Représentations de concepts médicaux apprises sur 3 millions de patients du SNDS (2008-2016)

Matthieu Doutreligne[1], Aude Leduc[1], Dinh-Phong Nguyen[1], Albert Vuagnat[1]

(1) Direction de la recherche, des études, de l'évaluation et des statistiques, 75350 Paris 07, France. Contact : matthieu.doutreligne@sante.gouv.fr

Introduction

Les bases de données médico-administratives sont des sources d'information très riches sur les systèmes de soin. Cependant, leur utilisation est complexe car elles sont constituées d'une accumulation d'événements de nature et de temporalité très diverses. En appliquant aux séquences de soins une méthode similaire à l'approche word2vec [3] ayant révolutionné le traitement automatique du langage, nous proposons des représentations vectorielles riches reflétant les interactions (co-occurrences) au cours des parcours de soins entre les codes ou événements de quatre grandes terminologies médicales françaises.

Nous avons construit une application web (disponible ici : **http://snds2vec.health-data-hub.fr:8051**) afin de visualiser ces concepts et d'effectuer des requêtes de proximité entre des codes spécifiques. Nous espérons ainsi donner une intuition sur la nature de nos représentations.

Données

Un événement est défini comme une consommation de soin accompagnée d'un code médical : diagnostic CIM10 hospitalier et ALD, acte CCAM hospitaliers et en consultations externes, médicament ATC 5e niveau en ville, actes de biologie NABM en ville. 950 millions d'événements de soin ont été extraits du SNDS dans les parcours d'un échantillon aléatoire de 3 112 565 patients âgés de 18 à 120 ans de 2008 à 2016. Les événements ont été groupés par individu et triés par date de soin, formant des séquences de codes. On note la séquence de soin d'un patient i, $x_i = [c_0, .., c_t, .., c_{T_i}]$ où $c_t \in V$, le vocabulaire de concepts médicaux de taille $|V| = 14450$.

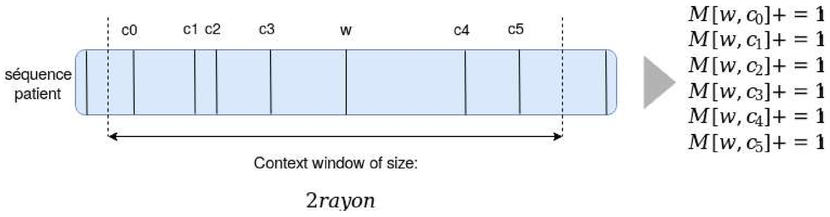
Table 1. Nomenclatures utilisées pour les représentations

Nomenclatures	Champ	Nombre de codes uniques	Nombre de codes total (millions)
CIM10	hospitalier, ALD	8013	40.1
CCAM	hospitalier, ambulatoire, ville	4547	110.4
ATC	ville	1133	559.6
NABM	ville	757	298.1

Méthodologie

Cette méthode a été récemment appliquée au domaine médical par [1]. L'hypothèse sous-jacente est que deux concepts (ici événements médicaux) sont d'autant plus similaires qu'ils apparaissent dans des contextes similaires. En choisissant une durée de contexte r, nous construisons la matrice de co-occurrence des événements $M \in \mathbb{R}^{|V|x|V|}$ où la cellule $M_{k,l}$ correspond au nombre de fois où les concepts k et l apparaissent dans une fenêtre de taille $2r$ dans la population étudiée.

Figure 1. Calcul de la matrice de co-occurrence pour un concept w dans un parcours patient

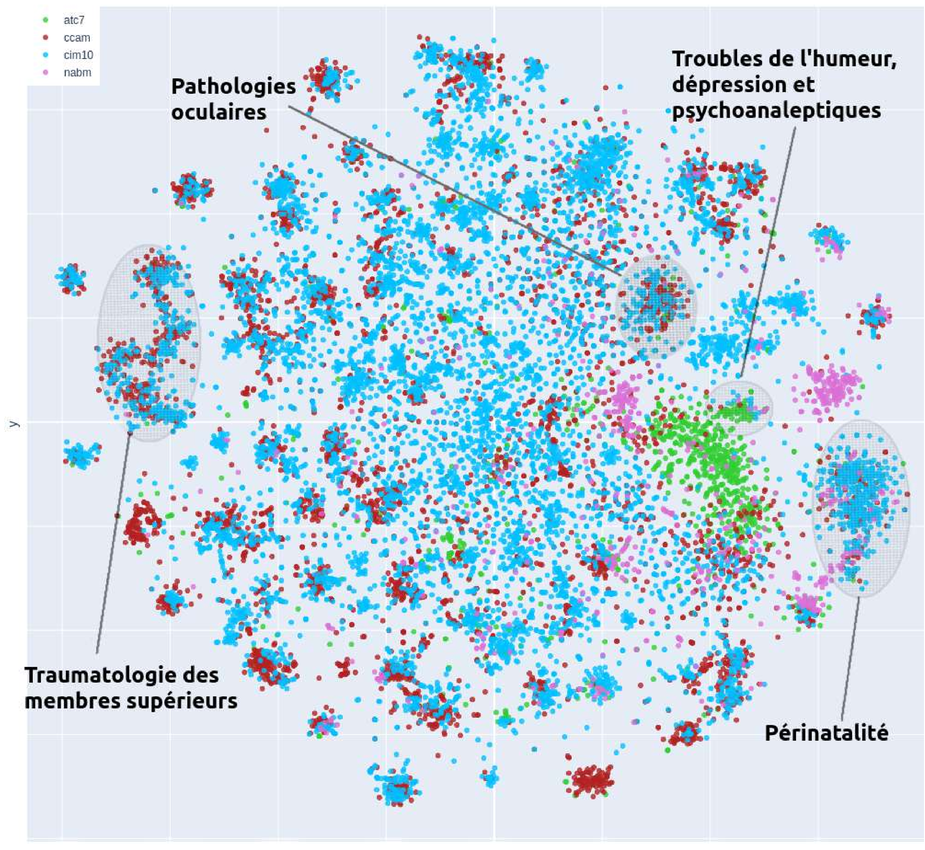


En factorisant par Décomposition en Valeurs Singulières la matrice M comme détaillée dans [2], on obtient pour chaque code k dans le vocabulaire (ex: I51, infarctus du myocarde) un vecteur $\Phi(c_k) \in \mathbb{R}^d$ où $d << |V|$. Les résultats présentés sont obtenus avec un contexte de $r = 90$ jours et une réduction de dimension à $d = 150$. Cette compression d'information est équivalente à entraîner un modèle prédisant quel code w est le plus pertinent pour un contexte donné, puis de prendre comme représentation les paramètres de ce modèle.

Résultats

En projetant en 2 dimensions les représentations, de grands groupes de pathologies ressortent avec, par exemple, la traumatologie, les pathologies oculaires, les troubles de l'humeur, la périnatalité (cf. figure 2).

Figure 2. Projection par TSNE en 2D des représentations médicales



En particulier, pour 4 concepts médicaux couvrant différents champs du système de soin, on regarde leurs 3 plus proches voisins pour les terminologies ATC, CCAM, CIM10 dans l'espace en dimension 150 (distance cosine entre vecteurs) :

Table 2. Plus proches voisins selon la terminologie pour 4 concepts médicaux

Concept (code)	néphrite tubulo-interstitielle aiguë (N10)	entorse et foulure de la cheville (S934)	insuline (humaine) (A10AB01)	fracture du col du fémur (S720)
Terminologie	CIM10	CIM10	ATC	CIM10
Prévalence (compte total)	30303 (41597)	7465 (10175)	3137 (27691)	4964 (19003)
3 voisins CIM10	- hydronéphrose avec obstruction de la jonction pyélo-urétérale (N130) - pyonéphrose (N136) - pyélonéphrite associée à un reflux (N110)	- douleur articulaire de la cheville et du pied (M2557) - luxation de la cheville (S930) - fracture fermée de l'astragale (S9210)	- diabète sucré de type 1 (E10) - présence d'implants endocriniens (Z964) - glomérulopathie au cours du diabète sucré (N083)	- fracture du trochanter (S721) - fracture fermée du col du fémur (S7200) - fracture fermée du trochanter (S7210)
3 voisins CCAM	- Pose d'une endoprothèse urétérale, par une néphrostomie déjà en place (JCLD001) - Fragmentation intrarénale de calcul avec ondes de choc ou laser par urétéronéphroscopie (JANE005) - Scintigraphie rénale glomérulaire ou tubulaire avec épreuve pharmacologique (JAQL003)	- Confection d'une orthèse non articulée cruropédieuse (ZEMP003) - Radiographie de la cheville selon 1 à 3 incidences (NGQK001) - Ostéosynthèse de fracture bimalléolaire simple, à foyer ouvert (NCCA016)	- Séance de destruction de lésion choriorétinienne par photocoagulation avec laser (BGNP003) - Séance de photocoagulation choriorétinienne du pôle postérieur, avec laser monochromatique ou laser à colorants (BGNP001) - Rétinographie par stéréophotographie, clichés composés de la périphérie rétinienne ou cliché grand champ supérieur à 60 (BGQP006)	- Remplacement de l'articulation coxofémorale par prothèse fémorale cervicocéphalique et cupule mobile (NEKA011) - Ostéosynthèse de fracture extracapsulaire du col du fémur (NBCA010) - Radiographie de l'articulation coxofémorale (NEQK010)
3 voisins ATC	- benzethonium chloride, combinations (D08AJ58) - epinephrine (C01CA24) - colecalciferol (A11CC05)	- naftazone (C05CX02) - niclosamide (P02DA01) - hydroxyzine (N05BB01)	- insulin aspart (A10AD05) - insulin lispro (A10AD04) - insulin aspart (A10AB05)	- ibuprofen (M02AA13) - megestrol (L02AB01) - phenylbutazone (M01AA01)

Discussion et conclusion

Ces premières représentations des informations médicales du SNDS semblent prometteuses pour décrire les liens entre les codes et l'utilisation effective de la codification. De nombreux cas d'application existent :

- Aide au phénotypage dans les bases médico-administratives,
- Reconstruction et comparaison avec les hiérarchies des nomenclatures (ces informations ne sont pas données en entrée de notre méthode),
- Suivi des évolutions des pratiques et des effets de substitution,
- Etudes de co-morbidités.

Plus généralement, nous pensons que ce type de représentations peut se révéler efficace pour distinguer des populations hétérogènes avec peu d'a priori sur les critères de recherche initiaux. Pour ce faire, nous poursuivons des travaux afin de recombiner ces vecteurs de manière pertinente au niveau de l'individu.

Références

[1] Andrew L. Beam, Benjamin Kompa, Allen Schmaltz, Inbar Fried, Grin Weber, Nathan P. Palmer, Xu Shi, Tianxi Cai, and Isaac S. Kohane. \Clinical Concept Embeddings Learned from Massive Sources of Multimodal Medical Data». en. In: arXiv:1804.01486 [cs, stat] (Apr. 2018). arXiv: 1804.01486. (Visited on 09/27/2019).

[2] Omer Levy and Yoav Goldberg. \Neural Word Embedding as Implicit Matrix Factorization». en. In: Neurips Process 2014 (2014), p. 9.

[3] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Je Dean. \Distributed Representations of Words and Phrases and their Compositionality». In: Advances in Neural Information Processing Systems 26. Curran Associates, Inc., 2013, pp. 3111{3119.