) is increasing in healthcare. Indeed AI allows to handle data with a large number of dimensions (features) and units (feature vectors) efficiently with a high precision. AI techniques offer benefits in estimation of health indicators both at individual and population levels (i.e., improving social and health policy process). Machine learning (ML) is an application of AI that provides systems the ability to learn automatically and improve from experience without being explicitly programmed [4]. Supervised learning algorithms build on a mathematical model of a set of data that contains both the inputs and the desired outputs [5]. This approach is based on the prior knowledge of what the output values for a given sample should be [6]. ML techniques have been applied for the diagnosis of certain conditions as well as outcome prediction and prognosis evaluation with high precision [7-9].

This study was carried out under the InfAct (Information for Action) project [10], which is a joint action of Member States aiming to develop a more sustainable European health information system through improving the availability of comparable, robust and policy-relevant health status data and health system performance information. InfAct gathers 40 national health authorities from 28 Member States. This study is part of a work package (WP9) focused on innovation in health information system (i.e., using data linkages and/or AI) to improve public health surveillance and health system performance for health policy process. As a first step, we have explored the current usage of these innovative techniques (i.e., data linkages and/or AI) in European countries and very few countries apply AI to estimate health indicators in their public health activities [11]. Therefore, the next step was to develop a generic approach by applying these innovative techniques to estimate the health indicators of chronic conditions for improved surveillance.

We used diabetes as a case study due to several reasons. First, it is one of the leading cause of morbidity in the world [12] and its prevalence is increasing among all ages in the European region, mostly due to increase in overweight and obesity, unhealthy diet and physical inactivity [13]. Second, a training data set using CONSTANCE cohort was already developed and used to answer various research questions for diabetes. Third, as this study is part of the InfAct project with a limited period to be completed. Fourth, estimation of incidence of diabetes cases is important to develop the prevention strategies to reduce its burden. Therefore, we decided to use this training dataset to develop a generic ML-approach.

The main objective of this study was to develop for the first time a generic ML-algorithm to estimate the incidence of diabetes based on the number of reimbursements over the last 2 years.

# Methods

### Development of the ML-algorithm

To develop ML-algorithm, we adopted a supervised ML approach. Following steps were performed: i. selection of final data set, ii. target definition, iii. coding of variables for a given window of time, iv. split

final data into training and test data sets, v. variables selection, vi. training model, vii. validation of model with test data set and viii. selection of the model.

### *i. Selection of final data set*

We selected a training data set from a population-based epidemiological cohort (i.e., CONSTANCES) to develop an algorithm to estimate the incidence of diabetes. The participants were recruited by CONSTANCES between January 1, 2012, and December 31, 2014. This cohort comprises after final completion a national representative randomly selected sample of 50,954 aged between 18 and 69 years (inclusive) and living in France [14, 15]. The participants are randomly selected from the beneficiaries of the National Health Insurance Fund (i.e. CNAM [Caisse Nationale d'Assurance Maladie]). In this cohort, data are collected using a self-administered questionnaire (SAQ) and a medical examination (MQ) and were used to define the known diabetes cases and pharmacologically-treated diabetes [16]. For known diabetes cases, in the SAQ, 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. For the pharmacologically-treated diabetes, two questions in the SAQ 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?"[16].*

After fulfilling a SAQ on health status, life style factors, socioeconomic and demographic characteristics, the participants attend to their related health screening center for a medical examination which includes: medical questionnaire, physical examination and blood sampling. This information previously collected was linked with the French National Health Data System (i.e., SNDS). We excluded pregnant women, women who declared being already diagnosed of gestational diabetes mellitus and participants without SNDS data.

### *ii. Target definition*

The diabetes status was defined according to CONSTANCES as described above. The diabetes cases treated for the first time over the 12 months before the date of SAQ were defined as incident cases (target 1). These diabetes cases included both type 1 and 2 diabetes. No diabetes cases treated over the 12 months before the date of SAQ, were defined as non-diabetes cases (target 0). The rest of diabetes cases were excluded (see Figure 1).

### *iii. Coding of variables for a given window of time*

In CONSTANCES, we only coded those variables, which were also available in the SNDS to apply the potential ML-algorithm on SNDS to estimate the incidence of diabetes. A total of 3,483 continuous variables were coded and standardized (mean= 0, standard deviation=1) over the last 24 months before the date of SAQ. The rational to have a time window of 24 months before the SAQ was to provide a long duration to study changes in diagnostic procedures, hospitalizations and drug consumption that allows

to estimate the incidence of diabetes with high accuracy. Following were the main categories of variables: number of medical consultations (50 variables), drug dispensed coded using the 5$^{th}$ level of the Anatomical Therapeutic code [ATC 05] (461 variables), biological test (747 variables), medical acts (i.e., X-ray, surgery, etc.) (2135 variables), all hospitalizations (5 variables), hospitalizations with a procedure (i.e., dialysis, radiotherapy, etc.) (5 variables), hospitalizations without a procedure (5 variables), hospitalizations related to following associated health conditions: diabetes, heart failure, stroke, heart attack, foot ulcer, lower limb amputation, ischemic heart disease, transient ischemic attack, end-stage renal failure, diabetic coma, diabetic ketoacidosis and cancer (75 variables).

### *iv. Split final data set into training and test data sets*

The final data set was randomly split into 80% as a training data set and 20% as test data set. There was a significant imbalance of number of positive target (i.e., target 1 = diabetes treated cases) over the number of negative target (i.e., target 0 = non-diabetes cases) in the training dataset. To avoid the bias in ML-algorithm and skew in class distribution, we performed a random under sampling in the target 0 group to achieve the same number of individuals in both target groups. The selection of variables and the model was performed using the training data. The test data was used solely to test the final model performance.

### *v. Variables selection*

First, we removed all variables with a variance equal to zero and then the ReliefF exp score was estimated, based on the relevance of each variable, to differentiate between target 1 and target 0. The ReliefF expRank method is noise tolerant and is not affected by features interactions [17-19]. All the variables were ranked according to the ReliefF exp score. For continuous variables, the score values range from 0 to 1 [18]. The cutoff score was 0.01 and was selected based on the visual inspection of the ordered plot of ReliefF values for all variables, called “elbow plot” approach. The variables that had a ReliefF exp score equal or more than 0.01 were included to train different models and the variables less than 0.01 were excluded.

### *Steps vi to viii Model selection and validation of the model with test data set*

The four following models [i.e., 1. Linear discriminant analysis (LDA), 2. Logistic regression (LR), 3. Flexible discriminant analysis (FDA) and 4. Decision tree model (C5)] were applied to the training data set. For each model, we compared the performance in terms of Area under the Receiver Operating Characteristics (AROC) curve. The first validation of the models was performed using k-fold [three repeats of five-fold] cross-validation on training data set. After that, the models’ performances were assessed using the testing data set. Then, we automated the model selection process by giving the computer a specific metric including sensitivity, specificity, positive predictive value, negative predictive value, F1-score and kappa. Finally, a single model was retained based on its performance and its transferability to other databases.

# Results

### *1. Final data set*

The final data set to develop the algorithm included 44,659 participants, with 81 incident diabetes cases (target 1) and 44,578 participants without diabetes (target 0) (Fig.2). The general characteristics of the final data set is described in table 1. The incident diabetes group was included older, with a higher percentage of men, treated hypertension and dyslipidemia, former smokers, a higher body mass index and a family history of diagnosed diabetes as compared to non-diabetes group.

**Table 1: General characteristics of final data set (i.e., study population)**

| Study Population | Total (N = 44,659) | N = 44,659 | |
|---|---|---|---|
| | | Target 0 (Non-diabetes cases = 44578) | Target 1 (Incident diabetes cases = 81) |
| Age in years, mean (±SD) | 47.8, ±13.2 | 47.8, ±13.2 | 57.0, ±8.2 |
| Gender, men % (n) | 46.9 (20946) | 46.9 (20896) | 61.7 (50) |
| Smoking status, % (n) | | | |
| Never smoked | 43.2 (19296) | 43.2 (19271) | 30.9 (25) |
| Former smoker | 33.1 (14772) | 33.1 (14741) | 38.3 (31) |
| Current smoker | 18.6 (8320) | 18.6 (8307) | 16.0 (13) |
| Missing | 5.1 (2271) | 5.1 (2259) | 14.8 (12) |
| Body mass index, kg/m² (mean, ±SD), (n) | 25.0, ±4.4 (43668) | 25.0, ±4.4 (43588) | 31.8, ±6.0 (80) |
| Treated hypertension, yes, % (n) | 11.3 (5031) | 11.2 (4996) | 43.2 (35) |
| Treated dyslipidemia, yes, % (n) | 8.1 (3635) | 8.1 (3609) | 32.1 (26) |
| Mother/father diagnosed with diabetes, yes, % (n) | 15.1 (6764) | 15.1 (6730) | 42.0 (34) |
| Education[i] % (n) | | | |
| No education - primary education | 3.1 (1374) | 3.1 (1366) | 9.9 (8) |
| Lower secondary education | 6.9 (3060) | 6.8 (3042) | 22.2 (18) |
| Upper secondary education | 33.5 (14942) | 33.4 (14911) | 38.3 (31) |
| Lower tertiary education | 33.0 (14728) | 33.0 (14714) | 17.3 (14) |
| Upper tertiary education | 21.7 (9709) | 21.8 (9699) | 12.3 (10) |
| Missing or other category | 1.9 (846) | 1.9 (846) | 0 (.) |
| Geographical origin, % (n) | | | |
| Metropolitan France | 87.9 (39249) | 87.9 (39177) | 88.9 (72) |
| FOT[ii] | 0.9 (381) | 0.9 (379) | 2.5 (2) |
| Europe | 4.2 (1861) | 4.2 (1859) | 2.5 (2) |
| North Africa | 2.8 (1260) | 2.8 (1257) | 3.7 (3) |
| Sub-Saharan Africa | 1.1 (503) | 1.1 (502) | 1.2 (1) |
| Asia | 0.7 (326) | 0.7 (326) | . (.) |
| Others | 1.0 (433) | 1.0 (433) | . (.) |
| Missing or don't want to answer | 1.4 (646) | 1.4 (645) | 1.2 (1) |
| Professional activity, % (n) | | | |
| Employed | 65.2 (29123) | 65.3 (29093) | 37.0 (30) |
| Unemployed | 6.1 (2721) | 6.1 (2712) | 11.1 (9) |
| Retired | 21.8 (9753) | 21.8 (9720) | 40.7 (33) |
| Student | 1.5 (653) | 1.5 (653) | 0 (.) |
| Unemployed due to disability | 0.9 (390) | 0.9 (385) | 6.2 (5) |
| No professional activity | 1.4 (603) | 1.4 (602) | 1.2 (1) |
| Missing or other category | 3.2 (1416) | 3.2 (1413) | 3.7 (3) |

SD: Standard Deviation
i. Based on the International Classification ISCED
ii. French overseas territories

## *2. Variables selection*

Out of 3,468 continuous variables coded, 23 variables (0.7%) had a ReliefF exp Score above 0.01 and were therefore selected (Fig.3).

The 23 selected variables were ranked based on their ReliefF exp score (Table 2). The first variable was the "age". The following nine were related to "number of reimbursements of biological tests performed in

last 2 years" (i.e., Alkaline Phosphatase test, Gamma Glutamyle Transferase test, Transaminases (ALAT and ASAT, TGP and TGO) blood test, Uric Acid (Uricemia) blood test, glucose blood, Creatinine level blood test, Exploration of a Lipid Anomaly (ELA) blood test, HbA1c test and C-Reactive Protein test). The next seven were related to "number of reimbursements of various non-diabetes drugs in last 2 years" (i.e., Proton pump inhibitors drugs, antidiarrheal drugs, Penicillin with broad spectrum drugs, bacterial and viral vaccines, Acetic acid derivatives, Propionic acid derivatives and Anilides (Paracetamol). The following five were related to "number of reimbursements of various medical acts" (i.e., fundus examination by biomicroscopy with contact lens, functional examination of ocular motricity, binocular vision examination, mammography and X-ray for thorax). The last one is "the total number of hospitalization without a procedure (i.e., dialysis, chemotherapy) in last 2 years".

**Table 2: List of selected variables ranked based on their ReliefF Exp Score**

| Ranked # | CATEGORIES | Independent Variables |
|---|---|---|
| 1 | AGE | Age in years |
| | | **Diabetes related variables** |
| 6 | BIOLOGICAL TESTS | Nb. of reimbursement of Glucose blood test in last 2 years |
| 9 | BIOLOGICAL TESTS | Nb. of reimbursement of HbA1c tests in last 2 years |
| 18 | MEDICAL ACTS | Nb. of reimbursement of Fundus examination by biomicroscopy with contact lens in last 2 years |
| 19 | MEDICAL ACTS | Nb. of reimbursement of Functional examination of the ocular motricity in last 2 years |
| 20 | MEDICAL ACTS | Nb. of reimbursement of Binocular vision examination in last 2 years |
| | | **Non-diabetes related variables** |
| 2 | BIOLOGICAL TESTS | Nb. of reimbursement of Alkaline Phosphatase test in last 2 years |
| 3 | BIOLOGICAL TESTS | Nb. of reimbursement of Gamma Glutamyle Transferase test in last 2 years |
| 4 | BIOLOGICAL TESTS | Nb. of reimbursement of Transaminases (ALAT and ASAT, TGP and TGO) blood test in last 2 years |
| 5 | BIOLOGICAL TESTS | Nb. of reimbursement of Uric Acid (Uricemia) blood test in last 2 years |
| 7 | BIOLOGICAL TESTS | Nb. of reimbursement of Creatinine level blood test in last 2 years |
| 8 | BIOLOGICAL TESTS | Nb. of reimbursement of Exploration of a Lipid Anomaly (ELA) blood test in last 2 years |
| 10 | BIOLOGICAL TESTS | Nb. of reimbursement of C-Reactive Protein test in last 2 years |
| 11 | DRUGS | Nb. of reimbursement of Proton pump inhibitors drugs in last 2 years |
| 12 | DRUGS | Nb. of reimbursement of other antidiarrheal drugs in last 2 years |
| 13 | DRUGS | Nb. of reimbursement of Penicillin with broad spectrum drugs in last 2 years |
| 14 | DRUGS | Nb. of reimbursement of bacterial and viral vaccines, combined (diphtheria-haemophilus influenza B-pertussis-tetanus-hepatitis B-meningococcal A + C) in last 2 years |
| 15 | DRUGS | Nb. of reimbursement of Acetic acid derivatives and related substances in last 2 years |
| 16 | DRUGS | Nb. of reimbursement of Propionic acid derivatives in last 2 years |
| 17 | DRUGS | Nb. of reimbursement of Anilides (Paracetamol) in last 2 years |
| 21 | MEDICAL ACTS | Nb. of reimbursement of Mammography, in last 2 years |
| 22 | MEDICAL ACTS | Nb. of reimbursement of X-ray thorax in the previous 2 years in last 2 years |
| 23 | HOSPITALIZATION | Total number of hospitalizations without a procedure (i.e. dialysis, chemotherapy) in last 2 years |

### *3. Algorithm to estimate the incidence of diabetes*

After the selection of variables, four different models [i.e., 1. Linear discriminant analysis (LDA), 2. Logistic regression (LR), 3. Flexible discriminant analysis (FDA) and 4. Decision tree model (C5)], were trained with the training dataset. The results of k-fold [three repeats of five-fold] cross-validation graph on training data set, were plotted under the AROC curve (Fig. 4). After that, we compared the performances of these four models using test data set to select the one based on the performance metrics (Table 3). We kept the LDA model since it showed a better performance with an accuracy of 67% with the test data set as compared to other models (Table 3).

**Table 3: Model performance evaluation with test data set, "A case study performed in 2019-20 to develop a Machine Learning-algorithm to estimate the incidence of Diabetes Mellitus in France"**

| | LDA | LR | FDA | C5 |
|---|---|---|---|---|
| Accuracy : | 0,67 | 0,65 | 0,66 | 0,64 |
| 95% CI : | (0,66-0,68) | (0,64-0,66) | (0,65-0,67) | (0,63-0,65) |
| No Information Rate : | 0,998 | 0,998 | 0,998 | 0,998 |
| P-Value [Acc > NIR] : | 1,000 | 1,000 | 1,000 | 1,000 |
| Kappa | 0,003 | 0,004 | 0,002 | 0,003 |
| McNemar's Test P-Value | <2e-16 | <2e-16 | <2e-16 | <2e-16 |
| Sensitivity | 0,625 | 0,750 | 0,563 | 0,625 |
| Specificity | 0,673 | 0,650 | 0,661 | 0,640 |
| Pos Pred Value | 0,003 | 0,004 | 0,003 | 0,003 |
| Neg Pred Value | 0,999 | 0,999 | 0,999 | 0,999 |
| F1-statistics | 2,50 | 3,0 | 2,252 | 2,50 |
| Detection Rate | 0,001 | 0,001 | 0,001 | 0,001 |
| Balanced Accuracy | 0,649 | 0,700 | 0,612 | 0,633 |

### *4. Distribution of means of selected variables in test data set*

After the selection of LDA model, the 23 selected variables were trained with the test data set (20% of final data set 44,659 = 8,931). We compared the distribution of means of these continuous variables among two groups: incident diabetes cases (i.e., 2,921) and non- diabetes cases (i.e., 6,010) using LDA algorithm in the test data set (Table 4). The mean distribution of all selected variables related to the number of reimbursements of biological tests, medicines not used for diabetes treatment and medical acts performed in last 2 years, was significantly higher in the incident diabetes group than in non-diabetes group. For example, the age was the first ranked variable with 0.04 ReliefF exp score among 23 selected variables and was highly discriminant in the incident diabetes group. The mean age of patients in diabetes group was 57 years old as compared to 44 years old in non-diabetes group (Table 4).

**Table 4: Distribution of means of variables selected for the final algorithm in test data set using Linear Discriminant Analysis (LDA) model**

| Ranked ≠ | Categories | Variables | Mean (incident diabetes group) n= 2921 | Mean (non-diabetes group) n=6010 | P value (Student's t-test) |
|---|---|---|---|---|---|
| 1 | AGE | Age in years | **56.70*** | 43.69 | <0.0001 |
| 2 | BIOLOGICAL TESTS | Nb. of reimbursement of Alkaline Phosphatase test in last 2 years | **0.36*** | 0.20 | <0.0001 |
| 3 | BIOLOGICAL TESTS | Nb. of reimbursement of Gamma Glutamyle Transferase test in last 2 years | **1.05*** | 0.40 | <0.0001 |
| 4 | BIOLOGICAL TESTS | Nb. of reimbursement of Transaminases (ALAT and ASAT, TGP and TGO) blood test in last 2 years | **1.33*** | 0.63 | <0.0001 |
| 5 | BIOLOGICAL TESTS | Nb. Of reimbursement of Uric Acid (Uricemia) blood test in last 2 years | **0.47*** | 0.22 | <0.0001 |
| 6 | BIOLOGICAL TESTS | Nb. of reimbursement of Glucose blood test in last 2 years | **1.29*** | 0.74 | <0.0001 |
| 7 | BIOLOGICAL TESTS | Nb. Of reimbursement of Creatinine level blood test in last 2 years | **1.09*** | 0.39 | <0.0001 |
| 8 | BIOLOGICAL TESTS | Nb. of reimbursement of Exploration of a Lipid Anomaly (ELA) blood test in last 2 years | **1.20*** | 0.65 | <0.0001 |
| 9 | BIOLOGICAL TESTS | Nb. Of reimbursement of HbA1c tests in last 2 years | **0.20*** | 0.04 | <0.0001 |
| 10 | BIOLOGICAL TESTS | Nb. Of reimbursement of C-Reactive Protein test in last 2 years | **1.16*** | 0.38 | <0.0001 |
| 11 | DRUGS | Nb. Of reimbursement of Proton pump inhibitors drugs in last 2 years | **3.67*** | 0.54 | <0.0001 |
| 12 | DRUGS | Nb. Of reimbursement of other antidiarrheal drugs in last 2 years | **0.21*** | 0.03 | <0.0001 |
| 13 | DRUGS | Nb. of reimbursement of Penicillin with broad spectrum drugs in last 2 years | **0.74*** | 0.23 | <0.0001 |
| 14 | DRUGS | Nb. Of reimbursement of bacterial and viral vaccines, combined (diphtheria-haemophilus influenza B-pertussis-tetanus-hepatitis B-meningococcal A + C) in last 2 years | **0.23*** | 0.13 | <0.0001 |
| 15 | DRUGS | Nb. of reimbursement of Acetic acid derivatives and related substances in last 2 years | **0.57*** | 0.19 | <0.0001 |
| 16 | DRUGS | Nb. of reimbursement of Propionic acid derivatives in last 2 years | **1.53*** | 1.07 | <0.0001 |
| 17 | DRUGS | Nb. of reimbursement of Anilides (Paracetamol) in last 2 years | **3.65*** | 1.69 | <0.0001 |
| 18 | MEDICAL ACTS | Nb. of reimbursement of Fundus examination by biomicroscopy with contact lens in last 2 years | **0.30*** | 0.03 | <0.0001 |
| 19 | MEDICAL ACTS | Nb. of reimbursement of Functional examination of the ocular motricity in last 2 years | **0.19*** | 0.08 | <0.0001 |
| 20 | MEDICAL ACTS | Nb. of reimbursement of Binocular vision examination in last 2 years | **0.32*** | 0.10 | <0.0001 |
| 21 | MEDICAL ACTS | Nb. of reimbursement of Mammography in last 2 years | **0.12*** | 0.10 | 0.0062 |
| 22 | MEDICAL ACTS | Nb. Of reimbursement of X-ray thorax in the previous 2 years in last 2 years | **0.20*** | 0.07 | <0.0001 |
| 23 | HOSPITALIZATION | Total number of hospitalizations without a procedure (i.e. dialysis, chemotherapy) in last 2 years | **0.89*** | 0.26 | <0.0001 |
| * Highest mean | | | | | |

Following the age variable, nine other features selected, related to the mean number of reimbursements of biological tests, were more discriminant in incident diabetes group than in non-diabetes group. These biological tests were performed to measure the normal values of certain enzymes, proteins, glucose and uric acid in the blood to check the normal functions of liver, kidney, pancreas and other organs. For example, the mean number of reimbursement of blood glucose test in last two years was 1.82 times more discriminant in diabetes group than in non-diabetes group. The following group of features was the mean number of reimbursements of drugs. There were seven drugs and their mean number of reimbursements

in last 2 years was more discriminant in incident diabetes group than in non-diabetes group. In the category of medical acts, there were three following features more discriminant in incident diabetes group: mean number of reimbursements of examination of fundus by biomicroscopy with contact lens, ocular motricity and binocular vision in last 2 years.

There were seven unusual features selected by the ML-algorithm and were discriminant in incident diabetes group: mean number of reimbursements of broad-spectrum penicillin, vaccines, propionic acid, Anilides (Paracetamol), mammography, X-ray for thorax and mean number of hospitalizations without any procedure.

# Discussion

We have developed an algorithm based on the supervised ML approach to estimate the incidence of diabetes using a training data set from a cohort study. This algorithm (i.e., LDA model) was built on 23 selected variables from the CONSTANCES based on the number of reimbursements over the last 2 years to estimate the incidence of diabetes. This algorithm showed a moderate performance in predicting the incidence of diabetes cases with a sensitivity of 62% and an accuracy of 67%. Among 23 selected variables, six were related to diabetes that were expected, such as age and Glucose blood test. Whereas 17 other variables were not directly related to diabetes and were more discriminant in incident diabetes group than in non-diabetes group such as Proton pump inhibitors drug.

### *Main limitations of the ML-algorithm*

Our study has highlighted that there were two discriminant features related to diabetes in LDA model i.e., mean number of reimbursements of glucose blood and HbA1c tests, which could potentially characterize the incident diabetes cases. In France, the screening recommendations for diabetes are based on the glucose blood test. HbA1c is only recommended for the management of diabetes but not for diagnoses. In 2009 and 2010, the WHO has introduced HbA1c as an alternative method to diagnose diabetes that has been adopted by many countries since this date. The ophthalmologic problems such as glaucoma, cataract, ocular movement disorders, etc., are the main complications of diabetes. Therefore, the increase frequency of medical acts performed as a result of diabetes related complications such as visual functions allowed to better characterize incident diabetes cases. Moreover, the increased use of non-diabetic drugs along with mentioned biological tests in incident diabetes group may explain potentially the pre-existing comorbidity of cardiovascular or gastrointestinal diseases.

### *Implications and perspectives for future research*

This innovative approach has been applied to two further studies: i. to classify and to estimate the prevalence of type 1 and type 2 diabetes cases [21] and, ii. to identify the number of undiagnosed diabetes cases ML algorithms in the SNDS (on going). For the first study, ML-algorithm developed has a sensitivity of 100% and specificity of 97%, and for the second study, the sensitivity is 71% and specificity is 61%.

The next step is to apply this algorithm on SNDS to estimate the incidence of type 2 diabetes cases. We recommend further research for following perspectives using ML-techniques: *first* to use different time windows (for example 6 months, 12 months or 16 months) to code variables and to explore their impact on estimates, *second* to predict the trend of diabetes over time and *third,* to estimate the contribution of determinants of diabetes such as BMI, dietary habits and physical activity, on developing type 2 diabetes using ML approaches.

# Conclusions

The use of MLT to analyze large administrative databases (health and non-health related data sources) is increasing across European countries in order to improve the public health surveillance and health policy process. Supervised machine learning is an innovative methodology for the development of algorithms to exploit large health administrative databases. It was the first step that we have developed a generic ML-algorithm with a moderate performance to estimate the incidence of diabetes in a training data set. The results of this study have highlighted important methodological steps to apply MLTs and their implications on large health administrative databases. The next step is to apply this algorithm on SNDS to estimate the incidence of type 2 diabetes cases. More research is needed to apply various MLTs to estimate the incidence of various health conditions and to calculate the contribution of various risk factors on developing type 2 diabetes.

# List Of Abbreviations

ML: Machine Learning

SNDS: Système National de Données Santé: French National Health Database

CONSTANCE: A population-based epidemiological cohort

AI: Artificial Intelligence

InfAct: Information for Action i.e., a joint action of Member States to establish a sustainable European health information system.

WP: Work Package

CNAMTS: Caisse Nationale de l'Assurance Maladies des Travailleurs Salaries

SAQ: Self-administered Questionnaire

ATC: Anatomical Therapeutic Code

LDA: Linear Discriminant Analysis

LR: Logistic Regression

FDA: Flexible Discriminant Analysis

C5: Decision Tree

AROC: Area under Receiver Operating Characteristics

Pos Pred Value: Positive Predictive Value

Neg Pred Value: Negative Predictive Value

95%CI: 95% Confidence Interval

BMI: Body Mass Index

WHO: World Health Organization

# Declarations

### Ethics approval and consent to participate

Not applicable

### Consent for publication

All authors gave the consent for publication.

### Availability of data and materials

Not applicable

### Competing interests

All other authors declare that they have no competing interests related to the work.


### Funding

This research has been carried out in the context of the project ‘801553 / InfAct’ which has received funding from the European Union’s Health Programme (2014-2020).


### Authors’ contributions

Conceived and designed the survey: RH SF RHrzic AG. Performed the study: RH SF SK RHrzic. Analyzed the data: RH SF SK. Interpretation of the results: All authors contribute to the interpretation of the results. Contributed to the writing of the manuscript: All authors contributed to the writing of the manuscript. All authors read and approved the final manuscript.


### Acknowledgements

We acknowledge Le Marie Zins (Responsible for Constance cohort) for her kind support to access and use the data from this cohort.

# References

1. Harron K, Dibben C, Boyd J, Hjern A, Azimaee M, Barreto ML, Goldstein H: Challenges in administrative data linkage for research. *Big Data Soc* 2017, 4(2):2053951717745678-2053951717745678.
2. Tuppin P RJ, Constantinou P et al Value of a national administrative database to guide public decisions: From the. *Rev Epidemiol Sante Publique* 2017, 65(4):S149-S167.
3. Bradley CJ, Penberthy L, Devers KJ, Holden DJ: Health Services Research and Data Linkages: Issues, Methods, and Directions for the Future. *Health Services Research* 2010, 45(5p2):1468-1488.
4. Machine Learning: https://www.expertsystem.com/machine-learning-definition/. 2017.
5. Russell SJ: Artificial Intelligence: A Modern Approach: https://ifarus.com/artificial-intelligence-stuart-russell. *University Text Book (Third Edition)* 2009.
6. Soni D: Supervised vs Unsupervised Learning: https://towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8d. 2018.
7. Jha S, Topol EJ: Adapting to Artificial Intelligence: Radiologists and Pathologists as Information Specialists. *JAMA* 2016, 316(22):2353-2354.
8. Patel VL, Shortliffe EH, Stefanelli M, Szolovits P, Berthold MR, Bellazzi R, Abu-Hanna A: The coming of age of artificial intelligence in medicine. *Artificial Intelligence in Medicine* 2009, 46(1):5-17.
9. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I: Machine Learning and Data Mining Methods in Diabetes Research. *Comput Struct Biotechnol J* 2017, 15:104-116.
10. Joint Action on Health Information: https://www.inf-act.eu/. 2018.
11. Haneef R, Delnord M, Vernay M, Bauchet E, Gaidelyte R, Van Oyen H, Or Z, Pérez-Gómez B, Palmieri L, Achterberg P *et al*: Innovative use of data sources: a cross-sectional study of data linkage and artificial intelligence practices across European countries. *Archives of Public Health* 2020, 78(1):55.
12. Cho NH, Shaw JE, Karuranga S, Huang Y, da Rocha Fernandes JD, Ohlrogge AW, Malanda B: IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes Research and Clinical Practice* 2018, 138:271-281.
13. WHO-Europe: The challenges of diabetes: http://www.euro.who.int/en/health-topics/noncommunicable-diseases/diabetes/data-and-statistics.
14. CONSTANCES: http://www.constances.fr/index_EN.php#assets. 2019.
15. Zins M, Goldberg M, team C: The French CONSTANCES population-based cohort: design, inclusion and follow-up. *European Journal of Epidemiology* 2015, 30(12):1317-1328.
16. Fuentes S, Cosson E, Mandereau-Bruno L, Fagot-Campagna A, Bernillon P, Goldberg M, Fosse-Edorh S, Group C-D: Identifying diabetes cases in health administrative databases: a validation study based on a large French cohort. *International Journal of Public Health* 2019, 64(3):441-450.

17. Chaix B, Kestens Y, Bean K, Leal C, Karusisi N, Meghiref K, Burban J, Fon Sing M, Perchoux C, Thomas F *et al*: Cohort profile: residential and non-residential environments, individual activity spaces and cardiovascular risk factors and diseases–the RECORD Cohort Study. *Int J Epidemiol* 2012, 41(5):1283-1292.
18. Kononenko MR-SI: An adaption of Relief for attribute estimation in regression: http://www.clopinet.com/isabelle/Projects/reading/robnik97-icml.pdf. 1997.
19. Devaney M, Ram A: Machine Learning: Proceedings of the Fourteenth International Conference, Nashville, TN, July 1997 (to appear). 2004.
20. Çalişir D, Doğantekin E: An automatic diabetes diagnosis system based on LDA-Wavelet Support Vector Machine Classifier. *Expert Syst Appl* 2011, 38(7):8311–8315.
21. Fuentes S, Hrzic R, Haneef R, Kab S, Fosse-Edorh S, Cosson E: Development of type 1/type 2 classification algorithm through machine learning methods and its application to surveillance using a nationwide database in France In: *Diabetologia.* 2020.

# Figures

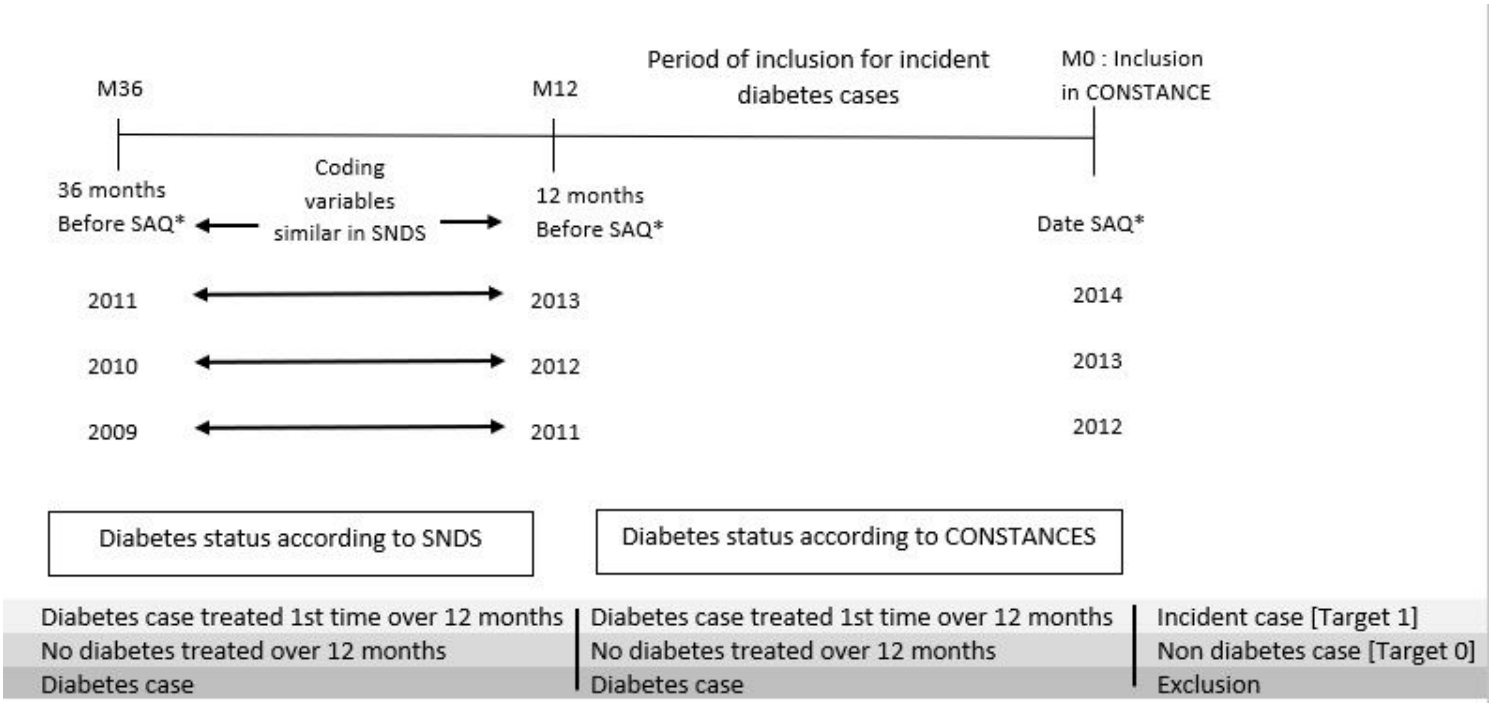
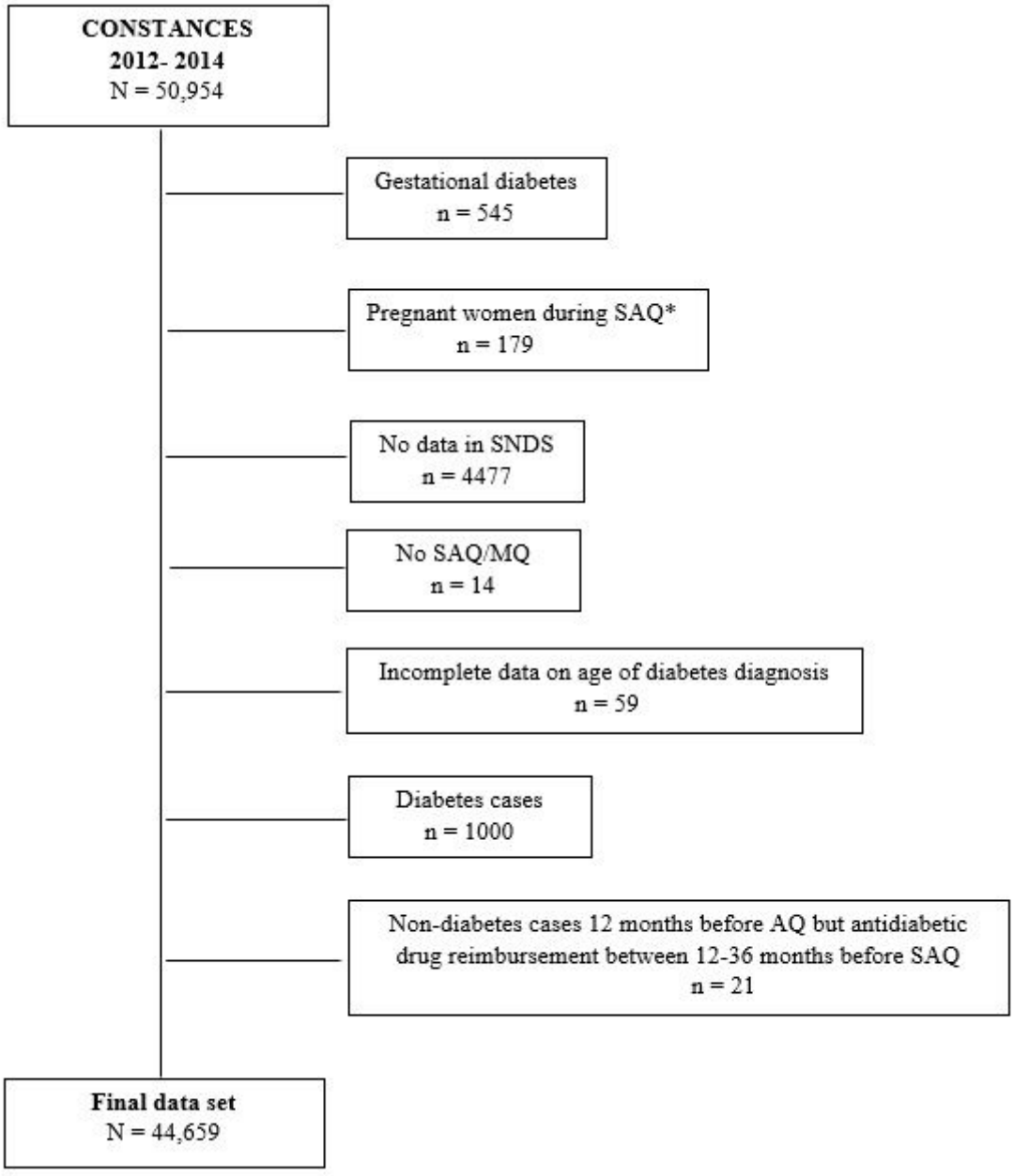



**Figure 1**

Target definition in CONSTANCES Cohort, “A case study performed in 2019-20 to develop a Machine Learning-algorithm to estimate the incidence of Diabetes Mellitus in France”




## Figure 2

Flow chart for the selection of the final data set from CONSTANCE Cohort, “A case study performed in 2019-20 to develop a Machine Learning-algorithm to estimate the incidence of Diabetes Mellitus in France”

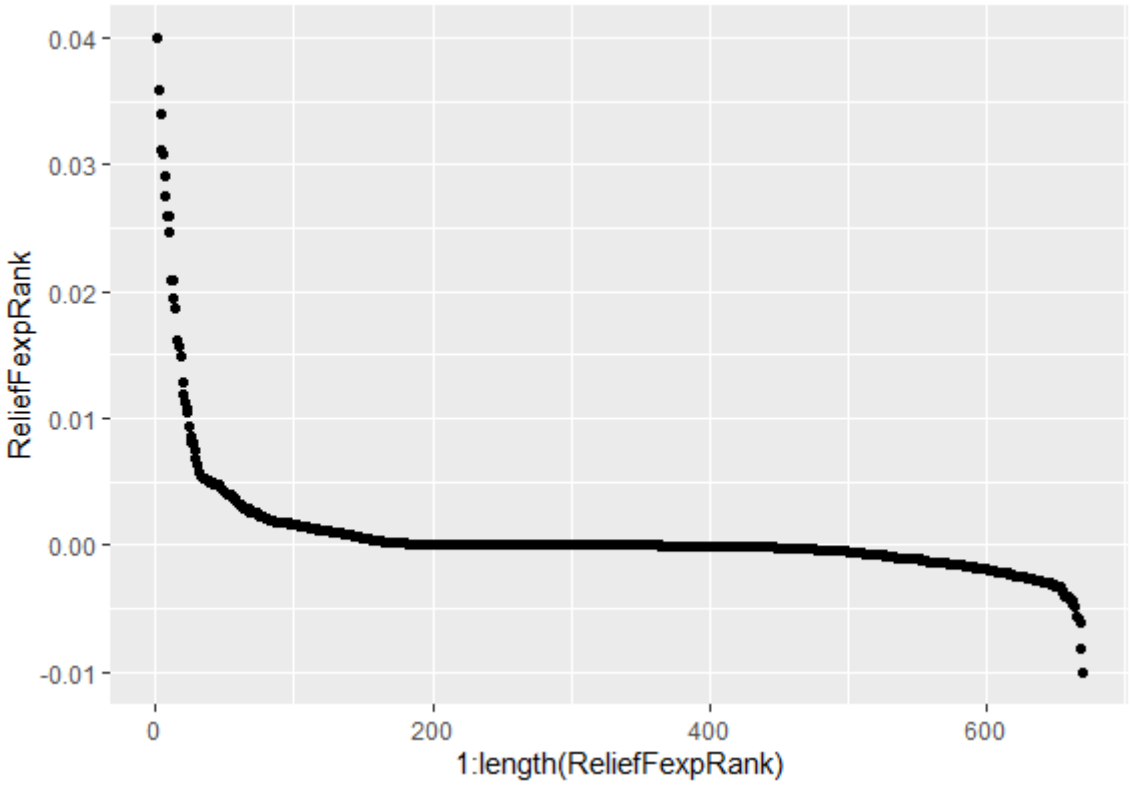



**Figure 3**

Variables selection based on ReliefF Exp Score , “A case study performed in 2019-20 to develop a Machine Learning-algorithm to estimate the incidence of Diabetes Mellitus in France”

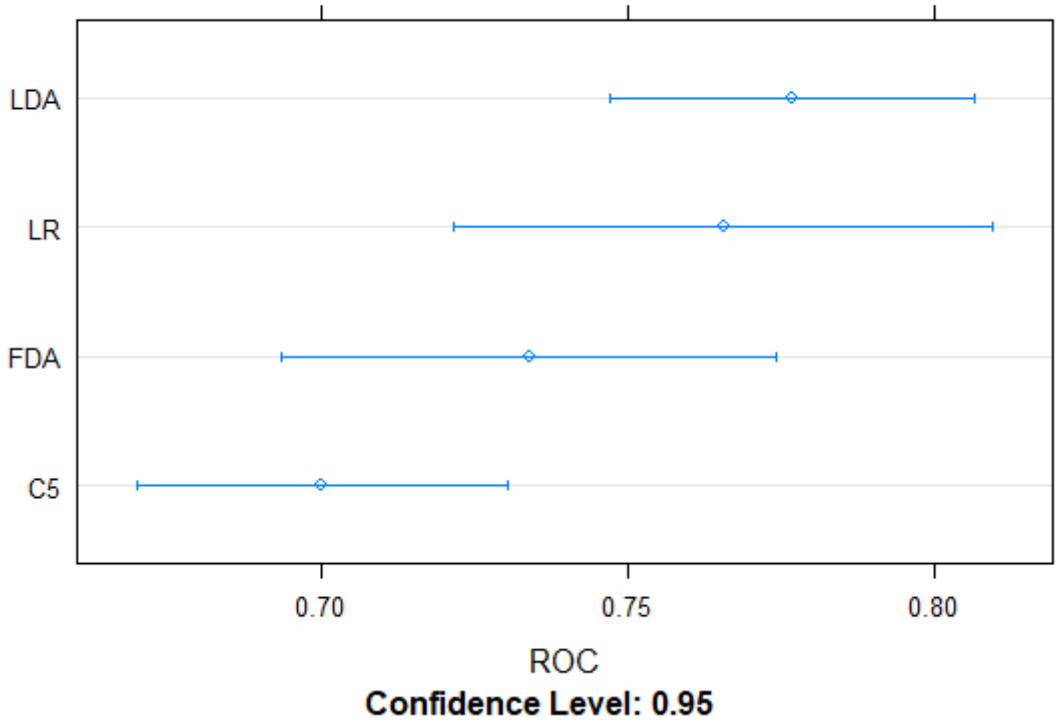



**Figure 4**

k-fold cross-validation graph using training data set under AROC curve, “A case study performed in 2019-20 to develop a Machine Learning-algorithm to estimate the incidence of Diabetes Mellitus in France”