Development and validation of diabetes case definition algorithms from health administrative databases using data from the Constances cohort and its application to the study of the evolution of diabetes prevalence and incidence based on the SNDS

Sonsoles Fuentes Gutierrez

Development and validation of diabetes case definition algorithms from health administrative databases using data from the Constances cohort and its application to the study of the evolution of diabetes prevalence and incidence in France based on the SNDS

UNIVERSITY PARIS-EST - DOCTORAL THESIS

Doctoral School Nº570 Public Health (EDSP)

Doctoral specialization: Public health - Epidemiology

Thesis presented and defended in Saint Maurice, the 17th October 2020, by

SONSOLES FUENTES

Composition of Jury :

Pierre Fontaine (Université de Lille)	President
Annick Fontbonne (Université Paris-Saclay/Université Versailles Saint-Quentin)	Reporter
Alfred Penfornis (Université Paris-Saclay)	Reporter
Fabrice Bonnet (Université Rennes1)	Examiner
Bruno Detournay (Université Paris VI - Pierre et Marie Curie)	Examiner
Guy Fagherazzi (Université Paris-Saclay, Université Paris-Sud)	Examiner
Sébastien Czernichow (Université Paris Descartes)	Examiner
Emmanuel Cosson (Université Paris 13)	Thesis Director

Invited guests:

Sandrine Fosse-Edorh

UNIVERSITÉ PARIS-EST

ABSTRACT

Context: The National health data system (*Système National de Données Santé*, SNDS) is a health-administrative database (HAD) comprising information on reimbursements of dispensed out-of-hospital health care, on public and private hospital stays and on deaths for the whole population living in France. It is one of the main data sources of the diabetes epidemiological surveillance system in France. As Big Data source of information, it offers a huge potential in terms of epidemiological surveillance and it can only be exploited through specific tools which faced in 2016 several methodological challenges.

Objectives: The objectives of this thesis were to use the data from the CONSTANCES cohort to improve the classical tools for diabetes surveillance in the SNDS and to develop new tools through Machine Learnings methods.

Results: First, the validation of the diabetes case definition algorithms using the CONSTANCES cohort showed they had excellent performances in identifying diagnosed cases and pharmacologically treated cases. After retaining the most suitable algorithm relative to our purpose, it was applied to the entire SNDS to study the evolution of the diabetes epidemic in France. Between 2010 and 2017, prevalence rates slightly increased while incidence rates decreased over the period 2012-2017, among adults aged 45 years or older.
Machine Learning methods were applied to the data from the CONSTANCES cohort to develop a high performant type1/type 2 classification algorithm. A linear discriminant model based on the number of reimbursements over the last 12 months of fast-acting insulin, long-acting insulin and biguanides was retained. Another two algorithms for identifying undiagnosed diabetes cases and prediabetes cases were developed with the same methodology. Both algorithms were logistic regression models. The undiagnosed diabetes algorithm was based on 5 variables (age, sex and number of reimbursement in the last 12 months of tests for lipid profile, screening tests for glucose and general practitioner consultations) and the prediabetes algorithm on 6 variables (age, sex and number of reimbursements in the last 12 months of specific antigen screening tests, HbA1c screening tests, tests for lipid profile and screening tests for glucose).

Conclusion: HADs such as the SNDS represent an opportunity for diabetes surveillance which is a key element for the development of prevention programs and public health policies.

Keywords: diabetes, epidemiological surveillance, big-data, public health, algorithm and prevention

RÉSUMÉ

Contexte: Le Système National de Données Santé (SNDS) est une base de données médico-administratives (BDMA) comprenant des informations sur les remboursements de soins en ville, sur les hospitalisations en secteur public et privé et sur les décès de l'ensemble de la population résidant en France. Il s'agit d'une des sources de données majeures du dispositif de surveillance épidémiologique du diabète en France. Cette source d'informations de type *Big data* offre un vaste potentiel en termes de surveillance épidémiologique qui ne peut être développé qu'après avoir levé les défis méthodologiques associés au recours à ces outils.

Objectifs: Les objectifs de cette thèse sont d'utiliser les données de la cohorte Constances pour améliorer le système de surveillance du diabète basé sur le SNDS et de développer de nouveaux outils en appliquant la méthodologie *Machine Learning.*

Résultats: Dans un premier temps, l'étude de validation des algorithmes d'identification des cas de diabète, à partir de la cohorte Constances, a montré qu'ils avaient d'excellentes performances que ce soit pour le diabète connu ou traité pharmacologiquement. L'algorithme basé sur les remboursements de traitements antidiabétiques a été retenu pour l'étude de l'évolution de l'épidémie du diabète en France dans le SNDS. Entre 2010 et 2017, une légère augmentation de la prévalence et une diminution de l'incidence sur la période 2012-2017, ont été observées chez les adultes âgées 45 ans ou plus. Ensuite, une méthodologie de type *Machine Learning* a été appliquée aux données de la cohorte Constances afin de développer un algorithme de typage du diabète. Un modèle d'analyse discriminante linéaire a été retenu, basé sur le nombre de remboursements d'insuline à action rapide, d'insuline de longue durée et de biguanides au cours des 12 mois. En utilisant la même méthodologie, deux autres algorithmes ont été développés pour identifier les cas de diabète non diagnostiqué et les cas du prédiabète. Ces deux algorithmes étaient basé sur des modèles de régression logistique. Le premier algorithme retenait 5 variables (âge, sexe et nombre de remboursements sur 12 mois de bilans lipidiques, de dosages de glycémie et de consultations d'un médecin généraliste) et le deuxième retenait 6 variables (âge, sexe et nombre de remboursements de dosages d'antigène prostatique spécifique, de glycémie et d'HbA1C et de bilans lipidiques).

Conclusion: Les BDMAs, telles que le SNDS, représentent une opportunité pour la surveillance épidémiologique du diabète, élément central pour le déploiement des programmes de prévention et des politiques de santé publique.

Mots clés: diabète, surveillance épidémiologique, Big-data, santé publique, algorithme et prévention.

ACKNOWLEDGEMENTS

This manuscript is the culmination of three years of work and I would like to acknowledge the following people and institutions involve in this project.

First, I would like to thank to Santé Publique France for funding the phD scholarship program that has enabled me to complete this thesis as well as for its financial support to attend to courses and congress in France and abroad which have enhanced my academic experience.

I want to express my gratitude to my thesis supervisors Ms. Sandrine Fosse-Edorh and Prof. Emmanuel Cosson for their continuous guidance and encouragement. They have shared with me their incredible knowledge and enthusiasm, especially in our priceless brain-storming sessions. Also, they have wisely pushed me out of my comfort zone to achieve always the highest standards of research. I really feel very lucky because I had the opportunity of working with you.

My truthful thanks to the *Département maladies non transmissibles et traumatismes* (DNMTT) directed by Dr. Anne Gallay and Dr. Emmanuelle Bauchet for their warmly welcoming and their professional and personal support. I want specially thanks to my colleague Dr. Clara Piffaretti for all the good times we have shared together in our nice office and in our academical trips to Cambridge and Luxembourg.

This work would not be possible without the fantastic team of the CONSTANCES cohort, leaded by Prof. Marie Zins and Prof. Marcel Goldberg. They have not only provided me with high quality data and technical support, they have also integrated me as a member of their team. Special thanks to Dr. Sofiane Kab who had patiently solved all my questions and helped me to deal with the SNDS data.

I am also very thankful to the *Direction Appui, Traitements et Analyses des données* for their relevant support on the methodology of this thesis, particularly to Ms. Laurence Mandereau-Bruno. She has become my SNDS "guru" and I deeply admire her thoroughness, her patience and her positive attitude. Also thanks to Dr. Rok Hrzic from Maastricht University for his generosity helping me with the Machine Learning methods.

My sincerely acknowledgement goes to Prof. Dr Martine Bellanger, director of the Masters of Public Health Program at EHESP, for her strong commitment to all her students and for trusting in me when she recommended me for this phD position.

I would like to state my recognition to the *Ecole Doctorale en Santé Publique* (EDSP), specially to its former director Prof. Jean Bouyer, its current director Prof.

Florence Menegaux as well as Ms. Audrey Bourgeois and Ms. Fabienne Renoirt for their dedicated engagement with all doctoral students.

I want to thank the Endocrinology ReDSiam Working Group where the algorithms were discussed for its methodological support.

My appreciation also goes to the components of my thesis committee: the rapporteurs Dr. Annick Fontbonne and Prof. Pierre Fontaine, the examiners Prof. Alfred Penfornis, Prof. Fabrice Bonnet, Prof. Bruno Detournay, Dr. Guy Fagherazzi and Prof. Sébastien Czernichow and the president of my thesis committee Prof. Pierre Fontaine. Thanks to all of them for accepting to participate in this committee.

This thesis was a long and winding road and I would have been lost without my family and my friends. I wish to express my gratitude to my aunt Mª Angeles, my sister-in-law María, my nieces Martina and Carlota, my nephew Félix, Pupi, my aunt Mª Luisa and my uncle Ignacio for their love and their unconditional support through all these years. I want also to thank my second family, my friends Beatriz, Paco, Javy, Chus, Manu, Ana, Elena and Zu for listening to all my problems, and helping me not to give up, you are my precious treasure. Also, thanks to Michael, Bérénice and Ornella for their affection and empathy. Thanks to my Parisian friends Jamila and Maria, last year was really difficult and you were the best support. I have to include in these acknowledgements *mon petit* Chicho for cheering me up when I came back home after a hard-working day.

I want to remember those who are no longer with us: my parents Consuelo and Félix, my aunt Antonia, my uncle Antonio, my grandparents and my dear friend Silvia. While walking this road, I have felt in my heart that you have taken care of me from heaven.

Last but not least, thanks to my dear brothers Félix and Fernando because they have been always by my side even in the darkest times. Your wisdom, your integrity, your courage and your love always will be my guiding lights.

SCIENTIFIC PRODUCTION

Articles accepted for publication

- Fuentes S., Mandereau-Bruno L., Fagot-Campagna A., Bernillon P., Goldberg M., Fosse-Edorh S and Cosson E. *Identifying diabetes cases in health administrative databases: a validation study based on a large French cohort*. Int J Public Health. 2019 Apr;64(3):441-450. doi: 10.1007/s00038-018-1186-3.

- Fuentes S, Mandereau-Bruno L., Regnault N., Bernillon P., Bonaldi C., Cosson E. and Fosse-Edorh S. *Is the type 2 diabetes epidemic plateauing out in France? A nationwide population-based study*. Diabetes Metab. 2020. Available online 7 January. doi.org/10.1016/j.diabet.2019.12.006.

Articles submitted for publication

- Fuentes S, Hrzic R, Haneef R, Kab S, Cosson E and Fosse-Edorh S. *Artificial intelligence for diabetes research: development of type 1/type 2 classification algorithm and its application to surveillance using a nationwide population-based medico-administrative database in France.* Submitted to Diabetes Care.

Oral Communications

- Fuentes S, Hrzic R, Haneef R, Kab S, Cosson E and Fosse-Edorh S.. *L'intelligence artificielle au service de la surveillance du diabète : développement d'un algorithme de typage du diabète à partir de la cohorte Constances et application aux données du Système National des Données Santé*. French-Speaking Diabetes Society Conference (Belgium, Brussels 2020).

- Fuentes S, Hrzic R, Haneef R, Kab S, Fosse-Edorh S and Cosson E. *Apports de l'intelligence artificielle dans la prévention du diabète : comment cibler les personnes ayant un diabète méconnu dans le Système National des Données Santé : Etude basée sur les données de la cohorte CONSTANCES*. French-Speaking Diabetes Society Conference (Belgium, Brussels 2020).

- Fuentes S., Mandereau-Bruno L., Bernillon P., Bonaldi C., Fosse-Edorh S and Cosson E.. *Évolution de la prévalence et de l'incidence du diabète en France entre 2010 et 2017*. Santé Publique France Meetings (Paris, France, June 2019).

- Fuentes S, , Mandereau-Bruno L, Goldberg M, Fosse-Edorh S and Cosson E. *Validation d'algorithmes d'identification des cas de diabète dans les bases médico-administratives à partir des données de la cohorte Constances*. Santé Publique France Meetings (Paris, France, May 2018).

- Fuentes S . *Apport de la cohorte Constances dans l'épidémiologie du diabète en France. 5TH CONSTANCES and Gazel cohort scientific sessions* (Paris, France, May 2018).

- Fuentes S, Mandereau-Bruno L , Regnault N , Santin G, Fosse-Edorh S and Cosson E *Prévalence du pré-diabète, du diabète non-diagnostiqué et du diabète diagnostiqué chez les personnes âgées de 18 à 70 ans en France en 2013 à partir de la cohorte CONSTANCES*. French-Speaking Diabetes Society Conference (Nantes, France, May 2018).

Poster presentations

- Fuentes S, Mandereau-Bruno L, Bernillon P, Bonaldi C, Fosse-Edorh S and Cosson E. *Trends on prevalence and incidence of type 2 diabetes in France between 2010 and 2017: a nationwide population-based study*. 59th Annual Meeting of the European Association for the Study of Diabetes (EASD) (Barcelone, Spain, September 2019).

- Fuentes S, Bernillon P, Bonaldi C, Fosse-Edorh S and Cosson E. Trends on prevalence and incidence in France: a nationwide study. 79th Scientific Sessions of the American Diabetes Association (ADA) (San Francisco, the US, June 2019)

- Fuentes S, Mandereau-Bruno L, Bernillon P, Bonaldi C, Fosse-Edorh S and Cosson E. *Trends on prevalence and incidence in France: a nationwide study.* 54th Annual Meeting of the European Diabetes Epidemiology Group (EDEG). (Luxembourg, Luxembourg, May 2019).

- Fuentes S, Fosse-Edorh S, Regnault N, Goldberg M, Fosse-Edorh S and Cosson E. *Prevalence of pre-diabetes, undiagnosed and diagnosed diabetes among adults aged 18 to 70 years in France: the CONSTANCES cohort*. 58th Annual Meeting of the European Association for the Study of Diabetes (Berlin, Germany, October 2018)

−Fuentes S, Regnault N, Goldberg M, Fosse-Edorh S and Cosson E. *Prevalence of pre-diabetes and undiagnosed diabetes among adults aged 18 to 70 years in France: the CONSTANCES cohort*. 78th Scientific Sessions of the American Diabetes Association (ADA) (Orlando, the US, June 2018).

- Fuentes S, Mandereau-Bruno L, Goldberg M., Fosse-Edorh S and Cosson E. *Diabetes case definitions in health administrative databases: a validation study in France based on the CONSTANCES cohort*. 53th Annual Meeting of the European Diabetes Epidemiology Group (EDEG) (Elsinore, Denmark, April 2018).

Others

- Jury prize “My thesis in 180 seconds” .French-Speaking Diabetes Society Conference (Marseille, France, Mars 2019)
- “MT180” Université Paris Est
https://www.youtube.com/watch?v=QziqFicd_R8&feature=emb_logo

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ADA: American diabetes association

ALD: *Affection de longue durée*

ARS: *Agence regional de santé*

ATC: The Anatomical Therapeutic Chemical Classification System

ATIH: *Agence technique d'information sur l'hospitalisation*

BDMA: *Base de données médico-administratives*

BMI: Body Mass Index

BRFSS: Behavioral Risk Factor Surveillance System

CCAM: *classification commune des actes médicaux*

CDC: Center for Disease Control Prevention

CDK: chronic kidney disease

CépiDc: *Centre d'épidémiologie sur les causes médicales de décès*

CGM: Continuous interstitial glucose monitoring

CMUc: *Couverture maladie universelle complémentaire*

CNAMTS: *Caisse National d'Assurance Maladie des Travailleurs Salariés*

CNAV: *Caisse Nationale d'Assurance Vieillesse*

CTLA-4: Cytotoxic T lymphocyte associated-4

DCIR: *Données de consommation inter-régimes*

DKA: diabetic ketoacidosis

DRG: diagnoses-related group

EMR: Electronic medical records

ENNS: *Etude national nutrition santé*

Entred: *Echantillon National Témoin Représentatif des personnes diabétiques*

Esteban: *Etude de santé sur l'environnement, la biosurveillance, l'activité physique et la nutrition*

EUR: euro

FDEP: French area deprivation index

FOT: French Overseas Territories are the French

FPG: Fasting plasma glucose

GAJ: *Glycémie à jeun*

HAD: Health-administrative database

HAS: *Haute Autorité Santé*

HbA1C: glycated haemoglobin

HHS: hyperosmolar hyperglycemic syndrome

HLA: Human Leukocyte Antigen

HSC: Health screening center

ICD: International Statistical Classification of Diseases and Related Health Problems

IDF: International Diabetes Federation

IFG: Impaired fasting glucose

IGT: Impaired glucose tolerance

INSEE: *Institut national de la statistique et des études économiques*

kg: kilograms

LADA: latent autoimmune diabetes in adults

m: meters

MENA: Middle East-North African region

mg: miligrams

mmol: milimol

MODY: Maturity-onset diabetes of the young

MSA: *Mutualité Social Agricole*

NCDR: Norwegian Childhood Diabetes Registry

NHANES: The National Health and Nutrition Examination Survey

NHIS: National Health Interview Survey in the US

NIR: *numéro d'identification au répertoire*

NPV: negative predictive value

OGTT: Oral glucose tolerance test

PMSI: *Programme de médicalisation des systèmes d'information*

PPV: positive predictive value

py: person years

RG: *Régime Général*

RSI: *Régime social des travailleurs indépendants*

SLM: *Section locales mutualiste*

SMBG: Patient self-monitoring of blood glucose

SML: Supervised Machine Learning

SNDS: *Système National de Données Santé*

SNIIRAM: *Système d'information inter-régime de l'assurance de maladie*

US: United States

USD: United States dollar

WHO: World Health Organization

INTRODUCTION

We are living in the Digital Era, where the development of information technologies has allowed us to collect, store, transmit and manipulate huge amounts of data. This is what we called Big Data, characterized by the three V: Volume, Velocity and Variety. Thanks to Big Data, many fields of knowledge like economics or genetics advance quickly and achieve incredible objectives. Big Data can also make epidemiology evolve by using exhaustive and updated data not from a sample of population, as it has been done before, but from the entire population. Can you imagine how easily had associated John Snow the cholera outbreak with the public water pumps in London if he would have used Big Data sources?

I became fascinated by epidemiology while doing my master of public health in the *Ecole des Hautes Etudes en Santé Publique*. The second year of my master, I decided to specialized in biostatistics in order to learn the latest methodologies for epidemiological research like factor analysis or cluster analysis which I would latter apply in my master thesis for studying the relationship between psychosocial factors and obesity. In 2016, I granted a scholarship from *Santé Publique France* to do a phD in epidemiology on diabetes surveillance. It was stimulating doing my phD in *Santé Publique France*, since the results of the studies conducted there are widely use by the French Government for developing and evaluating policies on public health. But what I found really exciting was the chance of working with a Big Data source comprising information from the 66 million people living in France, the French national health insurance information system (*Système National de Données Santé*, SNDS). The SNDS is a unique data source but the tools applied for diabetes surveillance presented certain limitations. The objective of this thesis was to solve these limitations for making the SNDS an optimal data source for diabetes epidemiology.

1. Diabetes disease

Diabetes is a complex disorder related to disruptions of insulin metabolism. In this section, we aim to explain the pathogenesis, the natural history, the diagnosis, the related complications and the treatment of diabetes in order to better understand the surveillance of this important non-communicable disease. We summarize the content of this section in **Figure 1**, where the different stages of diabetes disease are represented.

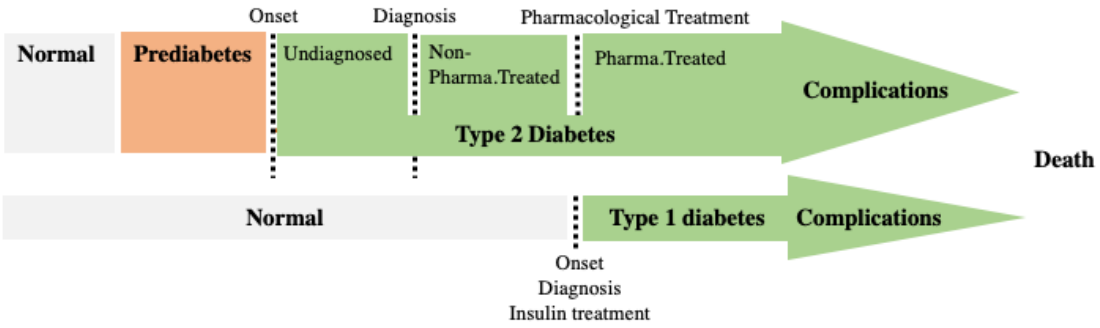


Figure 1. Stages of diabetes disease

1.1 History of diabetes

The first reference to diabetes appears in an Egyptian papyrus from c.1500 BCE[1], where it is described as a condition of "passing too much urine"[1] . The name of diabetes mellitus – after the Latin word *mel* meaning honey - was given to the sweetness of diabetic urine by the 17th century British physician, Thomas Willis. He also noticed that, although the disease was rare in ancient times, its frequency was increasing. The cause of diabetes remained unknown until the late 19th century when Joseph von Mering and Édouard Hedon found the association between diabetes and pancreas [2]. Later, in 1893, the French pathologist Gustave-Édouard Languesse finally discovered that the regulation of blood sugar levels depends on a hormone secreted by the pancreas, the Insulin.

1.2 Diabetes definition and pathogenesis

In 1999, the World Health Organization (WHO) defined the term diabetes mellitus as "a metabolic disorder of multiple aetiology characterized by chronic hyperglycemia with disturbances of carbohydrate, fat and protein metabolism resulting from defects in insulin secretion, insulin action, or both" [3]. Insulin is a hormone produced in the β cells of the islets of Langerhans from the pancreas which is responsible – together with another hormone produced by α pancreatic cells, glucagon - of the regulation of blood glucose levels [4]. High levels of glucose in blood induce the release of insulin which activates through specific receptors in muscles, adipose tissue and liver the absorption of carbohydrates, especially glucose (**Figure 2**).

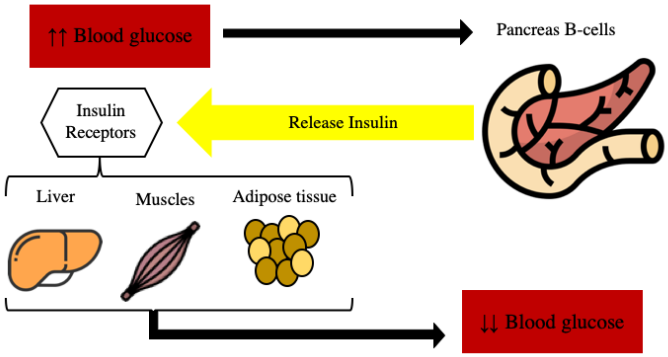


Figure 2.Regulation of blood glucose through insulin metabolism

Two main diabetes pathogenic pathways drive to chronic hyperglycemia: (i) the destruction of the β cells and no insulin production and (ii) deficient insulin action due to impairment of insulin secretion and/or defects in insulin action. The continued high levels of glucose in blood damages different organs such as eyes, kidneys, nerves, heart and

[1] BCE: Before Christian Era

blood vessels.

1.3 Classification of diabetes mellitus

The current WHO classification is represented in **Table 1.** Diabetes has been classified into five clinical categories [5]:

a) Type 1 diabetes is due to β-cell destruction, commonly leading to absolute insulin deficiency. Usually, the destruction of β-cells is an immune-mediated process (identified as Type 1A) but a small group of cases present an idiopathic form of the disease (identified as Type 1B). The clinical classical characteristics of type 1 cases are : acute onset at young age – before 35 years-, normal BMI[2], use of insulin within 12 months of diagnosis, and high risk of diabetic ketoacidosis [6]. This form of diabetes accounts for 5 to 10% of diabetes cases.

b) Type 2 diabetes is due to a β-cell dysfunction resulting into a gradual loss of insulin secretion on the background of insulin resistance. Type 2 diabetes differs much from Type 1 in terms of clinical characteristics since the onset of disease is slow and usually at older ages, most of the cases are overweight or obese, they are less likely to require insulin treatment in the 12 months after diagnosis and commonly they do not present ketoacidosis [6]. It accounts for between 90% and 95% of diabetes cases [5].

c) Gestational diabetes mellitus (GDM) is a type of diabetes diagnosed during pregnancy –usually in the second or third trimester- in women not previously diagnosed with diabetes. Normally, it does not persist after delivery but some cases of type 2 diabetes are discovered during pregnancy [5]. Overweight, older age, family history of diabetes or personal history of GDM are the most common risk factors. Lifestyle interventions and if necessary insulin injections protect from adverse pregnancy outcomes, such as macrosomic infant and preeclampsia [7].

d) Specific types of diabetes due to other diseases not included in the three previous categories are included in this category, such as diseases of the exocrine pancreas (pancreatitis, cystic fibrosis haemochromatosis), endocrine disorders (Cushing's syndrome, acromegaly, hyperthyroidism), drug- or chemical-induced diabetes (due to glucocorticoid use or pentamidine for example), infections (such as congenital rubella), monogenic defects of β-cell function (maturity-onset diabetes of the young –MODY- or transient neonatal diabetes –TNDM-), monogenic defects in insulin action

[2] BMI: Body Mass Index. It is estimated by dividing the weight by the square of the body height (m2). In adults, four categories are defined: underweight - less than 18.5 kg/m2-, normal weight -18.5 to 25 kg/m2-, overweight -25 to 30 kg/m2-, or obese -over 30 kg/m2-

(Leprechaunism or Rabson-Mendenhall syndrome) and other genetic syndromes associated to diabetes (Down syndrome or Klinefelter's syndrome) [6].

e) Hybrid forms of diabetes is a new category recently added by the WHO comprising clinical forms of diabetes which combine type 1 and type 2 characteristics [5]. The Slowly evolving immune-mediated diabetes (former latent autoimmune diabetes in adults –LADA-) is included, since its clinical characteristics are similar to type 2 diabetes, but the individuals present pancreatic autoantibodies. Another example is the Ketosis-prone type 2 diabetes.

Table 1. Classification of diabetes mellitus

a) Type 1
- Immune mediated
- Idiopathic
b) Type 2
c) Gestational diabetes mellitus
d) Specific types of diabetes due to other causes
- Diseases of the endocrine pancreas
- Endocrine disorders
- Drug- or chemical induced diabetes
- Infections
- Monogenetic defects of β-cell function
- Monogenic defects in insulin action
- Other genetic syndromes associated to diabetes
e) Hybrid forms of diabetes
- Slowly evolving immune-mediated diabetes
- Ketosis-prone type 2 diabetes

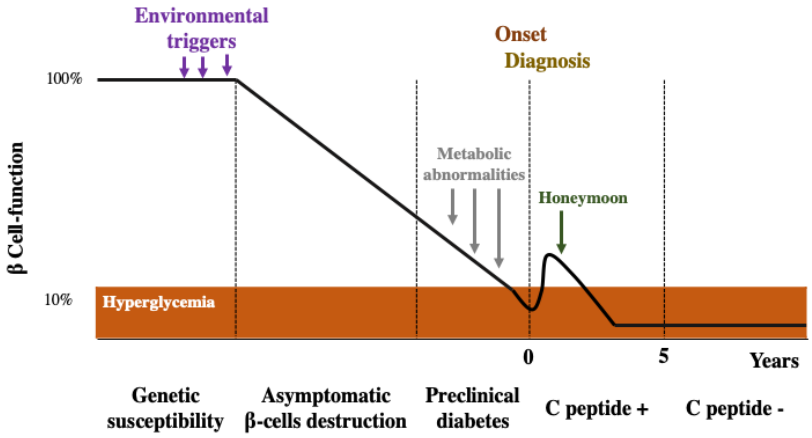
1.4 Natural history and risk factors

We have seen that most of the diabetes cases accounts for type 2 and type 1 diabetes. These forms of diabetes differ not only in terms of clinical characteristics but also on natural history and risk factors.

Individuals with type 1 diabetes present a genetic susceptibility for the disease activated by environmental triggers [8]. Mutations in different genes related to insulin have been associated with a higher risk of developing type 1 diabetes. Among them the Human Leukocyte Antigen (HLA) gene, Insulin gene/IDDM2 locus or the cytotoxic T lymphocyte associated-4 (CTLA-4) gene [9]. Environmental factors trigger the process of β-cells destruction by the immune system. Different triggers have been suggested i.e.

infections of Enteroviruses or Mycobacterium avium, albumin component of cow's milk, wheat proteins or low serum concentrations of vitamin D [9, 10]. The prevention of type 1 diabetes is complicated since the role of these factors in the development of the disease remains unclear.

The natural history of type 1 diabetes is *represented in* **Figure 3**. *Four* phases have been described. First the destruction of β-cells happens without clinical symptoms. Then, a preclinical phase is observed with certain metabolic abnormalities, rapidly followed by a clinical symptomatic onset when around 90% of β-cells are destructed. Finally a complete loss of β-cell function appears, with no C-peptide (a marker of insulin secretion) anymore found in blood [6]. Before this last stage and after the clinical onset, some patients present the so called "Honeymoon phase" which corresponds to a transient stage during which they recover enough production of insulin to keep normal blood glucose values without exogenous insulin injection



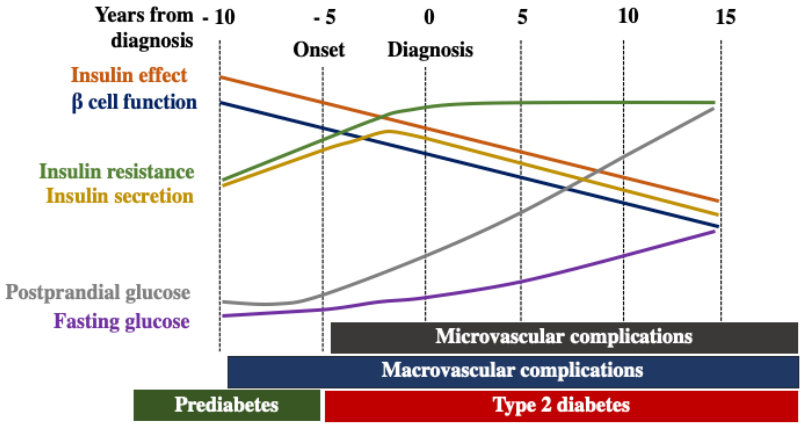
Adapted from Raverot et al 2005 [6]

Figure 3. Natural history of type 1 diabetes

Contrary to type 1 diabetes, type 2 diabetes is a highly preventable disease. A genetic susceptibility is described, since first degree relatives of type 2 diabetes patients are at higher risk of developing the disease [11]. However most of the risk factors are related to health behaviors like lack of physical activity, sedentary lifestyle, sleep duration, smoking, alcohol consumption and poor diet (low-fiber, high glycemic index and high saturated fat) [12]. Many of these risk factors lead to overweight and obesity, which are highly associated with type 2 diabetes, especially abdominal obesity [13].

In the natural history of type 2 diabetes, an increase in insulin resistance is described, first compensated by an increase of insulin secretion (**Figure 4**). However, β-cells function is progressively insufficient (relative insulin deficiency), leading first to an increase in post-prandial glucose levels (impaired glucose tolerance -IGT-) and then in

fasting blood glucose (impaired fasting glucose -IFG-). These conditions are asymptomatic and named prediabetes. Individuals with prediabetes are not only at higher risk of developing diabetes but also other diseases such as cardiovascular diseases (especially those with IGT). The next stage is type 2 diabetes onset, with many patients remaining undiagnosed due to the mild symptoms of the disease. In the following years, the β-cells function continues to decrease from a relative to an absolute insulin deficiency.



Adapted from American Association of Endocrinologists 2019 [14]
Figure 4. Natural History of type 2 diabetes

1.5 Symptoms and Diagnosis

1.5.1 Symptoms

The classical symptoms diabetes symptoms are polyuria, polydipsia, fatigue and weakness. Type 1 diabetic patients also present weight loss despite increased appetite and sometimes blurred vision. Also in type 1 diabetes, symptoms emerge usually in days or weeks so it is unlikely that type 1 cases are diagnosed due to a routine medical screening [6]. On the contrary, the onset of type 2 diabetes is frequently not associated with clinical signs so patients are usually diagnosed during a routine check-up. In addition to classical diabetes symptoms, type 2 cases can also present other conditions such as skin infections or healing problems. It is considered that around one third of patients with type 2 diabetes present chronic diabetes-related complications (**See page 32**) at onset.

1.5.2 Blood test for diabetes diagnosis

Four types of blood test can be used for diabetes and prediabetes diagnosis (**Table 2**) [15]:

Table 2. Criteria for prediabetes and diabetes diagnosis

	Prediabetes				**Diabetes**	
	Impaired glucose tolerance IGT		Impaired fasting glucose IFG			
	mmol/l	mg/dl	mmol/l	mg/dl	mmol/l	mg/dl
Fasting Plasma Glucose (FPG)					≥ 7	≥ 126
[Criteria before 1999]					≥ 7.8	≥ 140
-WHO criteria			6.1 to 6.9	110 to 125		
-ADA criteria			5.6 to 6.9	100 to 125		
Oral Glucose Tolerance Test (OGTT)	7.8 to 11	140 to 199			≥11.1	≥200
Random Plasma glucose with previous symptoms *					≥11.1	≥200
	%				%	
Glycated Hemoglobin (HbA1c)	5.7 to 6.4				6.5	

*Symptoms of hyperglycemia or hyperglycemia crisis

a) Fasting plasma glucose (FPG) test: the test measures the levels of venous plasma glucose after 8 hours of fasting. Following the French guidelines, diabetes is diagnosed when there are two consecutive test measures with levels of FPG equal or higher than 7 mmol/l (≥126 mg/dl) [16]. These diagnosis criteria were implemented by the WHO in 1999 because the previous FPG threshold for diabetes diagnosis (7.8 mmol/l (140 mg/dl)), was considered too high [17]. To note, IFG is defined as FPG between 6.1 and 6.9 mmol/l (110 to 125 mg/dl) by the WHO or FPG between 5.6 and 6.9 mmol/l (100 to 125 mg/dl) by the American diabetes association (ADA) [18]

b) Plasma glucose value at any time: diabetes is diagnosed when the levels of plasma glucose (no previous fasting required) are equal or higher than 11.1 mmol/l (≥ 200 mg/dl) and individuals present diabetes-related symptoms [16].

c) Oral glucose tolerance test (OGTT): first FPG level is measured; then the individual drinks a syrup solution containing 75 grams of glucose and finally the level plasma glucose is measured 2 hours after syrup intakes [15]. If the level of 2-hours plasma glucose is equal or higher than 11.1 mmol/l (200 mg/dl), then diabetes is diagnosed. To note, IGT is defined when the levels of venous plasma glucose are between 7.8 and 11 mmol/l (140 – 199 mg/dl) [18].

d) Glycated hemoglobin (HbA1c) test: The test measures the percentage of glucose attached to hemoglobin. It does not require previous fasting and whereas this test is used for evaluation of diabetes management after diagnosis, it is also used in some countries like in the United States (US) for primary diagnosis of diabetes and prediabetes [19]. On one hand, the test is easier to perform, it has less pre-analysis instability and, contrary to glycemia, the levels of HbA1c are less affected by the stress [20]. On other hand, the proportion of HbA1c is only a proxy of blood glucose levels and non-glycemic factors can impact the results such as anemia (a common disorder in women) or alcohol consumption [21]. The criteria for diabetes and prediabetes diagnosis are percentage of HbA1c greater or equal to 6.5 and between 5.7 and 6.4, respectively [18]. It is not recommended to use HbA1c as diagnosis criteria in France, mainly because of its high cost **.**

1.6 Diabetes complications

The main diabetes complications are represented in **Table 3**. They can be classified into two groups: acute complications and chronic complications.

Table 3. Diabetes complications

1. Acute complications
Diabetic ketoacidosis
Hyperosmolar hyperglycemic syndrome
2. Chromic complications
2.1 *Microvascular*
Retinopathy
Neuropathy
Peripheral neuropathy
Autonomic neuropathy
Nephropathy
Diabetic foot
2.2 *Macrovascular*
Coronary artery diseases
Cerebrovascular diseases
2.3 *Others*
Infections
Pregnancy disorders

1.6.1 Acute complications

One of the most common acute complications of type 1 diabetes is diabetic ketoacidosis (DKA) [22]. When the levels of insulin are very low, glucose does not get into cells and the body starts to produce metabolites called ketones. If DKA is not treated,

it could produce multiorgan failure and death. Another acute complication is hyperosmolar hyperglycemic syndrome (HHS) [23]. When polyuria, due to high levels of glucose in blood, is not compensated by polydipsia, serum osmolarity increases, damaging different organs including the brain which could lead to coma. The HHS is more common in type 2 than in type 1 diabetes [6].

1.6.2 Chronic complications

The chronic diabetic complications are usually categorized in microvascular and macrovascular complications. Among the microvascular complications, we can cite retinopathy, nephropathy, neuropathy (peripheral and autonomic) and diabetic foot [24].

Long-term hyperglycemia damages retinal vessels causing retinopathy. This can lead to mild vision problems and finally blindness; diabetic retinopathy is the most frequent cause of new cases of blindness in adults [25]. Some type 2 diabetic patients already have this pathology when they are diagnosed [6].

Unlike retinopathy, neuropathy is usually diagnosed long time after diabetes diagnosis [6]. Depending on the type of nerve system affected, we can differentiate peripheral neuropathy (peripheral nervous system) and autonomic neuropathy (autonomic nervous system) [26]. The most common peripheral nervous system damage leads to abnormal limb sensitivity. In autonomic neuropathy, the vagus nerve and other nerves from the sympathetic system are damaged. The symptoms of autonomic neuropathy usually remain invisible and they include sexual dysfunction, constipation or resting tachycardia [27].

Nephropathy is one of the leading causes of mortality among diabetic patients [28]. The glomerulus responsible for renal function is harmed by hyperglycemia. When a high percentage of glomeruli are affected, the individual develops chronic kidney disease (CDK) and renal failure [29]. Diabetic nephropathy can be accelerated by other conditions such as hypertension [27].

Diabetic foot is another microvascular complication of diabetes [24]. A combination of peripheral neuropathy, peripheral arterial disease and skin disorders like hyperkeratosis participate in the development of foot ulcers [30]. If foot ulcers are not treated properly, they can lead to amputation of lower limbs [31].

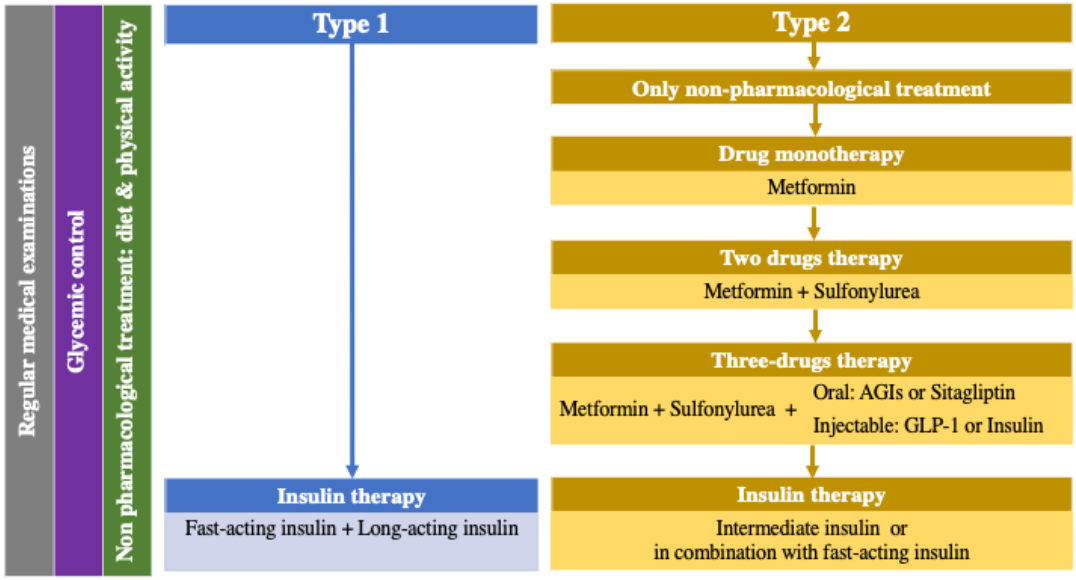
Individuals with diabetes and also IGT are at higher risk of developing cardiovascular diseases: coronary artery diseases like myocardial infarction or heart failure, and cerebrovascular diseases like stroke [18, 32]. Most of them are life-threatening disorders and they represent the most important cause of morbidity and

mortality among diabetic patients [33].

Finally, there have been described other diabetes complications such as infections or pregnancy disorders [34-36].

1.7 Diabetes management

Diabetes management should be based in four pillars: non pharmacological treatment, pharmacological treatment (if required), glycemic control and regular medical examinations (**Figure 5**).



Adapted from HAS guidelines 2007 and 2013 [37, 38]

Figure 5. Type 1 and type 2 diabetes management

1.7.1 Non pharmacological treatment

Healthy lifestyle plays an essential role in the prevention of diabetes and diabetes-related complications [6]. Diabetic patients are recommended to have a balanced and varied diet, including a source of carbohydrates in each meal (taking into account food's glycemic index) and avoiding excessive fat intake and alcohol consumption [39]. They also must have easy access to food for controlling possible hypoglycemia crises (due to hypoglycemic treatment).

Physical activity in diabetic patients is fundamental not only to control blood glucose levels but also for reducing cardiovascular risk factors, for weight control and for improving well-being and mental health [27]. It has to be adapted to the age and disabilities of the patient. In general, the WHO recommends 150 minutes/week of moderate-intensity or 75 minutes/week of vigorous physical activity [40].

Smoking is the main risk factor for cardiovascular diseases, premature death and microvascular complications though diabetic patients should not use tobacco products [39].

1.7.2 Pharmacological treatment

1.7.2.1 Pharmacological treatment of type 1 diabetes

People with type 1 diabetes require insulin treatment for survival [37]. There are different kinds of insulin based on the onset, the peak time and the duration of the effect [41]:

a) Fast-acting insulin (ATC[3]: A10AB): the action begins about 15 minutes after injection, with a peak time at 1 hour and effects during 2 to 4 hours
b) Intermediate-acting insulin (ATC: A10AC): its onset starts between 2 to 4 hours after injection, the highest action is reached in 4 to 12 hours, and is effective for 12 to 18 hours
c) Long-acting insulin (ATC:A10AE): it reaches the bloodstream within several hours after injection with action for 24-hours or more, without peak
d) Combinations of intermediate- or long-acting with fast-acting (ATC: A10AD)

Insulins are administered through subcutaneous injection. There is also a code (ATCA10AF) for inhaled insulin through nasal route. However, these insulins are not used nowadays.

Pharmacological treatment of type 1 diabetes is based only on insulin therapy [37]. The insulin therapy guidelines for type 1 diabetes recommend the combination of injections of fast-acting insulin before each meal with an additional injection of long-acting insulin to control glycemia between meals [42]. Insulin might be injected or administrated with continuous subcutaneous injection with pump [43].

1.7.2.2 Pharmacological treatment of type 2 diabetes

Type 2 diabetes can be treated only through diet and lifestyle changes. When the glycemic goals are not achieved, it should start a pharmacological treatment with metformin (ATC: A10BA02), an oral lowering drug from the group of biguanides [27]. The French guidelines from the *Haute Autorité Santé* (HAS) defines this stage as "Monotherapy" [38]. According to these guidelines, practitioners should add another drug ("Bi-therapy") if the patient does not reach the glycemic goals; at this stage a combination of metformin with sulfonylurea (ATC: A10BB) is advised. When glycemic levels continue to worsen, in addition to metformin and sulfonylureas, a third type drug

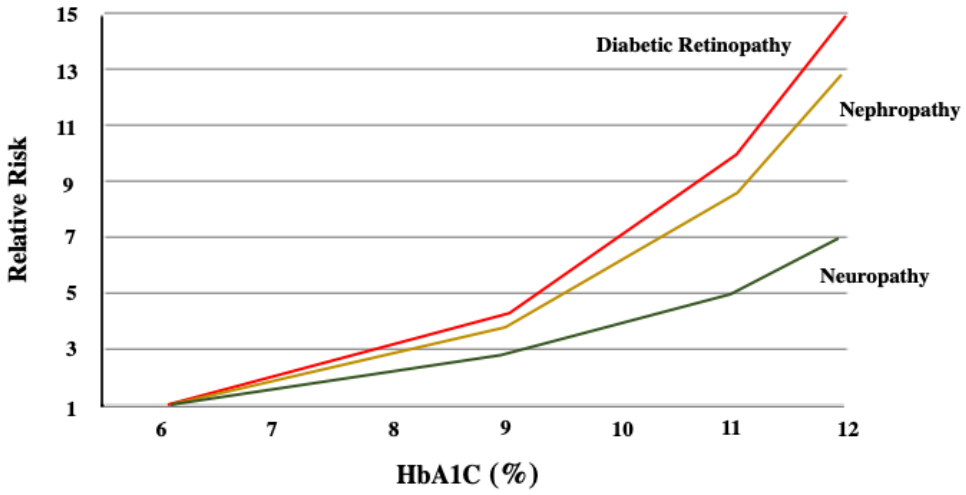
[3] ATC: The Anatomical Therapeutic Chemical Classification System is an international drug classification system with five different levels. The first level is the anatomical/pharmacological group (i.e. A- Alimentary tract and metabolism), the second level is the pharmacological/therapeutic group (A10-Drugs used in diabetes), the third and the fourth levels are the chemical /pharmacological /therapeutic subgroups (A10A- Insulins and analogues; A10AB – Insulins fast-acting) and the fifth level is the chemical substance (A10AB04- insulin lispro-).

is delivered, an oral glucose lowering drug (alpha glucosidase inhibitors - ATC: A10BF- or sitagliptin - ATC: A10BH01-) or an injectable drug (insulin or Glucagon-like peptide-1 (GLP-1) analogues - ATC: A10BJ-). In the last stage, patients are treated with a combination of intermediate or long-acting insulin and fast-acting insulins.

1.7.3 Glycemic control

The glycemic control is usually evaluated through different ways: HbA1c laboratory test and patient self-monitoring of blood glucose (SMBG) or of continuous interstitial glucose monitoring (CGM) [27].

As previously seen, the HbA1c test measures the amount of glucose attached to hemoglobin [21]. It is a good measure of average levels of blood glucose over the last two to three months and it has been shown as an excellent predictor of diabetes complications (**Figure 6**) [44].



Adapted from Sklyer et al. 1996 [44]

Figure 6. HbA1C levels and relative risk of developing diabetes complications

The test must be performed at least every three months but the frequency can be increased if glycemic goals are not achieved or if the patient presents any special conditions [38]. Generally, the glycemic goals for non-pregnant adults should be HbA1c lower than 7%; in elderly patients or patients with advanced complications, a less rigorous HbA1c goal may be required (HbA1c < 8%).

People with type 1 diabetes and type 2 diabetes treated by insulin particularly need to self-monitor their blood glucose levels to prevent hypo or hyperglycemia and to adapt their treatment (diet, physical activity and medication) to their glycemic levels during the day [27]. The measurements are done in capillary blood extracted using a lancet device to stick the finger and disposed in a test strip in order to be read by a glycemic reader or glucose-meter [45]. The CGM devices are a type of self-monitoring systems which do

not require the procedure previously described since the levels of blood glucose are continuously measured by a subcutaneous sensor inserted in the upper gluteal area [46]. The glycemic goals under CGM are in France are between 70 and 120 mg/dl before meals and less than 160 mg/ dl after meals for type 1 and less than 180 mg/dl for type 2 diabetes [45].

1.7.4 Regular medical examinations

To prevent and to manage diabetes complications, are advised to follow different regular medical examinations [27]. The guidelines of HAS regarding these examinations are represented in **Table 4** [47].

Table 4. The HAS Guidelines for regular medical examinations of diabetic patients

Medical examinations	Frequency
GP/ Diabetologist consultation (check-up, BMI, blood pressure)	3 times/year
HbA1c test	3 times/year
Microalbumin test in urine	1 time/ year
Blood creatinine test	1 time/year
Blood lipid profile	1 time/ year
Electrocardiogram (or cardiologist consultation)	1 time/year
Eye examination by an ophthalmologist or optometrist	1 time/ 2 years
Foot examination by a healthcare professional	1 time/year
Dental examination by a dentist	1 time/year

The patient must visit her/his general practitioner or diabetologist in order to have a complete medical evaluation including measurement of BMI and blood pressure. As we have seen before, HbA1c test must be performed at least three times a year to assess glycemic control. Other laboratory tests are required at least once a year: blood creatinine level and microalbuminuria (to assess renal function), and lipid profile in blood. Diabetic patients also need to perform an electrocardiogram or to visit the cardiologist annually. They also must have annually a foot and a dental examination by a healthcare professional. Finally, they must visit the ophthalmologist or optometrist once every two years, for eye examination (specially examination of the fundus of the eye to diagnose retinopathies).

1.8 Conclusion

Diabetes is a complex metabolic disease, with heterogeneous natural history depending on the type of diabetes. The two main types of diabetes are the following:

a) Type 1 diabetes is related to destruction of pancreatic ß-cells through a combination of genetic susceptibility and environmental triggers, with an acute symptomatic

onset at young age because of insulin deficiency. It accounts for 5-10% of diabetes cases.

b) Type 2 diabetes is associated with different risk factors – mainly lifestyle factors- that lead to insulin resistance and relative impairment of β-cells function. It is more common in elderly population and many cases remain undiagnosed. It represents 90 to 95% of diabetes cases.

Diabetes can be diagnosed through different ways: measures of FPG or glucose level 2 hours after OGTT or random glucose value when symptoms are present or in some countries HbA1c levels. Diabetes is responsible for various acute and chronic (microvascular and macrovascular) complications and the patients requires an holistic management to prevent these complications including: non-pharmacological treatment (diet and physical activity), pharmacological treatment (oral glucose lowering drugs and/or GLP1 agonists or insulin), glycemic monitoring (HbA1c, SMBG, CGM) and regular medical examinations.

2. Diabetes epidemiology

The WHO defines epidemiology as “the study of the distribution and determinants of health-related states or events (including diseases), and the application of this study to the control of diseases and other health problems”. After having described diabetes from an individual perspective in the previous section, we want to move to a population perspective, in order to understand the global burden of diabetes. More specifically, we are going to study the descriptive epidemiology of diabetes by presenting its main indicators: the incidence and the prevalence of type 1 and type 2 diabetes ,the prevalence of the different diabetes stages, the morbidity and mortality due to complications and the total cost of diabetes , first in the World and then in France (**Figure 7**).

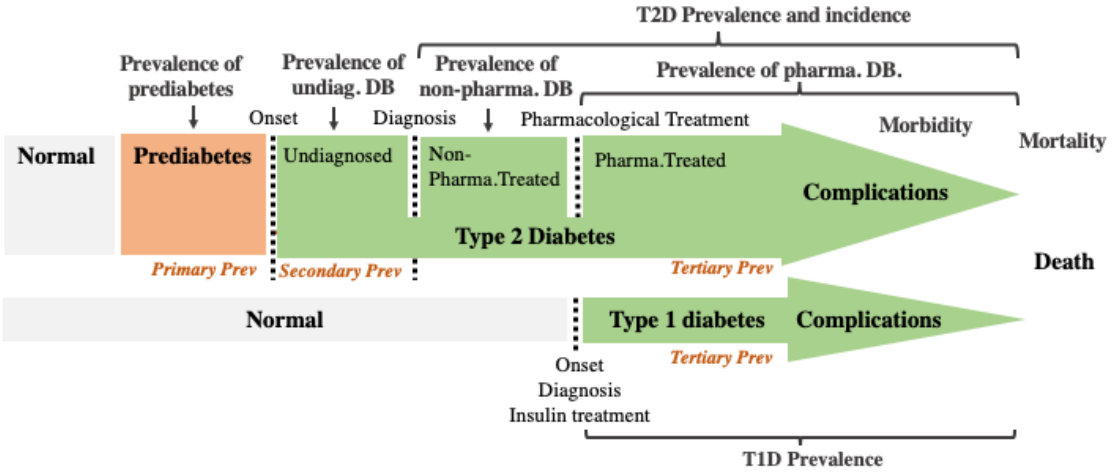
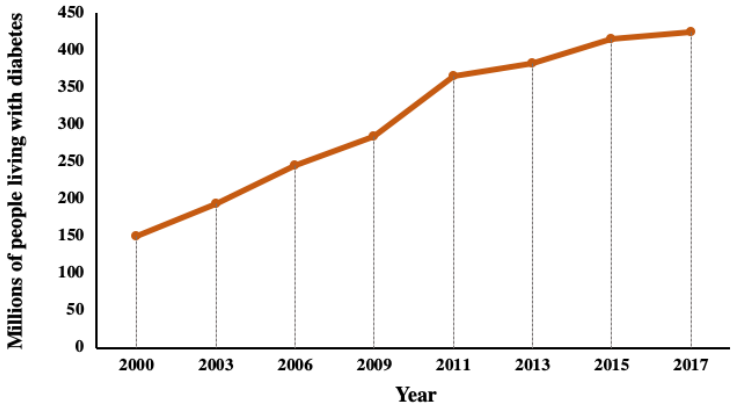


Figure 7. Main indicators in diabetes descriptive epidemiology

2.1 Diabetes descriptive epidemiology

In 2017, the International Diabetes Federation [48] [48] estimated that the number of adults living with diabetes in the World was 451 millions, more than four times higher than its first estimation in 2000 (**Figure 8**) [49]. The number of cases is increasing due to different factors like ageing population, increasing prevalence of risk factors such as obesity, increasing life-expectancy of patients due to better healthcare or more availability of data [50]. Almost 42% of cases live in China or India; the third country with the highest number of cases is the US, followed by Brazil and Mexico [49]. In high income countries, between 87 and 91% of diabetes cases are type 2 diabetes, between 6-12% type 1 diabetes and 1-3% other diabetes subtypes [51].



Adapted from Cho et al.2018 [49]

Figure 8. Number of people living with diabetes in the World between 2000 and 2017

2.1.1 Descriptive epidemiology of type 1 diabetes

Compared to type 2 diabetes, the prevalence - or the proportion of cases among the general population -of type 1 diabetes is very low [52]. There are relevant differences on type 1 diabetes prevalence between countries as well as within countries or ethnicities

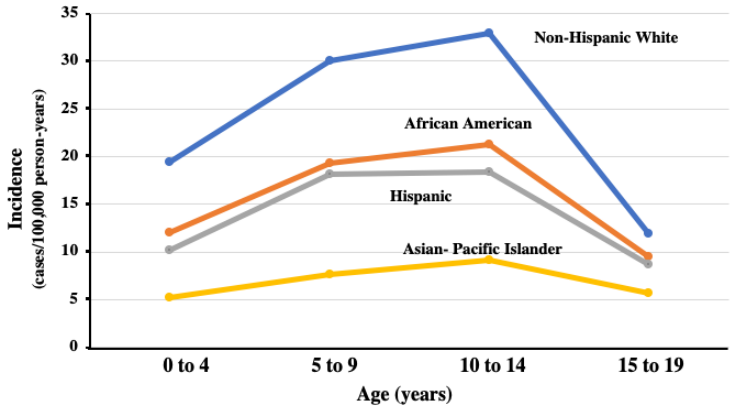
The Scandinavian countries are the regions with the highest prevalence of type 1 diabetes in the world [53]. There are approximately 500,000 children with type 1 diabetes in the World; among them 26% are from European region and 22% from North America and Caribbean region. Regarding ethnicity, the percentage of type 1 diabetes cases is higher among non-Hispanic white [51].

Incidence is the proportion of new cases over a period of time among the population at risk. It is a valuable indicator for public health [54]. Type 1 diabetes incidence is greater in children compared to adults, since 75% of type 1 diabetes cases are diagnosed before

the age of 18 years [52]. The EURODIAB[4] project showed that incident rate increases from birth to its highest point between 10 to 14 years and then it declines after puberty, remaining stable in adulthood [55]. However, the information on type 1 diabetes incidence in adults is scarce. A systematic review on incidence of type 1 diabetes only found 70 articles reporting data from population aged over 15 years [56]. This study also observed gender differences on type 1 diabetes incidence with a male to female ratio of 1.47. The excess of male cases is unusual in autoimmune diseases which are more common among women [25].

Regarding the geographical disparities, the lowest age-adjusted incidence rates in children are found in China and Venezuela (less than 0.2/100,000 person year (py)) while the highest rates are in Finland and Sardinia (more than 30/100,000 py) [57]. Nevertheless, the high incidence rate in the former Italian Island is an exception because a North-South gradient is observed in Europe, with low rates in the Mediterranean countries (Italy and Greece) and very high rates in the Nordic countries (Finland, Norway and Denmark) [58].

The SEARCH[5] study identified important differences on the incidence of type 1 diabetes in individuals aged less than 20 years depending on their ethnicity (**Figure 9**) [59]. Non-Hispanic Whites had the highest incidence rates through all age groups while Asian-Pacific Islanders had the lowest rates and African-American and Hispanic were placed between them with similar rates, especially in the youngest age-group.



Adapted from Mayer-Davis et al. 2009 [60]

Figure 9. Incidence of type 1 diabetes in children by age-group and ethnicity

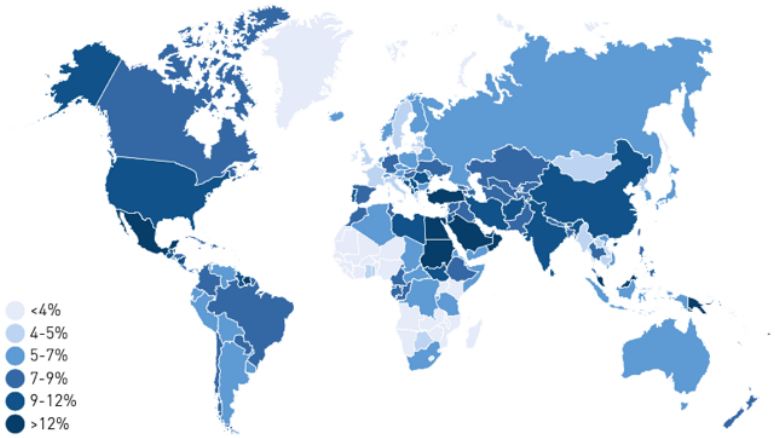
[4] EURODIAB is a European project established in 1989 which aims to study the epidemiology and the etiology of type 1 diabetes in children. A total of 24 centers distributed across Europe participate in the Study

[5] SEARCH is an national multicenter study launched in 2000 for studying diabetes among children and young adults in the US

Both SEARCH and EURODIAB studies have shown an increase of 2.4 to 3.4% per year on incidence rates over the last decades [52]. The strongest increase was described in the youngest age group, from 0 to 4 years (6.3%) [55]. Different hypotheses have tried to explain this trend like the "hygiene theory". Improving standards of hygiene in early life might have an impact in the development of the immune system, leading to an increase in the incidence of autoimmune disease like type 1 diabetes [61]. Other hypotheses are viral infections, toxins and numerous dietary factors.

2.1.2 Descriptive epidemiology of type 2 diabetes

Prevalence of type 2 diabetes varies from countries with very low prevalence like Benin –0.9%- to countries like Tubalu where more than 30% of the population live with type 2 diabetes [49]. Most of the countries with the highest prevalence in the world diabetes are in the Pacific region (**Figure 10**).

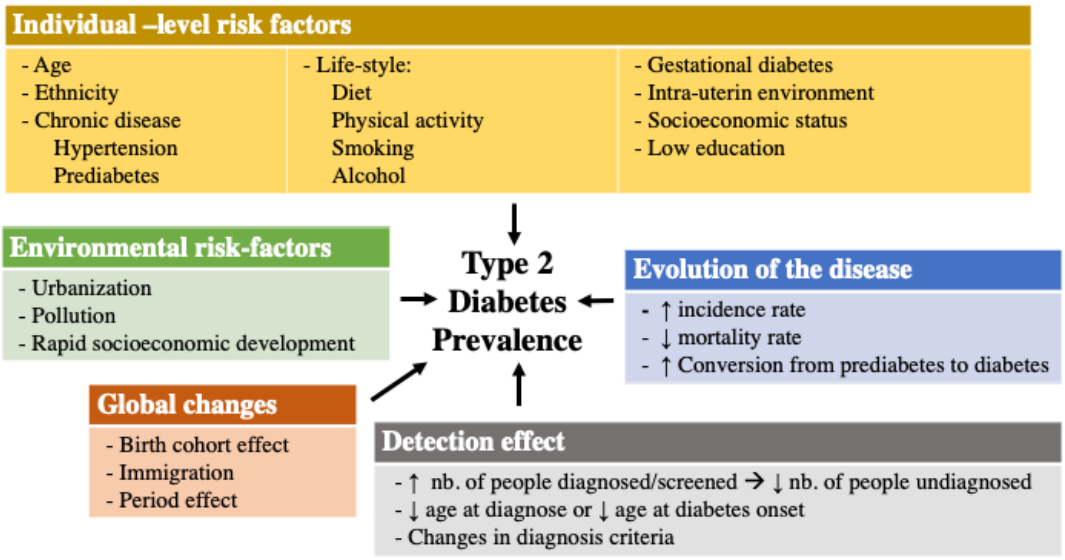


Adapted from Cho et al. 2018 [49]

Figure 10. Age-adjusted prevalence of diabetes in 2017 by country

Some studies have suggested a genetic predisposition of Pacific islanders to develop obesity and diabetes but it seems that the elevated rates of type 2 diabetes are associated to recent changes from traditional diet -fresh fruits, vegetables, poultry and seafood- to western diet -meat, dairy products and processed foods rich in salt and sugar- [62]. This nutrition transition together with an increase of sedentarism and rapid urbanization are also associated with the excessive rates of obesity and diabetes in other regions, like the Middle East-North African region (MENA) [63]. The average prevalence of the region is 9.2%, but in some countries like Saudi Arabia or Egypt, age-adjusted prevalence reaches 17% [64]. Different studies conducted in the US and in France have also observed these elevated rates of type 2 diabetes among immigrants from the MENA region. These studies also described that, conversely to general population, the age-adjusted prevalence rates were higher among women compared to men [65, 66].

Different factors are involved in the prevalence of diabetes. A systematic literature review on factors associated to an increase in prevalence of type 2 diabetes summarized them in five groups: individual level risk-factors, environmental risk factors, evolution of the disease, detection effect and global changes (**Figure 11**) [50].



Adapted from Thibault et al. 2016 **[50]**

Figure 11. Factors responsible for type 2 diabetes prevalence increase

The expansion of individual risk factors associated with diabetes is an important contributor for diabetes prevalence [67]. These factors are also associated with the prevalence of obesity rates which are highly correlated to diabetes rates [68]. There are also environmental factors such as rapid socioeconomic development and urbanizations that lead to obesogenic environments characterized by inhibiting physical activity, low accessibility to fresh food and high availability of junk food [69]. These obesogenic environments are also linked with "global changes" factors like birth cohort effect since it has been observed that the higher prevalence of diabetes in the younger cohort groups could be explained by early exposure in life to this type of environment [70]. Another environmental factors are related to the increasing exposure to pollutants associated with diabetes like bisphenol A [71].

Another group of factors, the "evolution of disease" factors, are directly linked to the dynamics of the disease, including in one hand the increase of incidence and the conversion from prediabetes to diabetes and in other hand the improvement of diabetic patients' survival rate because of better health care [72].

Detection factors can also influence prevalence rates. Active screening reduces the number of people undiagnosed and reduces the age at diagnosis, increasing therefore the

pool of diagnosed individuals and the prevalence of diabetes [73]. Furthermore, the change in diagnosis criteria in 1999, by lowering the level of FPG threshold from 7.8 to 7.0 mmol/L, has enhanced the number of people with diabetes [74].

Unlike type 1 diabetes, the age-distribution of type 2 incidence rates is shifted to the oldest age groups with its highest point between 65 and 85 years [75]. However, there is an increase in diabetes incidence among children and adolescents, especially among girls and in the countries with the highest prevalence of the disease [76, 77]. The incidence is also high in certain ethnic groups, such as Hispanic, African American, Pacific islanders, American and Australian Indigenous populations [74].

Since the first estimations in 1980, type 2 diabetes incidence rates have grown constantly [78]. However, this trend has recently changed in some high income countries where incidence remains stable (in Canada, Italy, Scotland and the United Kingdom (UK)) or even decreases (in the US, Israel, Switzerland, Hong Kong, Sweden, Norway and Korea) [54]. This shift could be due to the effectiveness of public health prevention programs or to other factors associated with screening and diagnosis [79].

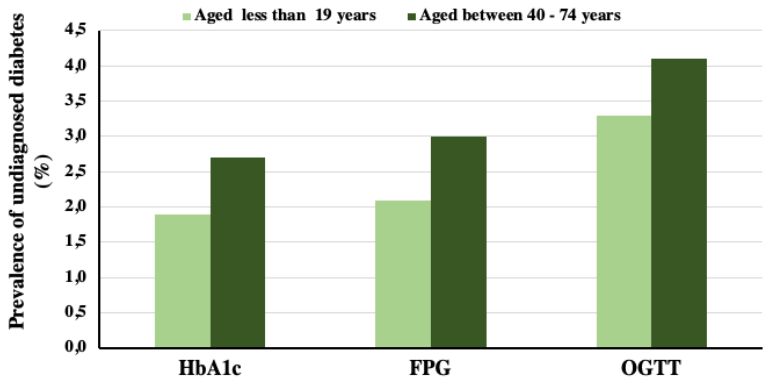
2.2 Descriptive epidemiology of undiagnosed diabetes and prediabetes

2.2.1 Prevalence of undiagnosed diabetes

Undiagnosed and therefore untreated diabetes cases are at very high risk of acute and chronic complications. They represent a major public health issue. The International Diabetes Federation (IDF) estimated that in 2017 at least 49% of all diabetes cases were undiagnosed [49]. However, nationwide information on undiagnosed diabetes are scarce and the results are influenced by the methodology used in the study. A study based on the NHANES[6] data found that the age-standardized prevalence of undiagnosed diabetes differs depending on the diagnostic test applied: when using the OGTT, the prevalence is higher than when using FPG and much higher than when using HbA1c (**Figure 12**) [80].

The latest data from the NHANES have shown that the global prevalence of undiagnosed diabetes in adults in the US was 5.2% [81]. This percentage is very high compared to the prevalence observed in European countries (between 1.6 and 2.0% in the UK, Germany or Denmark) [82-84].

[6] NHANES: The National Health and Nutrition Examination Survey is a program of studies designed to assess the health and nutritional status of adults and children in the US. The survey includes interviews and physical examinations. (**See page 59)**



Adapted from Geiss et al. 2018 [80].

Figure 12. Prevalence of undiagnosed diabetes in the US based on the NHANES data from 2011-2014 by age group, using different diagnostic tests

Prevalence of undiagnosed diabetes is greater in men and in older age-groups [84]. There are also relevant disparities according to socioeconomic status, with almost two-fold higher rates in less educated and most deprived population compared to high educated and less deprived groups [85, 86].

2.2.2 Prevalence of prediabetes

We have already seen that prediabetes is a risk factor of diabetes and cardiovascular diseases. The worldwide prevalence of prediabetes among adults aged 20 to 79 was 7.3% (4.8–11.9%) in 2017 [49]. The highest prevalence was observed in the North American and Caribbean region (14.1% age-adjusted prevalence) although comparison between countries is complicated because different criteria to define prediabetes are used in the studies. The latest NHANES study, applying the ADA criteria, assessed a prevalence of 38% between 2011 and 2014, while European studies conducted in Germany or Luxembourg found a prevalence of 21% and 25%, respectively [81, 83, 86]. In the UK, a national study based on data from the HSE[7] used the WHO criteria and found that the prevalence of prediabetes was 11% in 2009-2013 in adults aged 16 years or older [82]. This study also found important social inequalities: people from lower income, lower occupational class and lower education-level groups had more likely prediabetes than less deprived groups. Other characteristics associated with prediabetes were male gender, higher BMI and waist circumference, and older age (especially over 75 years).

2.3 Diabetes mortality and morbidity

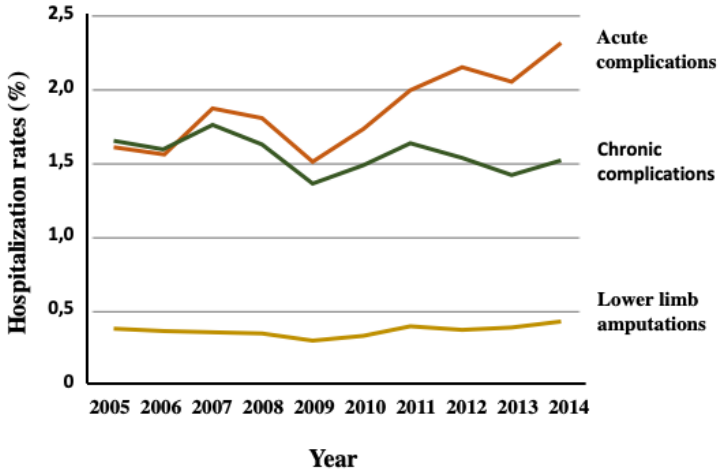
Diabetes accounts for 11% of global all-cause mortality among people aged

[7] HSE: Health Survey for England is an annual, cross-sectional, general population survey of individuals living in England. The survey is done at home and includes a questionnaire, a medical examination and collection of biological samples. The 2009/13 HSE included 24,254 participants

between 20 and 79 years [49]. In 2017, diabetes was the seventh leading cause of death in the US (3% of total death) [87]. However, this data must be interpreted with caution because diabetes is usually underreported in death certificates [88].

A study comparing 3-years mortality in four different waves of NHIS- 1997 to 1998, 1999 to 2000, 2001 to 2002 and 2003 to 2004- found a decrease in mortality rates from 20.3 in the first period to 15,1 per 1000 person-years in the last period (men from 24 to 18.4 and women from 17.7 to 12.1 person-years) [89]. The decrease was also observed in cardiovascular disease mortality –from 9.5 to 5.6 per 1000 py- and in cancer mortality, with a lower intensity -from 3.3 to 3.0 per 1000 py. This decrease has also been observed in other countries and it is attributed to earlier diagnosis and better health care in terms of treatment and regular medical examinations [75]. In the same period mortality rates also decreased in the general population, so it is required to assess the excess mortality due to diabetes [90]. In the previous cited study based on NHIS data, the excess mortality was estimated as the difference between mortality in diabetic population and mortality in general population [89]. As it was described for mortality, the excess mortality decreased over the study periods from 10.8 in the first period to 6.1 per 1000 py in the last period (men from 12.4 to 7.1 and women from 9.7 to 4.7). Also, mortality related to cardiovascular disease decreased from 5.8 to 2.5 (men from 7.5 to 2.5 and women from 4.8 to 1.8 per 1000 py).

Diabetes-related morbidity is linked to acute and chronic complications. There are several determinants, including individual (age at diagnosis, early onset…) and healthcare factors (universal coverage, screening programs…). An American study based on data from different hospitals in the US reported 5,399,199 hospitalizations related to diabetes complications between 2005 and 2014 [91]. The highest hospitalization rates were observed in the youngest age group (from 18 to 44 years). **Figure 13** illustrates the evolution of diabetes-related preventable hospitalization rates and associated conditions. To note, the age-adjusted rates for chronic complications and for lower-extremity amputations remained stable or decreased, while the hospitalizations rates due to acute complications increased over the last 5 years of the study.

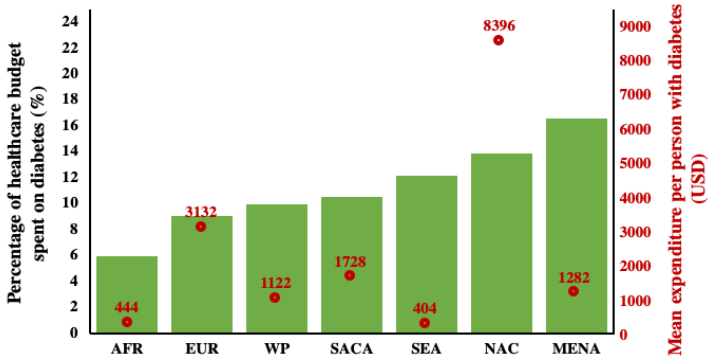


Adapted from Rubens et al. 2018 [91]

Figure 13. Age-adjusted hospitalization relates to diabetes complications rates by associated conditions from 2005 to 20014

2.4 Economic cost of diabetes

The IDF estimated that total global healthcare expenditure due to diabetes among adults was 850 USD billions in 2017.and it is expected to increase by 7% in 2045 [49]. **Figure 14** shows the percentage of healthcare budget spent on diabetes and the annual mean expenditure per adult with diabetes by WHO region.



AFR: Africa, EUR: Europe, WP: West Pacific, SACA: South and Central America, SEA: South East Asia, NAC: North America and Caribbean and MENA: Middle East North Africa Region.
Adapted from Cho et al 2018 [49]

Figure 14. Economic cost of diabetes by WHO region in 2017: percentage of healthcare budget spent on diabetes and annual mean expenditure per adult with diabetes

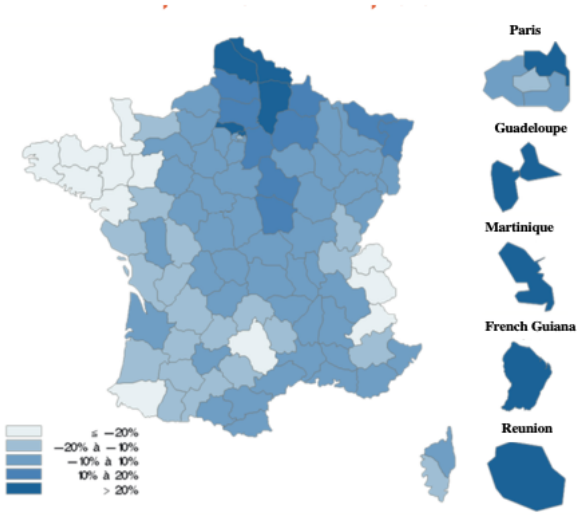
The region where the percentage is the lowest is the African (6%) and the region with the highest percentage is the MENA one (16.6%). These figures can be explained due to the elevated prevalence of diabetes in the MENA region as described before. Following the MENA region, we found the North America and Caribbean region (14% of all healthcare budget). In this region, the annual mean expenditure per adult with diabetes is the highest (8,396 USD) and is almost three times the one of Europe (3,132 USD).

However, these figures only estimated the medical cost of diabetes. In 2017, the total cost of diabetes in the US was 327 billion USD, 73% due to direct medical cost and 27% due to indirect cost in reduced productivity [92]. This indirect cost included: increased absenteeism (3.3 billion USD), reduced productivity for people in and out the labor force (26.9 billion USD and 2.3 billion USD), inability to work because of disability (37.5 billion USD), and lost productivity due to premature deaths (19.9 billion USD).

2.5 Descriptive epidemiology of diabetes in France

More than 3,3 million people were pharmacologically treated for diabetes in France in 2016, corresponding to a prevalence (both type 1 and type 2 diabetes) of 5% [93]. As previously described in other countries, the rates were highest in men and in older population. One in five men and one in seven women aged between 70 and 85 years were pharmacologically treated for diabetes.

Besides, there are also important territorial disparities, as represented in **Figure 15** [93]. In 2016, age-standardized rates were almost two times higher in the FOT[8] – Martinique 7,7%, French Guiana 7,8%, Guadeloupe 9,2% and Reunion 10,1%- than in France. Also, in the FOT, and conversely to Metropolitan France, the prevalence was highest in women. Both in Metropolitan and FOT, the prevalence of pharmacologically treated diabetes increased with age. However, there was an excess of diabetes rates in younger age-groups in FOT compared to metropolitan France [94].

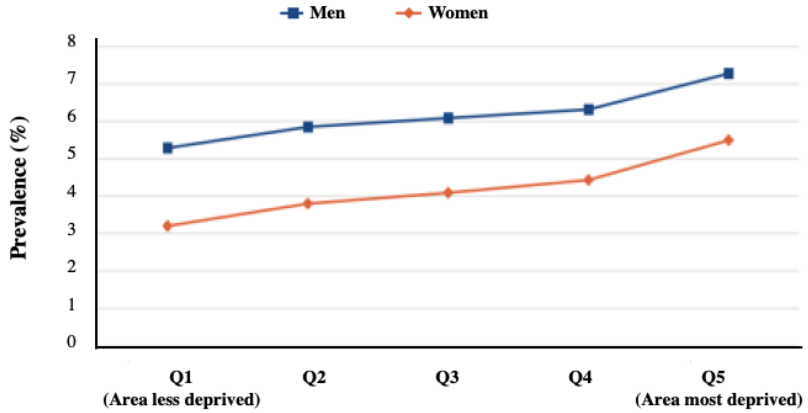


Adapted from Santé Publique France 2019 [95]

Figure 15. Prevalence of pharmacologically treated diabetes in France in 2016 by department

[8] FOT: French Overseas Territories are the French regions outside European continent: three in the Caribbean region (Guadeloupe, Martinique and French Guiana) and one in the Indian Ocean (Reunion)

There were also important socioeconomic inequalities on diabetes prevalence. Comparing the prevalence between people aged less than 60 years old who benefited from CMUc[9] and those who did not benefit, the rates were two times higher in the first group. **Figure 16** represents diabetes prevalence by quintile FDEP09[10] and by sex. The prevalence is associated with the level of deprivation of the living area, with the highest rates in the fifth quintile.



Adapted from Santé Publique France 2018 [93]

Figure 16. Pharmacologically treated diabetes prevalence in France in 2016 by FDEP quintile

There are no nationwide data on diabetes incidence in France except a study assessing the evolution of type 1 diabetes in children under 15 years old [96]. This study observed an incidence rate of 19.1 cases per 100,000 py in 2015 and an annual increase of 4% between 2010 and 2015.

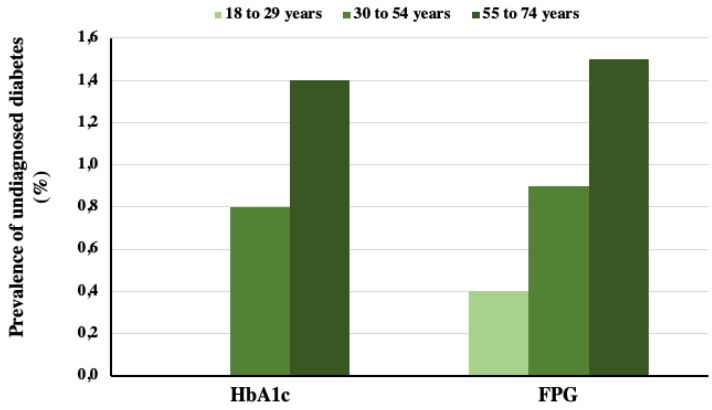
Concerning the prevalence of prediabetes and undiagnosed diabetes, the latest estimations were done in 2006 in the context of ENNS[11] [97]. The study used FPG and HbA1c data from a representative sample of adults aged 18 to 74 years and living in Metropolitan France. The prevalence of undiagnosed diabetes was 1% using FPG and 0.8 using HbA1c and 1.4% using both (1.6% in men and 1.1% in women). Based on the WHO criteria, the prevalence of prediabetes was 5.6% while applying the ADA criteria it was 15.5%. Both undiagnosed diabetes and prediabetes prevalence increase with age

[9] CMUc: *Couverture maladie universelle complémentaire* or universal complementary healthcare is a complementary healthcare insurance scheme available in France for people with low annual income. **See page 72**

[10] FDEP: French deprivation index is an ecological index of deprivation calculated at the municipality level which is based on 4 variables: percentage of blue-collar workers in the labor force, percentage of high school graduates in the population aged 15 years or older, unemployment rate in the labor force, and median income per household. So far, there have been two estimations, one using data from the census of the year 2009 (FDEP09) and another one using data from the census of 2013 (FDEP13)

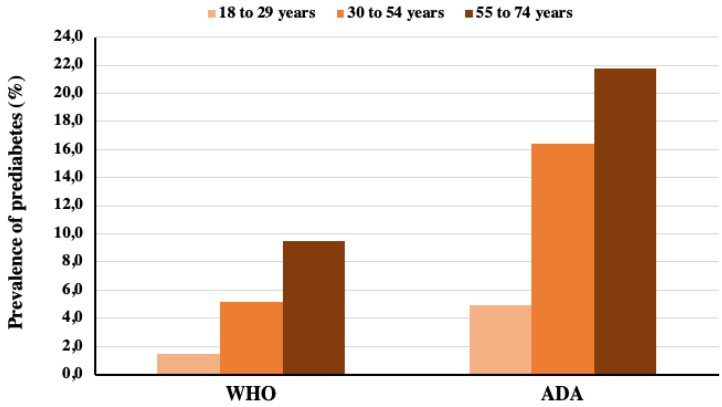
[11] ENNS: *Etude nationale nutrition santé* or the French Nutrition and Health Survey conducted between 2006 and 2007. **See page 69**

(**Figure 17** and **Figure 18**). The study also assessed the prevalence of diagnosed diabetes and pharmacologically treated diabetes based on data from a self-administered questionnaire. The prevalence of diagnosed diabetes was 4.6% and the prevalence of pharmacologically treated diabetes was 3.7%, meaning the prevalence of non-pharmacologically treated diabetes was 0.9%.



Adapted from Bonaldi et al. 2011 [97]

Figure 17. Prevalence of undiagnosed diabetes in France based on the ENNS data from 2006-2007 by age group, using different diagnostic tests



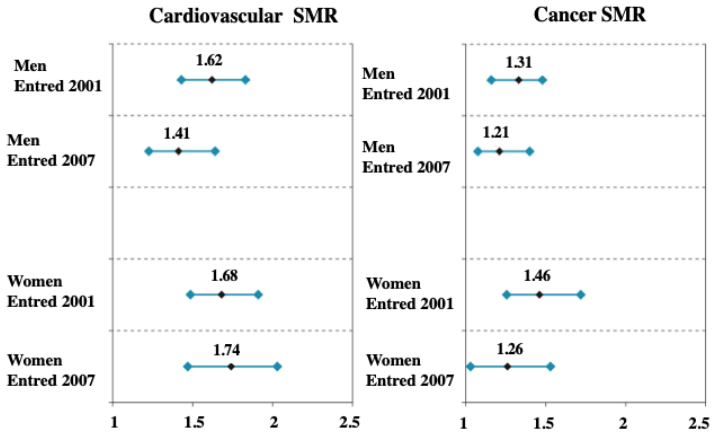
Adapted from Bonaldi et al. 2011 [97]

Figure 18. Prevalence of prediabetes in France based on the ENNS data from 2006-2007 by age group, using WHO or ADA FPG criteria

Regarding mortality in France, a recent study has compared 5-year mortality and excess mortality trends from two consecutive waves of the Entred[12] study (Entred 2001 and Entred 2007) [98]. Over 2002-2006 period, the mortality rate was 48.5 per 1000 in men and 30.5 per 1000 in women. They were respectively 35.8 and 37.1 per 1000 over the period 2007-2012, Therefore, the decrease observed during these two 5-years periods was 26% in men and 11% in women. During the second period, 2007-2012, the

[12] Entred study: *Echantillon National Témoin Représentatif des personnes Diabétiques* (National representative sample of people with diabetes. **See page 70**

percentage of death due to cardiovascular diseases was 27% in men and 33% in women and the percentage due to cancer was 32% in men and 22% in women; these rates have only decreased significantly for cardiovascular diseases in men.



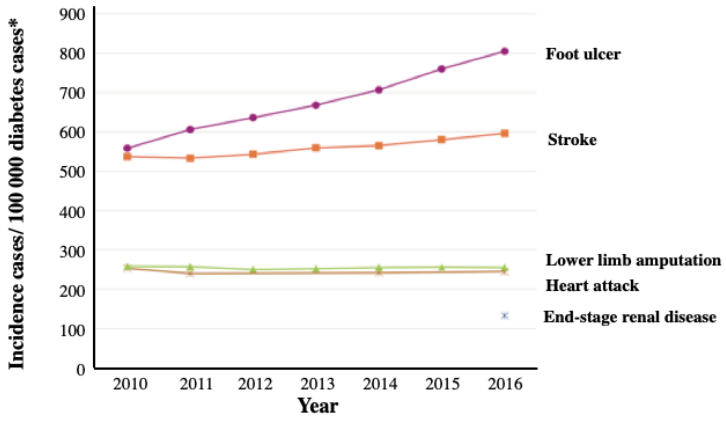
Adapted from Mandereau-Bruno 2007 [98]

Figure 19. Excess mortality rates due to cardiovascular and cancer causes in the 2002-2006 and 2007-2012 periods) by sex

Compared to the general population, the excess mortality rate measured using SMR[13] was 1.3 in men and 1,5 in women over the 2007-2012 period. It decreased in men and remained stable in women compared with the previous period. In **Figure 19** ,the cardiovascular disease' SMR and the cancer's SMR in the two Entred waves by sex are represented. There was a non-significant decrease for all groups, except for cardiovascular disease SMR in women where a non-significant increase was observed.

In 2016, 244 in 100,000 people pharmacologically treated for diabetes were hospitalized due to acute coronary syndromes [99], 596 in 100,000 due to stroke, 805 in 100,000 due to foot ulcer, 255 in 100,000 due to lower limb amputation and 133 in 100,000 due to an incident end-stage renal disease [93]. Men and people aged over 69 years were more likely to develop chronic complications. As described for diabetes prevalence, there are relevant regional disparities with the highest incidence rates of incident end-stage renal disease in the FOT. There were also important disparities concerning socioeconomic status, since the age-adjusted incidence of hospitalization were higher in people younger than 60 years old who benefit from the CMUc and in those living in the most deprived areas [100, 101]. Compared to previous data from 2010, the incidence rates of hospitalizations due ACS and amputations remained stable while those due to stroke or foot ulcer increased (**Figure 20**) [93].

[13] SMR: standardized mortality ratio is the ratio of observed deaths in the group of interest – i.e. diabetes patients- to expected deaths in the general population



Adapted from Santé Publique France [93]

Figure 20. Evolution of incidence of hospitalization due to diabetes complications in France

In 2010, the total cost of pharmacologically treated diabetes in France was 17.7 billion EUR [102] . At least 4.2 billion EUR was dedicated to treatment of diabetes-related complications and 2.5 billion EUR to treatment of diabetes and regular medical examinations. The mean expenditure was 6,930 EUR for type 1 diabetic patients and 4,890 EUR for type 2 patients, but this mean increased among type 2 patients treated by insulin to 10,400 EUR. These expenditure did not include indirect cost, which should include disability pensions (on average 7,060 EUR per person each year) and daily allowances (2,661 EUR per person each year).

2.6 Conclusion

Despite having different epidemiology dynamics, type 1 and type 2 diabetes represent an important burden for the society because of their prevalence, incidence, morbidity and mortality. Moreover, the proportion of undiagnosed diabetes and of prediabetes is high – enhancing the number of people at very high risk of developing diabetes and cardiovascular diseases. The total cost of diabetes does not includes only the direct cost due to medical care (treatment, hospitalizations …), it also comprises indirect costs related to loss of productivity.

Notwithstanding, although France presents lower prevalence rates compared to other countries, the burden of diabetes remains relevant with significant socioeconomic and territorial inequalities. There is an important gap on information since the prevalence of undiagnosed diabetes and prediabetes has not been updated since 2007 Additionally, there is no data on incidence of diabetes among adults in France and the trends in the diabetes epidemic have not been studied. Additionally, the national data do not differentiate type 1 and type 2 diabetes.

3. Diabetes surveillance

In the previous section, we presented the most relevant indicators in diabetes epidemiology. The Diabetes Surveillance Systems estimate these indicators using different data sources: national studies – representative sample of study population- , health-administrative databases – including [almost] all study population- and other sources like patient registries (**Figure 21**). We aimed to present in this section the definition of surveillance, the different data sources for diabetes surveillance and how they integrate in different surveillance systems over the World, with special focus on the French System.

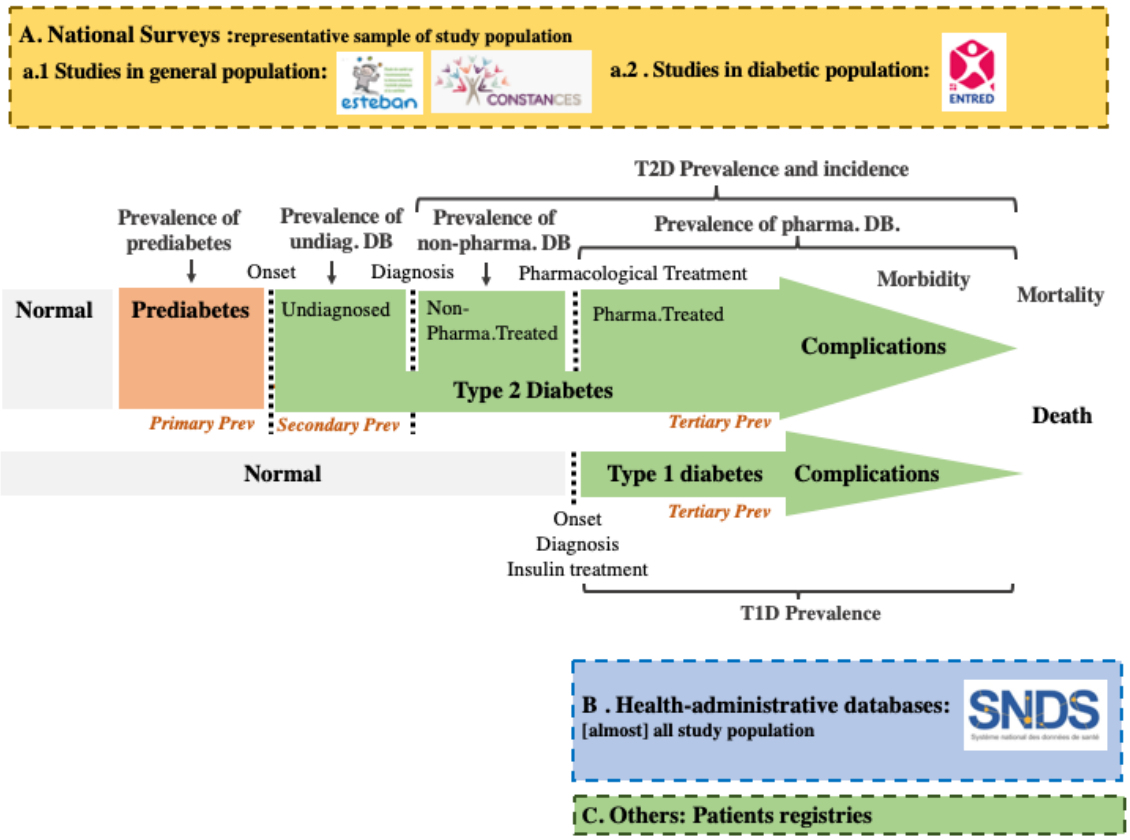


Figure 21. Data sources of the French Diabetes Surveillance System

3.1 Public Health Surveillance

The origin of the term surveillance is the French word *surveillance*, coming in the English language from the French Terror. At the end of the XVIII century, the *Comités de surveillance révolutionnaires* were in charge of monitoring the actions and movements of suspicious people in every French municipality [103]. In the XIX century, William Farr founded the concept of public health surveillance when he started to collect and to analyze routinely the data from vital registration of the General Register Office in England and Wales, to record the causes of death and to assess, for the first time, mortality rates by groups of population [104]. This information was later used by John Snow in his famous study of the 1854 Broad Street cholera outbreak.

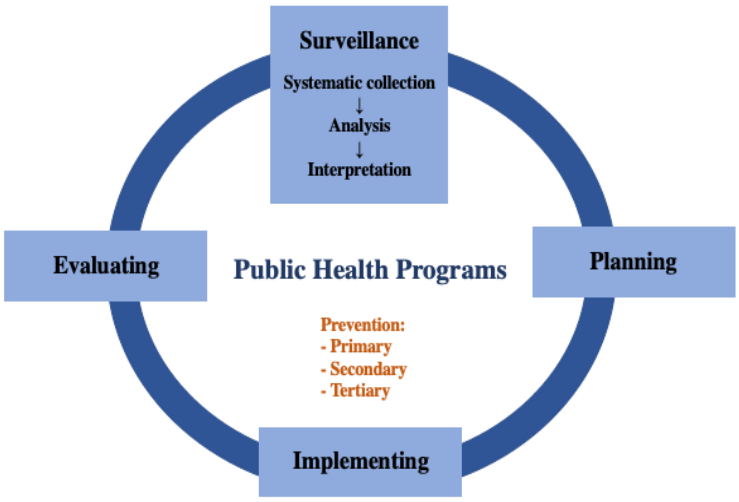


Figure 22. Public health programs

In 1986, the CDC[14] defined public health surveillance as: "the ongoing systematic collection, analysis, interpretation and dissemination of health data for the planning, implementation and evaluation of public health action" [105]. Surveillance is a key component of public health programs, together with planning, implementation and evaluation (**Figure 22**).

The indicators applied in public health surveillance are not only related to diseases but also to risk factors like tobacco use or alcohol consumption [104]. Communicable and non-communicable or chronic disease surveillance have common characteristics. Both of them analyze variations related to time, place and individual characteristics and their results are used for developing and evaluating prevention programs [106]. However, they present important differences summarized in **Table 5**.

[14] CDC: Center for Disease Control Prevention is the national public health institute of the US

Table 5. Main differences between chronic disease and communicable disease surveillance

	Chronic Diseases	Communicable Diseases
Study units	Rates	Number of cases
Observed changes in trends	Years	Days or weeks
Frequency of reporting results	Annually	Weekly or daily
Data sources	Existing databases	Reporting systems

Adapted from Gil et al. 2015 [106]

First, the basic study unit for communicable disease surveillance is usually the number of cases while for chronic disease surveillance, it is rates or units within general or specific population. Observing changes in trends requires shorter periods of time for communicable diseases. This has also an impact in the reporting frequency of results: the frequency for chronic diseases is annual or greater while it is daily or weekly for communicable diseases. Finally they differ in the data sources to estimate indicators. Surveillance of communicable diseases is based on reporting systems where new cases are actively reported to public health authorities. Though, chronic disease surveillance is based on already existing data sources like vital statistics or surveys.

3.2 Data sources for diabetes surveillance

We have seen that chronic disease surveillance uses different existing data sources to estimate health indicators. Now, we will describe the main population-based data sources currently used. These data sources can be classified in: health surveys, disease registers and health-administrative databases [107] (**Figure 23**).

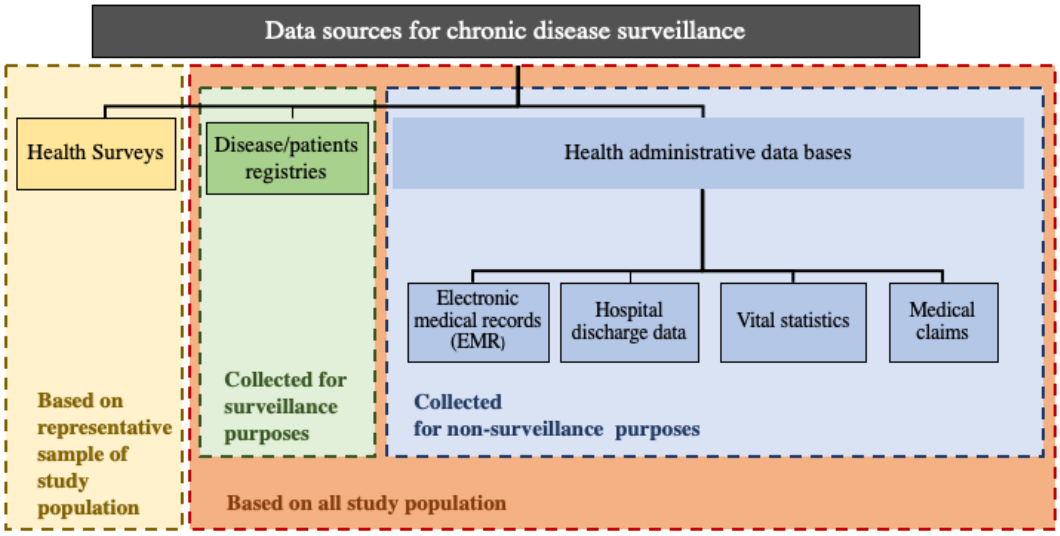


Figure 23. Data sources for chronic disease surveillance

3.2.1 Health surveys

Health surveys are based on a representative sample of a priori defined study population [108]. One example of health surveys is the NHANES. Since 1960, the NHANES has been conducted periodically to assess the health and nutritional status of adults and children in the US and it comprises personal household interviews, physical examination, and laboratory testing of nearly 10,000 adults and children [109]. The detail information collected in the NHANES has been used to study the prevalence of diagnosed diabetes, undiagnosed diabetes and prediabetes in the US and to assess inequalities among certain groups of population [80, 85].

The data collected in these studies not only include information on diabetes diseases but also on life-style, socioeconomic status and other individual characteristics that are very useful to identify target populations and to design effective public health programs [109]. Despite the size of the sample allows to calculate indicators at a national level, it is usually not large enough to do estimations at lower geographical levels like state or local levels. Therefore, these data cannot be used to compare rates across regions [104]. Another limitation is that the data recorded in the questionnaire can be subject of different bias such as recall bias or social desirability bias [110]. Also, the examination data and laboratory data are exposed to examiner effects and unexpected variations. There is also an important participation bias since people with poor health conditions or in severe stages of the disease are less likely to participate in these studies [111].

3.2.2 Disease registries

Disease or patient registries is a list of individuals who developed a specific disease or health conditions, or who have followed certain medical procedures [112]. They are collected through organized programs involving three different stages: case-finding, follow-up and statistical utilization of the information collected. Related to diabetes surveillance, we can cited the Norwegian Childhood Diabetes Registry (NCDR) [113]. Since 2006, the NCDR collects all new cases of diabetes in Norway in children reported from the pediatric health care departments (after informed consent from the child and/or the parents). The information recorded has been used to estimate incidence rates of diabetes among children aged from 0 to 14 as well as for other studies for improving diagnosis and diabetes management in this age-group [114, 115]. The NCDR also participates in the EURODIAB project [116].

Disease registries are better sources than health surveys to estimate incidence rates,

because they identify more accurately new diabetes cases [107]. Also, most of the information collected in these registries come from medical records so it is less exposed to response bias [117]. However, the recording of this information and the registration of new cases can be influenced by current medical practices and health care activities. For example, patients living in rural areas with low density of practitioners are less likely to be registered than those living in urban areas. The cost-efficiency ratio of these sources for Health Care Systems is unfavorable because to collect exhaustive and high quality data on one disease, they require important budgets and organizational structures [112].

3.2.3 Health administrative databases

Health administrative databases (HAD) include a heterogenous group of sources of massive data recorded routinely for non-surveillance purposes [118]. These sources are not based on representative samples, but comprise all study population. Hospitals, health maintenance organizations and health insurance organizations are in charge of maintaining these databases where the interesting information is hidden by big amounts of useless data. Since they are not recorded for surveillance purposes, case ascertainment must be done through identification algorithms.

The major HAD for chronic disease surveillance are: electronic medical records (EMR), hospital discharge databases, vital statistics databases and medical claims-health care reimbursement databases.

The development of new technologies have led to transform the patient's paper chart into a digital version, the EMR [119]. EMR includes updated information on diagnosis, results of prescribed tests and treatments. Shared individual EMRs from different health care centers across the country are a valuable source for surveillance. An example is the Primary Care Database in the UK [120]. Almost all primary care consultations in the UK are computerized with software platforms where practitioners can write the patient's information by coding systems. These data are shared for research purposes only if the patient gives her/his consent. The EMR can offer exhaustive and updated information for surveillance, but this quality of information can be affected by miscoding and misclassification [121]. They are also limited by the software used where users cannot put more than a certain number of diagnosis codes or they can only introduce measurements in integers.

Hospital discharge databases are the abstracted records associated with patients' hospitalizations [105]. These data include diagnosis, treatment and payment information. Public and private hospitals share through national or regional networks this information

and it is useful for chronic disease surveillance, especially for morbidity assessment. One of these networks is the Discharge Abstract Database and Hospital Morbidity Database where hospitalization data from different Canadian provinces like Ontario or Manitoba are stored [122]. Conversely to other sources, we can have access to individuals with poor health conditions or in severe stages of the disease through hospital discharge databases [123]. The main limitations of hospital discharge databases are difficulties to distinguish hospital admission patterns for a disease from actual patterns of disease occurrence and the error bias related to disease coding.

Governments record data at national population level on deaths, live births and other situations like marriages and divorces through civil registration administrative systems; these are the so called vital statistics databases [124]. As we have seen, vital statistics as death certificates were one of the first data sources for public health surveillance [104]. Death certificates give access to information related to the individuals characteristics (age, sex, place of birth), place and date of death and the cause of death [125]. The causes of death in death certificates are classified in immediate cause, or the disease or event directly leading to death, and in conditions contributing to death, or intermediate and contributory conditions related to death not included as the immediate cause of death field. It has been estimated that the National Vital Statistics System in the US analyzes 2.6 million deaths each year and it is used to estimate mortality rates [126]. In contrast to disease registries, vital statistics use the same organizational structure in an homogeneous routine collection which can be used for surveillance not for one but for various diseases. The volume of data information allows to regional or local comparisons [127]. Also, the coding of causes of death is standardized internationally by the use of ICD[15] codes. Therefore, data from different countries can be easily contrasted. Digital systems have been developed for recording data from Vital statistics, but data treatment remains time consuming to validate and to update the information [125]. The percentage of papers based on death certificates remains very high but certain diseases such as diabetes are underreported in death certificates [128].

The administrative information recorded by public and private health insurance funds is a worthy source for surveillance [118]. It contains data on billing or reimbursement of provided health care services (physician consultations, dispensed drugs

[15] ICD: International Statistical Classification of Diseases and Related Health Problems is a medical classification list by the WHO. The latest version is the ICD 10, available since 1994

or test performed) as well as individual data on age, sex, residence or socioeconomic status [122]. In the US, people aged 65 years or more or living with disabilities or with end-stage renal disease benefit from the federal health insurance program, Medicare [129]. The administrative data from Medicare alone or combined with other sources is widely used to assess morbidity and mortality in the US [130]. Information from medical claims is frequently recorded automatically through electronic systems so it is updated faster and less exposed to miscoding than other sources [131]. Also, diseases with low prevalence or groups of people less likely to participate in surveys such as immigrants or deprive groups can be studied through this type of sources. The quality of data recorded is better than the data from health surveys because there are no at risk of response bias [132]. However, contrary to health surveys, there is no information on lifestyle characteristics. Moreover, there is a lack of information on the results of medical examinations or performed tests, meaning for example that obesity or undiagnosed diabetes prevalence cannot being estimated.

3.3 Diabetes surveillance systems

After having described different sources for chronic disease surveillance, we will now explain how they can be integrated in national or regional diabetes surveillance systems, with three examples across the World: the Scottish System, the Diabetes Registers from Nordic Countries and the US Diabetes Surveillance System.

3.3.1 Scottish Care Information-Diabetes

The Diabetes Surveillance System in Scotland is mainly based on the Scottish Care Information-Diabetes (SCI-Diabetes), a diabetes patient management system established in 2000 [133]. Every diabetes case identified at primary or secondary care is included in the system after patient's consent. Each patient has a personal identification number which allows the system to collect the EMR from different healthcare facilities such as general practitioners, hospitals and local laboratories. Therefore the system is able to follow up almost all people living with diabetes in Scotland [134]. It is an ideal surveillance system because it combines the positive characteristics of Disease Registries and EMR sources. However, the cost of the surveillance system is very high making it not suitable for countries with big populations or with large territories. In addition, the SCI-Diabetes has no information on prevalence of undiagnosed diabetes and prediabetes, so to estimate these indicators, other sources like the Scottish Health Survey are required [135]. Finally, the system alone does not allow transversal surveillance of other diseases or health conditions. For example, the percentage of diabetes cases among people who

are treated for hypertension cannot be estimated only using this system.

3.3.2 The Danish National Diabetes Register

In the Nordic countries, each citizen has a unique personal identification number used for all major health events and administrative purposes [136]. This is a valuable tool for public health surveillance since it allows to cross-reference information available in medical, social and other administrative data sources from all the population living in the country. The Danish National Diabetes Register is a population-based diabetes register which follows all diabetes cases in Denmark through different health-administrative registries [75]. Diabetes cases are ascertained in three different sources:

a) The National Patient Register (NPR) is a HAD where data from discharge data from hospitals and outpatient clinics can be found since 1994. Each diagnosis is coded in ICD-10.
b) The National Health Insurance Service Registry (NHISR) contains information on reimbursements of healthcare services from general and specialist practitioners since 1973. No information on diagnoses or test results is available in this registry.
c) The Register of Medicinal Product Statistics (RMPS) is a prescription register on all prescriptions dispensed at Danish pharmacies. Each drug is identified though the ATC code.

A new diabetes case is identified when she/he meets at least one the following criteria:

- Diagnosis code of diabetes in the NPR
- Feet examination for diabetic patients in NHISR
- The date of the fifth blood glucose measurements within one year in NHISR
- Two blood glucose measurements per year in five consecutive years in NHISR
- Second purchase of antidiabetic drugs (insulin or oral glucose-lowering drugs) in RMPS

The Danish surveillance system allows longitudinal surveillance of all diabetic patients in Denmark before and after diagnosis, including not health-related information such as work history or residential history [137]. Contrary to SCI-Diabetes, transversal surveillance of various diseases at the same time is possible. Due to the absence of test results, other sources like the Danish Health Examination Survey are needed to estimate the prevalence of undiagnosed diabetes and prediabetes [84]. In addition to these data sources, all children diagnosed with diabetes before the age of 15 years are included since

1996 in the Danish Registry of Childhood and Adolescent Diabetes (DanDiabKids) [138]. One important limitation is due to the lack of information on diagnoses or test results from the general and specialist practitioners which makes the complete classification of cases in type 1 and type 2 diabetes unfeasible. The characteristics of the country helps the Danish Diabetes Register to produce high quality indicators but this system would be difficult to implement in countries with larger populations or presenting relevant regional inequalities on healthcare.

3.3.3 The US Diabetes Surveillance System

The CDC supports the national- and state-level diabetes surveillance system by analyzing, interpreting and reporting data on diabetes across the US [139]. Chronic disease surveillance faces many challenges in the US such as absence of universal healthcare, a big population irregularly distributed through a large territory or difficulties on cross-referencing data sources due to the lack of a unique personal identification number [140]. To overcome these limitations, the US Diabetes Surveillance System combines the information from different data sources to build a complete picture of diabetes burden in the US [141].

National health surveys are the keystone of the US diabetes surveillance system [139]. Some of them have been already presented like the NHANES but we can also cite the National Health Interview Survey (NHIS) or the Behavioral Risk Factor Surveillance System (BRFSS). These surveys have been conducted periodically for more than 50 years, providing exhaustive information on the evolution of diabetes and its complications over the last decades [85].

Vital statistics on birth and mortality from the US Census Bureau are used to assess national and state total resident population which is secondly applied for the estimation of prevalence and incidence rates [139]. Also, data from death certificates is exploited to complete information on mortality indicators [142].

Some of the information on long- and short-term complications come from hospital discharge data [139]. The estimation of the number of hospitalizations for major cardiovascular diseases, lower-extremity amputation and diabetic ketoacidosis are based on the National Inpatient Sample (NIS) where data from more than seven million hospital stays is stored. The Nationwide Emergency Department Sample (NEDS) database is the source employed to calculate the number of emergency department visits for hypoglycemia and hyperglycemic crisis.

The US diabetes surveillance system also uses data from public health insurance

funds like Medicare or Medicaid [143]. Another example of data source from public health insurance funds is the National Data Warehouse (NDW) of the Indian Health Service (IHS) [139]. The NDW collects all the information from the different IHS facilities serving to the different federally recognized tribes of American Indian/ Alaska Native people distributed across the US. This information is relevant since these ethnicities are not well-represented in national health surveys.

3.4 The French diabetes surveillance system

We have seen that the SCI-Diabetes followed by the Danish Diabetes Register are highly efficient surveillance systems. However, they are not appropriate for countries with large population, with extended territories or with relevant disparities between regions. This is particularly the case of France, a country with a population of 67 million people unevenly distributed among 18 regions covering a territory of 643,801 square kilometers and including five regions situated overseas (in the Caribbean, the Pacific and the South African regions) [144].

Chronic surveillance system in France faces several limitations due to country characteristics but it also benefits from the structure of the French Healthcare System. In France there is universal health care coverage which is financed by the government through national public health insurance funds. Each beneficiary has a unique personal identification number and a personal smartcard (*carte vitale,* **Figure 24**), allowing information on health care utilization to be electronically recorded [145].

Figure 24. Example of personal smart-card from the French Health Care System

The French diabetes surveillance system is managed by *Santé Publique France*, the French institute of public health [146]. There are no diabetes disease registries in France except three covering the regions of Aquitanie, Franche-Comté and Languedoc-Roussillon which only include children diagnosed with type 1 diabetes [147, 148]. Therefore, the French national surveillance systems is mainly based on two types of data sources: health surveys and health administrative databases (**Figure 25**).

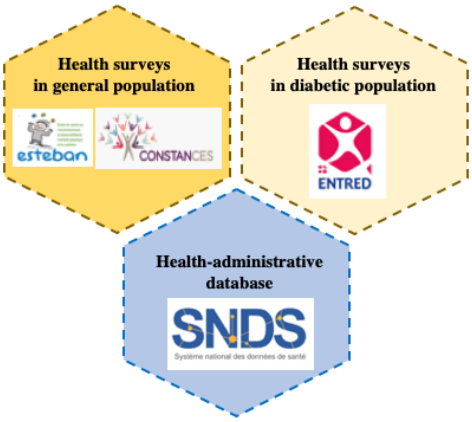


Figure 25. Main sources of the French diabetes surveillance system

3.4.1 Health surveys

We have seen that health surveys are based on a representative sample of the study population. This study population can be national general-based population, like in the national nutrition and health survey (*Etude national nutrition santé*, ENNS) and the health survey on environment, biomonitoring, physical activity and nutrition (*Etude de santé sur l'environnement, la biosurveillance, l'activité physique et la nutrition*, Esteban). It can also be specific populations like people with diabetes, such as in the national representative sample of people with diabetes *(Echantillon National Témoin Représentatif des personnes Diabétiques*, Entred).

The ENNS was a study conducted between 2006 and 2007 to describe dietary intake, nutritional status and physical activity in children and in adults [149]. A multistage, stratified random sample of non-institutionalized people living in Metropolitan France was used. The sample of adults included around 3000 individuals aged from 18 to 74 years. The study included a self-administered questionnaire, a health examination and a blood sample drawn to perform laboratory tests, such as FPG and HbA1c measurements. In the previous section, we have seen that the latest estimation on the prevalence of undiagnosed diabetes and of prediabetes were based on the results of this study [97].

The Esteban study was carried out between 2014 and 2016 using the same methodology [150]. Its objectives were to understand the main environmental and lifestyle risk factors associated with chronic diseases in children and in adults. The study design was analogous to the one of ENNS, including sample selection on non-institutionalized individuals living in metropolitan France (except Corsica), face-to-face interviews and self-administered questionnaire and finally medical examination with

collection of biological samples (blood, urine and hair). Also, the participants‘ data on hospitalizations and dispensed health care reimbursement were gathered from the SNDS[16]. Esteban recruited 3476 adults aged from 18 to 74 years. This sample allows to estimate the prevalence of undiagnosed diabetes and prediabetes in France. However, as ENNS' sample, it is not large enough to decline these results by individual characteristics like obesity status or socio-economic level.

Another data sources for the French diabetes surveillance system is Entred, which specifically included people with diabetes. So far, there have been three waves of Entred: in 2001, in 2007 and an ongoing wave in 2019 [151]. The sample is selected among all pharmacologically treated diabetes cases living in metropolitan France (the last wave has also included a subsample of people from FOT). After inception, the participants received by mail a questionnaire on socio-demographics, health status, quality of life and quality of care [152]. Then, they provided the name and the professional address of their healthcare providers, in order to mail them another questionnaire to gather data from clinical and biological measurements. Entred 2007 included a random sample of 8,926 adults treated for diabetes and Entred 2019 included 13,000 individuals. The information from Entred has been used to assess diabetes morbidity and mortality as well as to study quality of life and quality of care of diabetic patients in France [153-155]

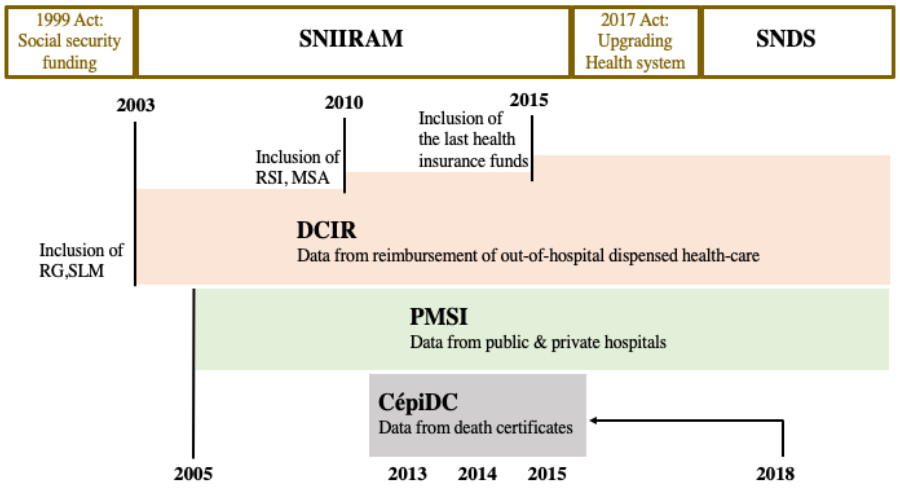
3.4.2 Health administrative databases

The French national health insurance information system (*Système National de Données Santé*, SNDS) is a digital warehouse of health-administrative data from almost all French population which is hosted and managed by the National Health Insurance Fund for Salaried workers (*Caisse National d'Assurance Maladie des Travailleurs Salariés*, CNAMTS) [156]. It was created in 2003 following the directions of the 1999 social security funding law under its previous name, the National Health Insurance Inter-Scheme Information System (*Système d'information inter-régime de l'assurance de maladie,* SNIIRAM) with three objectives: to improve the quality of healthcare, to contribute to public health politics and to inform health practitioners.

Each insured individual has a unique personal identification number named *numéro d'identification au répertoire* (NIR) which is transformed into an anonymous code to preserve the identity of the patients [131]. This anonymous code enables to cross-reference the main health-administrative data sources in the SNDS: the inter-scheme

[16] SNDS: The French national health insurance information system (*Système National de Données Santé*). See page 65

consumption datamark (*Données de consommation inter-régimes*, DCIR), the program for the medicalizations of information systems (*Programme de médicalisation des systèmes d'information*, PMSI) and the French national registry of specific mortality causes (*Centre d'épidémiologie sur les causes médicales de décès*, CépiDc) (**Figure 26**).



RG Insurance fund for salaried workers; SLM Insurance fund for students and civil servants; RSI Insurance fund for self-employed people; MSA Insurance fund for agricultural workers

Figure 26. Evolution of the French national health insurance information system (SNDS)

3.4.2.1. The DCIR

The DCIR gathered from the different Public Health Insurance Funds in France data from reimbursements of out-of-hospital healthcare, like medical consultations, medical procedures, treatments and tests performed, . These funds cover 70% of healthcare cost except in specific cases like having a low-income (*Couverture maladie universelle complémentaire*, CMUc) or having certain diseases (*Affection de longue durée*, ALD) when they cover 100% of the cost [157]. In the first case, the CMUc beneficiaries must have an annual household income under the poverty threshold [158]. In the second case, people being diagnosed with certain chronic diseases (such as diabetes or hypertension or Alzheimer disease), can benefit from ALD full reimbursement after practitioner request and insurance's physician validation [159]. The information available in the DCIR not only refers to healthcare reimbursement but also to sociodemographic characteristics, socioeconomic status and disabilities [160].

Table 6 presents the main public French health insurance regimes and the proportion of French population covered by them. It also shows the year when data were included in the DCIR [161].

Table 6. Main Public French Health Insurance Funds

	French Public Health Insurance Funds	Covered[a]	Year DCIR[b]
RG	Insurance fund for salaried workers (*Régime Général*)	76%	2003
SLM	Insurance fund for students and civil servants (*Section locales mutualiste)*	10%	2003
RSI	Insurance fund for self-employed people (*Régime social des travailleurs indépendants*)	5%	2010
MSA	Insurance fund for agricultural workers (*Mutualité Social Agricole*)	5%	2010
CAMIEG	Insurance fund electricity and gas industries workers (*Caisse d'assurance maladie des industries électriques et gazières*)	< 1%	2010
CNMSS	Insurance fund for military personnel (*Caisse national militaire de sécurité social*)	< 1%	2010
CANSSM	Insurance fund for mineworkers (*Caisse autonome nationale de la sécurité sociale dans le mines*)	< 1%	2015
CRPCEN	Insurance fund for clerks and their employees (*Caisse de retraite et de prévoyance des clercs et employés notaires*)	< 1%	2010
CAVIMAC	Insurance fund for priest and other religious workers (*Caisse d'assurance vieillesse, invalidité et maladie des cultes*)	< 1%	2010
ENIM	Insurance fund for sailors (*Etablissement national des invalides de la marine*)	< 1%	2015
CANSSM	Insurance fund for mineworkers (*Caisse autonome nationale de la sécurité sociale dans le mines*)	< 1%	2015
CCAS RATP	Insurance fund for Paris public transport workers (*Caisse de coordination aux assurance sociales de la RATP*)	< 1%	2015
CPRP SNCF	Insurance fund for French National Railway Company (*Caisse de prévoyance et de retraite du personnel de la SNCF*)	< 1%	2015
CPPAB	Insurance fund for Bordeaux port workers (*Caisse de prévoyance du port autonome de Bordeaux*)	< 1%	2015

a Percentage of the French population covered by the Insurance fund. b Year when the data were included in the DCIR

First, only the data from the Insurance fund for salaried workers (RG) and the Insurance fund for students and civil servant (SLM) were accessible in the DCIR; these Insurance funds cover 86% of the French population [158]. Then in 2010, the data from almost all insurance funds were incorporated including the insurance fund for self-employed people (RSI) and the insurance fund for agricultural workers (MSA). The last group of insurance funds included in the DCIR covers less than 2 % of the population.

3.4.2.2 The PMSI

The collection of data from public and private hospitals (including military hospitals) in the PMSI is coordinated by the Agency for information on hospital care (*Agence technique d'information sur l'hospitalisation*, ATIH) [158]. The PMSI covers data from stays in hospitals, psychiatric institutions or rehabilitation centers and data from

ambulatory care. It started in the late 90's and its data were incorporated into the SNDS in 2005. These data contain information on [162]:

- Patient's characteristics: gender, age and city/town of residence,
- Patient's condition and management: pathologies (primary, related and diagnosis coded with the ICD-10), medical procedures (in CCAM[17] coding) and drugs dispensed (only those with a high cost like cancer treatments),
- Hospital information,
- Hospital stays: month and year of discharge, length of stay, type of admission and discharge (home, referral), or DRG[18]s.

The unique personal identification number enables to link the data from PMSI with other data sources in the SNDS.

3.4.2.3 The CépiDC database

Since 1968, the French national institute of health and medical research (*Institut national de la santé et de la recherche médical*, Inserm) has been responsible for the compilation, the validation and the transmission of the information from death certificates in France [163]. The CépiDC database contains information from all death certificates regarding date, location and cause of a person's death and his/her sociodemographic characteristics [164]. The causes of death are coded using the ICD-10. The CépiDC database is the latest data source included in the SNDS in 2018 when the mortality data from 2013, 2014 and 2015 were linked to the DCIR's and the PMSI's data [156].

The SNDS offers exhaustive and high-quality data on the entire French population, also for those living in the FOT. To update the information requires less time than health surveys and the cost-efficiency ratio is more favorable. However, certain types of information such as lifestyle factors or individuals perception on health status cannot be assessed only with the SNDS data.

3.4.3 Other surveillance sources: the CONSTANCES cohort

The described limitation can be overcome using data sources where behavioral and environmental data are combined with data from the SNDS [165]. This is the case of the CONSTANCES, a general-purpose cohort launched in 2012 and composed by a randomly selected sample of 200,000 adults living in Metropolitan France and aged 18-

[17] CCAM coding of Standard classification of medical procedure (*classification commune des actes médicaux*)

[18] DRG: diagnoses-related group is a patient classification system that standardizes prospective payment to hospitals in groups of diagnosis

69 years at inclusion [166].

CONSTANCES aims to be an epidemiological research tool to study the combined effects of lifestyle, environment, genetic background and other risk factors with the onset of different diseases. Another objective of the cohort is to provide useful information to public health actors to improve the knowledge of the health status of the French population and the utilization of healthcare resources [167]. In this context, the CONSTANCES cohort can become a key data source for diabetes surveillance.

The first step is the sampling of participants among those beneficiaries from RG and SLM insurance funds [167]. A sample of non-participants are included on a "parallel cohort", with health-administrative data from the SNDS and the CNAV[19] are prospectively collected. These data are essential to estimate the coefficients of adjustment for attrition. After inclusion, the participants receive a questionnaire to be completed at home. This self-administered questionnaire is constituted by questions on behaviors, health status, occupational factors and socio-demographic and socio-economic characteristics. The next step is a complete medical examination in one of the 17 Health Screening Centers (HSCs) distributed in different regions in metropolitan France. The medical examination includes a medical questionnaire and physical measurements (weight, height, waist-hip ratio, blood pressure or electrocardiogram). Blood and urine samples are collected to perform laboratory tests such as measurement of blood sugar, liver enzymes or creatinine. Around 50% of these samples are stored in a biobank for further analysis.

Once a year, the data recorded in the self-administered questionnaire and in the medical examination are linked to the data from different health-administrative databases. The CNAV database is used to cross-reference information on socio-demographic characteristics and occupational status. Data from out-of-hospital dispensed health-care reimbursement, ALD-chronic conditions, hospitalizations and vital status are collected from the SNDS.

The follow-up of participants is done through " active" and "passive" active procedures. The "active" follow-up comprises an annual self-administered questionnaire and health examination every 5 years. The "passive" procedure is based on the annual linkage with the health-administrative databases.

[19] CNAV: French national retirement pension fund (*Caisse Nationale d'Assurance Vieillesse*).

3.5 Conclusion

Diabetes surveillance allows developing and evaluating public health programs. It is based on three types of data sources: health surveys, disease/patient registries and health-administrative databases. We have presented different surveillance systems across the World, with special focus on the French diabetes surveillance system. The French system is based on three types of sources: health surveys in general population (ENNS, Esteban), health surveys in diabetic population (Entred) and health-administrative databases (SNDS). The SNDS collects individual information from the entire French population, being one of the largest health-administrative databases in Europe. Using the unique individual identification number, it is possible to cross-reference the databases that composed the SNDS:

a) The DCIR, with information on reimbursement of out-of-hospital dispensed health care from the different French health insurance funds,
b) The PMSI, with discharge data from all public and private hospitals in France,
c) The CépiDC database, with data collected from death certificates.

The SNDS is a valuable source for diabetes surveillance in France, due to its exhaustive and updated information on the entire French population but it also presents some limitations such as the absence of data on lifestyle factors or results of medical test. We can overcome the limitations of the SNDS through a new source of data, the CONSTANCES cohort. Indeed, in this cohort, longitudinal data from 200,000 participants are recorded in self-reported questionnaires and in medical examinations (including biological test) were linked to their SNDS data.

OBJECTIVES

This chapter aims to expose the challenges for diabetes surveillance, regarding the limitations of the current tools for the exploitation of the SNDS data in France. The objective of this thesis is to tackle them.

1. Tools for diabetes surveillance in France

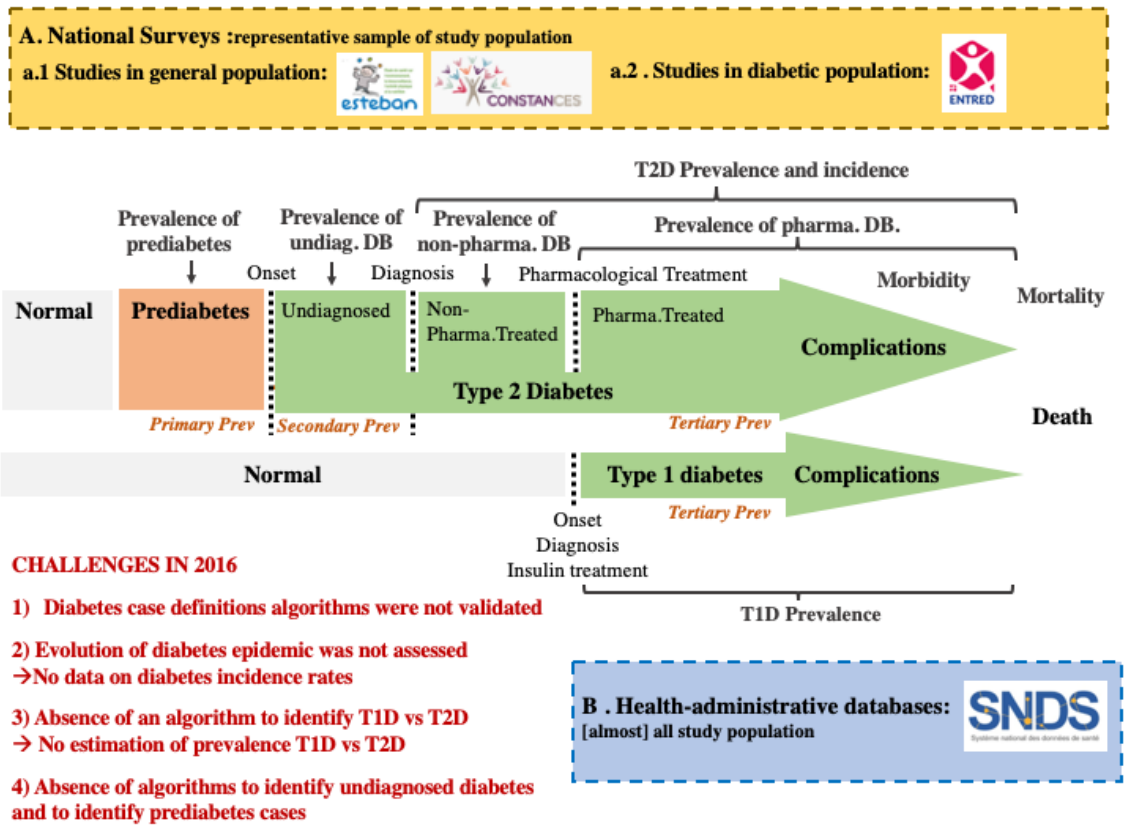
We have seen that diabetes represents an important burden and how surveillance is essential to develop and to evaluate diabetes prevention programs. One of the most important data sources of diabetes surveillance in France is the SNDS. The ReDSiam [20] network includes users of the SNDS and aims to construct and to improve the methodology for the exploitation of the French health-administrative databases [169]. The ReDSiam working group on endocrine, nutritional and metabolic diseases has recorded three algorithms to identify diabetes cases in the SNDS [170]:

- Algorithm A: positive case if the individual benefits during the given year from an ALD with a ICD-10 code of diabetes (E-10 or E14),
- Algorithm B: positive case if the individual has a reimbursement of an antidiabetic drug (class A10 from ATC code - except Benfluorex-) on at least three different dates in a given year or two dates if at least one large package of antidiabetic drugs was dispensed,
- Algorithm C: positive case if the individual meets at least one of following conditions: (a) is registered as having ALD-Diabetes during the given year; (b) is reimbursed for an antidiabetic drug on at least three different dates in the previous 2 years, or on two dates if at least one large package of antidiabetic drugs was dispensed; (c) was hospitalized with a principal or related diagnosis of diabetes (E10–E14) or with a principal or related diagnosis of a diabetes-related complication (G59.0*,G63.2*, G73.0*, G99.0*, H28.0*, H36.0*, I79.2*, L97, M14.2*, M14.6*, N08.3) and an associated diagnosis of diabetes (E10-E14) in the previous 2 years.

[20] REDSIAM : Network for the improvement of exploitation of the SNDS, *Résau pour mieux utilser les données du SNDS*

2. Challenges for diabetes surveillance in France using the SNDS

The previous diabetes case definition algorithms had been widely used in diabetes research but their application for diabetes surveillance faced some challenges in 2016 when we began our thesis (**Figure 27**).



T1D: Type 1 diabetes; T2D: Type 2 diabetes

Figure 27. Challenges on diabetes surveillance based on the SNDS in 2016

First, the performances of the different algorithms in identifying diabetes cases had not been validated.

Additionally, although algorithm B had been used to assess diabetes prevalence in France, there was no study on the evolution of the diabetes epidemic due to the absence of data on incidence rates.

None of the algorithms were capable of differentiating between type 1 and type 2 diabetes hence there was a lack of information on the prevalence of type 1 and type 2 diabetes separately among adults in France.

Finally, there were no algorithms to ascertain undiagnosed diabetes and prediabetes based on SNDS data.

3. Objectives of the thesis

The principal objective of this thesis was to address the challenges of diabetes surveillance based on health administrative databases which have been described in the previous section. These objectives were the following :

a) To improve the classical surveillance tools based on the SNDS data by (i) the validation of the diabetes case definition algorithms registered by the ReDSiam network and (ii) the application of the most suitable algorithm to study the evolution of the diabetes epidemic in France, assessing the prevalence and the incidence of diabetes among adults aged 45 years or higher between 2010 and 2017

b) To develop new tools for diabetes surveillance through two steps: (i) the development of a type 1/type 2 classification algorithm using a generic method based on Machine Learning and (ii) to use the previous generic method to develop an algorithm to identify undiagnosed diabetes cases and prediabetes cases using the SNDS data.

MATERIALS

1. The SNDS

We have already introduced the SNDS in section 3 of the introduction. More detailed information on the access, the data sources and the structure of the SNDS will be developed in this section [171].

1.1 SNDS Data warehouse: Access requirements

CNAMTS manages the access to the SNDS, giving different profiles depending on the user's institution and the project proposed. Our profile included access to PMSI data since 2009 and to DCIR data since 2006. We also had access to sensible variables or those allowing an individual indirect re-identification: city/town of residence, date of birth (month and year), date of death (day, month and year) and date of delivered health care (day, month and year). Though, cross-referencing these variables for indirect re-identification was completely forbidden.

1.2 SNDS Data warehouse: Data collection

In **Figure 28,** the data sources of the SNDS are represented, as well as its structure: the DCIR, the PMSI and the CépiDC database.

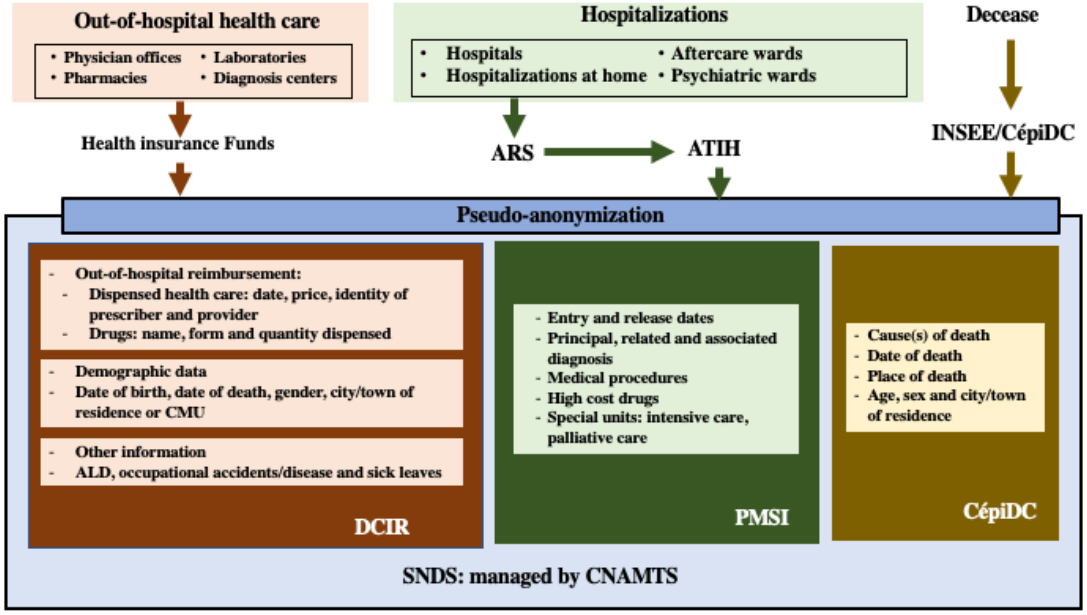


Figure 28. The SNDS: data sources and structure

The DCIR information on out-of-hospital healthcare dispensed in physician offices, laboratories, diagnosis centers and pharmacies is recorded by each health insurance fund and it is submitted once a month to the CNAMTS. However, as we have seen in the section 3 of the introduction (**See page 64**), the inclusion of the different health insurance funds has been done gradually since 2006, and the information from certain funds might

be not accessible depending on the year of study. Moreover, certain information from some funds might be unreliable during the first years after its input.

The hospitalization data are recorded in public and private hospitals, after care and rehabilitation wards, hospitalizations at home and psychiatric wards. They are submitted every six weeks to the Regional Health Agency (*Agence regional de santé*, ARS) concerned where data are collected and validated. A second validation is done by the ATIH before injecting the data in the SNDS. The validation process lasts at least two months.

The collection, the validation and the input in the SNDS of the information recorded in the death certificates is on charge of the Inserm unit CépiDC. In 2018, the first package of information from CépiDC was included in the SNDS with data from 2013, 2014 and 2015.

1.3 SNDS data warehouse: Data structure

This work is based on the DCIR and the PMSI since the CépiDC database was not accessible in the SNDS when the thesis work started. We will explain in detail how data are structured in these two datamarts, the DCIR and the PMSI.

1.3.1 The DCIR datamart

The DCIR datamart has a complex structure based on different data tables that can be linked through a combination of 9 variables. The central data table is ER_PRS_F where each line corresponds to a reimbursement of dispensed healthcare. Depending on the type of healthcare dispensed, this data table has to be linked to other data tables to complete the information. For example, if the reimbursement is associated with a medical act, we should linked the table ER_PRS_F with the table ER_CAM_F. The reimbursement of medical devices such as self-monitoring glucose test kit or pens for insulin injection are coded in the table ER_TIP_F. The codes for laboratory tests like HbA1c or glucose tests can be found in the table ER_BIO_F. To complete the information on drug reimbursements, the table ER_PHA_F has to be linked with the table ER_PRS_F so the ATC code of the dispensed drug, as well as the quantity and the form can be ascertained.

The IR_BEN_R table has to be linked with the table ER_PRS_F through two variables in order to collect information on the reimbursements' beneficiary. This table contains individual data on sex, city/town of residence, health insurance scheme associated, date of birth or date of death. Date of death is not reliable for certain health insurance schemes like RSI so unfortunately, it cannot be used to assess mortality rates.

In this table, there is also a unique anonymized identification number allowing to aggregate different reimbursements from the same individual and to cross-reference information with the PMSI and the CepiDC data marts. The deceased individuals or those without a reimbursement in the last four years are moved to the table IR_BEN_ARCH.

1.3.2 The PMSI datamart

The PMSI is composed by four main data files: PMSI-MCO (data from medicine, surgery, obstetrics and odontology wards), PMSI-SSR (data from aftercare and rehabilitations wards), PMSI-HAD (data from hospitalizations at home) and PMSI-PSY (data from psychiatric wards). In our study, we use the information from the first data file, PMSI-MCO. The data on the datafile are structured in different databases which can be linked using the variables ETA_NUM and RSA_NUM, corresponding to the legal number of the institution and the number of the anonymized hospital discharge summary. The MCO_B is the central table where each line refers to one hospital stay; in this table, we can find information on the patient (age, sex and city/town of residence) and on the principal and related diagnosis. The associated diagnosis are in the table MCO_D. Specific information on the hospital stay related to the medical acts performed, the stays in special units or the establishment can be found in the data tables MCO_A, MCO_UM and MCO_E. The admission and discharge dates are in the table MCO_C where there are also the individual unique anonymized identification numbers for linking data with the DCIR datamart.

2. The CONSTANCES cohort

In the introduction, we explained why the **CONSTANCES** cohort represents an excellent opportunity for overcoming the limitations of diabetes surveillance based on the SNDS data. This section aims to show in detail what kind of data are recorded in the CONSTANCES cohort and how they are collected .

2.1 The CONSTANCES cohort's protocol

The CONSTANCES cohort attained the milestone of 200,000 participants in 2019. These figures could not be achieved without an exhaustive protocol comprising the interoperability of different stakeholders: the CONSTANCES unit, the CNAV, the CNAMTS and the HSCs (**Figure 29**).

2.1.1 Sampling

The CONSTANCES unit sends the list of the partner HSCs and the sampling rates to obtain a representative sample of the RG and SLM beneficiaries aged from 18 to 70 years. The sampling rates are estimated on the basis of three criteria: age, sex and socioeconomic status. They are used by the CNAV to oversampling those groups less likely to participate in the cohort studies such as deprived population.

The list of selected individuals is sent to the CNAMTS since it is the only stakeholder with access to the complete personal address, compulsory for sending the invitation mail. This information, together with individual data on sex, birth and place of residence and associated health insurance fund are received by the CONSTANCES unit.

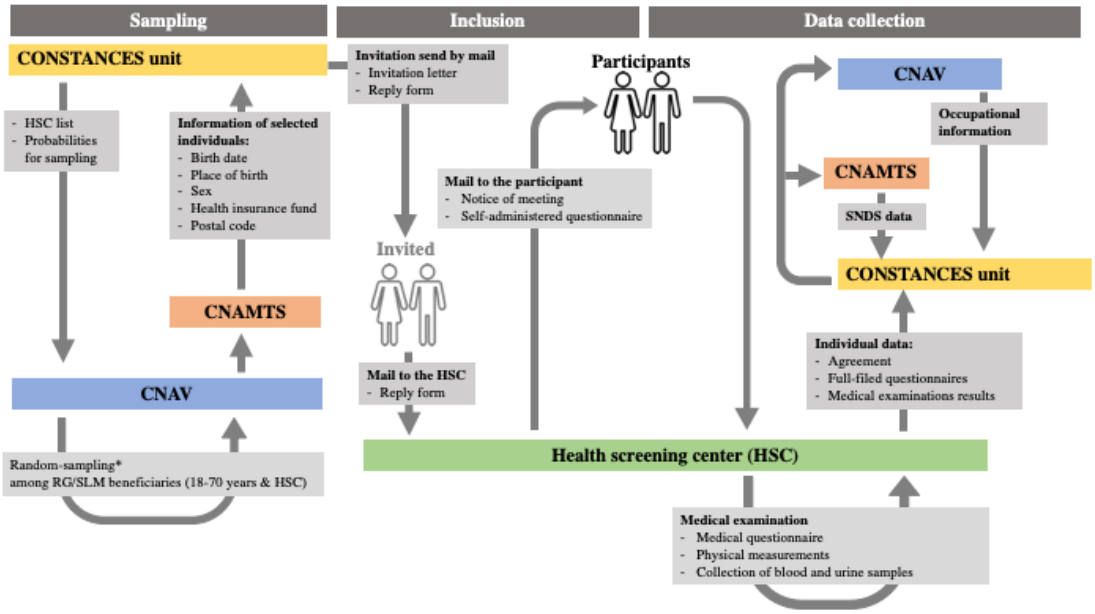


Figure 29. The CONSTANCES cohort protocol

2.1.2 Inclusion

Based on the information submitted by the CNAMTS, the invitation is sent to each selected individual comprising an invitation letter and a reply form. Those willing to participate must send the reply form to the reference HSC while those willing not to participate should follow the same procedure indicating their refusal to participate. In the absence of an answer, the individuals could be included in the non-participants cohort consisting in a passive follow up through their SNDS and CNAV data which is used by the CONSTANCE cohort for the coefficients of adjustment for attrition.

Once the agreement is received in the HSC, a convocation letter with the medical examination requirements and a questionnaire is sent to the participants.

2.1.3 Data collection

The participants must fulfill two self-administered questionnaires, one on health

and life-style factors and another on occupational history. They can be web-based or paper-based questionnaires.

The participant can completea the paper questionnaire at home and give it back the day of the medical examination. Other questionnaires are completed this day in the HSC: a medical questionnaire, a questionnaire on occupational exposures and also a questionnaire on women's health and a cognitive test, which are dispensed depending on the sex and the age of the participants. Skilled medical staff from the HSC is on charge of the paraclinical examination including the collection of biological samples. Finally, the participant must complete an informed agreement on the utilization of the previous collected data and other agreement for the access to their health administrative data on the SNDS and the CNAV data.

The agreements, the questionnaires and the results of the medical examinations are submitted from the HSCs to the CONSTANCES unit on charge of validating, collecting and storing the data. The list of the individuals who agree to give access to their data from health administrative databases is received by the CNAMTS and the CNAV, which then provide the SNDS data and the occupational data to the CONSTANCES team.

2.1.4 Follow up

Each year, the participants of the CONSTANCES cohort receive a short questionnaire in order to update personal information related to their health, socioeconomic and occupational status. They also attend to their reference HSC for a medical examination every five years .

A passive follow-up of the participant is performed through the information from the health-administrative databases. In the first quarter of the year, the participants' SNDS and occupational data are submitted from the CNAMTS and from the CNAV to the CONSTANCES unit.

2.2 Data in the CONSTANCES cohort

A great variety of data are collected for the CONSTANCES cohort. They are recorded through self-administered questionnaires, medical examination and HAD.

2.2.1 Self-administered questionnaires

Each participant of the CONSTANCES cohort completes different questionnaires. The longest is on life style characteristics and health status. This questionnaire includes sections on physical activity, nutrition, alcohol and tobacco use, disabilities, socioeconomic status and health status. In the health status section, there are various questions related to diabetes diagnosis, treatment, glycemic control and medical

examinations.

There are also questionnaires on occupational history and on present and past occupational exposure such as chemical products, noise or extreme temperatures. Women fulfill a specific questionnaire with questions on pregnancies, menstruation or contraceptive use. Another specific questionnaire is given to people aged over 45 years to assess their cognitive function.

2.2.2 Medical examination

The medical examination is performed in the HSC by health professionals and it comprises a questionnaire on participants medical history, a physical examination and a blood test.

The information on personal and family medical history is recorded through a questionnaire which includes four questions on the diagnosis of type 1 and type 2 diabetes and the age of diabetes onset.

The physical examination is composed of different tests like hearing and vision tests, blood pressure test, electrocardiogram or spirometry test. Height, weight and waist and hip circumference are measured to evaluate corpulence indicators such as BMI or waist-to-hip ratio.

Finally, there is a collection of biological samples. Participants must have fasted for at least 8 hours before the medical examination so FPG can be measured in the blood sample. Together with FPG, complete blood count and creatinine, cholesterol, triglycerides levels are also assessed in the laboratory using the collected samples.

2.2.3 The SNDS and the CNAV data

At the beginning of the year, the SNDS data and the CNAV data are received by the CONSTANCES team, which validates and integrates the information to the CONSTANCES cohort database. This information is accessible to the researchers in different ways. In may 2017, the CONSTANCES team built a hub with the “raw” SNDS data of the participants recruited between 2012 and 2015 in order to tackle the specific needs of our study. The SNDS data of CONSTANCES participants were structured in data tables linked between them through different variables as we have previously described in the SNDS section.

BASELINE METHOD

1. Baseline method

The baseline method of this thesis is represented in **Figure 30**. It was composed of a central core step comprising the definition of the references of diabetes stages followed by four steps, each of them associated to one objective of the thesis: the validation of diabetes case definition algorithms, the study of the evolution of the diabetes epidemic in France, the development of a type 1 / type 2 classification algorithm and the development of an algorithm to identify undiagnosed diabetes cases and an algorithm to identify prediabetes cases.

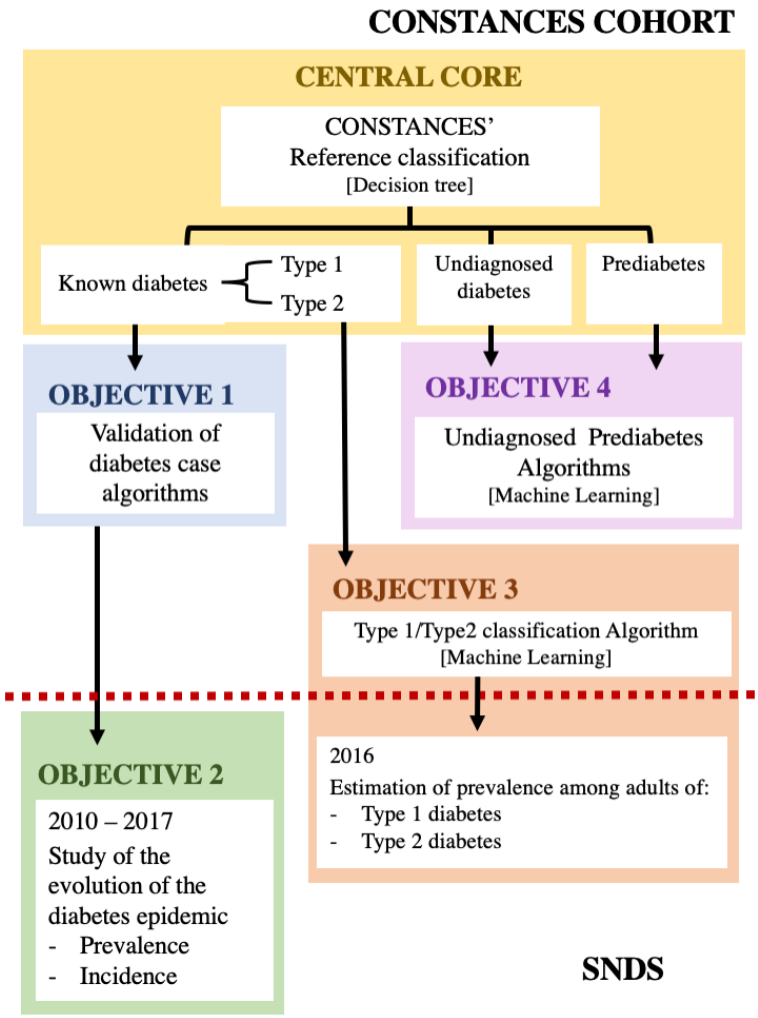


Figure 30. The baseline method of the thesis

The central core step of the baseline method was the definition of the reference classification categories. The participants recruited by the CONSTANCES cohort between 2012 and 2014 were classified in the following groups: known diabetes, undiagnosed diabetes, prediabetes and non-diabetes. Among the known diabetes cases, we performed two subclassifications: pharmacologically treated *vs*. non-

pharmacologically treated cases and type 1 *vs*. type 2 diabetes cases. Also, the WHO and then the ADA criteria were used to define prediabetes cases. The classification was done through different decision trees based on the information recorded in the self-reported questionnaire, in the medical questionnaire and in the laboratory test, more specifically the FPG test.

The classification defined in the central core was the gold standard to assess the performances of the three diabetes case definition algorithms identified by REDSIAM working group. Two different gold-standards were evaluated: known diabetes cases and pharmacologically treated diabetes cases. These performances were also reported by sex and age group. Besides, each element of the algorithms was evaluated with the two gold standards.

Based on the results of the previous stage, we selected the most suitable case definition algorithm to assess the evolution of the diabetes epidemic in France between 2010 and 2017. Applying this algorithm to the entire SNDS, we constructed a retrospective cohort with all diabetes cases in France during the study period in order to characterize prevalent and incident cases. Then, prevalence and incidence rates were estimated for each year by sex, age and region. Since the algorithms were not able to differentiate between type 1 and type 2 diabetes cases, we restricted the analysis to the population aged over 45 years in order to focus on type 2 diabetes. Finally, the evolution of prevalence and incidence along the study period was estimated through negative binomial models.

To overcome the limitation of the diabetes case definition algorithms in differentiating between type 1 and type 2 diabetes, we developed a type 1 / type 2 classification algorithm using a Supervised Machine Learning (SML) method. All the pharmacologically treated diabetes cases in the CONSTANCES cohort constituted the final data set for developing the classification algorithm. The type 1 and type 2 diabetes cases were already characterized in the central core through the Entred decision tree based on the age of diagnosis and the delay between the diagnosis and the onset of insulin treatment reported in the self-administered questionnaire and the medical questionnaire. This characterization of type 1 and type 2 diabetes cases was used as reference target in the SML. Then, almost the whole SNDS data of these type 1 and type 2 diabetes cases were coded into variables for being selected as components of the selected algorithm. The classification algorithm was applied to the entire SNDS data from 2016 to estimate the prevalence of type 1 and type 2 diabetes among adults aged between 18 and 70 years and

living in France, by sex and age.

The last stage of the methodology was the development of an algorithm to identify undiagnosed diabetes cases and an algorithm to identify prediabetes cases in the SNDS. We use the same SML method described for the type 1/type 2 classification algorithm. The final dataset for both algorithms was selected from the CONSTANCES population, and the target reference used where the undiagnosed diabetes cases and prediabetes cases characterized in the central core stage of the baseline methodology.

2. The central core

The central core was an essential step in the baseline method of this thesis. After selecting the CONSTANCES population among the individuals recruited by the cohort between 2012 and 2014, it was classified in different diabetes stages through three successive classification trees. This classification was used as gold standard for the validation of the three diabetes case definition algorithms and as reference categories in the SLM methodology for developing the type 1 / type 2 classification algorithms, the algorithm to identify undiagnosed diabetes cases and the algorithm to identify prediabetes cases.

2.1 The CONSTANCES population

In 2016, we had access to data from 81,997 participants of the CONSTANCES cohort but certain groups were excluded of the final population used in the thesis, henceforth named “the CONSTANCES population” (**Figure 31**).

The data of the participants recruited in 2015 were not yet linked to the SNDS. These participants were excluded because we needed the SNDS data from at least one year after the date of the self-administered questionnaire in order to confirm the undiagnosed diabetes cases. Also, due to issues related to the interpretation of the question on gestational diabetes mellitus in the two first versions of the self-administered questionnaire, we decided to exclude the women who declared already being diagnosed of gestational diabetes mellitus or those who declared being pregnant.

The last group of participants excluded from the CONSTANCES population were those without data linked to the SNDS or those without data from the self-administered questionnaire and from the medical questionnaire.

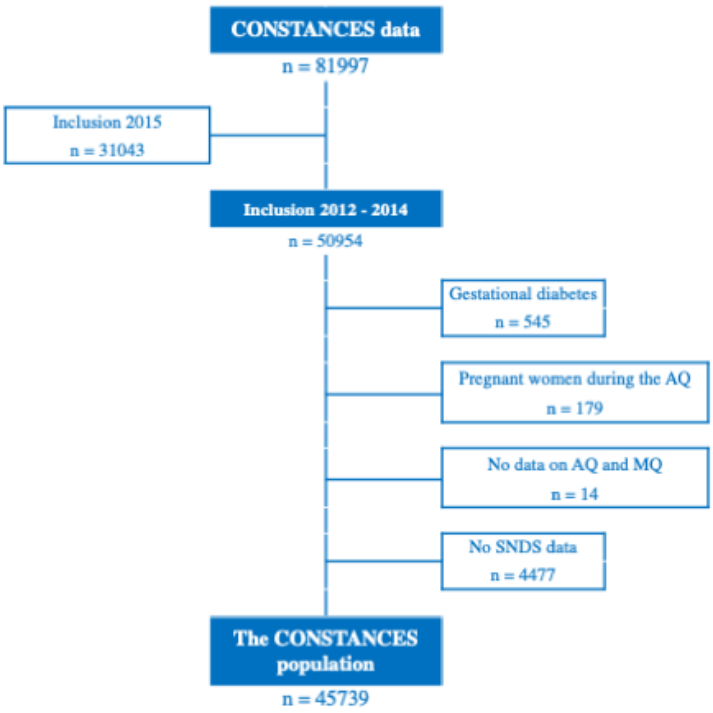


Figure 31. Flow chart of the selection of the CONSTANCES population

2.2 Stage 1: First decision tree

Once the CONSTANCES population was selected, all its components were classified in different diabetes stages through three decision trees. The first tree was an initial stage for classifying individuals in three categories: diabetes, non-diabetes and inconsistent. These categories were not definitive and they were reviewed through a second decision tree.

The first decision tree was based on the information recorded in the self-administered questionnaire and in the medical questionnaire (**Figure 32**). Two main variables were coded from the former questionnaire: "declared diabetes diagnosis" and "declared diabetes follow-up". First, based on the answer to the question: "Have you ever been told by a physician or a healthcare professional that you have diabetes? *(Est-ce qu'un médecin ou un professionnel de santé vous a déjà dit que vous étiez atteint(e) de diabète* ?)", three values were defined for the variable "declared diabetes diagnosis"(yes, no and missing values). After, the variable "declared diabetes follow-up" was determined as positive if the participant had at least answered yes to one of the following questions:

- "Do you regularly visit a physician for monitoring your diabetes? (*Consultez-vous régulièrement un médecin pour le suivi de votre diabète* ?) »

- «Currently, are you treated for diabetes with oral medication? (*Actuellement*

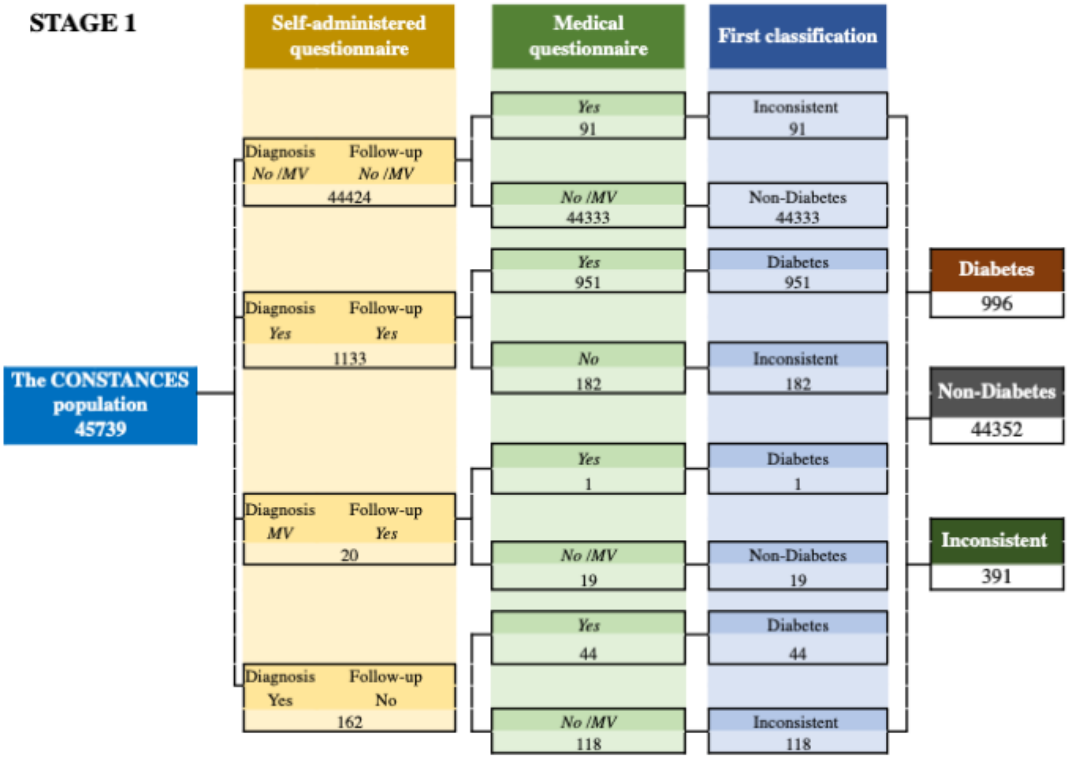
êtes-vous traité(e) pour votre diabète par des comprimés ?)"

- "Currently, are you treated for diabetes with one or more insulin injections? (*Actuellement, êtes-vous traité(e) pour votre diabète par une ou plusieurs injections d'insuline* ?)"
- « Have you ever had a screening test of glycated hemoglobin (HbA1c)? (*Avez-vous déjà eu un dosage d'hémoglobine glyquée* (HbA1c)?) »

During the medical examination, the physician demands to the participants if they have a type 1 diabetes or a type 2 diabetes. In our validation sample, 1087 participants were identified as having a type1 diabetes or a type 2 diabetes or both from the medical questionnaire. This group of participants was categorized as "yes" in the constructed variable "diabetes medical questionnaire".

In some cases, the information observed among these variables, "declared diabetes diagnosis", "declared diabetes follow-up" and "diabetes medical questionnaire" was not consistent. Some participants were classified as "yes" in the last variable but they declared neither a diabetes diagnose nor a monitored diabetes in the self-administered questionnaire. Also, inconsistencies in the opposite sense were observed: participants declaring to be diagnosed of diabetes with a monitoring, but not identified as type 1 or type 2 diabetes in the medical questionnaire. Both cases were categorized as "Inconsistent" at the end of stage 1.

Two more categories were defined at this stage: "Diabetes" and "Non-diabetes". In the "Diabetes" category were included not only those participants positively consistent with the three variables, but also those declaring living with diabetes but not having been followed up for diabetes and those with a missing value for diabetes diagnosis but declaring any diabetes follow-up (treatment, medical consultations for diabetes or HbA1c test), as far as they were categorized as having diabetes in the medical questionnaire. However, when he/she was categorized as negative and the variable "declared diabetes diagnosis" from the self-administered questionnaire was a missing value, even if they had declared to have any kind of diabetes monitoring, the participant was categorized as "Non-diabetes" at this stage. When the participant had only answered "yes" to the question on diabetes diagnosis but he/she neither declared having been followed up for diabetes nor was classified as positive in the medical questionnaire, his/her category at stage 1 was " Inconsistent". The rest of the cases were added to the "Non-diabetes" group.



MV: Missing value

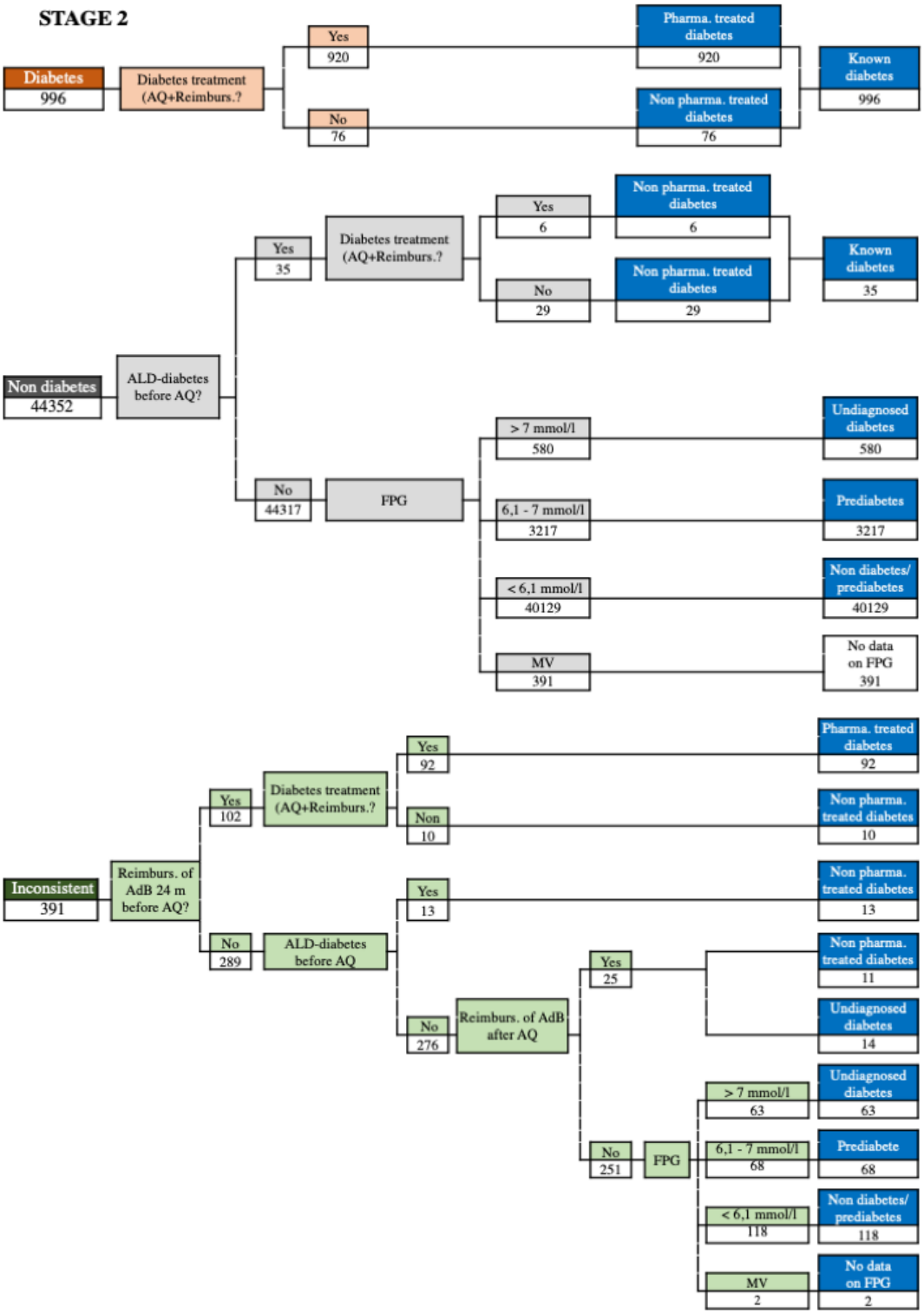
Figure 32. First decision tree of the central core step

After applying the decision tree, the CONSTANCES population was divided into three groups: “Diabetes” (996 participants), “Non diabetes” (44352 participants) and “Inconsistent” (391 participants).

2.3 Stage 2: Second decision tree

The second decision tree of the central core step comprised three different branches, corresponding to the three groups classified in the previous stage (**Figure 33**).

All the individuals in the branch “Diabetes” were characterized as having “Known diabetes”. They were divided into “Pharmacologically treated diabetes” and “Non-pharmacologically treated diabetes” depending on their answers to the two questions from the self-administered questionnaire on antidiabetic treatment (oral agents or insulin). However, the declared information was further confirmed with the data on drug reimbursements; those who declared not taking antidiabetic treatment and the missing values with at least one reimbursement of antidiabetic drug in the previous 200 days were reclassified as “Pharmacologically treated”.



AdB: Antidiabetic drugs; AQ: self-administered questionnaire; MV: Missing value

Figure 33. Second decision tree of the central core step

Another branch was for the individuals previously categorized as "Non-diabetes". We wanted to confirm their status by cross-referencing the information on ALD-diabetes. When a participant was beneficiary of an ALD-diabetes, she/he was reclassified as "Known diabetes", and it was also characterized as "Pharmacologically treated diabetes"

or “Non-pharmacologically treated diabetes” based on the diabetes treatment variable explained in the previous branch. Those not benefiting of an ALD-diabetes were divided into four categories based on their FPG: “ Undiagnosed diabetes” (FPG > 7 mmol/l), “Prediabetes” (FPG from 6.1 to 7.0 mmol/l, WHO definition), “Non diabetes/prediabetes” (FPG < 6.1 mmol/l, WHO definition) and “No data on FPG” (when FPG was a missing value).

The last branch corresponded to the classification of the “Inconsistent” group. Those who had at least one reimbursement of antidiabetic drug during the 24 months before the self-administered questionnaire were included in the category of “Known diabetes”, and they were also stratified into “Pharmacologically treated diabetes” or “Non-pharmacologically treated diabetes” using the variable of diabetes treatment previously exposed.

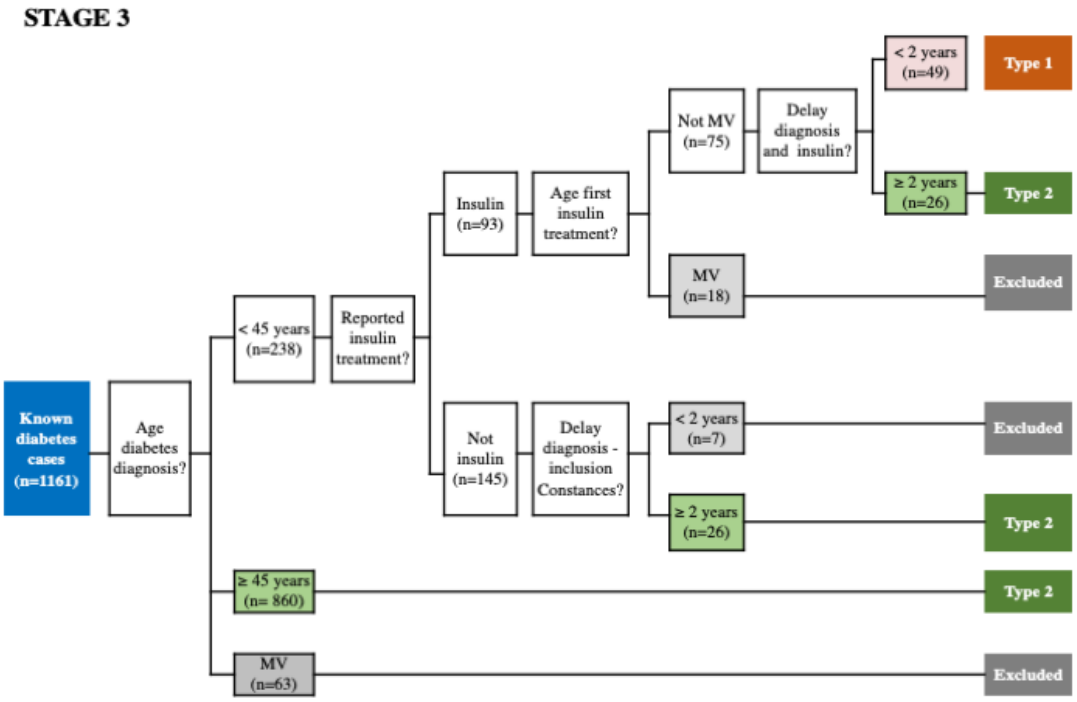
For those not having any reimbursement of antidiabetic drugs but benefiting from ALD-diabetes, their category was “Non-pharmacologically treated diabetes”. If the “Inconsistent” individual had neither reimbursement in the last 24 months nor ALD-diabetes, but she/he had at least one reimbursement of antidiabetic drugs 12 months after the self-questionnaire, the individual was classified as “Undiagnosed diabetes” or “Non-pharmacologically treated diabetes” depending on their answers to the diabetes section of the self-administered questionnaire and the results of the medical questionnaire. On the basis of the FPG, the rest of the individuals were classified into four categories as we have previously described in the “Non-diabetes” branch.

2.4 Stage 3: Entred classification tree

The last stage was the classification of the “Known diabetes” group into type 1 diabetes and type 2 diabetes (**Figure 34**). For this purpose, we applied the Entred decision tree [152] based on three items: age at diabetes diagnosis, current insulin treatment, and the delay between diabetes diagnosis and first insulin treatment. This information was collected in the self-administered questionnaire and validated with the data from the medical questionnaire and from the reimbursement of antidiabetic drugs.

The diabetes case was classified as type 2 diabetes if the age of diagnosis was 45 years old or more. Individuals diagnosed with diabetes before 45 years, not currently having an insulin treatment and with a delay between diabetes diagnosis and inclusion in the CONSTANCES cohort of 2 years or higher were classified as type 2 diabetes. The last group of type 2 diabetes were the cases diagnosed with diabetes before 45 years, currently having an insulin treatment and with a delay between diabetes diagnosis and

first insulin treatment of 2 years or higher. Those with the same characteristics as the last group but with a delay between diabetes diagnosis and first insulin treatment of less than 2 year were characterized as having type 1 diabetes.



MV: missing value

Figure 34. Application of the Entred decision tree

3. CONSTANCES' reference classification

3.1 The CONSTANCES population characteristics

After excluding the pregnant women, the women who declared having been diagnosed with gestational diabetes mellitus and those individuals without SNDS data, the CONSTANCES population comprised 45,739 participants. Their main characteristics are shown in **Table 7**.

The mean age was 49 years old and the proportion of women was higher than the one of men. Almost half of the individuals declared not being smokers, their mean BMI was 25 kg/m^2 and 13% of them were treated for hypertension and 10% for dyslipidemia. Most of the individuals had a French or an European origin, and around 3% of the population had an North-African origin. Since the participants of the CONSTANCES cohort were recruited from the beneficiaries of the RG and the SLM, the population was mainly composed of employed or retired individuals; there was also 6.2 % of unemployed and only 1.8 % of students. Regarding education, at least half of the population had at

least tertiary education level.

Table 7. Characteristics of the CONSTANCES population

	CONSTANCES population
Age (mean, ±sd)	49.1 ±13.2
Gender (men, %)	47.4
Current smoking status (%)	
Never smoked	45.3
Former smoker	19.5
Current smoker	35.1
Body mass index, kg/m^2 (mean, ±sd)	25.1 ±4.5
Treated hypertension (yes, %)	13.2
Treated dyslipidemia (yes, %)	10.6
Mother/father diagnosed with diabetes (yes,%)	15.7
Socioeconomic status	
Education [i] (%)	
No education - primary education	3.2
Lower secondary education	7.1
Upper secondary education	34.3
Lower tertiary education	33.3
Upper tertiary education	21.9
Geographical origin (%)	
Metropolitan France	89.0
FOT[ii]	0.9
Europe	4.2
North Africa	2.9
Sub-Saharian Africa	1.2
Asia	0.8
Others	1.0
Professional activity (%)	
Employed	65.1
Unemployed	6.2
Retired	23.4
Student	1.8
Unemployed due to disability	1.6
No professional activity	1.9

sd: standard deviation
i Based on the International Classification ISCED
ii French overseas territories

3.2 Reference classification

The reference classification obtained after applying of the three decision trees is represented in **Figure 35**. The CONSTANCES population comprised 45,739 individuals. Most of them (88.8%) were classified as “Non-diabetes/prediabetes”. The percentage of undiagnosed diabetes cases and prediabetes (using the WHO criteria) was 1.4% and 7.2% respectively. Only 2.4% of the population was classified as `Known diabetes”. Among them, the percentage of “Non-pharmacologically treated diabetes” was 12%. Through the Entred decision tree, a total of 49 type 1 diabetes cases and 1024 type 2 diabetes cases were identified (corresponding to 4.6% and 95.4% of all known diabetes cases respectively).

As we will see, this classification was essential in further steps of the baseline method. It was used as gold-standard for the validation of the diabetes case definition algorithms. Then, the different categories were applied for the definition of the target 1 and target 0 when applying SML methods for developing the type 1 / type 2 classification algorithm, and finally the algorithms to identify undiagnosed and prediabetes cases in the SNDS.

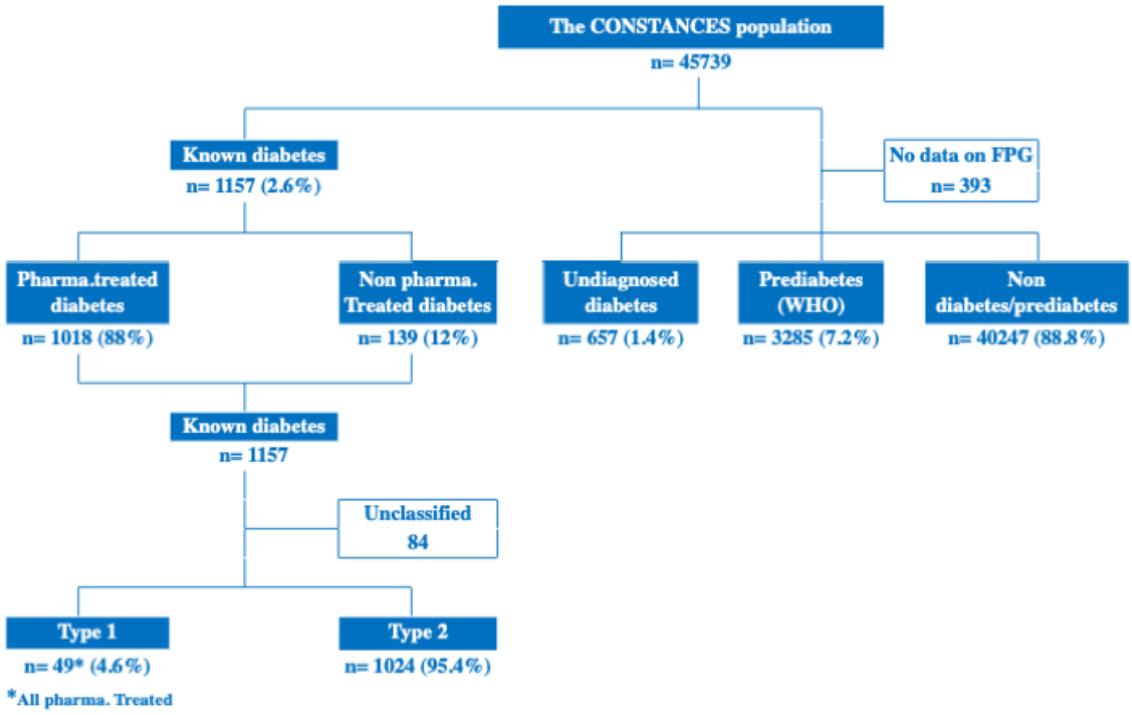


Figure 35. Reference classification

RESULTS

1. Validation of diabetes case definition algorithms

1.1 Introduction

One of the crucial challenges for diabetes surveillance based on the SNDS data was the validation of the case definition algorithms used to identify diabetes cases (**Figure 36**).

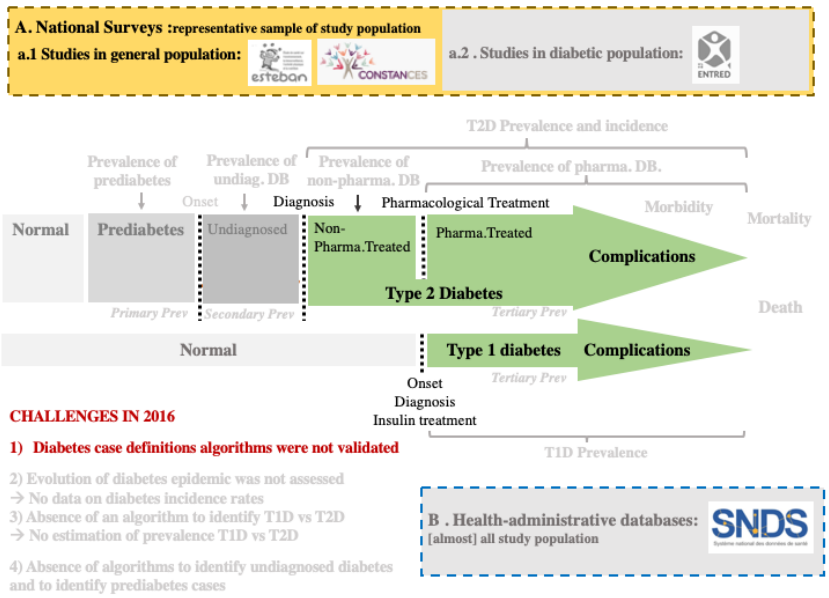


Figure 36. Challenges faced in the results' section 1

Validation studies of case definition algorithms from other countries have used different sources to define their gold standard reference, such as primary care medical charts or patients registries [172]. Since the CONSTANCES cohort links individual data from self-administered questionnaire and from medical examinations with their SNDS data, it represented an excellent source for the validation of the algorithms.

1.2 Objectives

The objective was to assess the test performance of the three diabetes case definition algorithms introduced above (**See page 72**) in identifying both "known diabetes" and "pharmacologically-treated diabetes".

1.3 Methods

Figure 37 represents the method of this section in the context of the thesis' baseline method. Previously in the central core step of the baseline method, we defined the references for "known diabetes" and "pharmacologically treated diabetes" cases through a complex decision tree based on data from the self-administered questionnaire and from the medical examination. This reference classification was used as the gold-standard.

Then, using the SNDS data from the participants of the CONSTANCES cohort, we applied the algorithms recorded by ReDSiam for identifying diabetes cases (**See page 72**). Then, we cross-referenced the results with the gold-standard in order to evaluate the test performances of each algorithm.

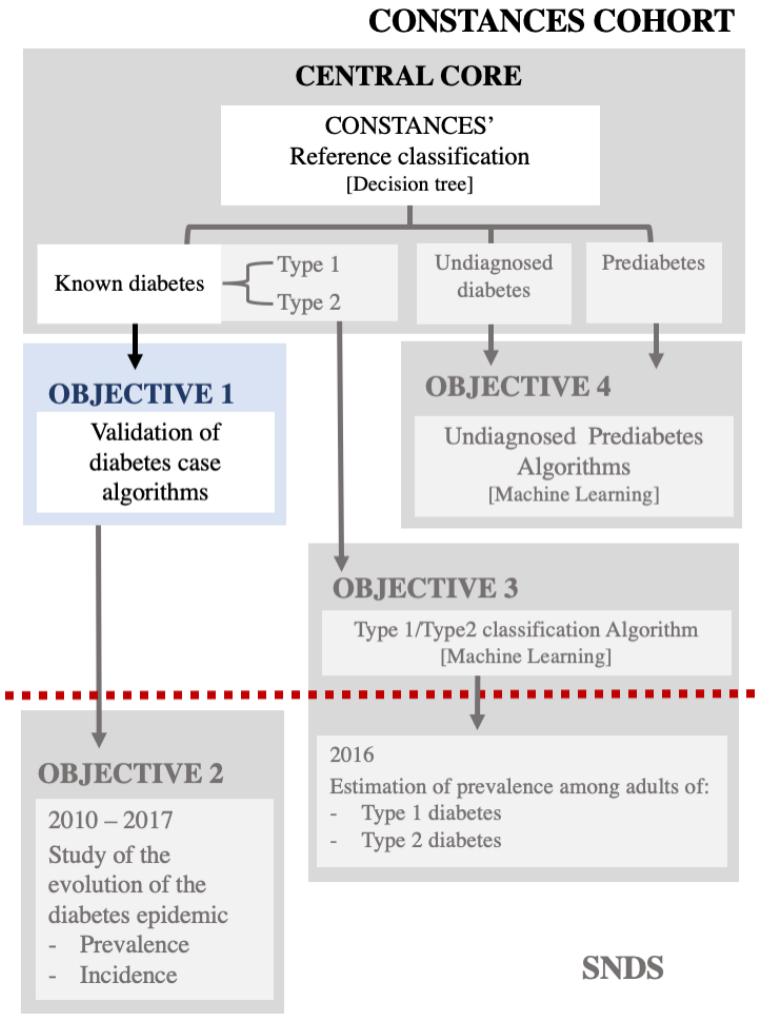


Figure 37. Methods of the results' section 1

Seven test characteristics were assessed : sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy, kappa and F1 score (**Figure 38**).

The analyses were done in all the study population and then stratified by sex and age group (from 18 to 29 years, from 30 to 54 year and 55 years or more). These stratified analyses were relevant because if test performances presented significant differences depending on sex or age, the algorithm would not being suitable for diabetes surveillance.

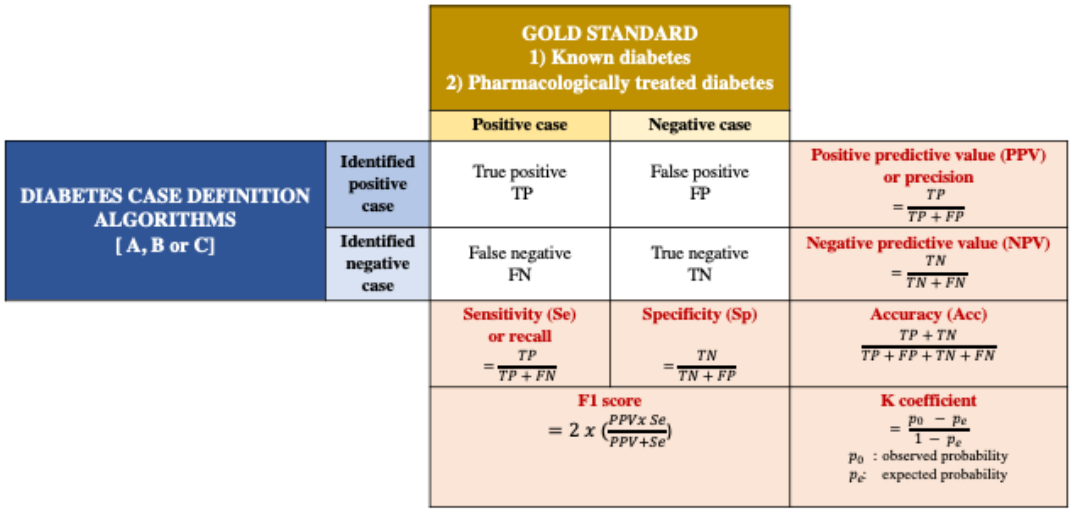


Figure 38. Test performances assessed for the validation of the diabetes case definition algorithms

Finally, the performances of each component of algorithm C were also studied separately as case definition:

- Component 1 : positive if the individual is beneficiary of an ALD-diabetes,
- Component 2: positive case if the individual has a reimbursement of an antidiabetic drug on at least three different dates in the two previous years or two dates if at least one large package of antidiabetic drugs was dispensed,
- Component 3: positive case if the individual was hospitalized with a principal or related diagnosis of diabetes or with a principal or related diagnosis of a diabetes-related complication and an associated diagnosis of diabetes in the previous 2 years.

1.4 Results

1.4.1 Gold standard “known diabetes”

We describe in

Table 8 the test characteristics of the diabetes case definition algorithms recorded by ReDSiam. The algorithm with the highest sensitivity was algorithm C (93.8%), followed by algorithm B and then algorithm A (85.8% and 73.7%). Each algorithm presented very high specificity, reaching 100% in algorithm A, due to the absence of false positives. The NPVs and the PPVs were also high; again, algorithm A had a PPV of 100% because no false positives were found. The accuracy, the kappa coefficient and the F1 scores provides an overall estimation of the performances of each algorithm. Based on these values, the algorithm with the highest performance in identifying “known diabetes” was algorithm C. Algorithm B had also good performances in identifying “known

diabetes" cases.

Table 8. Test characteristics of three diabetes case definition algorithms using known diabetes as the gold standard

	Algorithm A		**Algorithm B**		**Algorithm C**	
True positives (n)	853		993		1085	
False positives (n)	0		19		31	
True negatives (n)	44582		44563		44551	
False negatives (n)	304		164		72	
Sensitivity (95%CI)	73.73	(71.09, 76.24)	85.83	(83.68, 87.79)	93.78	(92.23, 95.10)
Specificity (95%CI)	100.0	(99.99, 100.0)	99.96	(99.93, 99.97)	99.93	(99.90, 99.95)
PPV (95%CI)	100.0	(99.57, 100.0)	98,12	(97.08, 98.87)	97.22	(96.08, 98.11)
NPV (95%CI)	99.32	(99.24, 99.40)	99.63	(99.57, 99.69)	99.84	(99.80, 99.87)
Accuracy (95%CI)	99.34	(99.26,99.41)	99.60	(99.54,99.66)	99.77	(99.73,99.82)
K coefficient	0.85	(0.83, 0.86)	0.91	(0.90, 0.93)	0.95	(0.94, 0.96)
F1score	0,85		0,92		0,95	

PPV: Positive predictive value; NPV: negative predictive value; K: Kappa coefficient

When the results were stratified by sex and age, we observed that algorithm A had better characteristics in the women group and among people aged from 30 to 55 years, while algorithms B and C had improved performances in the men group and in the oldest age group (**Table 9**). However, the differences between subgroups were irrelevant, with an overlap of the 95% confidence intervals.

Table 9. Test characteristics of three diabetes case definition algorithms applied using known diabetes as the gold standard by sex and age

	Se (%) (95% CI)	Sp (%) (95% CI)	PPV (%) (95% CI)	NPV (%) (95% CI)	Acc (%) (95% CI)	Kappa (95% CI)
Algorithm A						
Men	72.99	100.0	100	99	99.02	0.84
	(69.74,76.07)	(99.98,100.0)	(99,36,100,0)	(98.85,99.13)	(98.88,99.15)	(0.82,0.86)
Women	75.27	100.0	100	99.61	99.62	0.86
	(70.56,79.57)	(99.98,100.0)	(98,69,100,0)	(99.53,99.69)	(99.53,99.69)	(0.83,0.88)
Age 18-30	71.43	100.0	100	99.91	99.91	0.83
	(41,90,91,61)	(99.92,100.0)	(69,15,100,0)	(99.77,99.98)	(99.77,99.98)	(0.67,0.99)
Age 30-55	76.86	100.0	100	99.76	99.76	0.87
	(71.03,82.02)	(99.98,100.0)	(98,04,100,0)	(99.69,99.82)	(99.69,99.82)	(0.83,0.90)
Age 55 +	72.92	100.0	100	98.57	98.62	0.84
	(69.89,75.80)	(99.98,100.0)	(99,44,100,0)	(98.38,98.74)	(98.44,98.79)	(0.82,0.86)
Algorithm B						
Men	85.86	99.96	98.83	99.47	99.45	0.92
	(83.22,88,22)	(99.92,99.98)	(97.70,99.49)	(99.36,99.57)	(99.34,99.55)	(0.90,0.93)
Women	85.75	99.95	96.67	99.78	99.73	0.91
	(81.78,89.14)	(99.92,99.98)	(94.11,98.32)	(99.71,99.83)	(99.66,99.80)	(0.88,0.93)
Age 18-30	78.57	99.98	91,67	99.93	99.91	0.85
	(49.2,95.34)	(99.87,100)	(61.52,99,79)	(99.80,99.99)	(99.77,99.98)	(0.70,0.99)
Age 30-55	81.82	99.97	96.12	99.81	99.78	0.88
	(76.37,86.47)	(99.93,99.99)	(92.49,98.31)	(99.75,99.86)	(99.71,99.84)	(0.85,0.91)
Age 55 +	87.01	99.94	98.74	99.31	99.28	0.92
	(84.,64,89.14)	(99.89,99.97)	(97.70,99.39)	(99.17,99.43)	(99.15,99.40)	(0.91,0.93)
Algorithm C						
Men	93.89	99.93	98.14	99.77	99.71	0.96
	(91.97,95.46)	(99.89,99.96)	(96.89,98.98)	(99.70,99.83)	(99.63,99.78)	(0.95,0.97)
Women	99.39	99.83	88.77	99.99	99.82	0.94
	(97.80,99.93)	(99.77,99.88)	(85.07,91.82)	(99.97,100)	(99.76,99.87)	(0.92,0.96)
Age 18-30	78.57	99.98	91.67	99.93	99.91	0.85
	(49.20,95.34)	(99.87,100)	(61.52,99.79)	(99.80,99.99)	(99.77,99.98)	(0.70,0.99)
Age 30-55	93.39	99.95	95.36	99.93	99.89	0.94
	(89.49,96.17)	(99.92,99.98)	(91.85,97.66)	(99.89,99.96)	(99.83,99.92)	(0.92,0.96)
Age 55 +	94.12	99.89	97.81	99.68	99.59	0.96
	(92.38,95.56)	(99.82,99.93)	(96.60,98.68)	(99.59,99.76)	(99.49,99.68)	(0.95,0.97)

Se sensitivity; Sp specificity; PPV positive predictive value; NPV negative predictive value; Acc accuracy

1.4.2 Gold standard "pharmacologically treated diabetes"

The test characteristics of the algorithms for identifying "pharmacologically treated diabetes" were slightly higher than in the previous analyses (**Table 10**).The sensitivity of algorithm A, algorithm B and algorithm C were 77.2%, 97.3% and 99.3%, respectively. The specificities and the NPVs of all algorithms were over 99%. Algorithm B had the highest PPV (97.9%) and algorithm C the lowest (90.6%.) Algorithm B had the most important values for accuracy, kappa coefficient and F1 in identifying "pharmacologically-treated diabetes, followed by algorithm C and algorithm A.

Table 10. Test characteristics of three diabetes case definition algorithms using pharmacologically treated diabetes as the gold standard

	Algorithm A		**Algorithm B**		**Algorithm C**	
True positives (n)	786		991		1011	
False positives (n)	67		21		105	
True negatives (n)	44654		44700		44616	
False negatives (n)	232		27		7	
Sensitivity (95%CI)	77.21	(74.51, 79.75)	97,35	(96.16, 98.25)	99,31	(98.59, 99.72)
Specificity (95%CI)	99.85	(99.81, 99.88)	99.95	(99.93, 99.97)	99.77	(99.72, 99.81)
PPV (95%CI)	92,15	(90.13, 93.86)	97,92	(96.85, 98.71)	90,59	(88.73, 92.24)
NPV (95%CI)	99.48	(99.41, 99.55)	99.94	(99.91, 99.96)	99.98	(99.97, 99.99)
Accuracy (95%CI)	99.35	(99.27,99.42)	99.90	(99.86,99.92)	99.76	(99.71,99.80)
K coefficient	0.84	(0.82, 0.86)	0.98	(0.97, 0.98)	0.95	(0.94, 0.96)
F1score	0,84		0,98		0,95	

PPV: Positive predictive value; NPV: negative predictive value; K: Kappa coefficient

The results of the analyses stratified by sex and age were similar for "pharmacologically treated diabetes" than for "known diabetes". Algorithm A had better characteristics in the women group and in the group aged from 30 to 54 years; algorithms B and C in the men group and in the oldest age group . Once again, no relevant differences between the subgroups were observed (**Table 11**).

Table 11. Test characteristics of three diabetes case definition algorithms applied using pharmacologically treated diabetes as the gold standard by sex and age

	Se (%) (95% CI)	Sp (%) (95% CI)	PPV (%) (95% CI)	NPV (%) (95% CI)	Acc (%) (95% CI)	Kappa (95% CI)
Algorithm A						
Men	76.30	99.79	92.15	99.22	99.04	0.83
	(72.95,79.42)	(99.71,99.84)	(89.63,94.21)	(99.10,99.34)	(98.90,99.16)	(0.81,0.85)
Women	79.14	99.91	92.14	99.71	99.63	0.85
	(74.32,83.42)	(99.86,99.94)	(88.35,95.01)	(99.64,99.78)	(99.54,99.70)	(0.82,0.88)
Age 18-30	90.91	100	100.0	99.98	99.98	0.95
	(58.72,99.77)	(99.92,100.0)	(69.15,100.0)	(99.87,100.0)	(99.87,100.0)	(0.86,1.00)
Age 30-55	80.77	99.92	90.32	99.83	99.76	0.85
	(74.75,85.89)	(99.85,99.95)	(85.14,94.16)	(99.77,99.88)	(99.68,99.81)	(0.82,0.89)
Age 55 +	76.1	99.71	92.54	98.88	98.64	0.82
	(72.98,79.01)	(99.62,99.79)	(90.26,94.43)	(98.71,99.03)	(98.46,98.81)	(0.81,0.85)
Algorithm B						
Men	97.11	99.95	98.53	99.9	99.86	0.98
	(95.57,98.23)	(99.91,99.98)	(97.32,99.29)	(99.85,99.94)	(99.80,99.91)	(0.97,0.98)
Women	97.85	99.95	96.67	99.97	99.93	0.97
	(95.63,99.13)	99.92,99.98)	(94.11,98.32)	(99.94,99.99)	(99.88,99.96)	(0.96,0.98)
Age 18-30	100.0	99.98	91.67	100.0	99.98	0.96
	(71.51,100.0)	(99.87,100.0)	(61.52,99.79)	(99.92,100)	(99.87,100.0)	(0.87,1.00)
Age 30-55	95.19	99.97	96.12	99.96	99.92	0.96
	(91.34,97.67)	(99.93,99.99)	(92.49,98.31)	(99.92,99.98)	(99.88,99.95)	(0.96,0.98)
Age 55 +	97.87	99.93	98.49	99.90	99.84	0.98
	(96.62,98.76)	(99.88,99.96)	(97.37,99.22)	(99.84,99.94)	(99.76,99.89)	(0.97,0.99)
Algorithm C						
Men	99.28	99.7	91.48	99.98	99.68	0.95
	(98.32,99.76)	(99.61,99.77)	(89.25,93.38)	(99.94,99.99)	(99.60,99.75)	(0.94,0.96)
Women	99.39	99.83	88.77	99.99	99.82	0.94
	(97.80,99.93)	(99.77,99.88)	(85.07,91.82)	(99.97,100.0)	(99.76,99.87)	(0.92,0.96)
Age 18-30	100.0	99.98	91.67	100.0	99.98	0.96
	(71.51,100.0)	(99.87,100.0)	(61.52,99.79)	(99.92,100.0)	(99.87,100.0)	(0.87,1.00)
Age 30-55	99.04	99.87	86.92	99.99	99.86	0.93
	(96.57,99.88)	(99.81,99.91)	(81.95,90.94)	(99.97,100.0)	(99.80,99.90)	(0.90,0.95)
Age 55 +	99.37	99.57	91.58	99.97	99.56	0.95
	(98.55,99.80)	(99.46,99.66)	(89.53,93.34)	(99.93,99.99)	(99.45,99.65)	(0.94,0.96)

Se sensitivity; Sp specificity; PPV positive predictive value; NPV negative predictive value; Acc accuracy

1.4.3 Analysis of the components of algorithm C

Among the three components of algorithm C, the one based on the information of antidiabetic drug reimbursement over the last two years had the best performances using both gold standards while the one based on diabetes hospitalizations had the lowest performances (**Table 12**). Indeed, this component had a sensitivity of 10.11% and 11.39%

and a kappa coefficient of 0.17 and 0.20 using “known diabetes” and “pharmacologically treated diabetes” as gold standard respectively.

Table 12. Test characteristics of different components of algorithm C

	Gold standard known diabetes			Gold standard pharmacologically treated diabetes		
	Component 1 ALD	**Component 2 Antidiabetic reimburs. (2 years)**	**Component 3 Diabetes hospitaliza. (2 years)**	**Component 1 ALD**	**Component 2 Antidiabetic reimburs. (2 years)**	**Component 3 Diabetes hospitaliza. (2 years)**
Se (%)	73.74	87.04	10.11	77.21	97.84	11.39
(95% CI)	(71.09,76.24)	(84.96,88.92)	(8.44, 12.00)	(74.51,79.75)	(96.7,98.64)	(9.51,13.51)
Sp (%)	100.0	99.94	99.99	99.85	99.92	99.99
(95% CI)	(99.99, 100)	(99.91,99.96)	(99.98, 100)	(98.59,99.72)	(99.88,99.94)	(99.97,100)
PPV(%)	100.0	97.39	96.69	92.15	96.32	95.87
(95% CI)	(99.57, 100)	(96.22,98.27)	(91.75,99.09)	(90.13,93.86)	(94.99,97.39)	(90.62,98.64)
NPV(%)	99.33	99.66	97.72	99.48	99.95	98.02
(95% CI)	(99.24,99.40)	(99.61,99.72)	(97.58,97.86)	(99.41,99.55)	(99.93,99.97)	(97.89,98.15)
Acc	99.35	99.61	99.72	99.35	99.87	98.02
(95% CI)	(99.26,99.41)	(99.55,99.67)	(97.58,97.85)	(99.27,99.42)	(99.83,99.90)	(97.89,98.14)
K	0.86	0.91	0.17	0.84	0.97	0.20
(95% CI)	(0.83, 0.86)	(0.90, 0.93)	(0.15,0.20)	(0.82, 0.86)	(0.96,0.98)	(0.17,0.23)

Se sensitivity; Sp specificity;PPV positive predictive value;NPV negative predictive value;Acc accuracy

1.5 Discussion

The objective of this section was to validate the three diabetes case definition algorithms using the data from the CONSTANCES cohort. In general, the algorithms presented very good performances in identifying known diabetes and pharmacologically treated diabetes cases. Irrespective of the algorithm applied, the specificity was above 99.0%. This high proportion of true negatives among those not having the disease is a common characteristics of the algorithms used to ascertain chronic diseases in health administrative databases [132]. On the contrary, these algorithms usually present less important sensitivities. A meta-analysis on the case definition algorithms from the Canadian National Diabetes Surveillance system which is based on the different regional health administrative databases found a pooled sensitivity of 82.3% [172]. The lowest sensitivity in identifying known diabetes cases estimated in our study was 74% for the algorithm A while the algorithms B and C had sensitivities values over 86 % for both

gold-standards (known diabetes and pharmacologically treated diabetes).

1.5.2 Algorithm A: ALD diabetes

Many French authors had used algorithm A to identify diabetes cases in their studies due to its computational simplicity [173, 174]. The PPV of this algorithm was 100% due to the absence of false positives. This fact could be explained by the complex and controlled administrative procedure that the patient and her/his practitioner must follow in order to benefit from an ALD diabetes status. However, the other performances of algorithm A were lower than those of algorithms B and C. Also, this algorithm has important limitations when it is applied to assess temporal trends. Before 2014, the information related to ALD-diabetes was poorly recorded or not accessible for certain Health Insurance Funds, such as MSA or RSI [170]. Also, a study conducted in 2014 observed that 21% of all pharmacologically treated diabetes cases did not benefit from an ALD-diabetes [170] and this rate is higher in FOT and in patients benefiting from CMUc making the algorithm not suitable for assessing socioeconomic or regional disparities.

1.5.3 Algorithm B: antidiabetic drug reimbursements

Initially, an algorithm based on antidiabetic drug reimbursement was developed by the CNAMTS at the beginning of the 2000's. It was based on two reimbursements per year [175]. Then, it evolved in order to better take into account drug practices and packaging, resulting in the definition of the Algorithm B. Nowadays, the algorithm B is routinely applied to assess diabetes prevalence in France by the French diabetes surveillance system, because of the absence of socioeconomic or regional disparities on the recording of antidiabetic drug reimbursement [170]. Moreover, the quality of this information does not vary by year of study or health insurance fund. Due to these characteristics, algorithm B appears appropriate for the study of trends and as well as socioeconomic or regional inequalities.

Because algorithm B requires at least three reimbursements (out hospital) to classify individuals as positive cases, those diabetic patients who were hospitalized for long periods of time due to complications or those who died before having three reimbursements could be false negative cases. This fact could represent a limitation for studies on diabetes morbidity or mortality which need to include severe diabetes cases.

1.5.3 Algorithm C: ALD diabetes, antidiabetic drug reimbursement and diabetes hospitalizations

In our study, algorithm C presented the best performances, especially in identifying known diabetes cases. The case definition is broader, captures non-pharmacologically treated diabetes cases as well as severe cases hospitalized for long periods, leading to a limited number of false negatives.

When applying the algorithm C in the SNDS, certain practical implications must be regarded. First, since part of the algorithm is based on ALD-diabetes information, the limitations described for algorithm A are also applicable for algorithm C [170]. Also, the algorithm is very computationally expensive, since it requires a large number of variables from two different SNDS sources: the DCIR and the PMSI.

In the last part of the analyses, we compared the performances of the different components of the algorithm C. The component based on the hospitalizations related to diabetes over the last two years had a reduced sensitivity for both gold standards because of a high proportion of false negatives. Concerning other test characteristics, all the elements had specificities and PPV over 99 %, due to the reduced proportion of false negatives, a common feature of the chronic disease algorithms as we have previously described.

1.6 Conclusion

In this section, we have assessed the test characteristics of the three algorithms used in the SNDS for identifying diabetes cases. All the algorithms had excellent performances without relevant differences between age groups or by gender. However, when selecting an algorithm for the ascertainment of diabetes cases in HAD, other factors must be taken into account like the objective of the study because each algorithm has specific limitations related to heterogeneity and the quality of the variables composing the algorithm. For surveillance purposes and particularly for this thesis for which the next objective was the assessment of the evolution of diabetes epidemic in France (including the analysis of temporal trends and regional disparities), we considered that the algorithm B based on the information on antidiabetic drugs reimbursement was the most suitable for that purpose.

Further information on the validation of the three case diabetes definition algorithms used in the SNDS is available in the article published in the journal International Journal of Public Health **(See Annex II: Article 1)**

International Journal of Public Health
https://doi.org/10.1007/s00038-018-1186-3

ORIGINAL ARTICLE

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

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

Received: 18 June 2018 / Revised: 3 September 2018 / Accepted: 26 November 2018

2. Evolution of the diabetes epidemic in France

2.1 Introduction

We have seen that since the first estimations, the number of people with type 2 diabetes has never ceased to increase worldwide, leading to define it as "the diabetes epidemic" (**See page 34**). Notwithstanding, the dynamics of the diabetes epidemic might shift in certain Western countries where a leveling off or a decrease in diabetes incidence rates have been observed [72, 80, 176]. In France, the evolution of prevalence has already been studied [175] but not the evolution of the diabetes epidemic due to the absence of data on incidence (**Figure 39**).

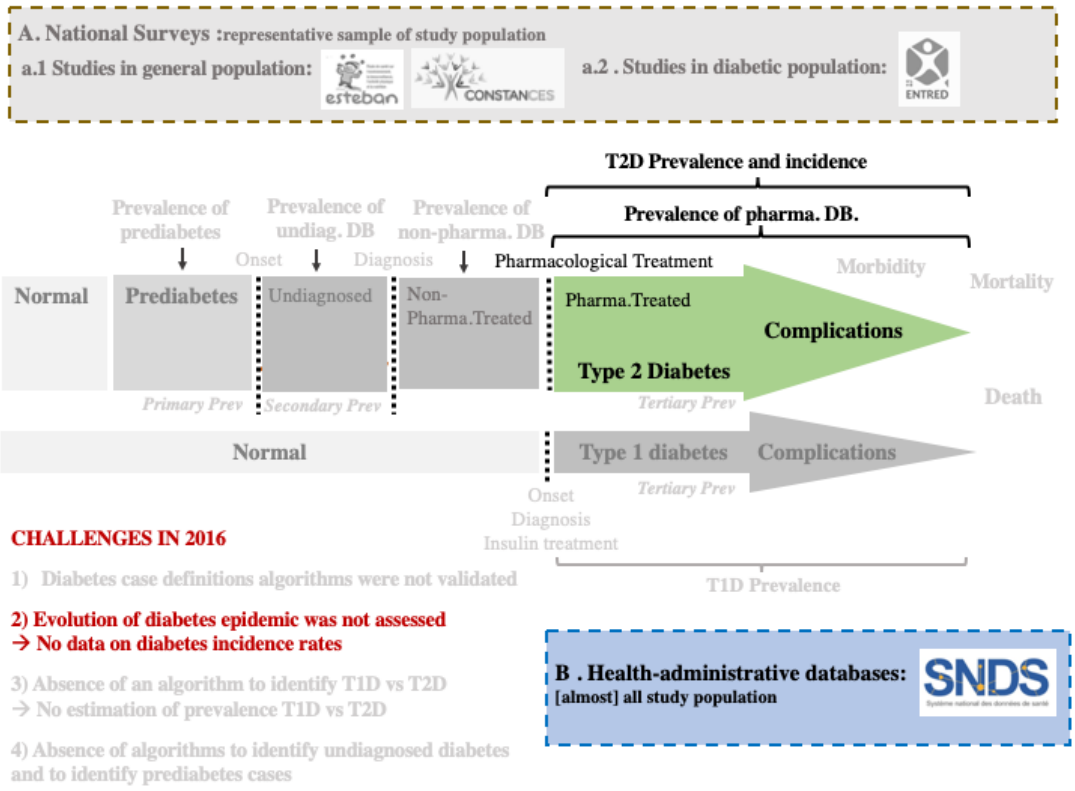


Figure 39. Challenges faced in the results' section 2

The validated algorithms in the previous section offered an opportunity to identify prevalent and incident cases in the SNDS, an exhaustive nationwide data source including the 66 million people living in France (metropolitan France and FOT).

2.2 Objectives

The objective was to describe the evolution of diabetes epidemic in France through the estimation of the prevalence and incidence rates between 2010 and 2017 by sex, age and region and the study of their annual time trends over the study period.

2.3 Methods

As described in **Figure 40**, this stage of the thesis corresponded to the third step in the baseline method.

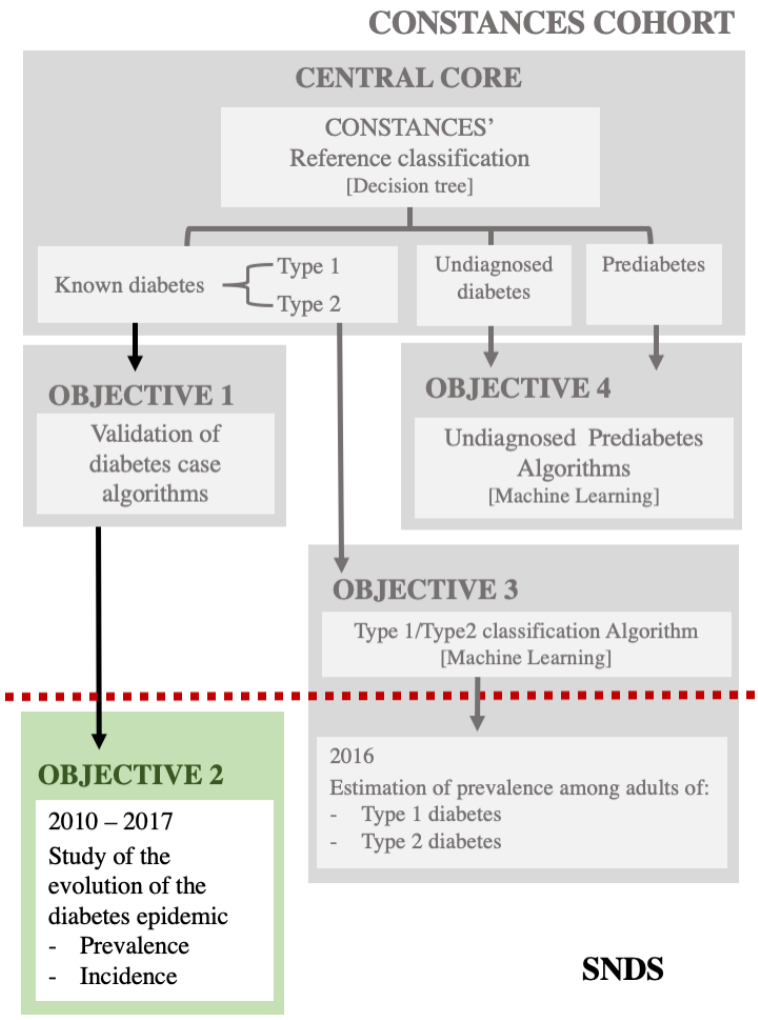


Figure 40. Methods of the results' section 2

In the previous section, we considered that the most suitable algorithm for surveillance purposes in general and for studying temporal trends and regional disparities in particular, was algorithm B, identifying as positive case those individuals with a reimbursement of antidiabetic drugs on at least three different dates over the last year or two dates if at least one large package of antidiabetic drugs was dispensed. The algorithm was applied on the entire SNDS to construct a retrospective cohort, which was used to estimate the prevalence and incidence rates over the study period.

2.3.1 The retrospective cohort of diabetes cases

Since for certain important health insurance schemes, such as MSA or RSI, the DCIR data were not available in the SNDS before 2010 (**See page 64**), the study period was restricted from year 2010 to year 2017. The information on antidiabetic drug

reimbursements was extracted for each year in order to apply the diabetes case definition algorithm. The unique identification number allowed to construct a retrospective cohort of cases by cross-referencing the individuals identified as diabetes cases in different years.

2.3.2 Prevalence and incidence rates

A diabetes case for a given year was classified as a prevalent case. The prevalence rates were calculated by dividing the number of prevalent cases by the mean French population for each year. The French population was the mean number of residents between 1st January and 31st December estimated by the National Institute for Statistics (*Institut national de la statistique et des études économiques*, Insee).

When an individual was a diabetes case in a given year and not a diabetes cases in the two previous years, it was classified as an incident case. The number of incident case was divided by the population at risk to estimate the incidence rates for each year between 2012 and 2017. The population at risk was defined as the mean number of the total population free of diabetes at the beginning (the French population minus the number of prevalent cases in the previous year) and at the end of the year (the French population minus the number of prevalent cases in the year of study).

The estimation of crude prevalence and incidence rates was stratified by sex, age (one year class) and region (17 regions because Mayotte region was excluded due to the lack of exhaustivity of data recorded on consumption). The 2013 European population was used to assess standardized rates by sex and region [177]. In order to focus on the study of the evolution of type 2 diabetes, the analyses were restricted to population aged 45 years old and older.

2.3.3 Analysis of trends

The number of prevalent and incident cases were modeled through generalized linear regressions to assess the annual time trends. They were analyzed by modelling the number of prevalent cases and incident cases with, respectively, the log- French population and the log- French population at risk as offsets.

Models with a log link and negative-binomial distribution were applied due to the overdispersion in both outcomes. The independent variables of the model were: calendar year, age (as a continuous parametric fractional polynomial function) and region [178]. In the variable “region”, Occitanie was the reference because its prevalence and incidence rates over the study period were the closest to the national rates. Also in the model, an interaction term between region and calendar year was included in order to study the

regional trends.

2.4 Results

In 2017, a total of 3,333,741 diabetes cases were identified in France based on the SNDS (1,836,410 men and 1,497,331women) (**Figure 41**). More than 94% of these cases were aged 45 years or more with a mean age of 69 years and a ratio of males to females of 1.24.

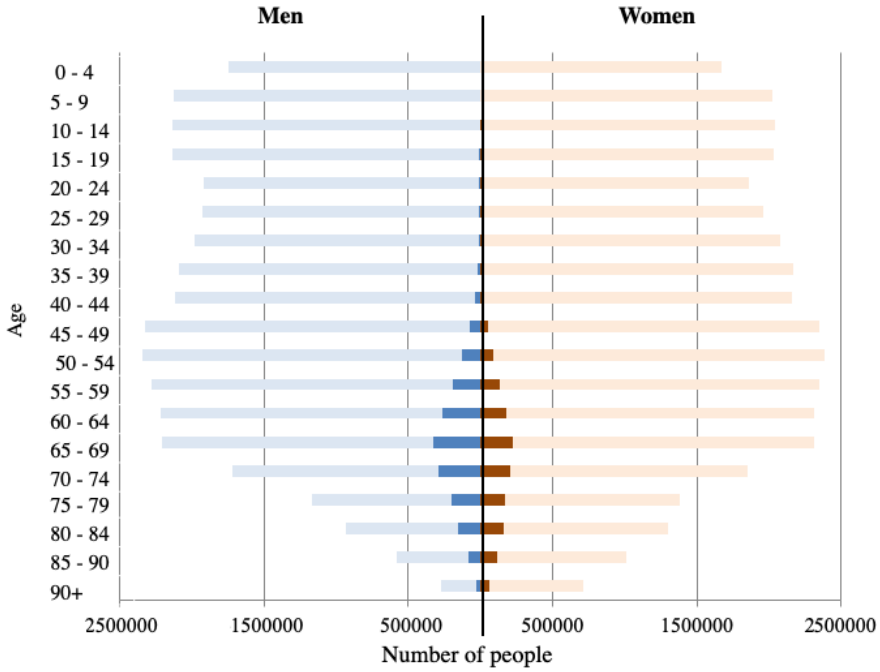
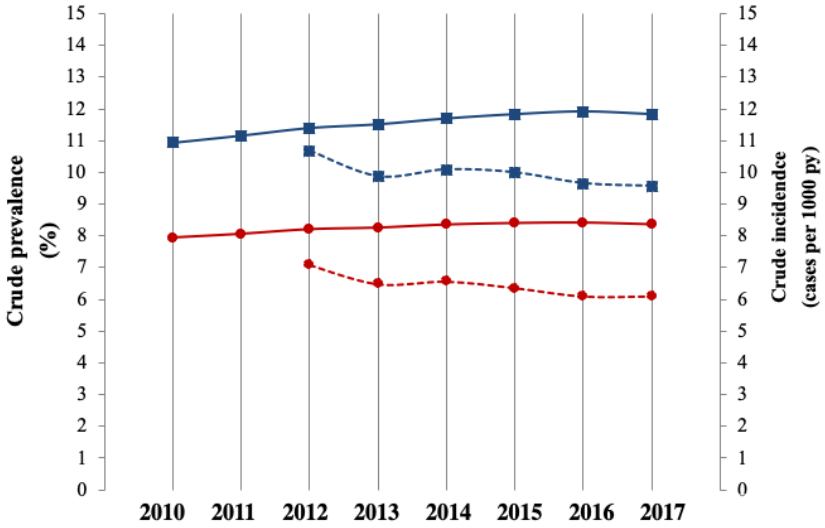


Figure 41. Pyramid of general population and diabetic population in France in 2017

2.4.1 Diabetes prevalence and incidence rates

In **Figure 42** the evolution of crude prevalence and incidence rates in adults aged 45 years or more is represented. Among men, the crude prevalence of diabetes was 10.9% in 2010 and 11.8% in 2017 while among women it was 7.9% and 8.4%, respectively. Regarding crude incidence rates, it was 10.7 cases per 1000 person-years (py) in 2012 and 9.6 cases per 1000 py in 2017 in men and it was 7.1 and 6.1 cases per 1000 py, respectively in women.



Prevalence men: blue solid line; incidence men: blue dot line; prevalence women: red solid line; incidence women: red dot line;

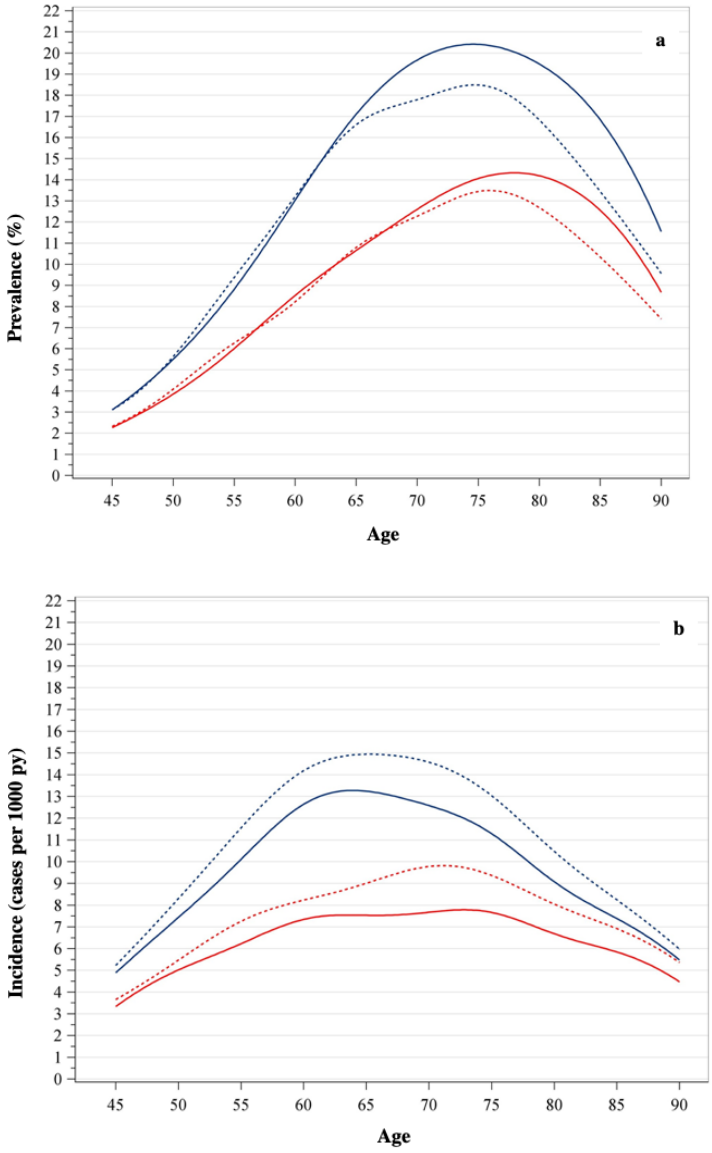
Figure 42. Evolution of crude prevalence and incidence of diabetes between 2010 and 2017 among adults aged 45 years or more by sex

The evolution of age-standardized prevalence and incidence rates were similar to those described for crude rates and they are shown in **Table 13**.

Table 13. Age-standardized prevalence and incidence of diabetes between 2010 and 2017 by sex among adults aged 45 years or more

	Age-standardized prevalence (%)							
	2010	2011	2012	2013	2014	2015	2016	2017
Men	11.5	11.7	11.9	11.9	12.1	12.2	12.2	12.1
Women	7.9	8.0	8.1	8.1	8.2	8.2	8.2	8.1
	Age-standardized incidence (per 1000 py)							
			2012	2013	2014	2015	2016	2017
Men			11.0	10.1	10.4	10.2	9.8	9.7
Women			7.2	6.6	6.6	6.4	6.2	6.2

The prevalence rates were higher in men through all ages (**Figure 43**). The highest point for men was in the age group 75-79 years, 18.5% in 2010 and 20.4% in 2017. Among women, this point was also at 75-79 year in 2010 (13.4%), while in 2017 it was among those aged 80-84 years (14.2 %).



Men in 2010/12 men blue dotted line; men in 2017 blue solid line; women in 2010 /12 red dotted line; women in 2017 red solid line.

Figure 43. Age-specific prevalence (a) and incidence (b) in 2012 and 2017 in France among adults aged 45 years and over, stratified by sex

As described for prevalence, the age-specific incidence was higher among men in 2012 and in 2017 than among women. In men, both curves increased from the first point in the age group 45-49 years until their highest point in the group 65-69 years (14.9 cases per 1000 py in 2012 and 13.2 cases per 1000 py in 2017). The age-specific incidence in women reached its peak at 70-74 years (9.8 cases per 1000 py) and then decreased in 2010 but in 2017, the incidence rates plateaued out from 60 to 80 years (7.6 cases per 1000 py).

The **Figure 44** corresponds to the age-standardized prevalence of diabetes in 2017 by sex and region. In metropolitan France, a West to North-East gradient was observed. The highest prevalence in both sex was found in the FOT (Reunion, Guadeloupe,

Martinique and French Guiana). To note, prevalence rates of diabetes were more important in women than in men only in FOT.

2.4.2 Annual time trends

After adjusting for age and geographical region, the annual time trend for diabetes prevalence among men aged 45 years and over was +0.9 % (95% CI +0.7, +1%) and it was -2,6% (95% CI -3.1, -2%) for incidence. The same dynamics were observed in women with an increasing annual time trend for prevalence of +0.4 % (95%CI +0.2, +0.6%) and a decreasing annual time trend for incidence of -3.9% (95%CI -4.5, -3.4%).

2.4.3 Regional disparities

The smallest age-adjusted prevalence rates were in Brittany (7.3% in 2010 to 8% in 2017) and the greatest rates in Reunion (19.4% in 2010 and 19.3% in 2017). The annual time trend in these two regions were +1.3 % and –0.3%, respectively (**Figure 45**a). An increasing time trend was observed in all regions except in Reunion. However, this decreasing annual time trend in Reunion was not significantly different from 0%. In women, the age-adjusted prevalence also increased between 2010 and 2017 in all regions, except in Martinique and Reunion were it decreased (17 to 16.1% and 21.9 to 20.2%, respectively) (**Figure 45**b).The annual time trend over the study period in Reunion and in Martinique were significantly different from 0% (-1.1 % and -0.9%) confirming these decreasing trends in prevalence.

Age-adjusted incidence rates in men decreased between 2012 and 2017 in every French region between 2012 and 2017, especially in the regions with the highest prevalence rates like Guadeloupe (from 16.5 to 12.8 cases per 1000 py) and Reunion (from 17.5 to 13.3 cases per 1000 py) (**Figure 46**a). These regions also presented the highest decrease in annual time trends: Guadeloupe -3.8% and Reunion -4.4%. Similar results were observed in women (**Figure 46**b) ,with the largest decrease in incidence rates in FOT: Martinique, from 13.3 to 10.8, Reunion from 16.4 to 10.7, Guadeloupe from 16.8 to 12.6 and French Guiana from 21.5 to 15 cases per 1000 py. Their decreasing annual time trends were also the largest: Martinique: -4.5%, Guadeloupe -4.8%, French Guiana -5.3% and Reunion -7.5%.

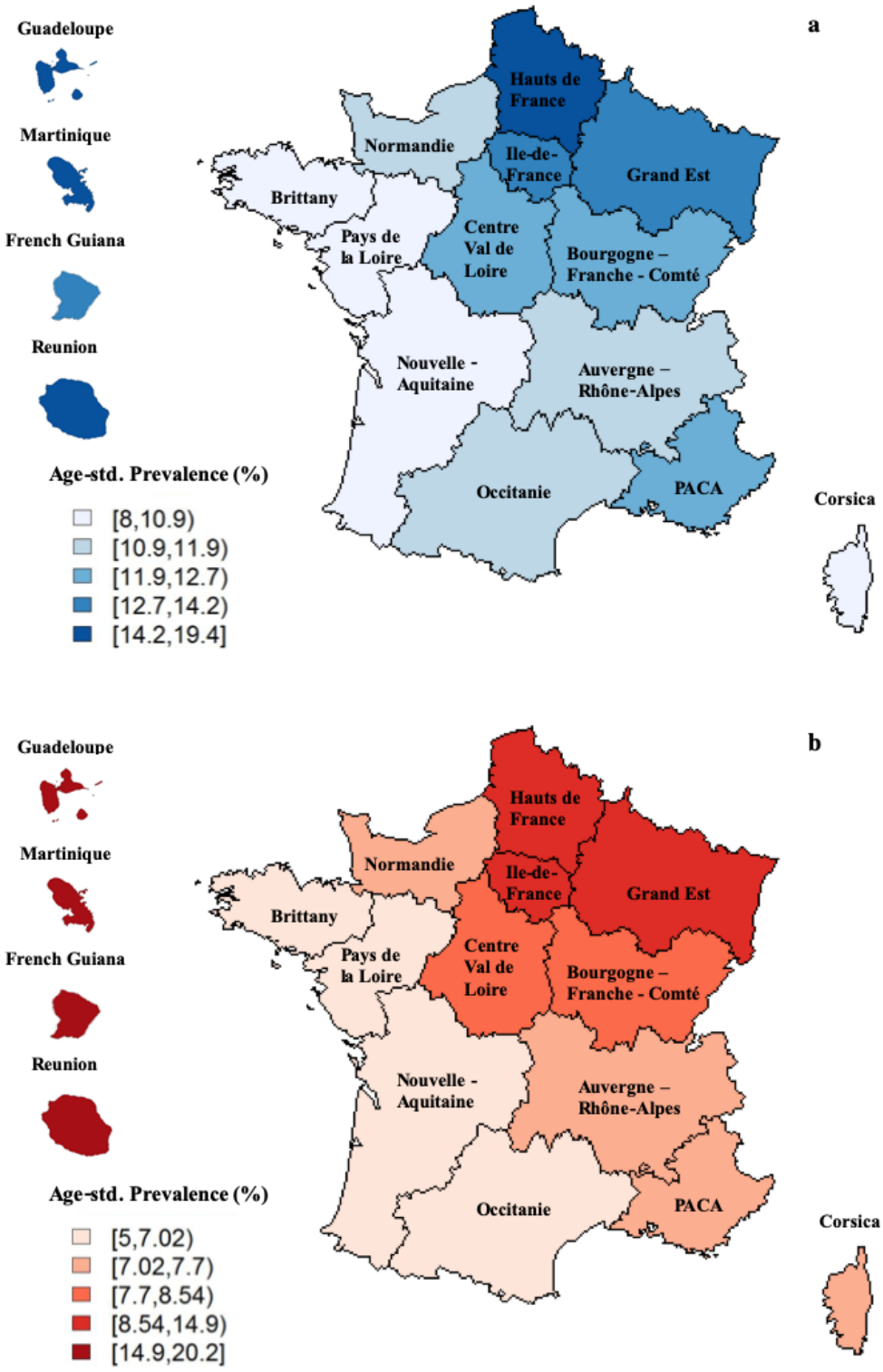
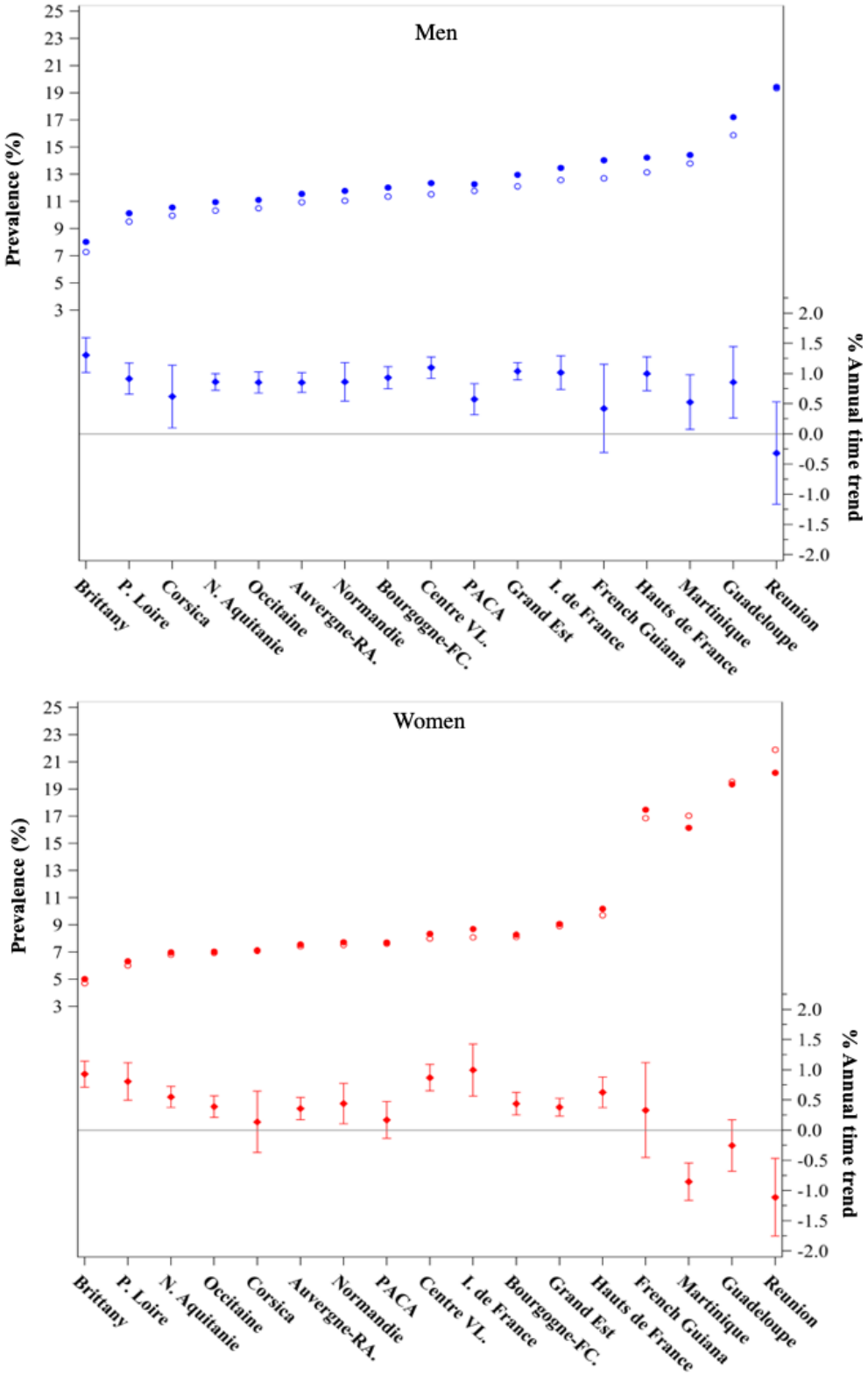
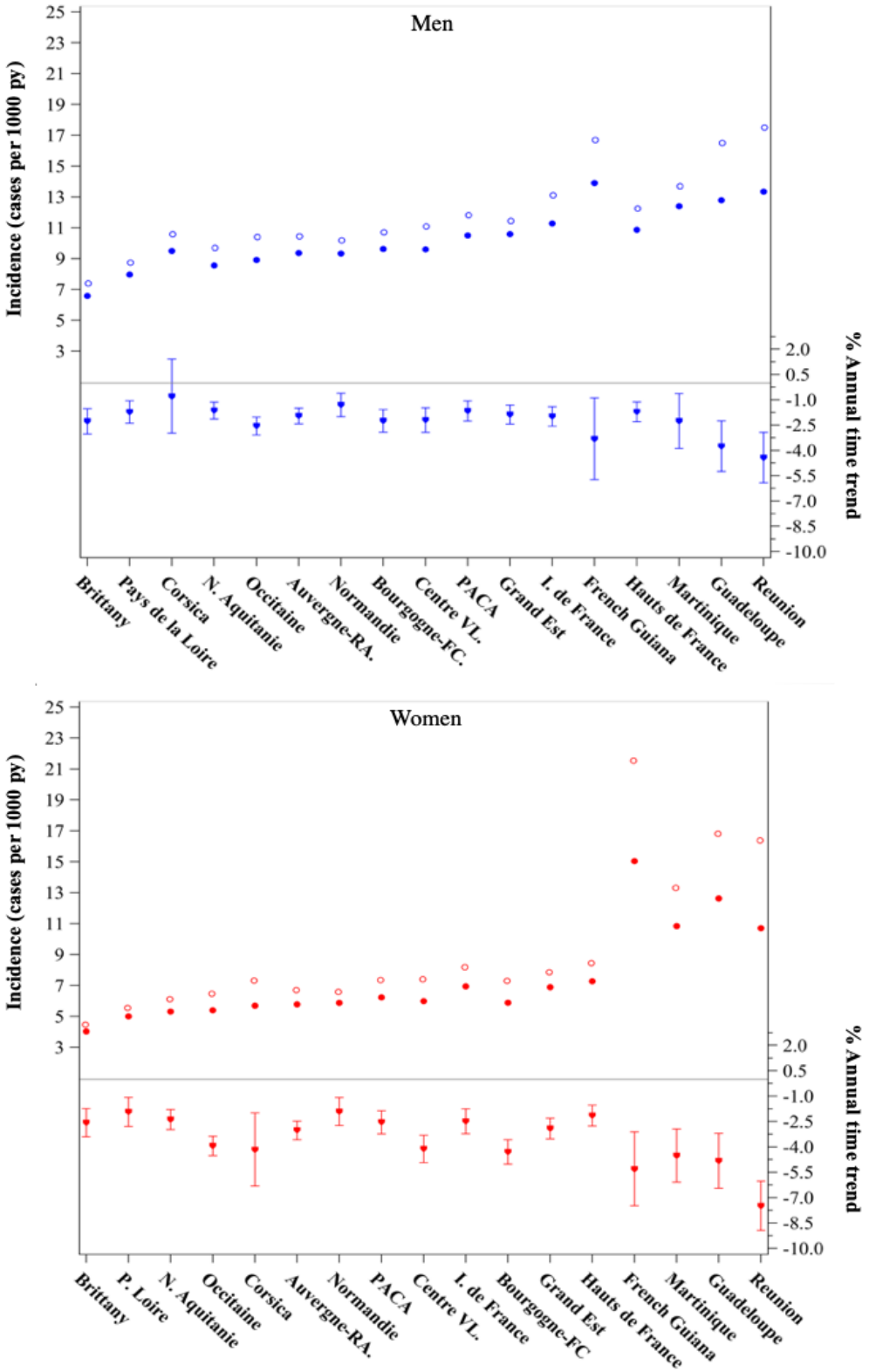


Figure 44. Age-standardized prevalence of type 2 diabetes in France in 2017 among men (a) and women (b) aged 45 years and over by geographical region



Top plot: age-standardized prevalence (%) in 2010 (void points and 2017 (solid points)
Bottom plot- percentage annual time trends

Figure 45. Evolution of diabetes prevalence in France among men and women stratified by geographical region



Top plot: age-standardized incidence (cases/1000py) in 2012 (void points) and 2017 (solid points)
Bottom plot- percentage annual time trends

Figure 46. Evolution of diabetes incidence in France among men and women stratified by geographical region

2.5 Discussion

One of the challenges of diabetes surveillance in France was the absence of information on the evolution of diabetes epidemic. Previous studies have estimated an important increase in prevalence over an earlier period [94, 175, 179]. However, the complete dynamic of diabetes has not been studied due to the lack of data on incidence rates. This limitation was overcame using a retrospective cohort with all the diabetes cases identified in the SNDS between 2010 to 2017. We observed a slight increase on the prevalence trends but a decrease on the trends of incidence rates over the study period. These dynamics have already been described in other countries such as Norway, Sweden or the US [72, 80, 176].

2.5.1 Understanding the dynamics of the diabetes epidemic

There are different hypotheses about the reduction on incidence rates. On one hand, these results could be explained by the reduction in the pool of undiagnosed diabetes cases due to previous efforts in screening [176]. Since our algorithm is based on antidiabetic drug reimbursement, a rise on the number of non-pharmacologically treated diabetes cases could be also responsible for the diminution of incidence rates. However, the prevalence of undiagnosed diabetes and non-pharmacologically treated diabetes in France did not increase, based on the comparison of the results of the ENNS in 2006 and the results of the CONSTANCES cohort [97, 180].

Another hypothesis is related to the slowing down on the prevalence of diabetes risk factors. Despite the important prevalence rates of obesity described in the US, they have become stable over the last decade [181]. In France, the comparison of two similar studies conducted in 2006 and 2015 -ENNS and Esteban- showed stable prevalence rates of obesity (17%) and of overweight (49%) [97, 182]. This plateau on overweight/obesity could be due to the implementation of national prevention programs as the Nutrition and Health National Plan (*Plan national nutrition santé*, PNNS) in France in the beginning of the 2000's. Finally, the use of bariatric surgery in people with a high risk of developing a diabetes could be another possible explanation for the reduction on diabetes incidence rates.

2.5.2 Understanding the regional inequalities on the diabetes epidemic

Regional disparities were observed. The prevalence of diabetes had a West to North-East gradient in metropolitan France. The same gradient has been described for various diabetes risk factors such as obesity or low socioeconomic status [183, 184].

The FOTs had the highest rates of diabetes prevalence. In these regions, conversely

to mainland France, the rates were more elevated among women than among men. Various factors could explain this inverse male to female ratio. Ethnicity and/or cultural factors may play a role as it has been described in other Caribbean countries and in South Africa countries [185, 186]. Also, higher prevalence of diabetes risk factors such as obesity/overweight have been described among women in these regions [186, 187].

Both the highest prevalence rates and the largest decrease in incidence were observed in the FOTs. The decreasing trends can be related to changes in the prevalence of undiagnosed diabetes or non-pharmacologically treated diabetes as we have noticed before for all France. However, based on the results of two national surveys conducted in these regions in 2002 and in 2014, the prevalence of obesity have leveling off in Guadeloupe and Martinique and even decreased in Reunion [188, 189]. These favorable trends could be due to the increasing awareness of the burden of diabetes and obesity in these regions among the public health actors and general population. A shift in the age of diagnosis in FOT to younger age groups under 45 years among the new diabetes could be another possible explanation of this decrease.

2.6 Conclusion

The results of this section are coherent with the favorable dynamics on the diabetes epidemic described in other Western countries and they could represent a glimmer of hope in the reduction of the burden of type 2 diabetes in France.

However, the case definition algorithm was not able to differentiate between type 1 and type 2 diabetes. Thus, we had to narrow the analysis to the population aged 45 years or more where the incidence and the prevalence of type 1 diabetes are minimal. In the next section, we aimed to develop an algorithm to classify separately type 1 and type 2 diabetes in order to assess their specific prevalence among adults in France.

Further details on the study of diabetes epidemic in France are available in the article published in the journal Diabetes and Metabolism: “Is the type 2 diabetes epidemic plateauing out in France? A nationwide population based study”**) (See Annex III: Article 2)**

Diabetes & Metabolism
Available online 7 January 2020
In Press, Corrected Proof

Original article

Is the type 2 diabetes epidemic plateauing in France? A nationwide population-based study

S. Fuentes [a], L. Mandereau-Bruno [a], N. Regnault [a], P. Bernillon [a], C. Bonaldi [a], E. Cosson [b, c], S. Fosse-Edorh [a]

Show more

https://doi.org/10.1016/j.diabet.2019.12.006

3. Development of a type 1/type 2 diabetes classification algorithm

3.1 Introduction

In the last section, we had to restrict the study population to individuals aged 45 years or older to assess the prevalence and the incidence of type 2 diabetes due to the inability of the algorithm applied to differentiate type 1 and type 2 diabetes cases (**Figure 47**).

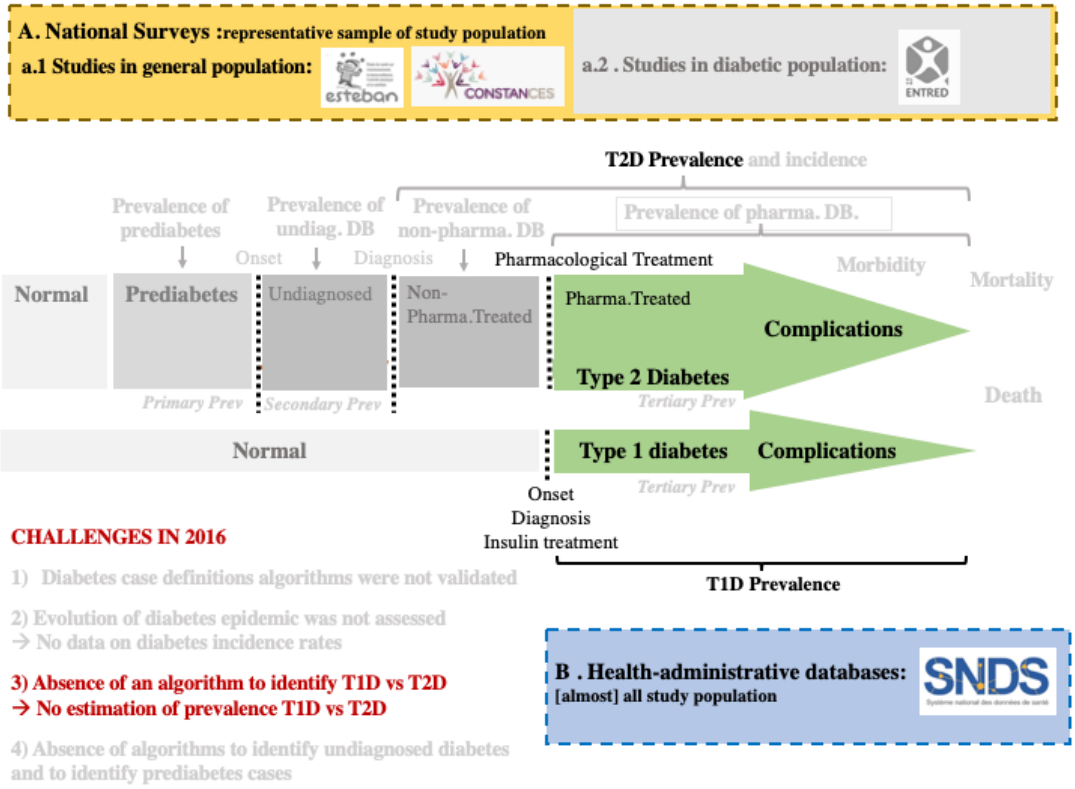


Figure 47. Challenges faced in the results' section 3

The algorithms to ascertain health conditions on the SNDS were usually based on the criteria of professional experts in the field. Recently, artificial intelligence and more specifically SML has opened new perspectives in the development of case definition algorithms. SML consists in different methods based on datasets where known variables and targets are linked (final data set) for developing algorithms to predict or characterized these targets in the study population.

3.2 Objectives

The objective of this section was to develop an algorithm to classify type 1 and type 2 diabetes based on the SNDS information through SML and to apply it for assessing the prevalence of type 1 and type 2 diabetes in France among adults in 2016.

3.3 Methods

The known diabetes identified in the CONSTANCES population constituted the final data set used for the development of the classification algorithm (**Figure 48**).

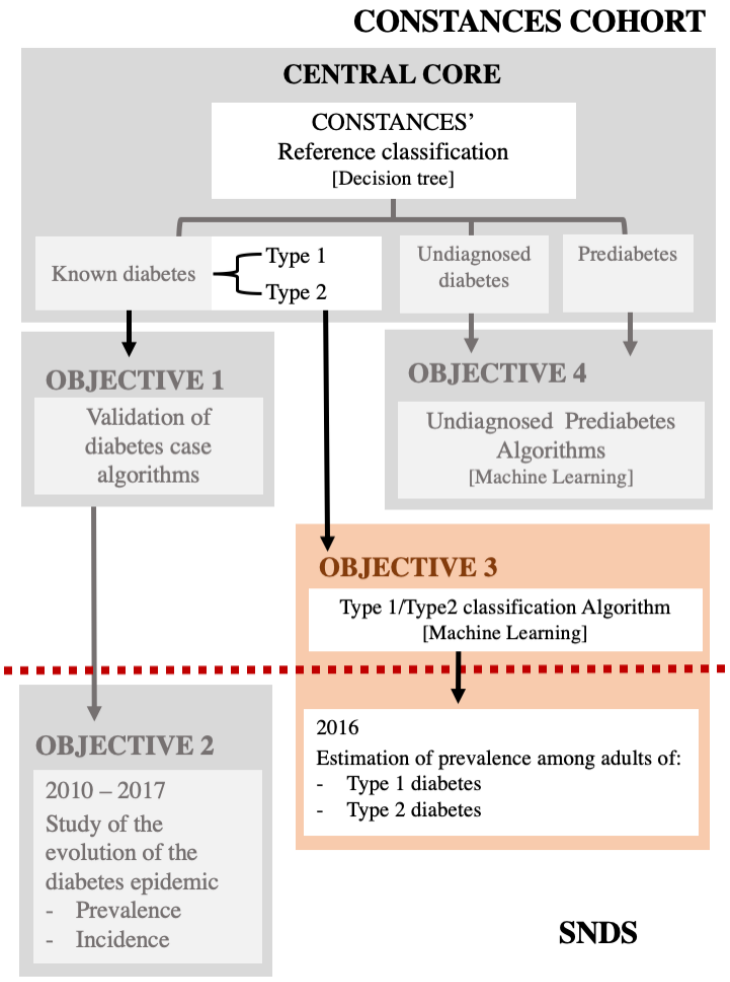


Figure 48. Methods of the results' section 3

These diabetes cases were classified in type 1 and type 2 through the Entred decision tree in the central core step of the baseline method (**See page 93**). This classification was used to define the target in the SML method. Different algorithms were developed and the most suitable were selected to assess the prevalence of type 1 and type 2 diabetes among adults in France using the entire SNDS database.

3.3.1 Development of a classification algorithm using SML

The SML method applied to develop the classification algorithm was based on eight steps (**Figure 49**) : (1) selection of the final data set, (2) target definition, (3) coding variables for a given time window, (4) splitting final data set into training and testing data sets, (5) variables selection, (6) training algorithms, (7) algorithms validation and (8) final algorithm selection [190].

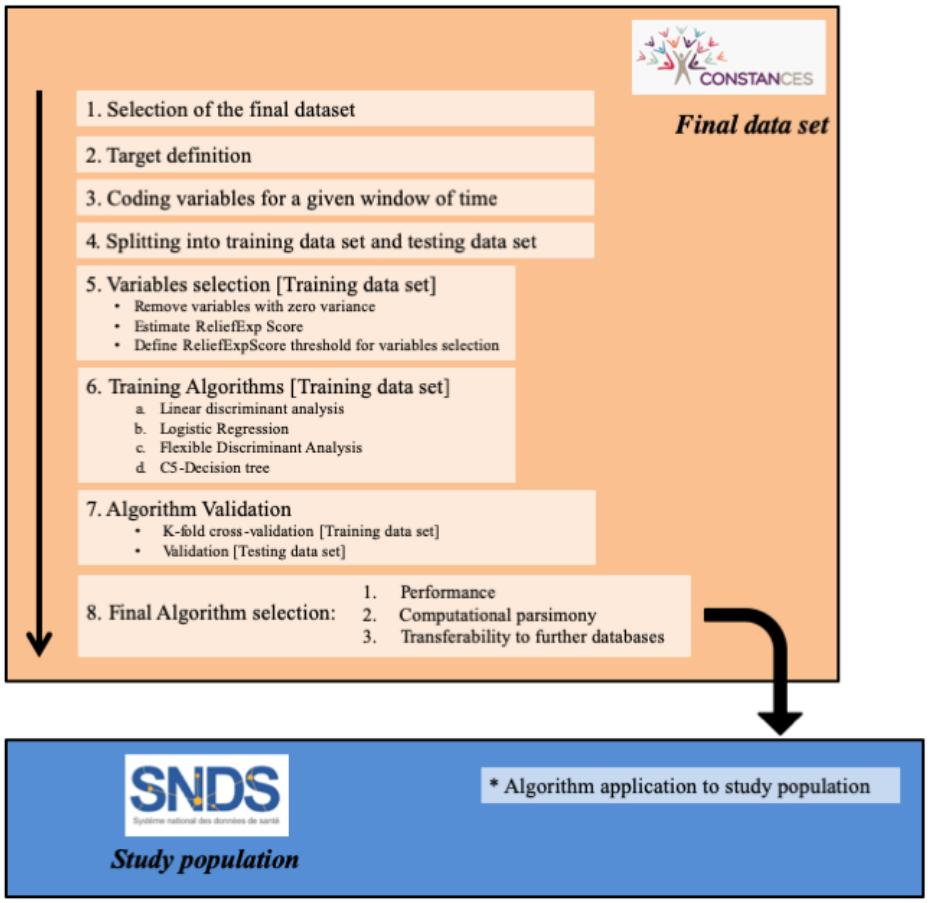


Figure 49. Supervised machine learning method for developing algorithms

Step 1: Selection of the final data set

The final data set was composed by the pharmacologically treated diabetes cases from the CONSTANCES population after excluding pregnant women, participants without accessible data in the SNDS and diabetes cases without complete data on their diabetes diagnosis and treatment.

Step 2: Target definition

In the central core step of the thesis , we classified diabetes cases through the Entred decision tree based on three items: age at diabetes diagnosis, current insulin treatment and delay between diabetes diagnosis and first insulin treatment (**See page 93**). This classification was used to define the target of the algorithm: target 1 were all type 1 diabetes cases and target 0 were all type 2 diabetes cases.

Step 3: Coding variables for a given window of time

Based on the SNDS data from the CONSTANCE participants included in the final data, a total of 3481 continuous variables were coded. These variables included:

a) Number of out-of-hospital health care reimbursement in the 12 months before the date of the self-administered questionnaire: using the information from the DCIR

tables, we coded variables related to the reimbursement of medical consultations, biological test and medical acts performed and drugs (using the classification ATC-05) and medical devices for self-monitoring of blood glucose and self-treatment medical dispensed.

b) Number or days of hospitalizations in the 24 months before the date of the self-administered questionnaire: the information from the different tables in the PMSI was used to code the number of hospitalizations (total hospitalizations, less than 1 day hospitalizations and hospitalizations between 1 and 7 days) and the days of hospitalizations of overall hospitalizations, hospitalizations related to a specific treatment such as radiotherapy or dialysis, hospitalizations not related to a specific treatment or hospitalizations associated to a specific diagnosis; the diagnosis considered were heart failure, stroke, foot ulcer, lower limb amputation, ischemic heart disease, transient ischemic attack, end stage renal failure, diabetic coma, diabetic ketoacidosis, cancer and diabetes.

c) Demographics variables: the coded variables were age, sex (value 1 for men and 2 for women) and characteristics of the city/town of residence (FDEP and rural/urban status).

Step 4: Training data set and testing data set

The final dataset was divided into training dataset (80%) and testing dataset (20%). In the training dataset, a random down-sampling was performed in the target negative group due to the substantial imbalance between the number of type 1 and type 2 diabetes cases.

Step 5: Variables selection

All the variables with a variance equal to zero were removed. Then, the ReliefFExp score was estimated for the variables retained. This score is calculated using the Relief algorithm, which evaluates the capability of each variable to differentiate between target 1 and target 0 [191-193]. The Relief algorithm is commonly used as a filter-method approach since it is noise-tolerant and not affected by features interactions. All the variables were ranked based on their ReliefFExp score. Only the variables with a ReliefFExp score equal to 0.05 or higher were used for the development of the algorithms.

Step 6 and 7: Training and validation of algorithms

Four models were applied to the training dataset: linear discriminant analysis (LDA), logistic regression (LR), flexible discriminant analysis (FDA) and C.5 decision tree (C5). For each model, three algorithms were trained using three, nine and fourteen

variables. The variables were selected with three different thresholds for the ReliefFExp score: 0.35, 0.01 and 0.05.

The twelve algorithms were validated using first the training dataset (through k-fold cross validation). A second validation of the algorithm's performances was done on the testing data set. The test characteristics assessed were sensitivity, specificity, accuracy, kappa and F1 score.

Step 8: Final algorithm selection

Finally, one out of the twelve algorithms was selected based on the following criteria: performance, computational parsimony and applicability to other health administrative databases.

3.3.2 Assessment of the prevalence of type 1 and type 2 diabetes in France in 2016

To define the study population for applying the classification algorithm, the diabetes cases in 2016 were ascertained in the entire SNDS using the case definition algorithm based on the reimbursement of antidiabetic drugs. Then, women identified as diabetes cases who gave birth in 2016 and individuals aged lower than 18 years or higher than 70 years were excluded. The classification algorithm previously selected was applied in the study population to characterize each case as type 1 or type 2 diabetes.

The prevalence of type 1 and type 2 diabetes was estimated using the mean French population in 2016 estimated by the INSEE as denominator. The results were stratified by sex and age (1-year class). The prevalence of type 1 and type 2 diabetes in all study population was adjusted by the performances of the algorithms (PPV and NPV) [172].

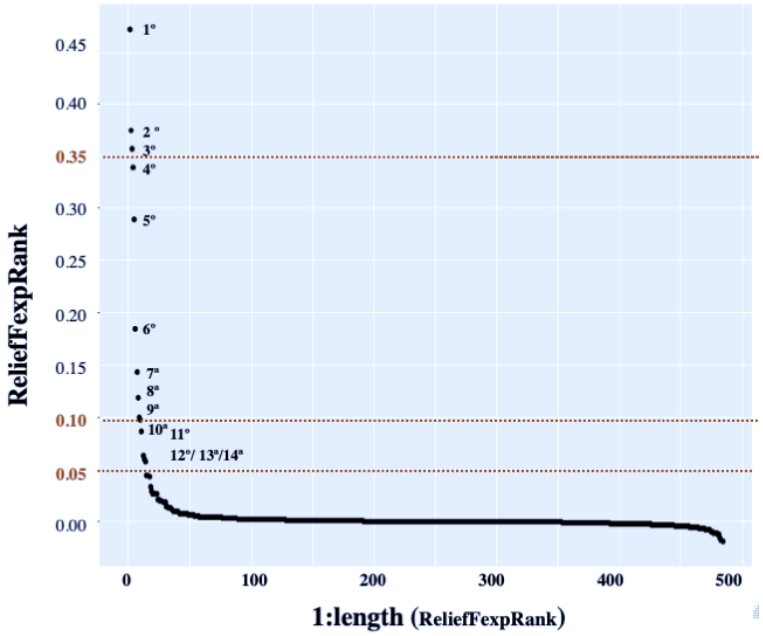
3.4 Results

3.4.1 Variables selected for the type1/type2 diabetes classification algorithm

The variables were ranked on the basis of their ReliefFExp score (**Figure 50**). The variable with the highest score was the number of reimbursements for fast-acting insulin, followed by the number of reimbursements for long-acting insulin and the number of reimbursements for biguanides. Their scores were higher than 0.35.

Other eleven variables had scores higher or equal to 0.05. They included variables related to 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) and out-of-hospital biological tests (glucose, microalbuminuria and prostate-specific antigen tests), variables related to hospitalizations associated with diabetes (total number of hospitalizations and number of hospitalizations with a duration between 1 and 7 days) and age. The mean in the type 2

group from the final dataset was greater than the mean in the type 1 group only in the four following variables: the number of reimbursement of biguanides, out-of-hospital glucose tests, prostate-specific antigen measurement and age.



Rank	Variables	Highest mean
1º	No. reimb. of fast-acting insulin in the last 12 months	T1D
2º	No. reimb. of long-acting insulin in the last 12 months	T1D
3º	No. reimb. of biguanides in the last 12 months	T2D
4º	No. reimb. of test strip for blood glucose test for self-monitoring in the last 12 months)	T1D
5º	No. reimb. of test strip for blood prothrombin for self-monitoring in the last 12 months	T1D
6º	No. reimb. of glucose test kits for self-monitoring in the last 12 months	T1D
7º	Age	T2D
8º	No. of hospitalizations related to diabetes in the last 24 months	T1D
9º	No. reimb. of devices for blood glucose test for self-monitoring in the last 12months	T1D
10º	No. reimb. of prostate-specific antigen screenings in the last 12 months	T2D
11º	No. reimb. of test strip for urine glucose and ketone bodies test for self-monitoring in the last 12 months	T1D
12º	No. reimb. de screening test for blood glucose in the last 12 months	T2D
13º	No. reimb.of screening tests for microalbuminuria in the last 12 months	T1D
14º	No. of hospitalizations related to diabetes with a duration 1 to 7 days in the last 24 months	T1D

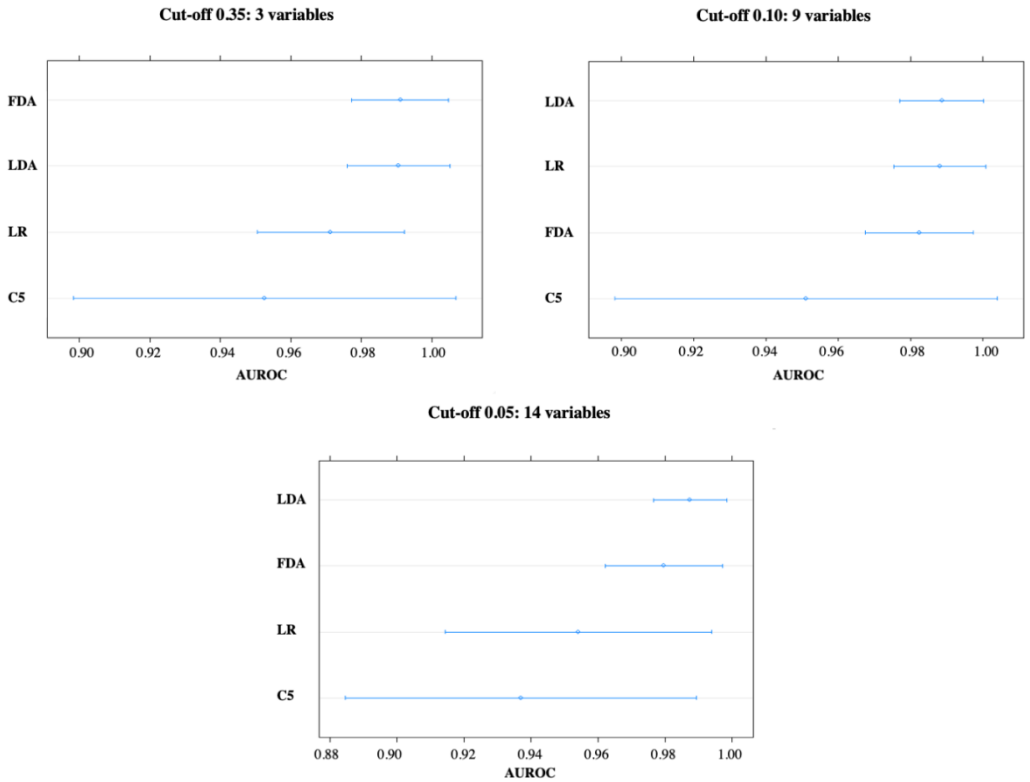
No.:number; reimb.: reimbursement; T1D: type 1 diabetes: TD2: type 2 diabetes

Figure 50. Variable selection for developing the type1/type 2 classification algorithm based on their ReliefFexp Score using three different thresholds (0.35, 0.1 and 0.05)

3.4.2 Validation of trained algorithms

In **Figure 51**, the results of the k-fold cross validation using the testing dataset is represented, more specifically the Area under the Receiver Operating Characteristics

Curve (AUROC)



LDA: Linear Discriminant Analysis; FDA: Flexible Discriminant Analysis; LR: Logistic regression; C5: C5 decision tree; AUROC: area under the ROC curve

Figure 51. Results of k-fold cross validation of different type1 /type 2 classification algorithms from training data set

The upper limit of six out of twelve trained algorithms included the value 1: three LDA (with 3, 9 and 14 variables), two FDA (with 3 and 14 variables) and one LR (9 variables). Regardless of the number of variables, C5 algorithms presented the lowest AUROCs. However, 95% confidence intervals of the twelve AUROCs overlapped.

Table 14 shows that the LR algorithm with 9 variables had less sensitivity than the other algorithms. Specificities and accuracies were above 93%. There were slight differences in the F1 score and the Kappa coefficients between algorithms, but in general LDA and FDA models had the highest values.

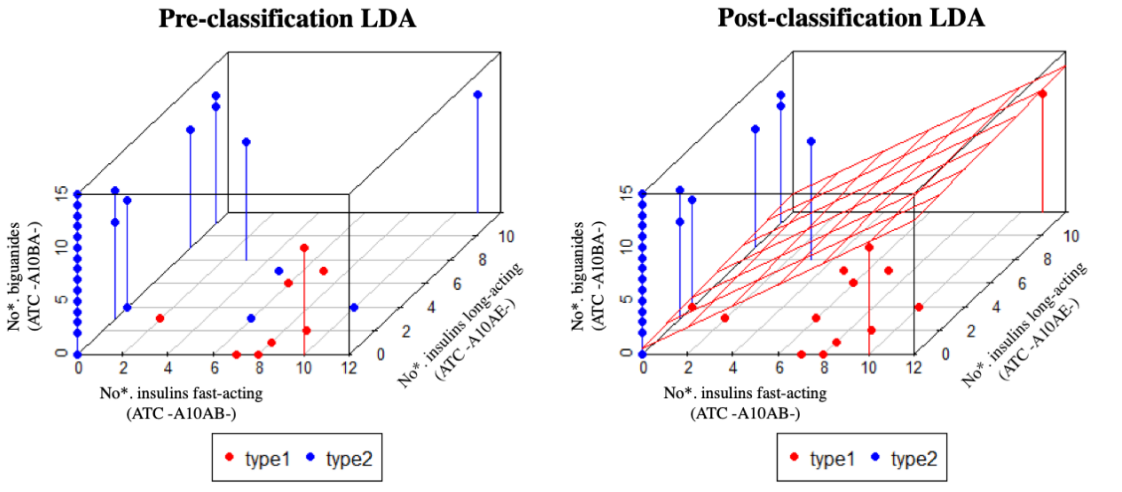
Table 14. Results of validation of twelve type 1/type 2 classification algorithms (three different thresholds of ReliefExp score for variables with four models)

		Acc	Sens	Spec	K	F1
Threshold: 0.35 (3 variables)	**LDA**	97,40%	100%	97,20%	0,8	0,8
	LR	95,20%	100%	95,00%	0,6	0,7
	FDA	96,80%	100%	96,70%	0,7	0,8
	C5	93,70%	100%	93,30%	0,6	0,6
Threshold: 0.1 (9 variables)	**LDA**	97,90%	100%	97,80%	0,8	0,8
	LR	94,20%	88,9%	94,40%	0,6	0,6
	FDA	97,40%	100%	97,20%	0,8	0,8
	C5	96,30%	100%	96,10%	0,7	0,7
Threshold: 0.05 (14 variables)	**LDA**	97,90%	100%	97,80%	0,8	0,8
	LR	95,20%	100%	95,00%	0,6	0,7
	FDA	97,40%	100%	97,20%	0,8	0,8
	C5	94,70%	100%	94,40%	0,6	0,6

LDA: Linear Discriminant Analysis; FDA: Flexible Discriminant Analysis; LR: Logistic regression; C5: C5 decision tree; Acc: accuracy; Sens: sensitivity; Spec: specificity: K kappa coefficient; F1: F1 score

3.4.3 The selected type 1 / type 2 classification algorithm

The selected algorithm was the LDA algorithm based on the number of reimbursements of fast-acting insulin, long-acting insulin and biguanides (**Figure 52**). Indeed, compared to the other algorithms which had high performances in distinguishing between type 1 and type 2 diabetes cases, this algorithm also had the highest parsimony and applicability to further health administrative databases.



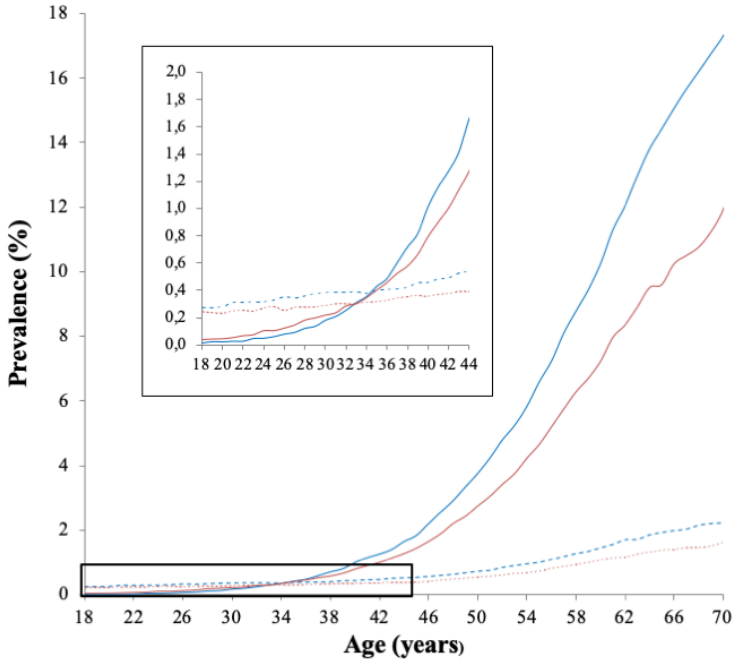
No. Number of reimbursement over the last 12 months

Figure 52. Selected algorithm Linear discriminant analysis (LDA)with 3 variables (Relief Exp Score for variables selection of 0.35)

3.4.4 Prevalence of type 1 and type 2 diabetes in France in 2016

The number of people pharmacologically treated for diabetes aged from 18 to 69 years in 2016 was 1,844,329 (after excluding 7,248 pregnant women). The prevalence of type 1 and the prevalence of type 2 is represented by sex and 1-year age group in **Figure 53**. In both gender, the curve of prevalence of type 1 diabetes was above the one of prevalence of type 2 diabetes from 18 until 32-34 years when the prevalence of type 2 diabetes increased sharply with age. The curve for type 1 was more elevated among men across all age groups, while the curve of type 2 diabetes was higher in women until the age of 32 years. Then there was a shift and men started to have higher prevalence rates than women.

The percentage of type 1 cases among all diabetes cases was 6.9%, after adjusting for the PPV and NPV of the algorithm. In 2016, the prevalence of type 1 diabetes in France 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).



Men type 1: blue dot line; men type 2 solid blue; women type 1: red dot line; women type 2 red solid line

Figure 53. Distribution of type 1 and type 2 diabetes prevalence (%) in France among adults aged 18 to 70 years by sex and age

3.5 Discussion

Through innovative methodology based on SML, we developed a parsimonious algorithm to differentiate type 1 and type 2 diabetes cases based on the number of reimbursements of three antidiabetic drugs: fast-acting insulin, long acting insulin and biguanides. This algorithm was applied in the entire SNDS for the study of the prevalence

of type 1 and type diabetes in France.

Data on type 1 diabetes prevalence in adults population are scarce. The observed rates in our study were similar to those described in the UK and in the US but the prevalence of type 2 diabetes in France was lower, especially compared to the US (8.5%) [51, 194].

3.5.1 Variables selection: from 3,481 to 14 variables

Almost all of the variables selected for constituting the twelve algorithms were expected because they were correlated with type 1 or type 2 diabetes treatment and follow-up. The three features with the highest ReliefExp scores were related to diabetes treatment. As we have seen in the introduction (**See page 34**) , type 2 diabetes is usually treated with biguanides like metformin while the most common treatment for type 1 diabetes is a combination of fast-acting and long-acting insulin [195]. Another group of variables expected was the variables associated with devices for self-monitoring of glucose levels such as test strips for blood glucose or for urine glucose since they are more frequently used by type 1 diabetes cases. The hospitalization variables were also discriminant because type 1 individuals are more likely to be hospitalized, especially because they experience more usually acute complications than those with type 2 diabetes [22]. Two selected variables were related to screening test for follow-up: the number of reimbursements of tests for the urinary albumin excretion rate and the number of reimbursements of out-of-hospital glucoses test. The mean of last variable was higher in the type 2 group, because conversely to type 1 cases, they are less likely to self-monitor blood glucose levels [196].

Nevertheless, certain highly discriminant variables were unexpected. One of them was the number of reimbursements of prostate-specific antigen test, which also had higher mean in the type 2 group compared to type 1. Its discriminant ability may be explained because this type of screening is usually recommended for older men—a group more likely to have type 2 diabetes [196]. The other variable was the number of reimbursements of test strips for self-monitoring of blood prothrombin, which had higher mean in the type 1 group. Since type 1 individuals are more concerned about self-monitoring, they might be more likely to self-monitor blood characteristics related to cardiovascular complications [197].

3.5.2 The applicability of the type 1 / type 2 classification algorithm in other HAD

One of the criteria for selecting the classification algorithm was its applicability to HAD from other countries. The final algorithm is highly applicable because the

information on reimbursement of antidiabetic drugs is available in many countries. This fact is relevant because any HAD containing information on reimbursements of dispensed drugs could be exploited as the main source for type 1 diabetes surveillance or used to complete the information from other sources like disease registries.

Likewise, most of the treatment guidelines in Europe and in the US are coherent with the selected algorithm [195, 198]. These guidelines recommend metformin monotherapy as starting pharmacological treatment for type 2 diabetes cases and multiple daily injections of fast-acting insulin with meals combined with daily basal insulin.

However, the algorithm presented also certain limitations. First, it was develop using a dataset composed by diabetes cases aged from 18 to 70 year and not likely to present severe stages of the diseases. As we have seen in the introduction section, type 2 severe cases are usually treated by insulin, however the combination of fast-acting with long acting is more common in type 1 treatment while type 2 diabetes cases usually are treated by intermediate acting insulin [38]. Secondly, we need to adjust for NPV and PPV of the algorithm, but maybe these values are not stable across all age groups though the algorithm requires to be validated stratifying the results by age group and gender as we have done in the section 5.1.

3.6 Conclusion

The type 1/ type 2 classification algorithm developed through SML allowed to evaluate the prevalence of type 1 diabetes among adults in France for the first time. This algorithm has good performances in distinguishing type 1 and type 2 diabetes cases. Furthermore, it can be used to identify type 1 and type 2 diabetes cases in HAD from other countries.

The SML methodology opens new perspectives in surveillance and can be applied for developing algorithms with different targets such as undiagnosed diabetes cases or prediabetes cases. These applications are presented in the following section.

Further details on the development of the type 1 / type 2 classification algorithm can be found in the article submitted to Diabetes Care entitled: "Artificial intelligence for diabetes research: development of type 1/ type 2 classification algorithm and its application to surveillance using a nationwide population-based medico-administrative database in France" **(See Annex IV: Article 3)**

4. Development of an algorithm to identify undiagnosed diabetes cases and an algorithm to identify prediabetes cases

4.1 Introduction

Studies on the prevalence of undiagnosed diabetes and prediabetes are usually based on National Surveys. In section 3 of the introduction (**See page 55**), we cited the limitations of these data sources including small sample size preventing its use for calculating prevalence rates at subnational level, high cost or lack of access to certain groups of population. These limitations could be overcome if the analyses were based on HAD such as the SNDS. However, no laboratory test results are recorded in the SNDS, so it is impossible to estimate these prevalence rates using a direct approach.

The SML presented in the previous section can be applied to develop algorithms to exploit the SNDS for assessing the prevalence of undiagnosed diabetes and prediabetes in France (**Figure 54**).

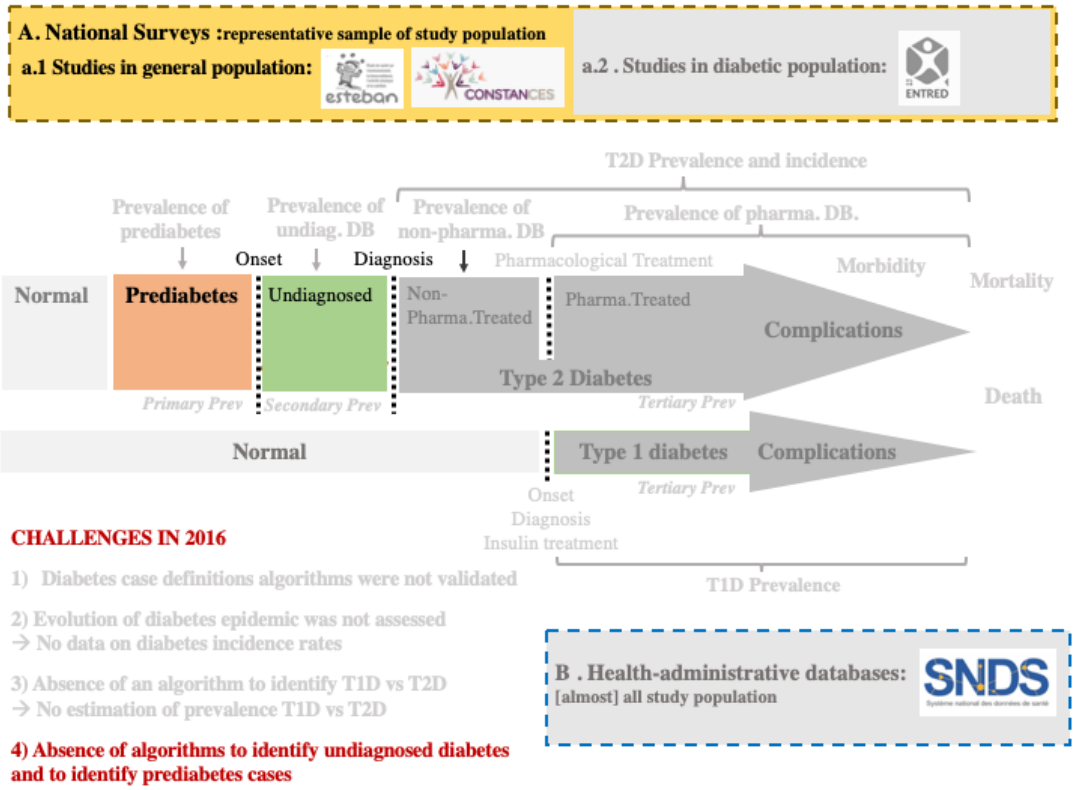


Figure 54. Challenges faced in the results' section 4

4.2 Objectives

The objectives of this section were to develop an algorithm to identify undiagnosed diabetes cases and an algorithm to identify prediabetes cases in the SNDS using the SML methods introduced in the previous section.

4.3 Methods

In the **central core** of the thesis baseline method (**See page 88)** , the undiagnosed diabetes cases and the prediabetes cases were identified in the CONSTANCES population using a decision tree based on linked information from the self-administered questionnaire, the medical examination and the results of the FPG measurement (**Figure 55**).

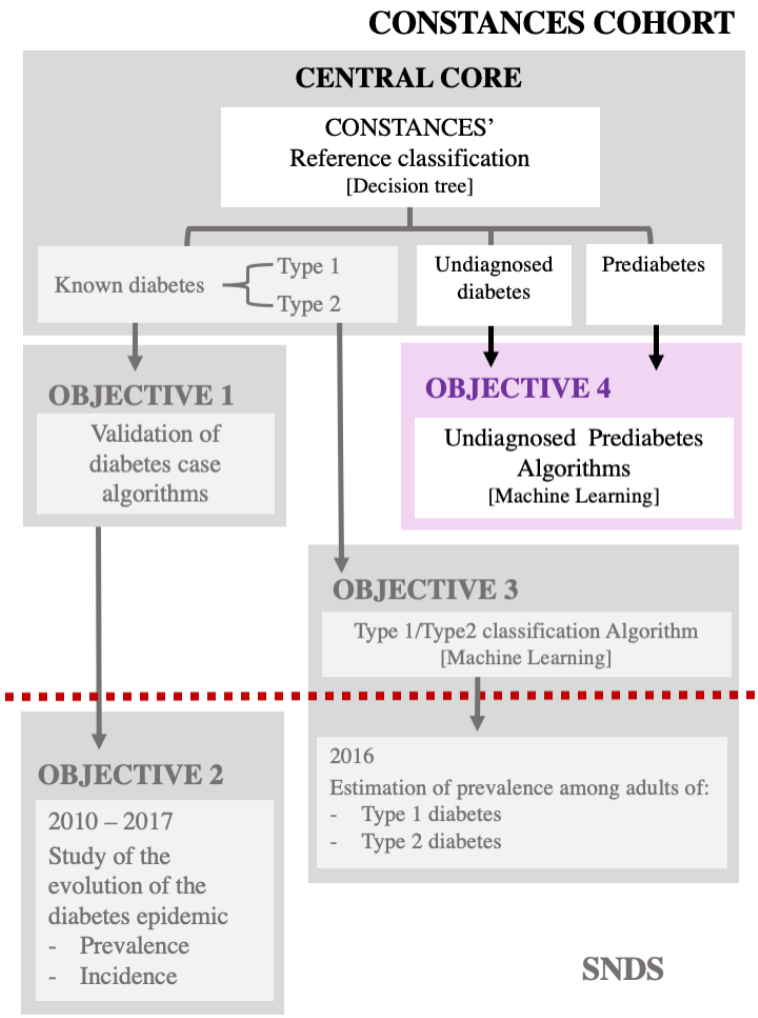


Figure 55. Methods the results' section 4

They were used to define the final dataset and to characterize the target in the SML method. In the development of each algorithm, we used the same methodology with different final datasets, target classification, variables selection and algorithm trained (**See page 127**).

4.3.1 Selection of the final datasets

The final dataset to develop the algorithm to identify undiagnosed diabetes cases was composed by the individuals from the CONSTANCES population not classified as having "known diabetes" and having data on FPG measurement. For developing the

algorithm to identify prediabetes cases, we used the former final dataset after having excluded all the undiagnosed diabetes cases.

4.3.2 Target classification

The target of both algorithms was defined in the central core stage of the thesis methodology (**See page 93**). The target 1 of the first algorithm was unknown diabetes (FPG equal or higher than 7 mmol/l) and target 0 individuals with a FPG lower than this value. For the prediabetes algorithm, target 1 group was composed by individuals from the final dataset having a FPG equal or higher than 6.1 mmol/l, and target 0 group by individuals with FPG level < 6.1 mmol/l.

4.3.3 Variables selection

The 3481 SNDS variables coded for the analyses of the previous section (number of reimbursements of out-of-hospital dispensed healthcare, number of hospitalizations and sociodemographic variables) were also used in the development of undiagnosed and prediabetes algorithms. The Reliefexp score for each variable was estimated, assessing their capability to distinguish between target 1 and target 0 and then they were ranked.

For the selection of the variables comprising the different algorithms to identify undiagnosed diabetes cases, the threshold established was a ReliefFExp score higher than 0.005 while for the prediabetes algorithm the threshold was 0.002.

4.3.4 Algorithms trained

Twelve algorithms to identify undiagnosed diabetes cases were trained and validated: four models (LDA, LR, FDA or C5) with 3, 5 and 12 variables (corresponding to 0.015, 0.010 and 0.005 thresholds, respectively).

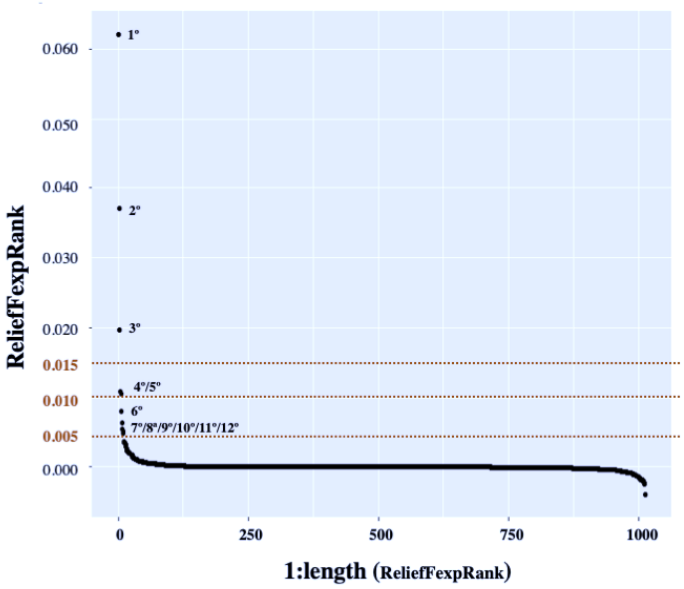
Since the variable ReliefFExp score for prediabetes target was very low, only two thresholds were applied: 0.005 with 6 variables and 0.002 with 16 variables. Therefore eight algorithms were trained and validated.

Finally, as we have described in the previous section, the selection of the most suitable algorithm to identify undiagnosed cases and the algorithm to identify prediabetes cases was based on three criteria: performances, computational parsimony and applicability to further databases.

4.4 Results

4.4.1 Undiagnosed diabetes algorithm

Figure 56 shows the selection of the variables included in the different algorithms to identify undiagnosed diabetes cases.



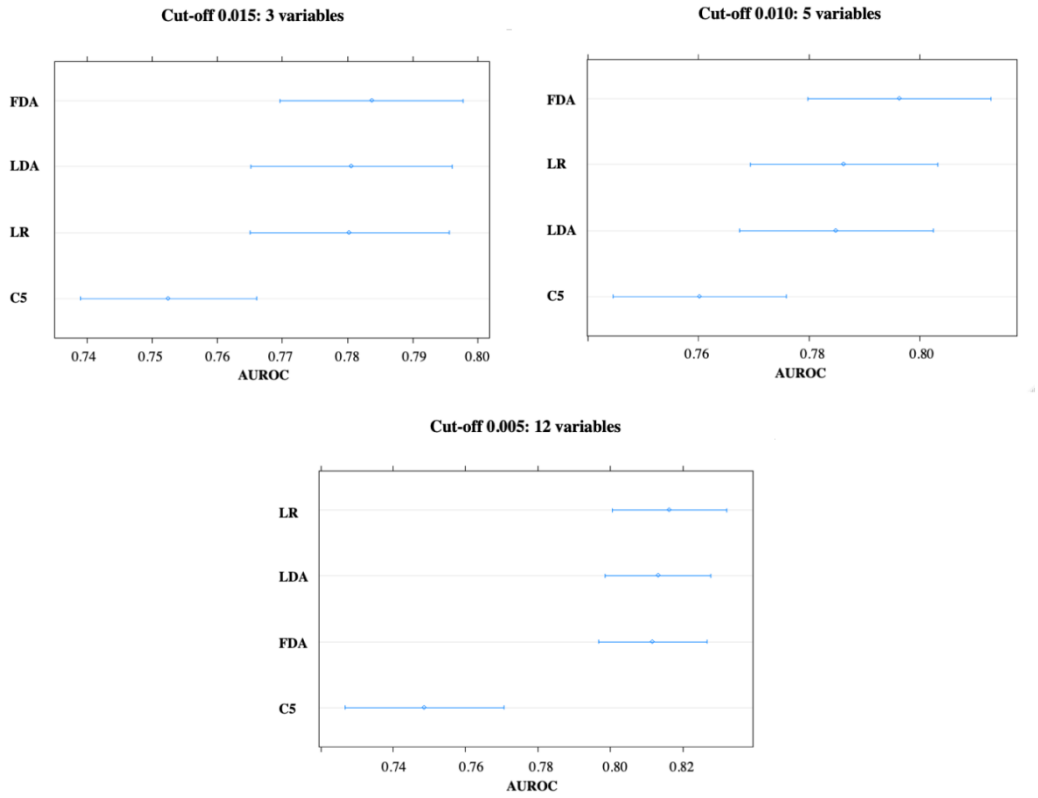
Rank	Variables	Highest mean
1º	Age	Undiag
2º	Sex	Non-Undiag
3º	No. reimb. test for lipid profile in the last 12 months	Undiag
4º	No. reimb. general practitioner consultation in the last 12 months	Undiag
5º	No. reimb. screening tests for glucose in the last 12 months	Undiag
6º	No. reimb. hemmogramme blood test in the last 12 months	Undiag
7º	No. reimb. erythrocyte sedimentation rate blood test in the last 12 months	Undiag
8º	No. reimb. transaminases test in the last 12 months	Undiag
9º	Deprivation index (2009) of commune of residence in the last 12 months in the last 12 months	Undiag
10º	No. reimb. of prostate-specific antigen screenings in the last 12 months	Undiag
11º	No. reimb. electrolytes blood test (sodium + potassium + chlore) in the last 12 months	Undiag
12º	No. reimb. . HbA1c screening test in the last 12 months	Undiag

No.: number of; reimb.: reimbursement of.Undiag: undiagnosed diabetes. Non-undiag: non-undiagnosed

Figure 56. Variable selection for developing the algorithm to identify undiagnosed diabetes cases based on their ReliefFexp Score using three different thresholds (0.015, 0.01 and 0.005)

The variable with the highest ReliefFeExp score was age. The next variable was male gender. Then, different variables related to reimbursement of out-of-hospital biological tests in the last 12 months were observed such as test for lipid profile, screening tests for glucose, hemograms, transaminases, specific antigen screening, electrolytes and HbA1c. Another variable included in the selection was the number of reimbursement for a general practitioner consultation. Finally, the deprivation index of the city/town of residence was also selected as variables with good capability to differentiate between undiagnosed diabetes cases and non-undiagnosed cases. All of these variables had a highest mean in the undiagnosed group.

The results of the k-fold cross-validation of the twelve algorithms based on the training data set showed elevated AUROC with values between 0.74 and 0.80 (**Figure 57**). As we have described in the results of the type 1 / type 2 classification algorithms, the 95% confidence intervals of the twelve algorithms overlapped. After a closer look of these results, we can observe that LR and FDA had the highest values of the AUROC.



LDA: Linear Discriminant Analysis; FDA: Flexible Discriminant Analysis; LR: Logistic regression; C5: C5 decision tree; AUROC: area under the ROC curve

Figure 57. Results of k-fold cross validation of different algorithms to identify undiagnosed diabetes cases from training data set

Regarding the validation with testing dataset, the values of the accuracy, the sensitivity and specificity were moderate (between 59% and 73%). Kappa coefficients and F1 scores were very low, not reaching the value 0.1 **(Table 15)**

The retained algorithm to identify undiagnosed diabetes cases in the SNDS was the LR with 5 variables: age, sex, number of reimbursements in the last 12 months of test for lipid profile and for blood glucose and of general practitioner visits. Its sensitivity was 71% and its specificity was 69%.

Table 15. Results of validation of twelve algorithm to identify undiagnosed diabetes cases (three different thresholds of ReliefExp score for variables with four models)

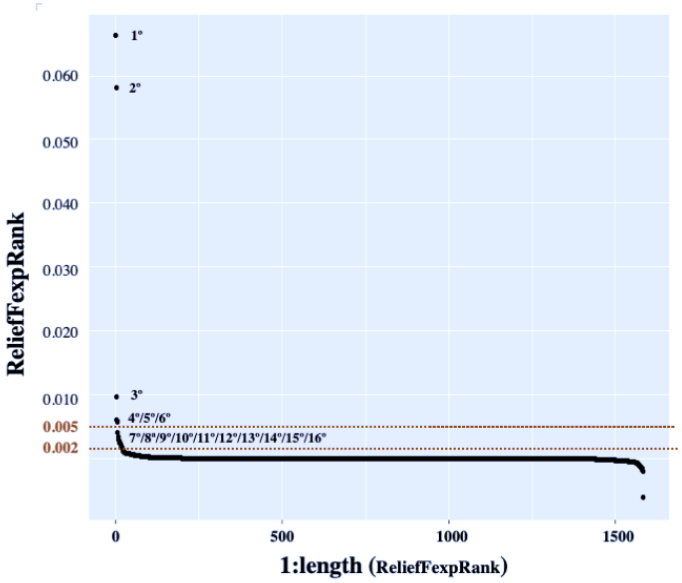
		Acc	Sens	Spec	K	F1
Threshold: 0.015 (3 variables)	**LDA**	64,54%	73,28%	64,40%	0,03	0,06
	LR	66,21%	72,52%	66,12%	0,03	0,06
	FDA	68,88%	67,18%	68,91%	0,03	0,06
	C5	64,54%	73,28%	64,40%	0,03	0,06
Threshold: 0.010 (5 variables)	**LDA**	67,55%	71,76%	67,48%	0,03	0,06
	LR	69,20%	70,99%	69,17%	0,04	0,06
	FDA	72,52%	68,70%	72,58%	0,04	0,07
	C5	71,69%	68,70%	71,73%	0,04	0,07
Threshold: 0.005 (12 variables)	**LDA**	69,07%	71,76%	69,03%	0,04	0,06
	LR	70,86%	66,41%	70,93%	0,04	0,06
	FDA	67,74%	71,76%	67,68%	0,03	0,06
	C5	71,30%	58,78%	71,49%	0,03	0,06

LDA: Linear Discriminant Analysis; FDA: Flexible Discriminant Analysis; LR: Logistic regression; C5: C5 decision tree; Acc: accuracy; Sens: sensitivity; Spec: specificity: K kappa coefficient; F1: F1 score

4.4.2 Prediabetes algorithm

Sixteen variables had ReliefFexpScores above the threshold defined for variables selection of the algorithm to identify prediabetes cases (0.002) (**Figure 58**).

As for undiagnosed diabetes, the variables with the highest score were sex and age. Another variable shared with the previous algorithms is the number of reimbursements for a general practitioner consultation. Most of the selected variables were related to the number of reimbursements of different biological test performed in the last 12 months: specific antigen screening test, HbA1c, lipid profile, Papanicolau, vitamin D, gamma glutamyl transferase or creatinine. Two variables were associated to the number of reimbursements of mammography for screening or for diagnosis. The only variable related to dispensed drugs was the number of reimbursements of influenza vaccines.

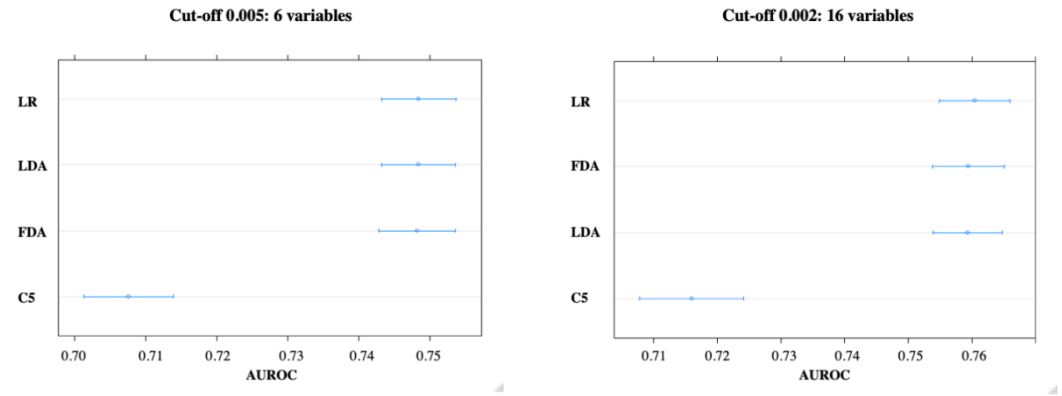


Rank	Variables	Highest mean
1º	Age	Prediab
2º	Sex	Non-Prediab
3º	No. reimb. of prostate-specific antigen screenings in the last 12 months	Prediab
4º	No. reimb. HbA1c screening test in the last 12 months	Prediab
5º	No. reimb. test for lipid profile (EAL) in the last 12 months	Prediab
6º	No. reimb. screening tests for glucose in the last 12 months	Prediab
7º	No. reimb. Influenza vaccine (J07BB) in the last 12 months	Prediab
8º	No. reimb. Papanicolau test in the last 12 months	Non-Prediab
9º	No. reimb. 25-hydroxy vitamin D blood test in the last 12 months	Non-Prediab
10º	No. reimb. blood sample for complete profile in the last 12 months	Prediab
11º	No. reimb. pre-analytical in the last 12 months	Prediab
12º	No. reimb. diagnosis mammograms in the last 12 months	Prediab
13º	No.reimb. gamma glutamyl transferase (GAMMA GT,CGT) test in the last 12 months	Prediab
14º	No. reimb. general practitioner consultation in the last 12 months	Prediab
15º	No. reimb. screening mammograms in the last 12 months	Non-Prediab
16º	No. reimb. creatinine test in the last 12 months	Prediab

No.: number of; reimb.: reimbursement of. Prediab: prediabetes. Non-prediab: non-prediabetes

Figure 58. Variable selection for developing the algorithm to identify prediabetes diabetes cases based on their ReliefFexp Score using two different thresholds (0.005 and 0.002)

Since all the variables presented very low ReliefFExp scores, only two thresholds were applied to define the number of variables used for each model (LDA, FDA, LR and C5), 0.005 including 6 variables and 0.002 including 16 variables. The results of the k-fold cross-validation of the eight algorithms trained are represented in **Figure 59**. The C5 decision tree had the lowest AUROC both using 6 or 16 variables. Then LR, LDA and FDA had similar AUROCs. The models using 16 variables had slightly higher AUROC values. Especially, the LR model had an AUROC between 0.755 and 0.765.



LDA: Linear Discriminant Analysis; FDA: Flexible Discriminant Analysis; LR: Logistic regression; C5: C5 decision tree; AUROC: area under the ROC curve

Figure 59. Results of k-fold cross validation of different algorithms to identify prediabetes cases from training dataset

A second validation was performed using the testing dataset (**Table 16**). The performances of the algorithms in identifying prediabetes cases were modest. Most of the accuracy, sensitivity and specificity values ranged from 63% to 74 % and the most common values for K-coefficient and for F1score were 0.16 and 0.26, respectively.

Table 16. Results of validation of eight algorithm to identify prediabetes cases (two different thresholds of ReliefExp score for variables with four models)

		Acc	Sens	Spec	K	F1
Threshold: 0.005 (6 variables)	**LDA**	67,38%	76,12%	66,65%	0,16	0,26
	LR	68,00%	74,48%	67,46%	0,16	0,26
	FDA	68,74%	74,18%	68,29%	0,16	0,27
	C5	63,92%	75,52%	62,95%	0,13	0,24
Threshold: 0.002 (16 variables)	**LDA**	67,74%	74,78%	67,16%	0,16	0,26
	LR	68,26%	73,73%	67,80%	0,16	0,26
	FDA	69,59%	70,60%	69,51%	0,16	0,26
	C5	74,41%	62,99%	75,36%	0,18	0,27

LDA: Linear Discriminant Analysis; FDA: Flexible Discriminant Analysis; LR: Logistic regression; C5: C5 decision tree; Acc: accuracy; Sens: sensitivity; Spec :specificity

Taking into account performance, computational parsimony and applicability to further databases, we considered the most suitable algorithm was the LR model based on 6 variables: age, sex and number of reimbursements in the last 12 months of specific antigen screening test, HbA1c, lipid profile and glucose.

4.5 Discussion

The SML methodology exposed in the previous section, allowed us to develop two different algorithms based on the SNDS data: an algorithm to identify undiagnosed diabetes cases and an algorithm to identify prediabetes cases. However, the results of the

validation step showed more moderate performances in ascertaining target 1 (undiagnosed cases or prediabetes cases) compared to the performances described for the type 1 / type 2 classification algorithm.

4.5.1 Variable selection

In step 5 of the SML methodology, we selected the variables with the highest capacity in differentiating between target 1 and target 0 *i.e.* their highest ReliefFExp score. Most of the variables selected to develop the undiagnosed diabetes algorithms were associated with health care provided to people at high risk of diabetes. The profile of the undiagnosed diabetes cases is an individual who frequently visits the general practitioner and for whom different biological tests have been prescribed such as lipid profile test, complete hemogram or transaminases test. The number of screening tests for glucose and HbA1c tests in the last 12 months had also elevated ReliefFExp scores suggesting the undiagnosed diabetes cases were already identified as individuals at high risk of developing diabetes by their practitioner and previous test for measuring blood glucose levels had been prescribed.

Five variables were risk factors for diabetes. Deprivation index of the town of residence had high capability differentiating undiagnosed diabetes cases from non-diabetes cases, with a higher mean in the former group. This fact is coherent with the studies showing diabetes prevalence is associated with low socioeconomic status [94]. Other variables were age, sex and the number of reimbursements of specific antigen screening tests which as we have exposed in the previous section is a proxy of the combination of age and sex since this test is usually prescribed in older men, a group at high risk of developing diabetes.

Regarding the variables selected for the prediabetes algorithms, most of them were related to age, sex or both of them. In fact the two variables with the highest ReliefFExp score were age and sex. Then, we found different tests, medical acts or drugs usually prescribed in older populations such as the already presented prostate specific antigen screenings, diagnosis mammograms or influenza vaccine. In the other hand, there were selected variables related to age and sex but with higher means in the non-prediabetes group like the number of reimbursements of Papanicolaou tests, 25-hydroxy vitamin D blood test or screening mammograms. These tests and medical acts are more frequently performed in women of reproductive age, a group at low risk of developing prediabetes [84, 86].

Variables related to the follow up of people at high risk of developing diabetes were also observed in the prediabetes algorithms like the number of reimbursements for lipid profile tests, gamma glutamyl transferase tests, HbA1c tests, screening test for glucose or general practitioner consultations. So, for both algorithms, the profile of cases identified corresponds to people who benefited from medical care and who are not excluded of the care system.

4.5.2 Performances of the undiagnosed diabetes and prediabetes algorithms

The observed performances in the validation of both algorithms were far poorer than those of the type 1 / type 2 classification algorithm. This small ReliefFExp score of the SNDS variables reflected a low capability in distinguishing between target 1 (undiagnosed diabetes or prediabetes) and target 0. We hypothesized these two groups were more similar in terms of care consumptions coded in the SNDS variables than the type 1 and the type 2 diabetes groups of the previous algorithm. A new coding of the SNDS information might be necessary to improve the performances of the algorithms by reducing the window of time. May be the window of time of 12 months for the number of reimbursements of dispensed health care is too wide.

Nevertheless, even if the performances of these algorithms are low, it is important to note that some crucial risk factors for diabetes and prediabetes are not included in the SNDS data, such as BMI and familial history of diabetes [84, 199]. To note, hypertension, high triglycerides and low HDL cholesterol levels are usual determinants of hyperglycemia. However, antihypertensive and lipid lowering treatment were not found as discriminant variables in algorithms. We speculate that people with unknown hyperglycemia might not adhere to treatment, although they seem to be explored and to visit by their general practitioners.

4.6 Conclusion

Applying the SML methodology introduced in the previous section, we succeed in developing an algorithm to identify undiagnosed diabetes cases and an algorithm to identify prediabetes cases. However, both algorithms had lower performances compared to the type 1 / type 2 classification algorithm.

Further analyses are required to develop high-performing algorithms since they represent an excellent opportunity for studying undiagnosed diabetes and prediabetes at national level overcoming the limitations from the national surveys through the exploitation of the SNDS.

SYNTHESIS, PERSPECTIVES AND CONCLUSION

1. Main results

The exhaustive information on healthcare reimbursements and hospitalizations collected from all French population makes the SNDS a valuable source of data for diabetes surveillance. However, there were relevant limitations related to its use, which we aimed to overcome in this work.

First, we assessed the performances of the three diabetes case definition algorithms described by REDSIAM working group in identifying known diabetes cases and pharmacologically treated diabetes cases. We used the data form the CONSTANCES cohort where information recorded in self-administered questionnaires and in medical examinations were linked to the SNDS data. All algorithms presented excellent performances and no relevant differences were observed when analyses were stratified by sex and age group. In light of their limitations and in light of our objectives, we consider the algorithms based on antidiabetic drug reimbursements as the most suitable for the study of the trends of diabetes prevalence and incidence in the whole France, our second objective.

The absence of data on diabetes incidence prevented the assessment of the evolution of the diabetes epidemic in France. A retrospective cohort of diabetes patients between 2010 and 2017 was built with the cases identified in the entire SNDS by applying the case definition algorithm selected in the previous stage. The results of annual diabetes prevalence and incidence rates dynamics were coherent with those observed in other countries like the US or the Nordic countries where the prevalence rates increase slightly and the incidence rates decreased. The decreasing trends on incidence rates were more important in the regions where the highest prevalence rates were observed, the FOT.

One important limitation for diabetes surveillance based on the SNDS was the inability of the case definition algorithms to differentiate between type 1 and type 2 diabetes. That is the reason why we had to limit our population of interest to age 45 years or more when we studied diabetes epidemic in France -limiting thus our results to type 2 diabetes. Therefore, we developed a type 1 / type 2 classification algorithm using SML methods. The algorithm was a LDA model based on the number of reimbursements over the last year of three antidiabetic drugs: fast-acting and long-acting insulin and biguanides. It showed good performances in identifying type 1 diabetes cases and also high applicability to HAD from other countries. Through the application of this algorithm to the entire SNDS, we achieve to study for the first time in France the prevalence of type

1 and type 2 diabetes separately among adults. In 2016, in adults aged 18 to 70 years, the prevalence of type 1 diabetes was 0.32% (0.36% in men and 0.29% in women), and the prevalence of type 2 was 4.36% (5.03% in men and 3.72% in women).

The SML method was applied to develop two more algorithms based on the SNDS data: an algorithm to identify undiagnosed diabetes cases and an algorithm to identify prediabetes cases. The former algorithm was a LR model based on 6 variables: age, sex and number of reimbursements in the last 12 months of specific antigen screening test, HbA1c screening test, lipid profile test and glucose screening test. The prediabetes algorithm was also a LR model, but it used 6 variables: age, sex and number of reimbursements in the last 12 months of specific antigen screening test, HbA1c screening test, lipid profile test and glucose screening test.

2. Research perspectives

New perspectives on the materials and on the tools applied for diabetes research using health administrative databases in France are opened by this work.

A retrospective cohort with more than 4,5 million cases between 2010 and 2017 was constructed to assess the evolution of diabetes epidemic. The cohort could be expanded with further years in order to confirm the trends described in our work. The extension of this study using the same methodology should be possible as the therapeutics included in the algorithm are not likely to change in the next future. Moreover, it is crucial to decline these results by socioeconomic status, in order to assess possible inequalities in the observed trends, as it has been previously described for other indicators. Also, the evolution of mortality and its role in the dynamics of diabetes epidemic could be studied thanks to the data from death certificates recently included in the SNDS. This cohort can be also used for developing prediction models on prevalence and incidence of diabetes which could be useful for developing interventions at national, regional or local level.

The ReDSiam working groups has recorded several algorithms to identify different diseases in the SNDS like hypertension or depression. As shown in our work, the CONSTANCES cohort (comprising actually 200,000 participants) can be an excellent source of data the validation of these algorithms. Also, new algorithms can be developed with CONSTANCES data by using the SML methodology for example an algorithm to identify obese or overweight individuals.

Furthermore, this SML methodology can be applied to other datasets such as the third wave of ENTRED study. We have seen that the ENTRED study is one of the main

sources for diabetes surveillance in France. Its third wave has recruited 13,000 diabetic patients living in France, including an important sample from the FOT. By applying SML methodology to ENTRED data new algorithms to estimate diabetes complication epidemiology could be developed. Indeed, some complications, like retinopathy or diabetes monitoring exams like eye examination are difficult to identify in the SNDS due to the lack of specific acts.

Likewise, The ENTRED data could be used to validate the type 1/type 2 classification algorithm. The algorithm's performances could be evaluated by age and gender ,for example , to make sure that they can be applied in younger age groups. The algorithm could also be evaluated in a type 2 diabetic population with severe complications which lead to an intensified treatment with insulin, because we have seen that these profile of patients are less likely to be included in a generalist cohort like CONSTANCES.

Finally, we aimed to develop two algorithms for identifying undiagnosed and prediabetes cases. They need to be improved since they could be very useful in prevention programs, not only to assess the prevalence of undiagnosed diabetes or prediabetes but also to characterized target populations.

3. Conclusion

Diabetes surveillance in France has achieved new milestones in recent years thanks to HAD like the SNDS. Our work tries to go beyond the limitations imposed by classical methods and to open new horizons in epidemiological research through the improvement and the development of tools for exploiting Big Data sources.

ANNEXES

Annex I: Résumé en français

1. Introduction

1.1 Le diabète

Le diabète est une maladie chronique causée par un dysfonctionnement du métabolisme de l'insuline, l'hormone responsable de la régulation du glucose dans le sang [3].

1.1.1 Types du diabète

Les deux principaux types de diabète en termes de fréquence sont le diabète du type 1 et le diabète du type 2 [5] :

- le diabète de type 1 est causé par la destruction des cellules ß du pancréas due à une combinaison de facteurs génétiques et environnementaux. Il représente de 5 à 10% de l'ensemble des cas de diabète,

- le diabète de type 2 est dû à une résistance des cellules à l'action de l'insuline associée à une détérioration de la fonction des cellules ß. Les principaux facteurs de risque du diabète de type 2 sont l'âge et les antécédents familiaux mais également des facteurs liés au mode de vie. Entre 90 et 95% de l'ensemble des cas de diabète correspond à un diabète de type 2. Le diabète de type 2 est précédé d'une période de prédiabète caractérisée par une glycémie élevée. Les personnes ayant un prédiabète sont à haut risque de développer un diabète et également des maladies cardiovasculaires.

1.1.2 Symptomatologie et diagnostic du diabète

Les principaux symptômes du diabète de type 1 sont la polyurie, la polydipsie et l'asthénie. Le diabète de type 1 survient de façon aigüe avec des symptômes qui apparaissent au cours des jours ou des semaines précédents [6]. En revanche, le diabète de type 2 est parfois asymptomatique et il est fréquemment diagnostiqué au cours d'un dépistage fortuit.

Le diagnostic du diabète peut être établi à partir de la mesure de la glycémie [16] :

a) Après au moins 8 heures de jeûne (glycémie à jeun (GAJ)) :
   Deux résultats de GAJ supérieurs ou égaux à 7 mmol/l (≥126 mg/dl)
b) Non à jeun :
   Un résultat supérieur ou égal à 11.1 mmol/l (≥ 200 mg/dl) associé à des symptômes
c) À jeun puis deux heures après l'ingestion de 75 g de glucose (test d'hyperglycémie provoquée par voie orale) :
   Un résultat supérieur ou égal à 11.1 mmol/l (200 mg/dl) de la glycémie à deux

heures post-charge.

Dans certains, une hémoglobine glyquée (HbA1c) supérieure ou égale à 6,5% définit également un diabète mais en France, la Haute Autorité Santé (HAS) recommande un dépistage du diabète basé sur la glycémie, l'HbA1c étant recommandée comme examen de suivi uniquement.

Il existe deux types de prédiabète :

- l'hyperglycémie modérée à jeun : le critère de l'Organisation Mondial du Santé (OMS) – GAJ entre 6.1 et 6.9 mmol/l (110 to 125 mg/dl) - et le critère de l'Association Américaine du Diabète (American Diabetes Association, ADA) – GAJ entre 5.6 et 6.9 mmol/l (100 to 125 mg/dl).

- l'intolérance au glucose : glycémie 2 heures post charge entre 7,8 et 11,0 mmol/l (140 et 199 mg/dl).

1.1.3 Complications liées au diabète

Les personnes diabétiques non prises en charge sont à haut risque de développer des complications. Les complications liées au diabète sont groupées en deux types : les complications aigües et les complications chroniques. Dans le premier groupe se trouvent des complications causées par une baisse ou une hausse très rapide de la glycémie, comme l'acidocétose diabétique ou le syndrome d'hyperglycémie hyperosmolaire. Les complications chroniques sont provoquées par une hyperglycémie à long terme [24, 32]. Elles sont classifiés en complications microvasculaires (rétinopathie, neuropathie, neuropathie, néphropathie et plaies du pied) et complications macrovasculaires (infarctus ou accident vasculaire cérébral).

1.1.4 Prise en charge du diabète

La prise en charge du diabète comprend une intervention sur le mode de vie, le recours au traitement pharmacologique et la surveillance biologique et médicale.

Le traitement de première intention du diabète de type 2 et la prévention des complications passent par un régime alimentaire plus sain, une activé physique régulière et la diminution de la consommation d'alcool et de tabac [27]. Les recommandations de la HAS pour le traitement pharmacologique du diabète de type préconisent de débuter par une monothérapie par metformine, un antidiabétique oral de la famille des biguanides [38]. En cas de non atteinte des objectifs glycémiques, l'intensification du traitement passe par une bithérapie puis une poursuite de l'intensification jusqu'à l'instauration d'une combinaison d'insulines d'action intermédiaire et de longue durée.

Le traitement par insuline est vital pour les personnes atteintes d'un diabète de type

1 [37]. Les principaux types d'insuline sont [41] : insulines et analogues d'action rapide, insulines et analogues d'action lente et insulines et analogues d'action intermédiaire.

Le contrôle glycémique des personnes diabétiques de type 1 ou de type 2 est établi à partir des dosages d'HbA1c. La HAS recommande un dosage trimestriel. Il est également recommandé la mise en place d'un autocontrôle glycémique pour les personnes insulinotraitées. Cet autocontrôle est effectué à partir de lecteurs de glycémie de sang capillaire (obtenu l'extrémité d'un doigt grâce à un autopiqueur) ou des appareils de mesure du glucose interstitiel [45].

Un suivi biologique et médical régulier est également recommandé afin de suivre l'évolution de la maladie [47]. Les examens médicaux incluent des bilans biologiques (HbA1c, bilan rénal et bilan lipidique), des bilans cardiovasculaires, des examens de la rétine ainsi que des bilans podologiques et bucco-dentaires.

1.2. Épidémiologie descriptive du diabète

En 2017, la Fédération Internationale du diabète (International Diabetes Federation, IDF) a estimé que 451 millions de personnes étaient diabétiques dans le monde [49]. Ce chiffre n'a pas cessé d'augmenter depuis les premières estimations faites en 2000 et cela et dû principalement au vieillissement de la population, à l'augmentation de la prévalence des facteurs de risque (obésité ou sédentarité), à l'augmentation de l'espérance de vie des diabétiques ou une meilleure disponibilité des données [50].

1.2.1 Épidémiologie descriptive du diabète de type 1

Un gradient Nord-Sud a été décrit en Europe, avec des taux de prévalence plus faibles dans les pays méditerranéens comme l'Italie ou la Grèce et des taux plus élevées dans les pays du Nord comme la Norvège, le Finlande ou le Danemark [58]. Par exemple, l'incidence ajustée sur l'âge chez les enfants en Finlande est de plus de 30/100,000 personnes-années [57]. L'incidence du diabète est plus important chez les enfants et les adolescent, plus de 75% des cas sont diagnostiqués avant 18 ans [52]. Les données d'incidence chez l'adulte sont plus rares [56]. Le ratio homme/femme est de 1.47 ce qui diffère du ratio fréquemment retrouvé dans les maladies auto-immunes qui sont généralement plus fréquentes chez les femmes. Différentes études ont observé une augmentation de l'incidence du diabète de type 1 entre 2,4 et 3,4% au cours des dernières décennies [52]. Parmi les hypothèses évoquées, l'hypothèse hygiéniste est fréquemment avancée ainsi que l'exposition à certaines infections virales ou certains facteurs environnementaux [61].

1.2.2 Épidémiologie descriptive du diabète de type 2

Comme évoqué précédemment, le diabète de type 2 est le plus fréquent (> 90% des cas). Sa prévalence est très élevée dans les pays de la région Pacifique et la région du Moyen-Orient et de l'Afrique du Nord (MENA), où plus de 12% de la population est atteinte [62, 63]. La prévalence du diabète de type 2 est généralement plus élevée chez l'homme, sauf dans certaines régions comme la région MENA ou les Caraïbes où la prévalence est plus importante chez les femmes [65, 66].

La prévalence ne cesse pas d'augmenter depuis le début des années 2000. Différents facteurs y sont associés, des facteurs de risque individuels mais également environnementaux. Par ailleurs, l'évolution de la prise en charge et des pratiques de dépistage peuvent avoir un impact sur l'augmentation de la prévalence [50].

Contrairement au diabète de type 1, le diabète de type 2 survient à un âge plus avancé. Le pic de la prévalence est atteint entre 65 et 85 ans [75]. Néanmoins, l'incidence chez les enfants a commencé à augmenter, particulièrement chez les filles dans les pays à forte prévalence [76, 77]. Après avoir observé une augmentation de l'incidence dans la plupart des pays, récemment un plateau a été observé dans certains pays tels que le Canada, l'Italie, l'Ecosse ou le Royaume-Uni. Une diminution de l'incidence a même été observée aux États Unis, en Norvège, en Israël ou en Suède [54]. Ce changement de dynamique de l'incidence du diabète peut être expliqué par l'efficacité des politiques de prévention primaire sans que l'impact des facteurs associés au dépistage et au diagnostic du diabète ne puisse être écarté [79].

1.2.3 Épidémiologie descriptive du diabète non-diagnostiqué et du prédiabète

Selon les dernières estimations de l'IDF, 49% de l'ensemble des cas de diabète, dans le monde, ne sont pas encore diagnostiqués. Aux États Unis, la prévalence du diabète non diagnostiqué en 2014 était de 5,2%, une chiffre très élevée par rapport aux pays européens ou la prévalence varie entre 1,6 et 2,0% [81] [82-84]. La prévalence est plus élevée chez les hommes et chez les plus âgés [84].

Environ 7,3% de la population mondiale a un prédiabète [49]. Les taux les plus élevés sont observés en Amérique du Nord et dans les Caraïbes (14,1%, prévalence standardisée sur l'âge) mais les comparaisons entre pays sont délicates à cause des différents critères de diagnostic utilisés. En appliquant le critère de l'ADA, la prévalence du prédiabète aux États Unis était de 38% entre 2011 et 2014, alors qu'en Allemagne et au Luxembourg, la prévalence était de 21% et 25%, respectivement [81, 83, 86]. Une étude conduite au Royaume-Uni a estimé une prévalence de 11% entre 2009 et 2013 chez les adultes en utilisant le critère de l'OMS [82]. Cette étude a observé aussi que les

caractéristiques associées au prédiabète étaient un faible niveau socio-économique, le fait d'être un homme, d'être âgé de 75 ou plus et d'être obèse.

1.2.4 Épidémiologie descriptive du diabète en France

En France, la prévalence du diabète traité pharmacologiquement était de 5% en 2016. Comme décrit dans d'autres pays, elle est plus élevée chez les hommes et augmente avec l'âge [93]. Par ailleurs, d'importantes inégalités régionales sont observées, avec des taux très élevés dans les départements et régions d'outre-mer [200] : Martinique, Guadeloupe, Guyane et la Réunion.

Une étude récente a estimé que l'incidence du diabète de type 1 était 19,1 cas pour 100 000 personnes-années, chez les enfants âgés de moins de 15 ans, en France en 2015. Une augmentation annuelle du 4% sur la période 2010-2015 était également observée [96]. Mais aucune estimation de l'incidence dans la population adulte n'est disponible au niveau national.

Les dernières estimations de la prévalence de prédiabète et du diabète non-diagnostiqué datent de 2006 [97]. Ces estimations sont basées sur l'Étude nationale nutrition santé (ENNS) qui portait sur un échantillon représentatif de la population en France métropolitaine âgées entre 18 et 74 ans. Cette étude a rapporté une prévalence du diabète non-diagnostiqué de 1% basé sur une GAJ et une prévalence du prédiabète de 5,6% à partir du critère de l'OMS et de 15.5% en appliquant le critère de l'ADA. Dans cette même étude la prévalence du diabète traité pharmacologiquement et du diabète non traité pharmacologiquement étaient estimées à 3,7% et 0,9%, respectivement.

1.3 Surveillance épidémiologique du diabète

La surveillance épidémiologique a un rôle majeur dans le dispositif de santé publique car elle permet d'estimer des indicateurs nécessaires au déploiement et à l'évaluation des programmes de prévention [105]. La surveillance épidémiologique correspond au recueil systématique, à l'analyse et à l'interprétation des données sur des maladies et ses facteurs du risque.

1.3.1 Sources de données pour la surveillance épidémiologique du diabète

Les sources de données pour la surveillance épidémiologique du diabète peuvent être classées en trois groupes : les enquêtes de santé, les registres de patients et les bases de données médico-administratives.

Les enquêtes de santé sont basées sur des échantillons représentatifs d'une population d'étude [108]. Les données sont collectées en combinant des auto-questionnaires, des questionnaires médicaux ou des examens médicaux. Les données

recueillies incluent des informations sur le diabète et autres maladies, le mode de vie ou le niveau socio-économique [109]. Elles permettent d'établir des estimations au niveau national mais souvent la taille de l'échantillon n'est pas suffisante pour effectuer des déclinaisons à des niveaux géographiques inférieurs (région au département) [104]. Aussi, la qualité de l'information peut être affectée par différents biais comme le biais de participation ou biais de mémoire [110].

Les registres de patients consistent en des listes exhaustives d'individus atteints d'une maladie ou ayant suivi certains actes médicaux [112]. A traves de programmes organisés, les individus sont identifiés et suivis afin de collecter information sur la maladie ou le procès d'intérêt [113]. Les registres constituent une source de données idéale pour estimer l'incidence d'une pathologie [107]. Néanmoins, le recueil d'informations ou l'identification des cas peut être impacté par des pratiques médicales hétérogènes ou l'accès à certains établissements de santé. Les registres de patients ne peuvent pas être utilisés comme la seule source de surveillance transversale des différentes maladies. Par ailleurs, son coût élevé de mise en place et de fonctionnement est une limite majeure [112].

Les bases de données médico-administratives incluent une variété de sources de données qui recueillent de grands volumes d'information à visée autre que de surveillance [118]. Contrairement aux enquêtes, l'information n'est pas recueillie sur un échantillon mais de façon exhaustive sur l`ensemble de la population. Comme exemple de bases de données médico-administratives, peuvent être cités les systèmes de dossiers médicaux électroniques, les systèmes d'enregistrement des données relatives aux naissances et décès, les données de séjours hospitaliers ou les données de remboursements de soins des régimes d'assurance maladie. Grâce à leur volume et à leur exhaustivité, des analyses à un niveau infranational peuvent être réalisées. Par ailleurs, elles permettent l'accès à des informations portant sur groupes de population qui ne sont pas bien représentées dans les enquêtes, comme les personnes atteintes de maladies rares ou les personnes défavorisées. Contrairement aux registres de patients, elles permettent de réaliser une surveillance transversale de plusieurs maladies associées. En outre, le rapport coût-efficacité est généralement plus avantageux [105] [127]. Cependant, elles ont également des limites. Par exemple, le recueil de données sur le mode de vie y est rare. En outre, ces sources n'étant pas à visée de surveillance, l'information d'intérêt est parfois « cachée » au milieu d'un grand volume de données non utiles. Des outils spécifiques pour leur exploitation sont donc nécessaires.

1.3.2 Le système de surveillance du diabète en France

Le système de surveillance du diabète en France est développé par Santé publique France et repose sur trois types de sources : les enquêtes en population générale, les enquêtes en population diabétique et les bases de données médico-administratives [146].

1.3.2.1 Les enquêtes en population générale

Des études en population générale comme l'Etude Nationale Nutrition Santé (ENNS), menée en 2006, ou Esteban, menée en 2015, sont basées sur un échantillon représentatif de la population âgé de 18 à 74 ans et résidant en France Métropolitaine [97, 150]. Elles incluent un auto-questionnaire, recueillant des données sur l'état de santé, le mode de vie, les caractéristiques socioéconomiques, et un examen médical, comprenant un bilan sanguin avec une mesure de la GAJ et l'HbA1c. L'information recueillie a permis d'estimer la prévalence du diabète non-diagnostiqué et du prédiabète en France métropolitaine. Toutefois, la taille de l'échantillon a limité les estimations au niveau national.

1.3.2.2 Les enquêtes en population diabétique

L'étude Échantillon national témoin représentatif des personnes diabétiques [168] a été réalisée en 2001, en 2007 et en 2019 [151]. L'étude inclus un échantillon représentatif des personnes diabétiques traitées pharmacologiquement adultes. Cette étude permet de répondre aux enjeux de surveillance de la mortalité et des complications liées au diabète ainsi que de la qualité de vie des personnes diabétiques [153-155].

1.3.2.2 Base de données médico administratives

Le Système national de données santé (SNDS) est une base de données médico-administratives qui constitue une des principales sources de données pour la surveillance épidémiologique du diabète en France [156]. Depuis la création du SNDS en 2003, la Caisse Nationale d'Assurance Maladie des Travailleurs Salariés (CNAMTS) a collecté, anonymisé, traité et mis à disposition des données sur les remboursements de soins en ville, les hospitalisations et récemment les causes médicales de décès.

Grâce aux données du SNDS, le système de surveillance épidémiologique du diabète a accès à des informations exhaustives, mises à jour et de qualité sur l'ensemble de la population résidant en France (y compris les personnes résidant dans les DROM). En revanche, des données sur les diagnostics non hospitaliers, sur les résultats de tests biologiques et les actes médicaux ou sur les modes de vie ne sont pas accessibles dans le SNDS.

1.3.2.3 La cohorte Constances

CONSTANCES est une cohorte épidémiologique « généraliste » constituée d'un échantillon représentatif d'adultes âgés de 18 à 70 ans résidant en France Métropolitaine (en 2019 200,000 participants étaient inclus) [167]. Les données concernant l'état de santé des participants, leur mode de vie et leurs caractéristiques socio-économiques sont recueillies à partir d'auto-questionnaires et elles sont couplées à des données recueillies lors d'examens médicaux durant lesquels des prélèvements sont réalisés (incluant une mesure de GAJ). Postérieurement, ces données sont croisées avec les données du SNDS.

2. Objectifs de la thèse

2.1. Outils disponibles pour la surveillance du diabète en France

Trois algorithmes de repérage des cas de diabète dans le SNDS ont été recensés par le Réseau pour l'utilisation des Données du Système national des données de santé (ReDSiam) [170]:

- Algorithme A: le cas est positif si la personne bénéficie d'une affection de longue durée (ALD) avec un code CIM-10 (classification internationale des maladies) pour diabète (E-10 ou E14),
- Algorithme B: le cas est positif si la personne a eu un remboursement de médicament antidiabétique –code Anatomique, thérapeutique et chimique - ATC classe A10- à l'exception du benfluorex) à au moins trois dates différentes, ou deux en cas de délivrance d'au moins un grand conditionnement, au cours d'une année calendaire,
- Algorithme C: le cas est positif si la personne répond à au moins une des conditions suivantes: (a) être bénéficiaire d'une ALD diabète dans l'année précédente (b) avoir eu un remboursement de médicament antidiabétiques -code ATC class A10- (à l'exception du benfluorex) au moins dans trois dates différentes, ou deux en cas de délivrance d'au moins un grand conditionnement, au cours des deux années précédentes (c) avoir eu une hospitalisation avec un diagnostic principal (DP) ou relié (DR) de diabète (E10–E14) ou une complication du diabète en DP ou DR (G59.0*,G63.2*, G73.0*, G99.0*, H28.0*, H36.0*, I79.2*, L97, M14.2*, M14.6*, N08.3) et un diagnostic associé de diabète (E10–E14).

2.2 Défis pour la surveillance du diabète en France basé sur le SNDS

Au début de cette thèse en 2016, le système de surveillance du diabète en France était confronté à plusieurs défis.

Tout d'abord, aucun des algorithmes de repérage de cas de diabète dans le SNDS n'était validé. Néanmoins, l'algorithme B a déjà été utilisé pour estimer la prévalence du

diabète en France mais il n'y avait pas d'étude sur la dynamique épidémiologique du diabète du fait de l'absence de données nationales sur l'incidence du diabète.

Par ailleurs, les algorithmes de repérage de cas de diabète ne permettaient pas de différentier le diabète de type 1 du diabète de type 2. Les estimations de la prévalence portaient sur les deux types confondus.

Enfin, l'absence de résultats biologiques et de diagnostics dans le SNDS ne permettait pas le repérage des personnes non diagnostiquées ou prédiabétiques.

2.3. Objectifs de la thèse

L'objectif principal de la thèse était d'améliorer la surveillance épidémiologique du diabète en France à partir des données SNDS, en soulevant les défis précédemment exposés. Plus spécifiquement, les objectifs étaient :

a) D'améliorer les outils classiques de surveillance basés sur le SNDS (i) en validant les algorithmes de repérage de cas de diabète identifiés par le ReDSiam et (ii) en appliquant l'algorithme le plus pertinent pour étudier la dynamique épidémiologique du diabète en France entre 2010 et 2017.

b) De développer de nouveaux outils pour la surveillance du diabète (i) en développant un algorithme de typage du diabète à partir d'une méthode générique basée sur la méthodologie *Machine Learning* et (ii) en appliquant cette méthode pour développer un algorithme de repérage de cas de diabète non diagnostiqué et de prédiabète dans le SNDS.

3. Matériels et méthodes

3.1 Le SNDS

Le SNDS est une des plus grandes bases de données médico-administratives au monde comprenant des données individuelles de santé et de recours aux soins anonymisés sur l'ensemble de la population résidant en France (66 millions de personnes) [156, 161, 201].

Le SNDS est composé de trois principales sources de données:

a) Les données de consommation inter-régimes (DCIR) qui contiennent des informations sur le remboursement de soins dispensés en ville (actes médicaux, examens biologiques, consultations, dispositifs médicaux et médicaments), les diagnostics d'ALD et des données sociodémographiques (âge, sexe, commune de résidence et couverture maladie universelle complémentaire). De 2006 à 2010, seules les données des bénéficiaires du Régime Général (RG) et des Sections locales mutualistes (SLM) étaient accessibles dans le DCIR. Après 2010, la plupart des régimes étaient inclus dans le DCIR

incluant le Régime Social des travailleurs Indépendants (RSI) et la Mutualité Sociale Agricole (MSA). Enfin, les autres régimes (représentant moins de 2% de la population) ont été inclus progressivement dans le DCIR.

b) Le programme de médicalisation des systèmes d'information (PMSI) qui contient les informations médicales et administratives de chaque séjour hospitalier en France comme par exemple les diagnostics principaux, reliés et associés, les actes et certains traitements dispensés pendant le séjour. Ces données sont accessibles depuis 2005 dans le SNDS.

c) La base de données du Centre d'épidémiologie sur les causes médicales de Décès (CépiDC, géré par l'Inserm) qui contient les informations enregistrées dans les certificats de décès (date, lieu et cause de décès et informations sur la personne décédées). En 2018, pour la première fois, les données du CépiDc ont été accessibles dans le SNDS pour les années 2013, 2014 et 2015.

Dans le SNDS, les personnes sont identifiées par un numéro d'inscription unique et anonymisé qui permet de chainer les différentes bases de données.

3.2 La cohorte CONSTANCES

CONSTANCES est une cohorte « généraliste » lancée en 2012 et basée sur un échantillon d'adultes âgés de 18 à 70 ans et résidant en France Métropolitaine [166]. Un des objectifs généraux de la cohorte CONSTANCES est de fournir des informations utiles aux acteurs de santé publique sur l'état de santé de la population française et son recours aux services de soins [167]. Dans ce contexte, cette cohorte constitue l'un des piliers de cette thèse.

Les participants sont sélectionnés parmi les bénéficiaires du RG et des SLM, âgés de 18 à 69 ans et inscrits à l'un des 17 centres d'examen de santé (CES) participants. Après l'inclusion, les participants remplissent un auto-questionnaire. Ensuite, ils ont un examen médical dans le CES de référence. Cela comprend un questionnaire médical, un examen clinique et un analyse de sang (dont une GAJ). Si le participant a donné son accord, l'ensemble de ses données sont appariées avec ses données extraites du SNDS et de la Caisse nationale d'assurance de vieillesse (données sociales et professionnelles) et celles du CépiDC (pour la période précédant leur mise à disposition dans le SNDS).

3.3 Méthodologie de base

La méthodologie de base de la thèse reposait sur une étape centrale suivie de quatre étapes répondant aux différents objectifs de la thèse. Dans l'étape centrale, « la population CONSTANCES » était sélectionnée parmi les participants de la cohorte CONSTANCES

inclus entre 2012 et 2014. Ensuite, différents groupes étaient constitués : les personnes non-diabétiques, prédiabétiques, diabétiques non diagnostiquées et diabétiques diagnostiquées. Dans ce dernier groupe, deux distinctions étaient effectuées : les personnes diabétiques traitées pharmacologiquement vs non traitées pharmacologiquement et les personnes diabétiques de type 1 vs de type 2.

L'étape suivante était la validation des algorithmes de repérage de cas de diabète à partir de la population CONSTANCES en utilisant les catégories de référence «diabète diagnostiqué» et «diabète traité pharmacologiquement» comme références ou *gold-standard*. Une fois l'algorithme le plus pertinent sélectionné, il a été appliqué à l'ensemble du SNDS pour estimer la prévalence et l'incidence du diabète entre 2010 et 2017 afin d'étudier l'évolution de la prévalence et de l'incidence du diabète en France. Finalement, la méthodologie *Machine Learning* a été utilisée sur les données de la population CONSTANCES pour développer des algorithmes de typage du diabète (type 1 ou type 2) et de repérage du diabète non diagnostiqué et du prédiabète dans le SNDS.

3.4 Étape centrale

Nous avons sélectionné la population CONSTANCES et avons constitué différents groupes de référence. Sur l'ensemble des participants de la cohorte CONSTANCES recrutés entre 2012 et 2014, nous avons exclus les femmes enceintes au moment des réponses à l'auto-questionnaire, les femmes ayant déclaré avoir eu un diabète gestationnel dans l'auto-questionnaire, les participants sans donnée d'auto-questionnaire et d'examen médical et les participants sans données SNDS.

Les informations recueillies dans l'auto-questionnaire, dans le questionnaire médical et le résultat de la mesure de GAJ ont été combinés pour classifier les patients. En appliquant les deux premiers arbres de décision, nous avons classé les participants en «non-diabète», «prédiabète», «diabète non-diagnostiqué», «diabète traité pharmacologiquement» et «diabète non-traité pharmacologiquement». Ces deux derniers groupes réunis étaient aussi classés comme «diabète diagnostiqué». Le troisième arbre de décision, basé sur l'algorithme Entred, a utilisé l'information sur l'âge au diagnostic et l'âge à la mise sous insuline pour catégoriser le groupe de diabète diagnostiqué en diabète de type 1 et de type 2.

3.5 La population CONSTANCES

Dans la population CONSTANCES (n=45,739 entre 2012 et 2014), nous avons exclu les femmes enceintes (n=179), celles ayant déclaré avoir eu un diabète gestationnel (n=545), les participants sans données d'auto-questionnaire et d'examen médical (n=14)

et ceux sans données SNDS (n=4477).

L'âge moyen de la population CONSTANCES était de 49 ans et la proportion des femmes était légèrement supérieure à la proportion d'hommes. La plupart des individus étaient originaires de France Métropolitaine, employés ou retraités et leur niveau d'éducation était secondaire ou supérieure. Ils avaient en général un bon état de santé (IMC moyen de 25 kg/m^2, 65% de non-fumeurs et 13% traités par hypertension).

3.6 Catégories de référence

En appliquant les différents arbres de décision, la population CONSTANCES incluse était composée de : 88% de « non-diabète » (n=40,247), 7,2% de « prédiabète » (n=657), 1,4% de « diabète non diagnostiqué » (n=139) et de 2,6% de « diabète diagnostiqué » (n=1,157). Dans le groupe de diabète diagnostiqué, la proportion de traité pharmacologiquement ou non était de 88% et 12% respectivement, tandis que la proportion de diabète de type 1 et de diabète de type 2 était de 95,4% et 4,6% respectivement.

Cette classification était une pièce fondamentale (gold standard) pour la validation des algorithmes de repérage de cas de diabète et pour le développement des algorithmes de typage de diabète et d'identification des cas de diabète non diagnostiqué et de prédiabète.

4. Résultats

4.1 Validation des algorithmes de repérage des cas de diabète dans le SNDS

4.1.1 Contexte

L'objectif de cette étape était d'étudier les performances des algorithmes de repérage des cas de diabète en utilisant deux gold standards - diabète diagnostiqué et diabète traité pharmacologiquement - à partir des données de la cohorte CONSTANCES.

4.1.2 Méthodes

Dans l'étape centrale, nous avons sélectionné et classifié la population CONSTANCES selon deux gold standard: «diabète diagnostiqué» et «diabète traité pharmacologiquement». Nous avons utilisé les données du SNDS de la population CONSTANCES pour appliquer les trois algorithmes de repérage des cas de diabète. Ces cas de «diabète» étaient croisés avec les deux gold-standard pour estimer les caractéristiques de performance des trois algorithmes: sensibilité, spécificité, valeur prédictive positive (VPP), valeur prédictive négative (VPN), coefficient kappa (K) et F1 score. Les résultats étaient déclinés par âge et sexe.

4.1.3 Résultats

4.1.3.1 Gold standard « diabète diagnostiqué »

Tous les algorithmes ont des performances optimales en identifiant les cas de «diabète diagnostiqué». L'algorithme C présentait la sensibilité la plus haute (93,8%) suivi par l'algorithme B et l'algorithme A (85,8% et 73,7% respectivement). L'ensemble des algorithmes avaient des spécificités, VPP et VPN très élevées, spécialement l'algorithme A avec une spécificité de 100% en l'absence de faux positifs. L'algorithme C avait les coefficients kappa et les scores F1 les plus élevés (0,95 et 0,95 respectivement). L'algorithme B présentait aussi un score F1 et un coefficient kappa supérieurs à 0,9. Après déclinaison par âge et sexe, les différences entre les catégories n'étaient pas significatives.

4.1.3.2 Gold standard « diabète traité pharmacologiquement »

La sensibilité pour identifier les cas du « diabète traité pharmacologiquement » des algorithmes A, B et C était de 77,2, 97,3 et 99,3%, respectivement. Les spécificités et VPNs de tous les algorithmes étaient supérieures à 99%. L'algorithme B avait la VPN la plus élevée (97,9%) et l'algorithme C la moins élevée (90,6%) L'algorithme B avait le coefficient kappa et le score F1 les plus élevés, suivi par l'algorithme C et l'algorithme A. Comme observé pour le gold standard « diabète diagnostiqué », il n'existait pas de différences significatives selon le sexe et les groupes d'âge.

4.1.4 Discussion

Nous avons montré que les algorithmes de repérage des cas de diabète (diagnostiqué ou traité pharmacologiquement) utilisés dans le SNDS ont d'excellentes performances.

Pour sélectionner un algorithme de repérage, il est nécessaire de ne pas considérer uniquement les performances des algorithmes mais aussi les objectifs de l'étude. Si l'objectif de l`étude est d'étudier des tendances temporelles ou des inégalités entre régions, les algorithmes qu'utilisent l'information sur ALD-diabète ne sont pas recommandés [170]. En effet, l'information sur l'ALD diabète n'était pas accessible pour certains régimes avant 2014 et sa qualité varie beaucoup selon les régions et le niveau socioéconomique. En revanche, l'algorithme C apparaît le plus indiqué pour étudier des indicateurs de morbidité ou la mortalité car il identifie bien les cas de diabète sévère.

4.2 Évolution de l'épidémie du diabète en France

4.2.1 Contexte

L'objectif de cette étape était d'étudier l'évolution de la prévalence et de l'incidence du diabète en France chez les adultes âgés de plus de 45 ans par sexe, âge et région sur la

période 2010-2017.

4.2.2 Méthodes

D'après les résultats du travail précédent, nous avons sélectionné l'algorithme B basé sur les remboursements de médicaments antidiabétiques comme étant le plus pertinent pour étudier les évolutions temporelles. Nous avons appliqué cet algorithme à l'ensemble du SNDS pour la période 2010-2017. Nous avons construit une cohorte rétrospective à partir de tous les cas de diabète identifiés pour caractériser les cas prévalents et incidents de diabète. Nous avons défini un cas incident comme étant identifié comme diabétique une année donnée mais pas au cours des deux années précédentes. Afin de se focaliser sur le diabète de type 2, les analyses ont été restreintes aux personnes âgées de plus de 45 ans (chez qui la proportion de diabète de type 1 est plus faible qu'avant cet âge).

Nous avons estimé les taux de prévalence et d'incidence pour chaque année en utilisant les données du Institut national de la statistique et des études économiques (INSEE) pour estimer les dénominateurs. Nous avons décliné les résultats par sexe, âge et régions (17 régions en excluant Mayotte). Finalement, nous avons appliqué des modèles négatifs binomiaux, stratifiés par sexe et ajustés sur l'âge et la région pour estimer les taux d´évolution annuelle.

4.2.3 Résultats

En 2017, un total de 3,333,741 cas traités pour un diabète étaient identifiés en France dans le SNDS (1,836,410 hommes et 1,497,331 femmes). Environ 94% de tous ces cas avaient plus de 45 ans. Dans ce groupe, l'âge moyen était de 69 ans et le ratio hommes-femmes était de 1,24.

4.2.3.1. Évolution de l'épidémie de diabète entre 2010 et 2017

Durant la période d'étude, la prévalence brute a augmenté (pour les hommes: de 10,9 à 11,8% et pour les femmes: de 7,9 à 8,4%) alors que l'incidence a diminué (pour les hommes : de 10,7 à 9,6 cas par 1000 personnes-années et pour les femmes de 7,1 à 6,1 cas par 1000 personnes-années). Nous avons observé des résultats similaires après avoir standardisé les taux sur l'âge.

Les modèles multivariés ont montré un taux d´évolution annuelle croissant de la prévalence sur la période 2010-2017 (hommes : +0,9% [IC95% +0,7,+1,0%] et femmes : +0,4% [IC95% +0,2, +0,6%]) et un taux d´évolution annuelle décroissant de l'incidence sur la période 2012-2017 (hommes: -2,6% [IC95% -3,1,-2,0%] et femmes: -3,9% [IC95% -4,5, -3,4%]).

4.2.3.2 Évolution de l'épidémie de diabète entre 2010 et 2017 par âge

Chez les hommes, la prévalence du diabète atteignait son point culminant dans le groupe d'âge 75-79 ans : elle était de 18.5% en 2010 et de 20.4% en 2017. Chez les femmes, le pic était de 13.4% en 2010 pour le groupe d'âge 75-79 ans et de 14.2% pour le groupe d'âge 80-84 ans en 2017. Dans toutes les tranches d'âge, l'incidence était plus élevée en 2012 qu'en 2017. Le point culminant chez les hommes était retrouvé dans le groupe d'âge 65-69 ans en 2012 (14.9 cas pour 1000 individus) et en 2017 (13.2 cas pour 1000 individus). Chez les femmes, l'incidence atteignait son taux le plus élevé pour le groupe d'âge 70-74 ans (9.8 cas pour 1000 individus) alors qu'en 2017 un plafonnement était observé de 60 à 80 ans (7.6 cas pour 1000 individus).

4.2.3.3 Évolution de l'épidémie de diabète entre 2010 et 2017 par région

La prévalence standardisée pour l'âge était plus élevée en 2017 qu'en 2010 chez les hommes, tant dans les régions avec les taux les plus faibles comme la Bretagne (7.3% en 2010 vs 8% en 2017) que dans les régions avec les taux les plus importants comme la Réunion (19.4 vs 19.3%, respectivement). Les taux d´évolution annuelle de la prévalence dans toutes les régions étaient supérieurs à 0%, sauf pour La Réunion où il était de – 0.3% (diminution non significative). La même dynamique était observée chez les femmes, sauf pour la Martinique et la Réunion où la prévalence en 2010 était supérieure à celle en 2017 (de 17 à 16.1% et de 21.9 à 20.2%, respectivement). Les taux d'évolution annuelle confirmaient la diminution significative de la prévalence dans ces deux régions (La Réunion -1.1% et Martinique -0.9%).

L'incidence standardisée par l'âge chez les hommes diminuait dans toutes les régions entre 2012 et 2017, spécialement dans les régions avec les valeurs les plus élevées en 2012, comme la Guadeloupe (diminution de 16.5 à 12.8 cas pour 1000 individus) et la Réunion (diminution de 17.5 à 13.3 cas pour 1000 individus). Les régions avec un taux annuel décroissant plus importantes étaient la Guadeloupe (-3.8%) and la Réunion (-4.4%.). L'incidence standardisée par âge diminuait aussi chez les femmes, spécialement dans les DOM: en Martinique, de 13.3 à 10.8 ; à la Réunion, de 16.4 à 10.7, en Guadeloupe de 16.8 à 12.6 et en Guyane de 21.5 à 15 cas pour 1000 individus.

4.2.4 Discussion

La prévalence du diabète continue d'augmenter en France tandis que son incidence diminue, avec une diminution plus importante dans les DOM.

Une diminution de l'incidence du diabète a déjà été observée dans d'autre pays comme la Norvège, la Suède ou les États Unis [72, 85, 176]. Cette dynamique a été

expliquée par différentes hypothèses. D'un côté, suite aux efforts dans le dépistage du diabète, le pool de personnes non-diagnostiqués a diminué [176]. D'un autre côté, comme l'algorithme appliqué pour identifier les cas de diabète était basé sur le remboursement de médicament antidiabétiques, une augmentation des cas non traités pharmacologiquement pourrait impacter les taux d'incidence. En outre, les actions de prévention, notamment dans les régions à très forte prévalence de diabète, pourraient porter leurs fruits.

En parallèle, les prévalences de diabète non diagnostiqué et de diabète non traité pharmacologiquement semblent ne pas avoir augmenté quand on compare les résultats de l'ENNS en 2006 et les résultats de la cohorte CONSTANCES en 2013 [97, 180]. En France, la comparaison des études ENNS 2006 et Esteban 2014 montre que les taux d'obésité et de surpoids sont restés stables (17% et 49%) [97, 182].

4.3. Développement d'un algorithme de classification du diabète de type 1/de type 2

4.3.1 Contexte

Les objectifs de cet étape étaient le développement d'un algorithme de typage du diabète en utilisant une méthodologie Machine Learning avec un apprentissage supervisé (Supervised Machine Learning , SML) et son application pour estimer la prévalence du diabète de type 1 et du diabète de type 2 chez l'adulte en France.

4.3.2 Méthodes

4.3.2.1 Développement d'un algorithme de typage du diabète

Nous avons développé l'algorithme à partir d'une base de données de référence composée par les participants inclus dans la population CONSTANCES présentant un « diabète traité » .

Nous avons appliqué une méthodologie SML en huit étapes pour le développement de l'algorithme [190] :

1) Sélection de la base de données de référence, composée des participants classifiés comme ayant un «diabète traité»,
2) Identification de la cible en appliquant la classification diabète de type 1 (cible 1) vs de type 2 (cible 0) réalisée dans l'étape centrale à partir de l'arbre de décision Entred,
3) Codification des variables SNDS : le nombre de remboursements de médicaments, consultations, actes médicaux en ville, dispositifs d'auto-surveillance et auto traitement sur les 12 mois qui précédaient l'auto-questionnaire ; le nombre d'hospitalisations ou nombre de jours hospitalisés

dans les 24 mois qui précédauent l'auto-questionnaire et les caractéristiques socio-demographiques (âge, sexe et indice de désavantage social de la commune de résidence - Fdep),

4) Séparation de la base de référence en base d'entrainement (80%) et base de test (20%),
5) Sélection des variables pour constituer les algorithmes à partir de la base d'entrainement. Une fois enlevées les variables avec une variance nulle, nous avons estimé le ReliefExp score de chaque variable SNDS [191, 202]. Le Relief Exp score évalue la capacité des variables à discriminer la cible 1 de la cible 0. Nous n'avons retenu que les variables avec un Relief Exp score supérieur ou égal à 0,005,
6) Entraînement des algorithmes à partir de la base d'entrainement. Nous avons entraîné douze algorithmes: quatre types de modèles (*linear discriminant analysis*, *flexible discriminant analysis*, régression logistique et C5 arbre de décision) avec 3, 9 ou 14 variables (correspondant au trois seuils de ReliefExp score 0,35, 01,1 and 0,05),
7) Validation des algorithmes. Nous avons évalué les performances des algorithmes pour identifier la cible 1 (diabète de type 1) en utilisant la base d'entraînement puis la base de test,
8) Sélection de l'algorithme. Nous avons sélectionné l'algorithme final selon sa performance, sa parcimonie computationnelle et sa transférabilité (ou sa capacité à être utilisée) dans d'autres bases des données médico-administratives.

4.3.2.2 Estimation de la prévalence du diabète de type 1 et diabète de type 2

Nous avons identifié tous les cas de diabète en 2016 dans le SNDS en appliquant l'algorithme basé sur les remboursements de médicaments antidiabétiques après exclusion des femmes ayant accouché en 2016 et des personnes de moins de 18 ans ou de plus de 70 ans. Finalement, l'algorithme de typage a été appliqué pour estimer la prévalence du diabète de type 1 et de type 2 par âge et sexe, et la prévalence totale a été ajustée à partir des VPN et VPP de l'algorithme.

4.3.3 Résultats

L'algorithme sélectionné était un modèle de *Linear discriminant analysis* basé sur le nombre de nombre de remboursements, dans l'année précédente, d'insuline à action rapide, d'insuline de longue durée et de biguanides. Les performances de l'algorithme pour identifier les cas de diabète de type 1 étaient les suivantes : spécificité 97,2 %,

sensibilité de 100% et F1 score de 0,8.

Après avoir exclus les femmes enceintes, le nombre total de cas de diabète traité pharmacologiquement et âgées de 18 à 70 ans était de 1,844,329 en 2016 dans le SNDS. Nous avons décliné les prévalences de diabète de type 1 et de diabète de type 2 par sexe et âge. La prévalence des deux types de diabète était plus importante chez les hommes que chez les femmes, sauf pour le diabète de type 2 dans le groupe d'âge de 18 à 34 ans. Dans les groupes d'âge plus jeunes, jusqu'à 32 ans, la courbe de prévalence de diabète de type 1 était supérieure à celle de diabète de type 2. Ensuite, les courbes étaient inversées.

Le pourcentage de diabète de type 1 parmi l'ensemble des diabètes traités pharmacologiquement était de 6,9%. La prévalence du diabète de type 1 en 2016 en France était 0,32% (hommes 0,36% et femmes 0,29%), et la prévalence du diabète de type 2 était de 4,36% (hommes 5,03% et femmes 3,72%), après ajustement sur la VPN et la VPP de l'algorithme.

4.3.4 Discussion

Nous avons développé, à partir d'une méthodologie SML, un algorithme du typage du diabète. Cela nous a permis d'estimer pour la première fois en France la prévalence de diabète de type 1 et de diabète de type 2 séparément chez les adultes.

L'algorithme de typage est basé sur le remboursement d'insuline à action rapide, d'insuline de longue durée et de biguanides. Ces résultats sont cohérents avec les recommandations de la HAS pour le traitement des deux types de diabète en France [37, 47]. Ces mêmes recommandations de prise en charge thérapeutique sont indiquées par les autres organismes au niveau international comme l'EASD ou l'ADA [195, 198]. Ainsi, cet algorithme n'est pas seulement très performant, il peut être appliqué dans d'autres bases de données médico-administratives utilisées dans d'autres pays.

Néanmoins, cet algorithme présente des limites liées aux caractéristiques de la population CONSTANCES, utilisée pour son développement. En effet, elle n'inclut que des personnes âgées de 18 à 70 ans et ayant probablement une faible probabilité d'avoir un diabète sévère. En outre, la récente prise en charge à 100% des appareils de mesure du glucose interstitiel par les régimes d'assurance de maladie, pour les personnes avec un traitement intensif d'insuline [46], survenue en 2017, pourrait modifier la sélection des variables incluses dans l'algorithme final s'il avait été développé après cette date.

4.4. Développement d'un algorithme de repérage de cas du diabète non-diagnostiqué et d'un algorithme de repérage du cas du prédiabète

4.4.1 Contexte

L'objectif de cette étape était de développer un algorithme de repérage de cas de diabète non diagnostiqué et un algorithme de repérage de cas de prédiabète dans le SNDS en utilisant la méthodologie SML.

4.4.2 Méthodes

Nous avons appliqué la même méthodologie SML basée sur huit étapes. Nous avons tout d'abord sélectionné la base de référence. Pour l'algorithme d'identification des cas de diabète non diagnostiqué, nous avons considéré la population CONSTANCES en excluant les cas de diabète diagnostiqué ; pour l'algorithme d'identification des cas de prédiabète, nous avons, en outre, exclus tous les cas de diabète (diagnostiqué ou non). Ensuite, nous avons utilisé la classification réalisée dans l'étape centrale de la méthodologie de la thèse pour définir les cibles, « diabète non diagnostiqué » et « prédiabète ».

Nous avons ensuite caractérisé les variables SNDS, séparé la base de référence en base d'entrainement et base de test et sélectionné les variables pour constituer les algorithmes. Le seuil du ReliefFExp score pour l'algorithme de diabète non diagnostiqué était de 0,005 et celui pour l'algorithme de prédiabète de 0,002.

Nous avons entraîné douze algorithmes de « diabète non diagnostiqué » selon les quatre modèles (*linear discriminant analysis*, *flexible discriminant analysis*, régression logistique et C5 arbre de décision) avec 3, 5 et 12 variables (correspondant au trois seuils de ReliefExp score 0,015, 0,010 and 0,005). Les algorithmes d'identification du prédiabète avaient un ReliefExp score faible et nous n'avons entraîné que huit algorithmes: les quatre modèles avec 6 et 16 variables (ReliefExp seuils 0,005 et 0,002).

Nous avons validé les algorithmes en utilisant la base d'entrainement et la base de test et avons finalement sélectionné l'algorithme le plus pertinent en tenant en compte de sa performance, sa parcimonie computationnelle et sa transférabilité dans des bases de données différentes du SNDS.

4.3.3Résultats

L'algorithme de repérage de cas de «diabète non diagnostiqué» était un modèle de régression logistique basé sur les variables suivantes: sexe, âge et nombre de remboursements dans les 12 derniers mois d'explorations d'une anomalie lipidique, de consultations de médecins généralistes et dosages de glycémie en laboratoire de ville. La sensibilité, la spécificité et le score F1 de l'algorithme étaient 73,3%, 64,4% et 0,06, respectivement.

L'algorithme retenu pour l'identification des cas de « prédiabète » dans le SNDS

était aussi un modèle de régression logistique utilisant 6 variables : âge, sexe, nombre de remboursements dans les 12 dernières mois d'explorations d'une anomalie lipidique, de dosages d'antigène prostatique spécifique, de dosages d'HbA1c et de dosages de la glycémie dans un laboratoire de ville. La sensibilité, la spécificité et le score F1 de l'algorithme étaient 74.48%, 67.46% and 0,26, respectivement.

4.4.4 Discussion

L'utilisation du SNDS pour étudier la prévalence du diabète non diagnostiqué et de prédiabète permet de contourner les limites imposées par le recours aux enquêtes de santé, avec notamment des tailles d'échantillon ne permettant pas de décliner les analyses par région ou d'étudier des populations particulières. En appliquant la méthodologie SML, nous avons développé deux algorithmes pour identifier les cas de diabète non diagnostiqué et du prédiabète dans le SNDS.

Nonobstant, les performances pour identifier ces cibles sont inférieures comparées aux résultats observés pour l'algorithme du typage du diabète. Ces performances modérées sont probablement liées aux faibles RelifeExp scores, ou pouvoir discriminant, des variables SNDS. Un nouveau codage des variables du SNDS en utilisant une période du temps plus courte, par exemple 6 mois, pourrait améliorer les performances des algorithmes.

5. Perspectives et conclusions

Le SNDS est une source de données majeure pour la surveillance du diabète en France, grâce à l'exhaustivité des remboursements de soins et des hospitalisations de l'ensemble de la population française. Néanmoins, quand cette thèse a commencé en 2016, nous étions confronté à certains défis concernant les outils de surveillance de diabète basé sur le SNDS. Au cours de cette thèse, nous avons surmonté certains de ces défis, en :

- Validant les algorithmes de repérage du diabète,
- Mesurant l'évolution de l'épidémie de diabète, en étudiant en particulier la prévalence et l'incidence du diabète chez les adultes âgés 45 ans ou plus,
- Développant un algorithme de typage du diabète, permettant pour la première fois d'estimer les prévalences du diabète de type 1 et de type 2 chez les adultes,
- Développant des algorithmes de repérage de diabète non diagnostiqué et de prédiabète.

Les résultats obtenus dans cette thèse ont ouverts plusieurs perspectives. D'un côté, la cohorte rétrospective construite avec tous les cas de diabète repérés dans le SNDS entre

2010 et 2017 pourra être enrichie par les cas repérés les années suivantes pour vérifier que les dynamiques observées dans cette thèse persistent. En outre, grâce à l'algorithme du typage, ces dynamiques pourraient être étudiées spécifiquement pour le diabète de type 1 et le diabète de type 2. L'estimation, pour la première fois en France, de l'incidence du diabète pourrait être utilisée aussi pour développer des modèles prédictifs et pour évaluer l'impact de programmes de prévention au niveau national, régional ou local.

Le réseau ReDSiam a identifié plusieurs algorithmes de repérage de différentes maladies comme l'hypertension ou la dépression. Comme montré dans cette thèse, la cohorte CONSTANCES est une excellente source de données pour la validation des algorithmes utilisés dans le SNDS. Nous pensons qu'elle pourrait être utilisée pour développer et valider de nouveaux algorithmes en appliquant la méthodologie Machine Learning utilisée dans cette thèse.

Cette méthodologie Machine Learning peut également être utilisée dans d'autres bases de données appariées au SNDS comme l'étude Entred 3 pour développer des algorithmes d'identification des complications du diabète ou du recours aux soins. Les données d'Entred 3 nous permettrons également de valider l'algorithme de typage développé.

Ainsi, notre travail a inspiré de nouveaux champs d'investigation pour l'épidémiologie en améliorant l'existant et en développant de nouveaux outils pour exploiter les sources de données de type *Big Data* comme le SNDS.

Annex II: Article 1

International Journal of Public Health
https://doi.org/10.1007/s00038-018-1186-3

ORIGINAL ARTICLE

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

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

Received: 18 June 2018 / Revised: 3 September 2018 / Accepted: 26 November 2018

Abstract

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

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

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

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

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

Introduction

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

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

✉ Sonsoles Fuentes
sonsoles.fuentes@santepubliquefrance.fr

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

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

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

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

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

Published online: 04 December 2018

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

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

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

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

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

Methods

The French national health insurance information system

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

Diabetes case definition algorithms

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

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

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

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

The CONSTANCES cohort

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

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

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

Study population

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

Gold standard: 'known diabetes'

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

Gold standard: 'pharmacologically-treated diabetes'

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

Statistical analysis

The three diabetes case definition algorithms A, B and C outlined above were applied and their test characteristics

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

Results

Validation study population

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

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

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

Gold standard: 'known diabetes'

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

Gold standard: 'pharmacologically-treated diabetes'

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

Discussion

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

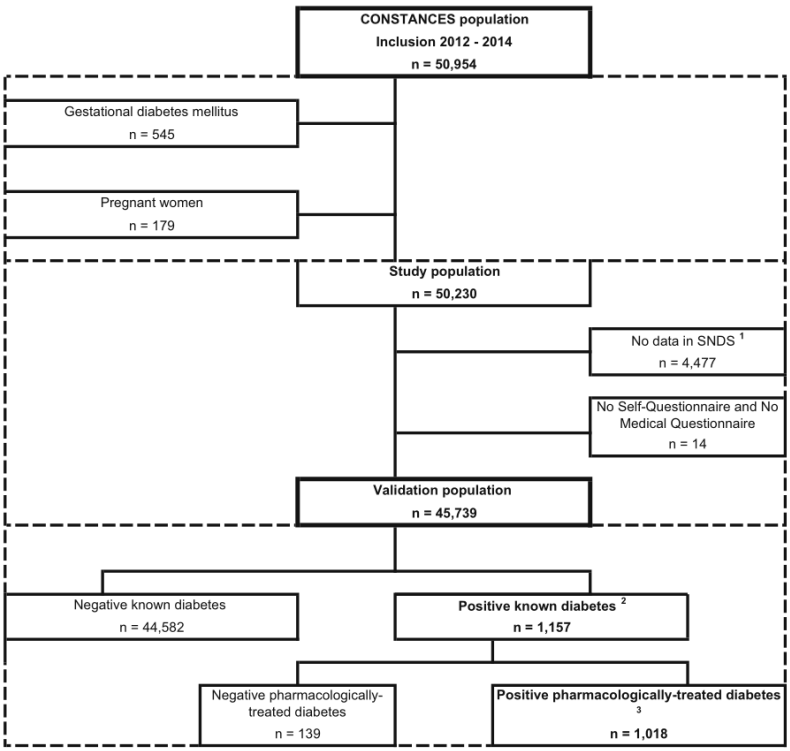


Fig. 1 Flow chart and number of people with known diabetes and with pharmacologically treated diabetes in the validation population. 1 *Système National des Données de Santé* (SNDS) French national health insurance information system, 2 Gold standard "Known Diabetes" and 3 gold standard "pharmacologically-treated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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

The highest sensitivity in both validation analyses was observed in algorithm C, which combined data on ALD-diabetes, drug reimbursements and hospitalization diagnoses. Since its case definition was broader, both a lower number of false negatives and a higher number of false

positives were expected. Because part of this algorithm is based on ALD-diabetes information, the practical considerations of algorithm A must be acknowledged (Fosse-Edorh et al. 2017). Beyond the objectives of the present study, we have validated the different components of algorithm C and their combinations; the results of these supplementary analyses are described in the tables C and D in the EMS. When removing the component related to ALD-diabetes from algorithm C, the test characteristics are similar to those of algorithm B. In addition, algorithm C is more complex, requiring the combination of a large number of variables from two databases (DCIR and PMSI) in the French national health insurance information system. This makes this algorithm more computationally expensive.

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

Conclusion

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

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

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

Compliance with ethical standards

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

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

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

References

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

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

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

Annex III: Article 2

Diabetes and Metabolism xxx (xxxx) xxx-xxx

Contents lists available at ScienceDirect

Diabetes and Metabolism

journal homepage: www.elsevier.com

Original article

Is the type 2 diabetes epidemic plateauing in France? A nationwide population-based study

S. Fuentes [a, ȳ], L. Mandereau-Bruno [a], N. Regnault [a], P. Bernillon [a], C. Bonaldi [a], E. Cosson [b, c], S. Fosse-Edorh [a]

[a] *Santé publique France, the French National Public Health Agency, 12, rue du Val d'Osne, 94415 Saint-Maurice, France*
[b] *Department of diabetology, endocrinology and metabolism, CRNH-IdF, CINFO, Paris 13 university, Sorbonne Paris cité, Avicenne hospital, AP–HP, 93000 Bobigny, France*
[c] *UMR U1153 Inserm, U1125 Inra, Cnam, Paris 13 university, Sorbonne Paris cité, 93000 Bobigny, France*

ARTICLE INFO

Article history:
Received 12 September 2019
Received in revised form 19 December 2019
Accepted 27 December 2019
Available online xxx

Keywords:
Algorithm
Diabetes
French overseas territories
Incidence
National Health Data System
Prevalence
Time trends

ABSTRACT

Aim

Nationwide data on the evolution of diabetes incidence and prevalence are scarce in France. For this reason, our objectives were to determine type 2 diabetes prevalence and incidence rates between 2010 and 2017, stratified by gender, age and region, and to assess annual time trends over the study period in adults aged □ 45 years.

Methods

Diabetes cases in the National Health Data System (SNDS), which covers the entire French population (66 million people), were identified through a validated algorithm. Gender- and age-specific prevalence and incidence rates were estimated. Negative binomial models, adjusted for gender, age and region, were used to assess annual time trends for prevalence and incidence throughout the study period.

Results

During 2017, 3,144,225 diabetes cases aged □ 45 years were identified. Over the study period, prevalence increased slightly (men from 11.5% to 12.1%, women from 7.9% to 8.4%) whereas incidence decreased (men from 11 to 9.7, women from 7.2 to 6.2 per 1000 person-years). In only four groups did prevalence rates decrease: men aged 45□65 years; women aged 45□60 years; women in Reunion; and women in Martinique. An increasing annual time trend was observed for prevalence (men: +0.9% [95% CI: +0.7%, +1%]; women: +0.4% [95% CI: +0.2%, +0.6%]) with a decreasing annual time trend for incidence in both genders (men: −2.6% [95% CI: −3.1%, −2.0%]; women: −3.9% [95% CI: −4.5%, −3.4%]).

Conclusion

Further efforts towards diabetes prevention are required to ensure that incidence rates in France continue to diminish, as the disorder continues to represent an important public-health burden.

Introduction

With an increasing number of cases worldwide reaching a total of 451 million in 2017, type 2 diabetes (T2D) has attained global epidemic status [1]. Various factors have been put forward as being responsible for this epidemic: increasing prevalence of risk factors (such as obesity and physical inactivity); an ageing population; and environmental factors [2,3].

Over recent decades, some studies have observed a plateauing of the diabetes epidemic, characterized by a slowing of prevalence growth and a levelling off of incidence [4,5]. Indeed, more recent studies have described a diminishing trend in diabetes incidence [6□8]. However, these results were based on surveys using samples from national populations or, in the case of population-based studies, on national health administrative databases, with no assessment of regional heterogeneity in the evolution of prevalence and incidence.

The French National Health Data System (*Système national des données de santé*, SNDS) is one of the largest health administrative databases in the world [9,10]. It covers the whole of the French population (approximately 66 million people), including people living in the French overseas territories (FOT), three in the Caribbean region (Guadeloupe, Martinique, French Guiana) and one in the Indian Ocean (Reunion) [11,12]. Nevertheless, nationwide data on diabetes prevalence for France are scarce and, to our knowledge, no data on diabetes incidence in adults in France have so far been reported.

Abbreviations: FOT, French overseas territories; SNDS, *Système national de données santé* (National Health Data System)

[C] Corresponding author.
Email address: sonsoles.FUENTES@santepubliquefrance.fr, sfuegut@gmail.com (S. Fuentes)

https://doi.org/10.1016/j.diabet.2019.12.006
1262-3636/ © 2020.

Thus, the objectives of the present study were:

- to describe T2D prevalence and incidence rates in France between 2010 and 2017, and 2012 and 2017, respectively, according to gender, age and geographical region;
- to assess annual time trends for T2D prevalence and incidence over those study periods.

Materials and methods

French National Health Data System

The SNDS collects individual and anonymized data from beneficiaries of the various different health insurance schemes available in France [13]. It is composed of two main databases: Inter-Scheme Consumption Data (*Données de consommation inter-régimes*, DCIR); and the Medical Information System Programme (*Progamme de médicalisation des systèmes d'information*, PMSI). The DCIR contains information on outpatient reimbursement ofdispensed healthcare, and demographic data such as age, gender and township of residence. The PMSI comprises inpatient information recorded from public and private hospitals on hospital stays (for example, admission and discharge dates; primary, related and associated diagnoses; and certain medical procedures). However, neither database includes clinical information regarding diagnoses made outside of hospital, the results of biological tests or patientsb anthropometric data. Linkage between the two databases is done only through anonymized identification of each individual.

In 2010, the SNDS also started including available data from the major health insurance schemes in France the general scheme for salaried workers (*Caisse nationale d'assurance maladie des travailleurs salariés*, CNAMTS), the scheme for agricultural workers and farmers (*Mutualité sociale agricole*, MSA) and the scheme for self-employed workers (*Régime social des indépendants*, RSI) along with other schemes. A small group of health insurance schemes covering < 1.5% of French population was finally added in 2014. In order to remain consistent throughout the study period, it was decided to restrict the study to only those beneficiaries of health insurance schemes included in 2010, which together cover almost all (98.5%) of the French population.

Diabetes case definition

Diabetes cases were identified using a validated algorithm that considered a case positive if the subject was reimbursed for any antidiabetic drug(class A10 in the Anatomical Therapeutic Chemical [ATC] classification system), except benfluorex, on at least three different dates in a given year, or on two dates if at least one large package of antidiabetic drugs was dispensed [14]. However, as our diabetes case definition algorithm was unable to differentiate between type 1 diabetes (T1D) and T2D, it was decided to study only adults aged 45 years in order to focus on the latter type of diabetes. In addition, all analyses were performed separately for women and for men.

Estimation of prevalence rate

Starting in 2010, the prevalence rate was calculated by dividing the total number of diabetes cases by the mean French population for a given year. The mean French population was assessed as the mean number of residents between 1 January and 31 December for the given study year, as estimated by the National Institute for Statistics (*Institut national de la statistique et des études économiques*, INSEE). These estimations were based on census data that are updated yearly using civil status statistics (births, deaths) and an estimate of net migration [15].

Estimation of incidence rate

Subjects were categorized as incident cases if they were identified as a diabetes case by the study algorithm in a given year and not identified as a diabetes case in the two preceding years. The population at risk for a given year was estimated as the mean number of the total population free of diabetes at the beginning (the French population minus the number of prevalent cases in the previous year) and at the end of the same year (the French population minus the number of prevalent cases in the study year).

Prevalence and incident rates

Crude prevalence and incidence rates were calculated according to gender, age (1-year prevalence) and region. Standardized rates stratified by gender and by 17 of France's 18 regions at the time (the Mayotte region was not included in our analyses) were estimated using 2013 European population data [16].

Annual time trends

Using generalized linear regression, annual time trends were analyzed by modelling the number of prevalent cases and incident cases with the log French population and the log French population at risk as offsets, respectively. Because of overdispersion of both outcomes, models with a log link and negative binomial distribution were applied robust variance estimators were used to correct standard errors and approximately estimated binomial variances. Independent variables were calendar year, region (with Occitanie as the reference region, as its prevalence and incidence rates over the study period were the closest to national rates) and age (as a continuous parametric fractional polynomial function) [17]. In the final model, interaction between calendar year and region was included because of its significance.

Results

Population characteristics

The French population in 2017 was estimated to be 66,815,984 people, with 46% of adults aged 45 years. In this age group, the mean age was 63 years and the ratio of men to women was 0.86. The number of diabetes cases aged 45 years was 3,143,225 (94% of all diabetes cases across all age groups) with a mean age of 69 years and a ratio of men to women of 1.24.

Diabetes prevalence and incidence

The evolution of diabetes prevalence and incidence in adults aged 45 years and stratified by gender is depicted in Fig. 1 . In 2010, the crude prevalence of diabetes in men was 10.9% whereas, in 2017, it was 11.8%. As regards incidence, crude rates were 10.7 cases per 1000 person-years (py) in 2012 and 9.6 cases per 1000 py in 2017. In women, the crude prevalence in 2010 was 7.9% vs. 8.4% in 2017. Crude incidence rates decreased from 7.1 to 6.1 cases per 1000 py from 2012 to 2017. Age-standardized prevalence and incidence rates for both genders between 2010 and 2017 arepresented in Table 1.

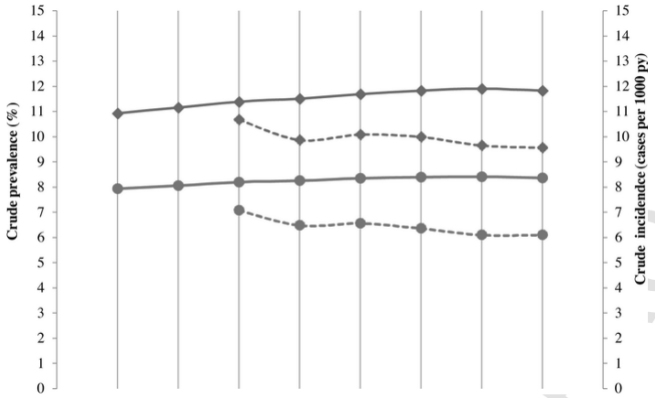


Fig. 1. Evolution of the crude prevalence (solid line) and incidence (dotted line) of diabetes from 2010 to 2017 in adults aged □ 45 years by gender (blue: men; red: women).

Table 1
Age-standardized (age-std) prevalence and incidence of diabetes from 2010 to 2017 by gender in adults aged □ 45 years and relative annual evolution rates (RaeR).

	2010	2011	2012	2013	2014	2015	2016	2017
Age-std prevalence(%)/RaeR								
Men	11.5	11.7	11.9	11.9	12.1	12.2	12.2	12.1
		+1.8%	+1.7%	+0.6%	+1.1%	+0.7%	+0.3%	−1.1%
Women	7.9	8.0	8.1	8.1	8.2	8.2	8.2	8.1
		+1.2%	+1.4%	+0.3%	+0.7%	+0.3%	−0.3%	−1.0%
Age-std incidence (cases per 1000py)/RaeR								
Men			11.0	10.1	10.4	10.2	9.8	9.7
				−8.1%	+2.5%	−1.7%	−3.6%	−1.0%
Women			7.2	6.6	6.6	6.4	6.2	6.2
				−8.9%	+0.6%	−2.7%	−4.4%	0.0%

Age-specific prevalence rates stratified by gender in 2010 and in 2017 are shown in Fig. 2a, while age-specific incidence rates stratified by gender in 2012 and in 2017 are shown in Fig. 2b. In both graphs, the curves representing men are always higher than those representing women. In 2010 and in 2017, prevalence increased with age, reaching its highest point (18.5% and 20.4%, respectively) in men aged 75□79 years before decreasing. Likewise, the age-specific curve of diabetes prevalence in women in 2010 reached its highest point in those aged 75□79 years (13.4%), but occurred at age 80□84 years (14.2%) in 2017.

In both 2012 and 2017, diabetes incidence increased in men from age 45□49 up to 65□69 years (14.9 and 13.2 cases/1000 py, respectively) before then starting to decrease. In 2012, age-specific incidence rates in women increased from age 45□49 years up to 70□74 years (3.7 and 9.8 cases per 1000 py, respectively) and then decreased whereas, in 2017, incidence rates plateaued from age 60 to 80 years (around 7.6 cases per 1000 py).

Age-standardized prevalence of diabetes in 2017 stratified by gender and region is presented in Fig. 3. Of the 17 French regions, the FOT (Reunion, Guadeloupe, Martinique, French Guiana) as well as the northern mainland region of Hauts-de-France had the highest age-adjusted prevalence rates. However, only in the FOT were the prevalence rates of diabetes higher in women than in men.

Annual time trends

Throughout the study period and after adjusting for age and geographical region, the prevalence of diabetes in men aged □ 45 years increased at an annual time trend of +0.9% (95% confidence interval [CI]: +0.7%, +1%), whereas incidence decreased at an annual time trend of −2.6% (95% CI: −3.1%, −2%). In women, the increasing annual time trend for prevalence was lower than in men (+0.4% [95% CI: +0.2%, +0.6%]) whereas, for incidence, the decreasing annual time trend was higher at −3.9% (95% CI: −4.5%, −3.4%).

Regional disparities

The evolution of diabetes prevalence and incidence in adults aged □ 45 years according to gender and region is shown in Fig. 4. Between 2010 and 2017, the age-adjusted prevalence in men increased in all regions (Fig. 4a), including those with the lowest rates, such as Brittany (from 7.3% to 8%) and Pays de la Loire (from 9.5% to 10.1%), and those with the highest rates, such as Guadeloupe (from 15.9% to 17.2%) and Reunion (from 19.3% to 19.4%). The annual time trend of diabetes prevalence adjusted for age ranged from +1.3% in Brittany to −0.3% in Reunion, although in the latter region, as in French Guiana, the trends were not significantly different from 0%. Also, just as was observed for men, the lowest age-adjusted prevalence rates for women (Fig. 4b) were in Brittany and Pays de la Loire

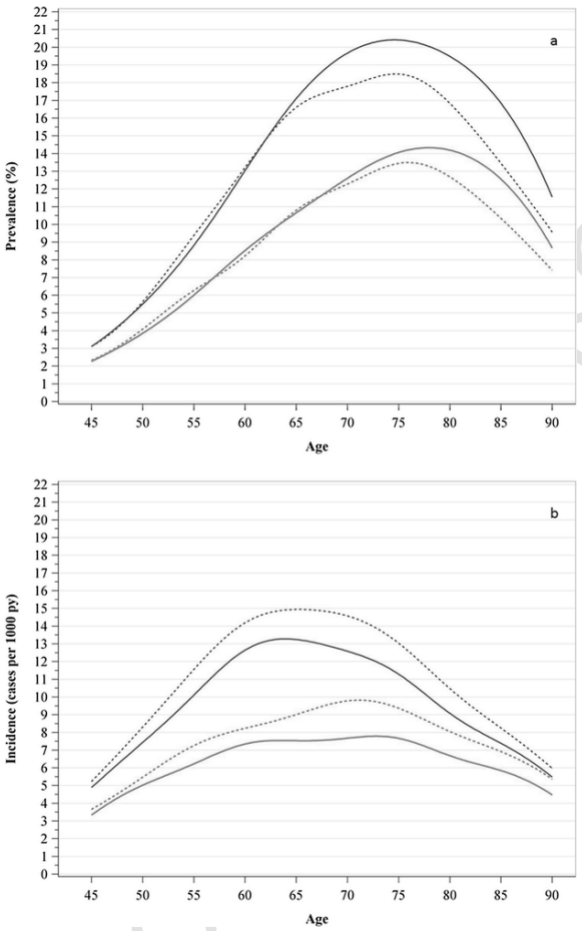


Fig. 2. Age-specific (a) prevalence in 2010 (dotted line) and in 2017 (solid line) and (b) incidence in 2012 (dotted line) and in 2017 (solid line) in France in adults aged □ 45 years stratified by gender (blue: men; red: women).

in both 2010 and 2017 (from 4.7% to 5% and from 6% to 6.3%, respectively). However, in contrast to men, the age-adjusted prevalence in women was lower in 2017 than in 2010 in two regions: Martinique (from 17% to 16.1%) and Reunion (from 21.9% to 20.2%). These trends were confirmed in the age-adjusted model, wherein both these regions saw significant decreases in annual time trends: Reunion, 1.1% (95% CI: −1.7%, −0.5%); Martinique, −0.9% (95% CI: −1.2%, −0.5%).

Age-adjusted incidence rates in men diminished in all regions between 2012 and 2017, with a greater decrease in regions where the age-adjusted prevalence was very high (Fig. 4c), for example, Guadeloupe (from 16.5 to 12.8 cases per 1000 py) and Reunion (from 17.5 to 13.3 cases per 1000 py). These regions also saw the largest decreases in annual time trends after adjusting for age and region (Guadeloupe −3.8% and Reunion −4.4%). In fact, significant decreases in annual time trends were observed for all regions studied except Corsica.

The evolution of diabetes incidence between 2012 and 2017 in women is depicted in Fig. 4d. Once again, Brittany (from 4.5 to 4 cases per 1000 py) and Pays de la Loire (from 5.5 to 5 cases per 1000 py) had the lowest incidence rates, while the FOT had both the highest age-adjusted incidence rates and largest decrease in incidence from 2012 to 2017: Martinique, from 13.3 to 10.8; Reunion, from 16.4 to 10.7; Guadeloupe, from 16.8 to 12.6; and French Guiana, from 21.5 to 15. These decreases were confirmed by the annual time trends, estimated using statistical models adjusted by age and region,

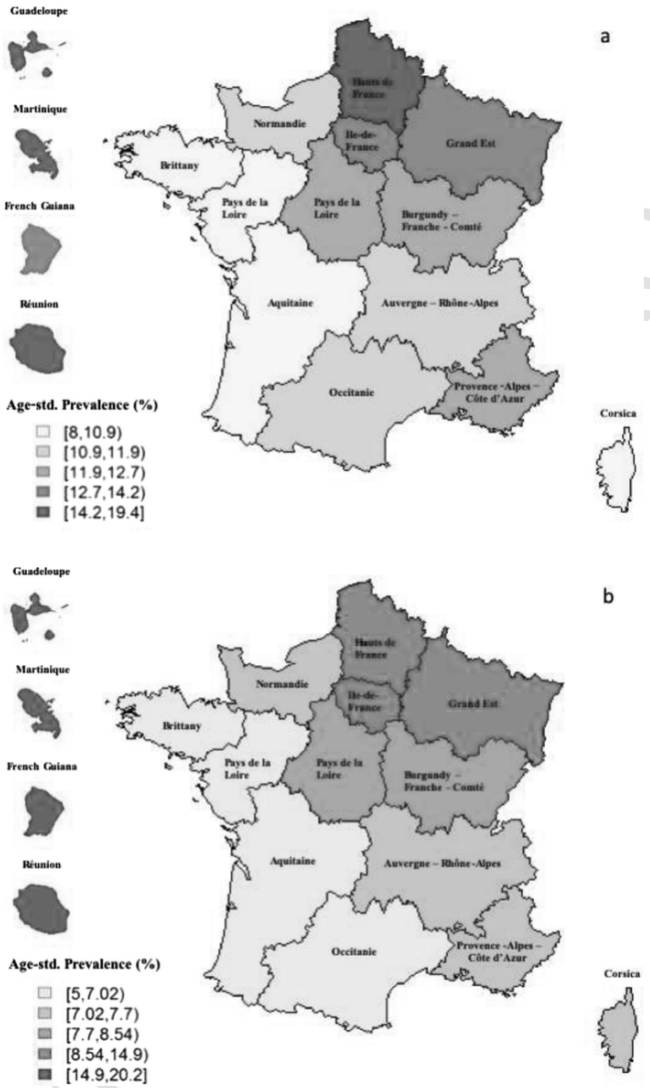


Fig. 3. Age-standardized (age-std.) prevalence of type 2 diabetes in France in 2017 in (a) men and (b) women aged □ 45 years according to geographical region.

which also showed the biggest decreases: Martinique, −4.5% (95% CI: −6.1%, −2.9%); Guadeloupe, −4.8% (95% CI: −6.4%, −3.2%); French Guiana, −5.3% (95% CI: −7.5%, −3.1%); and Reunion, −7.5% (95% CI: −8.9%, −6.2%).

Discussion

Using French nationwide population-based data, our present study was the first to determine diabetes incidence rates and time trends for diabetes prevalence and incidence in adults living in France. The

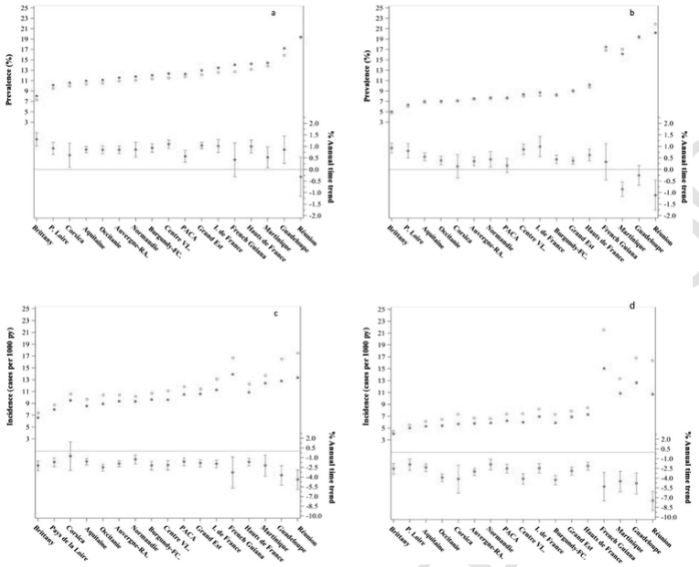


Fig. 4. Evolution of diabetes prevalence in France in (a) men and (b) women stratified by geographical region showing (upper plots) age-standardized prevalence (%) in 2010 (solid circles) and in 2017 (open circles), and (lower plots) annual time trends (%); and in (c) men and (d) women stratified by geographical region showing (upper plots) age-standardized incidence (cases per 1000 person-years, py) in 2012 (solid circles) and in 2017 (open circles), and (lower plots) annual time trends (%).

study hypothesis of a possible plateauing of the T2D epidemic in this country was not confirmed by our findings, as an increasing trend in prevalence rates was observed, although incidence rates decreased between 2010 and 2017 in adults aged □45 years. These results suggest that the mortality rate is decreasing among patients with T2D, a fact previously described in a French study based on a comparison of mortality rates observed over a 5-year follow-up of two waves in a large national cohort of diabetes patients aged □45 years [18]. In that study, the age-standardized mortality rates estimated in the first wave (2002□2007) were much higher than those estimated in the second wave (2007□2012). Such a reduction in mortality rates among patients with diabetes has also been observed in other European countries and in the United States (US), and has been explained by various factors, such as improvement of diabetes care management, and earlier diagnosis and treatment of the disorder [7,19□21].

The main strength of our present study is its comprehensiveness, as it was based on one of the largest health administrative databases in the world, including data for the entire French population (>66 million people) [13]. In the present nationwide study, geographical differences were also assessed. In certain regions such as the FOT, studying health outcomes through surveys and medical examinations presents many constraints in terms of accessibility and costs [22,23]. This challenge was overcome by using data from the SNDS, wherein health information together with data on age, gender and township/city of residence are available. On the other hand, the database lacks information on some other risk factors strongly associated with diabetes, such as ethnicity, family history of diabetes and body mass index scores [24□26]. The absence of such information prevents, for example, any study of the evolution of the diabetes epidemic in certain ethnic groups with a high prevalence of diabetes, including people born in North Africa [27].

A recent report validated the three algorithms commonly used in the SNDS to ascertain diabetes cases [14]. Based on the results of that study, the algorithm using antidiabetic drug reimbursement was selected for our present study as its performance was good in identifying diabetes cases (specificity 99%, sensitivity 86%, kappa coefficient 0.9), and these characteristics revealed no relevant differences when the validation analyses were stratified by gender and age group. Moreover, in contrast to the other two algorithms, the quality of information required to apply the selected study algorithm did not vary depending on the region or health insurance scheme, making it the most suitable for assessing trends and regional disparities.

Nevertheless, this algorithm was not able to identify those adult diabetes cases treated only through dietary changes and physical activity. However, the prevalence of non-pharmacologically treated diabetes in France is very low, as was pointed out in two nationwide studies conducted in 2006□2007 and in 2013 (0.9% and 0.7%, respectively) [28,29]. Furthermore, the algorithm could not differentiate between T1D and T2D, although our study overcame this limitation by focusing on those aged □45 years in whom the T1D prevalence is marginal [30].

Decreasing trends in diabetes incidence have been described previously in various studies. Recently, a study of the evolution of prevalence and incidence of diagnosed diabetes in US between 1980 and 2017 in adults aged 18□79 years found that, between 2008 and 2017, its incidence decreased significantly with an annual percentage change of −3.1% [8]. That study was based on cross-sectional survey data from the United States National Health Interview Survey (NHIS), but other studies based on national health administrative

databases, as was the case for our study, have also observed this trend for a diminishing diabetes incidence. In Sweden, time trends of pharmacologically treated diabetes incidence were observed to decrease from 2005 to 2013 in both men (−0.6%) and women (−0.7%) [6]. Recently, between 2009 and 2014, a significant 10% annual reduction in T2D incidence (whether treated pharmacologically or non-pharmacologically) was reported in residents of Norway aged 30–89 years [7].

Various factors may be responsible for such reductions in diabetes incidence rates. Some authors have suggested a decrease in the pool of undiagnosed cases due to a previous increase in screening programmes [2,4,7]. However, this would be unlikely in France, where prevalence rates of undiagnosed diabetes have not decreased over time, as shown by a comparison of two similar nationwide studies conducted in 2006–2007 (1%) and in 2013 (1.9%) [28,29]. Based on those same studies, the diminishing trend in incidence would also not be fully explained by a switch from pharmacologically treated to non-pharmacologically treated diabetes, as the prevalence rates of the latter were 0.9% in 2006–2007 and 0.7% in 2013. Another plausible explanation might be a slowdown in the increase of risk-factor prevalence: trends for obesity prevalence in the United States [31,32] and in France have become stable in recent years [33]. One study comparing two waves of the French Nutrition and Health Survey (ENNS) in 2006–2007 and in 2014–2016 (Health Study on Environment, Biomonitoring, Physical Activity and Nutrition, ESTEBAN) found that the prevalence of overweight and obesity had not increased over the preceding decade (both were stable at approximately 49% and 17%, respectively). On the other hand, the study did reveal an increase in other risk factors, such as a lack of physical activity and dietary factors.

The decrease in incidence rates was greater among women, which could result in a more favourable evolution of the diabetes epidemic compared with men. However, previous studies have discovered that the effects of socioeconomic inequality on diabetes prevalence and excess mortality among diabetes cases were more important in women than in the general population [34,35].

Our nationwide study has allowed us to assess regional differences in the prevalence and incidence of T2D. In continental France, a gradient from the western to northeastern parts of the country was observed for prevalence of diabetes. Such a gradient has also been described for various diabetes risk factors, such as obesity and low socioeconomic status [36,37]. Nevertheless, the highest prevalence was observed in the FOT with one noteworthy observation: in contrast to regions in mainland France, prevalence rates of T2D were higher in women than in men in the FOT. This inverse male-to-female ratio for diabetes prevalence has also been described in other studies of Caribbean and South African countries [38,39]. One possible explanation is the considerably higher rate of obesity in women in those regions: a meta-analysis of 27 studies from Caribbean countries found that obesity prevalence was three times higher in women [38]. In addition, a cross-sectional study conducted in 2252 subjects aged 18–74 years in Guadeloupe in 2014 found that 55.3% of women had a waist circumference at or above the United States National Cholesterol Education Program thresholds (≥102 cm in men, ≥89 cm in women) vs. only 14% of men [40].

In the present study, it was also noted that the FOT regions saw the largest decrease in incidence over the study period. This may partially be explained by a decrease in the prevalence of obesity. In 2014–2016 in Reunion, the ESTEBAN survey found that the prevalence of obesity among adults aged ≥ 40 was much lower than that described in a similar study conducted in 2002 [33,41]. Moreover, the prevalence of obesity among adults appears to have stabilized over recent years, especially in Guadeloupe and Martinique [22]. This favourable dynamic observed in the FOT regions could be the result of an increased awareness of the burden of diabetes and obesity in those regions. More specifically, prevention campaigns have been carried out there together with other public-health interventions, such as the 2013 Sugar Act. When independent studies revealed that many commercialized products, such as soft drinks and dairy, sold in the FOT regions had higher levels of sugar than the same products in mainland France, the French Parliament approved the 2013 Sugar Act, banning these practices with a view to ensuring the quality of food products distributed in those regions. Given the progressive implementation of this Sugar Act, it may perhaps be too early to assess its effectiveness and influence on our present results. However, such interventions may well have raised awareness in the general population of the benefits of following healthy lifestyles.

Conclusion

Thanks to the exhaustive data from the SNDS database, which covers the whole of the French population, our study was able to assess, for the first time, the incidence as well as time trends of both the prevalence and incidence of diabetes in France. Future studies should aim to determine the observed trends of T2D in younger populations and the role of different influences (such as socioeconomic and lifestyle factors) in the dynamics described in our present study. These results, based on data collected over a long time period, suggest a change in diabetes epidemic dynamics in France, and highlight the value of further efforts towards the prevention of diabetes, a disease that continues to represent a major public-health burden.

Prior presentation

Parts of this study were presented in poster form at the 79th Annual Scientific Sessions of the American Diabetes Association, 7–11 June 2019, San Francisco, CA, US (1618-Poster) and at the 55th Annual Meeting of the European Association for the Study of Diabetes, 16-20 September 2019, Barcelone, Spain (316-Poster).

Funding

This research was supported by *Santé publique France*.

Contributions of authors

The research question was formulated by S. F.-E. and E. C.; S. F.-E., E. C. and S. F. designed the study; S. F. analyzed the data; L. M.-B., N. R., C. B. and P. B. provided the analytical tools; S. F., S. F.-E., E. C., L. M.-B. and C. B. were involved in data interpretation; S. F. wrote the manuscript; S. F., S. F.-E., E. C., L. M.-B., N. R., C. B. and P. B. critically revised the manuscript; and S. F.-E. secured funding for the study and acquired the data.

S. F. is the guarantor of this work and, as such, had full access to all the data in the study, and takes responsibility for the integrity of the data and accuracy of the data analysis.

Disclosure of interest

The authors declare that they have no competing interest.

Acknowledgements

We thank Clara Piffaretti from the Directorate of Non-Communicable Disease and Trauma of *Santé publique France* for her help in analyzing the data and developing the figures presented.

References

[1] N.H. Cho, J.E. Shaw, S. Karuranga, Y. Huang, J.D. da Rocha Fernander, A.W. Ohlrogge, et al., IDF Diabetes Atlas: global estimates of diabetes prevalence for 2017 and projections for 2045, Diabetes Res Clin Pract 138 (2018) 271–281.
[2] V. Thibault, M. Bélanger, E. LeBlanc, L. Babin, S. Halpine, B. Greene, et al., Factors that could explain the increasing prevalence of type 2 diabetes among adults in a Canadian province: a critical review and analysis, Diabetol Metab Syndr 8 (2016) 71.
[3] L.S. Geiss, K. Kirtland, J. Lin, S. Shrestha, T. Thompson, A. Alright, et al., Changes in diagnosed diabetes, obesity, and physical inactivity prevalence in US counties, 2004–2012, PLoS One 12 (2017) e0173428.
[4] L.S. Geiss, J. Wang, Y.J. Cheng, T.J. Thompson, L. Barker, Y. Li, et al., Prevalence and incidence trends for diagnosed diabetes among adults aged 20 to 79 years, United States, 1980–2012, JAMA 312 (2014) 1218–1226.
[5] T.M. Abraham, K.M. Pencina, M.J. Pencina, C.S. Fox, Trends in diabetes incidence: the Framingham Heart Study, Diab Care 38 (2015) 482–487.
[6] S.P. Jansson, K. Fall, O. Brus, A. Magnuson, P. Wändell, C.J. Östgren, et al., Prevalence and incidence of diabetes mellitus: a nationwide population-based pharmaco-epidemiological study in Sweden, Diab Med 32 (2015) 1319–1328.
[7] P.L.D. Ruiz, L.C. Stene, I.J. Bakken, S.E. Haberg, K.I. Birkeland, H.L. Gulseth, Decreasing incidence of pharmacologically and non-pharmacologically treated type 2 diabetes in Norway: a nationwide study, Diabetologia 61 (2018) 2310–2318.
[8] S.R. Benoit, I. Hora, A.L. Albright, E.W. Gregg, New directions in incidence and prevalence of diagnosed diabetes in the USA, BMJ Open Diab Res Care 7 (2019) e000657.
[9] A. Palmaro, G. Moulis, F. Despas, J. Dupouy, M. Lapeyre-Mestre, Overview of drug data within French health insurance databases and implications for pharmacoepidemiological studies, Fundam Clin Pharmacol 30 (2016) 616–624.
[10] G. Moulis, M. Lapeyre-Mestre, A. Palmaro, G. Pugnet, J.L. Montastruc, L. Sailler, French health insurance databases: what interest for medical research?, Rev Med Interne 36 (2015) 411–417.
[11] A. Filipovic-Pierucci, A. Rigault, A. Fagot-Campagna, P. Tuppin, Health status of populations living in French overseas territories in 2012, compared with metropolitan France: an analysis of the national health insurance database, Rev Epidemiol Sante Publique 64 (2016) 175–183.
[12] P. Tuppin, P. Ricci-Renaud, C. de Peretti, A. Fagot-Camapagna, F. Alla, N. Danchin, et al., Frequency of cardiovascular diseases and risk factors treated in France according to social deprivation and residence in an overseas territory, Intern J Cardiol 173 (2014) 430–435.
[13] P. Tuppin, J. Rudant, P. Constantinou, C. Gastaldi-Ménager, A. Rachas, L. de Roquefeuil, et al., Value of a national administrative database to guide public decisions: from the Système national d'information interregimes de l'Assurance Maladie (SNIIRAM) to the Système national des données de santé (SNDS) in France, Rev Epidemiol Sante Publique 65 (4) (2017) S149–S167.
[14] S. Fuentes, E. Cosson, L. Mandereau-Bruno, A. Fagot-Campagan, P. Bernillon, M. Goldberg, et al., Identifying diabetes cases in health administrative databases: a validation study based on a large French cohort, Int J Public Health 64 (2019) 441–450.
[15] M.J. Saurel-Cubizolles, J.F. Chastang, G. Menvielle, A. Leclerc, D. Luce, EDISC group, Social inequalities in mortality by cause among men and women in France, J Epidemiol Community Health 63 (2009) 197–202.
[16] M. Pace, G. Lanzieri, M. Glickman, E. Grande, T. Zupanic, B. Wojtyniak, et al., Revision of the European standard population. Report of the Eurostat's task force. 2013, Publications Office of the European Union, Luxembourg, 2013.
[17] P. Royston, G. Ambler, W. Sauerbrei, The use of fractional polynomials to model continuous risk variables in epidemiology, Int J Epidemiol 28 (1999) 964–974.
[18] L. Mandereau-Bruno, A. Fagot-Campagna, G. Rey, A. Latouche, S. Fosse-Edorth, CO-03 : évolution de la surmortalité des personnes diabétiques traitées pharmacologiquement entre 2002–2006 et 2007–2012 – cohortes Entred2001 et Entred2007, Diab Metab 42 (2016) A1–A2.
[19] B. Carstensen, J.K. Kristensen, P. Ottosen, K. Borch-Johnsen, Steering Group on the National Diabetes Register, The Danish National Diabetes Register: trends in incidence, prevalence and mortality, Diabetologia 51 (2008) 2187–2196.
[20] S.H. Read, J.J. Kerssens, D.A. McAllister, H.M. Colhoun, C.M. Fishbacher, R.S. Lindsay, et al., Trends in type 2 diabetes incidence and mortality in Scotland between 2004 and 2013, Diabetologia 59 (2016) 2106–2113.
[21] D. Kim, A.A. Li, G. Cholankeril, S.H. Kim, E. Ingelsson, J.W. Knowles, et al., Trends in overall, cardiovascular and cancer-related mortality among individuals with diabetes reported on death certificates in the United States between 2007 and 2017, Diabetologia 62 (2019) 1185–1194.
[22] J.L. Daigre, A. Atallah, J.L. Boissin, G. Jean-Baptoste, P. Kangambega, H. Chevalier, et al., The prevalence of overweight and obesity, and distribution of waist circumference, in adults and children in the French Overseas Territories: the PODIUM survey, Diab Metab 38 (2012) 404–411.
[23] M.C. Riddle, L. Blonde, H.C. Gerstein, E.W. Gregg, R.R. Holman, J.M. Lachin, et al., Editorsbexpert forum 2018: managing big data for diabetes research and care, Diab Care 42 (2019) 1136–1146.
[24] L.S. Geiss, K.M. Bullard, R. Brinks, A. Hoyer, E.W. Gregg, Trends in type 2 diabetes detection among adults in the USA, 1999–2014, BMJ Open Diab Res Care 6 (2018) e000487.
[25] T.P. Candler, O. Mahmoud, R.M. Lynn, A.A. Majbar, T.G. Barrett, J.P.H. Shield, Continuing rise of Type 2 diabetes incidence in children and young people in the UK, Diab Med 35 (2018) 737–744.
[26] K.N. Turi, D.M. Buchner, D.S. Grigsby-Toussaint, Predicting risk of Type 2 diabetes by using data on easy-to-measure risk factors, Prev Chronic Dis 14 (2017) E23.
[27] S. Fosse Edorh, A. Fagot-Campagna, B. Detournay, H. Bihan, A. Gautier, M. Dalichampt, et al., Type 2 diabetes prevalence, health status and quality of care among the North African immigrant population living in France, Diab Metab 40 (2014) 143–150.
[28] C. Bonaldi, M. Vernay, C. Roudier, B. Salanave, A. Oleko, A. Malon, et al., A first national prevalence estimate of diagnosed and undiagnosed diabetes in France in 18- to 74-year-old individuals: the French Nutrition and Health Survey 2006/2007, Diabet Med 28 (2011) 583–589.
[29] S. Fuentes, S. Fosse-Edorh, N. Regnault, M. Goldberg, E. Cosson, Prevalence of prediabetes and undiagnosed diabetes among adults aged 18 to 70 years in France – The CONSTANCES Cohort, Diabetes 67 (1) (2018), [1657-P].
[30] P.A. Diaz-Valencia, BougneresP, A.J. Valleron, Global epidemiology of type 1 diabetes in young adults and adults: a systematic review, BMC Public Health 15 (2015) 255.
[31] C.C. Imes, L.E. Burke, The obesity epidemic: the United States as a cautionary tale for the rest of the world, Curr Epidemiol Rep 1 (2014) 82–88.
[32] M. Wabitsch, A. Moss, K. Kromeyer-Hauschild, Unexpected plateauing of childhood obesity rates in developed countries, BMC Med 12 (2014) 17.
[33] C. Verdot, M. Torres, B. Salanave, V. Deschamps, Corpulence des enfants et des adultes en France métropolitaine en 2015. Résultats de l'étude Esteban et évolution depuis 2006, Bull Epidemiol Hebd 13 (2017) 234–241.
[34] B.L. Mandereau, S. Fosse Edorh, Prévalence du diabète traité pharmacologiquement (tous types) en France en 2015. Disparités territoriales et socioéconomiques, Bull Epidemiol Hebd 28 (2017) 586–591, [Prevalence of pharmacologically-treated diabetes (all types) in France in 2015. Territorial and socio-economic disparities]..
[35] L. Mandereau-Bruno, et al., Surmortalité sur la période 2002–2011 des personnes diabétiques traitées pharmacologiquement en France métropolitaine par rapport à la population générale. Cohorte ENTRED 2001, BMC Pregnancy Childbirth 38 (2016) 676–680.
[36] E. Eschwege, S.A. Charles, A. Basdevant, C. Moisan, G. Bonnélye, C. Touboul, et al., ObEpi Roche 2012 – Enquête épidémiologique nationale sur le surpoids et l'obésité, 2012.
[37] Santé publique France, L'état de santé de la population en France : rapport 2017, 2017.
[38] N. Sobers-Grannum, M.M. Murphy, A. Nielsen, C. Guelle, T.A. Samuels, L. Bishop, et al., Female gender is a social determinant of diabetes in the Caribbean: a systematic review and meta-analysis, PLoS One 10 (2015) e0126799.
[39] E.H. Hilawe, C. Chiang, H. Yatsuya, C. Wang, E. Ikerdeu, K. Honjo, et al., Prevalence and predictors of prediabetes and diabetes among adults in Palau: population-based national STEPS survey, Nagoya J Med Sci 78 (2016) 475–483.
[40] P. Carrere, C. Fagour, D. Spotouch, F. Gane-Troplent, J. Hélène-Pelage, T. Lang, et al., Diabetes mellitus and obesity in the French Caribbean: a special vulnerability for women?, Women Health 58 (2018) 145–159.
[41] F. Favier, et al., Comportement alimentaire et activité physique des Réunionnais, 200283, [Étude RECONSAL, I.-O.L. Réunion, editor].

Annex IV: Article 3

TITLE PAGE

Full title:

Artificial intelligence for diabetes research: development of type 1/type 2 classification algorithm and its application to surveillance using a nationwide population-based medico-administrative database in France

Short running title (47 characters):

AI and the Diabetes Classification Algorithm

Authors:

Sonsoles Fuentes[1], Rok Hrzic[2], Romana Haneef [1], Sofiane Kab [3], Sandrine Fosse-Edorh*[1] and Emmanuel Cosson MD PhD*[4,5]

1. Santé Publique France (SpF), Saint Maurice, France

2. Maastricht University, Fac. Health, Medicine and Life Sciences , International Health, School for Public Health and Prim Care, Maastricht, The Neetherlands

3. Population-Based Epidemiological Cohorts Unit, Inserm UMS 011, Villejuif, France

4. Department of Endocrinology-Diabetology-Nutrition, AP-HP, Avicenne Hospital, Paris 13 University, Sorbonne Paris Cité, CRNH-IdF, CINFO, Bobigny, France

5. Sorbonne Paris Cité, UMR U1153 Inserm/U1125 Inra/Cnam/Université Paris 13, Bobigny, France.

* E.C. and S.F-E. contributed equally to this work.

ABSTRACT

Objective : 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). The objectives of this study were to develop a type 1/type 2 diabetes classification algorithm using artificial intelligence and to estimate the prevalence of type 1 and type 2 diabetes in France.

Research Design and 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 3,481 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, type 1 and type 2 diabetes prevalence in France was 0.3% and 4.4%.

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.

MAIN TEXT

Diabetes is a leading cause of morbidity and mortality worldwide (1), and public health surveillance is fundamental in decreasing the global burden of diabetes (2). In recent decades, big data have emerged and offered new opportunities for surveillance (3, 4). Big data refers to massive volumes of information collected from different sources, as characterized by the "three Vs": volume, velocity and variety (5).

One example of a big data source for public health surveillance is the French national health data system, the SNDS (6) (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). This algorithm has very good performance (sensitivity 97.3%, specificity 99.9% and accuracy 99.9%) but cannot distinguish type 1 from type 2 diabetes. Differentiating type 1 and type 2 diabetes is crucial in diabetes surveillance, because the two types of diabetes 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 pharmacologically treated 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 (5).

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 study population database, the SNDS, to assess the prevalence of type 1 and type 2 diabetes in France among adults in 2016.

RESEARCH DESIGN AND METHODS

(i) To develop a type 1/ type 2 classification algorithm

The CONSTANCES cohort

The CONSTANCES population-based general-purpose cohort was used to develop an algorithm for distinguishing type 1 from type 2 diabetes on the basis of SNDS data. Since 2012, the CONSTANCES 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 (*Caisse Nationale d'Asssurance Maladie des travailleurs salaries*, CNAMTS), including all active or retired workers and their families, *i.e.*, approximately 86 % of the French population. First, participants

complete a self-administered questionnaire on health status, health-related behaviors, socioeconomic and demographic characteristics, and occupational information. They then attend a HSC 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.

Supervised Machine Learning

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, *(14)* 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 (15). 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 (Student's t test for continuous variables and Fisher's exact test for categorical variables).

- 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 (16-18).

- 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 downsampling 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 (19, 20). 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: LDA, logistic regression, flexible discriminant analysis (FDA) and C.5 decision tree (C5) (21). For each model, the features were selected with three different thresholds of ReliefExp scores: 0.35, 0.1 and 0.05 (Figure 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.

(ii) To assess the prevalence of type 1 and type 2 diabetes among adults in France

Study population: The French national health insurance information system (SNDS)

The French health care system has universal coverage, and all beneficiaries have a unique identification number and a personal smartcard (*carte vitale*) allowing information on health care utilization to be electronically recorded (7, 22). This information is collected and anonymized by the SNDS, which comprises two main databases: inter-scheme consumption data (*Données de consummation inter-régimes*, DCIR) and the French national hospital discharge database (*Progamme de médicalisation des systèmes d'information*, PMSI). The DCIR contains 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, such as admission and discharge dates, diagnoses (primary, related and associated) and medical procedures.

Prevalence of type 1 and type 2 diabetes in France in 2016

All pharmacologically treated diabetes cases in France in 2016 were ascertained in the SNDS using a validated algorithm that identifies a diabetes case if an individual had a reimbursement for an antidiabetic drug (class ATC A10, except Benfluorex) on at least three different dates in a given year or on two dates if at least one large package of antidiabetic drugs was dispensed (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) (23). Data management and analysis of the SNDS were performed with SAS 7.1, and supervised machine learning was performed with the R packages COREIearn and caret.

RESULTS

Final data set

Among the 50,954 participants recruited by the CONSTANCES cohort between 2012 and 2014, a total of 1,161 diabetes cases, were identified. The final data set for developing the algorithm was composed of 951 pharmacologically treated diabetes cases after exclusion of 88 cases with incomplete data on diabetes diagnosis and treatment and 122 cases not pharmacologically treated (all of which were type 2 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 in the final study population are presented. The type 2 group contained a higher percentage of men, current smokers and currently obese individuals than the type 1 group. Regarding socioeconomic factors, only 6% of type 1 cases had a low education level (lower secondary, primary or no education), as compared with 22.7% of type 2 cases. Most type 2 cases were retired (58.5%), whereas most type 1 cases were employed (68.2%).

Feature selection for the type1/type2 diabetes classification algorithm

All 3,481 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). Fig. 3 provides 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.

Assessing type 1 and type 2 diabetes prevalence in 2017 using data from 66 million people living in France

In 2016, a total of 1,844,329 diabetes cases ranging in age from 18 to 69 years (after excluding 7,248 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).

DISCUSSION

We utilized information in SNDS and an innovative method based on supervised machine learning to develop an algorithm to distinguish type 1 from type 2 diabetes cases in France. This algorithm is based on the number of reimbursements for fast-acting insulin, long-acting insulin and biguanides over the prior 12 months. It has very good performance in identifying type 2 diabetes cases, with a sensitivity of 100% and an accuracy of 97%, as well as high transferability to other databases. We applied this classification algorithm to SNDS data and estimated the prevalence of type 1 and type 2 diabetes among the 66 million adults living in France in 2016. The prevalence of type 1 diabetes was higher than that of type 2 diabetes until 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 3,481 to 14 features

From the 3,481 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, because they were highly correlated with usual diabetes onset, treatment, management and complications. 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 (24). 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 (25). 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 (26). The second feature was the greater number of tests for the urinary albumin excretion rate in type 1 than type 2 cases (27).

However, some unexpected features were highly discriminant between type 1 and type 2 diabetes. One such feature was the 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 (28). 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 (29), 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 an 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 ICD 9/10 diagnostic codes (30). 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 (31). 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, 32). 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 (30). 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 (23).

The combination of therapeutic features constituting the algorithm was consistent with treatment guidelines in France, other European countries and the US (24, 33). 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 (34-36).

The estimations of type 1 diabetes prevalence among adult population are scarce (28, 37). 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) (28, 37).

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). In the final data set, the characteristics of age, socioeconomic status, lifestyle factors and anthropometric measures for type 2 and type 1 groups were consistent with those observed in previous studies (30, 38). Regarding socio-economic characteristics, type 2 diabetes is more frequent than type 1 diabetes among populations with low education levels (39), as observed in this data set, in which one-quarter of type 2 cases had a lower secondary education level or below, as compared with only 6% of type 1 cases. Finally, 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 CONSTANCES 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 (40). 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 CONSTANCES' 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. We could overcome these two limitations with the information recorded in the third wave of the ENTRED study, a national cross-sectional survey on a large representative sample of pharmacologically treated diabetes cases in France launched in 2019 (41).

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.

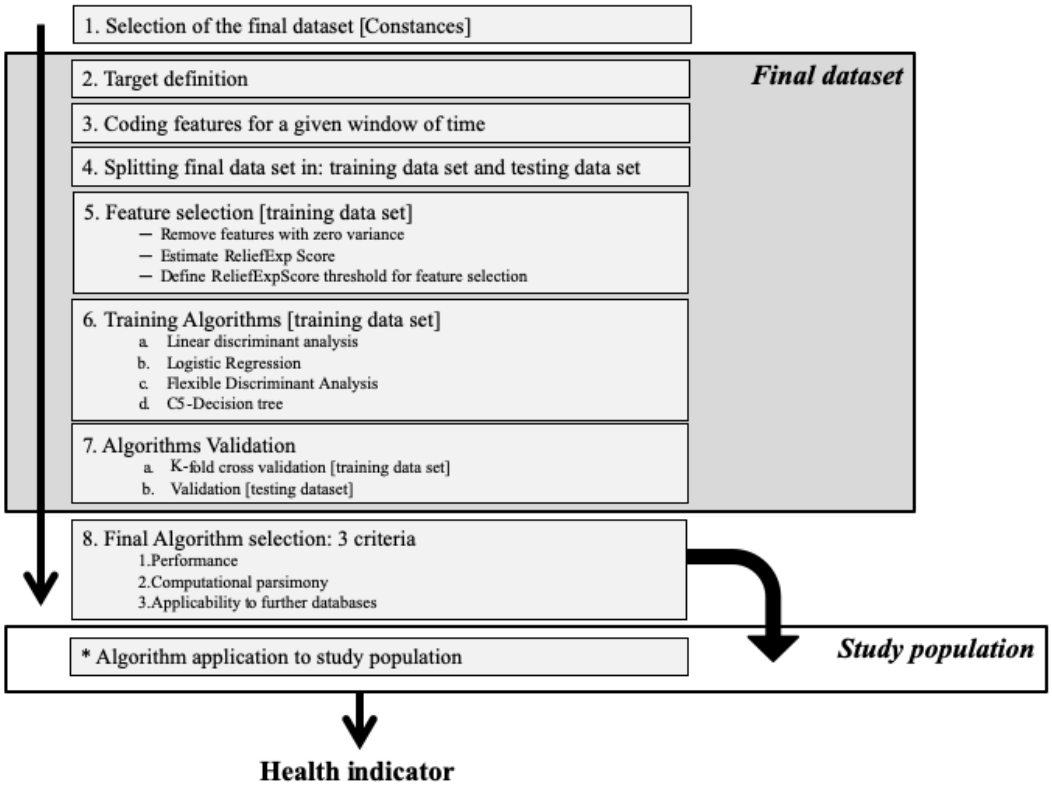


Fig. 1. Supervised machine learning for developing a classification/prediction algorithm

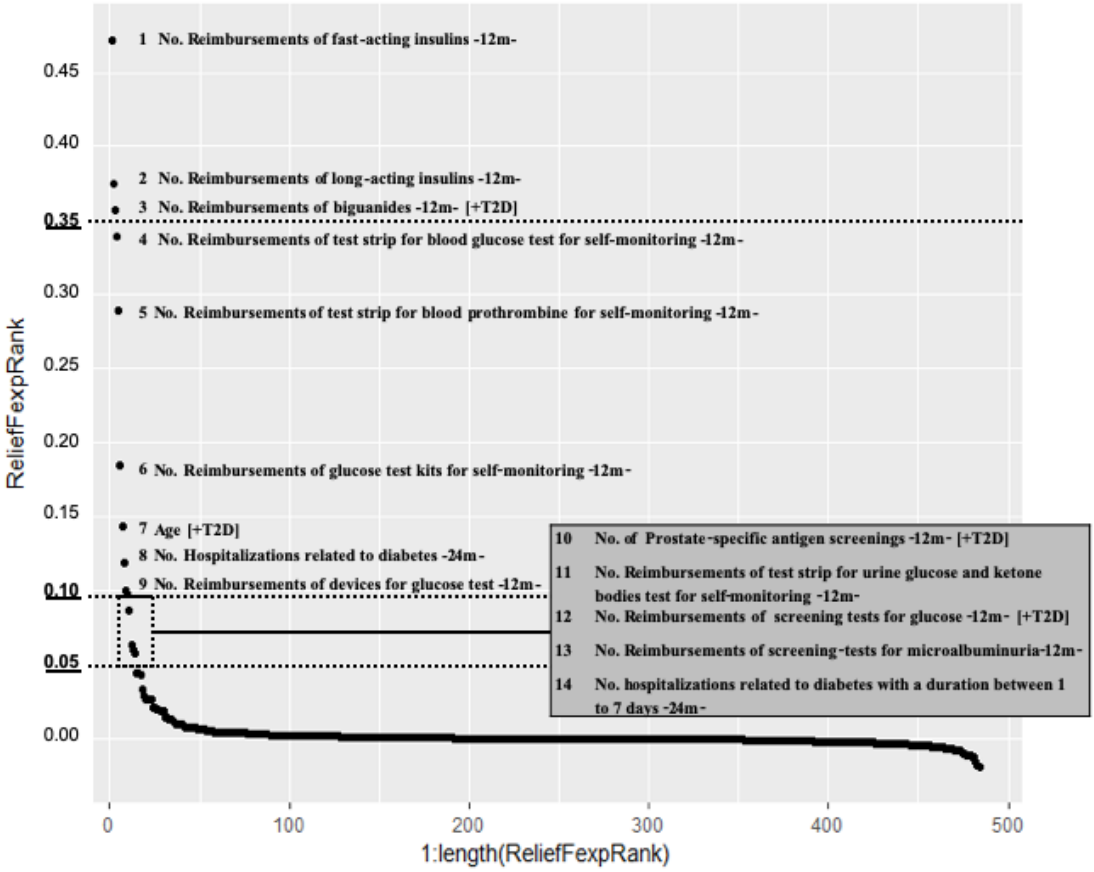


Fig. 2. Feature selection based on ReliefExp score using three different thresholds (0.35, 0.1 and 0.05)
No: number of; -12m-: over the prior 12 months; -24m-: over the prior 24 months; [+T2D]: the mean in the type 2 diabetes group is higher than that in the type 1 diabetes group

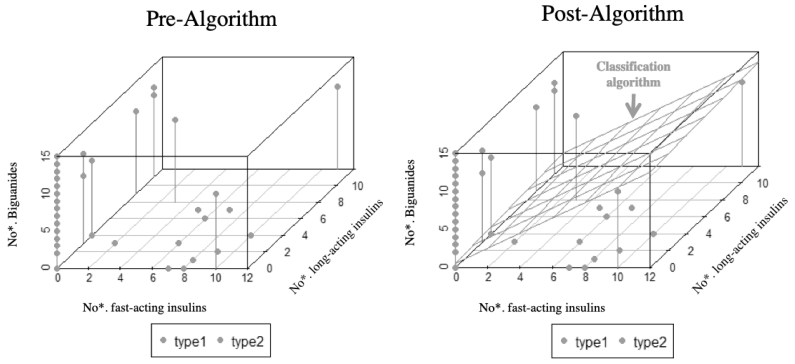


Fig. 3. Validation using data from the testing data set (n=189) of the type 1/type 2 diabetes classification algorithm (linear discriminant model with ReliefExp score threshold for feature selection of 0.35–3 features)
Nb.: number of reimbursements over the prior 12 months

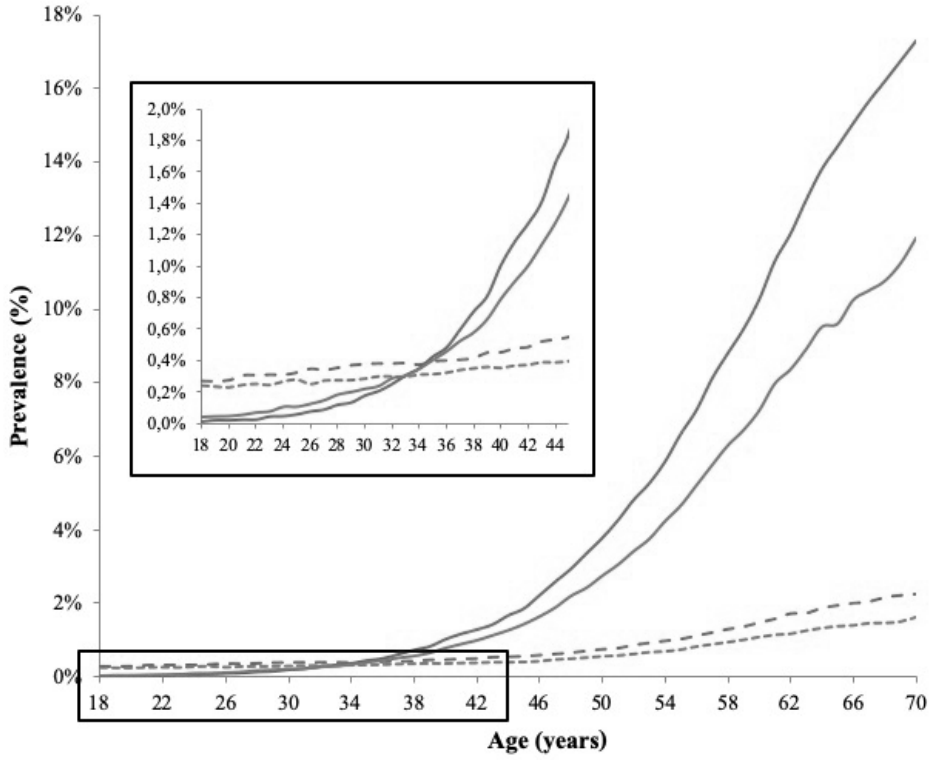


Fig. 4. Distribution of type 1 and type 2 diabetes prevalence (%) in France among adults 18–70 years of age, by sex and age
Dotted lines: type 1 diabetes; solid lines: type 2 diabetes; blue lines: men; red lines: women

REFERENCES

1. Karamanou, M., et al., *Milestones in the history of diabetes mellitus: The main contributors.* World J Diabetes, 2016. **7**(1): p. 1-7.
2. Wass, J.A.H., et al., *Oxford Textbook of Endocrinology and Diabetes*. 2 edition ed. Vol. 1. 2016: Oxford University Press.
3. World Health Organization, *Definition, diagnosis and classification of diabetes mellitus and its complications: report of a WHO consultation. Part 1, Diagnosis and classification of diabetes mellitus*. 1999, Geneva: World health organization.
4. Pfeifer, M.A., J.B. Halter, and D. Porte Jr, *Insulin secretion in diabetes mellitus.* The American journal of medicine, 1981. **70**(3): p. 579-588.
5. World Health Organization, *Classification of diabetes mellitus.* 2019.
6. Raverot, G., *Diabète sucré de types 1 et 2 de l'enfant et de l'adulte.* La collection Hippocrate: Endocrinologie Métabolisme Réanimation-urgences, 2005.
7. World Health Organization, *Diagnostic criteria and classification of hyperglycaemia first detected in pregnancy*. 2013, World Health Organization: Geneva.
8. van Belle, T.L., K.T. Coppieters, and M.G. von Herrath, *Type 1 diabetes: etiology, immunology, and therapeutic strategies.* Physiol Rev, 2011. **91**(1): p. 79-118.
9. Atkinson, M.A., *The pathogenesis and natural history of type 1 diabetes.* Cold Spring Harb Perspect Med, 2012. **2**(11).
10. American Diabetes, A., *2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes-2019.* Diabetes Care, 2019. **42**(Suppl 1): p. S13-S28.
11. Kobberling, J., *Empirical risk figures for first degree relatives of non-insulin dependent diabetes.* The genetics of diabetes mellitus, 1982. **201**.
12. Kolb, H. and S. Martin, *Environmental/lifestyle factors in the pathogenesis and prevention of type 2 diabetes.* BMC Med, 2017. **15**(1): p. 131.
13. Wu, Y., et al., *Risk factors contributing to type 2 diabetes and recent advances in the treatment and prevention.* Int J Med Sci, 2014. **11**(11): p. 1185-200.
14. American Association of Clinical Endocrinologists. *Clinical Presentation of Type 2 Diabetes Mellitus*. 2019 [cited 2019 November]; Available from: https://www.aace.com/disease-state-resources/diabetes/depth-information/clinical-presentation-type-2-diabetes-mellitus.
15. World Health Organization, *Definition and diagnosis of diabetes mellitus and intermediate hyperglycaemia: report of a WHO/IDF consultation.* 2006.
16. Assurance Maladie. *Symptômes et diagnostic du diabète*. 2019 [cited 2019; Available from: https://www.ameli.fr/assure/sante/themes/diagnostic/diagnostic-diabete.
17. Gabir, M.M., et al., *The 1997 American Diabetes Association and 1999 World Health Organization criteria for hyperglycemia in the diagnosis and prediction of diabetes.* Diabetes care, 2000. **23**(8): p. 1108-1112.
18. Bansal, N., *Prediabetes diagnosis and treatment: A review.* World J Diabetes, 2015. **6**(2): p. 296-303.
19. Kilpatrick, E.S., Z.T. Bloomgarden, and P.Z. Zimmet, *International Expert Committee report on*

the role of the A1C assay in the diagnosis of diabetes: response to the International Expert Committee. Diabetes Care, 2009. **32**(12): p. e159; author reply e160.

20. Cheng, Y.J., et al., *Association of A1C and fasting plasma glucose levels with diabetic retinopathy prevalence in the U.S. population: Implications for diabetes diagnostic thresholds.* Diabetes Care, 2009. **32**(11): p. 2027-32.
21. World Health Organization, *Use of glycated haemoglobin (HbA1c) in diagnosis of diabetes mellitus: abbreviated report of a WHO consultation.* 2011.
22. Fazeli Farsani, S., et al., *Incidence and prevalence of diabetic ketoacidosis (DKA) among adults with type 1 diabetes mellitus (T1D): a systematic literature review.* BMJ Open, 2017. **7**(7): p. e016587.
23. Adeyinka, A. and N.P. Kondamudi, *Hyperosmolar Hyperglycemic Nonketotic Coma (HHNC, Hyperosmolar Hyperglycemic Nonketotic Syndrome)*, in *StatPearls*. 2019: Treasure Island (FL).
24. American Diabetes Association, *10. Microvascular Complications and Foot Care: Standards of Medical Care in Diabetes-2018.* Diabetes Care, 2018. **41**(Suppl 1): p. S105-s118.
25. American Diabetes Association, *2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes-2019.* Diabetes Care, 2019. **42**(Suppl 1): p. S13-S28.
26. Vinik, A.I., et al., *Diabetic neuropathy.* Endocrinol Metab Clin North Am, 2013. **42**(4): p. 747-87.
27. American Diabetes Association, *Standards of medical care in diabetes--2014.* Diabetes Care, 2014. **37 Suppl 1**: p. S14-80.
28. Papadopoulou-Marketou, N., et al., *Diabetic nephropathy in type 1 diabetes.* Minerva Med, 2018. **109**(3): p. 218-228.
29. Meza Letelier, C.E., et al., *Pathophysiology of diabetic nephropathy: a literature review.* Medwave, 2017. **17**(1): p. e6839.
30. Mishra, S.C., et al., *Diabetic foot.* BMJ, 2017. **359**: p. j5064.
31. Jimenez, S., et al., *Trends in the incidence of lower limb amputation after implementation of a Multidisciplinary Diabetic Foot Unit.* Endocrinol Diabetes Nutr, 2017. **64**(4): p. 188-197.
32. American Diabetes Association, *9. Cardiovascular Disease and Risk Management: Standards of Medical Care in Diabetes-2018.* Diabetes Care, 2018. **41**(Suppl 1): p. S86-s104.
33. Romon, I., et al., *Prevalence of macrovascular complications and cardiovascular risk factors in people treated for diabetes and living in France : The ENTRED study 2001.* Diabetes Metab, 2008. **34**(2): p. 140-7.
34. Fisher, L., et al., *Prevalence of depression in Type 1 diabetes and the problem of over-diagnosis.* Diabet Med, 2016. **33**(11): p. 1590-1597.
35. Knapp, S., *Diabetes and infection: is there a link?--A mini-review.* Gerontology, 2013. **59**(2): p. 99-104.
36. Kulshrestha, V. and N. Agarwal, *Maternal complications in pregnancy with diabetes.* J Pak Med Assoc, 2016. **66**(9 Suppl 1): p. S74-7.
37. Haute Autorité de Santé, *Diabète de type 1 de l'adulte.* Guide affection de longue durée. Saint Denis La Plaine: HAS, 2007.
38. Haute Autorité de Santé, *Agence nationale de sécurité du médicament et des produits de santé*

(ANSM): Stratégie médicamenteuse du contrôle glycémique du diabète de type 2: Recommandation de bonne pratique. Recommandation de bonne pratique, 2013.

39. Assurance Maladie. *Les traitements non médicamenteux du diabète*. 2019 [cited 2019; Available from: https://www.ameli.fr/assure/sante/themes/diabete-traitement/traitements-non-medicamenteux.
40. American Diabetes Association, *Physical activity/exercise and diabetes.* Diabetes care, 2004. **27**(suppl 1): p. s58-s62.
41. World Health Organization, *ATC/DDD Index 2011*, in *WHO Collaborating Centre for Drug Statistics Methodology*. 2010, World Health Organization Oslo: Oslo.
42. American Diabetes Association, *8. Pharmacologic Approaches to Glycemic Treatment: Standards of Medical Care in Diabetes-2018.* Diabetes Care, 2018. **41**(Suppl 1): p. S73-S85.
43. American Diabetes Association, *Continuous subcutaneous insulin infusion.* Diabetes Care, 2004. **27 Suppl 1**: p. S110.
44. Skyler, J.S., *Diabetic complications. The importance of glucose control.* Endocrinol Metab Clin North Am, 1996. **25**(2): p. 243-54.
45. Assurance Maladie. *Comprendre l'autosurveillance de la glycémie*. 2019 [cited 2019; Available from: https://www.ameli.fr/assure/sante/themes/autosurveillance-glycemie/autosurveillance-glycemie.
46. Borot, S., et al., *Practical implementation, education and interpretation guidelines for continuous glucose monitoring: A French position statement.* Diabetes Metab, 2018. **44**(1): p. 61-72.
47. Haute Autorité de Santé, *La prise en charge de votre maladie, le diabète de type 2 de l'adulte*, in *Vivre avec un diabète de type*. 2010.
48. IDF, I., *IDF diabetes atlas seventh edition.* 2015.
49. Cho, N.H., et al., *IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045.* Diabetes Res Clin Pract, 2018. **138**: p. 271-281.
50. Thibault, V., et al., *Factors that could explain the increasing prevalence of type 2 diabetes among adults in a Canadian province: a critical review and analysis.* Diabetol Metab Syndr, 2016. **8**: p. 71.
51. Xu, G., et al., *Prevalence of diagnosed type 1 and type 2 diabetes among US adults in 2016 and 2017: population based study.* Bmj, 2018. **362**: p. k1497.
52. Maahs, D.M., et al., *Epidemiology of type 1 diabetes.* Endocrinol Metab Clin North Am, 2010. **39**(3): p. 481-97.
53. Tamayo, T., et al., *Diabetes in Europe: an update.* Diabetes Res Clin Pract, 2014. **103**(2): p. 206-17.
54. Magliano, D.J., et al., *Trends in incidence of total or type 2 diabetes: systematic review.* BMJ, 2019. **366**: p. l5003.
55. EURODIAB ACE Study Group, *Variation and trends in incidence of childhood diabetes in Europe. EURODIAB ACE Study Group.* Lancet, 2000. **355**(9207): p. 873-6.
56. Diaz-Valencia, P.A., P. Bougneres, and A.J. Valleron, *Global epidemiology of type 1 diabetes in young adults and adults: a systematic review.* BMC Public Health, 2015. **15**: p. 255.
57. Soltesz, G., et al., *Worldwide childhood type 1 diabetes incidence–what can we learn from*

epidemiology? Pediatric diabetes, 2007. **8**: p. 6-14.

58. Patterson, C.C., et al., *Is childhood-onset type I diabetes a wealth-related disease? An ecological analysis of European incidence rates.* Diabetologia, 2001. **44 Suppl 3**: p. B9-16.
59. Mayer-Davis, E.J., et al., *The many faces of diabetes in American youth: type 1 and type 2 diabetes in five race and ethnic populations: the SEARCH for Diabetes in Youth Study.* Diabetes Care, 2009. **32 Suppl 2**: p. S99-101.
60. Mayer-Davis, E.J., et al., *Diabetes in African American youth: prevalence, incidence, and clinical characteristics: the SEARCH for Diabetes in Youth Study.* Diabetes Care, 2009. **32 Suppl 2**: p. S112-22.
61. Patterson, C.C., et al., *Trends in childhood type 1 diabetes incidence in Europe during 1989-2008: evidence of non-uniformity over time in rates of increase.* Diabetologia, 2012. **55**(8): p. 2142-7.
62. Johnson, R.J., et al., *Fat storage syndrome in Pacific peoples: a combination of environment and genetics?* Pac Health Dialog, 2014. **20**(1): p. 11-6.
63. Abuyassin, B. and I. Laher, *Diabetes epidemic sweeping the Arab world.* World J Diabetes, 2016. **7**(8): p. 165-74.
64. Majeed, A., et al., *Diabetes in the Middle-East and North Africa: an update.* Diabetes Res Clin Pract, 2014. **103**(2): p. 218-22.
65. Jaber, L.A., et al., *Epidemiology of Diabetes Among Arab Americans.* Diabetes Care, 2003. **26**(2): p. 308-313.
66. Fosse-Edorh, S. and A. Fagot Campagna, *Prévalence du diabète, état de santé et recours aux soins des personnes diabétiques originaires d'un pays du Maghreb et résidant en France métropolitaine. Numéro thématique. Santé et recours aux soins des migrants en France.* Bull Epidemiol Hebd, 2012(2-3-4): p. 35-6.
67. Deshpande, A.D., M. Harris-Hayes, and M. Schootman, *Epidemiology of diabetes and diabetes-related complications.* Phys Ther, 2008. **88**(11): p. 1254-64.
68. Andersson, T., A. Ahlbom, and S. Carlsson, *Diabetes Prevalence in Sweden at Present and Projections for Year 2050.* PLoS One, 2015. **10**(11): p. e0143084.
69. Townshend, T. and A. Lake, *Obesogenic environments: current evidence of the built and food environments.* Perspect Public Health, 2017. **137**(1): p. 38-44.
70. Fishman, E.I., A. Stokes, and S.H. Preston, *The dynamics of diabetes among birth cohorts in the U.S.* Diabetes Care, 2014. **37**(4): p. 1052-9.
71. Rancière, F., et al., *Exposure to Bisphenol A and Bisphenol S and Incident Type 2 Diabetes: A Case–Cohort Study in the French Cohort DESIR.* Environmental health perspectives, 2019. **127**(10): p. 107013.
72. Jansson, S.P., et al., *Prevalence and incidence of diabetes mellitus: a nationwide population-based pharmaco-epidemiological study in Sweden.* Diabet Med, 2015. **32**(10): p. 1319-28.
73. Karpati, T., et al., *Towards a subsiding diabetes epidemic: trends from a large population-based study in Israel.* Popul Health Metr, 2014. **12**(1): p. 32.
74. Chen, L., D.J. Magliano, and P.Z. Zimmet, *The worldwide epidemiology of type 2 diabetes mellitus--present and future perspectives.* Nat Rev Endocrinol, 2011. **8**(4): p. 228-36.
75. Carstensen, B., et al., *The Danish National Diabetes Register: trends in incidence, prevalence and*

mortality. Diabetologia, 2008. **51**(12): p. 2187-96.

76. Alyafei, F., et al., *Incidence of type 1 and type 2 diabetes, between 2012-2016, among children and adolescents in Qatar.* Acta Biomed, 2018. **89**(S5): p. 7-10.
77. Hernandez-Montoya, D., et al., *Variation in incidence of type 2 diabetes mellitus: time series of Mexican adolescents.* Ann Epidemiol, 2019. **30**: p. 15-21.
78. Geiss, L.S., et al., *Prevalence and incidence trends for diagnosed diabetes among adults aged 20 to 79 years, United States, 1980-2012.* JAMA, 2014. **312**(12): p. 1218-26.
79. Selvin, E. and M.K. Ali, *Declines in the Incidence of Diabetes in the U.S.-Real Progress or Artifact?* Diabetes Care, 2017. **40**(9): p. 1139-1143.
80. Geiss, L.S., et al., *Considerations in Epidemiologic Definitions of Undiagnosed Diabetes.* Diabetes Care, 2018. **41**(9): p. 1835-1838.
81. Menke, A., et al., *Prevalence of and Trends in Diabetes Among Adults in the United States, 1988-2012.* JAMA, 2015. **314**(10): p. 1021-9.
82. Moody, A., et al., *Social inequalities in prevalence of diagnosed and undiagnosed diabetes and impaired glucose regulation in participants in the Health Surveys for England series.* BMJ Open, 2016. **6**(2): p. e010155.
83. Heidemann, C., et al., *Temporal changes in the prevalence of diagnosed diabetes, undiagnosed diabetes and prediabetes: findings from the German Health Interview and Examination Surveys in 1997-1999 and 2008-2011.* Diabet Med, 2016. **33**(10): p. 1406-14.
84. Jorgensen, M.E., et al., *Estimates of prediabetes and undiagnosed type 2 diabetes in Denmark: The end of an epidemic or a diagnostic artefact?* Scand J Public Health, 2018: p. 1403494818799606.
85. Geiss, L.S., et al., *Trends in type 2 diabetes detection among adults in the USA, 1999-2014.* BMJ Open Diabetes Res Care, 2018. **6**(1): p. e000487.
86. Bocquet, V., et al., *Public health burden of pre-diabetes and diabetes in Luxembourg: finding from the 2013-2015 European Health Examination Survey.* BMJ Open, 2019. **9**(1): p. e022206.
87. Murphy, S.L., et al., *Mortality in the United States, 2017.* 2018.
88. I. D. F. Diabetes Atlas Group, *Update of mortality attributable to diabetes for the IDF Diabetes Atlas: Estimates for the year 2013.* Diabetes Res Clin Pract, 2015. **109**(3): p. 461-5.
89. Gregg, E.W., et al., *Trends in death rates among U.S. adults with and without diabetes between 1997 and 2006: findings from the National Health Interview Survey.* Diabetes Care, 2012. **35**(6): p. 1252-7.
90. Ma, J., et al., *Temporal Trends in Mortality in the United States, 1969-2013.* JAMA, 2015. **314**(16): p. 1731-9.
91. Rubens, M., et al., *Trends in Diabetes-Related Preventable Hospitalizations in the U.S., 2005-2014.* Diabetes Care, 2018. **41**(5): p. e72-e73.
92. American Diabetes Association, *Economic Costs of Diabetes in the U.S. in 2017.* Diabetes Care, 2018. **41**(5): p. 917-928.
93. Santé Publique France, *Le poids du diabète en France en 2016. Synthèse épidemilogique* S.P. France, Editor. 2018.
94. Mandereau Bruno, L. and S. Fosse Edorh, *Prévalence du diabète traité pharmacologiquement*

(tous types) en France en 2015. Disparités territoriales et socio-économiques .
Prevalence of pharmacologically-treated diabetes (all types) in France in 2015. Territorial and socio-economic disparities. Bull Epidémiologique Hebd, 2017. **(27-28)**: p. 586-91.
95. Santé Publique France. *GEODES (Géo Données en Santé Publique)*. 2019 [cited 2019 November]; Available from: https://geodes.santepubliquefrance.fr/#bbox=-1013725,6621568,2495144,1540257&c=indicator&f=0&i=diabete.diabete_tx_std&s=2010&t=a01&view=map1.
96. Piffaretti, C., et al., *Trends in childhood type 1 diabetes incidence in France, 2010-2015.* Diabetes Res Clin Pract, 2018.
97. Bonaldi, C., et al., *A first national prevalence estimate of diagnosed and undiagnosed diabetes in France in 18- to 74-year-old individuals: the French Nutrition and Health Survey 2006/2007.* Diabet Med, 2011. **28**(5): p. 583-9.
98. Mandereau-Bruno, L., et al., *Evolution de la mortalité et de la surmortalité à 5 ans des personnes diabétiques traitées pharmacologiquement en France métropolitaine: comparaison des cohortes ENTRED 2001 et ENTRED 2007.* Bull Epidémiologique Hebd, 2016. **37-38**: p. 668-675.
99. Dastani, Z., et al., *Novel loci for adiponectin levels and their influence on type 2 diabetes and metabolic traits: a multi-ethnic meta-analysis of 45,891 individuals.* PLoS Genet, 2012. **8**(3): p. e1002607.
100. Fosse-Edorh, S., L. Mandereau Bruno, and A. Hartemann Heurtier, *Les hospitalisations pour complications podologiques chez les personnes diabétiques traitées pharmacologiquement en France en 2013. Numéro thématique. Journée mondiale du diabète 2015. Suivi du diabète et poids de ses complications sévères en France.* Bull Epidemiol Hebd, 2015(34-35): p. 638-44.
101. Fosse-Edorh, S., L. Mandereau Bruno, and V. Olie, *Les hospitalisations pour infarctus du myocarde ou accident vasculaire cérébral chez les personnes diabétiques traitées pharmacologiquement, en France en 2013. Numéro thématique. Journée mondiale du diabète 2015. Suivi du diabète et poids de ses complications sévères en France.* Bull Epidemiol Hebd, 2015(34-35): p. 625-31.
102. Chevreul, K., K. Berg Brigham, and C. Bouche, *The burden and treatment of diabetes in France.* Global Health, 2014. **10**: p. 6.
103. Douglas Harper. *Online etimology dictionary*. 2001-2019 [cited 2019; Available from: https://www.etymonline.com/word/surveillance.
104. Choi, B.C., *The past, present, and future of public health surveillance.* Scientifica (Cairo), 2012. **2012**: p. 875253.
105. Centers for Disease Control Prevention, et al., *Indicators for chronic disease surveillance.* MMWR Recomm Rep, 2004. **53**(RR-11): p. 1-6.
106. Gil, G.P. and R. Gálvez, *Medicina preventiva y salud pública*. 2015, Barcelona: Masson.
107. Saydah, S. and G. Imperatore, *Emerging Approaches in Surveillance of Type 1 Diabetes.* Curr Diab Rep, 2018. **18**(9): p. 61.
108. Lix, L.M., et al., *Population-based data sources for chronic disease surveillance.* Chronic Dis Can, 2008. **29**(1): p. 31-8.
109. Centers for Disease Control Prevention. *National health and nutrition examination survey, survey*

methods and analytic guidelines. 2019 [cited 2019 November]; Available from: https://www.cdc.gov/nchs/nhanes/index.htm.

110. Stommel, M. and C.A. Schoenborn, *Accuracy and usefulness of BMI measures based on self-reported weight and height: findings from the NHANES & NHIS 2001-2006.* BMC Public Health, 2009. **9**(1): p. 421.
111. Grabovac, I., et al., *Association of depression symptoms with receipt of healthcare provider advice on physical activity among US adults.* J Affect Disord, 2019.
112. World Health Organization, *The current and future use of registers in health information systems*, D.o.h.i.a. statistics, Editor. 1974, World Health Organization: Geneva.
113. Skrivarhaug, T., *Norwegian Childhood Diabetes Registry: Childhood onset diabetes in Norway 1973-2012.* Norsk epidemiologi, 2013. **23**(1).
114. Skrivarhaug, T., et al., *Incidence of type 1 diabetes in Norway among children aged 0–14 years between 1989 and 2012: has the incidence stopped rising? Results from the Norwegian Childhood Diabetes Registry.* Diabetologia, 2014. **57**(1): p. 57-62.
115. Iversen, H.H., O. Bjertnaes, and T. Skrivarhaug, *Associations between adolescent experiences, parent experiences and HbA1c: results following two surveys based on the Norwegian Childhood Diabetes Registry (NCDR).* BMJ Open, 2019. **9**(11): p. e032201.
116. Green, A., E.A. Gale, and C.C. Patterson, *Incidence of childhood-onset insulin-dependent diabetes mellitus: the EURODIAB ACE Study.* Lancet, 1992. **339**(8798): p. 905-9.
117. Richesson, R.L., *Data standards in diabetes patient registries.* J Diabetes Sci Technol, 2011. **5**(3): p. 476-85.
118. Gavrielov-Yusim, N. and M. Friger, *Use of administrative medical databases in population-based research.* J Epidemiol Community Health, 2014. **68**(3): p. 283-7.
119. Zhong, V.W., et al., *Use of administrative and electronic health record data for development of automated algorithms for childhood diabetes case ascertainment and type classification: the SEARCH for Diabetes in Youth Study.* Pediatr Diabetes, 2014. **15**(8): p. 573-84.
120. Bradley, S.H., N.R. Lawrence, and P. Carder, *Using primary care data for health research in England - an overview.* Future Healthc J, 2018. **5**(3): p. 207-212.
121. Tate, A.R., et al., *Quality of recording of diabetes in the UK: how does the GP's method of coding clinical data affect incidence estimates? Cross-sectional study using the CPRD database.* BMJ Open, 2017. **7**(1): p. e012905.
122. Lix, L., et al., *The Canadian Chronic Disease Surveillance System: A model for collaborative surveillance.* International Journal of Population Data Science, 2018. **3**(3).
123. Huff, L., et al., *Using hospital discharge data for disease surveillance.* Public Health Reports, 1996. **111**(1): p. 78.
124. United Nations *Types of vital statistics available in different countries.* Bull World Health Organ, 1954. **11**(1-2): p. 177-99.
125. Rey, G., K. Bounebache, and C. Rondet, *Causes of deaths data, linkages and big data perspectives.* J Forensic Leg Med, 2018. **57**: p. 37-40.
126. Richards, C.L., M.F. Iademarco, and T.C. Anderson, *A new strategy for public health surveillance at CDC: improving national surveillance activities and outcomes.* Public Health Reports, 2014.

129(6): p. 472-476.

127. Lozano, R., et al., *Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010.* Lancet, 2012. **380**(9859): p. 2095-128.
128. Cheng, W.S., et al., *Sensitivity and specificity of death certificates for diabetes: as good as it gets?* Diabetes Care, 2008. **31**(2): p. 279-84.
129. Mues, K.E., et al., *Use of the Medicare database in epidemiologic and health services research: a valuable source of real-world evidence on the older and disabled populations in the US.* Clinical epidemiology, 2017. **9**: p. 267.
130. Riley, G.F., *Administrative and claims records as sources of health care cost data.* Med Care, 2009. **47**(7 Suppl 1): p. S51-5.
131. Moulis, G., et al., *French health insurance databases: What interest for medical research?* Rev Med Interne, 2015. **36**(6): p. 411-7.
132. Muggah, E., et al., *Ascertainment of chronic diseases using population health data: a comparison of health administrative data and patient self-report.* BMC Public Health, 2013. **13**: p. 16.
133. Read, S.H., et al., *Trends in type 2 diabetes incidence and mortality in Scotland between 2004 and 2013.* Diabetologia, 2016. **59**(10): p. 2106-13.
134. Livingstone, S.J., et al., *Estimated life expectancy in a Scottish cohort with type 1 diabetes, 2008-2010.* JAMA, 2015. **313**(1): p. 37-44.
135. Hamer, M., et al., *Temporal trends in diabetes prevalence and key diabetes risk factors in Scotland, 2003-2008.* Diabet Med, 2011. **28**(5): p. 595-8.
136. Carstensen, B. and K. Borch-Johnsen, *Register-based studies of diabetes.* Scand J Public Health, 2011. **39**(7 Suppl): p. 175-9.
137. Andersen, G.S., et al., *Diabetes among migrants in Denmark: Incidence, mortality, and prevalence based on a longitudinal register study of the entire Danish population.* Diabetes Res Clin Pract, 2016. **122**: p. 9-16.
138. Svensson, J., et al., *Danish Registry of Childhood and Adolescent Diabetes.* Clin Epidemiol, 2016. **8**: p. 679-683.
139. Centers for Disease Control Prevention, *National diabetes statistics report, 2017*. 2017, Centers for Disease Control and Prevention, US Department of Health and Human Service: Atlanta.
140. Desai, J., et al., *Public health surveillance of diabetes in the United States.* J Public Health Manag Pract, 2003. **Suppl**: p. S44-51.
141. Geiss, L.S., et al., *Surveillance for diabetes mellitus--United States, 1980-1989.* MMWR CDC Surveill Summ, 1993. **42**(2): p. 1-20.
142. Wang, J., et al., *Declining death rates from hyperglycemic crisis among adults with diabetes, U.S., 1985-2002.* Diabetes Care, 2006. **29**(9): p. 2018-22.
143. Andes, L.J., et al., *Diabetes Prevalence and Incidence Among Medicare Beneficiaries - United States, 2001-2015.* MMWR Morb Mortal Wkly Rep, 2019. **68**(43): p. 961-966.
144. Ciderova, D. and V. Repasova, *Geo-heterogeneity in the context of the EU.* European Scientific Journal, 2013. **9**(25).
145. Rodwin, V.G., *The health care system under French national health insurance: lessons for health*

reform in the United States. Am J Public Health, 2003. **93**(1): p. 31-7.

146. Fosse-Edorh, S., L. Mandereau Bruno, and C. Piffaretti, *Le poids du diabète en France en 2016. Synthèse épidémiologique.* 2018, Santé Publique France: Saint-Maurice.

147. Mauny, F., et al., *Increasing trend of childhood Type 1 diabetes in Franche-Comté (France): Analysis of age and period effects from 1980 to 1998.* European journal of epidemiology, 2005. **20**(4): p. 325-329.

148. Trellu, M., et al., *Epidemiology of diabetes in children in Languedoc-Roussillon (France).* Archives de pediatrie: organe officiel de la Societe francaise de pediatrie, 2015. **22**(3): p. 241-246.

149. Andreeva, V.A., et al., *Comparison of Dietary Intakes Between a Large Online Cohort Study (Etude NutriNet-Sante) and a Nationally Representative Cross-Sectional Study (Etude Nationale Nutrition Sante) in France: Addressing the Issue of Generalizability in E-Epidemiology.* Am J Epidemiol, 2016. **184**(9): p. 660-669.

150. Balicco, A., et al., *Protocole Esteban: une Étude transversale de santé sur l'environnement, la biosurveillance, l'activité physique et la nutrition (2014–2016).* Toxicologie analytique et clinique, 2017. **29**(4): p. 517-537.

151. Santé Publique France. *Lancement de la 3e édition de l'étude Entred (Echantillon National Témoin Représentatif des personnes Diabétiques)*. 2019 July 2019 [cited 2019 09/09/2019]; Available from: https://www.santepubliquefrance.fr/etudes-et-enquetes/entred-3

152. Pornet, C., et al., *Trends in the quality of care for elderly people with type 2 diabetes: the need for improvements in safety and quality (the 2001 and 2007 Entred surveys).* Diabetes Metab, 2011. **37**: p. 152-61.

153. Marant, C., et al., *French medical practice in type 2 diabetes: the need for better control of cardiovascular risk factors.* Diabetes Metab, 2008. **34**(1): p. 38-45.

154. Romon, I., et al., *The excess mortality related to cardiovascular diseases and cancer among adults pharmacologically treated for diabetes--the 2001-2006 ENTRED cohort.* Diabet Med, 2014. **31**(8): p. 946-53.

155. Romon, I., et al., *The burden of diabetes-related mortality in France in 2002: an analysis using both underlying and multiple causes of death.* Eur J Epidemiol, 2008. **23**(5): p. 327-34.

156. Système National de Données Santé. *Qu'est-ce que le SNDS ?* 2019 [cited 2019 November]; Available from: https://www.snds.gouv.fr/SNDS/Qu-est-ce-que-le-SNDS.

157. Bellanger, M.M. and L. Tardif, *Accounting and reimbursement schemes for inpatient care in France.* Health Care Management Science, 2006. **9**(3): p. 295-305.

158. Tuppin, P., et al., *Health care use by free complementary health insurance coverage beneficiaries in France in 2012.* Revue d'epidemiologie et de sante publique, 2016. **64**(2): p. 67-78.

159. Assurance Maladie. *Qu'est-ce qu'une affection de longue durée (ALD) ?* 2019 [cited 2019 November]; Available from: https://www.ameli.fr/medecin/exercice-liberal/presciption-prise-charge/situation-patient-ald-affection-longue-duree/definition-ald.

160. Tuppin, P., et al., *Characteristics and management of diabetic patients hospitalized for myocardial infarction in France.* Diabetes Metab, 2010. **36**(2): p. 129-36.

161. Tuppin, P., et al., *Value of a national administrative database to guide public decisions: From the systeme national d'information interregimes de l'Assurance Maladie (SNIIRAM) to the systeme*

national des donnees de sante (SNDS) in France. Rev Epidemiol Sante Publique, 2017. **65 Suppl 4**: p. S149-s167.

162. Boudemaghe, T. and I. Belhadj, *Data Resource Profile: The French National Uniform Hospital Discharge Data Set Database (PMSI).* Int J Epidemiol, 2017. **46**(2): p. 392-392d.
163. Inserm. *Le CépiDC: Une mission légale et historique de l'Inserm*. 2019 [cited 2019 November]; Available from: https://www.cepidc.inserm.fr/qui-sommes-nous/le-cepidc.
164. Aouba, A., et al., *Données sur la mortalité en France : principales causes de décès en 2008 et évolutions depuis 2000.* Bull Epidemiol Hebd, 2011(22): p. 249-55.
165. Goldberg, M., et al., *The opening of the French national health database: Opportunities and difficulties. The experience of the Gazel and Constances cohorts.* Rev Epidemiol Sante Publique, 2016. **64**(4): p. 313.
166. Zins, M., et al., *The CONSTANCES cohort: an open epidemiological laboratory.* BMC Public Health, 2010. **10**: p. 479.
167. Zins, M. and M. Goldberg, *The French CONSTANCES population-based cohort: design, inclusion and follow-up.* European journal of epidemiology, 2015. **30**(12): p. 1317-1328.
168. Pornet, C., et al., *Trends in the quality of care for elderly people with type 2 diabetes: the need for improvements in safety and quality (the 2001 and 2007 ENTRED Surveys).* Diabetes Metab, 2011. **37**(2): p. 152-61.
169. Quantin, C., et al., *Using algorithms to identify cases of depression in the SNIIRAM database by the REDSIAM network.* Revue française des affaires sociales, 2016(2): p. 201-225.
170. Fosse-Edorh, S., et al., *Algorithms based on medico-administrative data in the field of endocrine, nutritional and metabolic diseases, especially diabetes.* Rev Epidemiol Sante Publique, 2017. **65 Suppl 4**: p. S168-s173.
171. Assurance Maladie. *SNIIRAM: Formation DCIR/PMSI*. in *Formation SNIIRAM*. 2016. Paris.
172. Leong, A., et al., *Systematic review and meta-analysis of validation studies on a diabetes case definition from health administrative records.* PLoS One, 2013. **8**(10): p. e75256.
173. Ricci, P., et al., *Reimbursed health expenditures during the last year of life, in France, in the year 2008.* Rev Epidemiol Sante Publique, 2013. **61**(1): p. 29-36.
174. Perlbarg, J., et al., *Feasibility and practical value of statistical matching of a general practice database and a health insurance database applied to diabetes and hypertension.* Sante publique, 2013. **26**(3): p. 355-363.
175. Ricci, P., et al., *Diabète traité : quelles évolutions entre 2000 et 2009 en France ?* Bull Epidemiol Hebd, 2010(42-43): p. 425-31.
176. Ruiz, P.L.D., et al., *Decreasing incidence of pharmacologically and non-pharmacologically treated type 2 diabetes in Norway: a nationwide study.* Diabetologia, 2018. **61**(11): p. 2310-2318.
177. Pace, M., et al., *Revision of the European standard population. Report of the Eurostat's task force*. 2013, Publications Office of the European Union, 2013: Luxembourg.
178. Lin, S., et al., *Diabetes incidence and projections from prevalence surveys in Samoa over 1978-2013.* Int J Public Health, 2017. **62**(6): p. 687-694.
179. Mandereau-Bruno, L., et al., *Prévalence du diabète traité pharmacologiquement et disparités territoriales en France en 2012.* Bull épidémiologique Hebd, 2014: p. 30-31.

180. Fuentes, S., et al., *Prevalence of Prediabetes and Undiagnosed Diabetes among Adults Aged 18 to 70 Years in France—The CONSTANCES Cohort.* Diabetes, 2018. **67**(Supplement 1): p. 1657-P.

181. Imes, C.C. and L.E. Burke, *The Obesity Epidemic: The United States as a Cautionary Tale for the Rest of the World.* Curr Epidemiol Rep, 2014. **1**(2): p. 82-88.

182. Verdot, C., et al., *Corpulence des enfants et des adultes en France métropolitaine en 2015. Résultats de l'étude Esteban et évolution depuis 2006.* Bull Epidemiol Hebd, 2017(13): p. 234-41.

183. Eschwège, E., et al., *Enquête épidémiologique nationale sur le surpoids et l'obésité.* ObEpi Roche, 2012.

184. Santé Publique France *L'état de santé de la population en France: rapport 2017*. 2017.

185. Hilawe, E.H., et al., *Differences by sex in the prevalence of diabetes mellitus, impaired fasting glycaemia and impaired glucose tolerance in sub-Saharan Africa: a systematic review and meta-analysis.* Bull World Health Organ, 2013. **91**(9): p. 671-682D.

186. Sobers-Grannum, N., et al., *Female gender is a social determinant of diabetes in the Caribbean: a systematic review and meta-analysis.* PLoS One, 2015. **10**(5): p. e0126799.

187. Carrere, P., et al., *Diabetes mellitus and obesity in the French Caribbean: A special vulnerability for women?* Women Health, 2018. **58**(2): p. 145-159.

188. Favier, F., et al., *Prevalence of Type 2 diabetes and central adiposity in La Reunion Island, the REDIA Study.* Diabetes Res Clin Pract, 2005. **67**(3): p. 234-42.

189. Daigre, J.L., et al., *The prevalence of overweight and obesity, and distribution of waist circumference, in adults and children in the French Overseas Territories: the PODIUM survey.* Diabetes Metab, 2012. **38**(5): p. 404-11.

190. Kuhn, M., *Building Predictive Models in R Using the caret Package.* Journal of Statistical Software, 2008. **28**.

191. Kononenko, I. *Estimating attributes: Analysis and extensions of RELIEF*. 1994. Berlin, Heidelberg: Springer Berlin Heidelberg.

192. Robnik-Šikonja, M. and I. Kononenko. *An adaptation of Relief for attribute estimation in regression*. in *Machine Learning: Proceedings of the Fourteenth International Conference (ICML'97)*. 1997.

193. Kira, K. and L.A. Rendell, *The feature selection problem: Traditional methods and a new algorithm*, in *Proceedings of the Eleventh International Conference on Machine Learning*. 1992: San Jose, California.

194. Holman, N., B. Young, and R. Gadsby, *Current prevalence of Type 1 and Type 2 diabetes in adults and children in the UK.* Diabet Med, 2015. **32**(9): p. 1119-20.

195. American Diabetes Association, *9. Pharmacologic Approaches to Glycemic Treatment: Standards of Medical Care in Diabetes-2019.* Diabetes Care, 2019. **42**(Suppl 1): p. S90-s102.

196. Almutairi, N., H. Hosseinzadeh, and V. Gopaldasani, *The effectiveness of patient activation intervention on type 2 diabetes mellitus glycemic control and self-management behaviors: A systematic review of RCTs.* Prim Care Diabetes, 2019.

197. Bebu, I., et al., *Mediation of the Effect of Glycemia on the Risk of CVD Outcomes in Type 1 Diabetes: The DCCT/EDIC Study.* Diabetes Care, 2019. **42**(7): p. 1284-1289.

198. Czupryniak, L., *Guidelines for the management of type 2 diabetes: is ADA and EASD consensus more clinically relevant than the IDF recommendations?* Diabetes research and clinical practice, 2009. **86**: p. S22-S25.

199. Hilawe, E.H., et al., *Prevalence and predictors of prediabetes and diabetes among adults in Palau: population-based national STEPS survey.* Nagoya J Med Sci, 2016. **78**(4): p. 475-483.

200. Bitar, D., et al., *Mycoses invasives en France métropolitaine, PMSI 2001-2010 : incidence, létalité et tendances. Numéro thématique. Mycoses invasives en France : épidémiologie, enjeux diagnostiques et thérapeutiques.* Bull Epidemiol Hebd, 2013(12-13): p. 109-14.

201. Happe, A. and E. Drezen, *A visual approach of care pathways from the French nationwide SNDS database - from population to individual records: the ePEPS toolbox.* Fundam Clin Pharmacol, 2017.

202. Kira, K. and L.A. Rendell, *A practical approach to feature selection*, in *Machine Learning Proceedings 1992*. 1992, Elsevier. p. 249-256.

Icons copyrights:

Icons made by Eucalyp from www.flaticon.com

Icons made by surang from www.flaticon.com

Icons made by Smashicons from www.flaticon.com

Icons made by Freepik from www.flaticon.com