Representations of medical concepts learned from 3 millions patients in the French National Health Insurance Database, SNDS (2008-2016)

Matthieu Doutreligne [1] [2] [*], Aude Leduc [1], Dinh-Phong Nguyen [1], Albert Vuagnat [1]

[1]Direction de la Recherche, des Etudes, de l'Evaluation et des Statistiques, Paris
[2]Inria Paris Saclay / HAS ; [*]Work done at DREES

Introduction

Medico-administrative databases are rich sources of information on health care systems. However, their use is complex because they are made up of an accumulation of events of very diverse nature and temporality. By applying to care sequences a method similar to the word2vec approach of [4] which revolutionized automatic language processing, we propose useful vector representations reflecting the interactions (co-occurrences) during the course of care sequences between the codes or events of four major French medical terminologies. We have built a web application (available here : **http://snds2vec.health-data-hub.fr:8051**) in order to visualize these concepts and perform proximity queries between specific codes. In this way, we hope to give an intuition on the nature of our representations.

Data

An event is defined as a care consumption accompanied by a medical code and a timestamp. This corresponds to : ICD10 hospital and Long Term Diseases diagnosis (ALD in French), CCAM hospital and outpatient procedures, ATC drug 5th level in town, NABM biology procedures in town. 950 million care events were extracted from the French National Health Insurance Database (SNDS) in the pathways of a random sample of 3,112,565 patients aged between 18 and 120 years old from 2008 to 2016. We use the SNDS processing pipeline of [1] to which we have added biology to extract these events. The events have been grouped by individual and sorted by date of care, forming code sequences. We note the sequence of care of a patient i, $x_i = [c_0, .., c_t, .., c_{T_i}]$ where $c_t \in V$, the vocabulary of medical concepts of size $|V| = 14450$.

Table 1. Nomenclatures used for the representations

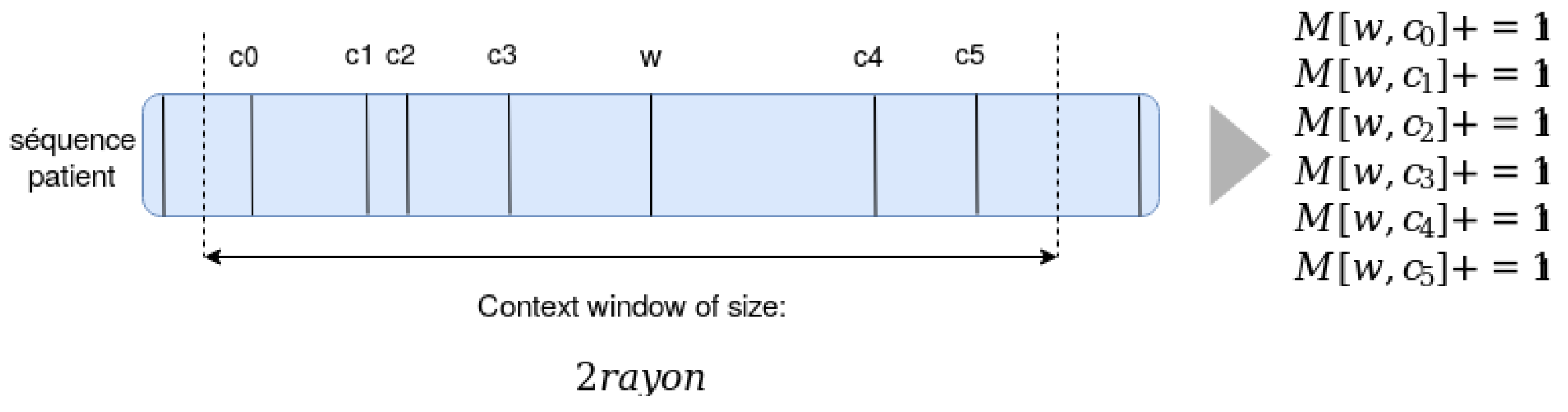
Nomenclatures	Source	Number of uniques codes	Total number of codes (millions)
ICD10	inpatient hospital, ALD	8013	40.1
CCAM	inpatient and outpatient hospital, town	4547	110.4
ATC	town	1133	559.6
NABM	town	757	298.1

Methods

This method has recently been applied to the medical domain by [2]. The underlying hypothesis is that two concepts (here medical events) are semantically similar if they share a common context.

Given a context time window of radius r, we build a co-occurrence event matrix $M \in \mathbb{R}^{|V|x|V|}$ where $M_{k,l}$ is the number of time that concepts k and l occure together in a $2r$ window in the population dataset.

Figure 1. Computation of the co-occurrence matrix for a concept w in a given patient healthcare trajectory

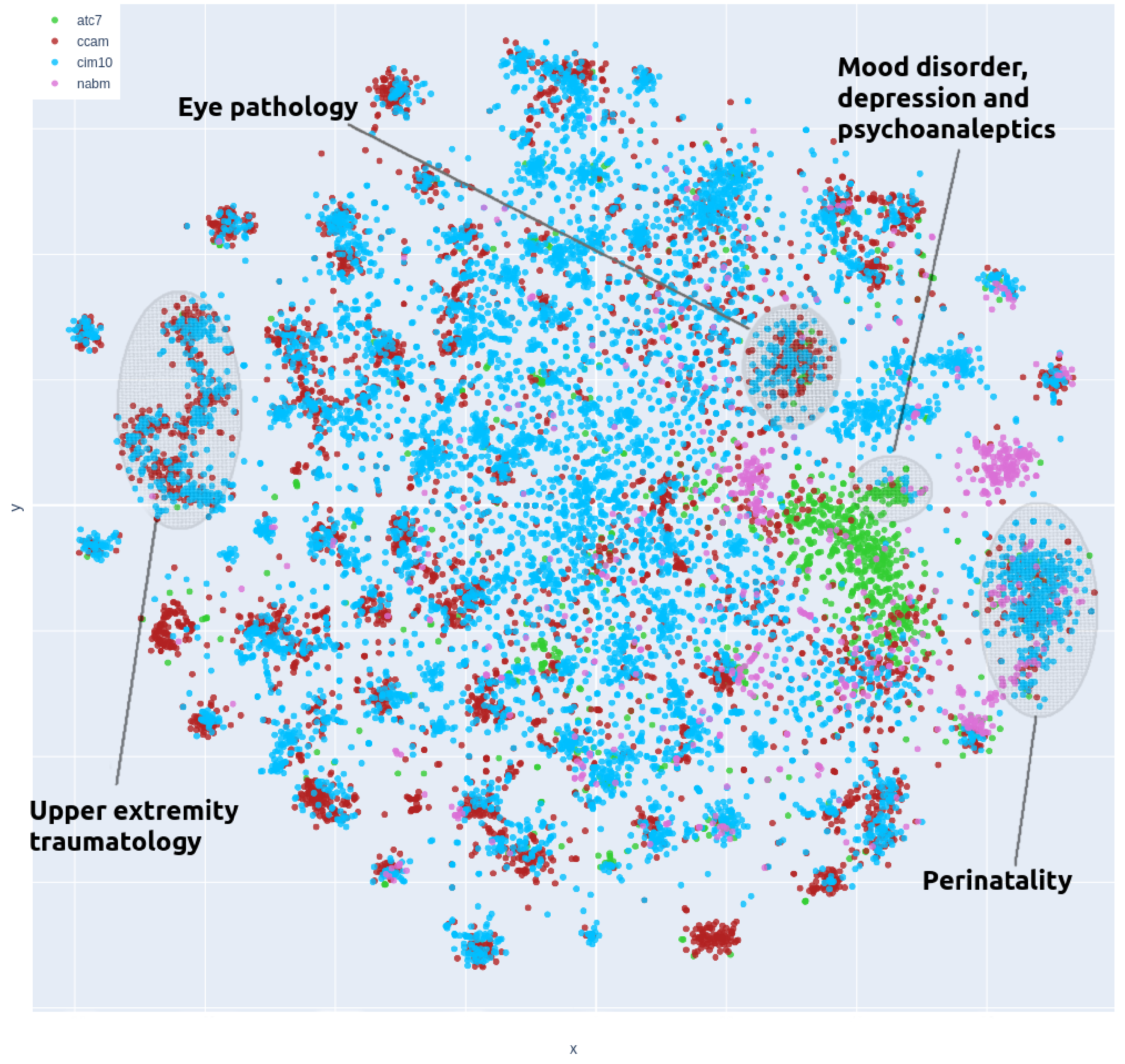


As detailed in [3], the SVD decomposition of matrix M, yields a vector $\Phi(c_w) \in \mathbb{R}^d$ where $d << |V|$ for each concept w in our vocabulary (ex: I50, heart failure). The parameters for the results shown here are: a context radius of $r = 90$ days, and a SVD reconstruction to $d = 150$ dimensions. This information compression is equivalent to the minization of a one-layer deep neural network model focusing on predicting wether a (word, context) pair (w, c) belongs to the training data (Skip-Gram with Negative Sampling). In this case, the concept representations are either the context or the input projection matrices.

Results

Projecting in 2 dimensions these embedings, we can distinguish some groups of pathology : eg. traumatology, eye pathology, mood disorders, perinatality (see. figure 2).

Figure 2. 2D TSNE projection of medical concepts representations



We detail 4 medical concepts covering various aspects of the healthcare system. With respect to the cosine distance in the representation space, we can compute their 3 nearest neighbours in the different terminologies : ATC, CCAM ICD10.

Table 2. 3-NN in each terminology for 4 medical concepts

Concept (code)	**acute tubulo-interstitial nerphritis (N10)**	**sprain and strain of the ankle (S934)**	**insulin (human) (A10AB01)**	**femoral neck fracture (S720)**
Terminology	ICD10	ICD10	ATC	ICD10
Prevalence (total count)	30303 (41597)	7465 (10175)	3137 (27691)	4964 (19003)
3-NN ICD10	- hydronephrosis with obstruction of the pyelo ureteral junction (N130) - pyonephrosis (N136) - pyelonephritis associated with reflux (N110)	- ankle and foot joint pain (M2557) - ankle dislocation (S930) - closed talus fracture (S9210)	- type 1 diabetes mellitus (E10) - presence of endocrine implants (Z964) - glomerulopathy during diabetes mellitus (N083)	- fracture of the trochanter (S721) - closed femoral neck fracture (S7200) - closed fracture of the trochanter (S7210)
3-NN CCAM	- Placement of a ureteral stent, by a nephrostomy already in place (JCLD001) - Intranarenal calculus fragmentation with shock waves or laser by ureteronephroscopy (JANE005) - Glomerular or tubular renal scintigraphy with pharmacological test (JAQL003)	- Manufacture of a non-articulated cruopedious orthosis (ZEMP003) -Radiography of the ankle at 1 to 3 incidences (NGQK001) - Simple bimalleolar fracture osteosynthesis with open focus (NCCA016)	- Session of destruction of chorioretinal lesion by photocoagulation using a laser. (BGNP003) - Posterior pole chorioretinal photocoagulation session, with monochromatic laser or dye laser. (BGNP001) - Retinography by stereophotography, compound images of the retinal periphery retinal or wide-field image greater than 60°. (BGQP006)	- Replacement of the coxofemoral joint with a cervicocephalic femoral prosthesis and mobile cup. (NEKA011) - Osteosynthesis of extracapsular femoral neck fractures (NBCA010) - Radiography of the coxofemoral joint (NEQK010)
3-NN ATC	- benzethonium chloride, combinations (D08AJ58) - epinephrine (C01CA24) - colecalciferol (A11CC05)	- naftazone (C05CX02) - niclosamide (P02DA01) - hidroxicina (N05BB01)	- insulin aspart (A10AD05) - insulin lispro (A10AD04) - insulin aspart (A10AB05)	- ibuprofen (M02AA13) - megestrol (L02AB01) - phenylbutazone (M01AA01)

Discussion

These first representations of medical information learned from the French National Health Care Insurance (SNDS) seem relevant to describe the relation between the concepts as they are coded in practice. Numerous applications could benefit from such concepts :

- Computational phenotyping ,
- Nomenclature and coding practice alignment and distributional testing,
- Comorbidity studies,
- Patient representation for causal discovery.

More generally, we think that this type of representation could be very efficient to distinguish heterogeneous population with a data-driven approaches and with little prior knowledge. We are currently cooperating with APHP to build patient level representations.

References

[1] Emmanuel Bacry, Stéphane Gaïffas, Fanny Leroy, Maryan Morel, Dinh Phong Nguyen, Youcef Sebiat, and Dian Sun. "SCALPEL3: a scalable open-source library for healthcare claims databases". In: *arXiv:1910.07045 [cs]* (Oct. 2019). arXiv: 1910.07045. URL: http://arxiv.org/abs/1910.07045 (visited on 10/18/2019).

[2] Andrew L. Beam, Benjamin Kompa, Allen Schmaltz, Inbar Fried, Griffin 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).

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

[4] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff 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.