Knowledge graphs based extension of patients' files to predict hospitalization

Raphaël Gazzotti

THÈSE DE DOCTORAT

Prédiction d'hospitalisation par la génération de caractéristiques extraites de graphes de connaissances

Raphaël GAZZOTTI

Laboratoire d'Informatique, Signaux et Systèmes de Sophia Antipolis (UMR7271)

Présentée en vue de l'obtention du grade de docteur en Informatique de l'Université Côte d'Azur
Dirigée par :
Mr. Fabien GANDON, Directeur de Recherche, INRIA
Dirigée par :
Mme. Catherine FARON ZUCKER, Maître de conférences, Université Côte d'Azur
Soutenue le : 30 Avril 2020

Devant le jury, composé de :
Président du jury :
Mr. Andrea TETTAMANZI, Professeur, Université Côte d'Azur
Rapporteurs :
Mme. Sandra BRINGAY, Professeure, Université Paul Valéry Montpellier 3
Mme. Sylvie DESPRES, Professeure, Université Sorbonne Paris Nord
Examinateur :
Mr. Olivier DAMERON, Maître de conférences, Université de Rennes 1
Invité :
Mr. David DARMON, Maître de conférences, Université Côte d'Azur

Prédiction d'hospitalisation par la génération de caractéristiques extraites de graphes de connaissances

Résumé

L'utilisation de dossiers médicaux électroniques (DMEs) et la prescription électronique sont des priorités dans les différents plans d'action européens sur la santé connectée. Le développement du DME constitue une formidable source de données ; il capture tous les épisodes symptomatiques dans la vie d'un patient et doit permettre l'amélioration des pratiques médicales et de prises en charge, à la condition de mettre en place des procédures de traitement automatique.

A ce titre nous travaillons sur la prédiction d'hospitalisation à partir des DMEs et après les avoir représentés sous forme vectorielle, nous enrichissons ces modèles afin de profiter des connaissances issues de référentiels, qu'ils soient généralistes ou bien spécifiques dans le domaine médical et cela, dans le but d'améliorer le pouvoir prédictif d'algorithmes de classification automatique. Déterminer les connaissances à extraire dans l'objectif de les intégrer aux représentations vectorielles est à la fois une tâche subjective et destinée aux experts, nous verrons une procédure semi-supervisée afin d'automatiser en partie ce processus.

Du fruit de nos recherches, nous avons ébauché un produit destiné aux médecins généralistes afin de prévenir l'hospitalisation de leur patient ou du moins améliorer son état de santé. Ainsi, par le biais d'une simulation, il sera possible au médecin d'évaluer quels sont les facteurs impliqués dans le risque d'hospitalisation de son patient et de définir les actions préventives à planifier pour éviter l'apparition de cet événement.

Cet algorithme d'aide à la décision a pour visée d'être directement intégré au logiciel de consultation des médecins et nous avons pour ce faire développé une interface graphique élaborée en collaboration avec de nombreux corps de métiers avec notamment les premiers concernés, des médecins généralistes.

Mots-clefs : Modèle prédictif, Dossier médical électronique, Graphe de connaissances.

Knowledge graphs based extension of patients' files to predict hospitalization

Abstract

The use of electronic medical records (EMRs) and electronic prescribing are priorities in the various European action plans on connected health. The development of the EMR is a tremendous source of data; it captures all symptomatic episodes in a patient's life and should lead to improved medical and care practices, as long as automatic treatment procedures are set up.

As such, we are working on hospitalization prediction based on EMRs and after having represented them in vector form, we enrich these models in order to benefit from the knowledge resulting from referentials, whether generalist or specific in the medical field, in order to improve the predictive power of automatic classification algorithms. Determining the knowledge to be extracted with the objective of integrating it into vector representations is both a subjective task and intended for experts, we will see a semi-supervised procedure to partially automate this process.

As a result of our research, we designed a product for general practitioners to prevent their patients from being hospitalized or at least improve their health. Thus, through a simulation, it will be possible for the doctor to evaluate the factors involved in the risk of hospitalization of his patient and to define the preventive actions to be planned to avoid the occurrence of this event.

This decision support algorithm is intended to be directly integrated into the physician consultation software. For this purpose, we have developed a graphical interface developed in collaboration with many professional bodies, including the first to be concerned, general practitioners.

Keywords: Predictive model, Electronic medical record, Knowledge graph.

Remerciements

Je tiens à remercier mes directeurs de thèse, Catherine Faron Zucker et Fabien Gandon, pour leur rigueur scientifique, leur encadrement et leurs conseils avisés, sans qui cette thèse n'aurait pas vu le jour.

Je souhaite exprimer ma gratitude aux membres du jury pour avoir accepté d'examiner et d'évaluer cette thèse.

Merci à Félix pour m'avoir offert l'opportunité de travailler au sein de son entreprise sur des thématiques et avec des partenaires stratégiques.

Merci à Virginie et à David pour nous avoir permis de travailler avec les données de leur département, pour avoir apporté leur expertise et pour avoir exprimé une problématique leur tenant à cœur.

Merci à Elisa qui a apporté une dimension plus concrète à ce travail, en premier lieu scientifique, de part ses illustrations.

Merci à Christine pour son support à toutes épreuves dans les méandres des démarches administratives.

Merci aux membres de l'équipe WIMMICS qui m'ont permis de grandir aussi bien d'un point de vue professionnel, intellectuel que humain.

Je tiens tout particulièrement à remercier Noëlle et Hervé pour m'avoir soutenu et avoir été auprès de moi au cours de toutes ces années.

Contents

CONTENTS

List of Figures

List of Tables

Chapter 1

Introduction

In 2017, the budget of the consumption of medical care and goods in France represented 199.3 billion euro or 8.7% of GDP with 92.8 billion for hospital care with an increase of 32% compared to 2016.[1]

Each year in France, the hospitalization rate of inhabitants is as high as 19.1 % (12.7 million patients for more than 3300 health facilities) and a full hospitalization in the public sector lasts on average 6.4 days in medicine, surgery and obstetrics, 36.4 days in follow-up care and rehabilitation, 44.5 days in home hospitalization and 57.0 days in psychiatry.[2] This represents a major societal impact and a concern about the patient's well-being both mental and physical.

General practitioners (GPs) are responsible for the care of 90% of population's health problems. However, the multitude of information and the constant increase of patients encountered complicate their tasks. For this reason, it is valuable to provide tools that offer them a summary and some feedback about the patients' file, including all the essential points to order out the therapeutic actions to be chosen as a priority.

In addition, the lack of dedicated recommendations for the treatment of polypathological patients[3] complicates the management of these patients with the risk of applying recommendations for isolated diseases BOYD et al. [2005]; TINETTI et al. [2004].

Moreover, 1 general practitioner out of 5 sees more than 50 patients per day which is twice the number recommended under European safety guidance and the current average is more than 40 patients per day,[4] which confirms for the general practitioner the

---

[1] https://www.insee.fr/fr/statistiques/3676713?sommaire=3696937
[2] https://www.atih.sante.fr/analyse-de-l-activite-hospitaliere-2017
[3] https://www.has-sante.fr/portail/upload/docs/application/pdf/2015-04/note_methodologique_polypathologie_de_la_personne_agee.pdf
[4] https://collegeofmedicine.org.uk/complementary-medicine-roundup-march-2018/

interest of being able to target the elements to be taken care of as a priority since the duration of consultations is necessarily shortened.

Computerization of general practitioners has been intensive over the last twenty years DE ROSIS et SEGHIERI [2015]. The use of Electronic Medical Records (EMRs) and e-prescribing are priorities in different European action plans on e-health and are reflected in the policies of several states STROETMANN et al. [2011]. EMR development constitutes a formidable source of information and large data collection networks in primary care exist in our European neighbors, notably in the United Kingdom and the Netherlands (Clinical Practice Research Datalink CPRD Ex GPRD,[5] Q research,[6] Netherlands Information Network in General Practice[7]). These development efforts have enabled the creation of voluminous databases that support many types of research that resulted in advances covered in multiple publications. Furthermore, secondary use of electronic medical records offer many perspectives including an improved quality of care DE LUSIGNAN et VAN WEEL [2005] or to enable public health surveillance BIRKHEAD et al. [2015]; HERSH [2007]. The authors of HILLESTAD et al. [2005] explain that the adoption of EMR systems can significantly reduce the cost in the healthcare domain, moreover, EMR systems can be used as a support for disease management and refer more easily higher-risk patients to a specialist. Prevention and early detection can significantly benefit from the use of EMR data.

Electronic medical records contain essential information about the different symptomatic episodes a patient goes through. They possess the potential to improve patient well-being and constitute, therefore, a potentially valuable source to artificial intelligence approaches.

This CIFRE Thesis started in the context of a partnership between the team WIM-MICS[8] (INRIA/CNRS) and the company SynchroNext.[9]

The initial project was a use case with the Allianz company on routing questions from customer service. Since most of this service is overloaded with customer emails and are therefore an excellent opportunity for artificial intelligence approaches to automate the processing of these requests. This first experience has shown us the need for expert feedback and care in annotating the data.

With this in mind, we applied our theories to a new case in the medical field and

---

[5] https://www.cprd.com/home/
[6] http://www.qresearch.org/
[7] http://www.ulb.ac.be/esp/emd/nl_debakker.htm
[8] https://team.inria.fr/wimmics/
[9] https://www.synchronext.com/

started the project HealthPredict.[10] This project became the core of this thesis where we aim at preventing the hospitalization of patients or at least at improving their health's condition whether physical or mental by prioritizing the different risk factors responsible for the hospitalization. The results of this research are intended to provide decision support tools for general practitioners to assist them in their daily practice. For that purpose, we trained a supervised machine learning to identify hospitalized patients from the others and to learn the risk factors involved in that decision. One of the purposes is to order the risk factors to be treated as a priority since it is complex to identify the best treatment plan and what is possible for some patients, as well as to take into account polypathology and adherence to treatment. And indeed, dealing with all the pathologies of a patient at the same time means dealing with drug interactions.

Our study focuses on general practice while most scientific papers rely on hospital data. However, it is more common for patients to go to visit their general practitioners rather than to the hospital. This situation is principally due to the lack of feedback on general practitioners' practices because of the non-existence of federated services using international standards for data collection from independent physicians.

Nowadays, it is fairly common to hear about great advances in the detection of a really specific pathology by applying machine learning technologies on images or recorded signal from electrical activity of the body. Our study, however, aims to propose a way to build a preventive decision aid tool based on Electronic Medical Record as a non-invasive procedure without specific prerequisites. The ideal would, of course, be the follow-up of all the medical trajectories of a patient (family physician, specialist physician, hospital...), but as a first step we study here the predictability of that decision to move a patient from general practitioners to hospitals.

Thus, we will focus on predicting the hospitalization of patients based on their medical records as well as those of other patients who have been used to train supervised classification algorithms.

This dissertation is composed of the following chapters:

- The second chapter introduces the PRIMEGE database and relevant works on representing textual and time-series data. Then, we propose two different vector representations with the sequential and non sequential modelling of EMRs.

---

[10]https://www.health-predict.com/

- The third chapter presents our study dataset, the supervised classification algorithms that we will feed with the vector representations from the first chapter and their evaluation in order to predict patient hospitalization.

- The fourth chapter studies relevant work on domain knowledge enrichment and entity linking, then explores the extraction of knowledge from general knowledge sources and domain specific ontologies and how we integrate them to vector representation of EMRs.

- The fifth chapter evaluates enrichment of vector representations of EMRs with domain knowledge.

- The sixth chapter shows the use case planned for our application and its interface with its evolution, then it displays the specificities applied on the predictive algorithm to use it in our application.

- The last chapter will review our contributions and highlight the perspectives of our work.

Chapter 2

Basic representation of raw data from electronic medical records

This chapter is dedicated to the foundations of knowledge representation and the characteristics of the electronic medical records (EMRs) relevant to the following chapters. We will first introduce the database of general practitioners consultations from which our dataset is derived in Section 2.1. Subsequently, we will present relevant works on vector-based representation of data on top of which we built our contributions in Section 2.2. Finally, we will propose in Section 2.4 and discuss 2.5 two vector representations of electronic medical records.

2.1 Characteristics of PRIMEGE: a database of general practitioner consultations

The national French Health Data System (SNDS) integrates data from several information systems[1] (see Figure 2.1):

- the national health insurance database, SNIIRAM[2] (Système National d'Information Inter-Régimes de l'Assurance Maladie),

- the national hospital discharge database, PMSI[3] (Programme de Médicalisation des Systèmes d'Information),

---

[1] Documentation referencing the data collected in the SNDS: https://www.snds.gouv.fr/download/ SNDS_Nomenclature_sous_produits.pdf.

[2] https://www.ameli.fr/l-assurance-maladie/statistiques-et-publications/sniiram/ finalites-du-sniiram.php

[3] https://www.epmsi.atih.sante.fr/
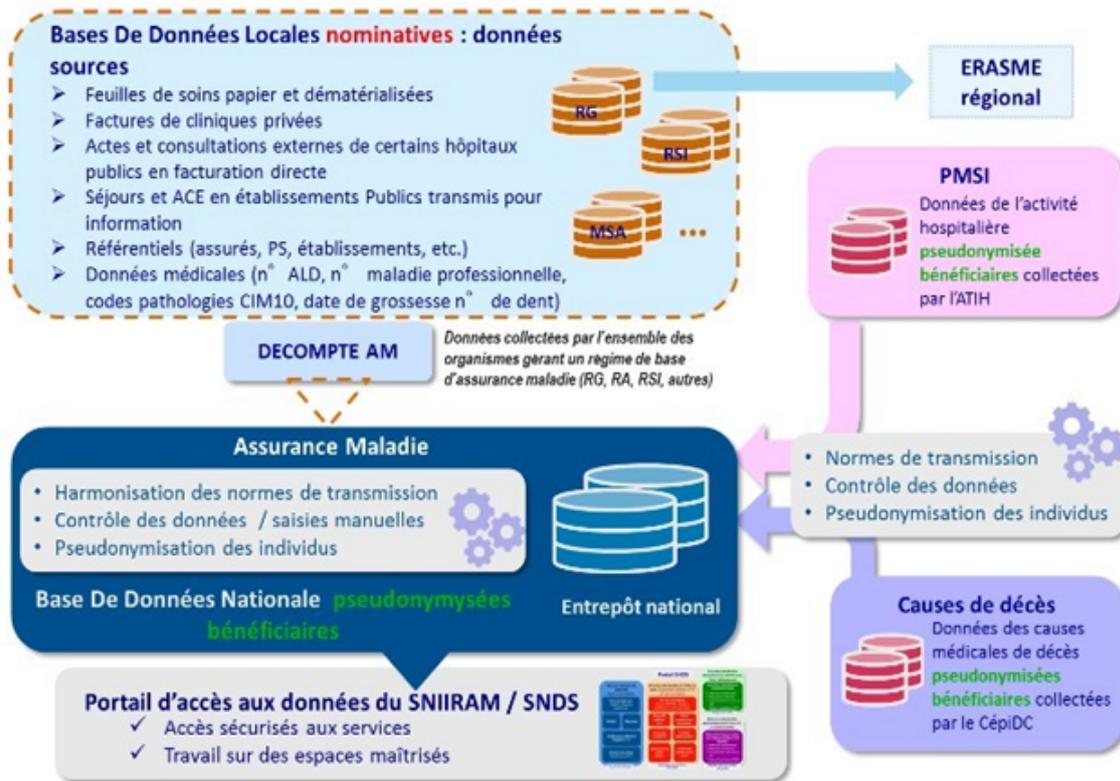
Figure 2.1 – Diagram illustrating the methods of collecting, processing and distributing health data with the SNDS. At the top right, there is data from the PMSI. At the bottom right, we find data from CépiDC. At the left, there is data from the SNIIRAM. Source: https://bit.ly/36R1FhY.

- the database on the medical causes of death, BCMD[4] (Base de données sur les Causes Médicales de Décès), managed by CépiDC,[5]

- the database from CNSA[6] (Caisse Nationale de Solidarité pour l'Autonomie) which integrates data related to disability from MDPH[7] (Maisons Départementales des Personnes Handicapées),

- data from OCAM (Organisme d'Assurance Maladie Complémentaire), OCAM includes complementary health insurance organizations like mutual insurance companies, pension funds and insurance companies.

Thus the SNDS essentially contains data related to hospital databases as well as the medicines dispensed, but it only gives a limited picture of general practice. The PRIMEGE database is the result of a preliminary study conducted to identify the data provided by

---

[4]https://epidemiologie-france.aviesan.fr/en/epidemiologie-france/fiches/
base-de-donnees-sur-les-causes-medicales-de-deces-en-france
[5]https://www.cepidc.inserm.fr/
[6]https://www.cnsa.fr/
[7]https://www.cnsa.fr/outils-methodes-et-territoires/mdph-et-departements

an EMR system in order to build a health-related data warehouse going beyond drug prescriptions. A list of data to collect was established from the recommendations of the ANAES (Agence Nationale d'Accréditation et d'Evaluation en Santé) BIRKHEAD et al. [2015], concerning medical records in general medicine in addition to a study led by the departments of General Medicine at the University of Nice Sophia-Antipolis and Lyon 1.[8] These data may concern episodes of care, but also include data valuable to evaluate and improve the practices such as the ICPC-2 code (International Classification of Primary Care) enabling standardization of the reasons for consultation and diagnoses. Data allowing patient identification (such as name, surname, address, etc.) were excluded.

Therefore, PRIMEGE differs from SNDS in that it contains, in addition to the prescribed drugs and the reasons for prescriptions, a great deal of additional information about patients. As shown in Table 2.3 and Table 2.4, we can find in PRIMEGE, the text descriptions written by general practitioners together with international classification codes of prescribed drugs, pathologies and reasons for consultations, as well as the numerical values of the different medical examination results obtained by a patient. PRIMEGE, also captures the history of the patient and his family. It has the merit of being exhaustive in its design, and the use of international codes provides an evident advantage in terms of risk monitoring and, in our case, for the application of artificial intelligence processes.

PRIMEGE contains the drugs prescribed to patients by their general practitioners. However, these drugs may not have been withdrawn by the patient, just as he may has withdrawn others from his pharmacist. In this sense, the SNDS and PRIMEGE are complementary.

This PRIMEGE dataset is also representative of the population met by general practitioners since no selection is made, unlike clinical studies with control and experimental groups, which are restricted to a narrow segment of the population. However, this allegation should be put into perspective in the sense that this database contains a significant sample of the population met by physicians in the Provence-Alpes-Côte d'Azur region where there is an overall ageing of the population.[9,10]

The PRIMEGE database LACROIX-HUGUES et al. [2017] contains more than 600,000 consultations carried out by 17 general practitioners across the Alpes-Maritimes depart-

---

[8]https://www.atih.sante.fr/analyse-de-l-activite-hospitaliere-2017
[9]https://connaissance-territoire.maregionsud.fr/fileadmin/user_upload/Annuaire/Etude/vieux2017.pdf
[10]https://www.insee.fr/fr/statistiques/2869942

ment in France. Table 2.1 describes the data collected in this database, Table 2.2 displays it statistics and Table A.1 its relational diagram.

Table 2.1 – Data collected in the PRIMEGE database.

Category	Data collected
GPs	Sex, birth year, city, postcode
Patients	Sex, birth year, city, postcode
	Socio-professional category, occupation
	Number of children, family status
	Long term condition -LTC- (Y/N)
	Personal history
	Family history
	Risk factors
	Allergies
Consultations	Date
	Reasons of consultation
	Symptoms related by the patient and medical observation
	Further investigations
	Diagnoses
	Drugs prescribed (dose, number of boxes, reasons of the prescription)
	Paramedical prescriptions (biology/imaging)
	Medical procedures

Table 2.2 – Data volume contained in the PRIMEGE database.

Element	Amount
Patients	68,415
Consultations	601,464
Past medical history	212,797
Biometric data	384,087
Reasons of consultation	345,626
Diagnoses	125,864
Prescribed drugs	1,089,470
Symptoms	33,273
Health care procedures	15,001
Additional examination	1,281,300
Paramedical prescription	25,910
Observations/notes	73,336

One of the observations that can be made on PRIMEGE is that the amount of information provided by physicians varies considerably from one physician to another, as well as from one patient to another, since physicians can fill out their consultation forms as they wish. Even so, they are encouraged to fill it in as accurately as possible, and offering them applications related to their data would be an additional motivation for that pur-

Table 2.3 – Example of consultation found in PRIMEGE. The translation and corrections of the texts are for 'Reason of consultation': "tetanus vaccination", for 'History': "Appendicitis" and for 'Observation': "In [First Grade], good general condition, clear pul[monary] auscul[tation], reg[ular] [heart sounds] without breathing, eardrums OK".

Birth date	...	Gender	LTC	Problem date	Number of visits	Reason	ICPC2	History	Medical procedure	Observation
2005	...	H	N	S17-2012	10	vaccin anti-tétanique	A44	Appendicite	VACCIN REVAXIS SER 0,5ML+2 AIG 1	EN CP - Bon état général -; auscult pulm libre; bdc rég sans souffle - tympans ok-

Table 2.4 – Example of consultation found in PRIMEGE. The translation and corrections of the texts are for 'Reason of consultation': "Results of specialized tests and examinations, depression", for 'History': "Type IIb Dislipidemia, non-toxic multinodular Goitre, Hypertension, Arthrosis of the knee, hemoc[c]ult" and for 'Active problem': "Non-toxic multinodular Goitre, Type IIb Dislipidemia, Hypertension, Arthrosis of the knee".

Birth date	...	Gender	LTC	Problem date	Number of visits	Reason	ICPC2	History	Active problem
1947	...	F	N	S30-2015	38	Résultats d'analyses et d'examens spécialisés, Dépression	S60, P76	Dyslipidémie type IIb, Goitre multinodulaire non toxique, Hypertension artérielle, Gonarthrose, hemocult	Goitre multinodulaire non toxique, Dyslipidémie type IIb, Hypertension artérielle, Gonarthrose

pose. Another point that can be noted is the apparent confusion between symptoms and diagnoses in the data provided by physicians.

One of the questions that arises when faced with the problem of predicting a medical event (e.g., prediction of hospitalization or prediction of a pathology) from a database such as PRIMEGE is how to represent medical records to apply machine learning algorithms. This question arises particularly with the number of text fields involved in a consultation, and with on one side permanent data and on the other side time-stamped data about the patient. Text modelling on such tasks is crucial since 79.3% of the phenotypes present in medical records can be identified within free texts ESCUDIÉ et al. [2017]. Of course, a medical record does not exclusively contain textual information, and it is also essential to address this specific point.

2.2 Relevant state of the art vector representations of textual data

Before going further into the representation we have chosen for EMR, we present a brief overview of state of the art of vector representations relevant to textual data.

The bag-of-words model HARRIS [1954] consists in representing a text as a list of tokens. Thus, words of a sentence are employed as features of a semantic distributional model (see Figure 2.2). Such model is also adopted outside of natural language process-

ing with image processing but this method is called, in this case, a bag-of-features. Either the occurrence or frequency based metrics are used in this model.

(a) Sample documents

Doc. 1	John likes football and eating apples.
Doc. 2	Mary doesn't like football, she also likes eating out.
Doc. 3	Mary likes eating out with John but not with Peter.
Doc. 4	Peter doesn't like eating out.

(b) Vocabulary, using all the distincts words in the document collection

Word 1	John	Word 9	like
Word 2	likes	Word 10	she
Word 3	football	Word 11	also
Word 4	and	Word 12	out
Word 5	eating	Word 13	with
Word 6	apples	Word 14	but
Word 7	Mary	Word 15	not
Word 8	doesn't	Word 16	Peter

(c) Word-document occurence matrix

$$\begin{pmatrix} 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}$$

Figure 2.2 – Example of generated vector representation by BOW from textual documents. Source: FONCUBIERTA RODRIGUEZ [2014].

Brown clustering BROWN et al. [1992] consists in generating clusters of words based on word classes through hierarchical clustering based on the context, the representation in the form of a tree (see Figure 2.3) provided by this method can then be used to perform different tasks in the scope of natural language processing.

Word embedding consists in capturing the relationships between words by generating a textual representation through an encoder (see Figure 2.4). An encoder is a deep learning algorithm that encodes input data and thereby performs dimensionality reduction by compressing the data. There are several notable models enabling to produce word embeddings, among which Word2Vec MIKOLOV et al. [2015], Glove PENNINGTON et al. [2014] and more recently Fastext BOJANOWSKI et al. [2016] (and the latest innovation brought to this project with MOE, Misspelling Oblivious word Embeddings PIKTUS et al. [2019]) There exists other variations of this model such as Doc2Vec LE et MIKOLOV [2014], which preserve the principle of dimensionality reduction with an encoder, but applied to documents.

The state of the art for textual representations has recently considerably changed and
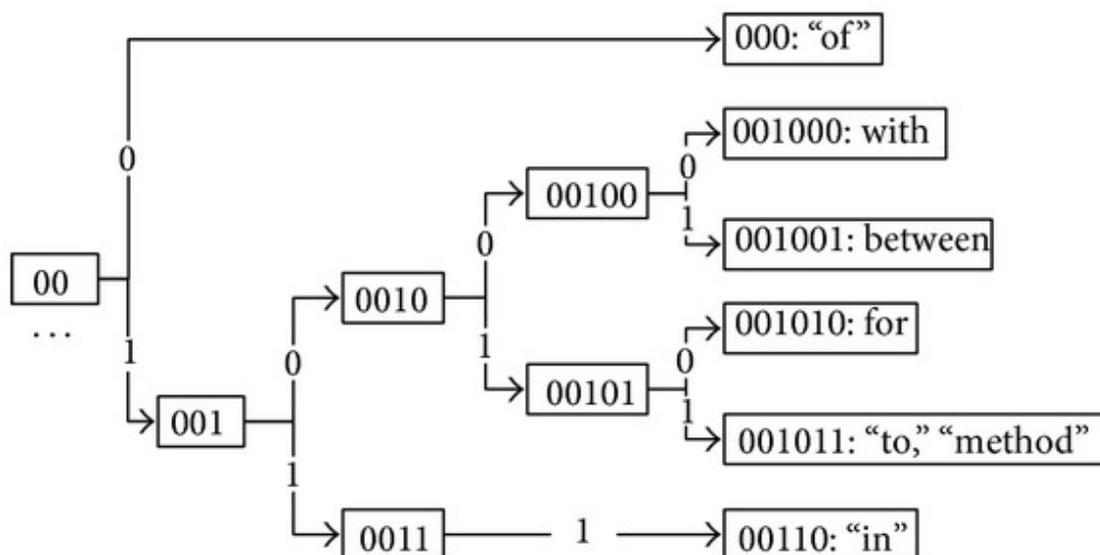
Figure 2.3 – Example of tree generated by Brown clustering for 7 words from the JNLPBA corpus. Source: TANG et al. [2014].

just like word embeddings, input data is compressed to generate a representation with the BERT model DEVLIN et al. [2018] that relies on a transformer. A transformer uses encoder-decoder attention layers to generate encodings that better capture the semantic relations of sequences' embeddings. This model learns to predict the words present in a sentence by using a mechanism called masked input tokens, some words are randomly not considered in the inputs of this algorithm, which improves the robustness of this algorithm.

To better address the specificities of biomedical source texts, BioBERT LEE et al. [2020] uses a pretrained BERT and retrains it on a biomedical corpora, so that it outperforms BERT and state of the art models on several biomedical natural language processing tasks such as named entity recognition, relation extraction and question answering.

MultiFiT EISENSCHLOS et al. [2019] which is based on ULMFiT HOWARD et RUDER [2018] outperforms other state of the art models for multilingual resources including MultiBERT,[11] a multilingual model for BERT, and requires lesser data for training. MultiFit uses for this purpose a quasi-recurrent neural networks, which combines the advantages of convolutional neural networks and recurrent neural networks. It uses as input subwords instead of unigrams (cf. Multiple Tile Encoding compression), and the labels of its instances are floating numbers. Assigning a floating number as a label is called label smoothing and this prevents a machine learning algorithm from becoming over-

---

[11]https://github.com/google-research/bert/blob/master/multilingual.md
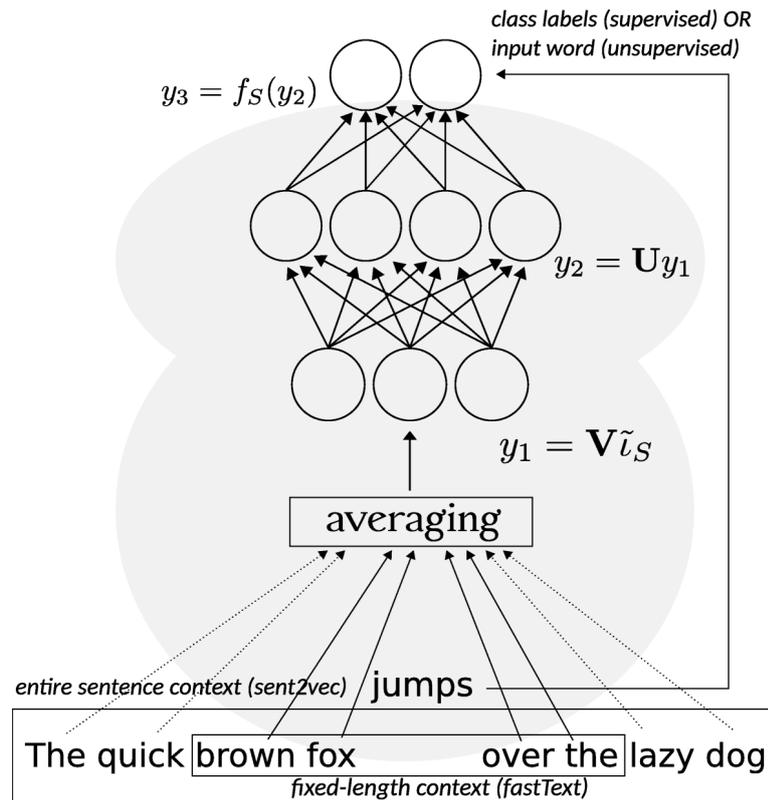
Figure 2.4 – Embedding vector learned through a shallow neural network. Source: AGIBETOV et al. [2018].

confident.

Actually, one of the negative aspects of more advanced models is that the data compression inherent in such models causes loss of information and therefore of some interpretability, although they are able to better capture the semantic relationships between terms. Conversely, the BOW representation is easily interpretable but does not capture the semantic relationships between words and generates huge models. The representations generated with a BOW are in the form of a sparse matrix, a matrix filled essentially with zeros, which makes it possible to accelerate the calculations performed in comparison to dense matrices. Each textual representation has pros and cons depending on the targeted use case.

Most of the time, preprocessing is required before feeding a textual representation, such as, at least, the steps of lemmatization or stemmatization that consist in dealing with inflected forms of words in order to obtain either lemma, the canonical form of a word, or stem, the main part of a word delivered by an algorithm. This process allows in a textual representation to reduce the textual variability that can be identified within a corpus. This can also be considered as data compression, although it is carried out

on a smaller scale and it is less opaque than using a deep learning algorithm or the PCA (Principal Component Analysis) method PEARSON [1901]. Other prepossessing steps can be considered such as spell-checking, removal of special characters or handling of named entities.

Without preprocessing, noise would be introduced with the generation of a textual representation. In particular without stemming or lemmatization, inflected forms of the same word will be present in a textual representation instead of being represented only once. The result is a distinction between inflected forms, whereas they have only one meaning.

It is crucial when using a domain-specific corpus to generate its own representation, since many terms may be omitted in a general representation or an ambiguous notion may be applied to a term when it admits a very precise definition in a given sector. We opted among existing textual representations for a model using a bag-of-words representation (BOW) HARRIS [1954] for different reasons: (i) the main information from textual documents is extracted without requiring a large corpus; (ii) the attributes are not transformed, which makes it possible to identify which terms contribute to a prediction, even if this implies to manipulate very large vector spaces; (iii) the integration of heterogeneous data is facilitated since it is sufficient to concatenate other attributes to this model without removing the meaning of the terms formerly represented in this way. As a result, the relatively simplistic BOW model allows us to confirm our theories as well as to analyze what a classifier has indeed learned.

2.3 Relevant work on the representation of time-series data

Time-series data are a list of values or events that occur over time. To represent these values, it is therefore necessary to model the time factor, which implies discretizing time into time windows. EMRs contain patient data that can cover several years and thus are suitable for a representation that takes into account time-series data.

Original early detection systems based on EMRs applied the aggregation of data from time windows and ignored the relations between events.

The improvements brought by different sequential approaches have been studied by SINGH et al. [2015] on the task of predicting deterioration of kidney function. They show

that a Stacked Temporal approach outperforms a Non-Temporal one but that this model is subject to overfitting. They also propose a competitive approach with a Multitask-Temporal model, though it implies to consider time-windows extracted from a same patient as independent. Another limitation of the approaches they propose is that textual information such as diagnoses, procedures and medications are represented as a binary variable. Thus, a variable is set to 1 if the textual information involved is encountered in any of the time-windows considered.

LIU et al. [2018b] propose to use event embeddings, considering time as a factor to handle lab tests, routine vital signals, diagnoses and drug administrations on the tasks of predicting death and abnormal lab tests. They applied a new algorithm based on the modification of a LSTM called Heterogeneous Event LSTM (HE-LSTM) that performs better than other LSTM approaches to these embeddings. However, textual information contained in their study is limited to the types of events.

HENRIKSSON et al. [2015] propose an extension to distributional semantics models to deal with heterogeneous data contained in EMRs on the task of detecting adverse drug events with the Word2Vec algorithm MIKOLOV et al. [2015]. They evaluated the performances of the representation with the random forest algorithm BREIMAN [2001] for adverse drug event detection and show that combining structured and unstructured data modeled in semantic space significantly improved predictive performances. However, the proposed representation loses in interpretability if we seek to provide the reasons behind the classifier's choices. Moreover, the issue of considering permanent information relative to the patient is not addressed.

Despite the constraints outlined by SINGH et al. [2015], we opted for an Stacked Temporal approach since it appears to be the most usual way to represent EMRs with a sequential representation. In addition, it does not require any modification of sequential machine learning algorithms. The Multitask-Temporal approach that they described is a special representation that works with non sequential machine learning algorithms. We excluded the other representations, although they take into account heterogeneous data, because they are not easily interpretable without drastic modification.

2.4   Vector representation of electronic medical records

In this section we propose two vector representations of EMRs, one non sequential and the other sequential. We first present the text preprocessing we perform for both repre-

sentations.

2.4.1 Preprocessing of texts in electronic medical records

We preprocessed the textual data in EMRs with a regular expressions tokenizer built with the NLTK library[12] LOPER et BIRD [2002] and with the TreeTagger[13] lemmatizer MÀRQUEZ et RODRÍGUEZ [1998] (we observed better results with this tool than with the existing French stemmers in NLTK). The number of occurrences of tokens was used as values of attributes in the different upcoming vector representations.

All text fields in the EMRs (see Table 2.1) are transformed into vectors. Just like in the structure of the PRIMEGE database, some textual data must be distinguishable from each other when switching to the vector representation of EMRs, e.g., a patient's personal history and his or her family history. To achieve this, we have introduced provenance prefixes during the creation of the bag-of-words to trace the contribution of the different fields.

We will provide more details on the text fields used, and which were prefixed in the creation of our vectors in Chapter 3.

2.4.2 Non sequential modelling of electronic medical records

Our base non sequential representation of EMR is as follows. Let $V^i = \{w_1^i, w_2^i, ..., w_n^i\}$ be the bag-of-words obtained from the textual data in the EMR of the ith patient.

When considering the task of predicting hospitalization, we aggregated all the consultations occurring before a hospitalization. For patients who have not been hospitalized, all their consultations are aggregated. We are in the presence of two classes, thus the labels y_i associated with V^i used for this representation are either 'hospitalized' or 'not hospitalized'.

This requires specific processing to take into account time-series values in non sequential representation. For instance, medical test for cholesterol levels may have been done on several consultations, and to consider it, it is necessary to discretize the time factor in one way or another. So, we can use the last known biomedical analysis or use a measure like the maximum, the mean, etc. of all the analysis related to cholesterol levels. This discretization allows to integrate this biomarker in a non sequential representation.

---

[12]http://www.nltk.org/
[13]https://cis.uni-muenchen.de/~schmid/tools/TreeTagger/

The principal characteristic that differentiates this representation from the sequential representation is that the information on consultations and permanent information are mixed up and not repeated.

2.4.3   Sequential modelling of electronic medical records

For a sequential modelling of EMRs, we chose to represent the different consultations of a patient as a sequence $(t_1, ..., t_n)$. This n-tuple contains all his consultations in chronological order, with t_1 his first consultation and t_n, his last consultation present in the database. Each consultation t_i contains both persistent patient data and data specific to the ith consultation.

Similarly to the non sequential representation of EMRs, for patients who have not been hospitalized, all their consultations are integrated in the sequential representation of EMRs whereas for patients who have been hospitalized only their consultations occurring before hospitalization are integrated.

Contrary to the vast majority of state of the art works that drop medical analysis to focus on textual information LIN et al. [2012] JIN et al. [2018], we propose an approach to include them with a way to handle permanent data such as the patient's history. Thus every $t_i = (x_i, y_i)$ where x_i contains two broad types of information about the patient. On one hand, it contains consultation notes on the reasons for the consultation, diagnoses, prescribed drugs, observations and the numerical data resulting from the medical tests. On the other hand, it contains textual information conveyed throughout the patient's life including, for instance, familial history, personal history, personal information, past problems, the environmental factors as well as allergies (see Figure 2.5). It means that textual information carried throughout the patient's life is repeated across all x_i of t_i. We also consider at the same level as permanent information the current health problems (like osteoporosis) and the past health problems (like warts) since they do not vary considerably over the years.

EMRs can contain multiple elements that occur at the same time (e.g., multiple reasons of consultation for one consultation) and different types of content for the same consultation. Conventional time-series data are not composed of heterogeneous data and use a succession of individual elements, e.g., succession of tokens in a text or succession of states traversed by an automaton. In EMRs, multiple reasons for consultations may occur during the same consultation, i.e., a patient may have fever and vomiting at the
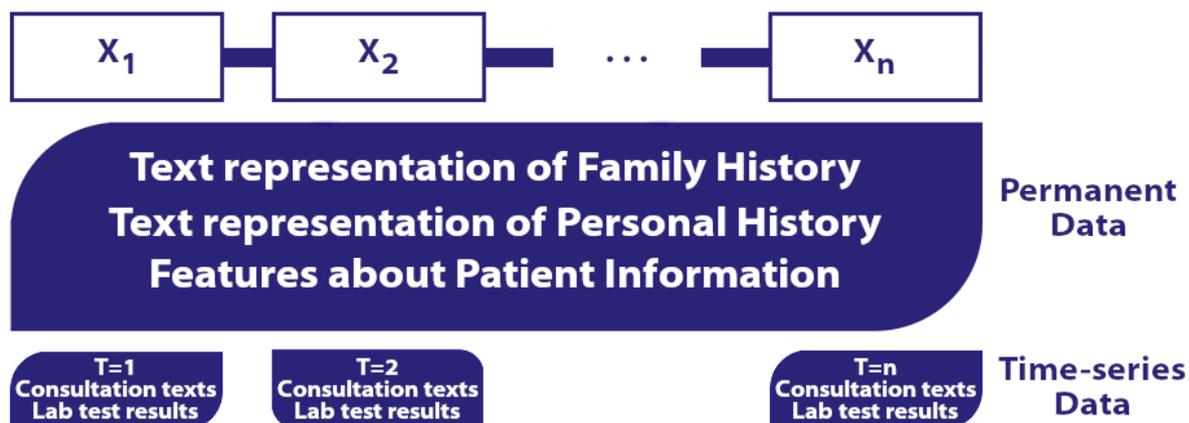
Figure 2.5 – Diagram illustrating the sequential representation of an electronic medical record.

same time, both conditions occur at the same time. The nature of these data prompted us to use the modelling described above.

In order to learn a prediction model, y_i contains the label to predict, in our case 'hospitalized' or 'not hospitalized'. In the experiments of Chapter 3, we attribute the same label to all the t_i of a patient indicating whether or not he was hospitalized.

2.5   Discussion

A major disadvantage of the chosen vector representation of the EMRs for sequential machine algorithm is that it is of considerable size. This is particularly true with regard to n-grams, which represent windows of words sequence. A textual representation that uses dense matrices also does not ensure better performance, in terms of disk space and time, since even if it is thinner, the fact that the matrices used are not sparse requires more resources in order to be able to compute. This is probably one of the reasons why most research focuses on international codes such as UMLS, rather than using the text in medical records, texts become a way to extract UMLS codes.

Despite this aspect, the representations that we will use to feed machine learning algorithms allow us to track the origin of an attribute from an EMR, since the attributes are not transformed with such model. Moreover, they will provide an effective feedback on what a machine learning algorithm learns from the inputted data.

2.6 Conclusion

In this chapter, we presented and compared our proposed sequential and non-sequential vector representations of EMRs.

In order to model the EMRs sequentially, we used a consultation as a time unit. In future work, we could investigate different time windows in order to better capture the notion of temporality present in EMRs.

Another point that is related to the modelling is the value to be assigned to the attributes. Indeed, we use in this work the number of occurrences to model attributes, however it is a questionable choice since the amount and the degree of completion of consultations varies from one patient to another. Future work on this matter could involve the normalization of attributes between patients.

The choice of another textual representation that takes into account the semantic relationships between words may also be discussed in the future, but this point should be considered with caution, as physicians must obtain the most accurate interpretation possible to assist them in their profession.

In the following chapter we evaluate the proposed representations by utilizing them with state of the art machine learning algorithms to predict patient's hospitalization.

Chapter 3

Predicting hospitalization on a basic representation of electronic medical records

In this chapter we consider the task of predicting patient hospitalization from electronic medical records. We review state-of-the-art supervised machine learning algorithms relevant for this task and we report the results of our first experiments aiming to predict patient hospitalization by applying these algorithms to the basic vector representations of electronic medical records introduced in Chapter 2.

We first present in Section 3.1 how we prepare our dataset to perform hospitalization predictions, then we introduce in Section 3.2 the relevant state of the art machine learning algorithms, whether sequential or non sequential. In Section 3.3 we describe our experimental protocol and present our results, and discuss them in Section 3.4. Finally, we will conclude and discuss our perspectives for future work in Section 3.5.

3.1 Prediction of hospitalisation from electronic medical records

The prediction of hospitalization can be formulated as a classification problem, which aims to separate hospitalized and non hospitalized patients.

To learn a prediction model for hospitalization from Electronic medical records (EMRs), we need to consider a training set that discriminates EMRs of hospitalized pa-

tients and EMRs of non hospitalized patients. Patients who bear an explicit label indicating a hospitalization in their records, within the fields related to the reasons for consultation and diagnoses, are the positive cases. The other patients are candidate negative cases; they were validated by a physician.

Generally, the EMRs of hospitalized patients contain an indication notifying the return of a patient from the hospital.

Consequently, in order to automatically identify the event of hospitalization (or return from hospitalization) from patient's EMRs, we used the following regular expression over reasons for consultation and diagnoses:

```
1   ((hospital[A-zÀ-ÿ]+)|(h[õo]pital)|(hospi)|(crh))
```

Where 'crh' is the French abbreviation for hospital report.

This regular expression allowed us to consider for hospitalized patients, only consultations that occurred before their hospitalization or before the decision to hospitalize them. Figure 3.1 displays the consultations thus considered for our vector representations.

Figure 3.1 – Diagram representing the consultations considered *before* (events from t_0 to t_{n-1}) the detection of a hospitalization (event t_n).

To construct V^i for our non sequential representation, as well as for x_i for our sequential representation, we consider the following EMR fields:

- sex

- birth year

- long term condition

- risk factors

- allergies

- reasons of consultation with their associated codes

- medical observations

- diagnosis with their associated codes

- care procedures

- the drugs prescribed with their associated codes

- current health problems

- reasons of the prescription

In addition to the previous fields we added a number of fields for which we prefix the terms and concepts in order to capture the fact they apply to different aspects e.g. feature of a patient vs feature of the family of the patient. These prefixed fields are:

- patient's history (prefix: '#history#')

- family history (prefix: '#family#')

- past problems (prefix: '#past_problem#')

- symptoms (prefix: '#symptom#')

- diagnosis of the patient with their associated codes (prefix: '#diagnosis#')

The inclusion of a prefix to these fields allows us to distinguish them from other textual data related to the patient's own record in the vector representation of EMRs.

To learn a prediction model, we extracted a balanced training set, DS_B, from the PRIMEGE database presented in Chapter 2. Classification algorithms are sensitive to unbalanced datasets, and in such a case, they will mostly assign to new instances the majority class label. Different methods exist to overcome this problem such as penalizing the cost function used, increasing or decreasing one of the two classes studied in a binary classification task. We use a balanced dataset in order to avoid going through this process.

This dataset is composed of 714 hospitalized patients and 732 patients who were not hospitalized over a 4-year period. These patients who were not hospitalized, were randomly selected from the PRIMEGE database. Random selection of not hospitalized patients was used to avoid human selection bias and thus, we obtained a representative sample of the population met by the general practitioner.

Finally our prediction task can be defined as: Let R be a representation of an EMR from the PRIMEGE Database P. Let L be the set of classes to predict L = {$Hospitalized, NotHospitalized$}. We try to learn the mapping M: M(R) = L. Where M is a classification algorithm that predicts a class L for a given EMR R.

In the next section we review state of the art supervised non-sequential and sequential classification algorithms that can be used to predict the hospitalization of patients.

3.2 Relevant state of the art machine learning algorithms

As we were able to develop a labeled dataset for our task of predicting patient hospitalization, we will rely on supervised machine learning algorithms to separate patients who require hospitalization from those who do not. Supervised learning consists in inferring a function from labeled training data to be able to classify new instances with the same labels learned from training examples.

In addition to the aspects related to machine learning, human factors must be taken into consideration. Interpretability is a crucial aspect for a physician since it is not enough to predict that a patient will be hospitalized, he must be provided with the factors involved in this prediction. In order to comply with this condition, we selected state of the art machine learning algorithms accordingly.

In addition, the limited size of our dataset excluded from the scope of our study neural networks algorithms since they require large amounts of data, unless pre-trained representations are used. However, adapting them to a specific domain and a completely different task can be complex or hardly feasible. The interpretability of such models is also questionable, although progress has been achieved in this direction with self-attention mechanisms and linear models on local approximations RIBEIRO et al. [2016].

3.2.1 Non sequential machine learning algorithms

For non sequential classification algorithms, we focus on three different machine learning algorithms which are frequently used in the literature. They will serve to evaluate the impact of sequential machine learning algorithm and modelling on the prediction of patients' hospitalization based on their EMRs.

Among the algorithms that give feedback on what they have learned from a classification problem, there are the logistic regression (LR) MCCULLAGH et NELDER [1989],

random forests (RF) BREIMAN [2001], and support vector machine (SVM) CHANG et LIN [2011] with a linear kernel. SVMs are often used as a basis for comparison in natural language processing tasks. Moreover, logistic regression and random forest algorithms are widely used in order to predict risk factors in EMRs GOLDSTEIN et al. [2017].

The principle behind the random forests algorithm is to learn in parallel, different decision trees trained with different subsets of features, which allow to avoid the limitations encountered with bagging trees that use all features. Another step of the random forest algorithm involves to aggregate by an ensemble learning method the learned decision trees. This implies that random forests do not use the same set of features at each new training phase if the number of features is large enough, which means that it is difficult to reproduce the same forest from one learning phase to another.

With a linear kernel, a support vector machine strives to determine the optimal hyperplane, margin, for the simplest case between two classes to differentiate them. This notion of optimal is perceived differently between logistic regression and support vector machine, since logistic regression mostly considers all points in the search for the hyperplane while support vector machine algorithm focuses mainly on points close to decision boundary. Of course, given the way these two algorithms operate, this implies the ability to linearly separate classes. More complex cases can be solved with a polynomial kernel for support vector machine, but this is at the expense of interpretability.

Initially, the logistic regression algorithm is used to perform binary classification, but some extensions allow it to be used with more classes. This algorithms assumes that there is a linear relationship between dependent and independent variables.

The logistic regression algorithm is defined as follows:

$$\mathrm{L}n(\frac{\mathrm{P}}{1-\mathrm{P}}) = \beta_0 + \beta_1 \mathrm{X}_1 + ... + \beta_n \mathrm{X}_n \tag{3.1}$$

Where P is the probability to hospitalize a patient, X_n the variables to model, i.e., the attributes used in our BOW, and β_n the coefficients used by the regression. The coefficients β_n, of each variables X_n are delivered after the training of the algorithm.

We highlight the logistic regression algorithm as it is the one that has obtained the best performance in predicting patient hospitalization, especially with the injection of knowledge (see Chapters 4 and 5).

3.2.2   Sequential machine learning algorithms

We will present in this section several Markovian models and what guided our choice towards one of them in particular. Markovian models share with the algorithms in the previous section the particularity of being interpretable since it is possible to obtain the weights of the state and transition characteristics.

Hidden Markov models (HMMs) use observations, labels about a classification problem, and allow to assume the sequence of states involved in order to generate such observations. Contrary to Markov chains, for hidden Markov models, the states involved to generate the observations are unknown.

Maximum entropy models (MEMMs) are discriminative models and thus consider a conditional probability distribution rather than a joint distribution. In a similar way to HMMs, the state of the current position in these models depends only on the state of the previous position. But they differ from HMMs in that they offer richer representations, where features of interest for natural language processing applications (like prefixes or suffixes) can be introduced.

CRFs Sutton et al. [2012], like MEMMs, are discriminative models. Contrary to hidden Markov models they do not require the assumption that the observations are independent especially because in many real-word cases we cannot assume that variables are not related to each other.

Among the existing sequential machine learning algorithms, we chose CRFs because HMMs are generative models, and MEMMs, which are discriminative models, have label bias issues: they proceed to a normalization at each state of the sequence whereas CRFs normalize the whole sequence.

Given s a sequence of consultations, i, the position of a consultation in a sequence of consultations and l_i, the current consultation, a linear-chain CRFs computes a probability:

$$P(l|s) = \frac{exp[\sum_{j=1}^{m} \sum_{i=1}^{n} f_j(s, i, l_i, l_{i-1})]}{\sum_{l'} exp[\sum_{j=1}^{m} \sum_{i=1}^{n} f_j(s, i, l'_i, l'_{i-1})]} \tag{3.2}$$

Where $f_j(s, i, l_i, l_{i-1})$ is a feature function that estimates the likelihood that the current consultation l_i belongs to a hospitalized patient given his previous consultation l_{i-1}.

CRFs can be compared to the logistic regression algorithm, as logistic regression is a log-linear model dedicated to classification and CRFs is a log-linear model dedicated to

sequential labels.

To assess the performance of the conditional random fields algorithm (CRFs) we relied
on the sklearn-crfsuite library,[1] a CRFsuite wrapper.

3.3 Experimental study

3.3.1 Experimental setting

We evaluated our representations following the K-Fold method, a cross-validation strat-
egy which allows us to test a classification algorithm across all the considered data. We
chose K = 10, which allows us to separate the data in DS_B into 10 partitions with approxi-
mately 70 patients for each of the classes "hospitalized" and "not hospitalized". It should
be noted that the BOW from one fold to another may differ since the data used during
training vary from one fold to another, and therefore a term may not be present in an-
other training partition.

The different experiments were conducted on a Precision Tower 5810, 3.7GHz, 64GB
RAM with a virtual environment under Python 3.5.4.

3.3.2 Search space for hyperparameters of machine learning algo-
rithms

We optimized the hyperparameters of the machine learning algorithms used in this
study with nested-cross validation CAWLEY et TALBOT [2010] (see Figure 3.2) in order to
avoid bias, and the exploration was done with random search BERGSTRA et BENGIO [2012].
The inner loop was executed with a L fixed at 2 over 7 iterations, which corresponds to 14
fits by machine learning algorithms. Optimizing with nested-cross validation allows to
optimize machine learning models by generalizing the search of best hyperparameters
on the training set instead of doing it on the test set. Different sets of hyperparameters
are thus tested to select the one with the best performance on the inner loop.

Other techniques exist to explore the research space of the hyperparameters like grid
search but this method is rather expensive in computation time and power efficiency. Op-
timization with bayesian searches SNOEK et al. [2012] is in the spotlight with the optimiza-
tion of neural networks and very often proves to obtain more convincing results than grid

---
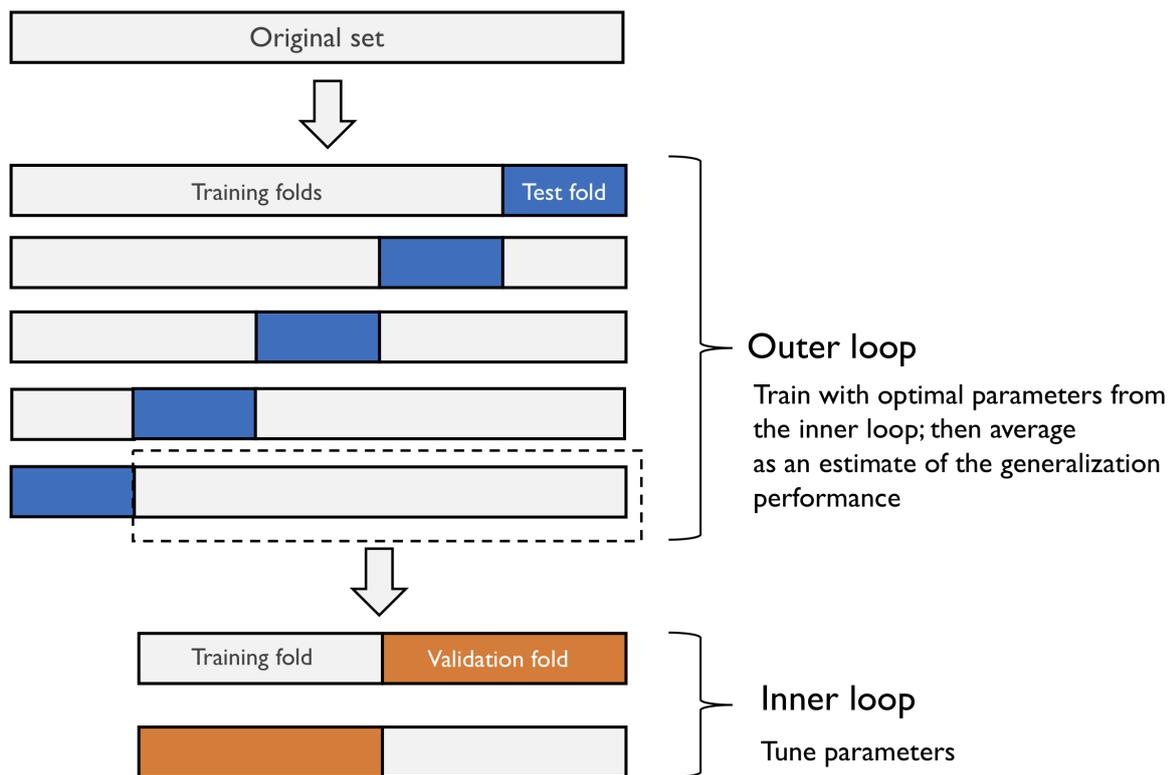
[1]https://github.com/TeamHG-Memex/sklearn-crfsuite

Figure 3.2 – Illustration of the nested-cross validation process. Source: https://bit.ly/2ClKLuX.

search or random search. Unlike these approaches, the iterative steps to optimize the machine learning algorithms are not random and do not explore all the search space. This leads to better configurations in fewer steps. But these methods, with the exception of random search, would also have been unsuitable for the CRFs in our case, since the computation time of this machine learning algorithm with our representation was very high. The protocol presented here is somewhat different from the following chapters, especially concerning the inner loop of nested cross validation since the training time of CRFs with this modellization was expensive (it required 22 hours to train a CRFs with this protocol).

We left the estimators for the CRFs algorithm by default: the sklearn-crfsuite implementation uses a gradient descent algorithm with the limited-memory BFGS method (L-BFGS). This method, although less efficient than the stochastic gradient descent (SGD) one,[2] is suited to our problem because it converges faster than the SGD algorithm and do not require a specific number of iterations as stopping criterion. The stopping criterion is

---

[2]http://www.chokkan.org/software/crfsuite/benchmark.html

either the minimization of the decrease in the objective function or the minimization of the projected gradient.

In order to evaluate the impact of taking into account time-series events in the prediction of hospitalization we performed our evaluation with non sequential state of the art algorithms described in Section 3.2.1 from the Scikit-Learn PEDREGOSA et al. [2011] library and with the CRFs algorithm described in Section 3.2.2 from the sklearn-crfsuite library,[3] a CRFsuite wrapper. The optimized hyperparameters determined by nested cross-validation are as follows:

- SVC, C-Support Vector Classifier, which implementation is based on libsvm CHANG et LIN [2011]: The penalty parameter C, the kernel used by the algorithm and the kernel coefficient gamma.

- RF, Random Forest classifier BREIMAN [2001]: The number of trees in the forest, the maximum depth in the tree, the minimum number of samples required to split an internal node, the minimum number of samples required to be at a leaf node and the maximum number of leaf nodes.

- L*og*, Logistic Regression classifier MCCULLAGH et NELDER [1989]: The regularization coefficient C and the penalty used by the algorithm.

- CRF*s*, Conditional Random Fields algorithm SUTTON et al. [2012]: The regularization coefficients $c1$ and $c2$ used by the solver L-BFGS.

3.3.3 Measure of test's accuracy

In order to assess the impact of the selected vector representations and machine learning algorithms, we evaluated the performance of the machine learning algorithms by using the $F_{tp,fp}$ metric FORMAN et SCHOLZ [2010]. Let TN be the number of negative instances correctly classified (True Negative), FP the number of negative instances incorrectly classified (False Positive), FN the number of positive instances incorrectly classified (False Negative) and TP the number of positive instances correctly classified (True Positive).

K represents the number of loops used to cross-validate (in our experiment K = 10), and the notation $_f$ is used to distinguish a fold related metric like the amount of true positives to the sum of true positives across all folds.

---

[3]https://github.com/TeamHG-Memex/sklearn-crfsuite

$$\text{TP}_f = \sum_{i=1}^{K} \text{TP}^{(i)} \qquad \text{FP}_f = \sum_{i=1}^{K} \text{FP}^{(i)} \qquad \text{FN}_f = \sum_{i=1}^{K} \text{FN}^{(i)}$$

$$\text{F}_{tp,fp} = \frac{2.\text{TP}_f}{2.\text{TP}_f + \text{FP}_f + \text{FN}_f}$$

The area under the curve (AUC) is another measure commonly used in medical studies. The AUC average is in particular used in a cross-validation context. However, this measure is mainly used for unbalanced datasets, which is not our case. That is why we will compare our algorithms with the $\text{F}_{tp,fp}$ measure.

3.3.4 Results

Table 3.1 presents the values of $\text{F}_{tp,fp}$ obtained with the above described state of the art machine learning algorithms on the dataset DS_B shaped with our sequential and non sequential representations.The sequential representation was used with CRF*s*, the non sequential representation with SVC, RF and L*og*.

Table 3.1 – $\text{F}_{tp,fp}$ of the selected classifiers on the balanced dataset DS_B.

SVC	RF	L*og*	CRF*s*
0.819	0.831	**0.850**	0.834

Logistic regression obtained the best performance on the prediction of hospitalization task with our experimental protocol.

3.4 Discussion

We proposed an approach to model temporal and persistent heterogeneous data in order to predict the advent of an event using the conditional random fields (CRFs) algorithm. We compared the sequential model of EMRs with a non sequential one and showed that, in the context of our model and task, and although the number of configurations tested was relatively small, there is no advantage on a performance side to opt for a sequential algorithm to predict hospitalization. One of the explanations could be that this algorithm is subject to overfitting with too large time-windows as shown by SINGH et al. [2015]. From an interpretability point of view, a sequential model allows to determine the influence of a factor and the time at which a risk factor appeared in the decision to hospitalize a person.

These results must also be put into perspective as the introduction of new data such as biological analysis can have an impact on the results obtained. However, the training of the CRFs algorithm is not optimized with the large amount of features introduced with a BOW.

3.5 Conclusion

As the experiments presented in this chapter did not demonstrate an advantage of sequential modelling over non sequential modelling, we will therefore conduct experiments in the following chapters with the non sequential modelling previously described. Also, the results obtained tend to prove that hospitalization prediction is a linearly separable problem.

As future work, we plan to further experiment on temporality models by exploring different time-windows, including different time spans before the hospitalization to evaluate their impact on the predictive power of each machine learning algorithms. By different temporality models, we also express different windows used by the model and not only consider a consultation as a time unit.

We also intend to test the improvement of the training of machine learning algorithms with the addition of masks, called erasing, on training data which allows to artificially increase a dataset and to improve the robustness of machine learning algorithms as shown by ZHONG et al. [2017]. Masks applied to training data is used by the BERT representation DEVLIN et al. [2018], which we have discussed in the state of the art of textual representations in Chapter 2.

On the subject of the integration of biomedical analysis, we will have to position ourselves on the discretisations to be used, and how to manage the outliers as well as how to deal with missing values. These are matters that we will need to address in order to be able to assess their impact on sequential and non sequential representations.

Chapter 4

Enrichment of representations of electronic medical records with domain knowledge

In this chapter we present the approach we proposed and evaluated to enrich the representation of electronic medical records with domain knowledge before learning things from them. Our goal was to show that the enrichment of the vector representations of electronic medical records (EMRs) with knowledge could improve the prediction of an event and, in particular, the hospitalization of a patient. In the next sections we will analyze and compare the impact of knowledge from different sources, whether separately incorporated or combined, on the vector representation of EMRs to predict hospitalization.

One of the main hypotheses we wanted to test was that the injection of domain knowledge into EMRs representations would have a positive impact on the predictions they could support. To extract the domain knowledge underlying the text descriptions written by general practitioners in EMRs, we search for medical entities in these texts and link them to the concepts of selected knowledge graphs. A knowledge graph, describes entities like objects, individuals from the real world as well as more abstract notions like ideas or events by representing their types, their attributes and the relationships between them and associating them with hierarchies of concepts (classes and relationships, ontologies, vocabularies). These knowledge graphs support not only interoperability and data integration, they also provide a shared reference and an extensible knowledge base from which to enrich other data sources and build other linked datasets. In our work we

consider both the well-known general knowledge graphs Wikidata and DBpedia and the health sector specific knowledge graphs such as those related to drugs.

Our work on the integration of knowledge from various knowledge graphs has been presented in GAZZOTTI et al. [2019a] and GAZZOTTI et al. [2019c]; our work on the extraction of relevant concepts from Wikipedia has been presented in GAZZOTTI et al. [2020].

In this chapter, we first present relevant works on the enrichment with knowledge graphs in Section 4.1, then we introduce Wikidata and DBpedia knowledge bases and detail the knowledge extracted from them in Section 4.2. We will also discuss the special case of knowledge obtained from DBpedia and, more precisely, we consider the question of filtering relevant domain knowledge from a general knowledge source and, in that context, the question of dealing with subjectivity in the annotation process of knowledge in Section 4.2.2. The Section 4.3.4 covers the extraction we performed on health sector specific knowledge graphs (ATC, ICPC-2 and NDF-RT). We detail how we integrate the extracted knowledge in the vector representation of EMRs in Section 4.4. Finally, we discuss the workflow for capturing knowledge from the different sources we used in Section 4.5 and we conclude this chapter in Section 4.6.

4.1 Relevant works on domain knowledge enrichment

In the following subsections, we will highlight the work carried out to exploit ontological knowledge with medical reports as well as work related to the semantic annotation of texts.

4.1.1 Relevant works on representations exploiting knowledge graphs

Approaches exploiting medical reports and biomedical knowledge graphs

Although the work cited here may be far from our task of predicting hospitalization, they present attempts to consider medical reports and biomedical knowledge graphs, whether from the literature or generated from specific corpus.

In MIN et al. [2017], the authors are focused on finding rules for the activities of daily living of cancer patients on the SEER-MHOS (Surveillance, Epidemiology, and End Results - Medicare Health Outcomes Survey) and they showed an improvement in the coverage of the inferred rules and their interpretations by adding 'IS-A' knowledge from the Unified

Medical Language System (UMLS[1]). They extract the complete sub-hierarchy of kinship and co-hyponymous concepts. Although their purpose is different from ours, their use of the OWL representation of UMLS with a machine learning algorithm improves the coverage of the identified rules. However, their work is based solely on 'IS-A' relationships without exploring the contributions of other kinds of relationships and they do not study the impact of this augmentation on different machine learning approaches: in their comparison, they only considered the AQ21 algorithm and the extension of this algorithm AQ21-OG.

In CHOI ET AL. [2017], to address data insufficiency and interpretation of deep learning models for the prediction of rarely observed diseases, the authors established a neural network with graph-based attention model that exploits ancestors extracted from the OWL-SKOS representations of ICD Disease, Clinical Classifications Software (CCS) and Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT). In order to exploit the hierarchical resources of these knowledge graphs in their attention mechanism, the graphs are transformed using the embedding obtained with Glove PENNINGTON et al. [2014]. The results show that the proposed model outperforms a standard recurrent neural network when identifying pathologies that are rarely observed in the training data and, at the same time, the model is also generalizing better when only few training instances are available.

In SALGUERO et al. [2018], to improve accuracy in the recognition of daily living activities, the authors extract knowledge from the dataset of ORDÓNEZ et al. [2013] and structure it with a knowledge graph developed for this purpose. The authors then propose an approach to automatically deduce new class expressions, with the objective of extracting their attributes to recognize activities of daily living using machine learning algorithms. The authors highlight better accuracy and results than with traditional approaches, regardless of the machine learning algorithm used for this task (up to 1.9% on average). Although they exploit solely the knowledge graph developed specifically for the purpose of discovering new rules, without trying to exploit other knowledge sources where a mapping could have been done, their study shows the value of structured knowledge in classification tasks. We intend here to study the same kind of impact but with different knowledge sources and for the task of predicting hospitalization.

---

[1]UMLS is a metathesaurus developed at the US National Library of Medicine
http://www.nlm.nih.gov/pubs/factsheets/umls.html

In FRUNZA et al. [2011], the authors show that combining bag-of-words (BOW), biomedical entities and UMLS improves classification results in several tasks such as information retrieval, information extraction and text summarizing, regardless of the classifier. We intend here to study the same kind of impact but from a more general repository like DBpedia and on a domain-specific prediction task. We also propose a method to select relevant domain knowledge in order to boost hospitalization prediction.

Graph embeddings and transformer based representations

The latest works on integrating knowledge in vector representations of texts is based on the combination of BERT (see Chapter 2) and graph embeddings.

Graph embedding methods consist in projecting the components of a knowledge graph in a vector space representation. One of them, TransE BORDES et al. [2013], consists in embedding entities in a vector space model; the property that links two entities in the knowledge graph are represented by a translation vector enabling to go from the embedding of the source entity to that of the target entity. However, their exists many more properties in knowledge graphs than those that link two entities together and this method is missing them.

ERNIE ZHANG et al. [2019], Enhanced Language Representation with Informative Entities, uses a pretrained model of TransE in order to combine the capabilities of transformers with a work derived from BERT and graph embeddings. It appears that ERNIE outperforms BERT on the English-language GLUE benchmark[2] (General Language Understanding Evaluation benchmark) WANG et al. [2018].

G-BERT SHANG et al. [2019], combines BERT and a graph neural network for the task of drug recommendation. The purpose of the neural network graph is to learn a representation of medical codes with the learning of medical ontology embedding. However, this approach suffers from the same shortcomings as most of the vector representations presented in the Section 4.1.1, because G-BERT only considers the hierarchical structure of knowledge graphs (hierarchical relations).

We must mention upfront that this part of the estate of the art is very recent and that methods combining BERT and graph embeddings did not exist at the beginning of this thesis. Moreover, representations like BERT require a large amount of data in order to train them, i.e. more data than with conventional neural network approaches. Graph em-

---

[2]https://gluebenchmark.com/

beddings, like neural networks, also lack of interpretability when it comes to explaining the decision of a machine learning algorithm to a physician. These were the main drawbacks that prevented us from using these models.

Our proposed approach does not require a large volume of data and exploits different types of relationships and various knowledge graphs. We also addressed the issue of selecting relevant knowledge for the task of predicting hospitalization with the particular case of DBpedia's knowledge.

4.1.2 Entity linking approaches for domain knowledge enrichment

Entity linking approaches aim to identify entities in free text and their related resources in a given knowledge graph. This kind of semantic annotation is of major interest in our approach since it allows us to extract information from knowledge graphs and from text.

The semantic annotators Dexter[3] CECCARELLI et al. [2013], EAGLET[4] JHA et al. [2017] and NERD[5] RIZZO et TRONCY [2011] are only dealing with the English language, therefore they are not relevant to our case study where we have to identify entities in French EMRs and exploit their corresponding resources from various knowledge graphs.

The SIFR Bioportal project TCHECHMEDJIEV et al. [2018] provides a web service based on NCBO BioPortal WHETZEL et al. [2011] to annotate clinical texts in French with biomedical knowledge graphs. This service is able to handle clinical notes involving negations, experiencers (the patient or members of his family) and temporal aspects in the context of the entity references. However, the adopted approach involves domain specific knowledge graphs, while general resources like EMRs require general repositories such as, for instance, DBpedia.

DBpedia Spotlight DAIBER et al. [2013] is a project dedicated to the automatic annotation of texts in eight different languages with DBpedia entities and proceed to their disambiguation through entity linking. The disambiguation of entities is performed by the generative model with maximum likelihood introduced by HAN et SUN [2011]. More refinements were obtained on the results by defining features induced from texts that helps to determine a threshold via linear regression. We will use DBpedia Spotlight to identify entities from the medical domain and link them to the DBpedia knowledge base.

---

[3]http://dexter.isti.cnr.it/
[4]https://github.com/dice-group/Eaglet
[5]http://nerd.eurecom.fr/

Multilingual ADGISTIS[6] (MAG) MOUSSALLEM et al. [2017] is a multilingual named entities disambiguator that uses a deterministic approach to link entities to their corresponding resources from a given knowledge graph (the online demonstrator only uses DBpedia). It relies on knowledge-base agnostic algorithms and outperformed all the publicly available state-of-the-art approaches on datasets present on the GERBIL platform[7] USBECK et al. [2015]. However, this method does not allow to automatically identify entities in a text.

Entity-fishing,[8] a software developed by science-miner[9] and Inria, is a named entity recognizer that disambiguates entities against Wikidata at a document level. It supports six languages and can handle PDF (Portable Document Format) and text. The authors use FastText word embeddings in order to generate candidates, then the entity candidates are ranked with gradient tree boosting and features derived from relations and context. Finally, these entities are selected by a random forest algorithm trained on an annotated corpus. We did not use Entity-fishing because were not able to compile it due to its complex architecture that involves different projects, however its development is still on-going and a lot of things are subject to change.

GATE[10] CUNNINGHAM [2002] (General Architecture for Text Engineering) is a language processing project developed at the Univerity of Sheffield in 1995 which relies on two mechanisms in order to proceed to the annotation of texts: the Ontology Annotation Tool (OAT[11]) and on JAPE (Java Annotation Patterns Engine) rules[12] that use a finite-state machine over annotations based on regular expressions. Thus, this framework allows to exploit both knowledge graphs and the power of regular expressions in order to automatically annotate text. Like SIFR Bioportal, GATE exploits labels from knowledge graph in a very simplistic way with a string matching approach to identify entities in text. As a result of the issues caused by the termination of the active development of SIFR Bioportal, we plan to use it in the future with domain specific knowledge graphs.

---

[6]http://aksw.org/Projects/AGDISTIS.html
[7]http://gerbil.aksw.org/gerbil/
[8]https://github.com/kermitt2/entity-fishing
[9]http://science-miner.com/
[10]https://gate.ac.uk/
[11]https://gate.ac.uk/sale/tao/splitch14.html#sec:ontologies:ocat
[12]https://gate.ac.uk/sale/tao/splitch8.html#chap:jape

4.2 Knowledge extraction based on general knowledge sources

In this section we present the knowledge bases that we have included in our study as well as the extraction procedures that we have used to obtain knowledge specific to the medical domain. In particular, Wikidata and DBpedia were chosen because general concepts can only be identified with general repositories.

4.2.1 Knowledge extraction based on Wikidata

Wikidata[13] is an open knowledge base, collaboratively edited, that centralizes data from various projects of the Wikimedia Foundation. We extracted drug-related knowledge by querying Wikidata's endpoint.[14] More precisely, we identified three properties of drugs relevant to the prediction of hospitalization: 'subject has role' (property `wdt:P2868`), 'significant drug interaction' (property `wdt:P2175`), and 'medical condition treated' (property `wdt:P769`).

In Wikidata, we identify the drugs present in EMRs using the ATC code (property `wdt:P267`) of the drugs present in the PRIMEGE database. The CUI UMLS (property `wdt:P2892`) and CUI RxNorm (property `wdt:P3345`) codes have been recovered using medical domain specific ontologies (ATC[15] and RxNorm[16] ontologies). Indeed, the codes from these three referentials are not necessarily all present to identify a drug in Wikidata, but at least one of them allows us to find the resource related to a given drug.

From the URI of a drug recognized in an EMRs, we extract property-concept pairs related to the drugs for the three selected properties (e.g. 'Pethidine' is a narcotic, 'Meprobamate' cures headache, 'Atazanavir' interacts with 'Rabeprazole'). To do this, we query Wikididata with the ATC, CUI UMLS and CUI RxNorm codes using the SPARQL query described in Listing 4.1.

---

[13] https://www.wikidata.org
[14] https://query.wikidata.org/sparql
[15] https://bioportal.bioontology.org/ontologies/ATC
[16] http://bioportal.bioontology.org/ontologies/RXNORM

Listing 4.1 – SPARQL query to extract property-concept pairs related to drugs from the Wikidata knowledge base.

```
1   SELECT ?property ?label where {
2       #ATC code, UMLS CUI code, RxNorm CUI code
3       {
4           SELECT ?y where {
5               ?y wdt:P267|wdt:P2892|wdt:P3345 ?code.
6               filter(?code in ("ATC_code", "CUI_UMLS_code", "CUI_RxNorm_code"))
7           } limit 1
8       }
9
10      # These are respectively the properties:  subject has role, medical condition treated, significant
            drug interaction.
11      VALUES ?property {
12          <http://www.wikidata.org/prop/direct/P2868>
13          <http://www.wikidata.org/prop/direct/P2175>
14          <http://www.wikidata.org/prop/direct/P769>
15      }
16      ?y ?property ?z.
17      ?z rdfs:label ?label.
18      filter(lang(?label) = 'en')
19  }
```

4.2.2   Knowledge extraction based on DBpedia

The knowledge base DBpedia structures knowledge pieces extracted from Wikipedia articles. This project is, like Wikipedia, multilingual and was initiated in 2007 by the Free University of Berlin[17] and the Leipzig University[18] in partnership with the company Open-Link Software.[19] DBpedia's applications are varied and can range from organizing content on a website to uses in the domain of artificial intelligence.

To detect in EMRs entities from the medical domain present in DBpedia, we used the semantic annotator DBpedia Spotlight DAIBER et al. [2013]. First, with the help of this annotator, we identify entities from text, then we proceed to the inspection of related resources in DBpedia. In order to improve DBpedia Spotlight's detection capabilities, words or abbreviated expressions within medical reports are added to text fields using a symbolic approach, with rules and dictionaries. For instance the abbreviation "ic" which means "heart failure" ("insuffisance cardiaque") is not recognized by DBpedia Spotlight, but is correctly identified by our rule-based approach. This dictionary contains more than 250 entries with terms and expressions dedicated to the medical domain.

As DBpedia is a general knowledge source, it is therefore necessary to perform a selection in order to extract relevant knowledge related to the medical field. The steps related to concept selection are described in the following subsections.

---

[17]https://www.fu-berlin.de/en
[18]https://www.uni-leipzig.de/en/
[19]https://www.openlinksw.com/

Table 4.1 – List of manually chosen concepts in order to determine a hospitalization, these concepts were translated from French to English (the translation does not necessarily exist for the English DBpedia chapter).

Speciality	Labels
Oncology	Neoplasm stubs, Oncology, Radiation therapy
Cardiovascular	Cardiovascular disease, Cardiac arrhythmia
Neuropathy	Neurovascular disease
Immunopathy	Malignant hemopathy, Autoimmune disease
Endocrinopathy	Medical condition related to obesity
Genopathy	Genetic diseases and disorders
Intervention	Surgical removal procedures, Organ failure
Emergencies	Medical emergencies, Cardiac emergencies

Preselected DBpedia concepts to be searched for in EMRs

Together with domain experts, we carried out a manual analysis of the named entities detected on a sample of approximately 40 consultations with complete information and determined 14 SKOS top concepts designating medical aspects relevant to the prediction of hospitalization, as they relate to severe pathologies. These are listed in Table 4.1.

For each EMR to model, from the list of resources identified by DBpedia Spotlight, we query the access point of the French-speaking chapter of DBpedia[20] to determine if these resources have as subject (property `dcterms:subject`) one or more of the 14 selected concepts.
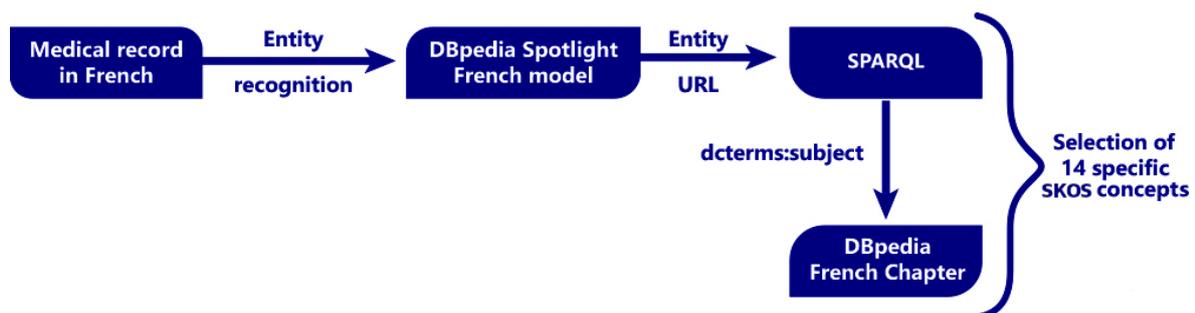
Figure 4.1 – Workflow diagram to extract DBpedia subjects from the list of 14 manually preselected subjects.

This manual selection of concepts can be seen as a preliminary step to the work we have conducted on the automatic selection of concepts from DBpedia, described in the following; the SPARQL query used to extract the concepts listed in Table 4.1 is a simpler variant of the one displayed in Listing 4.2.

---

[20]http://fr.dbpedia.org/sparql

Automatic selection of concepts from DBpedia

To further inspect the contribution of knowledge from DBpedia on the automatic selection of medical subjects, we have developed a new methodology to compensate the flaws of our first experiences with preselected concepts. This allows us to use a greater number of medical topics that can be found in DBpedia. Given the amount of general information available on DBpedia it is difficult to filter knowledge specific to the healthcare domain, and in particular, to identify notions relevant to the prediction of hospitalization. Moreover, the annotation process for determining concepts relevant to the analysis of risks related to hospitalization is complex and open to interpretation.

SPARQL query for medical concepts extraction   To ensure that the retrieved entities belong to the medical domain, we enforce two constraints on the resources identified by DBpedia Spotlight. The first constraint requires that the identified resources belong to the medical domain of the French chapter of DBpedia. The second one does the same with the English chapter in order to filter and select health domain-related subjects and to overcome the defects of the French version in which property `rdf:type` is poorly used. This involves calling two SERVICE clauses in the SPARQL query,[21] each one implementing a constraint according to the structure of the French and English chapter it remotely queries. The workflow is represented in Figure 4.2 and the query in Listing 4.2.

Listing 4.2 – SPARQL query to extract subjects related to the medical domain from DBpedia.

```
1  PREFIX dbo: <http://dbpedia.org/ontology/>
2  PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
3  PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>
4  PREFIX dcterms: <http://purl.org/dc/terms/>
5  PREFIX yago: <http://dbpedia.org/class/yago/>
6  PREFIX cat: <http://fr.dbpedia.org/resource/Catégorie:>
7
8  SELECT ?skos_subject WHERE {
9    SERVICE <http://fr.dbpedia.org/sparql> {
10     # Constraint on the medical domain
11     VALUES ?concept_constraint {
12       cat:Maladie             # disease
13       cat:Santé               # health
14       cat:Génétique_médicale  # medical genetics
15       cat:Médecine            # medicine
16       cat:Urgence             # urgency
```

---
[21] https://www.w3.org/TR/sparql11-query/

```
17        cat:Traitement          # treatment
18        cat:Anatomie            # anatomy
19        cat:Addiction           # addiction
20        cat:Bactérie            # bacteria
21      }
22      <link_dbpedia_spotlight> dbpedia-owl:wikiPageRedirects{0,1} ?page.
23      ?page dcterms:subject ?page_subject.
24      ?page_subject skos:broader{0,10} ?concept_constraint.
25      ?page_subject skos:prefLabel ?skos_subject.
26      ?page owl:sameAs ?page_en.
27      # Filter used to select the corresponding resource in the English Chapter of DBpedia
28      FILTER(STRSTARTS(STR(?page_en), "http://dbpedia.org/resource/"))
29    }
30
31    SERVICE <http://dbpedia.org/sparql> {
32      VALUES ?type_constraint {
33        dbo:Disease
34        dbo:Bacteria
35        yago:WikicatViruses
36        yago:WikicatRetroviruses
37        yago:WikicatSurgicalProcedures
38        yago:WikicatSurgicalRemovalProcedures
39      }
40      ?page_en a ?type_constraint
41    }
42  }
```

From the URIs of the identified resources, the first part of the query (lines 9-29) accesses the French chapter of DBpedia to check that the value of their property `dc-terms:subject`[22] belongs to one of the hierarchies of SKOS concepts (`skos:broader`, `skos:narrower`) having for roots the French terms for disease, health, medical genetics, medicine, urgency, treatment, anatomy, addiction and bacteria.

The second part of the query (lines 31-41) checks that the identified resources from the French DBpedia have for its English equivalent (`owl:sameAs`) at least one of the following types (`rdf:type`)[23]: dbo:Disease, dbo:Bacteria, yago:WikicatViruses, yago:WikicatRetroviruses, yago:WikicatSurgicalProcedures, yago:WikicatSurgicalRemovalProcedures. We do not consider some other types like `dbo:Drug`, `dbo:ChemicalCoumpound`, `dbo:ChemicalSubstance`, `dbo:Protein`, or `yago:WikicatMedicalTreatments`, as they generate answers related to chemical compounds: the retrieved resources can thus range from drugs to plants, to fruits. We do not consider either types referring to other living beings like `umbel-`

---

[22]Namespace: http://purl.org/dc/terms/
[23]Namespaces: http://dbpedia.org/ontology/, http://dbpedia.org/class/yago/

`rc:BiologicalLivingObject` or `dbo:Species` which are too general to return relevant results. We do not consider either many biomedical types in the yago namespace which URI ends by an integer (e.g. `http://dbpedia.org/class/yago/Retrovirus101336282`), which are too numerous and too close from each other. The type `dbo:AnatomicalStructure` is also non-relevant with this second constraint since it retrieves subjects related to different anatomical parts which are not human specific. The list of labels of concepts thus extracted allows us to construct a vector representation of EMR used to identify hospitalized patients.
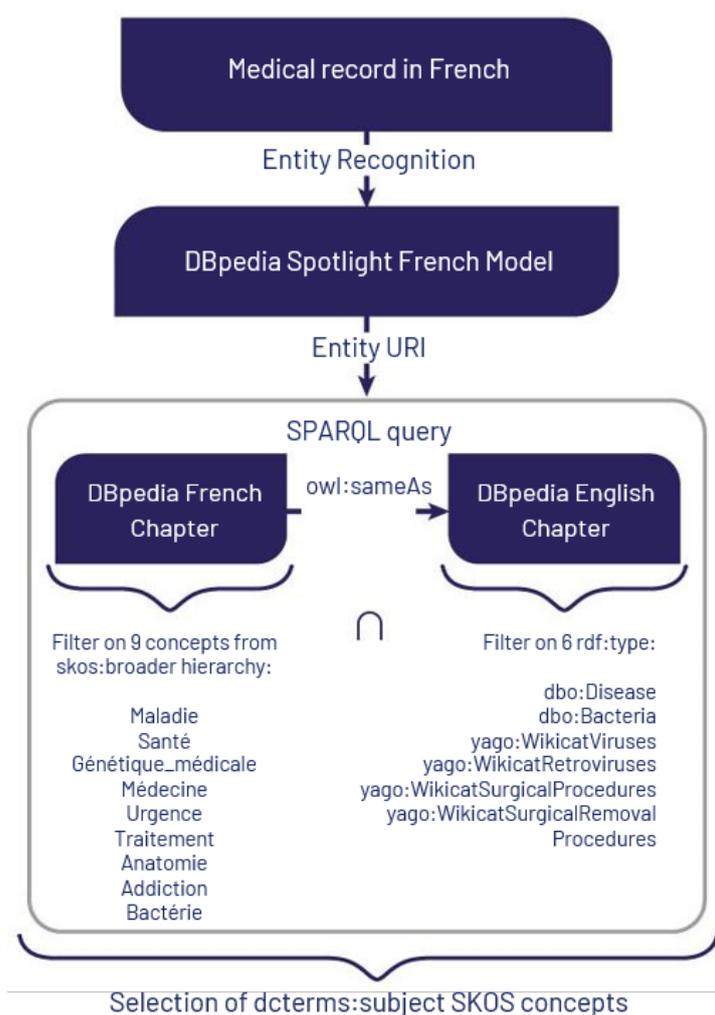
Figure 4.2 – Workflow used to extract candidate subjects from EMR.

Inter-rater reliability of DBpedia concept annotation To decide on the optimal vector representation of a patient's EMR, we considered further filtering the list of the labels of concepts extracted from DBpedia, depending on their relevancy for the targeted prediction task. We first submitted the list of the 285 extracted labels of concepts retrieved by

the SPARQL query in Listing 4.2 to human medical experts who were asked to assess the relevance of the labels of concepts for studying patients' hospitalization risks from their EMRs.

Two general practitioners and one biologist have independently annotated the 285 subjects extracted from DBpedia. The annotations were transformed in vectors with a size of 285. On average, among the 285 subjects proposed with the extraction based on the SPARQL query displayed in Listing 4.2, 198 subjects were annotated by experts as relevant to the study of patients' hospitalization risks (217 and 181 for the general practitioners and 196 for the biologist). We measured the inter-rater reliability with the Krippendorff's α metric Krippendorff [1970]. Figure 4.3 shows the workflow used to assess inter-rater reliability. We obtained a score of 0.51 when considering the three annotators, and 0.27 when considering only the annotation score between the two general practitioners.

Figure 4.3 – Workflow used to compute inter-rater reliability for both human and machine annotations.

We considered excluding some subjects involving a terminological conflict in their naming: if someone annotates the beginning of a label as relevant to predict the hospitalization of a patient (the opposite is also true) all the labels starting with the same expression will be annotated in the same way. The subjects excluded started by 'Biology', 'Screening and diagnosis', 'Physiopathology', 'Psychopathology', 'Clinical sign', 'Symptom' and 'Syndrome' which brings us back to a new total of 243 concepts. By doing so, the three annotators obtained a score of 0.66, and 0.52 for the inter-rater reliability between the two general practitioners.

As discussed by Artstein et Poesio [2008], a score within this range of values is insufficient to draw conclusions and it shows the difficulty of this task, both because identifying entities involved in patient hospitalization is subject to interpretation and because it is complex to find consensus in this task that could be seen at first sight as simplistic by an expert in the field.

Alternatively, we considered automatically selecting the concepts relevant for studying hospitalization by using a feature selection algorithm applied on a training set of vec-

Table 4.2 – Correlation metric ($1 - \frac{(u-\bar{u}).(v-\bar{v})}{\|u-\bar{u}\|_2 \|v-\bar{v}\|_2}$, with $\bar{u}$, the mean of elements of u, and respectively $\bar{v}$, the mean of elements of v) computed on the 285 subjects. A_1 to A_3 refers to human annotators and M_1 to M_{10} refers to machine annotators through feature selection annotation on the ζ approach (considering the 10 K-Fold). U_1 is the union of subjects from the sets M_1 to M_{10}. Cells in red are strictly superior to 0.5, cells in orange are between 0.25 and 0.5, cells in cyan are strictly inferior to 0.25.

	A_1	A_2	A_3	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9	M_{10}	U_1
A_1	\	0.6814	0.4180	1.1085	1.0688	1.1138	1.1399	1.0692	1.1166	1.1085	1.0688	1.1257	1.1363	1.1405
A_2	0.6814	\	0.2895	1.0618	1.1066	1.0072	1.0745	1.0534	1.1127	1.0618	1.0611	1.0904	1.0749	1.0737
A_3	0.4180	0.2895	\	1.0232	1.0807	1.0242	1.0721	1.0616	1.0708	1.0232	1.0320	1.0708	1.0520	1.0933
M_1	1.1085	1.0618	1.0232	\	0.2105	0.2635	0.2249	0.3410	0.3389	0.2116	0.2105	0.2031	0.2760	0.3293
M_2	1.0688	1.1066	1.0807	0.2105	\	0.2319	0.1605	0.1597	0.2037	0.1714	0.0724	0.2358	0.3019	0.2605
M_3	1.1138	1.0072	1.0241	0.2635	0.2319	\	0.1408	0.2700	0.2865	0.2249	0.1605	0.3346	0.2710	0.2472
M_4	1.1399	1.0745	1.0721	0.2249	0.1605	0.1408	\	0.2700	0.2527	0.1863	0.1248	0.2495	0.2710	0.2472
M_5	1.0692	1.0534	1.0616	0.3410	0.1597	0.2700	0.2700	\	0.2508	0.2379	0.1597	0.3595	0.4167	0.1200
M_6	1.1166	1.1127	1.0708	0.3389	0.2037	0.2865	0.2527	0.2508	\	0.2275	0.2037	0.3690	0.3495	0.2080
M_7	1.1085	1.0618	1.0232	0.2116	0.1714	0.2249	0.1863	0.2379	0.2275	\	0.1322	0.1565	0.3238	0.3293
M_8	1.0688	1.0611	1.0320	0.2105	0.0724	0.1605	0.1248	0.1597	0.2037	0.1322	\	0.2358	0.3019	0.2605
M_9	1.1257	1.0904	1.0708	0.2031	0.2358	0.3346	0.2495	0.3595	0.3690	0.1565	0.2358	\	0.2888	0.4030
M_{10}	1.1363	1.0749	1.0520	0.2760	0.3019	0.2710	0.2710	0.4167	0.3495	0.3238	0.3019	0.2888	\	0.4185
U_1	1.1405	1.0737	1.0933	0.3293	0.2605	0.2472	0.2472	0.1200	0.2080	0.3293	0.2605	0.4030	0.4185	\

tor representations of patients containing the concepts extracted from knowledge graphs. We generated different feature sets as subsets of the whole bag of concepts that can be used to represent a patient. They either follow the expert annotation or the machine annotation, and they consider different text fields from EMRs. These different feature sets will be detailed in Section 4.2.3.

The union of labels of concepts identified with the *sm* approach counts 51 different subjects (63 if we distinguish concepts according to their provenance fields in EMRs) and the intersection of labels of concepts identified with *sm* counts 14 different subjects (19 if we distinguish concepts according to their provenance fields in EMRs). Table 4.2 displays correlation metric values between experts and machine annotators (its value ranges from 0 to 2, meaning that 0 is a perfect correlation, 1 no correlation and 2 perfect negative correlation). This metric was computed by comparing among the 285 subjects, if they are deemed relevant, irrelevant or not annotated (in the case of human annotation) to study the patient's hospitalization risks from their EMR, thus vectors are compared in pairs in this table.

Table 4.2 shows up a wide variation between human annotators and machine annotators (maximum of 1.1399 between A_1 and M_4), whereas between annotators of a specific group this margin is not significant (maximum of 0.6814 for humans and maximum of 0.4185 for machines). The union of subjects U_1 retrieved by machine annotators is really similar to M_5, since they have a correlation score of 0.12.

4.2.3 Notations and feature sets using general knowledge sources

We introduce below the notation used to represent feature sets obtained from the general knowledge sources Wikidata and DBpedia. These notations will allow us to differentiate the many configurations we tested during the experiments on hospitalization prediction.

In order to use the knowledge from Wikidata, we used the ATC codes and links related to CUI UMLS and CUI RxNorm codes with the mapping to domain specific knowledge graphs.

- The notation $+wa$ refers to an approach using the enrichment of our representations with the property 'subject has role' (*wdt:P2868*) from Wikidata.

- $+wm$ indicates the usage of the property 'medical condition treated' (*wdt:P2175*) from Wikidata.

- $+wi$ refers to the usage of the property 'significant drug interaction' (*wdt:P769*) from Wikidata.

We used DBpedia Spotlight to identify entities in text fields, and we extracted DBpedia subjects related to the medical domain.

- The $+s*$ notation refers to an approach using the enrichment of representations with concepts among the list of manually selected concepts (see Table 4.1) from DBpedia. This approach does not exploit all text fields to extract knowledge from DBpedia, these fields are related to the patient's own record with: the patient's personal history, allergies, environmental factors, current health problems, reasons for consultations, diagnosis, medications, care procedures, reasons for prescribing medications and physician observations.

- The $+s$ notation refers to an approach using the enrichment of representations with concepts among the list of manually selected concepts (see Table 4.1) from DBpedia. This approach uses all text fields to identify entities with: the patient's personal history, family history, allergies, environmental factors, past health problems, current health problems, reasons for consultations, diagnosis, medications, care procedures, reasons for prescribing medications, physician observations, symptoms and diagnosis.

- $+s*$ T refers to an enrichment with the labels of concepts automatically extracted from DBpedia with the help of the SPARQL query in Listing 4.2, 285 concepts are thus considered with this approach. This approach uses the same text fields as $+s*$ to identify entities from DBpedia.

- $+s*\cap$ refers to an enrichment with a subset of the labels of concepts automatically extracted from DBpedia acknowledged as relevant by at least one expert human annotator. This approach uses the same text fields as $+s*$ to identify entities from DBpedia.

- $+s*\cup$ refers to an enrichment with a subset of the labels of concepts automatically extracted from DBpedia acknowledged as relevant by all the experts human annotators. This approach uses the same text fields as $+s*$ to identify entities from DBpedia.

- $+s*m$ refers to an enrichment with a subset of the labels of concepts automatically selected by using a feature selection algorithm. We chose the Lasso algorithm TIBSHIRANI [1996] and we executed it *within* the internal loop of the nested cross-validation in the global machine learning algorithm chosen to predict hospitalization. This approach uses the same text fields as $+s*$ to identify entities from DBpedia.

- $+sm$ uses the same enrichment procedure of $+s*m$ to automatically select a subset of the labels of concepts. This approach exploits the same text fields as $+s$ (all text fields) to identify entities from DBpedia.

4.3 Knowledge extraction based on domain specific ontologies

The majority of knowledge graphs related to the biomedical field can be found on the repositories OBO Foundry[24] and BioPortal.[25] Biomedical knowledge graphs can also be searched through the Ontology Lookup Service.[26] In addition to general knowledge graphs, we were also interested in the impact of contributions from domain specific

---

[24]http://www.obofoundry.org/
[25]http://bioportal.bioontology.org/
[26]https://www.ebi.ac.uk/ols/index

knowledge graphs, especially for text fields containing international drug codes from the ATC classification and codes related to the reasons for consulting a general practitioner with the International Classification of Primary Care (ICPC-2). We thus extracted knowledge based on three OWL representations specific to the medical domain: ATC,[27] NDF-RT[28] and ICPC-2.[29] The choice of OWL-SKOS representations of ICPC-2 and ATC in our study comes from the fact that the PRIMEGE database adopts these nomenclatures, while the OWL representation of NDF-RT provides additional knowledge on interactions between drugs, diseases, mental and physical conditions.

4.3.1   Knowledge extraction from ATC

The ATC classification was first published in 1976 and is maintained by the World Health Organization Collaborating Centre for Drug Statistics Methodology (WHOCC).[30] This classification is used to group the active ingredient of drugs according to the organ or system on which it interacts. It was originally used to improve the quality of drug use. This classification is composed of five hierarchical depth levels and 14 main groups (see Table A.1). Each letter or doublet of digits represents there a hierarchical level.

From the ATC OWL-SKOS representation, we extracted the labels of the superclasses of the drugs listed in the PRIMEGE database, using the properties `rdfs:subClassOf` and `member_of` on different depth levels thanks to SPARQL 1.1 queries with property paths.[31] For instance, the 'meprednisone' (ATC code: H02AB15) has as superclass 'Glucocorticoids, Systemic' (ATC code: H02AB) which itself has as superclass 'CORTICOSTEROIDS FOR SYSTEMIC USE, PLAIN' (ATC code: H02).

An example of SPARQL query to retrieve the first and second depth levels of superclasses from the ATC knowledge graph is presented in Listing 4.3.

---

[27] Anatomical Therapeutic Chemical Classification, https://bioportal.bioontology.org/ontologies/ATC

[28] National Drug File - Reference Terminology, https://bioportal.bioontology.org/ontologies/NDF-RT

[29] International Primary Care Classification, http://bioportal.lirmm.fr/ontologies/CISP-2

[30] https://www.whocc.no/

[31] https://www.w3.org/TR/sparql11-query/

Listing 4.3 – Example of SPARQL query to retrieve ATC superclasses.

```
1  PREFIX uatc: <http://purl.bioontology.org/ontology/UATC/>
2
3  SELECT ?label where {
4      ?x skos:notation  "ATC_code".
5      # Extraction of the first and second depth levels
6      ?x rdfs:subClassOf{1,2}|uatc:member_of|uatc:member_of/rdfs:subClassOf{1} ?y.
7      ?y skos:notation ?label
8      FILTER(STRSTARTS(STR(?x), "http://purl.bioontology.org/ontology/UATC/"))
9  }
```

4.3.2   Knowledge extraction from ICPC-2

ICPC-2 is a revised version in 1998 of the ICPC classification, developed in 1987 developed by the World Organization of Family Doctors International Classification Committee (WICC).[32] This classification lists the reasons of consultation, diagnoses and health care interventions. It is composed of only two hierarchical depth levels and 17 main groups (see Table A.2). The development of ICPC-3 started in 2018 and is still in progress.[33]

There are other classifications related to reasons of consultation and diagnoses such as ICD10 (International Statistical Classification of Diseases and Related Health Problems). However, the ICD10 is more complex to apply in the sense that you can find very (overly) specific events like W55.03XD entitled 'Scratched by cat, subsequent encounter' as well as many possible relevant codes, all associated to a condition like 'sinusitis' (there are 26 different results for sinusitis). Originally, this classification was not adopted for these reasons, since a small number of annotations with ICPC-2 codes were found for diagnoses and reasons for consultation in the PRIMEGE project. As a result, an automated coding procedure LACROIX-HUGUES [2016] has been implemented to increase the number of reasons of consultation and diagnoses annotated with the ICPC-2 classification in this database.

As we did with ATC, we extracted from the OWL-SKOS representation of ICPC-2 the labels of the superclasses, by exploiting property `rdfs:subClassOf`. However, given the limited depth of this representation, it is only possible to extract one superclass per diagnosed health problem or identified care procedure. For instance, 'Symptom and complaints' (ICPC-2 code : H05) has for superclass 'Ear' (ICPC-2 code : H).

The SPARQL query to extract the superclasses from the ICPC-2 knowledge graph is presented in Listing 4.4.

---

[32]http://wicc.news/
[33]http://www.icpc-3.info/about-project/

Listing 4.4 – SPARQL query to retrieve ICPC-2 superclasses.

```
1  select ?label where {
2      ?x skos:notation <ICPC-2_code>.
3      ?x rdfs:subClassOf ?y.
4      ?y skos:prefLabel ?label
5      FILTER(?y not in (<http://chu-rouen.fr/cismef/CISP-2#ARBO>)
6          && STRSTARTS(STR(?y), "http://chu-rouen.fr/cismef/CISP-2"))
7  }
```

4.3.3   Knowledge extraction from NDF-RT

NDF-RT is produced by the U.S. Departement of Veterans Affairs.[34] This classification organises drugs and models their characteristics such as ingredients, chemical structure, dose form, physiologic effect, mechanism of action, pharmacokinetics, and related diseases. The successor[35] (Medication Reference Terminology) of this project is MED-RT,[36] but there is not yet any transposition of it into Semantic Web standards.

In the OWL representation of NDF-RT, we selected three drug properties relevant to the prediction of hospitalization:

- 'may_treat' property (e.g. 'Tahor', which main molecule is 'Atorvastatin' (ATC code: C10AA05) can cure 'Hyperlipoproteinemias' (Hyperlipidemia)).

- 'CI_with' (e.g. 'Tahor' is contraindicated in 'Pregnancy').

- 'may_prevent' (e.g. 'Tahor' can prevent 'Coronary Artery Disease').

A dimension in our EMR vector representation will be a property-value pair. The drug Tahor is described in RDF as follows in NDF-RT:

```
1   @prefix : <http://evs.nci.nih.gov/ftp1/NDF-RT/NDF-RT.owl> .
2   :N0000022046 a owl:Class; rdfs:label "ATORVASTATIN"; :UMLS_CUI "C0286651";
3       owl:subClassOf [
4           rdf:type owl:Restriction; owl:onProperty :may_prevent;
5           owl:someValuesFrom :N0000000856 ];
6       owl:subClassOf [
7           rdf:type owl:Restriction; owl:onProperty :CI_with;
8           owl:someValuesFrom :N0000010195 ];
9       owl:subClassOf [
10          rdf:type owl:Restriction; owl:onProperty :may_treat;
11          owl:someValuesFrom :N0000001594 ].
12  :N0000000856 rdfs:label "Coronary Artery Disease [Disease/Finding]".
13  :N0000010195 rdfs:label "Pregnancy [Disease/Finding]".
14  :N0000001594 rdfs:label "Hyperlipoproteinemias [Disease/Finding]".
```

---

[34]https://www.va.gov/HEALTH/
[35]https://evs.nci.nih.gov/ftp1/MED-RT/Introduction%20to%20MED-RT.pdf
[36]https://evs.nci.nih.gov/ftp1/MED-RT/MED-RT%20Documentation.pdf

The SPARQL query in Listing 4.5 allowed us to extract property-concept pairs associated with the properties may_treat, may_prevent and CI_with from the NDF-RT knowledge graph. The ATC code allows to retrieve these properties for a given drug.

Listing 4.5 – SPARQL query to retrieve property-concept pairs associated with the properties may_treat may_prevent and CI_with from NDF-RT.

```
1   PREFIX ndfrt: <http://evs.nci.nih.gov/ftp1/NDF-RT/NDF-RT.owl#>
2
3   SELECT DISTINCT ?result WHERE {
4       ?x skos:notation <ATC_code>.
5       ?x <http://bioportal.bioontology.org/ontologies/umls/cui> ?cui.
6       ?w ndfrt:UMLS_CUI ?cui.
7       ?w rdfs:subClassOf ?y.
8       ?y owl:onProperty ?property.
9       ?y owl:someValuesFrom ?z.
10      ?z rdfs:label ?label
11      BIND(concat(strafter(?property, "NDF-RT.owl#"), "#", ?label) as ?result)
12      filter(?property in (ndfrt:may_treat, ndfrt:may_prevent, ndfrt:CI_with))
13  }
```

4.3.4   Notations and feature sets using domain specific ontologies

We introduce below the notation used to enrich our vector representation of EMRs with domain specific knowledge graphs. As we have seen above, in order to enrich our representation, we used the ATC and ICPC-2 codes. The ATC codes were used to extract concepts from the ATC and NDF-RT knowledge graphs, and ICPC-2 codes for concepts from ICPC-2.

- The notation $+c$ refers to an approach using the enrichment of vector representation with ATC and the number attached specifies the different depth levels used. For instance, $+c_{1-3}$ indicates that 3 superclass depth levels are integrated in the same vector representation.

- $+t$ indicates the enrichment of vector representations with ICPC-2.

- $+d$ indicates the enrichment of vector representations with NDF-RT. $+d$ is followed by indices CI if the property 'CI_with' is used, $prevent$ if property the 'may_prevent' is used and $treat$ if the property 'may_treat' is used. For instance, $+d_{\text{CI},prevent,treat}$ refers to the case where these three properties are used together in the same vector representation of EMRs.

4.4 Integrating ontological knowledge in vector representations of electronic medical records

Now that we have seen what kinds of knowledge can be extracted from different sources, here we detail how we have integrated it in the non-sequential representation of EMRs presented in Section 2.4.2.

Concepts from knowledge graphs are considered as a token in a textual message. When a concept is identified in a patient's medical record, this concept is added to a concept vector. This attribute will have as value the number of occurrences of this concept within the patient's health record. For instance, the concepts 'Organ Failure' and 'Medical emergencies' are identified for 'pancréatite aiguë', acute pancreatitis, and the value for these attributes in our concept vector will be equal to 1.

Similarly, if a property-concept pair is extracted from a knowledge graph, it is added to the concept vector. For instance, in vectors exploiting NDF-RT (enrichment with $+d$), we find the couple consisting of `CI_with` as a property - contraindicated with- and the name of a pathology or condition, for instance 'Pregnancy'.

Let $V^i = \{w_1^i, w_2^i, ..., w_n^i\}$ be the bag-of-words obtained from the textual data in the EMR of the ith patient. Let $C^i = \{c_1^i, c_2^i, ..., c_n^i\}$ be the bag of concepts for the ith patient resulting from the extraction of concepts belonging to knowledge graphs from semi-structured data of his consultations such as text fields listing drugs and pathologies with their related codes, and unstructured data from free texts such as observations. The different machine learning algorithms exploit the aggregation of these two vectors: $x^i = V^i \oplus C^i$.

For instance, in the following sentence:
"prédom à gche - insuf vnse ou insuf cardiaque - pas signe de phlébite - - ne veut pas mettre de bas de contention et ne veut pas augmenter le lasilix... -"

meaning:
"(predom[inates] on the left, venous or cardiac insuf[ficiency], no evidence of phlebitis, does not want to wear compression stockings and does not want to increase the lasix...)"

the expression 'insuf cardiaque', meaning 'heart failure', refers to two concepts listed in Table 4.1: 'Organ failure' and 'Cardiovascular disease', these concepts were retrieved by the property `dcterms:subject` from DBpedia. The concept vector of occurrences that represents the patient's EMR will therefore have a value of 1 for the attributes representing the concepts 'Organ Failure' and 'Cardiovascular Disease' (Figure 4.4).

	Patient 1:	Patient 2:
	prédom à gche - insuf vnse ou insuf cardiaque - pas signe de phlébite - - ne veut pas mettre de bas de contention et ne veut pas aumenter le lasilix...	procédure FIV - - transfert embryon samedi dernière - a fait hyperstimulation ovarienne; rupture de kyste - - asthénie, - - dleur abdo, doulleur à la palpation ++ - - voit gynéco la semaine prochaine pr controle betahcg, echo

	C[1]: Organ failure	C[2]: Cardiovascular disease	C[3]: Neoplasm stubs	...
Patient 1	1	1	0	...
Patient 2	0	0	1	...

Figure 4.4 – Concept vectors generated for two EMRs with the bag-of-words approach under the $+s$ configuration. The translation and correction of the texts are (a) for patient 1: "predom[inates] on the left, venous or cardiac insuf[ficiency], no evidence of phlebitis, does not want to wear compression stockings and does not want to increase the lasix". and (b) for patient 2: "In vitro fertilization procedure, embryo transfer last Saturday, did ovarian hyperstimulation, cyst rupture, asthenia, abdominal [pain], [pain] on palpation ++, will see a gyneco[logist] next week [for] a beta HCG, echo check-up".

Table 4.3 – Alternative concept vector representations resulting from the EMR of a patient under Tahor with the NDF-RT knowledge graph.

	C[1]: may_treat#Hyperlipoproteinemias	C[2]: CI_with#Pregnancy	C[3]: may_prevent#Coronary Artery Disease	...
+d_prevent	∅	∅	1	...
+d_CI	∅	1	∅	...
+d_treat	1	∅	∅	...
+d_CI,prevent,treat	1	1	1	...

As for the exploitation of NDF-RT, let us consider again the example description of the drug Tahor introduced in Section 4.3. It can be used to enrich the vector representation of the EMRs of patients under Tahor as detailed in Table 4.3. This table shows, in particular, how we have integrated property-concept pairs into our vector representation.

4.5 Discussion

The number of usable biomedical referentials is not limited to those presented above, we can therefore consider extending this study with other nomenclatures. Nevertheless, we presented different knowledge graphs and procedures related to data extraction for integration in vector representations of EMRs. In particular, the knowledge graphs of ICPC-2 and ATC were chosen to match the PRIMEGE relational database model.

The summarization of all mapping to knowledge graphs is represented in Figure 4.5. It

describes the links and entities used to query the different knowledge graphs considered during this study:

- With ICPC-2 codes to query ICPC-2.

- With ATC codes to query NDF-RT, ATC, Wikidata.

- With entities identified in free text via DBpedia Spotlight to query DBpedia.

As mentioned above, Wikidata does not necessarily contain the CUI RxNorm, CUI UMLS, or ATC codes for each of the drugs represented in it. That is why we had to do the mapping with the ATC knowledge graph to get the CUI UMLS codes, and with the RxNorm graph to get the CUI RxNorm codes. Other codes could have been obtained by linking information with other knowledge graphs such as Mesh codes, DrugBank codes, but the identifiers already obtained seem sufficient to request Wikidata on drugs.

Other entities and relations could have been extracted and identified in free text in EMRs but this requires the use of other semantic annotators to identify resources from knowledge graphs. On this subject, the first tests to use the SIFR annotator[37] TCHECHMED-JIEV et al. [2018] in our benchmark were not conclusive, because it was impossible to compile the project and we realized in the meantime, that this project is not currently maintained. Moreover, it is excluded for us to use the online portal of this project[38] because patient data are confidential. The transmission of these data to third parties is therefore prohibited, which prevents us from using online APIs or at least providing complete information on patient consultations. One possibility that we can exploit in the future is to rely on GATE[39] (General Architecture for Text Engineering) or another similar tool to identify entities in free text with domain specific knowledge graphs.

Our attempts were also unsuccessful with the semantic annotator entity-fishing,[40] an annotator dedicated to named entities linking with Wikidata resources. The complex architecture of this software makes it difficult to compile, especially since there are models to download and dependencies to other projects. The project entity-fishing is still in active development, which suggests that we will certainly be able to use it in the future. This would allow us to extract other information, not restricted to drugs with Wikidata, and thus to cover more broadly the possibilities offered by this knowledge base.

---

[37] https://github.com/sifrproject/docker-compose-bioportal
[38] http://bioportal.lirmm.fr/annotator
[39] https://gate.ac.uk/
[40] https://github.com/kermitt2/entity-fishing
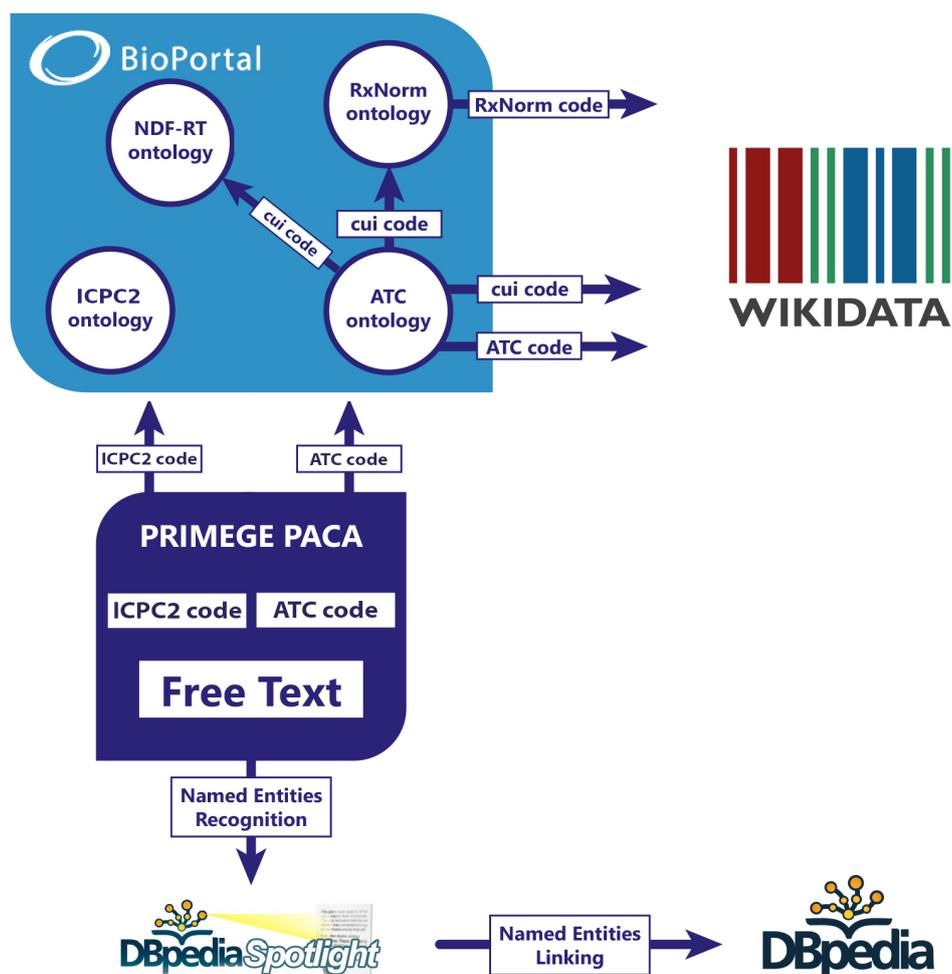
Figure 4.5 – Workflow of the mapping used to match ATC codes, ICPC2 codes with medical domain ontologies. The links used to proceed to the mapping with the knowledge bases Wikidata and DBpedia are also described.

4.6 Conclusion

The crucial part of this chapter was to show off how to link both structured and unstructured information from EMRs to knowledge graphs. We have been able to request knowledge graphs using SPARQL queries to retrieve the resources corresponding to entities identified within text. These entities can relate to expressions or international codes from the medical domain, allowing us to refer to knowledge graphs once these entities are linked to their corresponding resources. The knowledge extraction procedure is carried out upstream before injecting the knowledge retrieved in the vector representation of EMRs.

Related to the point discussed in Section 4.3.2 about the small number of annotations using international codes, the usage of semantic annotators and mapping with knowl-

edge graphs is another way to automatically increase the number of annotated data. This would improve the reasoning possibilities on EMRs, as we can enrich a dataset with new knowledge.

Now that we have presented how to extract knowledge from different knowledge graphs, and how to enrich EMRs vector representation with it, the next chapter will present our results on the evaluation of knowledge injection for the task of hospitalization risk prediction.

Chapter 5

Predicting hospitalization based on electronic medical records representation enrichment

In the previous chapter, we presented our approach to extract knowledge from the texts in patients' electronic medical records (EMRs) and to inject it into the vector representation of EMRs. This chapter is devoted to the use and evaluation of these enriched representations with different feature sets derived from knowledge graphs to predict hospitalization with different methods. More precisely we report the results of our study of which domain knowledge combined with which machine learning methods are the most suited to improve the prediction of a patient's hospitalization. We first present in Section 5.1 our experimental protocol and notation used for our features extracted from knowledge graphs in Section 5.2, then we present our results in Section 5.3 and discuss them in Section 5.4. Finally, we will conclude and discuss our perspectives for future work in Section 5.5.

563

5.1 Protocol & evaluation

We reuse the DS_B dataset we previously used in the experiments described in Chapter 3. It is composed of 714 hospitalized patients and 732 patients who were not hospitalized. The way we detect hospitalization events and the preprocessing steps remain unchanged.

Similarly to Chapter 2, since we use non sequential machine learning algorithms to assess the enrichment of ontological knowledge, we had to aggregate all patients' consul-

tations in order to overcome the temporal dimension inherent in symptomatic episodes occurring during a patient's lifetime. Thus, all consultations occurring before hospitalization are aggregated into a vector representation of the patient's medical file. For patients who have not been hospitalized, all their consultations are aggregated.

Just like in the experiments presented in Chapter 3, to construct V^i (the BOW) for our non sequential representation, we consider the following EMR fields:

- sex

- birth year

- long term condition

- risk factors

- allergies

- reasons of consultation with their associated codes

- medical observations

- diagnosis with their associated codes

- care procedures

- the drugs prescribed with their associated codes

- current health problems

- reasons of the prescription

In addition to the previous fields we added a number of fields for which we prefix the terms and concepts in order to capture the fact they apply to different aspects e.g. feature of a patient vs feature of the family of the patient. These prefixed fields are:

- patient's history (prefix: '#history#')

- family history (prefix: '#family#')

- past problems (prefix: '#past_problem#')

- symptoms (prefix: '#symptom#')

- diagnosis of the patient with their associated codes (prefix: '#diagnosis#')

5.1.1 Material and softwares

The different experiments were conducted on a HP EliteBook 840 G2, 2.6 GHz, 16 GB RAM with a virtual environment under Python 3.6.3 as well as a Precision Tower 5810, 3.7GHz, 64GB RAM with a virtual environment under Python 3.5.4. The creation of vector representations was done on the HP EliteBook and on this same machine were deployed DBpedia Spotlight as well as domain-specific knowledge graphs with the Corese Semantic Web Factory CORBY et ZUCKER [2010],[1] a software platform for the Semantic Web. It implements RDF, RDFS, SPARQL 1.1 Query & Update. Corese was also used to query Wikidata and DBpedia knowledge bases through their endpoints.

5.1.2 Search space for hyperparameters of machine learning algorithms

We evaluated vector representations enriched with knowledge graphs with state of the art algorithms from the Scikit-Learn PEDREGOSA et al. [2011] library and using nested cross-validation CAWLEY et TALBOT [2010]. The outer loop was executed with a K = 10, and the inner loop with a L = 3. The exploration of hyperparameters was performed by random search BERGSTRA et BENGIO [2012] over 150 iterations. Compared to Chapter 3 the folds have been re-shuffled. The optimized hyperparameters determined by nested cross-validation are the following:

- SVC, C-Support Vector Classification, which implementation is based on libsvm CHANG et LIN [2011]: The penalty parameter C, the kernel used by the algorithm and the kernel coefficient gamma.

- RF, Random forest classifier BREIMAN [2001]: The number of trees in the forest, the maximum depth in the tree, the minimum number of samples required to split an internal node, the minimum number of samples required to be at a leaf node and the maximum number of leaf nodes.

- L*og*, Logistic regression classifier MCCULLAGH et NELDER [1989]: The regularization coefficient C and the penalty used by the algorithm.

---

[1]http://corese.inria.fr

5.2   Notation and characteristics of candidate feature sets

We generate our different vector representations with a combination of features induced by various knowledge graphs. Thus, our vector representations are composed of V^i, the BOW, and subsets of C^i, the vector of concepts formed by subsets extracted from various knowledge graphs. We include below the notations used in the enrichment of our vectors with knowledge graphs.

- *baseline*: represents our basis of comparison where no ontological enrichment is made on EMR data, i.e. only text data in the form of bag-of-words: V^i.

- $+s$: refers to an enrichment with concepts from DBpedia among the list of the 14 concepts in Table 4.1. This approach involves the following text fields to identify entities: the patient's personal history, family history, allergies, environmental factors, past health problems, current health problems, reasons for consultations, diagnosis, medications, care procedures, reasons for prescribing medications, physician observations, symptoms and diagnosis.

- $+s*$: refers to an enrichment with concepts from DBpedia among the list of the 14 concepts in Table 4.1. When compared to $+s$, not all the text fields are exploited; concepts are extracted from the following text fields: patient's personal history, allergies, environmental factors, current health problems, reasons for consultations, diagnoses, medications, care procedures followed, reasons for prescribing medications and physician observations.

- $+t$: refers to an enrichment with concepts from the OWL-SKOS representation of ICPC-2.

- $+c$: refers to an enrichment with concepts from the OWL-SKOS representation of ATC, the number, or number interval indicates the different hierarchical depth levels used.

- $+wa$: refers to an enrichment with Wikidata's 'subject has role' property (`wdt:P2868`).

- $+wi$: refers to an enrichment with Wikidata's 'significant drug interaction' property (`wdt:P769`).

- $+wm$: refers to an enrichment with Wikidata's 'medical condition treated' property (`wdt:P2175`).

- $+d$: refers to an enrichment with concepts from the NDF-RT OWL representation, $_\text{prevent}$ indicates the use of the may_prevent property, $_\text{treat}$ the may_treat property and $_\text{CI}$ the CI_with property.

In addition to $+s$ and $+s*$, we considered several vector enrichments based on different bags of concepts extracted from DBpedia to study the selection of relevant concepts for knowledge enrichment. The notation is as follows:

- $+s*\text{T}$ refers to an enrichment with the labels of concepts automatically extracted from DBpedia with the help of the SPARQL query in Listing 4.2, 285 concepts are thus considered with this approach. This approach uses the same text fields as $+s*$ to identify entities from DBpedia.

- $+s*\cap$ refers to an enrichment with a subset of the labels of concepts automatically extracted from DBpedia acknowledged as relevant by at least one expert human annotator. This approach uses the same text fields as $+s*$ to identify entities from DBpedia.

- $+s*\cup$ refers to an enrichment with a subset of the labels of concepts automatically extracted from DBpedia acknowledged as relevant by all the experts human annotators. This approach uses the same text fields as $+s*$ to identify entities from DBpedia.

- $+s*m$ refers to an enrichment with a subset of the labels of concepts automatically selected by using a feature selection algorithm. We chose the Lasso algorithm TIBSHIRANI [1996] and we executed it *within* the inner loop of the nested cross-validation in the global machine learning algorithm chosen to predict hospitalization. For the Lasso algorithm, we chose the default parameters (and the number of folds used for cross-validating in that context, fixed at F = 3). This approach uses the same text fields as $+s*$ to identify entities from DBpedia.

- $+sm$ uses the same enrichment procedure of $+s*m$ to automatically select a subset of the labels of concepts. This approach exploits the same text fields as $+s$ (all text fields) to identify entities from DBpedia.

5.3 Evaluating the impact of ontological knowledge on prediction

Once again, in order to assess the value of ontological knowledge, we evaluated the performance of the machine learning algorithms by using the $F_{tp,fp}$ metric FORMAN et SCHOLZ [2010].

5.3.1 Evaluation of the enrichment with concepts extracted from knowledge graphs

We compared on the dataset DS_B the contribution of knowledge graphs on hospitalization prediction by considering the performance of the machine learning algorithms measured by the $F_{tp,fp}$ metric FORMAN et SCHOLZ [2010]. We considered the impact of knowledge from different sources, whether separately incorporated or combined, on the vector representation of patients' medical records to predict hospitalization.

Table 5.1 shows the values of $F_{tp,fp}$ for the chosen combinations. Despite the shallow OWL-SKOS representation of ICPC-2, the $+t$ configuration is sufficient to improve patient's hospitalization prediction, if we compare its results to those of the *baseline* (see Table 5.1). Surprisingly enough, a second level of superclass hierarchy with $+c_2$ from the ATC OWL-SKOS representation provides better results, while only one level of hierarchy with $+c_1$ seems to have a negative impact on the prediction of hospitalization. This may be explained by the fact that the introduction of a large number of attributes ultimately provides little information, unlike the second level of hierarchy.

Figure 5.1 shows the average F1 score (average between the different F1 scores obtained by cross-validation) and standard deviations associated to the vector sets considered in the Table 5.1. By comparing this figure with the above-mentioned table, it appears that, contrary to the trend shown in the table, there is no approach that performs better than another.

Table 5.2 and Table 5.3 show the confusion matrices for two machine learning algorithms: the comparison of performance on the *baseline* between random forest and the logistic regression algorithms is represented in Table 5.2. The improvements of adding $+s*$ to the combination of features $+t + c2 + wa + wi$ is displayed in Table 5.3, where we compared the results obtained with the logistic regression algorithm. We can see an im-

Table 5.1 – $F_{tp,fp}$ for the different vector sets considered on the balanced dataset DS_B.

Features set	SVC	RF	Log	Average
$baseline$	0.8270	**0.8533**	0.8491	0.8431
$+t$	0.8239	0.8522	**0.8545**	0.8435
$+s$	0.8221	0.8522	0.8485	0.8409
$+s*$	0.8339	0.8449	0.8514	0.8434
$+c_1$	0.8235	0.8433	0.8453	0.8245
$+c_{1-2}$	0.8254	0.8480	0.8510	0.8415
$+c_2$	0.8348	0.8522	0.8505	**0.8458**
$+d_{prevent}$	0.8254	0.8506	0.8479	0.8413
$+d_{treat}$	0.8338	0.8472	0.8481	0.8430
$+d_{CI}$	0.8281	0.8498	0.8460	0.8413
$+wa$	0.8223	0.8468	**0.8545**	0.8412
$+wi$	0.8149	0.8484	0.8501	0.8378
$+wm$	0.8221	0.8453	0.8458	0.8377
$+t+s+c2+wa+wi$	0.8258	0.8486	0.8547	0.8430
$+t+s*+c2+wa+wi$	0.8239	0.8494	0.8543	0.8425
$+t+c2+wa+wi$	0.8140	0.8531	**0.8571**	0.8414

Table 5.2 – Confusion matrix of the random forest algorithm (on the left) and the logistic regression (on the right) on the $baseline$ ('H' stands for Hospitalized and 'Not H' for 'Not Hospitalized').

	H	Not H
Predicted as 'H'	599	91
Predicted as 'Not H'	115	641

	H	Not H
Predicted as 'H'	588	83
Predicted as 'Not H'	126	649

provement in the number of true positives and false negative with features extracted from knowledge graphs. Thus, with knowledge graphs features the logistic regression outperforms random forest. Features from $+s*$ slightly penalize the results obtained, but this is a point that will be discussed using the results of the other figures.

As Table 5.4 shows it, the approach $+t+s+c2+wa+wi$ provides new information on the patient's file. For instance, a machine learning algorithm can identify with $+t+s+c2+wa+wi$ the use of antibiotics, whereas with the $baseline$ this information was implicit. The use of different names of antibiotics (i.e. Amoxicil, Cifloxan...) is considered by a machine learning algorithm as different features, which is certainly justified, but does not allow to apprehend the patient's file as a complete picture. More precisely, the generalization of antibiotics could be learned with more training data but can be avoided by the introduction of background knowledge to focus the learning on other aspects of the targeted prediction.
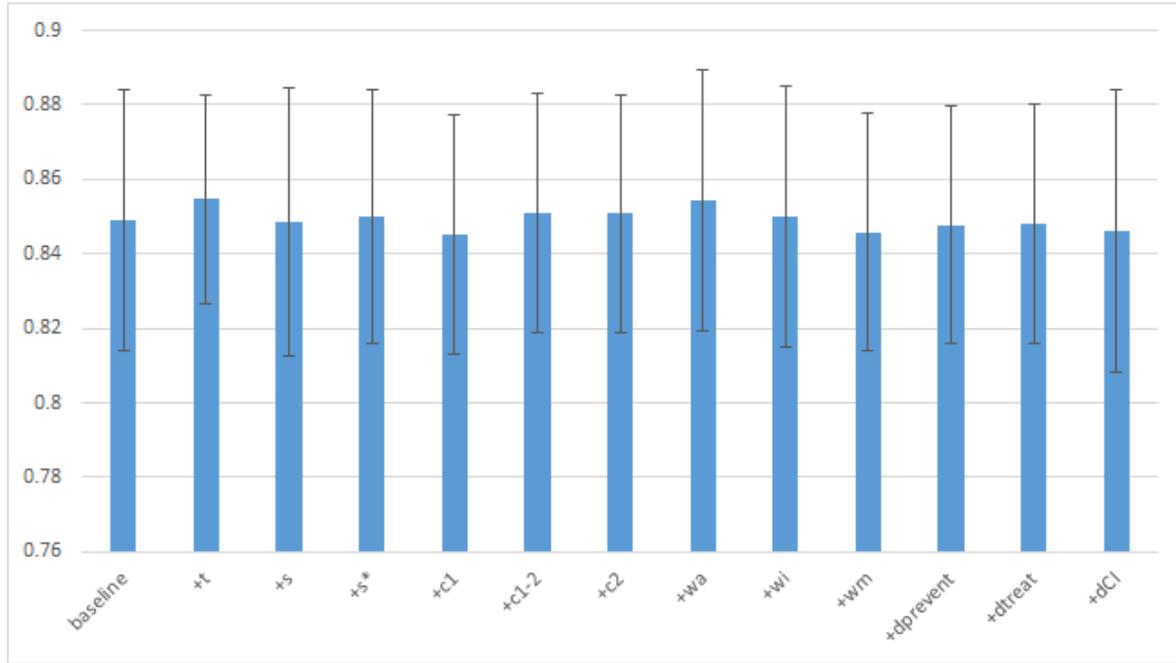
Figure 5.1 – Histograms that represent the average F1 score and standard deviations under logistic regression for the vector sets considered in the Table 5.1.

Table 5.3 – Confusion matrix of $+t+s*+c2+wa+wi$ (on the left) and $+t+c2+wa+wi$ (on the right) approaches under the logistic regression algorithm ('H' stands for Hospitalized and 'Not H' for 'Not Hospitalized').

	H	Not H
Predicted as 'H'	595	84
Predicted as 'Not H'	119	648

	H	Not H
Predicted as 'H'	597	82
Predicted as 'Not H'	117	650

Figure 5.2 displays the convergence curve with the $F_{tp,fp}$ measure on some configurations listed in Table 5.1. This curve shows the performance obtained with these configurations when training with less data. At first glance, in Figure 5.2, the approach in combination with $+s*$ does not achieve the best final results, it achieves the best overall performance among all the combined configurations tested with 0.858 under logistic regression when using 8 folds during the training phase. It also surpasses other combined methods under 3 folds partitions by exceeding the *baseline* by 0.9% and at 4 folds partitions by 0.7% $+t+s+c_2+wa+wi$ which suggests an improvement in classification results if we enrich a small dataset with attributes provided by knowledge graphs.

Figure 5.1 displays the average F1 score and standard deviations associated to different configurations of the Table 5.1. Similarly to Figure 5.1, Figure 5.3 points out that, contrary

Table 5.4 – Patient profiles correctly identified as being hospitalized (true positives) after injecting domain knowledge (the comparison of these two profiles was made on the baseline and the $+t+s+c2+wa+wi$ approaches with the logistic regression algorithm).

Patient profiles	Risk factors identified by knowledge graphs
Birth year: 1932 Gender: Female Without long-term condition 1 year of consultations before hospitalization No notes in the observations field	Usage of many antibacterial products noted by both ATC, and Wikidata (Amoxicil, Cifloxan, Orelox, Minocycline...) Different health problems affecting the digestive system noted by ICPC-2 (odynophagia 'D21', abdominal pain 'D06', vomiting 'D10')
Birth year: 1986 Gender: Male Without long-term condition 2 years of consultations before hospitalization	Within free text (contained in reasons of consultation and observations fields), daily chest pains are considered as 'Emergency' and a tongue tumor as 'Neoplasm stubs' by DBpedia

to the trend shown in the Table 5.1, there is no configuration that performs better than another.

5.3.2 Evaluation of the selection of concepts extracted from DBpedia for the enrichment of electronic medical records representations

Evaluation of manual vs. automatic selection of relevant subjects

Table 5.5 shows scores of $F_{tp,fp}$ obtained with different feature sets derived from DBpedia. These features were selected after human and machine annotations. The best results are those with the machine annotation approaches, i.e. $+s*m$ and $+sm$, and with the logistic regression algorithm. Also, the approach that uses the most fields, $+sm$, has the best results.

Table 5.5 – $F_{tp,fp}$ for the different vector sets considered on the balanced dataset DS_B.

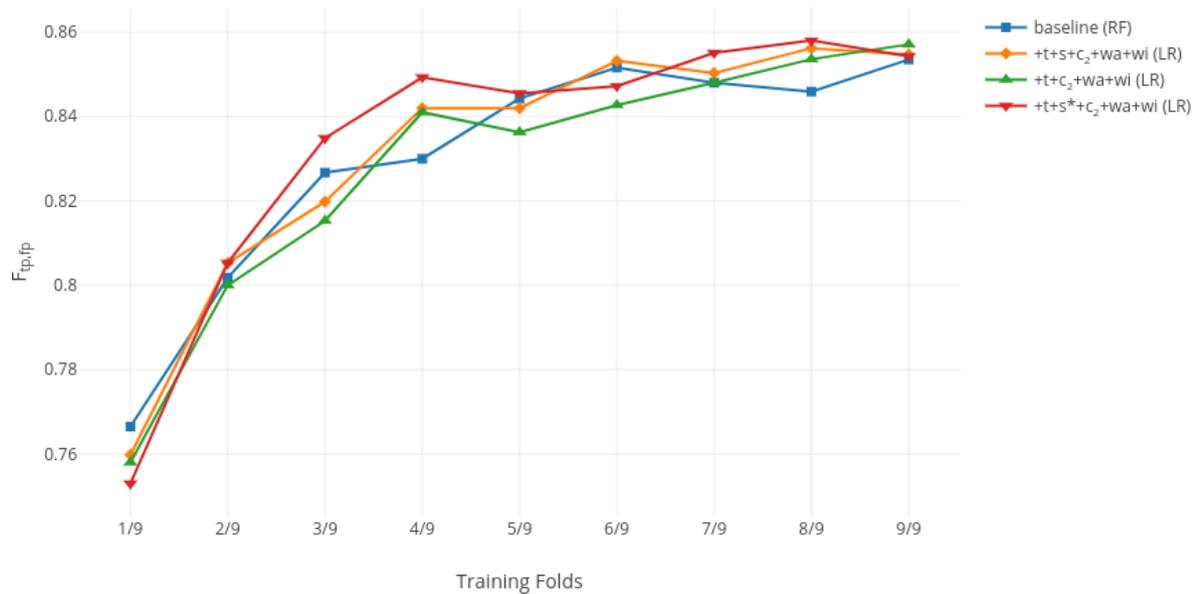| Features set	SVC	RF	Log	Average
$baseline$	0.8270	**0.8533**	0.8491	0.8431
$+s$	0.8221	0.8522	0.8485	0.8409
$+s*$	0.8339	0.8449	0.8514	0.8434
$+s*\mathrm{T}$	0.8214	0.8492	0.8388	0.8365
$+s*\cap$	0.8262	0.8521	0.8432	0.8405
$+s*\cup$	0.8270	0.8467	0.8445	0.8394
$+s*m$	0.8363	0.8547	**0.8642**	0.8517
$+sm$	0.8384	0.8541	**0.8689**	**0.8538**

Figure 5.2 – Convergence curve obtained following the training on n (x-axis) KFold partitions for different configurations of the Table 5.1.

Generalization of concepts vector

Following the list of concepts extracted for each fold with the $+sm$ approach, we evaluate the effect of a global vector of concepts (i.e. a same vector of concepts across all folds), since with our experimentation setup the features selected can be different from one fold to another. Thus, we generate different stable vector of concepts based on the number of intersections of concepts and union of concepts with the hyperparameters identified with the approach $+sm$.

Based on this observation, by using the logistic regression algorithm, the intersection of all the concepts gets a score of 0.8662 and the union of all concepts encountered for each fold obtains a score of 0.8714 which is better than the baseline (by more than 2%) and even better than the $+sm$ approach (score of 0.8689). Table 5.6 compares the confusion matrix obtained with the logistic regression algorithm on the $+sm$ configuration and the union of concepts present in each fold of the $+sm$ approach. This table indicates that the union of concepts increase the number of true positives, i.e. hospitalized patients correctly identified as such.

Figure 5.4 shows the average F1 score and standard deviations associated to the vector sets considered in the Table 5.5 and with the generalized concepts vectors approach described above. There are no significant differences between the different approaches except for the automated approaches for which there is a slight improvement in results.
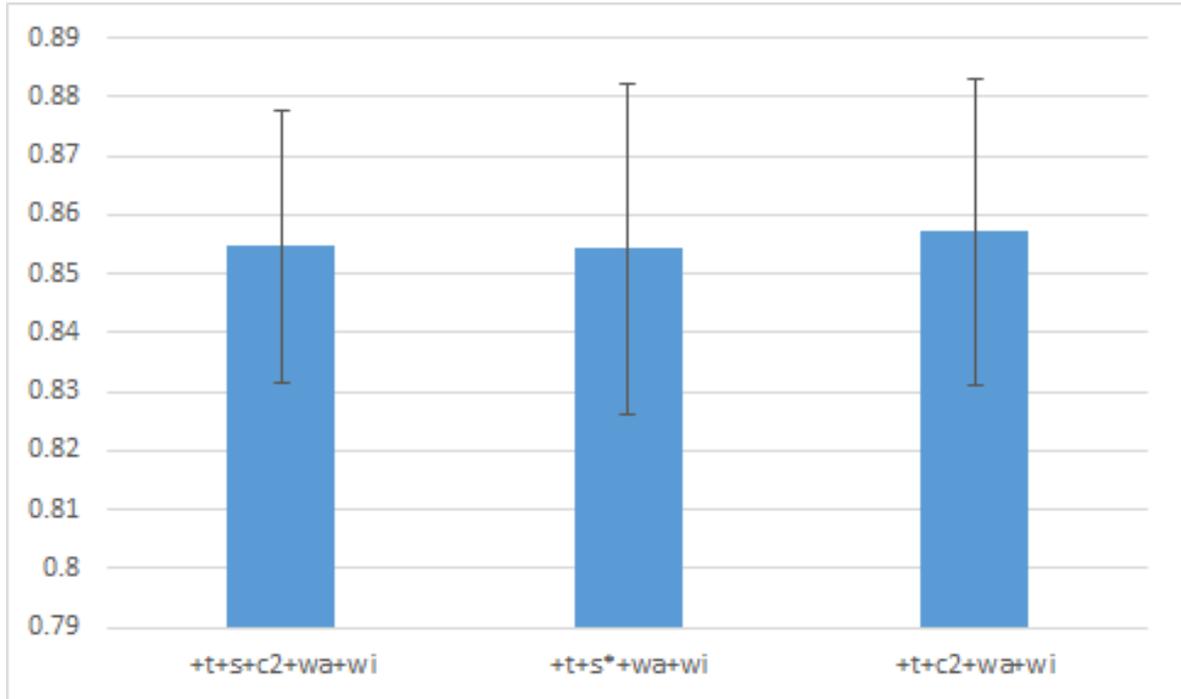
Figure 5.3 – Histograms that represent the average F1 score and standard deviations under logistic regression for different configurations of the Table 5.1.

Table 5.6 – Confusion matrix of $+sm$ (on the left) and the union of concepts under $+sm$ conditions (on the right) approaches under the logistic regression algorithm ('H' stands for Hospitalized and 'Not H' for 'Not Hospitalized').

	H	Not H
Predicted as 'H'	600	67
Predicted as 'Not H'	114	665

	H	Not H
Predicted as 'H'	603	67
Predicted as 'Not H'	111	665

5.3.3 Statistical hypothesis testing with concepts extracted from knowledge graphs

Different statistical tests exist in the literature DEMŠAR [2006] and we opted for the correction of dependent Student's t test NADEAU et BENGIO [2003] to test the null hypothesis against our vector sets. We used the dependent Student's t test because the training sets overlap in a cross-validation context, thus violating the independence assumption.

The formula for the corrected dependent Student's t test is as follows:

$$t = \frac{\frac{1}{n}\sum_{j=1}^{n} x_j}{\sqrt{(\frac{1}{n} + \frac{n_2}{n_1})\widehat{\sigma}^2}} \tag{5.1}$$
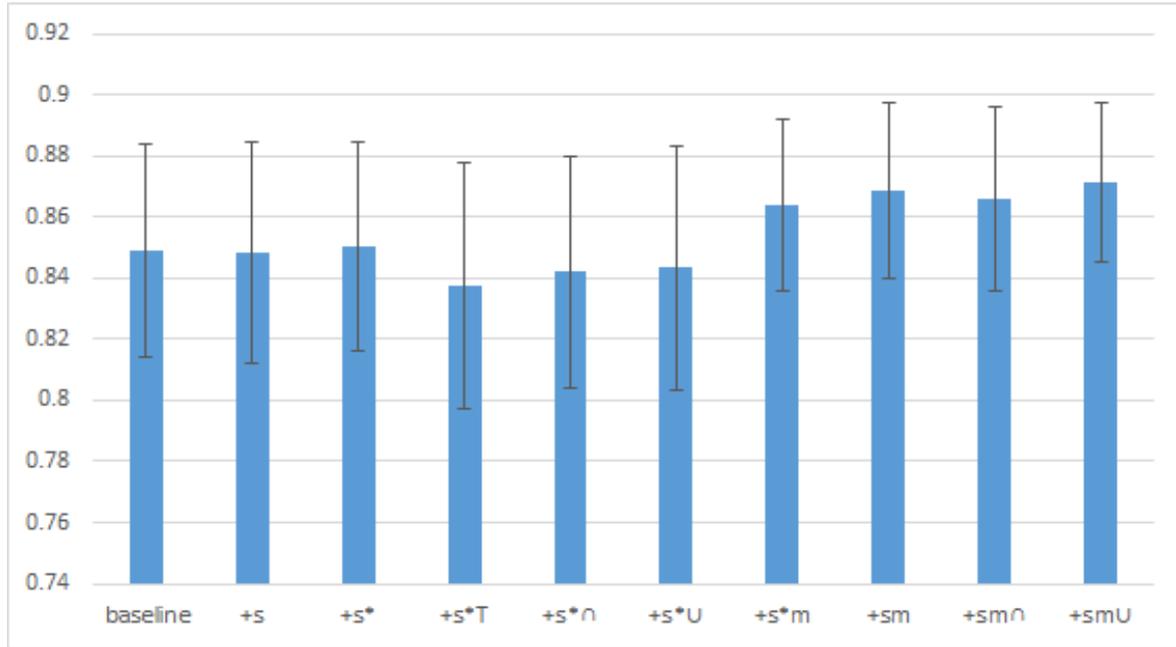
Figure 5.4 – Histograms that represent the average F1 score and standard deviations under logistic regression for the vector sets considered in the Table 5.5 and with the generalized concepts vectors approach described in Section 5.3.2.

Where x_j is equal to $A_j - B_j$, A and B are two sets of length n. n_2 is the number of testing folds, n_1 is the number of training folds and $\widehat{\sigma}^2$ represents the sample standard deviation on x.

Table 5.7 shows the t-value/p-value pairs obtained with the F1 and with the AUC obtained on each observation on different vector sets. The corrected Student's t test rejects the null hypothesis on the $+sm\cup$ approach, the approach that consists in using the union of concept of $+sm$, which has achieved the best results.

Table 5.7 – t-value/p-value pairs on F1 and on AUC for different vector sets considered on the balanced dataset DS_B.

Features set	t-value/p-value (on F1)	t-value/p-value (on AUC)
$+wa$	-1.06/0.32	0.11/0.92
$+t+s+c_2+wa+wi$	-0.47/0.65	0.02/0.98
$+t+s*+c_2+wa+wi$	-0.52/0.62	0.10/0.92
$+t+c_2+wa+wi$	-0.69/0.51	-0.16/0.87
$+sm$	-1.57/0.151	-0.77/0.46
$+sm\cap$	-1.62/0.139	-0.58/0.58
$+sm\cup$	**-2.23/0.05**	-0.81/0.44

5.4 Discussion

In general terms, knowledge graphs improve the detection of true positive cases (see Table
5.2 and Table 5.3). They provide a broader knowledge of the data present in patient files
like the type of health problem with ICPC-2 (see Table 5.4). We observe that using implicit
knowledge allows machine learning algorithms to better understand the content of EMRs.

Despite what may suggest the Figure 5.2, the automated selection approach $+sm$ implies that better results are obtained by considering all text fields when extracting concepts from DBpedia, and not only those related to the patient's own case, contrary to the
results suggested by $+s$ and $+s*$, as shown in the Table 5.5.

The best performing approach with knowledge extracted from DBpedia, $+sm$, selected a much smaller number of concepts with a feature selection process than those
selected by human annotation (approaches $+s*\cap$ and $+s*\cup$). This implies that the selected concepts are more precise to distinguish hospitalized patients from other ones (see
Table 5.6) by improving both the detection of true positives and true negatives, the union
of concepts seen in Section 5.3.2 also improves the number of true positives in comparison to the standard $+sm$ approach. That means that steps involving a feature selection
algorithm and the generalization of the concepts vector allow to retrieve the most relevant
concepts in a context where the training dataset is small and may help with annotation
procedures. Moreover, the corrected Student's t test rejects the null hypothesis on the
$+sm\cup$ approach.

Among the 51 concepts selected with the union of concepts, more generic knowledge was selected like 'Terme médical' (respectively 'Medical terminology'), one possibility could be that the general practitioner uses technical terms in a situation involving a
complex medical case. Numerous concepts related to patient's mental state (like 'Antidépresseur' -Antidepressant-, 'Dépression (psychiatrie)' -Major depressive disorder-, 'Psychopathologie' -Psychopathology-, 'Sémiologie psychiatrique' -Psychiatric assessment-,
'Trouble de l'humeur' -Mood disorder-) appear to be a cause of hospitalization. Different concepts related to the allergy ('Allergologie' -Allergology-, 'Maladie pulmonaire
d'origine allergique' -Lung disease of allergic origin-) and infectious diseases ('Infection
ORL' -ENT infection-, 'Infection urinaire' -Urinary tract infection-, 'Infection virale' -Viral
infection-, 'Virologie médicale' -Clinical virology-) were selected. Concepts related to
the cardiovascular system are widely represented within this set ('Dépistage et diagnos-

tic du système cardio-vasculaire' -Screening and diagnosis of the cardiovascular system-, 'Maladie cardio-vasculaire' -Cardiovascular disease-, 'Physiologie du système cardio-vasculaire' -Physiology of the cardiovascular system-, 'Signe clinique du système cardio-vasculaire' -Clinical sign of the cardiovascular system-, 'Trouble du rythme cardiaque' -Cardiac arrhythmia-). The unique concept retrieved in the family history of the patient at the exception of 'Medical Terminology' is 'Diabète' (respectively 'Diabetes'). Among the concepts selected by machine learning through feature selection, rare concepts considered irrelevant at first sight toward the problem of hospitalization such as 'Medical Terminology' could find an explanation. Also, a feature selection step helps to improve the prediction of hospitalization by adding knowledge indirectly related to the patient's condition, such as family history.

Although the number of concepts considered as relevant by experts is relatively high, 198 among 285 medical subjects, their integration into a vector representation reduced the performance obtained in comparison to the baseline, one of the possibilities for this result could be the limited size of our corpus.

Moreover, the qualitative analysis of the results indicates cases involving negation (e.g. 'pas de SC d'insuffisance cardiaque', meaning 'no symptom of heart failure') and poor consideration of several terms (e.g. 'brûlures mictionnelles', related to bladder infection, are associated with 'Brûlure', a burn, which, therefore, has as subject the concept 'Urgence médicale', a medical emergency). On this subject, Liu et al. [2018a]'s studies show improvements in congestive heart failure, kidney failure and stroke prediction by taking into account negation in medical reports, regardless of the algorithm used. Both cases are current limitations of our approach and we consider for our future work handling negation and complex expressions.

Another weakness of our enrichment method is that a knowledge base like DBpedia may be incomplete (incompleteness of properties `dcterms:subjects`, `owl:sameAs` and `rdf:type`). We may improve the results by curating the knowledge graph before extracting relevant concepts to represent EMRs.

The incompleteness of medical records implies a vast variety between patients. This degree of completion also varies from one consultation to another for the same patient. In the same line of thought, joint medical care provided by a fellow specialist with sometimes insufficient information about these cares is another negative factor for the degree of completion of EMRs. Moreover, the patient may not have been detected as being par-

ticularly at risk or may not be very observant and does not come frequently to consultations, this shows the interest of being able to work on patient trajectories and to set up a health data warehouse combining several sources.

Reports of the consultations contain abbreviations of experts and thus it would lead to notable improvements in the extraction of knowledge to be able to distinguish abbreviations with their meanings in a given medical context. We plan to detect negation and experiencer (the patient or members of his family) in future work since a pathology affecting a patient's relationship or the negation of a pathology does not carry the same meaning when it comes to predict a patient's hospitalization.

5.5 Conclusion

We generated different vector representations coupling concept vectors and bag-of-words and evaluated their performance for predicting hospitalization with different machine learning algorithms.

Deciding of the relevancy of some given concepts for a specific prediction task appeared to be quite difficult and subjective for human experts, with a significant variability in their annotations. To overcome this problem, we integrated an automatic step allowing annotators to confirm their thoughts on the case of DBpedia. This automated process to select concepts can be extended to other knowledge graphs to further improve our results.

The results of our work will be used in the development of a decision-support tool for physicians that we will present in the next chapter. The purpose of our tool is to define the risk factors to be treated as a priority for his patient in order to avoid his hospitalization and, if not, to improve his health condition.

Chapter 6

Decision support application

In the previous chapter, we demonstrated that features derived from knowledge graphs improve the efficiency of the prediction of hospitalization when added to the vector representation of electronic medical records (EMRs). In a last stage, and in order to propose a decision-making tool to help general practitioners (GPs), it is important to design an interface that meets their expectations and efficiently conveys the results we obtained. In particular, the binary prediction of the hospitalization of patients does not deliver any significant added value to physicians unless we are able to identify the factors on which they can act to prevent this outcome. Therefore, our goal in this chapter is to design interactions that provide predictions together with an intelligent synthesis of the patient's file and its features impacting the prediction.

The work we report here, led to the design of the interface of the decision-making tool HealthPredict, and it has been presented in GAZZOTTI et al. [2019b]. We designed our interface with Sketch[1] and Photoshop,[2] and we realized an interactive mockup with InVision.[3] In addition to the written description of the interface presented in this chapter, a video showing the features of HealthPredict is available via the link https://www.youtube.com/watch?v=3DoMn5KdpNk.

We will first state the requirements we defined through a focus group in Section 6.1 and then present other relevant medical applications we identified in Section 6.2. Then we introduce in Section 6.3 the scenarios we have considered for our decision support application to prevent hospitalizations. The Section 6.4 details how we exploit the predictive algorithm to order health problems and simulate the result of actions taken to

---

[1] https://www.sketch.com/
[2] https://www.photoshop.com/
[3] https://www.invisionapp.com/

address them. Section 6.5 presents an overview of the design of our interface and Section 6.6 shows the evolution of our interface and some perspectives. Then we will conclude this chapter in Section 6.7.

6.1 Specifying requirements with a focus group

During a focus group with a panel of 10 physicians, we defined different objectives to be achieved for the interface and the decision-making tool to meet their needs. These objectives were grouped in terms of priority into three sets corresponding to three terms: short term, medium term and long term needs.

Concerning the requirements with a maximum priority we identified the following ones:

- physicians indicated the need for some reliability indicators and for information related to the algorithm used and the quality of the predictions,

- they wanted an explicit way to display the diminution of the hospitalization risk,

- they wanted the screen dedicated to the patient to be as less anxiety-provoking as possible while remaining of course informative (therapeutic compliance),

- they wanted to automatically detect the outliers in biological analyses (this is particularly difficult to consider since it can be complex to distinguish a marginal value from an outlier),

- they needed explanations on the role of other risk factors and their participation in the evaluation of hospitalization risk.

Concerning the requirements for a medium term, we identified the following ones:

- propose to the general practitioner (GP) options to predict the hospitalization on different time scales (for instance up to 1, 3, 5 years),

- ability to take into consideration the socio-economic impacts in the algorithm with data such as the patient's health care system (CMU/AME),

- ability to be more restrictive on the dataset used to train the algorithm in order to exclude less relevant hospitalization cases (check-ups, emergency visits, ...),

- display the report of the analysis on some specific events present in the dataset such as the proportion of scheduled check-ups at the hospital,

- display evidence-based medicine (EBM) factors, such as smoking, even if they have a low weight in the prediction outcome.

Concerning the requirements for the long term, we identified the following ones:

- perform an evaluation of the impact of our application on hospitalizations and on morbidity/mortality,

- display a report on the physician's practice compared to his colleagues, because physicians have habits when treating some cases due to a lack of knowledge or because it is their specialty,

- evaluation of drugs' effect in order to propose the best options for treatment choice,

- evaluation of the impact of prescriptions for home care services (IDE -Diploma in Nursing-...),

- prediction of risks related to pathologies (cardiovascular risk, cancer risk...).

Before we focus on the scenario we implemented, the next section provides examples of related applications.

6.2 Related work and existing applications

Systems developed around the prediction of hospitalization are restricted to the prediction of 30 to 60 days re-admissions. Therefore these applications are of more interest to hospitals than to GPs responsible for general practice. Most of them also are not inclined to provide a simulation but only the current prediction result.

Health Catalyst,[4] is one of these systems. It provides feedback on the hospital population and is able to predict the re-admission of a patient (see Figure 6.1, and Figure 6.2 for a former version of the interface). However all the information displayed are not of equal importance. Displaying information in the standards does not provide added value and on the contrary distracts from other information that may be of interest, e.g. no history of dementia, depression, psychose.

---

[4] https://www.healthcatalyst.com/

Other systems are limited to the prediction of specific diseases or interventions and use a restricted number of parameters present in a patient's file. For instance, the tool developed by KHERA et al. [2019] is limited to 30-days re-admission prediction after transcatheter aortic valve replacement (see Figure 6.3).
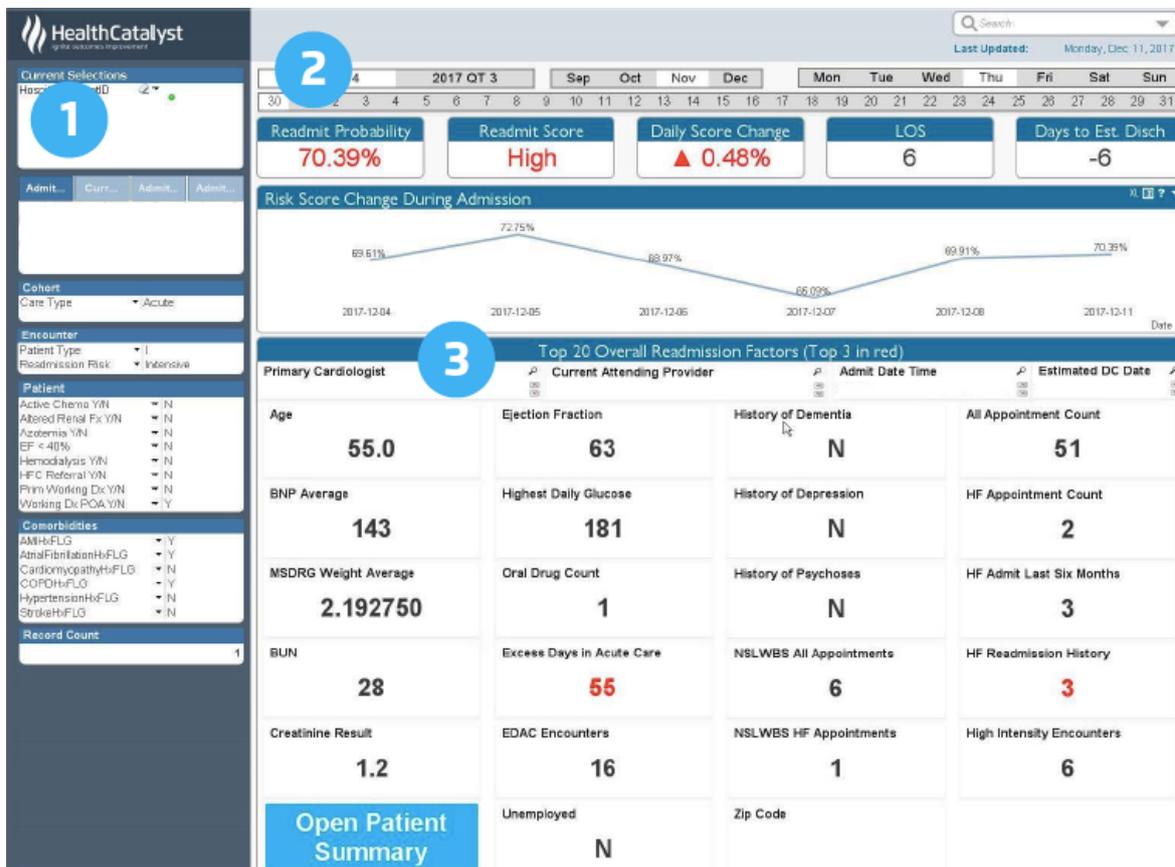

Figure 6.1 – Illustration of the 30-day re-admission prediction panel from the software developed by Health Catalyst. The area (1) allows to select a population of interest. The area (2) displays scores and probability related to the re-admission of the patient. The area (3) displays the top re-admission factors. Source: `https://bit.ly/2NwqSaF`.

In FLACH et al. [2018], the authors developed a tool[5] to predict cardiovascular risks with a simulator. Although it is true that the majority of EMRs system focus on text and dialog boxes, this tool is not intended to be directly integrated into an EMR system and therefore requires a physician to re-enter all information concerning his patient, thus wasting a considerable amount of time. The question of the choice of colours is also debatable for reasons of accessibility in particular for peoples with colour blindness. Figure 6.5 shows an overview of its interface.

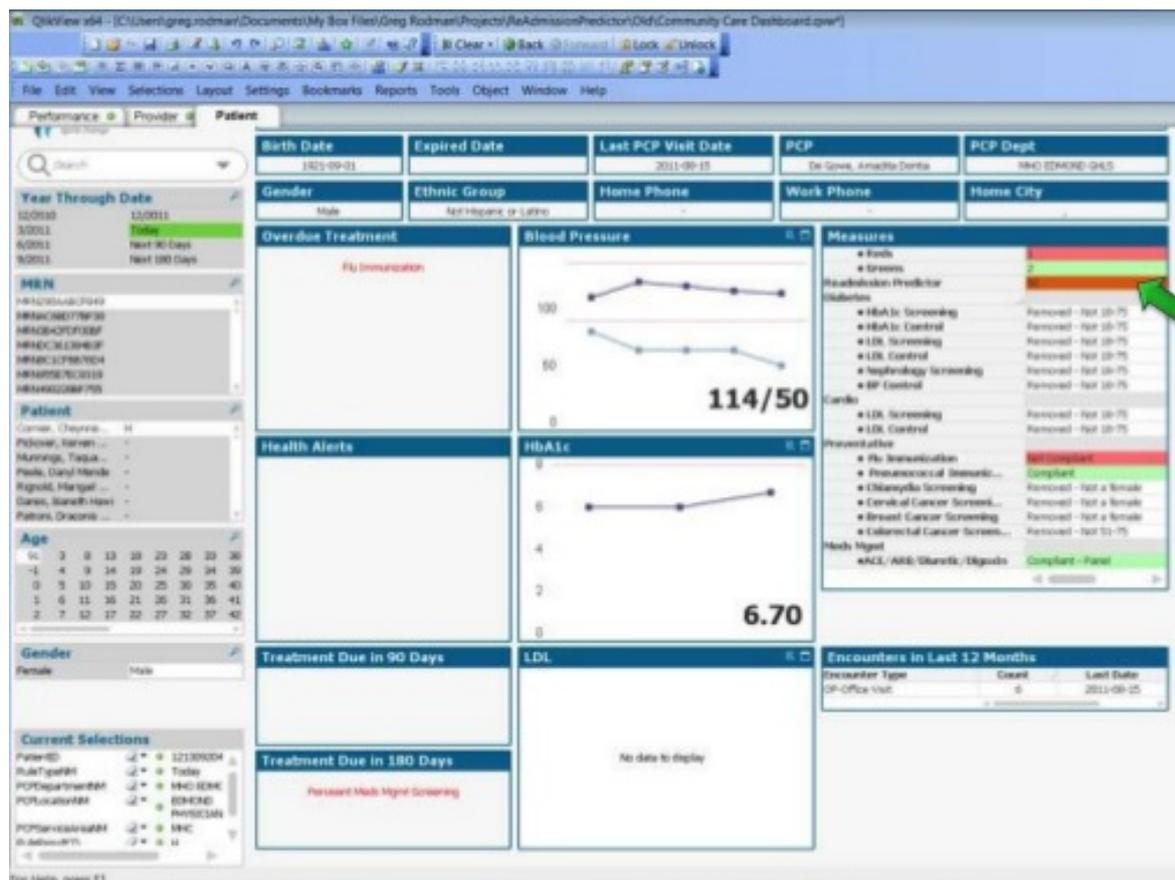Other tools exist, including CardioRisk[6] to determine cardiovascular risk based on the

---

[5]`https://mile-two.gitlab.io/CVDI/`
[6]`http://www.cardiorisk.fr/`

Figure 6.2 – Former interface of the 30-day re-admission prediction panel from the software developed by Health Catalyst. Source: https://bit.ly/2TwJNFV.

works of D'AGOSTINO et al. [2008] and CONROY et al. [2003] where they used Weibull and Cox regression models on a selection of risk factors. The interface of CardioRisk is much sober than the one of FLACH et al. [2018] and is limited to a small number of parameters, this software also proposes recommendations to improve the patient's health condition (see Figure 6.4).

There is a real expectation from the physicians to have the means to interpret and understand the results of machine learning algorithms they are provided with. For instance a deep learning algorithm was efficiently able to predict mortality based on electrocardiograms (EGCs),[7] however cardiologists have not been able in general to identify in EGCs abnormal signals for patients who have been classified as dying by the algorithm. This therefore confirms the importance of providing explanations so that physicians can appropriate the results of the analyses and take decisions and measures to prevent this kind of event from occurring.

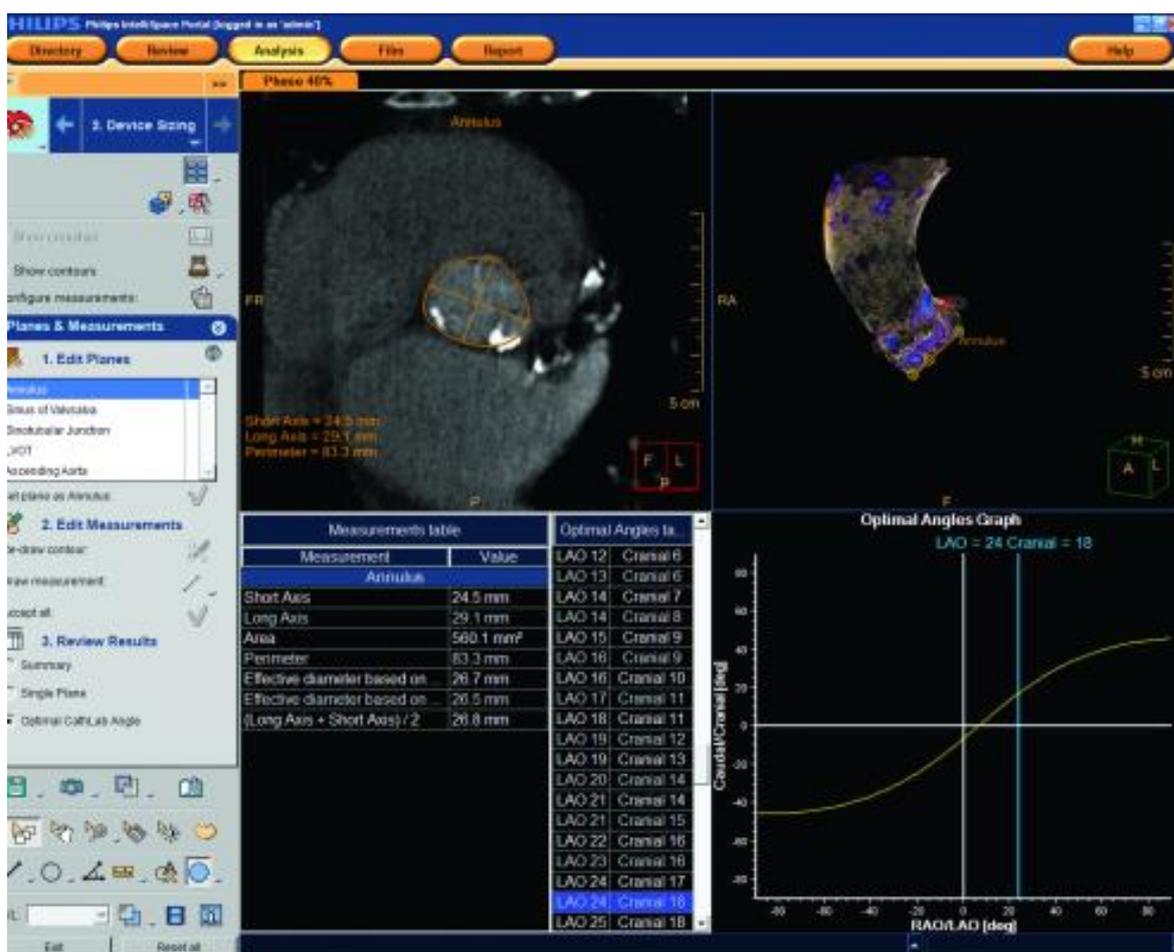---

[7]https://bit.ly/2NeJw6v

Figure 6.3 – Illustration of the re-admission prediction tool for patients undergoing transcatheter aortic valve replacement developed by KHERA et al. [2019]. Source: https://bit.ly/3075UnC.
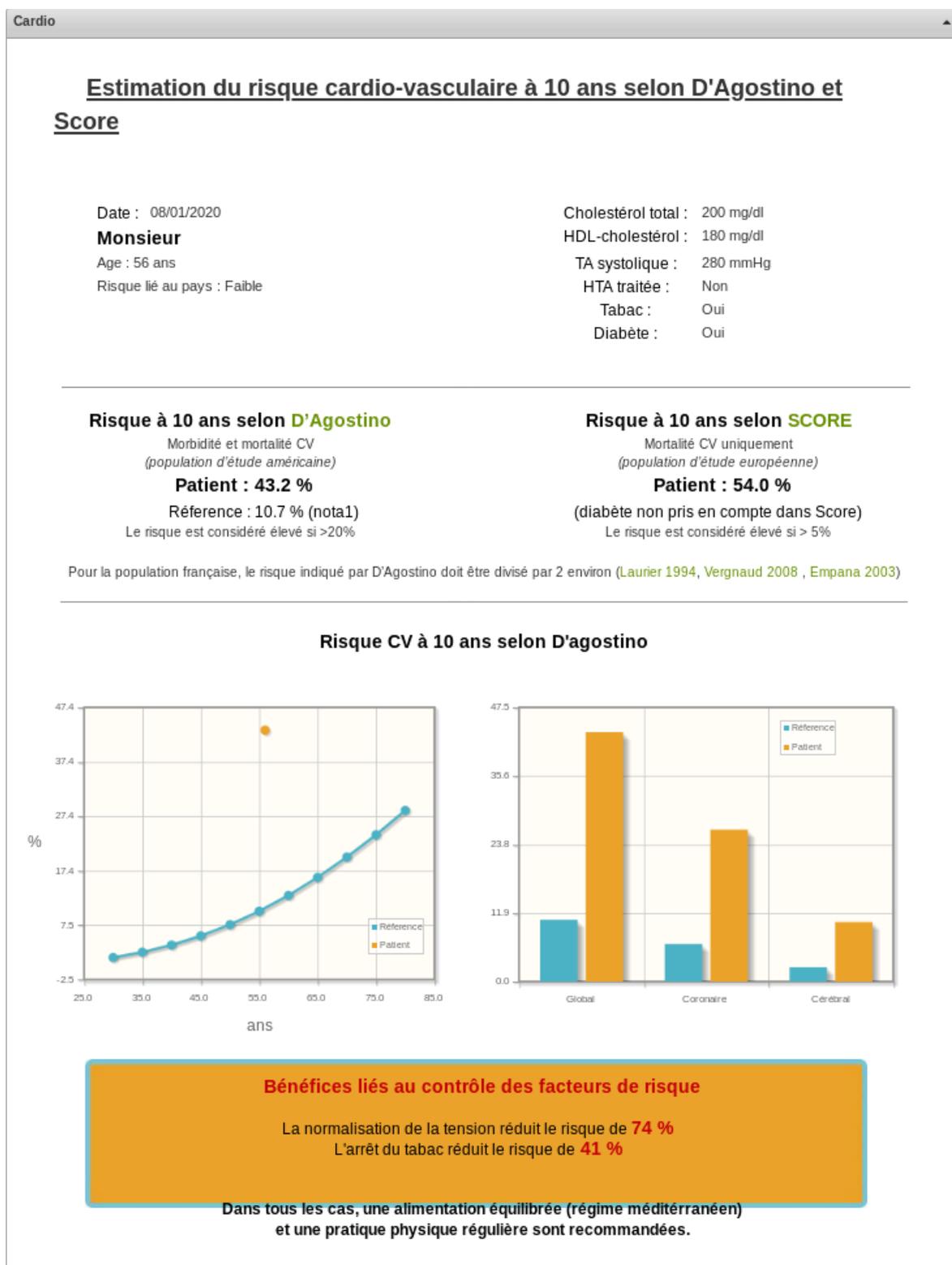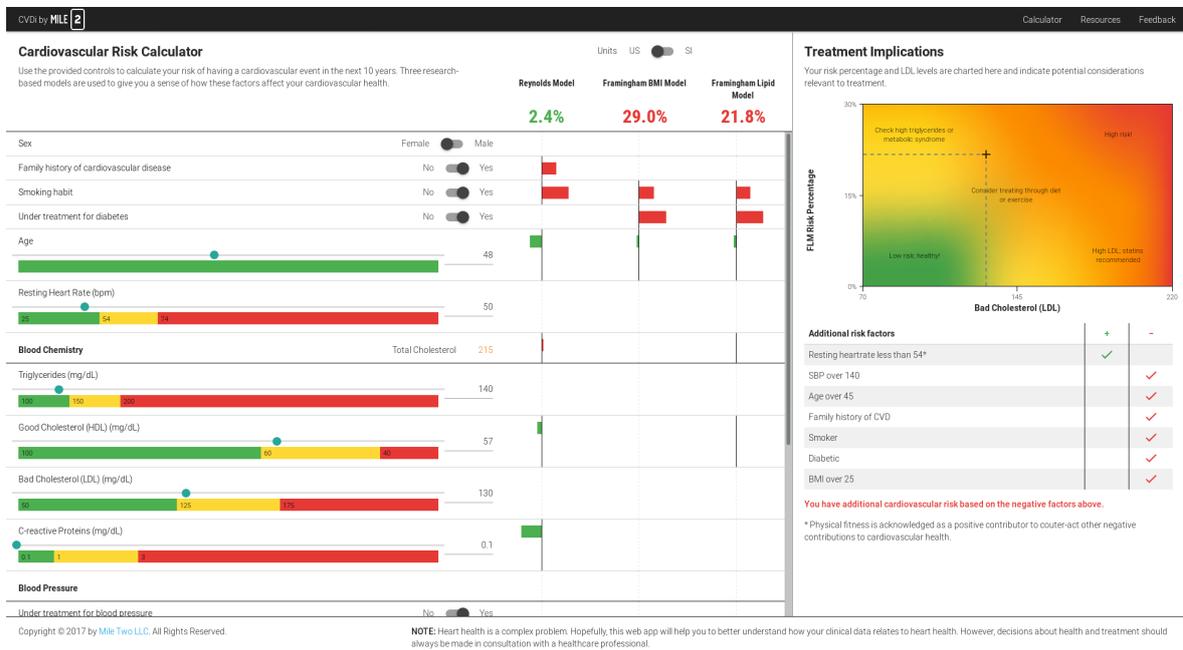
Figure 6.4 – Illustration of the CardioRisk tool. Source: http://www.cardiorisk.fr/.

Figure 6.5 – Illustration of the CDVI software, Cardiovascular Risk Calculator. Source: `https://mile-two.gitlab.io/CVDI/`.

6.3   Application scenarios

The interface of this project has been designed in order to present the outcome of the prediction of hospitalization to different categories of users.

Two different scenarios were considered in the use of the interface, one that applies to the GP and the other one which focuses on interactions with the patient in order to provide therapeutic education. Although they are not our only target population, poly-pathological patients are the first targeted audience of our application since it is complex to identify for them the actions to be taken and their priority, in order to improve their health condition.

6.3.1   General practitioner's perspective

During a consultation, a GP wants to know the hospitalization risk and the factors involved in the assessment of the hospitalization risk for this patient as soon as his consultation software indicates an alert or on demand when he is concerned about the health condition of his patient.

From his consultation software he can access his HealthPredict plugin that informs him about the current hospitalization risk of the patient. From there, it is possible for him to see the factors on which he can act in order to reduce the hospitalization risk of his patient.

He can also check the other symptoms of his patient to plan preventive actions. The software allows him to verify that a treatment is not harmful for a specific patient. If he is interested, he can also get some feedback on the records of the patient with his medical history, his abnormal medical tests, his family history and his allergies.

6.3.2   Patient's perspective

The GP wishing to rally the patient to his stance shows his monitor to his patient and indicates to him the potential gains of choosing different preventive actions to avoid hospitalization. He shows the different outcomes through simulations. This allows the patient to discuss with the GP about what it is possible for him to do such as: changes in eating habits, physical exercises, undergoing treatments, etc. This, in turn, improves his therapeutic compliance by relying on the report announced by the algorithm. However, this scenario implies that the GP wishes to indicate this outcome and the different results to

his patient. At his discretion, he can even show him the GP's display, if he feels his patient is psychologically strong enough to handle it.

6.4 Specificities of the application from the perspective of the predictive algorithm

6.4.1 Ordering of health problems

In a supervised scenario, after training, a machine learning algorithm is said to be 'interpretable' if it is able to provide the coefficients of its features that contribute to its decisions, i.e., ability to track down the features' weights involved in its decisions.

In the context of a binary classification, a positive coefficient for an attribute means that this attribute helps in the classification of class 1, a negative coefficient implies that an attribute participate in the prediction of class 0, while a coefficient close to 0 provides little or no information for a given classification task. In our case, the so-called class '1' corresponds to determining that a 'patient will be hospitalized' and the class '0' that a 'patient will not be hospitalized'.

A text representation using bag-of-words model contains as a value for these features either the occurrence of the words, or a computed value derived from the frequency of words such as the TF*IDF. Thus, for a new EMR provided as input to the machine learning algorithm, it is possible to isolate the features of interest since the vector representing the patient's record will have a value different from 0 for these features. This matrix is called 'sparse' because most values in such matrices are equal to 0, the features having a value different from 0 represent the information encoded for a given patient's medical record.

We performed the following work on the text fields related to the reasons for consultations, the patient's personal history, diagnoses, ongoing problems, allergies, environmental factors, and reasons for prescribing a drug. The observation's field being too diverse, we have not yet taken it into account in the calculation or the priority of the health problems to be treated because the task of parsing correctly text information complicates the operation.

For instance, the personal history 'Diabète de type 2' (meaning 'Type 2 diabetes') will be transformed into {'diabète':1, 'type':1, '@card@':1} (note the value at the right represents the number of occurrences and TreeTagger turns a number into '@card').

Table 6.1 – Coefficients learned on the expression 'Diabète de type 2' after training logistic regression on the prediction of hospitalization, the '#history#' means that the source of this expression comes from the personal history of the patient.

Feature	Coefficient
('#history#', 'diabète')	0.315144627498
('#history#', '@card@')	0.30715497576
('#history#', 'type')	0.158673598117

Thus, if we add the coefficients of the terms of an expression, we are able to determine with the sum of the coefficients the relative importance given by the machine learning algorithm to an expression. One other way to consider the relative importance of an expression would be to compute the arithmetic mean of an expression but this would have the effect of reducing the impact of an expression with several terms compared to an expression with only one term. However, the results obtained by this second methodology would not be accurate since it is not representative of how the coefficients are handled by logistic regression. Also, features with negative coefficients are kept in the sum performed for an expression, this results in the decrease or even in a negative sum of coefficients for some expressions. In our case, the presence of this 'negative expression' will mean that it does not contribute to the hospitalization of a patient.

According to Table 6.1, a coefficient of 0.781 can be assigned to the expression 'Diabète de type 2', this global coefficient allows to order expressions from the patient's record which are the most significant to predict the hospitalization's risk of a patient.

Regarding biological analyses, no additional work is required since a single coefficient is directly assigned to one analysis. Although the addition of biological analyses improves predictions of hospitalization, we did not go into detail about this point in the previous chapters since they are subject to a bias that deserves further study. Indeed, it is more common not to have a biological analysis than to have one in the standards, which has the side effect of associating the absence of analysis as better than having a bioassay in the standards.

Table 6.2 – Coefficients learned on the expression 'Absence de tabagisme' after training logistic regression on the prediction of hospitalization, the '#history#' means that the source of this expression comes from the personal history of the patient.

Feature	Coefficient
('#history#', 'absence')	-0.198578159717
('#history#', 'tabagisme')	0.141204157934

For instance, the personal history 'Absence de tabagisme' (meaning 'No smoking') will be transformed into {'absence':1, 'tabagisme':1} (note the value on the right represents the number of occurrences). According to Table 6.2, a coefficient of -0.0574 can be assigned to the expression 'Absence de tabagisme', which means that this expression found in a patient's file is significant in order to prevent hospitalization. By applying this method, we evaluate the importance of an expression according to the studied problem. Another way to proceed would have been to use better features (use of a chunker, features from a knowledge graph...), however, the two methods are not diametrically opposed and can be applied together, this is a perspective that we will consider in future work.

Thus, the ordering of health problems is made possible with the coefficients obtained after training a machine learning algorithm, in the case of a health problem we add the coefficients related to an expression (principle of logistic regression). It should be cautioned that an expression computed as negative may however become positive from one training to another, since it depends on the optimization performed by the machine learning algorithm on its features' weights. The injection of knowledge from knowledge graphs may negatively impact the results since we do not have any detection of negation at the moment, so it is crucial to take care of the features used in a vector representation. In addition, there is still room for improvement with the weight to assign on an expression, especially in the case where an expression is defined several times in the patient's file.

6.4.2 Simulation of the hospitalization prevention

We implemented in our solution the ability to remove a given medical problem from a patient's EMR in order to simulate the management of a health problem and thus to predict the patient's outcome. In order to simulate the management of a targeted health problem, we remove all its occurrences. That consists in assigning a zero value to all the features used to form a given expression. However, for bioassays, we must ensure that the new assigned values are inside the standards. It gives a global idea of the final decision to hospitalize a patient after taking into account one or more factors on which a GP can intervene.

This is a key feature expected by GPs since it allows them to plan preventive actions to avoid the hospitalization of a patient or at least to improve his health condition. Thus, the physician can have some feedback on the patient file, despite the fact that there may not be any recommendations associated to the patient's diseases from the HAS, la 'Haute

Autorité de Santé'[8] or they may be fragmented (cf. the management of polypathological patients discussed in Chapter 1).

6.5   Design of the interface

The major difficulty was to develop a user-friendly interface which provides all the information required by the GP to assist him in his decision-making process and to allow him to waste as little time as possible. In ASH et al. [2004], the authors reported that over-structured data cause physicians to lose attention. Moreover, having to navigate between too many screens in their patient care information system disrupts them and prevents them from identifying emerging health issues.

There are two different versions of the interface to achieve different goals: one must serve as a demonstrator to show the capabilities of the product and the other one must serve in real condition to assist the GP in his decision-making process to treat his patient. The main difference between the two versions of the interface remains the possibility to select a patient among a group of patients (see Figure 6.6) and to add a new bioassay or pathology associated to him (see Figure 6.8), with the objective for a GP to relate this specific case to a patient of his acquaintance.

The factors responsible for the prediction of patients' hospitalization are defined under two categories: factors strongly involved in predicting a patient's hospitalization and on which the GP can intervene (respectively 'Facteurs de risque modifiables') which are displayed in an inverted pyramid, and factors that have a lesser impact or on which the physician cannot intervene (respectively 'Autres facteurs', see Figure 6.9), all these factors are sorted out according to their importance in the decision made by the algorithm to hospitalize a patient. Displaying factors with less significant impact may allow the GP to take into account the different criteria interacting in a patient's file, in fact, the choice of a treatment may be contraindicated for certain pathologies or physical conditions. On the right-hand side are represented the percentages evaluated by the machine learning algorithm before the correction of a risk factor and the new percentage after the management of a given health problem (see Figure 6.7). On the top of each window, more generic information about the patient are included in a section with the gender, the age, if the patient has long-term conditions, a state medical assistance (AME)[9] or the complementary uni-

---

[8] https://www.has-sante.fr/
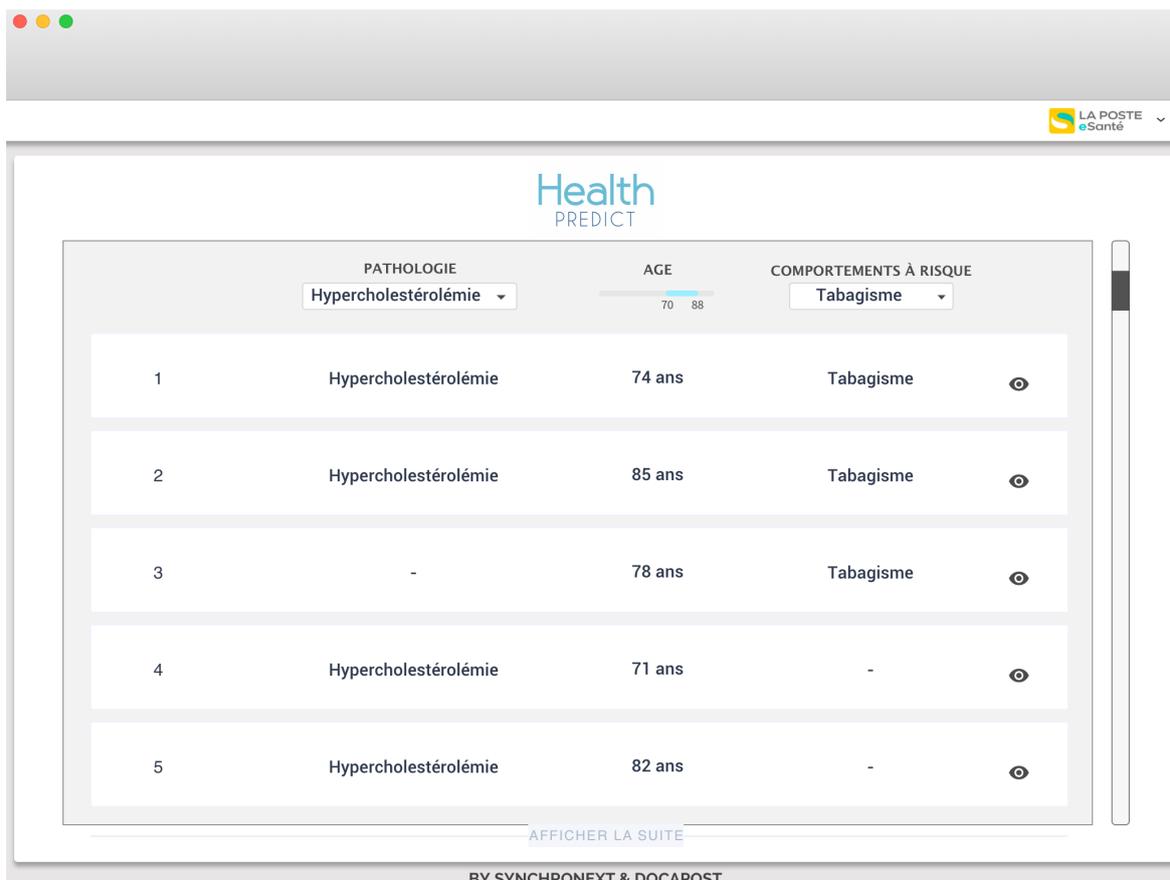[9] https://www.service-public.fr/particuliers/vosdroits/F3079

Figure 6.6 – View on the selection panel; this screen exists only in the demonstrator, which allows to select a patient according to different criteria. The patients searched here have hypercholesterolemia, aged between 70 and 88 years and are smokers.

versal health coverage (CMU-C),[10] concerning the patient in the example no long-term condition, AME or CMU-C are registered.

On the global overview of the patient's file, the risk factors are classified under different tabs such as 'History' (respectively 'Antécédents'), 'Biological examinations' (respectively 'Examens biologiques'), 'Risky behaviors' (respectively 'Comportements à risque'), other information completes the summary section on the patient's file with 'Allergies / Intolerances' (respectively 'Allergies / Intolérances') and his 'Family history' (respectively 'Antécédents familiaux'). On this component (see Figure 6.8), the factors determined as modifiable by the physician are represented with a pictogram symbolizing a magnifying glass, and only abnormal biological analyses are displayed under the biological examinations tab. Only the last biological analyses outside the standards are displayed in the panel corresponding to bioassays, we used for this purpose the reference ranges from different sources considering, when relevant, the age and the sex of the patient to apply the appro-

---
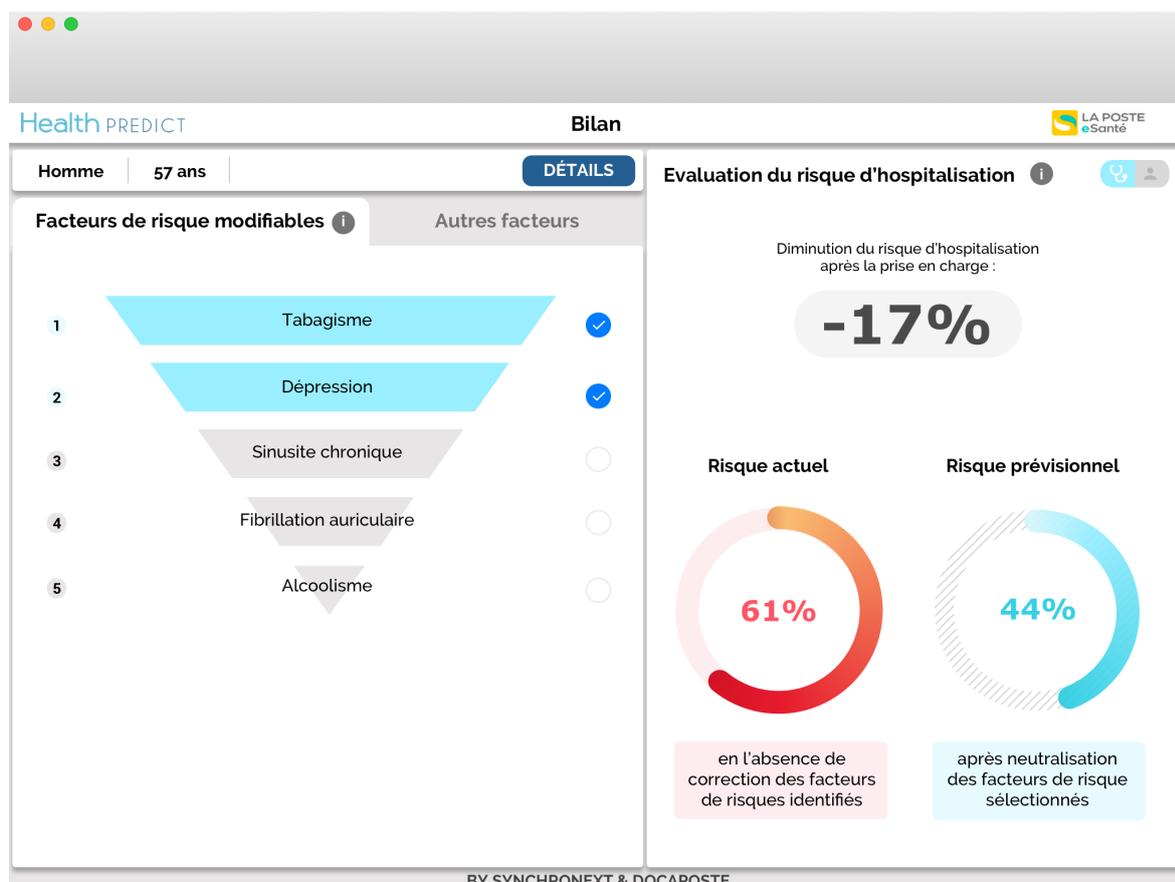
[10]https://www.service-public.fr/particuliers/vosdroits/F10027

Figure 6.7 – GP view with the expected hospitalization risk after management of the 'Smoking' and 'Depression' factors.

priate ranges and converting when necessary to the right unit. Some discretized values used for biological analyses can be found in the appendix Table A.3, for this purpose we used the CBC standards ranges from Mercy North Iowa,[11] standards ranges from VIDAL[12] and from the Lièges Teaching Hospital.[13]

Among all the points raised during our focus group with GPs, it was reported that showing up a less anxiety-provoking screen will be helpful to rally the patient to the GP's point. This is the so-called 'therapeutic compliance'. Thus, we responded to this request with a screen displaying only the total gain on the hospitalization risk in a half dial. The gain is calculated as a relative percentage (see Figure 6.10). A button on the upper right (icon representing a stethoscope and a patient) makes it easy to switch from the physician's to the patient's view.

The simulation aspect of this product (removal of factors involved in the prediction

---

[11] http://www.mercynorthiowa.com/cbc-normal-ranges
[12] https://web.archive.org/web/20150921080317/http://www.vidal.fr:80/infos-pratiques/id10442.htm
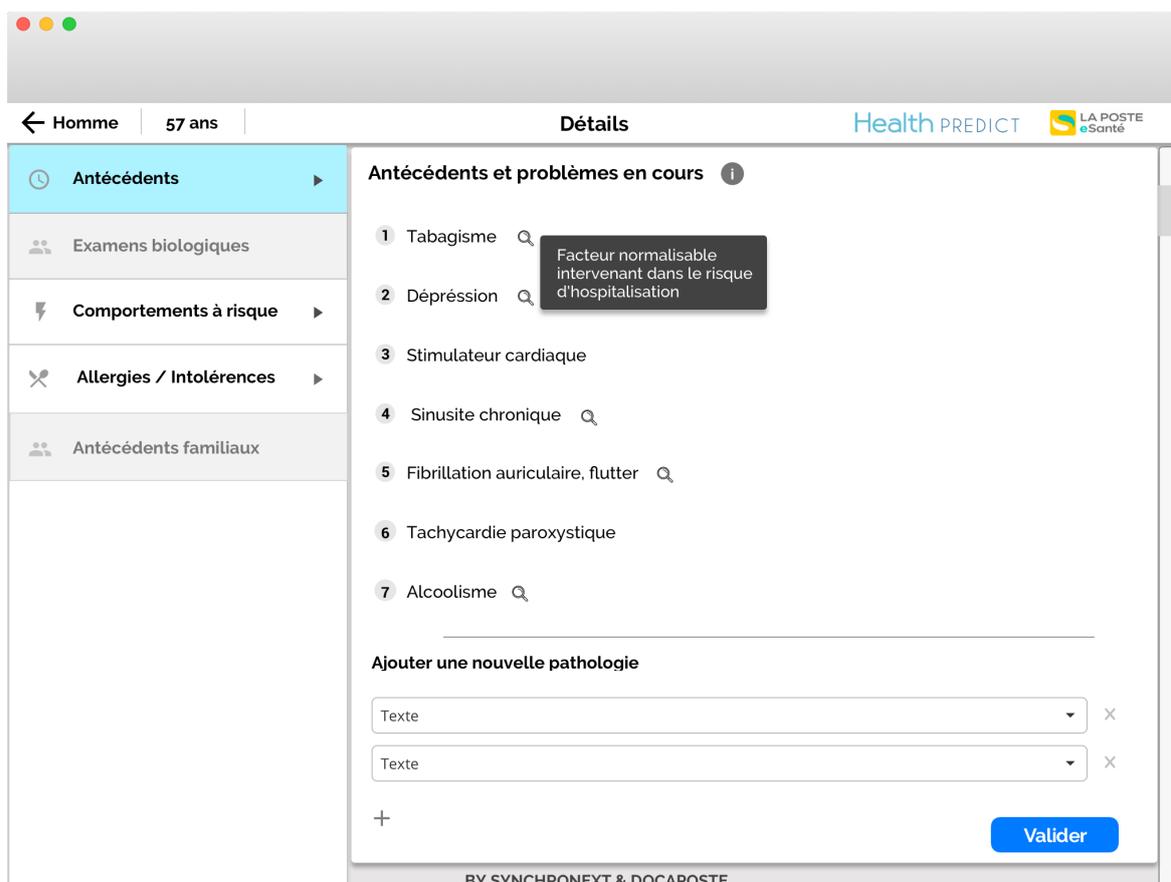[13] https://www.chu.ulg.ac.be/jcms/c_353640/

Figure 6.8 – Global overview panel on the patient's file, 'Details', under the tab 'History'. The demonstrator allows to add new pathologies.

risk) is due to the elimination of an expression or by normalizing a biological value as explained in Section 6.4.2.

Figure 6.9 – View on the tab that refers to lesser impact factors with details about methods and metrics involved in the software.
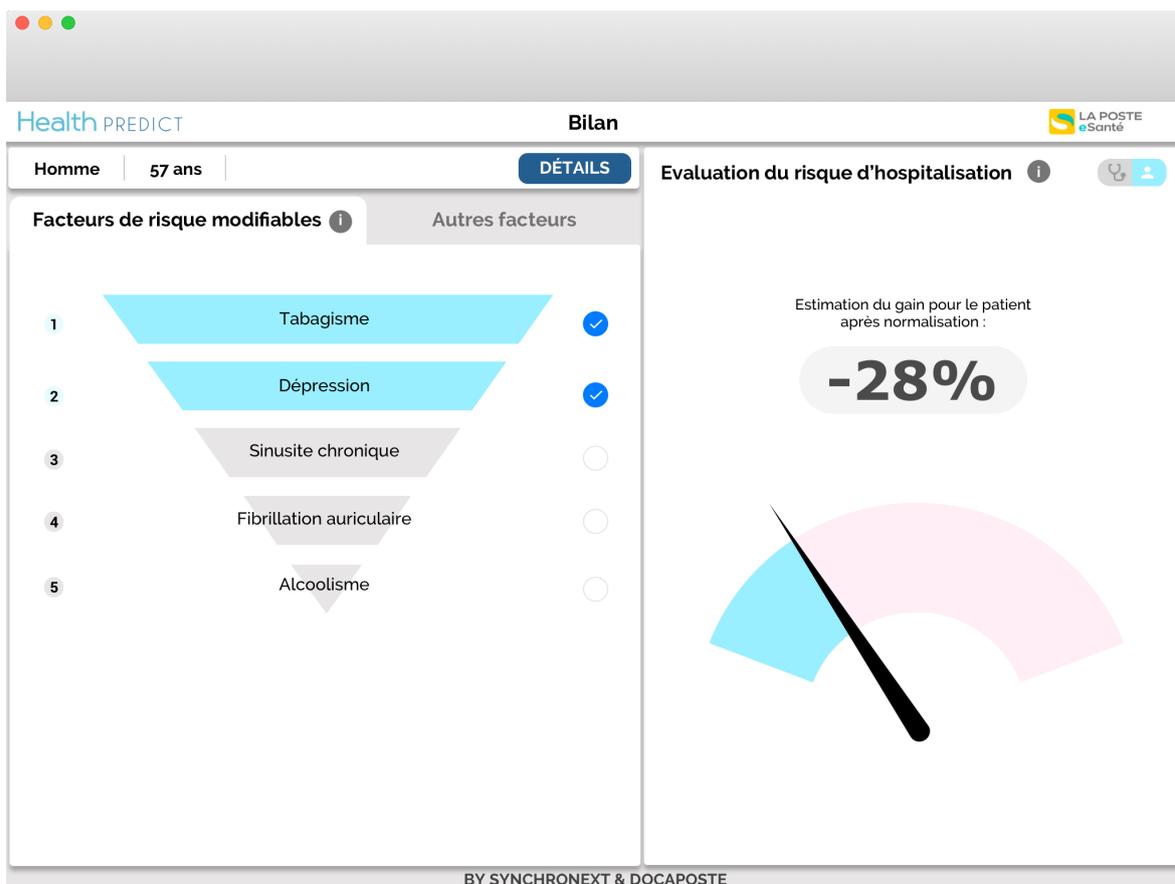
Figure 6.10 – Patient view with the total gain on the hospitalization risk after the management of the 'Smoking' and 'Depression' factors.

6.6 Perspectives of evolution of the application and interface

First of all, during the development of the interface it was difficult to position ourselves between providing the same services offered by a medical consultation software and to really deliver the heart of our studies with the hospitalization prediction since this prediction is only relevant if we have access to patient information. This has had an impact on the evolution of our interface and its different versions.

Indeed, whether we enter the patient's information from the interface of our product or not significantly changes the way we design it. This translates in providing physicians with fields relating to their patients that they must fill in.

In addition, at some point we had planned to conduct epidemiological monitoring (see Figure 6.11). This was a direction that could be envisaged with data collected in PRIMEGE, especially since the project will be extended in the coming years to the rest of France. Patient data will thus be collected from all over France and will therefore provide a greater representativeness of general practice.

However, among all the possible options we focused on hospitalization forecasting. We have especially highlighted in our interface the levers of action and the current and predictive percentages related to the hospitalization risk.

Before arriving at our current design we went through many steps to address how to present certain elements from patient's medical records and choices made for greater clarity in the essential information to be displayed (see Figure 6.12 and Figure 6.13).

We planned to add a direct link to medical recommendations issued by the HAS (Haute Autorité de Santé) on both the synthesis screen (see Figure 6.14) and the prediction screen with the DREFC (Diffusion des REcommandations Francophones En Consultation de Médecine Générale),[14] however since the SFMG (Société Française de Médecine Générale),[15] the holders of this application, launched a competing project to PRIMEGE we have abandoned this idea for the time being since it implies more development on our side. Still we can say that this initiative was perceived positively by the interviewed GPs.

---

[14] http://drefc.sfmg.org/
[15] http://www.sfmg.org/accueil/
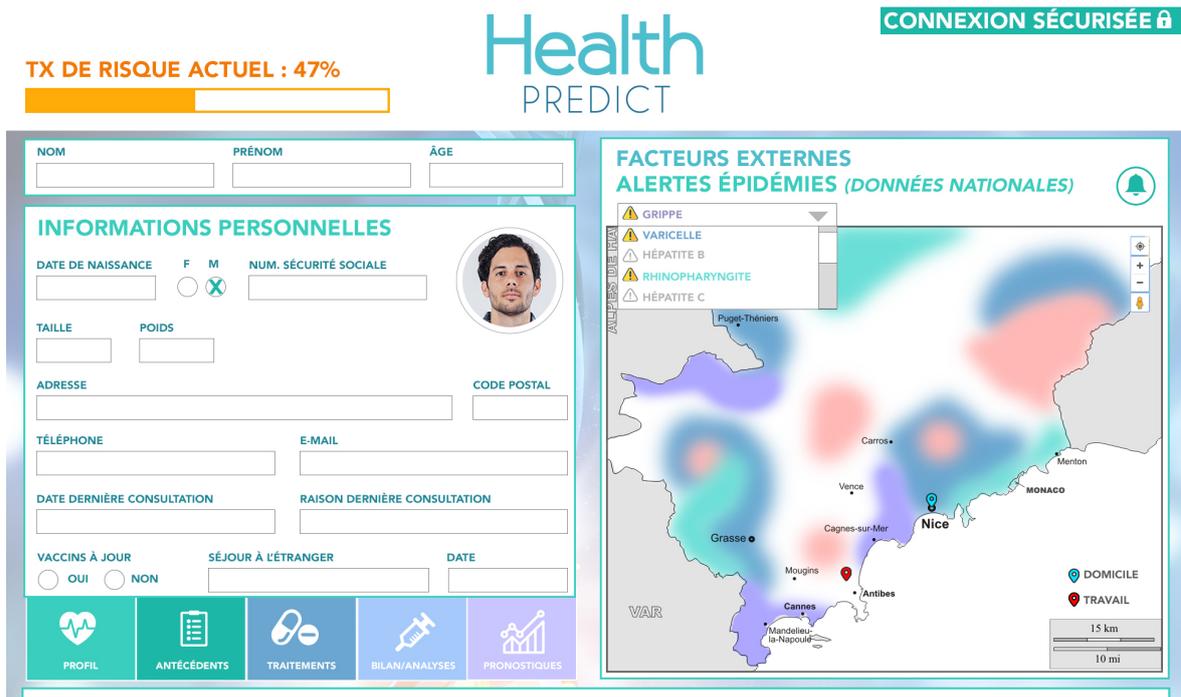
Figure 6.11 – Mock-up screen that contains patient information on the left part and on the right is displayed geolocalized epidemiological alerts.
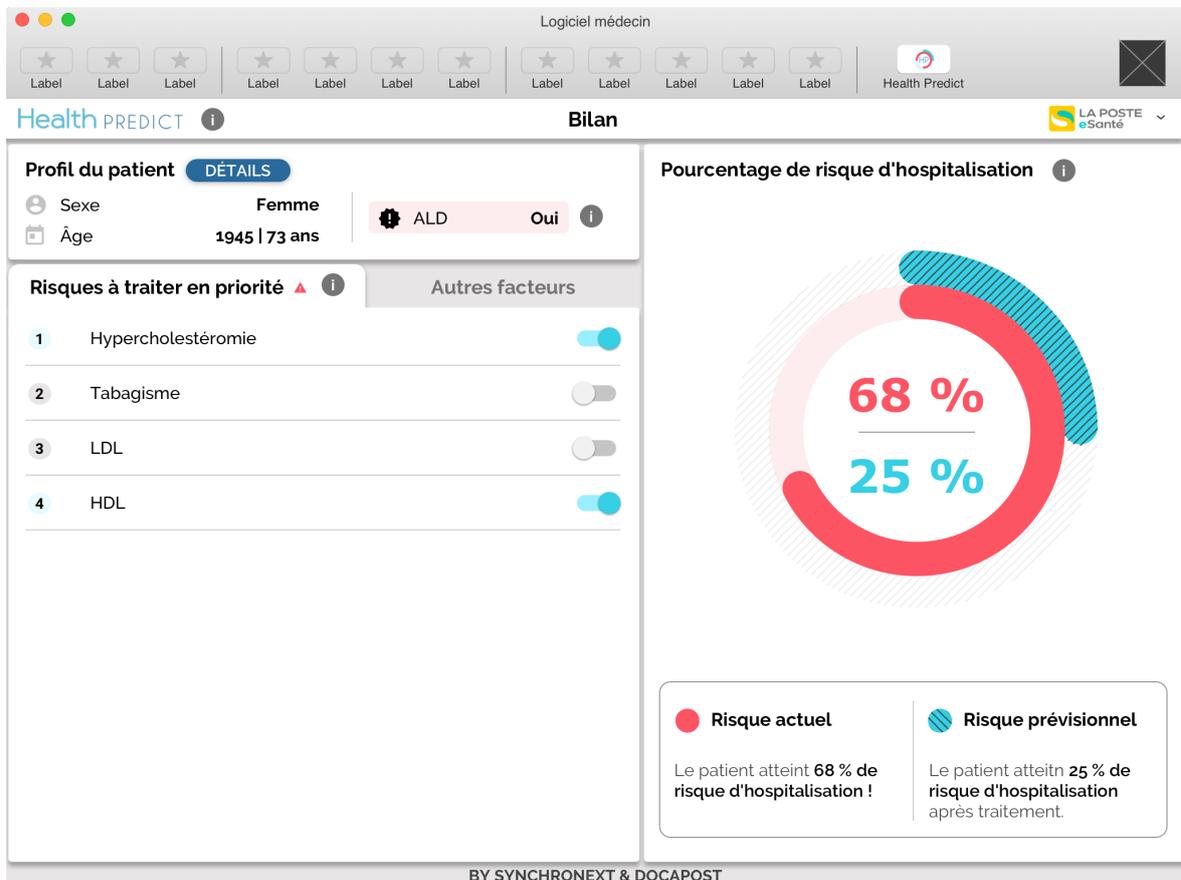
Figure 6.12 – First version of our interface with the risk factors and the prediction of hospitalization.
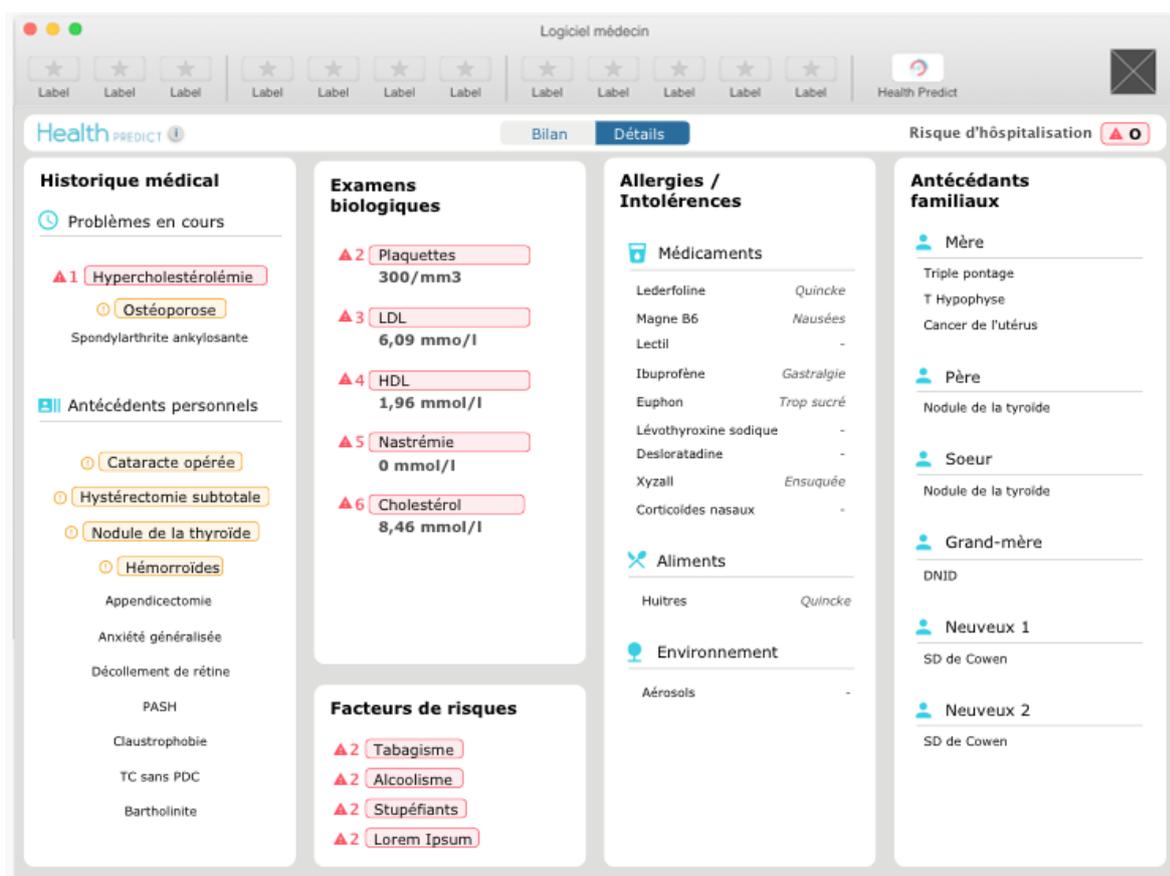
Figure 6.13 – Direction imagined for the synthesis screen on the patient's health condition.
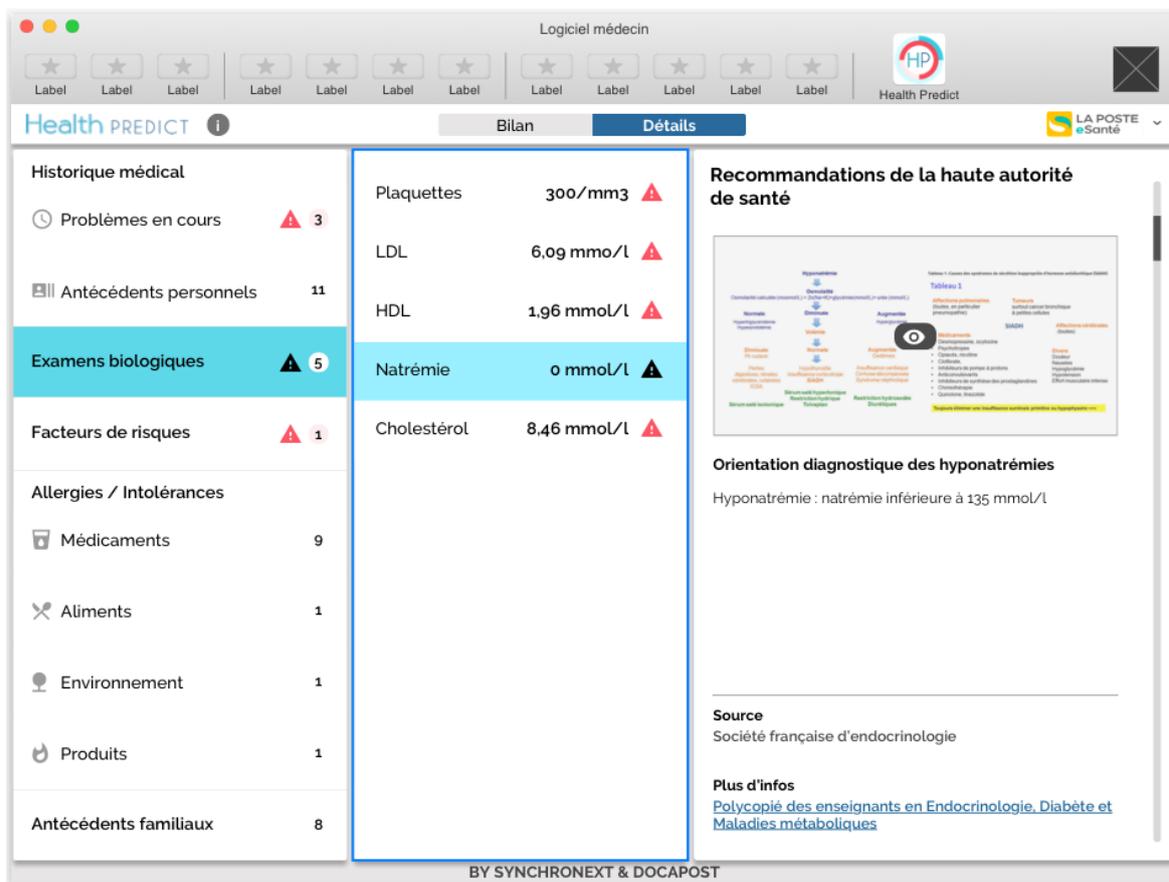
Figure 6.14 – First version of the synthesis screen on bioassays with a link to the recommendations of the HAS.

6.7   Conclusion and Future Work

Exploration, analysis and scheduling of factors responsible in the decision of hospitalization are crucial in order to prevent this event from happening and improving a patient's overall health. We propose an interface that meets these requirements and can be connected to the product we are developing.

The final goal of our application is to be integrated to a medical consultation software, and this will avoid the double entry of a patient's medical file and the multiplication of tools used by GP which would be a waste of time. The creation of tools to assist physicians in their practices may be an additional motivation for them to fill their patient records as accurately as possible, allowing them to obtain feedbacks on their practices and those of their colleagues.

As future work we plan to evaluate and adapt our interface with interviews of a representative panel of GPs once it will be fully integrated in a medical consultation software. The evaluation will focus in particular on the time spent on the interface and on the relevance of the provided information to the physician to plan an action. Finally, the aspects related to the impact of therapeutic patient compliance have to be studied.

Chapter 7

Conclusion

In Chapter 5, we evaluated the injection of domain knowledge in the vector representation of electronic medical records (EMRs) and our methodology to select automatically relevant knowledge to include. In the context of hospitalization prediction our method showed it improves the obtained categorization. Chapter 6 introduced the design of our interface and showed the steps we went through to display the results given by our algorithm while meeting the expectations of general practitioners (GPs). Our results are promising, they demonstrate that it is possible to improve the prediction results obtained by a machine learning algorithm with domain knowledge from different referentials and to make it a product dedicated to decision support to help the physician in the exercise of his practice. The enrichment of vectors with domain knowledge does not always imply improvement in the prediction of patient hospitalization, but we have outlined a way to evaluate and select the knowledge that contributes to this prognosis. In this way, the thesis showed the approaches, evaluations and evolution we achieved in order to develop HealthPredict, whether on aspects relating to research in artificial intelligence or on the development of an interface exploiting our work to assist GPs in the follow-up of their patients. Moreover, we applied and have been pre-selected as part of the Article 51[1] to assess the efficiency of our solution both in terms of reducing hospitalizations and in the management of mortality and morbidity.

Future work will be required to implement our solution, in particular to integrate it to a consultation software and conduct usability tests. Once the tests with GPs have been carried out, we will be able to change our interface according to the feedback received and focus on the development of new functionalities. We also need to investigate the

---

[1] https://bit.ly/2RkKMYw

contribution of biological analyses, but this implies to evaluate ways to discretize values, handle missing values and time-series. Further work needs to address the issue of biological analysis and biometric analyses, because these measures according to their type are not always represented in a structured way (e.g., blood presure...) and it would be of great interest to exploit them. To achieve this, they need to be properly extracted from free text beforehand.

Beyond hospitalization risks, the study of other medical risks (cardiovascular risk, cancer risk, mental health risk, rehospitalization...) may be analyzed in the same way as hospitalization prediction in our solution.

In relation to our experiments carried out in Chapter 5, we plan, in the short term, to evaluate the impact of new domain-specific knowledge graphs and identify other properties involved in predicting hospitalization of patients, as we focused principally on drugs in our study. We also plan to train our own model of DBpedia Spotlight in order to further avoid noise with named entities from other domains. We then intend to investigate different depth levels of subjects with DBpedia, since so far, we only integrated the knowledge on the direct subject, and to deal with the recognition of complex expressions, experiencer (the patient or members of his family) and entity negation. Finally, we want to work on a vector representation different from bag-of-words, thus coupling semantic relationships and textual data and to support the detection of negation alongside with the handling of complex expressions.

With knowledge graph and data from PRIMEGE, it is possible to inject data in a smarter way, by looking effectively at diseases cured by a specific drug and inject knowledge about drug interaction when the two drugs are actually present. The negation of a concept is misunderstood in our actual experimental setup since a GP often adds notes about potential symptoms that are not part of the final diagnosis. We expect to achieve better and more reliable results by handling these cases, but that involves detecting negation in the first place.

In terms of scalability, most text preprocessing could be performed on the client's side (on physician's computer) which can efficiently distribute the load on a server. For security reasons, it is necessary to encrypt the communications between the server where our application is running and the physicians' computers. A server should be able to handle the load for requests to check patient hospitalization risks, but we will still have to do some tests to verify the architecture to implement to support the deployment of our

solution first at the regional level and then at the national level.

The only calls we made on the Web are those involving DBpedia and Wikidata from which we have stored the temporary results, so that we do not have to run the same query over and over again. For instance, we query only once Wikidata for a given drug. Although for DBpedia and Wikidata we could have run them on a local server. A potential security breach of HealthPredict may come from the use of third-party programs in server mode (entities recognition, knowledge graphs management...), this is a point that we will need to address in our final architecture. There is one concern with the singular nature of these data, even if GDPR compliant (General Data Protection Regulation)[2] -after removing possible personal details in free text, or dropping free text-, it is still possible to cross-reference them to identify a patient NA et al. [2018]. Changing certain values is also not a sufficient guarantee to ensure the anonymity of patients and even the medical staff. A possible solution would be to split the data into different databases and to perform federated computing, but this point remains to be proven.

On a longer term and with a broader view, the collection of national data will allow us to build a better model, as all the specificities of the French population will be captured, which will increase the number of data used for training. However data curation will require a lot of work, since in our first experiments all hospitalization causes have been used which implies to use hospitalization that cannot be predicted. An effort on the part of the policies should also be invested to set up a healthcare system that captures the different patient trajectories (hospital, care home, specialist, drugstore, general practitioner and direct information from the patient...) and thus improve the results of collective prevention tools.

---

[2] https://gdpr-info.eu/

Appendix A

Appendix

A.1 Appendix figures
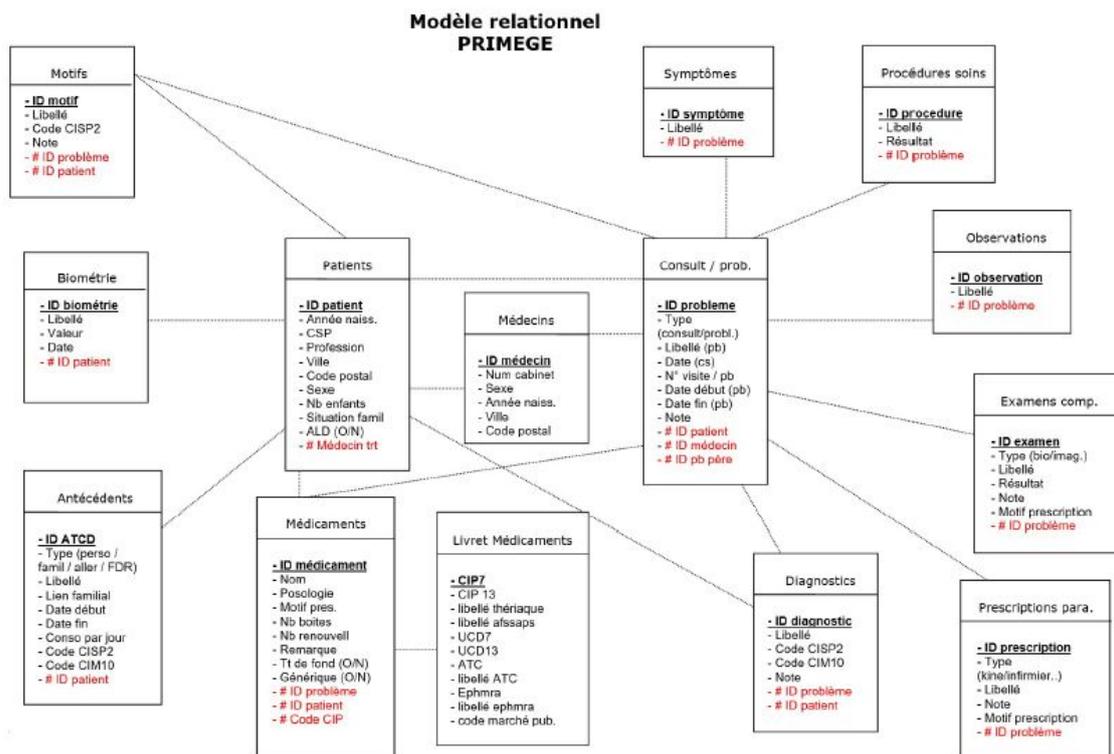
Figure A.1 – Relational diagram of the PRIMEGE database. Source: LACROIX-HUGUES [2016] and http://www.primege.org/.

A.2 Appendix tables

Table A.1 – The main groups of the ATC classification.

A	Alimentary tract and metabolism
B	Blood and blood forming organs
C	Cardiovascular system
D	Dermatologicals
G	Genito-urinary system and sex hormones
H	Systemic hormonal preparations, excluding sex hormones and insulins
J	Antiinfectives for systemic use
L	Antineoplastic and immunomodulating agents
M	Musculo-skeletal system
N	Nervous system
P	Antiparasitic products, insecticides and repellents
R	Respiratory system
S	Sensory organs
V	Various

Table A.2 – The main groups of the ICPC-2 classification. The complete specification[1] includes the subcategories belonging to these main groups.

A	General and unspecified
B	Blood, blood forming organs, lymphatics, spleen
D	Digestive
F	Eye
H	Ear
K	Circulatory
L	Musculoskeletal
N	Neurological
P	Psychological
R	Respiratory
S	Skin
T	Endocrine, metabolic and nutritional
U	Urology
W	Pregnancy, childbirth, family planning
X	Female genital system and breast
Y	Male genital system
Z	Social problems

---

[1] http://3cgp.docpatient.net/wp-content/uploads/2016/12/icpc_copydesk_en.pdf

Table A.3 – Medical tests values considered and discretized according to reference ranges. Examples are given between square bracket.

Category	Medical test (unit)
Blood cells	lymphocite percentage (%) $[\emptyset; < 18\%; 18 - 44\%; > 44\%]$,
	lymphocite number (\mm3) $[\emptyset; < 1250; 1250 - 7K; > 7K]$,
	neutrophil percentage (%) $[\emptyset; < 40\%; 40 - 70\%; 70 - 80\%; > 80\%]$,
	eosinophil percentage (%) $[\emptyset; 0 - 4\%; > 4\%]$,
	eosinophil number (\mm3) $[\emptyset; < 40; 40 - 650; > 650]$,
	basophil percentage (%) $[\emptyset; 0 - 2\%; > 2\%]$,
	percent monocytes (%) $[\emptyset; < 4.7\%; 4.7 - 12.5\%; > 12.5\%]$,
	percent band basophiles (%) $[\emptyset; 0 - 11\%; > 11\%]$,
	number of platelets (\mm3) $[\emptyset; < 160K; 160K - 350K; > 350K]$
Hemoglobin	mean corpuscular hemoglobin concentration (%) $[\emptyset; < 30\%; 30 - 35\%; > 35\%]$,
	mean corpuscular volume (fl) $[\emptyset; < 80; 80 - 100; > 100]$,
	mean corpuscular hemoglobin (pg) $[\emptyset; < 27; 27 - 32; > 32]$
Coagulation	sedimentation rates (mm) $[\emptyset; 0 - 15; > 15]$,
	prothrombin level (%) $[\emptyset; < 0.8; 0.8 - 1.2; 1.21 - 2; 2.1 - 3; > 3]$
Lipidemia	LDL cholesterol (mmol/l) $[\emptyset; < 2.85; 2.85 - 3.34; > 3.34]$,
	HDL cholesterol (mmol/l) $[\emptyset; < 1.06; 1.06 - 1.80; > 1.80]$,
	total cholesterol (mmol/l) $[\emptyset; < 3.87; 3.87 - 5.68; > 5.68]$,
	triglyceridemia (mmol/l) $[\emptyset; < 0.5; 0.5 - 2; > 2]$
Chemistry	albuminemia (µmol/l) $[\emptyset; < 650; 650 - 800; > 800]$,
	uremia (mmol/l) $[\emptyset; < 3; 3 - 7.5; > 7.5]$,
	proteinemia (g/l) $[\emptyset; < 60; 60 - 80; > 80]$,
	chloremia (mmol/l) $[\emptyset; < 100; 100 - 110; > 110]$,
	natremia (mmol/l) $[\emptyset; < 135; 135 - 145; > 145]$,
	calcium level (mmol/l) $[\emptyset; < 2.20; 2.20 - 2.75; > 2.75]$
Biology	serum glutamo-oxaloacetate transferase (IU/l) $[\emptyset; < 8; 8 - 30; > 30]$,
	glutamopyruvate transferase (IU/l) $[\emptyset; < 8; 8 - 35; > 35]$,
	C reactive protein (mg/l) $[\emptyset; \leq 6; > 6]$

References

AGIBETOV, A., K. BLAGEC, H. XU et M. SAMWALD. 2018, «Fast and scalable neural embedding models for biomedical sentence classification», *BMC bioinformatics*, vol. 19, nᵒ 1, p. 541. xi, 12

ARTSTEIN, R. et M. POESIO. 2008, «Inter-coder agreement for computational linguistics», *Computational Linguistics*, vol. 34, nᵒ 4, p. 555–596. 42

ASH, J. S., M. BERG et E. COIERA. 2004, «Some unintended consequences of information technology in health care: the nature of patient care information system-related errors», *Journal of the American Medical Informatics Association*, vol. 11, nᵒ 2, p. 104–112. 82

BERGSTRA, J. et Y. BENGIO. 2012, «Random search for hyper-parameter optimization», *Journal of Machine Learning Research*, vol. 13, nᵒ Feb, p. 281–305. 25, 57

BIRKHEAD, G. S., M. KLOMPAS et N. R. SHAH. 2015, «Uses of electronic health records for public health surveillance to advance public health», *Annual review of public health*, vol. 36, p. 345–359. 2, 7

BOJANOWSKI, P., E. GRAVE, A. JOULIN et T. MIKOLOV. 2016, «Enriching word vectors with subword information», *arXiv preprint arXiv:1607.04606*. 10

BORDES, A., N. USUNIER, A. GARCIA-DURAN, J. WESTON et O. YAKHNENKO. 2013, «Translating embeddings for modeling multi-relational data», dans *Advances in neural information processing systems*, p. 2787–2795. 33

BOYD, C. M., J. DARER, C. BOULT, L. P. FRIED, L. BOULT et A. W. WU. 2005, «Clinical practice guidelines and quality of care for older patients with multiple comorbid diseases: implications for pay for performance», *Jama*, vol. 294, nᵒ 6, p. 716–724. 1

BREIMAN, L. 2001, «Random forests», *Machine learning*, vol. 45, nᵒ 1, p. 5–32. 14, 23, 27, 57

BROWN, P. F., P. V. DESOUZA, R. L. MERCER, V. J. D. PIETRA et J. C. LAI. 1992, «Class-based n-gram models of natural language», *Computational linguistics*, vol. 18, nᵒ 4, p. 467–479. 10

CAWLEY, G. C. et N. L. TALBOT. 2010, «On over-fitting in model selection and subsequent selection bias in performance evaluation», *Journal of Machine Learning Research*, vol. 11, nº Jul, p. 2079–2107. 25, 57

CECCARELLI, D., C. LUCCHESE, S. ORLANDO, R. PEREGO et S. TRANI. 2013, «Dexter: an open source framework for entity linking», dans *Proceedings of the sixth international workshop on Exploiting semantic annotations in information retrieval*, ACM, p. 17–20. 34

CHANG, C.-C. et C.-J. LIN. 2011, «Libsvm: a library for support vector machines», *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, nº 3, p. 27. 23, 27, 57

CHOI ET AL. 2017, «Gram: graph-based attention model for healthcare representation learning», dans *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, p. 787–795. 32

CONROY, R., K. PYÖRÄLÄ, A. E. FITZGERALD, S. SANS, A. MENOTTI, G. DE BACKER, D. DE BACQUER, P. DUCIMETIERE, P. JOUSILAHTI, U. KEIL et al.. 2003, «Estimation of ten-year risk of fatal cardiovascular disease in europe: the score project», *European heart journal*, vol. 24, nº 11, p. 987–1003. 74

CORBY, O. et C. F. ZUCKER. 2010, «The kgram abstract machine for knowledge graph querying», dans *Web Intelligence and Intelligent Agent Technology (WI-IAT)*, vol. 1, IEEE, p. 338–341. 57

CUNNINGHAM, H. 2002, «Gate: A framework and graphical development environment for robust nlp tools and applications», dans *Proc. 40th annual meeting of the association for computational linguistics (ACL 2002)*, p. 168–175. 35

DAIBER, J., M. JAKOB, C. HOKAMP et P. N. MENDES. 2013, «Improving efficiency and accuracy in multilingual entity extraction», dans *Proceedings of the 9th International Conference on Semantic Systems (I-Semantics)*. 34, 37

DE ROSIS, S. et C. SEGHIERI. 2015, «Basic ict adoption and use by general practitioners: an analysis of primary care systems in 31 european countries», *BMC medical informatics and decision making*, vol. 15, nº 1, p. 70. 2

DEMŠAR, J. 2006, «Statistical comparisons of classifiers over multiple data sets», *Journal of Machine learning research*, vol. 7, nº Jan, p. 1–30. 65

DEVLIN, J., M.-W. CHANG, K. LEE et K. TOUTANOVA. 2018, «Bert: Pre-training of deep bidirectional transformers for language understanding», *arXiv preprint arXiv:1810.04805.* 11, 29

D'AGOSTINO, R. B., R. S. VASAN, M. J. PENCINA, P. A. WOLF, M. COBAIN, J. M. MASSARO et W. B. KANNEL. 2008, «General cardiovascular risk profile for use in primary care», *Circulation*, vol. 117, nº 6, p. 743–753. 74

EISENSCHLOS, J., S. RUDER, P. CZAPLA, M. KARDAS, S. GUGGER et J. HOWARD. 2019, «Multifit: Efficient multi-lingual language model fine-tuning», *arXiv preprint arXiv:1909.04761.* 11

ESCUDIÉ, J.-B., B. RANCE, G. MALAMUT, S. KHATER, A. BURGUN, C. CELLIER et A.-S. JANNOT. 2017, «A novel data-driven workflow combining literature and electronic health records to estimate comorbidities burden for a specific disease: a case study on autoimmune comorbidities in patients with celiac disease», *BMC medical informatics and decision making*, vol. 17, nº 1, p. 140. 9

FLACH, J. M., P. SCHANELY, L. KUENNEKE, B. CHIDORO, J. MUBASLAT et B. HOWARD. 2018, «Electronic health records and evidence-based practice: Solving the little-data problem», dans *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care*, vol. 7, SAGE Publications Sage India: New Delhi, India, p. 30–35. 73, 74

FONCUBIERTA RODRIGUEZ, A. 2014, *Description and retrieval of medical visual information based on language modelling*, thèse de doctorat, University of Geneva. xi, 10

FORMAN, G. et M. SCHOLZ. 2010, «Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement», *ACM SIGKDD Explorations Newsletter*, vol. 12, nº 1, p. 49–57. 27, 60

FRUNZA, O., D. INKPEN et T. TRAN. 2011, «A machine learning approach for identifying disease-treatment relations in short texts», *IEEE transactions on knowledge and data engineering*, vol. 23, nº 6, p. 801–814. 33

GAZZOTTI, R., C. FARON ZUCKER, F. GANDON, V. LACROIX-HUGUES et D. DARMON. 2019a, «Évaluation des améliorations de prédiction d'hospitalisation par l'ajout de connaissances métier aux dossiers médicaux», dans *EGC 2019 - Conférence Extraction*

et Gestion des connaissances 2019, *Revue des Nouvelles Technologies de l'Information (RNTI)*, vol. RNTI-E-35, Metz, France. URL https://hal.archives-ouvertes.fr/hal-01967586. 31

GAZZOTTI, R., C. FARON ZUCKER, F. GANDON, V. LACROIX-HUGUES et D. DARMON. 2020, «Injection of Automatically Selected DBpedia Subjects in Electronic Medical Records to boost Hospitalization Prediction», dans *SAC2020 - The 35th ACM/SIGAPP Symposium On Applied Computing*, Brno, Czech Republic, doi:10.1145/3341105.3373932. URL https://hal.archives-ouvertes.fr/hal-02389918. 31

GAZZOTTI, R., E. NOUAL, C. FARON ZUCKER, F. GANDON, A. GIBOIN, V. LACROIX-HUGUES et D. DARMON. 2019b, «Designing the Interaction with a prediction system to prevent hospitalization», dans *RJCIA 2019 - Rencontres des Jeunes Chercheurs en Intelligence Artificielle PFIA 2019*, Toulouse, France, p. 54–58. URL https://hal.archives-ouvertes.fr/hal-02157559. 70

GAZZOTTI, R., C. F. ZUCKER, F. GANDON, V. LACROIX-HUGUES et D. DARMON. 2019c, «Injecting domain knowledge in electronic medical records to improve hospitalization prediction», dans *The 16th Extended Semantic Web Conference (ESWC 2019)*, *Lecture Notes in Computer Science*, vol. 11503, Springer International Publishing, Cham, p. 116–130, doi:10.1007/978-3-030-21348-0_8. 31

GOLDSTEIN, B. A., A. M. NAVAR, M. J. PENCINA et J. IOANNIDIS. 2017, «Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review», *Journal of the American Medical Informatics Association*, vol. 24, nᵒ 1, p. 198–208. 23

HAN, X. et L. SUN. 2011, «A generative entity-mention model for linking entities with knowledge base», dans *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, Association for Computational Linguistics, p. 945–954. 34

HARRIS, Z. S. 1954, «Distributional structure», *Word*, vol. 10, nᵒ 2-3, p. 146–162. 9, 13

HENRIKSSON, A., J. ZHAO, H. BOSTRÖM et H. DALIANIS. 2015, «Modeling heterogeneous clinical sequence data in semantic space for adverse drug event detection», dans *Data*

Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on, IEEE, p. 1–8. 14

HERSH, W. R. 2007, «Adding value to the electronic health record through secondary use of data for quality assurance, research, and surveillance», *Clin Pharmacol Ther*, vol. 81, p. 126–8. 2

HILLESTAD, R., J. BIGELOW, A. BOWER, F. GIROSI, R. MEILI, R. SCOVILLE et R. TAYLOR. 2005, «Can electronic medical record systems transform health care? potential health benefits, savings, and costs», *Health affairs*, vol. 24, nᵒ 5, p. 1103–1117. 2

HOWARD, J. et S. RUDER. 2018, «Universal language model fine-tuning for text classification», *arXiv preprint arXiv:1801.06146*. 11

JHA, K., M. RÖDER et A.-C. N. NGOMO. 2017, «Eaglet–a named entity recognition and entity linking gold standard checking tool», dans *European Semantic Web Conference*, Springer, p. 149–154. 34

JIN, B., C. CHE, Z. LIU, S. ZHANG, X. YIN et X. WEI. 2018, «Predicting the risk of heart failure with ehr sequential data modeling», *IEEE Access*, vol. 6, p. 9256–9261. 16

KHERA, S., D. KOLTE, S. DEO, A. KALRA, T. GUPTA, D. ABBOTT, N. KLEIMAN, D. L. BHATT, G. C. FONAROW, O. KHALIQUE et al.. 2019, «Derivation and external validation of a simple risk tool to predict 30-day hospital readmissions after transcatheter aortic valve replacement.», *EuroIntervention: journal of EuroPCR in collaboration with the Working Group on Interventional Cardiology of the European Society of Cardiology*, vol. 15, nᵒ 2, p. 155–163. xii, 73, 75

KRIPPENDORFF, K. 1970, «Estimating the reliability, systematic error and random error of interval data», *Educational and Psychological Measurement*, vol. 30, nᵒ 1, p. 61–70. 42

LACROIX-HUGUES, V. 2016, *Utilisation des enregistrements médicaux électroniques, exemple dútilisation dans le cadre du projet PRIMEGE PACA ; quels sont les principaux motifs de recours, diagnostics et prescriptions en soins primaires.*, thèse de doctorat. URL https://dumas.ccsd.cnrs.fr/dumas-01450989. xii, 47, I

LACROIX-HUGUES, V., D. DARMON, C. PRADIER et P. STACCINI. 2017, «Creation of the first french database in primary care using the icpc2: Feasibility study.», *Studies in health technology and informatics*, vol. 245, p. 462–466. 7

LE, Q. et T. MIKOLOV. 2014, «Distributed representations of sentences and documents», dans *International conference on machine learning*, p. 1188–1196. 10

LEE, J., W. YOON, S. KIM, D. KIM, S. KIM, C. H. SO et J. KANG. 2020, «Biobert: a pre-trained biomedical language representation model for biomedical text mining», *Bioinformatics*, vol. 36, nº 4, p. 1234–1240. 11

LIN, C., H. CANHAO, T. MILLER, D. DLIGACH, R. M. PLENGE, E. W. KARLSON et G. K. SAVOVA. 2012, «Feature engineering and selection for rheumatoid arthritis disease activity classification using electronic medical records», dans *ICML Workshop on Machine Learning for Clinical Data Analysis*. 16

LIU, J., Z. ZHANG et N. RAZAVIAN. 2018a, «Deep ehr: Chronic disease prediction using medical notes», *arXiv preprint arXiv:1808.04928*. 68

LIU, L., J. SHEN, M. ZHANG, Z. WANG et J. TANG. 2018b, «Learning the joint representation of heterogeneous temporal events for clinical endpoint prediction», *arXiv preprint arXiv:1803.04837*. 14

LOPER, E. et S. BIRD. 2002, «Nltk: the natural language toolkit», *arXiv preprint cs/0205028*. 15

DE LUSIGNAN, S. et C. VAN WEEL. 2005, «The use of routinely collected computer data for research in primary care: opportunities and challenges», *Family practice*, vol. 23, nº 2, p. 253–263. 2

MÀRQUEZ, L. et H. RODRÍGUEZ. 1998, «Part-of-speech tagging using decision trees», dans *European Conference on Machine Learning*, Springer, p. 25–36. 15

MCCULLAGH, P. et J. A. NELDER. 1989, *Generalized linear models*, vol. 37, CRC press. 22, 27, 57

MIKOLOV, T., K. CHEN, G. S. CORRADO et J. A. DEAN. 2015, «Computing numeric representations of words in a high-dimensional space», US Patent 9,037,464. 10, 14

MIN, H., H. MOBAHI, K. IRVIN, S. AVRAMOVIC et J. WOJTUSIAK. 2017, «Predicting activities of daily living for cancer patients using an ontology-guided machine learning methodology», *Journal of biomedical semantics*, vol. 8, nº 1, p. 39. 31

MOUSSALLEM, D., R. USBECK, M. RÖEDER et A.-C. N. NGOMO. 2017, «Mag: A multilingual, knowledge-base agnostic and deterministic entity linking approach», dans *Proceedings of the Knowledge Capture Conference*, ACM, p. 9. 35

NA, L., C. YANG, C.-C. LO, F. ZHAO, Y. FUKUOKA et A. ASWANI. 2018, «Feasibility of reidentifying individuals in large national physical activity data sets from which protected health information has been removed with use of machine learning», *JAMA network open*, vol. 1, nᵒ 8, p. e186 040–e186 040. 95

NADEAU, C. et Y. BENGIO. 2003, «Inference for the generalization error», *Mach. Learn.*, vol. 52, nᵒ 3, doi:10.1023/A:1024068626366, p. 239–281, ISSN 0885-6125. URL https://doi.org/10.1023/A:1024068626366. 65

ORDÓNEZ, F. J., P. DE TOLEDO et A. SANCHIS. 2013, «Activity recognition using hybrid generative/discriminative models on home environments using binary sensors», *Sensors*, vol. 13, nᵒ 5, p. 5460–5477. 32

PEARSON, K. 1901, «Liii. on lines and planes of closest fit to systems of points in space», *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, nᵒ 11, p. 559–572. 13

PEDREGOSA, F., G. VAROQUAUX, A. GRAMFORT, V. MICHEL, B. THIRION, O. GRISEL, M. BLONDEL, P. PRETTENHOFER, R. WEISS, V. DUBOURG, J. VANDERPLAS, A. PASSOS, D. COURNAPEAU, M. BRUCHER, M. PERROT et E. DUCHESNAY. 2011, «Scikit-learn: Machine learning in Python», *Journal of Machine Learning Research*, vol. 12, p. 2825–2830. 27, 57

PENNINGTON, J., R. SOCHER et C. MANNING. 2014, «Glove: Global vectors for word representation», dans *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, p. 1532–1543. 10, 32

PIKTUS, A., N. B. EDIZEL, P. BOJANOWSKI, E. GRAVE, R. FERREIRA et F. SILVESTRI. 2019, «Misspelling oblivious word embeddings», dans *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 3226–3234. 10

RIBEIRO, M. T., S. SINGH et C. GUESTRIN. 2016, «Why should i trust you?: Explaining the

predictions of any classifier», dans *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, ACM, p. 1135–1144. 22

RIZZO, G. et R. TRONCY. 2011, «Nerd: A framework for evaluating named entity recognition tools in the web of data», dans *10th International Semantic Web Conference (ISWC'11), Demo Session, Bonn, Germany*, p. 1–4. 34

SALGUERO, A. G., M. ESPINILLA, P. DELATORRE et J. MEDINA. 2018, «Using ontologies for the online recognition of activities of daily living», *Sensors*, vol. 18, nᵒ 4, p. 1202. 32

SHANG, J., T. MA, C. XIAO et J. SUN. 2019, «Pre-training of graph augmented transformers for medication recommendation», *arXiv preprint arXiv:1906.00346.* 33

SINGH, A., G. NADKARNI, O. GOTTESMAN, S. B. ELLIS, E. P. BOTTINGER et J. V. GUTTAG. 2015, «Incorporating temporal ehr data in predictive models for risk stratification of renal function deterioration», *Journal of biomedical informatics*, vol. 53, p. 220–228. 13, 14, 28

SNOEK, J., H. LAROCHELLE et R. P. ADAMS. 2012, «Practical bayesian optimization of machine learning algorithms», dans *Advances in neural information processing systems*, p. 2951–2959. 25

STROETMANN, K. A., J. ARTMANN, V. N. STROETMANN, D. PROTTI, J. DUMORTIER, S. GIEST, U. WALOSSEK et D. WHITEHOUSE. 2011, «European countries on their journey towards national ehealth infrastructures», *Luxembourg: Office for official publications of the european communities.* 2

SUTTON, C., A. MCCALLUM et al.. 2012, «An introduction to conditional random fields», *Foundations and Trends® in Machine Learning*, vol. 4, nᵒ 4, p. 267–373. 24, 27

TANG, B., H. CAO, X. WANG, Q. CHEN et H. XU. 2014, «Evaluating word representation features in biomedical named entity recognition tasks», *BioMed research international*, vol. 2014. xi, 11

TCHECHMEDJIEV, A., A. ABDAOUI, V. EMONET, S. ZEVIO et C. JONQUET. 2018, «Sifr annotator: ontology-based semantic annotation of french biomedical text and clinical notes», *BMC bioinformatics*, vol. 19, nᵒ 1, p. 405. 34, 52

TIBSHIRANI, R. 1996, «Regression shrinkage and selection via the lasso», *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, nᵒ 1, p. 267–288. 45, 59

TINETTI, M. E., S. T. BOGARDUS JR, J. V. AGOSTINI et al.. 2004, «Potential pitfalls of disease-specific guidelines for patients with multiple conditions», *N Engl J Med*, vol. 351, nᵒ 27, p. 2870–4. 1

USBECK, R., M. RÖDER, A.-C. NGONGA NGOMO, C. BARON, A. BOTH, M. BRÜMMER, D. CECCARELLI, M. CORNOLTI, D. CHERIX, B. EICKMANN et al.. 2015, «Gerbil: general entity annotator benchmarking framework», dans *Proceedings of the 24th international conference on World Wide Web*, International World Wide Web Conferences Steering Committee, p. 1133–1143. 35

WANG, A., A. SINGH, J. MICHAEL, F. HILL, O. LEVY et S. R. BOWMAN. 2018, «Glue: A multi-task benchmark and analysis platform for natural language understanding», *arXiv preprint arXiv:1804.07461*. 33

WHETZEL, P. L., N. F. NOY, N. H. SHAH, P. R. ALEXANDER, C. NYULAS, T. TUDORACHE et M. A. MUSEN. 2011, «Bioportal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications», *Nucleic acids research*, vol. 39, nᵒ suppl_2, p. W541–W545. 34

ZHANG, Z., X. HAN, Z. LIU, X. JIANG, M. SUN et Q. LIU. 2019, «Ernie: Enhanced language representation with informative entities», *arXiv preprint arXiv:1905.07129*. 33

ZHONG, Z., L. ZHENG, G. KANG, S. LI et Y. YANG. 2017, «Random erasing data augmentation», *arXiv preprint arXiv:1708.04896*. 29