Injecting Domain Knowledge in Electronic Medical Records to Improve Hospitalization Prediction

Raphaël Gazzotti, Catherine Faron Zucker, Fabien Gandon, Virginie Lacroix-Hugues, David Darmon

Injecting Domain Knowledge in Electronic Medical Records to Improve Hospitalization Prediction

Raphaël Gazzotti[1 3][0000−0002−5618−9776], Catherine Faron-Zucker[1][0000−0001−5959−5561], Fabien Gandon[1][0000−0003−0543−1232], Virginie Lacroix-Hugues[2], and David Darmon[2][0000−0002−4425−4163]

[1] Université Côte d'Azur, Inria, CNRS, I3S, France
firstname.surname@unice.fr
[2] Université Côte d'Azur, Département de Médecine Générale, France
vhugues@outlook.fr, david.darmon@unice.fr
[3] SynchroNext, France

Abstract. Electronic medical records (EMR) contain key information about the different symptomatic episodes that a patient went through. They carry a great potential in order to improve the well-being of patients and therefore represent a very valuable input for artificial intelligence approaches. However, the explicit knowledge directly available through these records remains limited, the extracted features to be used by machine learning algorithms do not contain all the implicit knowledge of medical expert. In order to evaluate the impact of domain knowledge when processing EMRs, we augment the features extracted from EMRs with ontological resources before turning them into vectors used by machine learning algorithms. We evaluate these augmentations with several machine learning algorithms to predict hospitalization. Our approach was experimented on data from the PRIMEGE PACA database that contains more than 350,000 consultations carried out by 16 general practitioners (GPs).

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

1 Introduction

Electronic medical records (EMRs) contain essential information about the different symptomatic episodes a patient goes through. They have the potential to improve patient well-being and are therefore a potentially valuable source to artificial intelligence approaches. However, the linguistic variety as well as the tacit knowledge in EMRs can impair the prognosis of a machine learning algorithm.

In this paper, we extract ontological knowledge from text fields contained in EMRs and evaluate the benefit when predicting hospitalization. Our study uses a dataset extracted from the PRIMEGE PACA relational database [10] which contains more than 350,000 consultations in French by 16 general practitioners

(Table 1). In this database, text descriptions written by general practitioners are available 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.

Our initial observation was that the knowledge available in a database such as PRIMEGE remains limited to the specificities of each patient and in particular that the texts found in there are based on a certain amount of implicit information known to medical experts. Moreover, the level of detail of the information contained in a patient's file is variable. Therefore, a machine learning algorithm exploiting solely the information at its disposal in an EMR will not be able to exploit this specific knowledge implicit in the documents it analyzes or, at best, it will have to relearn this knowledge by itself, possibly in an incomplete and costly way.

In that context, our main research question is: *Can ontological augmentations of the features improve the prediction of the occurrence of an event?*. In our case study, we aim to predict a patient's hospitalization using knowledge from different knowledge graphs in the medical field. In this paper, we focus on the following sub-questions:

- How to integrate domain knowledge into a vector representation used by a machine learning algorithm?
- Is the addition of domain knowledge improving the prediction of a patient's hospitalization?
- Which domain knowledge combined with which machine learning methods provide the best prediction of a patient's hospitalization?

To answer these questions, we first survey the related work (section 2) and position our contribution. We then introduce the proposed method for semantic annotation and knowledge extraction from texts and specify how ontological knowledge is injected upstream into the vector representation of EMRs (section 3). Then, we present the experimental protocol and discuss the results obtained (section 4). Finally, we conclude and provide our perspectives for this study (section 5).

2 Related Work

In [12], 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). 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

Table 1. Data collected in the PRIMEGE PACA 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 (Y/N)
	Personal history
	Family history
	Risk factors
	Allergies
Consultations	Date
	Reasons of encounter
	Symptoms related by the patient and medical observation
	Further investigations
	Diagnosis
	Drugs prescribed (dose, number of boxes, reasons of the prescription)
	Paramedical prescriptions (biology/imaging)
	Medical procedures

of this augmentation on different machine learning approaches: they used the AQ21 and the extension of this algorithm AQ21-OG to compare.

In [5], 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 concepts of these knowledge graphs in their attention mechanism, the graphs are transformed using the embedding obtained with Glove. The results show that such a model better performs when identifying a pathology rarely observed in a training dataset than a recurrent neural network and it also better generalizes when confronted with less data in the training set. Again, this work also does not exploit other kinds of relationships, while we compare the impact of different kinds and sources of knowledge.

In [15] the authors extract knowledge from the dataset of [13] and structure it with an ontology developed for this purpose, then they 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 on which this task has been addressed (up to 1.9% on average). Although they exploit solely the ontology 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.

3 Enriching Vector Representations of EMRs with Ontological Knowledge

3.1 Extraction of Ontological Knowledge from EMRs

Our study aims to analyze and compare the impact of knowledge from different sources, whether separately incorporated or combined, on the vector representation of patients' medical records to predict hospitalization. To extract domain knowledge underlying terms used in text descriptions written by general practitioners, we search the texts for medical entities and link them to the concepts to which they correspond in Wikidata, DBpedia and health sector specific knowledge graphs such as those related to drugs. Wikidata and DBpedia were chosen because general concepts can only be identified with general repositories. In this section, we describe how these extractions are performed but we do not focus on this step as it is only a means to an end for our study and could be replaced by other approaches.

Knowledge Extraction based on DBpedia To detect in an EMR concepts from the medical domain present in DBpedia, we used the semantic annotator DBpedia Spotlight [7]. Together with the 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 subjects relevant to the prediction of hospitalization, as they relate to severe pathologies (Table 2).

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[4] to determine if these resources have as subject (property `dcterms:subject`) one or more of the 14 selected concepts.

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 (e.g., the abbreviation "ic" which means "heart failure" is not recognized by DBpedia Spotlight, but is thus correctly identified through this symbolic approach).

In the rest of the article, the $+s$ notation refers to an approach using the enrichment of representations with concepts from DBpedia according to the method previously described. The $+s*$ notation refers to an approach that does not exploit all text fields and extracts concepts from text fields related to the patient's personal history, allergies, environmental factors, current health problems, reasons for consultations, diagnoses, medications, care procedures, reasons for prescribing medications and physician observations. The $+s*$ approach focuses

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

Table 2. 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

on the patient's own record, not on his family history and past problems. Note that the symptom field is used by doctors for various purposes and this is the reason why we excluded this field in the DBpedia concept extraction procedure for this approach.

Knowledge Extraction based on Wikidata Wikidata[5] is an open knowledge base that centralizes data from various projects of the Wikimedia Foundation. Its coverage on some domains differs from that of DBpedia. We extracted drug-related knowledge by querying Wikidata's endpoint.[6] 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 codes (property `wdt:P2892`) and CUI RxNorm (property `wdt:P3345`) have been recovered using medical domain specific ontologies. Indeed, the codes from these three representations are not necessarily all present to identify a drug in Wikidata, but at least one of them allows us to find the resource about a given drug. From the URI of a drug, we extract property-concept pairs related to the drugs for three selected properties (e.g., 'Pethidine' is a narcotic, 'Meprobamate' cures headache, 'Atazanavir' interacts with 'Rabeprazole').

In the rest of the article, the notation $+wa$ refers to an approach using the enrichment of our representations with the property 'acts as such', $+wm$ indicates the usage of the property 'treated disease' and $+wi$ of the property 'drugs interacts with'.

Knowledge Extraction based on Domain Specific Ontologies We were interested in the impact of contributions from domain specific knowledge graphs

[5] https://www.wikidata.org
[6] https://query.wikidata.org/sparql

especially for text fields containing international drug codes from the Anatomical, Therapeutic and Chemical (ATC) classification and codes related to the reasons for consulting a general practitioner with the International Classification of Primary Care (CISP-2). We thus extracted knowledge based on three OWL representations specific to the medical domain: ATC,[7], NDF-RT[8] and CISP2.[9] The choice of OWL-SKOS representations of CISP2 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 states.

We extracted from the ATC OWL-SKOS representation 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[10] (e.g. '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)).

Similarly, we extracted from the OWL-SKOS representation of CISP2 the labels of the superclasses with 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 (e.g., 'Symptom and complaints' (CISP-2 code : H05) has for superclass 'Ear' (CISP-2 code : H)).

In the OWL representation of NDF-RT, we selected three drug properties relevant to the prediction of hospitalization: 'may_treat' property (e.g. 'Tahor', whose main molecule is 'Atorvastatin' (ATC code: C10AA05) can cure 'Hyperlipoproteinemias' (Hyperlipidemia), 'CI_with' (e.g. 'Tahor' is contraindicated in 'Pregnancy') and 'may_prevent' (e.g. 'Tahor' can prevent 'Coronary Artery Disease'). A dimension in our DME vector representation will be a property-value pair. Here is an example RDF description of the drug Tahor:

```
@prefix : <http://evs.nci.nih.gov/ftp1/NDF-RT/NDF-RT.owl> .
:N0000022046 a owl:Class; rdfs:label "ATORVASTATIN"; :UMLS_CUI "C0286651";
    owl:subClassOf [
        rdf:type owl:Restriction; owl:onProperty :may_prevent;
        owl:someValuesFrom :N0000000856 ];
    owl:subClassOf [
        rdf:type owl:Restriction; owl:onProperty :CI_with;
        owl:someValuesFrom :N0000010195 ];
    owl:subClassOf [
        rdf:type owl:Restriction; owl:onProperty :may_treat;
        owl:someValuesFrom :N0000001594 ].
:N0000000856 rdfs:label "Coronary Artery Disease [Disease/Finding]".
:N0000010195 rdfs:label "Pregnancy [Disease/Finding]".
:N0000001594 rdfs:label "Hyperlipoproteinemias [Disease/Finding]".
```

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

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

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

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

In the rest of the article, the notation $+c$ refers to an approach using the enrichment of vector representations with ATC and the number attached specifies the different depth levels used (e.g., $+c_{1-3}$ indicates that 3 superclass depth levels are integrated in the same vector representation). $+t$ indicates the enrichment of vector representations with CISP2. $+d$ indicates the enrichment of vector representations with NDF-RT, followed in indices by CI if property 'CI_with' is used, *prevent* if property 'may_prevent' is used, *treat* if property 'may_treat' is used. For example, $+d_{CI,prevent,treat}$ refers to the case where these three properties are used together in the same vector representation of DMEs.

3.2 Integrating Ontological Knowledge in a Vector Representation

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 has a very precise definition in a given sector. We opted for a model using a bag-of-words representation (BOW) 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 the distinction of patients to hospitalize or not, 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 previously represented in this way.

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 family history. To do this, we have introduced provenance prefixes during the creation of the bag-of-words to trace the contribution of the different fields.

Concepts from knowledge graphs are considered as a token in a textual message. When a concept is identified in a patient's medical record, it 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 (e.g., 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 example, in vectors exploiting the NDF-RT (vector representation d), we find the couple consisting of `CI_with` as a property - contraindicated with- and the name of a pathology or condition, for example '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 i^{th} patient. Let $C^i = \{c_1^i, c_2^i, ..., c_n^i\}$ be the bag of concepts for the i^{th} patient resulting from the extraction of concepts belonging to knowledge graphs after analysis of his consultations from semi-structured data such as text fields listing drugs and pathologies with their related codes, and

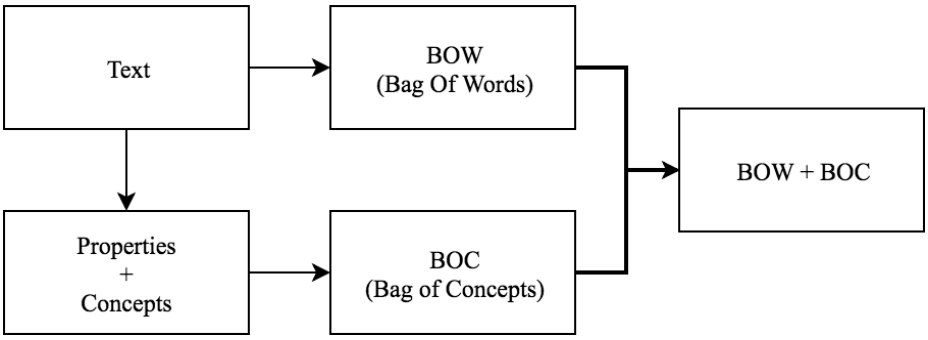


Fig. 1. Workflow diagram to generate vector representations integrating ontological knowledge alongside with textual information.

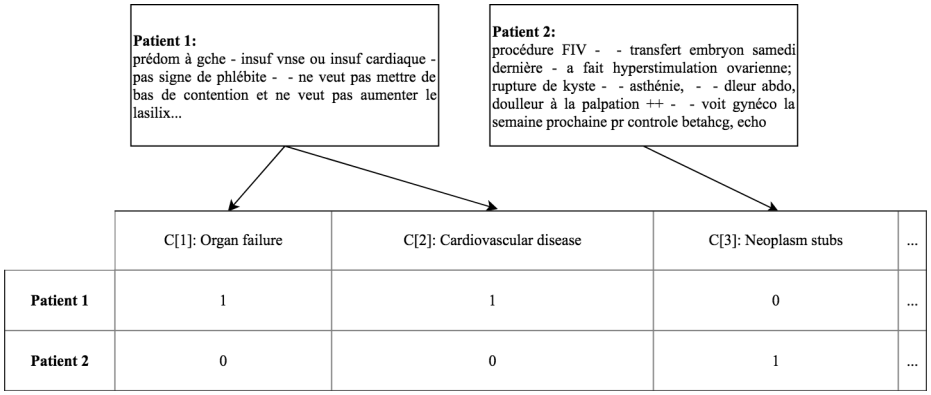
Table 3. Alternative concept vector representations of the EMR of a patient under Tahor generated using NDF-RT.

	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	...

unstructured data from free texts such as observations. The different machine learning algorithms exploit the aggregation of these two vectors: $x^i = V^i \oplus C^i$.

From 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... -" (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 2: 'Organ failure' and 'Cardiovascular disease', these concepts were retrieved by the property `dcterms:subject` from DBpedia. The concept (occurrence) vector 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 2).

As for the exploitation of NDF-RT, let us consider again the example description of the drug Tahor introduced in section 3.1. It can be used to enrich the vector representation of the EMR of patients under Tahor as detailed in Table 3.



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

Fig. 2. 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".

4 Experiments and Results

4.1 Dataset and Protocol

We tested and evaluated our approach for enriching the vector representation of EMRs with ontological knowledge on a balanced dataset DS_B containing data on 714 patients hospitalized and 732 patients not hospitalized. When the observation field is filled in by general practitioners, it can go from 50 characters to 300 characters on average. The best filled in fields concern prescribed drugs and reasons for consultations, then come the antecedents and active problems.

Since we use non-sequential machine learning algorithms to assess the enrichment of ontological knowledge, we had to aggregate all patients' consultations 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. Thus, the text fields previously described (Table 1) are transformed into vectors.

We evaluated the vector representations by nested cross-validation [3], with an external loop with a K fixed at 10 and for the internal loop a K fixed at 3 with exploration of hyperparameters by random search [1] over 150 iterations.

The different experiments were conducted on an HP EliteBook 840 G2, 2.6 hHz, 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 ontologies with Corese Semantic Web Factory [6],[11] a software platform for the Semantic Web. It implements RDF, RDFS, SPARQL 1.1 Query & Update, OWL RL.

4.2 Selected Machine Learning Algorithms

We performed the hospitalization prediction task with different state-of-the-art algorithms available in the Scikit-Learn library [14]:

- *SVC*: Support vector machine (SVC stands for 'Support Vector Classification') whose implementation is based on the libsvm implementation [4]. The regularization coefficient C, the kernel used by the algorithm and the gamma coefficient of the kernel were determined by nested cross-validation.
- *RF*: The random forest algorithm [2]. The number of trees in the forest, the maximum tree depth, the minimum number of samples required to divide an internal node, the minimum number of samples required to be at a leaf node and the maximum number of leaf nodes were determined by nested cross-validation.
- *Log*: The algorithm of logistic regression [11]. The regularization coefficient C and the norm used in the penalization were determined by nested cross-validation.

We opted for a bag-of-words model and the above cited machine learning algorithms since it is possible to provide a native interpretation of the decision of these algorithms, thus allowing the physician to specify the reasons for hospitalizing a patient with the factors on which he can operate to prevent this event from occurring. Moreover, logistic regression and random forest algorithms are widely used in order to predict risk factors in EHR [9]. Finally, the limited size of our dataset excluded neural networks approaches.

4.3 Results

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 [8]. 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).

$$TP_f = \sum_{i=1}^{K} TP^{(i)} \qquad FP_f = \sum_{i=1}^{K} FP^{(i)} \qquad FN_f = \sum_{i=1}^{K} FN^{(i)}$$

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

$$F_{tp,fp} = \frac{2.TP_f}{2.TP_f + FP_f + FN_f}$$

Table 4 summarizes the results for each representation and method combination tested on the DS_B dataset:

- *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.
- $+s$: refers to an enrichment with concepts from the DBpedia knowledge base.
- $+s*$: refers to an enrichment with concepts from the DBpedia knowledge base, unlike $+s$, not all text fields are exploited, thus, concepts from fields related to the 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 are extracted.
- $+t$: refers to an enrichment with concepts from the OWL-SKOS representation of CISP-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.

4.4 Discussion

In general terms, knowledge graphs improve the detection of true positives cases, hospitalized patients correctly identified as such (Table 5 and 6) and provide a broader knowledge of the data in patient files like the type of health problem with CISP-2 (Table 7).

Although the approach in combination with $+s*$ does not achieve the best final results, this approach achieves the best overall performance among all the approaches tested with 0.858 under logistic regression when using 8 K folds from the KFold during the training phase (Figure 3). It also surpasses other methods under 3 K partitions by exceeding the *baseline* by 0.9% and at 4 K 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 from the enrichment provided by knowledge graphs.

Despite the shallow OWL-SKOS representation of CISP2, the $+t$ configuration is sufficient to improve a patient's hospitalization predictions, if we compare

Table 4. $F_{tp,fp}$ 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
$+t+s+c_2+wa+wi$	0.8258	0.8486	0.8547	0.8430
$+t+s*+c_2+wa+wi$	0.8239	0.8494	0.8543	0.8425
$+t+c_2+wa+wi$	0.8140	0.8531	**0.8571**	0.8414

Table 5. 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

Table 6. 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

Table 7. 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 CISP2 (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 encounter and observations fields), daily chest pains are considered as 'Emergency' and a tongue tumor as 'Neoplasm stubs' by DBpedia

its results to those of the *baseline*. Surprisingly enough, a second level of super class 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 them, this can be explained by the fact that the introduction of a large number of attributes ultimately provides little information, unlike the second level of hierarchy.

However the results show that applying DBpedia expansion to fields indirectly related to the patient's condition, such as family history, can lead machine learning algorithms to draw wrong conclusions even if prefixes have been added to distinguish provenance fields. The text field related to symptoms has been poorly filled in by doctors – as it was an 'observation' field – and the majority of the remarks thus detected by DBpedia Spotlight are mostly false alerts. Moreover, the qualitative analysis of the results showed cases involving negation ('pas de SC d'insuffisance cardiaque', meaning 'no symptom of heart failure') and poor consideration of several terms ('brûlures mictionnelles', related to bladder infection, are associated with 'Brûlure', a burn, which, therefore, has as subject the concept 'Urgence médicale', a medical emergency). Both cases are current limitations of our approach and we consider for our future work the need for handling negation and complex expressions.

5 Conclusion and Future Work

In this paper, we have presented a method for combining knowledge from knowledge graphs, whether specialized or generalist, and textual information to predict the hospitalization of patients. To do this, we generated different vector representations coupling concept vectors and bag-of-words and then evaluated their performance for prediction with different machine learning algorithms.

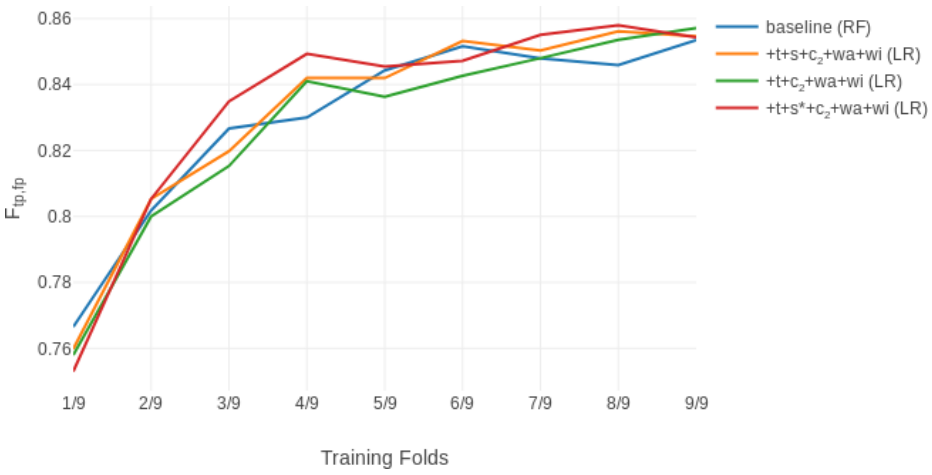


Fig. 3. Convergence curve obtained following the training on n (x-axis) KFold partitions for different configurations of the Table 4.

In the short term, we plan to identify additional concepts involved in predicting hospitalization of patients and to evaluate the impact of additional domain specific knowledge, as we focused mainly on drugs in our study. We also intend to propose an approach to automatically extract candidate medical concepts from DBpedia. Finally we plan to improve the coupling of semantic relationships and textual data in our vector representation, and support the detection of negation and complex expressions in texts.

Acknowledgement.

This work is partly funded by the French government labelled PIA program under its IDEX UCAJEDI project (ANR-15-IDEX-0001).

References

1. Bergstra, J., Bengio, Y.: Random search for hyper-parameter optimization. Journal of Machine Learning Research **13**(Feb), 281–305 (2012)
2. Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)
3. Cawley, G.C., Talbot, N.L.: On over-fitting in model selection and subsequent selection bias in performance evaluation. Journal of Machine Learning Research **11**(Jul), 2079–2107 (2010)
4. Chang, C.C., Lin, C.J.: Libsvm: a library for support vector machines. ACM transactions on intelligent systems and technology (TIST) **2**(3), 27 (2011)
5. Choi et al.: Gram: graph-based attention model for healthcare representation learning. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 787–795. ACM (2017)

6. Corby, O., Zucker, C.F.: The kgram abstract machine for knowledge graph querying. In: Web Intelligence and Intelligent Agent Technology (WI-IAT). vol. 1, pp. 338–341. IEEE (2010)
7. Daiber, J., Jakob, M., Hokamp, C., Mendes, P.N.: Improving efficiency and accuracy in multilingual entity extraction. In: Proceedings of the 9th International Conference on Semantic Systems (I-Semantics) (2013)
8. Forman, G., Scholz, M.: Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. ACM SIGKDD Explorations Newsletter **12**(1), 49–57 (2010)
9. Goldstein, B.A., Navar, A.M., Pencina, M.J., Ioannidis, J.: Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. Journal of the American Medical Informatics Association **24**(1), 198–208 (2017)
10. Lacroix-Hugues, V., Darmon, D., Pradier, C., Staccini, P.: Creation of the first french database in primary care using the icpc2: Feasibility study. Studies in health technology and informatics **245**, 462–466 (2017)
11. McCullagh, P., Nelder, J.A.: Generalized linear models, vol. 37. CRC press (1989)
12. Min, H., Mobahi, H., Irvin, K., Avramovic, S., Wojtusiak, J.: Predicting activities of daily living for cancer patients using an ontology-guided machine learning methodology. Journal of biomedical semantics **8**(1), 39 (2017)
13. Ordónez, F.J., de Toledo, P., Sanchis, A.: Activity recognition using hybrid generative/discriminative models on home environments using binary sensors. Sensors **13**(5), 5460–5477 (2013)
14. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
15. Salguero, A.G., Espinilla, M., Delatorre, P., Medina, J.: Using ontologies for the online recognition of activities of daily living. Sensors **18**(4), 1202 (2018)

6 Appendix

Among the three kernels tested for 'SVC' (RBF, linear and polynomial), the nested cross-validation selected RBF and a linear kernel equally. The ridge regression (L2 regularization) was overwhelmingly chosen by nested cross-validation for the logistic regression algorithm.