Development of an algorithm to identify the type of diabetes in the French administrative health care database "Système national des données de santé" (SNDS)

Bretin O[1], Casarotto E[1], Bessou A[1], Maurel F[1], Serusclat P[2], Joubert M[3], Fagherazzi G[4], Berteau C[5], Pouyet A[6], Maillard C[1]*

[1] IQVIA France, Courbevoie, France ; [2] Groupe Hospitalier Mutualiste Les Portes du Sud, Venissieux, France ; [3] CHU de Caen, Caen, France ; [4] Paris south Paris Saclay Univrsity, Villejuif, France ; [5] Roche Diabetes Care France, Myelan, 38, France ; [6] TIMKL, Montbonnot Saint Martin, France

**Corresponding author*

IMS Health & Quintiles are now IQVIA™

Roche

CONTEXT & OBJECTIVE

- **Diabetes** is a chronic disease characterized by elevated levels of blood glucose. There are **two main types**: type-1 diabetes (T1D) and type-2 diabetes (T2D). Epidemiological data describing these types of patients separately are limited.
- The French administrative health care database (SNDS) is a powerful tool for epidemiological and pharmacoeconomic studies. However, its **lack of clinical information** makes it difficult to accurately identify the type of diabetes.
- The objective was **to develop an accurate machine learning algorithm to determine the type of diabetes in the SNDS**, validated thanks to a **linkage with primary care clinical data**.

METHOD

► Design

- **Data source:** Electronic medical records (EMR) of a network of French general practitioners (GP) were probabilistically linked with the SNDS (Fig. 1), over the period 2010-2019.
- **Population:** A cohort of **adult patients with diabetes** was extracted from the **EMR** database. The **type of diabetes** of these patients was **known in the EMRs**, via the diagnoses provided by the physician during consultations.
- **Follow-up:** Patients were followed from the first identification of their diabetes in the SNDS, until death, loss of follow-up or December 31, 2019.

► A 4-step machine learning approach

- **Derivation of predictors:**

➢ After a literature review and a pre-selection based on experts' opinion, about **200 predictors** were derived from SNDS data to help discriminate between T1D and T2D.

➢ They included **socio-demographics**, **Long-Term Diseases** (LTD), **comorbidities**, **hospitalizations**, and reimbursements of **treatments**, **medical devices**, **biological tests**, **medical procedures** and **consultations**, collected up to 5 years before the end of patients' follow-up.

- **Data splitting:** The cohort was randomly divided into a training set (80%) and a test set (20%).
- **Model training and optimization:** Various machine learning algorithms (non-penalized and penalized logistic regressions, RandomForest, XGBoost) were trained and optimized by a 10-fold cross-validation procedure on the training set.
- **Evaluation of model performance:** The best model was selected for its ability to predict T1D on the test set, via the F1-score metric (harmonic mean of precision and sensitivity).

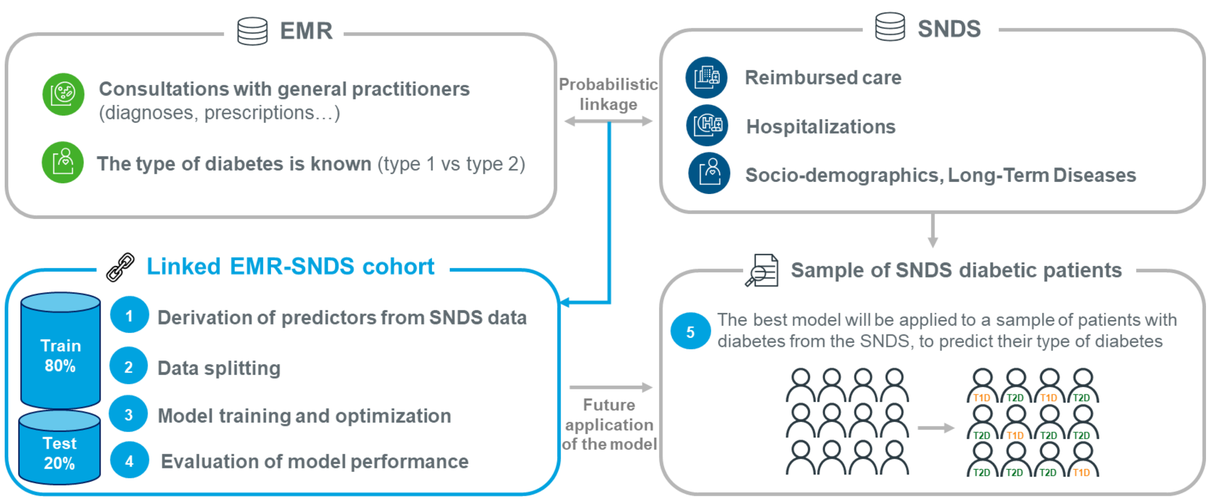

Figure 1. Study design

► Comparison with models from the literature

- After reviewing existing algorithms that classify types of diabetes in the SNDS, two complementary approaches[1,2] (one by decision tree and the other via machine learning) were replicated on our cohort.
- The decision tree used by Charbonnel et al[1] (Fig. 2) was reproduced in the 2-year period preceding the end of follow-up.
- Machine learning models from Fuentes et al[2] (logistic regressions and linear discriminant analyses using 3, 9 or 14 variables) were retrained on our training set and evaluated on the test set.

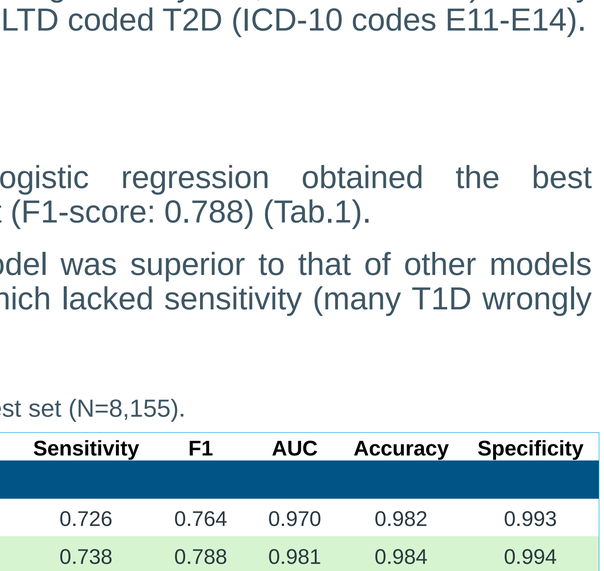

Figure 2. Decision tree to classify types of diabetes[1]

RESULTS

► Patients' characteristics

- The population consisted of **40,774 adult patients with diabetes**, including **39,122 (95.9%) T2D** (mean age: 70.7 years, males: 56.7%) and **1,652 (4.1%) T1D** (mean age: 55.6 years, males: 60.1%). Nearly 27% of T1D patients had an LTD coded T2D (ICD-10 codes E11-E14).

► Model performances

- The LASSO penalized logistic regression obtained the best performances on the test set (F1-score: 0.788) (Tab.1).
- The performance of this model was superior to that of other models identified in the literature, which lacked sensitivity (many T1D wrongly classified as T2D).

Table 1. Model performances on the test set (N=8,155).

Model	Precision	Sensitivity	F1	AUC	Accuracy	Specificity
Our models						
Logistic regression (LR)	0.807	0.726	0.764	0.970	0.982	0.993
Penalized LR - LASSO	0.846	0.738	0.788	0.981	0.984	0.994
Penalized LR - RIDGE	0.853	0.689	0.762	0.978	0.983	0.995
Penalized LR - ELASTIC NET	0.821	0.713	0.763	0.979	0.982	0.993
RandomForest	0.859	0.631	0.728	0.975	0.981	0.996
XGBoost	0.844	0.692	0.760	0.982	0.982	0.995
Models from the literature						
Charbonnel et al[1]						
Decision tree	0.688	0.585	0.633	-	0.973	0.989
Fuentes et al[2]						
LR with 9 variables*	0.756	0.482	0.588	0.936	0.973	0.993

*Only the model that obtained the best F1-score among the retrained models is displayed.

► Important predictors

- The LASSO penalized logistic regression selected 66 variables.
- The most important variables in our models were the **time between the first identification of diabetes and the first insulin**, the **age** at first identification of diabetes in the SNDS, the presence of an LTD coded T2D, and the **delivery of different anti-diabetic treatments**.

► Limits

- The analysis of model errors showed that patients with intermediate characteristics between T1D and T2D could be misclassified.

Conclusion

- **The innovative linkage between SNDS and EMR data enabled to develop a high-performance classification model that outperforms existing published algorithms to identify the type of diabetes in the SNDS.**
- **The methodology could be reused by the scientific community to conduct epidemiological and pharmacoeconomic studies on each type of diabetes in France.**
- **The linkage between SNDS and EMR could also be replicated to identify other chronic diseases in the SNDS, to facilitate epidemiological research in other disease areas.**

1. Charbonnel B, Simon D, Dallongeville J, Bureau I, Dejager S, Levy-Bachelot L, Gourmelen J, Detournay B. Direct Medical Costs of Type 2 Diabetes in France: An Insurance Claims Database Analysis. Pharmacoecon Open. 2018 Jun;2(2):209-219. doi: 10.1007/s41669-017-0050-3. PMID: 29623622; PMCID: PMC5972121.
2. Sonsoles Fuentes, R Hrzic, R Haneef, Sofiane Kab, Sandrine Fosse-Edorh, et al.. 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 de Santé. Congrès annuel de la Société Francophone du Diabète (SFD 2020), Sep 2020, Virtual conference, France. ⟨hal-03925742⟩. Available on: https://hal.science/hal-03925742

ISPOR Europe 2023, 12-15 November 2023, Copenhagen, Denmark

ISPOR Improving healthcare decisions