INFACT
Joint Action on Health Information

Deliverable 9.2 (Part A)

A generic method study using Machine Learning Technique

October 23, 2020

Santé publique France

HIGIENOS INSTITUTAS

Submission Date: October 23, 2020

WP9 Lead: Department of Non-Communicable Disease and Injury, Santé Publique France, France
WP9 Co-Lead: Health Information Centre and Institute of Hygiene, Lithuania
Sciensano | Rue Juliette Wytsmanstraat 14 |
1050 Brussels | Belgium| e-mail: infact.coordination@sciensano.be|
Website: www.inf-act.eu|Twitter: @JA_InfAct

This project is co-funded by the Health Programme of the European Union

Table of Contents

Executive summary

Background

The use of machine learning techniques is increasing in healthcare, which allows estimating and predicting 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. The following steps were performed: i. selection of final data set, ii. target definition, iii. coding features/variables for a given window of time, iv. split final data into training and test data sets, v. features/variables selection, vi. training model/algorithm, vii. validation of model/algorithm with test data set and viii. selection of the model/algorithm.

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 a 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 the context of InfAct project, this study have highlighted important methodological steps to apply MLTs. This was the first step that we have developed a generic ML-algorithm with a moderate performance to estimate the incidence of diabetes using a training data set. The next step is to apply this algorithm on SNDS (i.e., National health administrative database) 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 estimate the impact of various risk factors on developing type 2 diabetes.

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

Key points

- A generic ML-algorithm to estimate the incidence of diabetes for public health surveillance has been developed.

- More research is needed to apply various MLTs to estimate the incidence of various health conditions and to predict the impact of various risk factors on developing type 2 diabetes.

I. Part A: A generic method study using Machine Learning Technique

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

II. 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 updated, individual-level health information about health insurance claims, hospital discharge and mortality of the 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 handling data with a large number of dimensions (features) and units (feature vectors) efficiently with high precision. AI techniques offer benefits in the 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) [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 causes of morbidity in the world [12] and its prevalence is increasing among all ages in the European Region, mostly due to increases in overweight and obesity, unhealthy diet and physical inactivity [13]. Second, a data set using CONSTANCE cohort was already developed and used to answer various research questions for diabetes. Third was the time constraints. InfAct project has limited timelines and this study has to be completed within the project period. Fourth, the estimation of diabetes' incidence is important to develop prevention strategies to reduce its burden. Therefore, we decided to use this 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.

III. Method

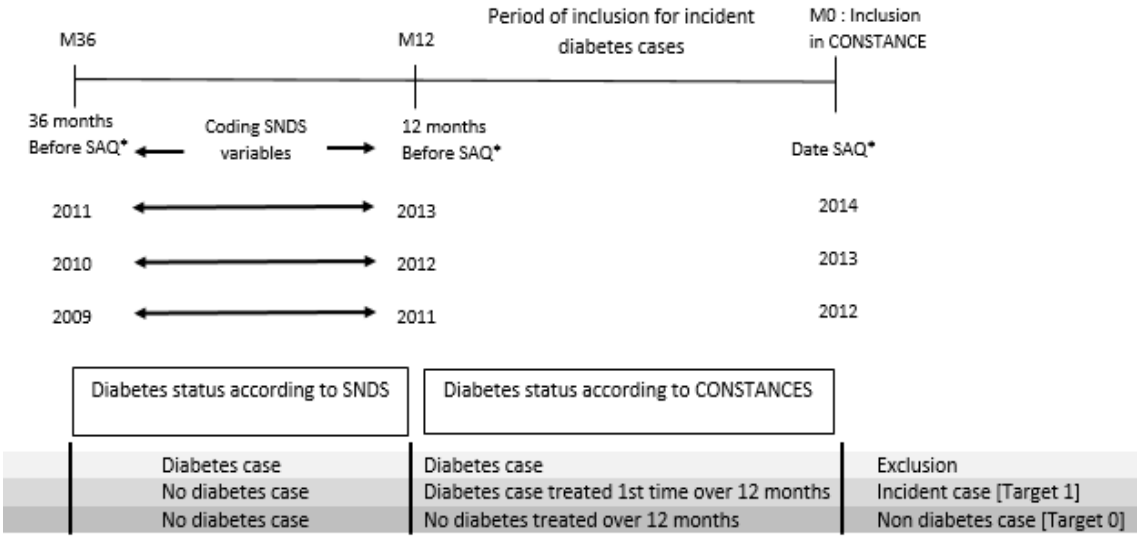
Development of the ML-algorithm

To develop ML-algorithm, we adopted a supervised machine learning approach. The following steps were performed: i. selection of final data set, ii. case/target definition, iii. coding of features/variables for a given window of time, iv. split final data into training and test data sets, v. features/variables selection, vi. training model/algorithm, vii. validation of model/algorithm with test data set and viii. selection of the model/algorithm.
We selected a final 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 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. CNAMTS [Caisse Nationale d'Assurance Maladie des travailleurs salaries]). In this cohort, data are collected using a self-administered questionnaire (SAQ) and a medical examination (MQ) and are used to define the known diabetes cases and pharmacologically-treated diabetes [16]. For known diabetes cases, in the SAQ, participants reported having 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 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 the SAQ on health status, lifestyle 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 with gestational diabetes mellitus and participants without SNDS data.

i. 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 treated over the 12 months before the date of SAQ, were defined as non-diabetes cases (target 0). The rest of the diabetes cases were excluded (see Figure 1).



**SAQ: Self-administered Questionnaire*

Figure 1: Target definition

ii. 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 rationale to have a time window of 24 months before the SAQ was to provide a long duration to evaluate the diagnostic procedures, hospitalizations and drug consumption that allows to take into account various changes over time and 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 5th 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).

iii. 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 a test data set. There was an imbalance of a 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 features/variables and the model was performed using the training data. The test data was used solely to test the final model performance.

iv. Features/variables selection

First, we removed all variables with a variance equal to zero and then the ReliefExp score was estimated, based on the relevance of each variable, to differentiate between target 1 and target 0. The ReliefExp method is noise-tolerant and is not affected by features interactions [17-19]. All the features were ranked according to the ReliefExp score.

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. After the first validation of the algorithms/models (k-fold [three repeats of

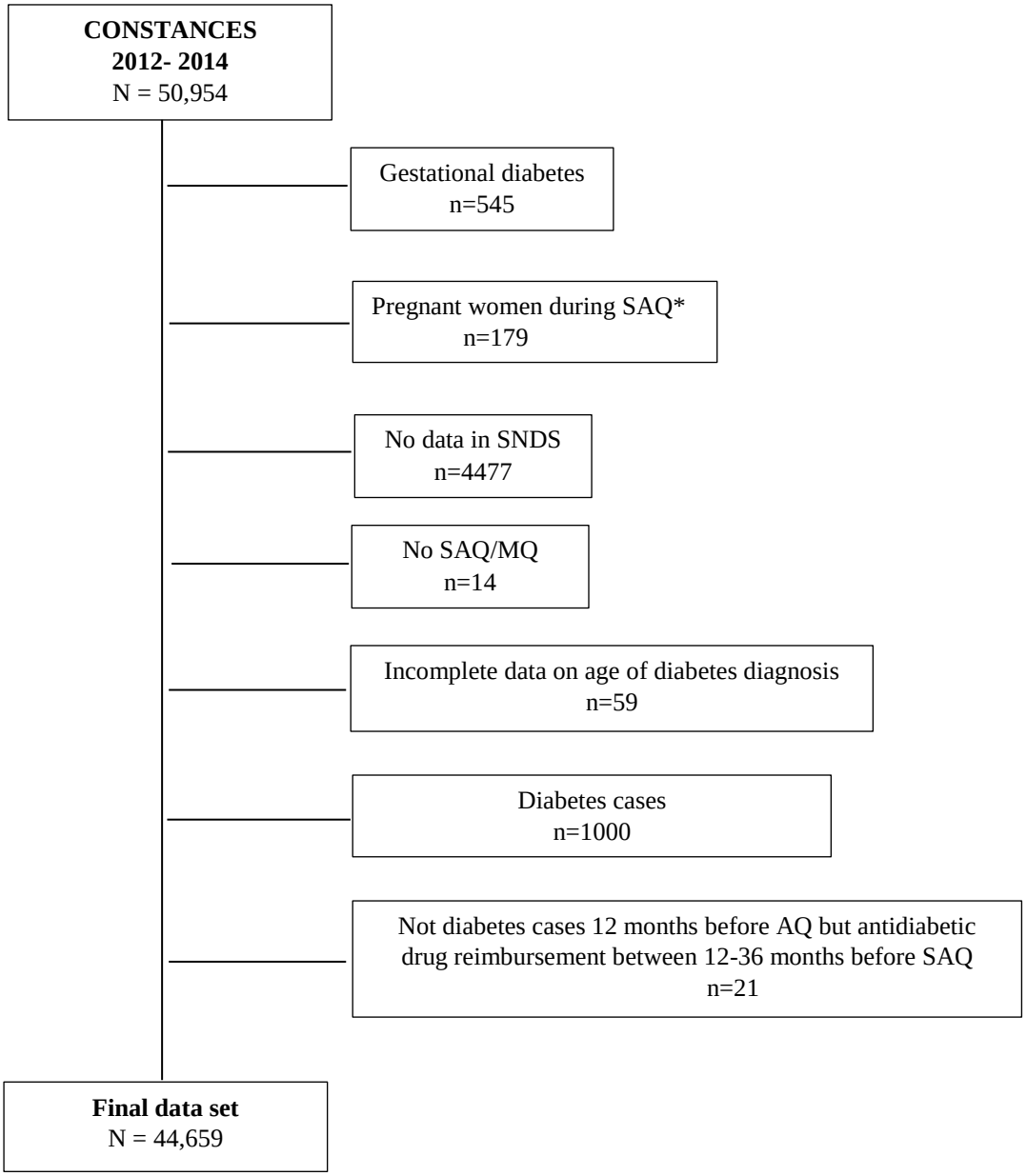
five-fold] cross-validation) using training data set, the algorithms' performances were assessed using the testing data set. Then, we automated the algorithm selection process by giving the computer a specific metric including sensitivity, specificity, positive predictive value, negative predictive value, F1-score and kappa. Finally, a single model was retained based on its performance and its transferability to other databases.

IV. 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 are described in table 1. The incident diabetes group was older, with a higher percentage of men, treated hypertension and dyslipidemia, former smokers, with higher body mass index and family history of diagnosed diabetes compared to non-diabetes group.

Fig. 2: Flow chart for the selection of the final data set



**SAQ= Self-administered Questionnaire*
MQ= Medical Questionnaire

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

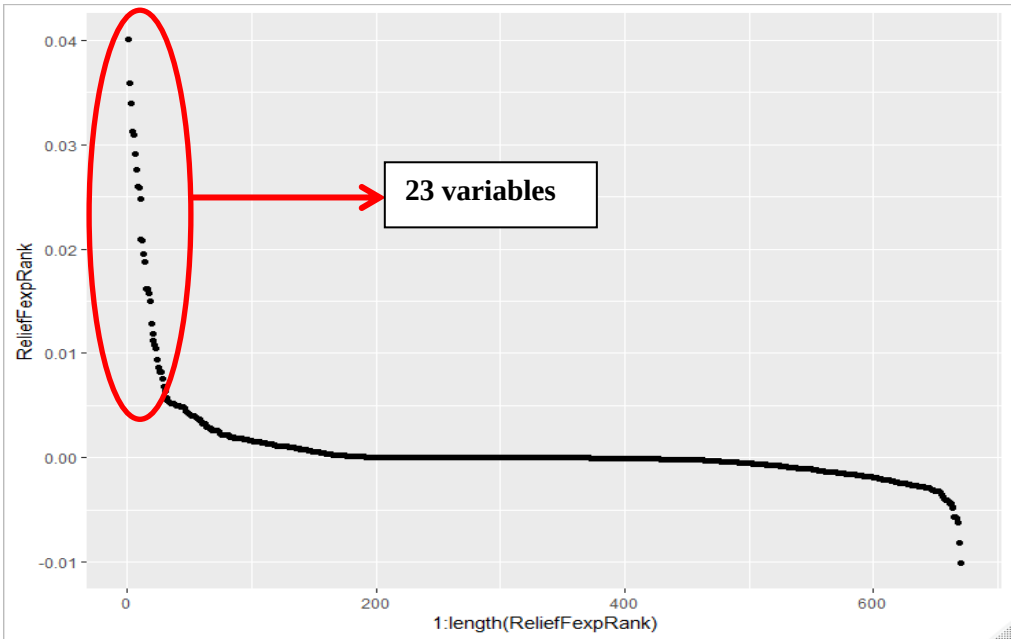
Study Population (i.e., CONSTANCES)	Total (N = 44,659)	N = 44,659	
		Target 0 (Non-incident diabetes cases = 44578)	Target 1 (Incident diabetes cases = 81)
Age, mean (±SD)	47.8, ±13.2	47.8, ±13.2	57.0, ±8.2
Gender, men % (n)	46.9 (20946)	46.9 (20896)	61.7 (50)
Smoking status, % (n)			
Never smoked	43.2 (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/m^2 (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. Features/variables selection for the prediction of diabetes cases

Out of 3468 variables coded, only 23 variables (0.7%) were selected because their ReliefExp Score was above 0.01 (Fig.3).

Fig. 3: Features/variables selection based on ReliefExp Score



The 23 selected variables were ranked based on their ReliefExp Score (Table 2). The first feature was "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 ReliefExp 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 (i.e., three repeats of five folds of training data set) cross-validation graph using training data set were plotted area under the ROC curve (Fig.3). We compared the performances of these four models to select the one based on the performance metrics using the test data set (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).

Fig. 3: k-fold cross-validation using training data set (area under the ROC curve)

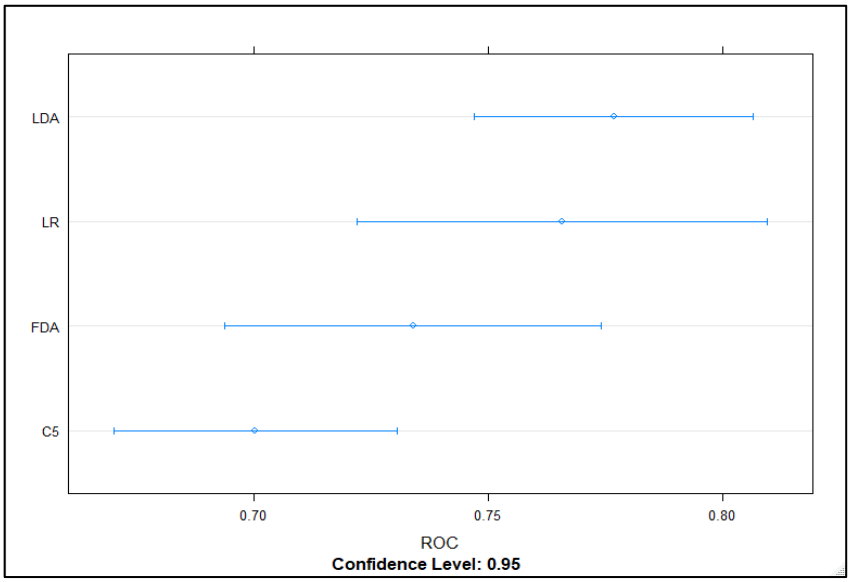


Table 3: Model performance evaluation with test data set

	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., 2,889) and non-diabetes cases (i.e., 6,042) 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 the last 2 years, was higher in the incident diabetes group than in non-diabetes group. For example, the age was the first feature with the highest ReliefExp Score among selected features and was highly discriminant in the incident diabetes group. The mean age of patients in the diabetes group was 56 years old as compared to 47 years old in the non-diabetes group (Table 4).

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

S/No	Categories	Variables	Mean (incident diabetes group)	Mean (non-incident diabetes group)
1	AGE	Age in years	**55,99***	47,40
2	BIOLOGICAL TESTS	Nb. of reimbursement of Alkaline Phosphatase test in last 2 years	**0,51***	0,31
3	BIOLOGICAL TESTS	Nb. of reimbursement of Gamma Glutamyle Transferase test in last 2 years	**1,09***	0,48
4	BIOLOGICAL TESTS	Nb. of reimbursement of Transaminases (ALAT and ASAT, TGP and TGO) blood test in last 2 years	**1,35***	0,78
5	BIOLOGICAL TESTS	Nb. Of reimbursement of Uric Acid (Uricemia) blood test in last 2 years	**0,63***	0,34
6	BIOLOGICAL TESTS	Nb. of reimbursement of Glucose blood test in last 2 years	**1,82***	0,89
7	BIOLOGICAL TESTS	Nb. Of reimbursement of Creatinine level blood test in last 2 years	**1,02***	0,54
8	BIOLOGICAL TESTS	Nb. of reimbursement of Exploration of a Lipid Anomaly (ELA) blood test in last 2 years	**1,38***	0,75
9	BIOLOGICAL TESTS	Nb. Of reimbursement of HBA1C tests in last 2 years	**0,88***	0,14
10	BIOLOGICAL TESTS	Nb. Of reimbursement of C-Reactive Protein test in last 2 years	**1,03***	0,42
11	DRUGS	Nb. Of reimbursement of Proton pump inhibitors drugs in last 2 years	**4,34***	0,69
12	DRUGS	Nb. Of reimbursement of other antidiarrheal drugs in last 2 years	**0,14***	0,03
13	DRUGS	Nb. of reimbursement of Penicillin with broad spectrum drugs in last 2 years	**0,48***	0,18
14	DRUGS	Nb. Of reimbursement of bacterial and viral vaccines, combined (diphtheria-haemophilus influenza B-pertussis-tetanus-hepatitis B-meningococcus A + C) in last 2 years	**0,15***	0,08

15	DRUGS	Nb. of reimbursement of Acetic acid derivatives and related substances in last 2 years	**0,86***	0,31
16	DRUGS	Nb. of reimbursement of Propionic acid derivatives in last 2 years	**1,25***	1,08
17	DRUGS	Nb. of reimbursement of Anilides (Paracetamol) in last 2 years	**4,43***	1,62
18	MEDICAL ACTS	Nb. of reimbursement of Fundus examination by biomicroscopy with contact lens in last 2 years	**0,18***	0,03
19	MEDICAL ACTS	Nb. of reimbursement of Functional examination of the ocular motricity in last 2 years	**0,17***	0,06
20	MEDICAL ACTS	Nb. of reimbursement of Binocular vision examination in last 2 years	**0,20***	0,08
21	MEDICAL ACTS	Nb. of reimbursement of Mammography in last 2 years	**0,31***	0,11
22	MEDICAL ACTS	Nb. Of reimbursement of X-ray thorax in the previous 2 years in last 2 years	**0,34***	0,05
23	HOSPITALIZATION	Total number of hospitalizations without a procedure (i.e. dialysis, chemotherapy) in last 2 years	**0,91***	0,35
* Highest mean				

Following the age variable, nine other features selected, related to the mean number of reimbursements of biological tests, were more discriminant in the 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 tests in the last two years was 1.82 times more discriminant in the 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 the last 2 years was more discriminant in the incident diabetes group than in non-diabetes group. In the category of medical acts, there were three following features more discriminant in the 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 the 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.

V. 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, such as age and Glucose blood test. Whereas 17 other variables were not directly related to diabetes and were more discriminant in the incident diabetes group than in the non-diabetes group such as Proton pump inhibitors drug.
The LDA model has been used for features selection and dimensionality reduction for diabetes diagnosis [20]. In our study, the algorithm has shown a moderate performance in accuracy and sensitivity metrics.

Main limitations of the ML-algorithm

This study has some limitations: *first,* a small number of diabetes-treated cases in the final data set, which could be related to the lack of older population in the final data set of this cohort CONSTANCES. Participation in a cohort like CONSTANCES is challenging and demands individuals' additional time to take part in health examinations. People in less good health and having co-morbidities (including both old and young people) require regular health check-ups, therefore maybe less motivated to participate in cohort studies. Thus, it is required to wait for volunteers to include various age groups to have more incident cases. The risk of developing diabetes increases with age, therefore by including a larger number of older people in the final dataset, the performance of this algorithm may be improved. *Second,* the time window of the previous two years used to code the variables was too long. We included a longer window to better evaluate the consumption of diagnostic procedures, number of hospitalizations and drug consumption that allows to take into account various changes over time and to estimate the incidence with high accuracy. More research is needed to explore different time windows and their impact on the accuracy level of estimates. The *third* is related to diabetes disease nature, which is a complex medical condition with two major clinical types of diabetes, type 1 diabetes and type 2 diabetes. The pathology and dynamics of developing these two types of diabetes are very different. Type 1 diabetes is thought to be due to autoimmunological destruction of the Langerhans Islets hosting pancreatic-ß-cells and it is diagnosed at a very early stage of life. Whereas the main causes of type 2 diabetes are due to lifestyle, physical activity, dietary habits and genetic, and develop usually at later than 50 years of life. In our study, we defined the pharmacologically treated diabetes cases as target 1 and non-diabetes cases as target 0. However, we did not explicitly define the pharmacologically treated diabetes cases to be further characterized as type 1 and type 2. With the inclusion of this information in the model, the accuracy level of the model could be enhanced. Despite these limitations of this ML-algorithm, this study has some strengths: *first, it* is using a supervised machine learning approach, we have developed an innovative methodology and could be applied to address other research questions. *Second,* this approach allows to reduce the dimensionally of a large number of variables (i.e.,

3,468) and identifies the most relevant variables (i.e., 23/3,468 = 0.7%) to the desired outcomes more efficiently. *Third,* it allows identifying new variables and enriching the information to estimate the health indicators.
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 increased 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 the mentioned biological tests in the 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 (ongoing). 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 the following perspectives using ML-techniques: *first* to help estimating and predicting the trend of diabetes over time and *second,* to improve the development the prevention strategies, using the information on determinants of diabetes such as BMI, dietary habits and physical activity, predict/estimate their impact on developing type 2 diabetes.

VI. Conclusions

The use of MLT to analyze large administrative databases (health and non-health related data sources) is increasing across European countries 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. The results of this study have highlighted important methodological steps to apply MLTs. This was the first step that we have developed a generic ML-algorithm with a moderate performance to estimate the incidence of

diabetes using a training data set. The next step is to apply this algorithm on SNDS (i.e., National health administrative database) 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 estimate the impact of various risk factors on developing type 2 diabetes.

VII. List of abbreviations

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

AI: Artificial Intelligence

ML: Machine Learning

DM: Diabetes Mellitus

MLTs: Machine Learning Techniques

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

WP: Work Package

CONSTANCE: A population-based epidemiological cohort

SAQ: Self-administered Auto Questionnaire

LDA: Linear Discriminant Analysis

LR: Logistic Regression

FDA: Flexible Discriminant Analysis

C5: Decision Tree

ROC: Receiver Operating Characteristics

Pos Pred Value: Positive Predictive Value

Neg Pred Value: Negative Predictive Value

95%CI: 95% Confidence Interval

BMI: Body Mass Index

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.

Authors' contributions

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

VIII. 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).

IX. Acknowledgements

We specially acknowledge Le Marie Zins (Responsible for Constance cohort) for her kind support to access and use the data from this cohort. We acknowledge Romana Haneef, Anne Gallay, Sonsoles Fuentes, Sandrine Fosse-Edorh (Department of Non-Communicable Diseases and Injuries, Santé Publique France, Saint-Maurice); Sofian Kab (Population-Based Epidemiological Cohorts Unit, INSERM UMS 011, Villejuif, France); Emmanuel Cosson (Department of Endocrinology-Diabetology-Nutrition, AP-HP, Avicenne Hospital, Paris 13, University, Sorbonne Paris Cité, CRNH-IdF, CINFO, Bobigny, France, Sorbonne Paris Cité, UMR U1153 Inserm/U1125 Inra/Cnam/Université Paris 13, Bobigny, France) and Rok Hrzic (Department of International Health, Care and Public Health Research Institute - CAPHRI, University[2] of Maastricht University, Maastricht, The Netherlands), for their great support to perform this study.

X. 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.

Sciensano | Rue Juliette Wytsmanstraat 14 |
1050 Brussels | Belgium| e-mail: infact.coordination@sciensano.be|
Website: www.inf-act.eu |Twitter: @JA_InfAct