UNIVERSITE DE LILLE – SECTEUR DROIT ET SANTE

FACULTE DE MEDECINE HENRI WAREMBOURG

Année 2020

THESE POUR LE DIPLOME D'ETAT
DE DOCTEUR EN MEDECINE

Apprentissage par transfert en épidémiologie : construction et évaluation d'un « diagnosis embedding » à partir de la base nationale médico-administrative du PMSI

Présentée et soutenue publiquement le 30 juin 2020
à 16h00 au pôle recherche

Par Margaux Riant

JURY

Président :
Monsieur le Professeur Philippe AMOUYEL

Assesseurs :
Monsieur le Professeur Emmanuel CHAZARD
Monsieur le Docteur Antoine LAMER
Monsieur le Docteur Pierre BALAYE

Directeur de thèse :
Monsieur le Docteur Grégoire FICHEUR

Avertissement

La Faculté n’entend donner aucune approbation aux opinions émises dans les thèses : celles-ci sont propres à leurs auteurs.

Liens d’intérêt

L’auteur et son directeur de thèse ne déclarent aucun lien d’intérêt en rapport avec le sujet traité au cours des trois années précédant la présentation de cette thèse.

Sigles

ATIH	Agence Technique de l'Information Hospitalière
AUC	*Area under the curve,* aire sous la courbe
CBOW	*Continuous Bag Of Words*
CCAM	Classification Commune des Actes Médicaux
CépiDC	Centre d'épidémiologie sur les causes médicales de décès
CIM-10	Classification Internationale des Maladies version 10
CMD	Catégorie Majeure de Diagnostic
DP	Diagnostic Principal
GHM	Groupe Homogène de Malades
HAD	Hospitalisation à Domicile
ICD10	*International Classification of Diseases, 10th revision* (voir CIM10)
INDS	Institut National des Données de Santé
MCO	Médecine Chirurgie Obstétrique (= court séjour)
PMSI	Programme de Médicalisation des Systèmes d'Information
ROC	*Receiver Operating Characteristic*
SNDS	Système National des Données de Santé
SNIIRAM	Système National d'Information Inter-Régime de l'Assurance Maladie
SSR	Soins de Suite et de Réadaptation (= moyen séjour)

Sommaire

Préambule

Le travail scientifique présenté dans cette thèse de médecine fait l'objet d'une publication d'article international en anglais. Il suit le plan suivant :

- Une introduction longue en français, qui poursuit deux objectifs : présenter le contexte médical avec une orientation principalement pédagogique, et présenter le contexte scientifique et l'objectif, comme le fait également l'introduction de l'article en anglais
- L'abstract en anglais, tel qu'il sera soumis en complément de l'article reproduit juste après.
- L'article en anglais, tel qu'il sera soumis à une revue scientifique internationale. Cet article suit le plan classique, dans le format imposé par le journal (introduction, matériel et méthodes, résultats, discussion)
- Une discussion en français, qui reprend pour l'essentiel la discussion en anglais de l'article

Le document est structuré ainsi en application de la circulaire Toubon[1].

Les références présentées en fin de document, ainsi que les listes de figures et tables, résultent de la fusion des parties en anglais et en français. La numérotation est donc incrémentée dans l'ensemble du document, que les parties soient anglophones ou francophones.

[1] Circulaire du 19 mars 1996 concernant l'application de la loi no 94-665 du 4 août 1994 relative à l'emploi de la langue française. JORF n°68 du 20 mars 1996 page 4258. NOR: PRMX9601403C

Introduction

1 Définitions

1.1 Définition de la réutilisation de données

La réutilisation de données, (« *data reuse* » ou « *secondary use of data* » en anglais) [1–3], se définit comme l'exploitation secondaire de données pour une finalité différente de celle pour laquelle elles ont été initialement recueillies.

Dans le domaine biomédical, la réutilisation de données est progressivement devenue une réalité en raison de la numérisation en routine d'un nombre croissant d'informations. Par exemple, les comptes rendus médicaux, la biologie, l'imagerie sont réunis dans le dossier patient informatisé (« *Electronic Health Record – EHR* » en anglais) dans une finalité de suivi et de prise en charge. D'autre part, des informations telles que l'identité d'un patient, les actes médicaux réalisés et les diagnostics posés lors d'une hospitalisation sont recueillies dans un but de facturation. L'ensemble de ces données peut ensuite de manière rétrospective servir à décrire et suivre l'activité d'un service ou d'un établissement, ou être utilisées dans un cadre de recherche. On parle alors de réutilisation de données.

La dissociation entre la finalité du recueil et la finalité de l'analyse entraîne des effets positifs comme négatifs. Elle permet de réaliser des études à bas coût, le coût du recueil des données étant économisé puisque les données concernées ont déjà été recueillies par ailleurs. Elle permet aussi d'étudier un plus grand nombre d'individus que pour les études épidémiologiques classiques, en facilitant l'identification des sujets à inclure, améliorant de fait la puissance statistique. Les principaux inconvénients de la réutilisation de données sont la possible non-adéquation entre la question analysée et les données disponibles, le design rétrospectif obligatoire de ce type d'étude (cohorte historique ou cas-témoin par exemple), et la difficulté de l'extraction de caractéristiques depuis les variables nativement disponibles [4].

1.2 Définition des Big Data ou données massives

Les « *Big Data* » ou données massives peuvent être définies comme des bases de données de si grande dimension qu'elles dépassent les capacités des outils classiques de gestion de base de données et pour lesquelles les techniques classiques d'analyse ne peuvent plus s'appliquer [5]. Elles sont caractérisées à travers 3 dimensions, les 3 « V » (volume, variété et vélocité) [6] auxquelles on peut ajouter 2 « V » (véracité et valorisation) dans un contexte de réutilisation de données [7] :

- Volume : un nombre élevé d'individus statistiques (nombre de lignes), de variables (nombre de colonnes), de modalités pour les variables qualitatives (les codes diagnostics de la Classification Internationale des Maladies version 10 – CIM-10 – par exemple), de mesures d'un même paramètre dans le temps, de tables et de relations.
- Variété : les sources et les types des données sont hétérogènes, certaines données, comme les images ou le texte libre, sont non structurées.

- Vélocité : les données sont enregistrées très fréquemment (par exemple une mesure du risque cardiaque) ou fréquemment mises à jour.
- Véracité : il est difficile de contrôler la qualité des données recueillies en routine et le grand nombre d'individus pose la question de l'interprétation de la significativité des résultats.
- Valorisation : l'extraction d'informations pertinentes depuis des bases de données massives demande des techniques de gestion, d'extraction et d'analyse spécifiques.

Les « *Big data* » peuvent être des données recueillies en routine (le plus souvent) ou des données recueillies expérimentalement. Parmi les données massives de routine, on peut citer par exemple les données recueillies par les applications de santé ou les bases de données nationales du Programme de Médicalisation des Systèmes d'Information (PMSI) : cette base de données contient les informations standardisées de plus de 28 millions de séjours pour ce qui est du seul secteur Médecine Chirurgie Obstétrique (MCO) sur l'année 2017. Dans ce cadre, les données sont considérées comme massives du fait du grand nombre d'individus statistiques. Parmi les données recueillies expérimentalement, on peut citer le champ des « *-omics* », qui regroupe notamment la génomique, la protéomique ou l'étude du métabolisme. Dans ce cadre, les données sont massives du fait du grand nombre de variables.

Il peut parfois exister une confusion entre les notions de données massives et de réutilisation de données. En effet, les données massives peuvent faire l'objet d'une réutilisation et une grande partie des études de réutilisation de données s'appuie sur des données massives.

1.3 Définition du Machine Learning

Le terme de *machine learning* (apprentissage automatique) concerne la conception et l'analyse d'algorithmes permettant aux ordinateurs d'extraire (« d'apprendre ») automatiquement des informations à partir de données [8]. Il s'agit de développer un modèle à partir de données d'entrainement, puis de l'utiliser par la suite sur d'autres données. On retrouve plusieurs types d'algorithmes d'apprentissage :

- **apprentissage supervisé** : il concerne des tâches d'explication et de prédiction. L'algorithme « apprend » à prédire une variable en comparant les résultats obtenus par son modèle et les résultats attendus.
- **apprentissage non supervisé** : aucune variable n'est à prédire, on ne connait pas *a priori* les catégories dans lesquelles l'algorithme doit classer les données. L'algorithme identifie sans *a priori* des relations entre variables (corrélations) ou entre individus (proximité, *clusters*, etc.). Ce sera le cas notamment pour des modèles dont le but est de trouver des informations sur la structure de données sans faire d'hypothèse *a priori*.
- **apprentissage par renforcement** : ce type d'algorithme apprend des actions pour maximiser une ressource (des points, de l'argent) au cours du temps. Contrairement à l'apprentissage supervisé où le modèle ne se compare aux données réelles qu'à la fin de sa tâche de prédiction, lors de l'apprentissage par renforcement, le modèle réalise une action, reçoit une récompense puis accomplit une nouvelle action en fonction de son état courant. Ce type d'approche ne s'entend que dans des environnements déterministes, dans lesquels il est possible de « calculer » la récompense et d'envisager une liste

limitée d'actions possibles (ex : jeux vidéo, jeux de société, calculs de trajectoires, etc.).

On peut enfin également définir **l'apprentissage par transfert** : ce type d'apprentissage consiste à transférer des connaissances acquises par une première tâche vers d'autres tâches. Ce type d'approche peut tirer parti des trois approches précédemment décrites.

1.4 Définition du *Word Embedding*

Le *word embedding* (« plongement de mots » en français) est un ensemble de méthodes de *machine learning* qui permettent d'obtenir des représentations vectorielles des mots d'un texte. Ces représentations (*embeddings*) sont construites de manière à avoir une qualité sémantique, c'est-à-dire que des termes ayant une représentation similaire doivent avoir une signification similaire. Il est possible d'utiliser des *embeddings* déjà construits et mis librement à disposition (GloVe, Google News) [9,10], ou d'en construire de nouveaux à partir de corpus généraux comme Wikipédia ou ou de corpus spécifiques à un champ d'activité comme par exemple des courriers médicaux. Les *embeddings* peuvent ensuite être utilisés pour différentes tâches de traitement automatique du langage naturel, par exemple pour des outils de traduction [11] ou de reconnaissance vocale [12]. Les embeddings construits à partir de corpus spécifiques permettent de mieux capturer les relations sémantiques entre les mots, mais ne sont pas forcément meilleurs pour les tâches de traitement du langage [13].

2 Contexte

2.1 Principales bases de données de santé réutilisables en France

La quantité et la variété des données médicales recueillies en routine est en augmentation du fait de la dématérialisation du dossier patient. De plus, en France, il existe un recueil centralisé d'informations médicales structurées dans un but de facturation pour les prises en charges ambulatoires et hospitalières.

2.1.1 Le Système National des données de Santé

Le Système National des Données de Santé (SNDS) a été créé en avril 2017 pour simplifier et contrôler l'accès aux données médicales. Il combine les données du Système National d'Information Inter-Régimes de l'Assurance Maladie (SNIIRAM), du PMSI et du Centre d'épidémiologie sur les causes médicales de décès (CépiDC). Un des intérêts principaux du regroupement de ces bases de données et qu'un patient dispose dans le SNDS du même numéro d'identification anonyme pour l'ensemble des données, ce qui n'est pas le cas si l'on prend les bases séparément. L'Institut National des Données de Santé (INDS), fait office d'accompagnant et de guichet unique pour les demandes d'utilisation du SNDS.

Le SNIIRAM est une base de données constituée et tenue par l'Assurance Maladie de la Sécurité Sociale. Cette base de données regroupe les données de facturation ambulatoires (telles que les consultations et actes externes, et la délivrance des médicaments sur ordonnance en pharmacie). Cette base de données contient aussi des données de mortalité, ce qui est obtenu par croisement les données du CépiDC.

La base nationale du PMSI, constituée de données envoyées par les hôpitaux et gérée l'Agence Technique de l'Information sur l'Hospitalisation (ATIH) contient les informations relatives à la facturation hospitalière. Cette base de données permet le financement des établissements de santé par l'Assurance Maladie de la Sécurité Sociale. Elle découpe les activités d'hospitalisation en 4 champs : MCO, Soins de Suite et de Réadaptation (SSR), Hospitalisation à Domicile (HAD) et psychiatrie. Pour le champ MCO, l'individu statistique est un séjour hospitalier. Les données recueillies incluent principalement un diagnostic principal (la raison qui a motivé le séjour), codé selon la CIM-10, les autres diagnostics ayant eu un impact sur le séjour, aussi codés en CIM-10, les actes médicaux thérapeutiques et diagnostiques codés selon la Classification Commune des Actes Médicaux (CCAM) et certaines informations démographiques. L'ensemble des informations du séjour est synthétisée (« groupée ») dans un code de Groupe Homogène de Malade (GHM). Par exemple, le GHM 04M081 correspond à un séjour pour bronchopneumopathies chroniques, de niveau de sévérité 1. Le dernier caractère du code correspond au niveau de sévérité, si on le tronque, on obtient la racine du GHM, ici 04M08 correspond aux bronchopneumopathies chroniques quelle que soit la sévérité. Enfin, si l'on prend uniquement les deux premiers caractères, on obtient la Catégorie Majeure de Diagnostic (CMD), ici 04 correspond aux affections de l'appareil respiratoire.

2.1.2 Les autres sources de données

Chaque établissement de santé dispose aussi des informations des dossiers médicaux informatisés de ses patients. Ces données peuvent être regroupées et homogénéisées dans le cadre d'entrepôts de données de santé.

Il existe des données « *open data* » (données « ouvertes ») qui sont mises à disposition du public. Ce sont des données qui doivent être [14] :

- accessibles : n'importe qui peut y accéder et leur support doit permettre leur diffusion, en général il s'agit d'un support numérique,
- évaluables : il doit être possible de juger de leur qualité et leur fiabilité,
- intelligibles : elles doivent être compréhensibles par les personnes qui y accèdent,
- utilisables : leur format doit permettre leur réutilisation.

Notamment, le site du gouvernement http://data.gouv.fr met à disposition des données qui permettent par exemple d'enrichir des données de santé recueillies par ailleurs avec des informations géographiques.

2.1.3 Le Health Data Hub

A terme, le « *Health Data Hub* » [15] a pour ambition de faciliter l'accès à l'ensemble des sources de données citée ci-dessus. Il permettrait notamment de simplifier la mise en relation des données issues d'entrepôts de données de santé avec les bases du SNDS ou de partager des outils d'analyse des données.

2.2 Exemples d'utilisation d'*embeddings* en santé

Dans le domaine biomédical, les *embeddings* peuvent être utilisés pour l'analyse du texte libre, notamment en en extrayant des données structurées. Ils ont été utilisés pour retrouver automatiquement à partir de courriers, de notes ou de comptes rendus des entités nommées (noms de personnes, d'entreprises, de lieux, de médicaments,

etc) [16,17], des synonymes [18], des relations (entre un médicament et une pathologie par exemple) [19], des informations médicales (un diagnostic, un antécédent, le statut tabagique, etc) [20], ou pour expliciter des abréviations [21].

Le principe du *word embedding* s'est aussi élargi à d'autres champs d'application que le traitement automatisé du langage et a été appliqué à d'autres objets que des mots. Dans le domaine des « *omics* », BioVec [22] et dna2vec [23] sont des modèles développés à partir ***d'embeddings*** de séquences biologiques (des séquences de nucléotides ou de protéines). Choi et al. ont créé des *embeddings* de codes médico-administratifs (diagnostics, examens de laboratoire et médicaments) à partir de données de facturation dans le but d'obtenir une représentation des concepts médicaux [24]. Ils montrent que leurs *embeddings* permettent de retrouver les relations sémantiques entre les différents codes. Farhan et al. ont eux créé un *embedding* à partir des dossiers de santé informatisés de la base MIMIC-III [25], qui regroupe des informations sur les séjours en soins intensifs, pour prédire les futurs codes diagnostics des patients [26].

3 Principales méthodes utilisées dans cette thèse

3.1 Réseau de neurones artificiels

Les réseaux de neurones artificiels sont un type de modèle informatique dont la construction est inspirée du fonctionnement des neurones biologiques. Ils sont classiquement utilisés en statistique comme une méthode d'apprentissage supervisé maris peuvent également être utilisés pour des tâches non supervisées. Un perceptron, ou neurone artificiel est la modélisation mathématique d'un neurone biologique. La Figure 1 illustre le fonctionnement d'un perceptron.

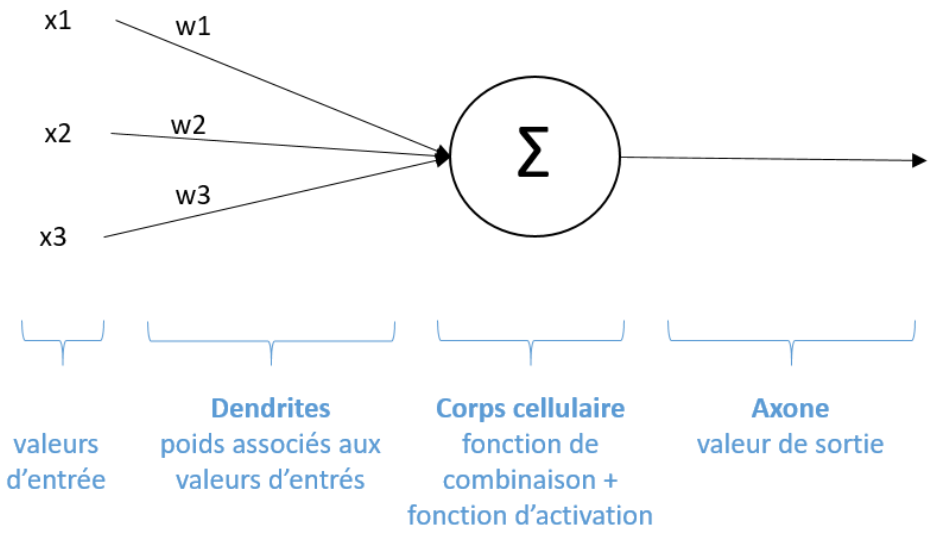


Figure 1 Schéma d'un perceptron

Le perceptron prend une série de valeur en entrée (x1 à x3 sur le schéma). Chaque valeur se voit associer un poids (w1 à w3). Le perceptron combine ces valeurs selon une fonction de combinaison, souvent une combinaison linéaire pondérée. Il applique ensuite une fonction d'activation, qui peut par exemple être la fonction identité (il n'y a

alors pas de modification) ou une fonction seuil (le résultat est 1 si la valeur d'entrée est supérieure au seuil, 0 sinon). Enfin, il retourne la valeur obtenue comme sortie.

Un réseau de neurones est une mise en réseau de plusieurs perceptrons, la valeur de sortie des premiers perceptrons devenant la valeur d'entrée des perceptrons suivants. Ils sont organisés en couches successives. Par convention, la première couche est appelée couche d'entrée (« *input layer* » en anglais), la dernière couche est appelée couche de sortie (« *output layer* » en anglais) et toutes les couches intermédiaires sont appelées des couches cachées (« *hidden layers* » en anglais). La Figure 2 illustre un exemple de réseau de neurones.

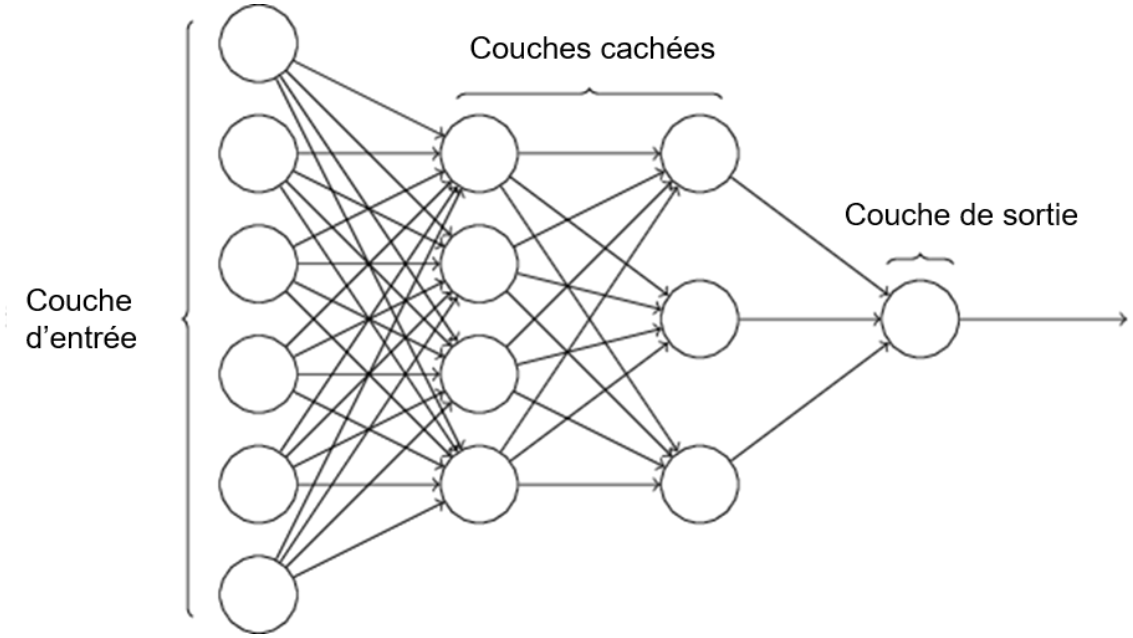


Figure 2 Réseau de neurones, traduit de [27]

3.2 Word2Vec

Word2Vec est une méthode de *word embedding* développés par Mikolov et al. [28,29] qui existe en deux versions, *Continuous Bag Of Words* (CBOW) et *Continuous Skip-gram*. Cette méthode est basée sur l'hypothèse distributionnelle, selon laquelle des mots qui apparaissent dans des contextes similaires partagent des significations similaires [30]. La version CBOW entraîne un réseau de neurones à deux couches (une couche cachée et une couche de sortie) à prédire un mot à partir de son contexte, alors que la version Skip-gram entraine un réseau de neurones à deux couches à prédire le contexte d'un mot donné. Selon Mikolov, *Skip-gram* est plus efficace sur les mots peu fréquents, mais plus lent que CBOW [31,32].

McCormick détaille le fonctionnement de la version *Skip-gram* de *Word2Vec* [33]. Le réseau de neurones correspondant est schématisé sur la Figure 3.

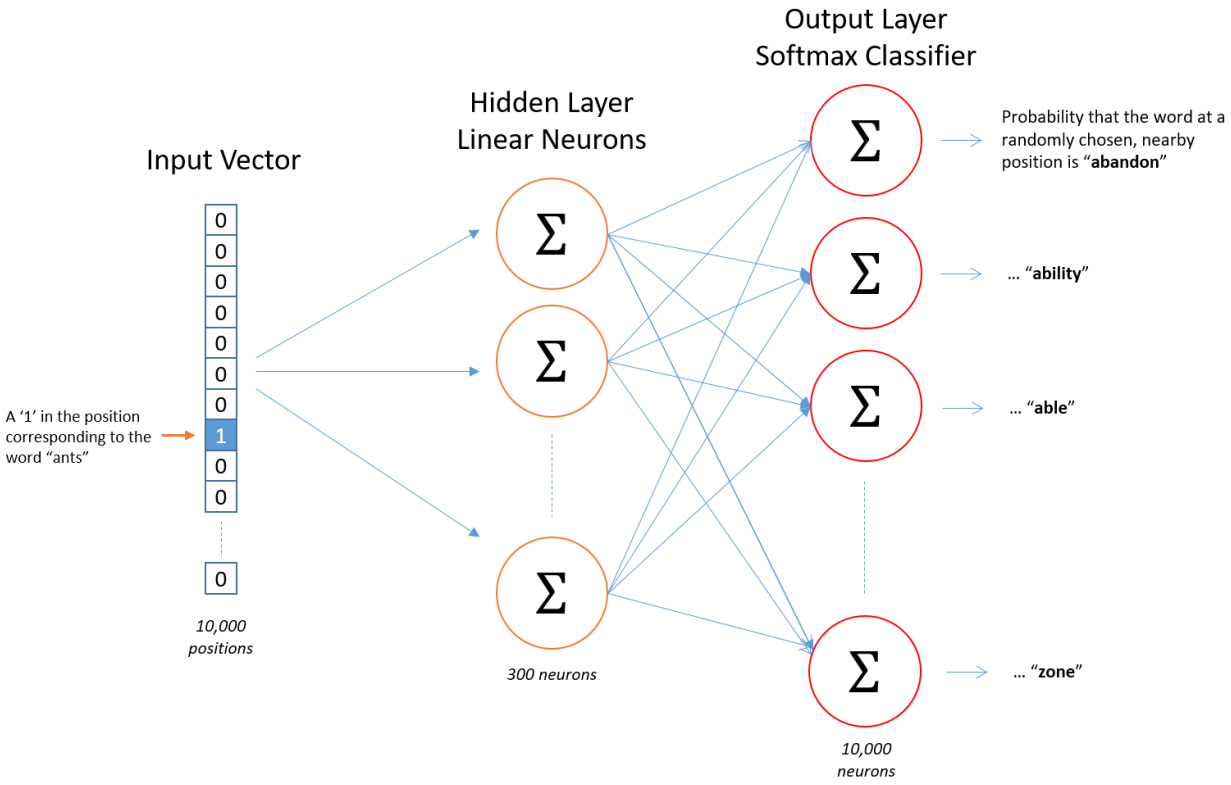


Figure 3 Schéma de l'architecture Skip-gram du modèle Word2Vec, selon McCormick [33]

Lors de l'entraînement du réseau de neurones, tous les mots du texte sont un à un passés dans le modèle en tant que « mot d'entrée » (« *input word* » en anglais). Le contexte de ce mot d'entrée est défini comme les n mots précédents et les *n* mots suivants au sein de la même phrase, *n* étant une fenêtre que l'on fixe *a priori*. McCormick prend en exemple le moment où le mot d'entrée serait « *ant* » (fourmi) dans un texte qui comporterait 10 000 mots différents.

La première étape, nécessaire pour pouvoir passer ce mot dans le réseau de neurones, est de le transformer en une information informatiquement intelligible. Le mot est transformé en un vecteur «*one-hot*» (ou vecteur « *1* parmi *n* »). Il s'agit d'un vecteur de la taille du vocabulaire du texte (10 000 ici) dont l'ensemble des valeurs sont *0*, sauf celle qui correspond à la position du mot d'entrée dans le vocabulaire. On obtient alors le vecteur d'entrée (*input vector* en anglais).

La « couche cachée » est une matrice qui contient autant de lignes que de mots dans le vocabulaire et un nombre prédéfini de colonnes, qui correspondra à la taille de *l'embedding*. Ce nombre est fixé à 300 dans l'exemple, les 300 neurones que l'on voit sur la Figure 3 correspondent aux 300 colonnes de la couche cachée. Chaque ligne contient en fait une représentation vectorielle d'un mot du vocabulaire selon un nombre choisi de dimensions. Dans l'exemple, le mot « *ant* » est le septième mot du vocabulaire, la septième ligne de la couche cachée contient donc un vecteur de taille 300 qui est la représentation du mot « *ant* ».

Le modèle multiplie le vecteur d'entrée par la matrice couche cachée. En pratique, le résultat de cette multiplication est un vecteur qui est la représentation sur 300 dimensions du mot « *ant* », et c'est ce qu'on retrouve en sortie de la couche cachée. En effet, le vecteur d'entrée ne contient que des 0, excepté un 1 en septième position.

Le résultat de la multiplication avec la matrice couche cachée est alors la 7ème ligne de cette matrice, qui est la représentation du mot « *ant* » en 300 dimensions.

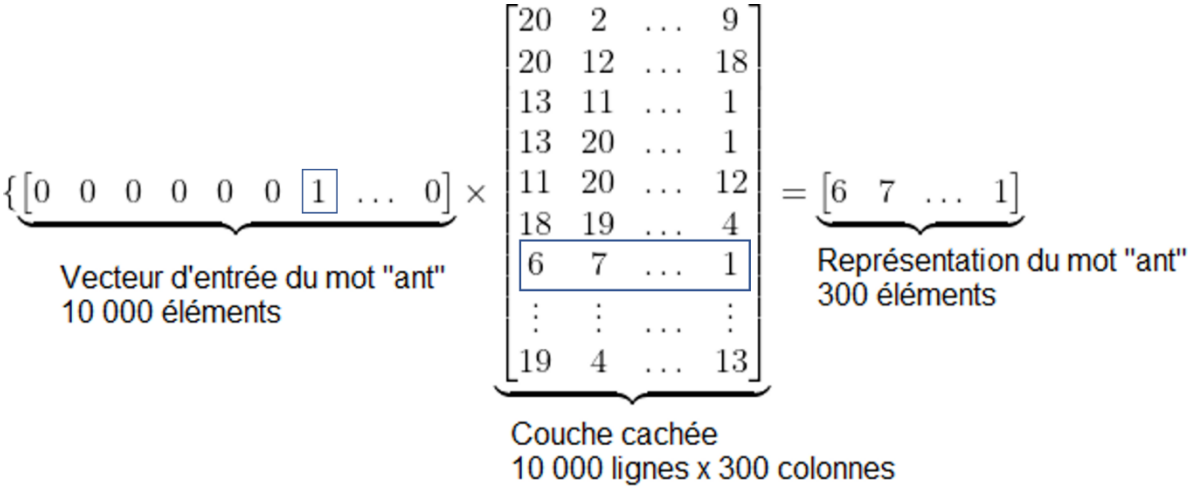


Figure 4 Projection du vecteur d'entrée sur la couche cachée, inspiré de McCormick [33]

La couche de sortie contient autant de neurones que de mots dans le vocabulaire. Chaque neurone contient un vecteur de poids associé à un mot du vocabulaire et donne en sortie la probabilité que ce mot soit un mot contexte du mot d'entrée après une transformation par la fonction *softmax*.

Le modèle multiplie le vecteur obtenu à la sortie de la couche cachée par chacun des vecteurs poids puis applique une transformation avec la fonction *softmax* pour obtenir les probabilités pour chaque mot d'être dans le contexte du mot d'entrée. La fonction *softmax* est une fonction qui prend un vecteur de nombre réels et renvoie un vecteur de réels entre 0 et 1 de somme 1 et qui peut donc être utilisé comme une distribution de probabilités. La Figure 5 illustre ce mécanisme pour l'obtention de la probabilité que le mot « *car* » (voiture) soit compris dans le contexte du mot « *ant* ».

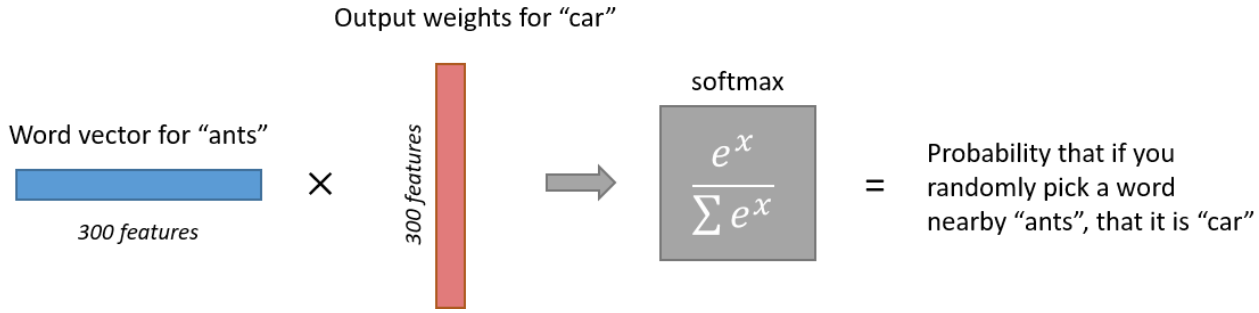


Figure 5 Effet de la couche de sortie du modèle Skip-gram, selon McCormick [33]

Cette probabilité calculée est comparée à une valeur cible (1 si le mot est effectivement dans le contexte du mot d'entrée, 0 sinon), puis il y a une modification des poids des vecteurs de la couche de sortie, puis de la matrice de la couche cachée par rétropropagation du gradient. Le modèle passe alors au mot suivant du texte. A la fin, la couche cachée contient une représentation vectorielle sur 300 dimensions pour chacun des mots du vocabulaire, c'est cette couche cachée qu'on récupère en tant qu'*embedding*. Par construction, les mots qui ont des contextes similaires ont des représentations similaires, et selon l'hypothèse distributionnelle, devraient avoir des significations proches.

Chia détaille aussi le fonctionnement de la version *Skip-gram* de *Word2Vec*, mais propose en plus un exemple à implémenter en *Python* et des *Googles Sheets* qui permettent de suivre les itérations du modèle de façon interactive [34].

La mise à jour de tous les poids du réseau de neurones, à chaque passage du modèle, sur chaque mot du texte, est en pratique extrêmement lourd. Deux variantes, appelées « *negative sampling* » (échantillonnage négatif en français) et « *hierarchical softmax* » (fonction *softmax* hiérarchique en français) ont été développées pour pallier ce problème [29]:

Dans le cas du *negative sampling*, au lieu de recalculer les poids de tous les mots, on ne le fait que pour les mots qui sont effectivement dans le contexte du mot d'entrée et une sélection aléatoire de 5 à 20 mots qui ne sont pas dans le contexte. Ces 5 à 20 mots sont les échantillons négatifs. Ils sont sélectionnés en fonction de leur fréquence d'apparition dans le texte, les mots les plus fréquents ayant plus de chance d'être sélectionnés. La variante *hierarchical softmax*, permet une accélération majeure du temps de calcul au prix d'une légère perte en précision [35]. L'idée est de construire un arbre binaire dont les feuilles terminales sont les mots du vocabulaire. Des poids sont associés aux branches de l'arbre de telle façon que si on parcoure l'arbre de la racine à une feuille terminale, la multiplication du poids des branches parcourues corresponde à la probabilité pour le mot associé à la feuille terminale.

3.3 Arbres de classification et forêts aléatoires

L'apprentissage par arbre de classification est une méthode non paramétrique et supervisée qui se base sur l'utilisation d'un arbre de décision comme modèle prédictif [36]. A l'origine, un arbre de décision est la représentation graphique d'un ensemble de choix sous forme d'un arbre. Les feuilles représentent les décisions finales et les embranchements (les nœuds) représentent les choix menant à ces décisions. Lorsqu'il est utilisé en tant que modèle prédictif, il met en relation une variable à prédire avec un ensemble de variables explicatives. On parlera d'arbre de classification si la variable à prédire est binaire ou qualitative, et d'arbre de régression si la variable à prédire est quantitative. Cette méthode segmente un échantillon d'apprentissage en sous-groupes, puis chacun des sous-groupes en sous-groupes, etc., jusqu'à obtention de sous-groupes le plus homogènes possible en leur sein (on minimise la variance intra-groupe) et distincts les uns des autres (on maximise la variance intergroupe) au regard de la variable à prédire. Chaque segmentation est faite en sélectionnant la variable explicative qui permet le mieux de réaliser cette opération. Les sous-groupes terminaux forment les feuilles de l'arbre et les embranchements (ou nœuds) correspondent aux segmentations successives.

Un des avantages de l'apprentissage par arbre de classification est qu'il est facilement interprétable, par simple visualisation de l'arbre. Il permet de retrouver simplement quelles ont été les règles de classification sélectionnées par l'apprentissage. Ces règles peuvent aisément être enregistrées sur disque (par exemple en XML) et automatiquement transcrites dans différents langages (par exemple, une formule Excel, du code R, du code SQL, etc.). En outre, ces règles peuvent également être filtrées et réorganisées par des experts [37,38]. Cependant, un désavantage majeur de ce type d'apprentissage est que la construction de l'arbre de classification dépend de la variable explicative choisie au rang 1 (à la première segmentation). Cela entraine une instabilité des arbres lorsqu'ils sont construits sur des échantillons ne différant que par quelques individus pouvant être drastiquement différents [39]. On peut noter

cependant que cette instabilité des règles produites n'altère cependant pas la qualité de la prédiction statistique, car elle s'observe justement quand les variables explicatives sont fortement corrélées entre elles.

La méthode des forêts aléatoires (*random forest* en anglais) a été développée pour compenser ce défaut de stabilité [40]. L'idée est de créer un grand nombre d'échantillons partiellement indépendants à partir du même échantillon d'apprentissage. Pour cela, on sélectionne de façon aléatoire non seulement les individus à inclure dans chaque échantillon, mais aussi les variables. Un arbre de décision classique est ensuite construit pour chaque échantillon. Le résultat final du modèle complet est un simple vote : la valeur retenue est celle qui a été retournée par la majorité des arbres. Cette méthode est surtout intéressante pour identifier une liste de variables d'intérêt. Inversement, on perd la possibilité de visualisation des règles de classification que l'on obtenait avec les arbres simples, et on perd également la faculté de « détecter des interactions », qui était la vocation originelle des arbres.

4 Objectifs de ce travail

Notre objectif principal était, dans le cadre d'un apprentissage par transfert, de construire un *embedding* des diagnostics de la 10ème Classification Internationale des Maladies à partir de la base medico-administrative du PMSI, puis de transférer cet *embedding* à un nombre restreint d'individus pour étudier son intérêt dans une tâche de prédiction.

Nos objectifs secondaires étaient de comparer les versions *CBOW* et *Skip-gram* des *embeddings* et d'expérimenter sur les effets de la modification des paramètres techniques des modèles lors de la construction de ces *embeddings*.

Abstract en Anglais

1 Introduction

In the biomedical field, the data reuse supplies data describing many patients on limited variables types, whereas classical epidemiological studies allow to describe many variables for fewer patients. Transferring representations built from the data reuse towards epidemiological studies could improve the quality of modelling in these studies.

The word embedding methods allow to build low-dimensional representations of objects and can be applied to medico-administrative codes.

We aimed to build a Word2Vec embedding of the 10th International Classification of Diseases codes from a national medico-administrative database, then to transfer this representation towards a study with a limited number of patients.

2 Material & methods

We computed diagnosis embeddings, using the Continuous Bag Of Words (CBOW) and the Skip-gram versions of Word2Vec, from 21.7 million inpatient medical claims in 2008 in France. We transferred these embeddings in two epidemiological studies and compared their performances with those obtained with bare diagnostic codes only. The outcomes studied were the predictions of 30 days readmission and in-hospital death, after a hospitalization for chronic obstructive pulmonary disease.

3 Results

The readmission predictive models were trained on 770 inpatient stays. The receiver operating characteristic area under the curve (ROC AUC) of the CBOW, the Skip-gram and the bare diagnoses based models were 0.548 (95% confidence interval: [0.532; 0.564]), 0.568 [0.552; 0.584] and 0.547 [0.531; 0.563], respectively.

The death predictive models were trained on 380 inpatients stays. The ROC AUC of the CBOW, the Skip-gram and the bare diagnoses based models were 0.862 [0.848; 0.877], 0.868 [0.854; 0.883] and 0.840 [0.824; 0.856], respectively.

4 Discussion & conclusion

Using an embedding built from a large medico-administrative database in an epidemiological context with fewer patients allowed to improve prediction tasks, demonstrating the possibility of transferring a representation built from big data to an epidemiological study context.

Article en anglais

1 Introduction

The term “word embedding” refers to several methods which aim to transform raw textual data into structured real valued vectors. They were originally developed to help with natural language processing (NLP) tasks, such as machine translation [11] or speech recognition [12]. In the biomedical field, they have been used for medical synonym extraction or medical abbreviation disambiguation [21], among other applications [13].

The Word2Vec method, developed by Mikolov et al. [28,29], is based on the distributional hypothesis: words that are used in a similar context tend to share a similar meaning [30]. Word2Vec exists in 2 versions: continuous bag of words (CBOW) and skip-gram. CBOW trains a 2-layers neural network (NN) to predict the context of a given word, whereas Skip-gram trains a 2-layers NN to predict a word of a given context. Another difference between both versions is that Skip-gram takes into account the proximity of a context word, whereas CBOW does not. For both versions, the output layer is discarded and the hidden layer is kept as the word embedding. Words that share a similar context share a similar representation in this hidden layer, so according to the distributional hypothesis, words that share a similar representation should share a similar meaning. According to Mikolov, Skip-gram is better for infrequent words than CBOW [31,32].

The principle of Word2Vec has been derived to other fields and applied on objects other than words. For example, in bioinformatics, BioVec [22] and dna2vec [23] embed biological sequences (such as DNA or proteins). Choi et al. embed medico-administrative data (diagnostic codes, laboratory results and drug prescriptions) from a private medical claims dataset to obtain a global low-dimensional representation of medical concepts [24]; They have showed that such embeddings are able to capture semantic relations between codes. Similarly, Farhan et al. embed diagnostic codes, laboratory results, drug prescriptions, medical conditions et symptoms from the intensive care MIMIC-III database [25] to predict the future diagnosis of a patient [26]. These studies highlight the potential value of using these embeddings in a medical data context. Nevertheless, these studies did not assess the interest of transferring these representations.

On the first hand, data reuse (e.g. electronic health records (EHR) data warehouses and national claims databases) brings data describing a large amount of patients on limited types of variables, and, on the other hand, classical epidemiological studies (with biobanks) allow to describe a high number of variables for well phenotyped small amount of patients. It is reasonable to question the possibility of transferring representations constructed from the reuse of data (with a large number of patients) towards contexts where the number of patients is reduced in order to improve the quality of modelling on these classic epidemiological studies.

1.1 Objective

We aimed, with a transfer learning approach, to build a Word2Vec embedding of the 10th International Classification of Diseases (ICD10) codes from a national medico-administrative database then to transfer this representation towards a study with a limited number of patients to assess its interest for a prediction task. Secondarily, we aimed to compare the performances of CBOW and Skip-gram versions of the embedding, and to evaluate the effects of parametrization for these models.

2 Material and methods

2.1 Data

We used data from the French « PMSI », which is the national comprehensive inpatient stay claims database, from 2008 to 2009. A statistical individual consists in one inpatient stay information, such as the dates of beginning and discharge, the list of diagnoses (10th International Classification of Diseases - ICD10) or the life status at discharge (CNIL authorization number 2049035). Different inpatient stays of a given unique patient are considered as independent records.

2.2 Embeddings' computing

For each of the 21,649,708 inpatient stays in 2008 in France, we extracted and randomly shuffled the diagnoses list. We then computed two main diagnostic embeddings, using the CBOW and the Skip-Gram versions of Word2Vec, with negative sampling and no hierarchical softmax. We used an embedding dimension of 100 and a context window which included all the other diagnoses of the stay.

As a secondary analysis, we studied the effects of varying the embedding parameters:

- To assess the effect of the size of embedding, we computed 200 dimensions embeddings.
- To assess the effect of the training sample size, we computed the diagnosis embeddings from 1 million stays and from 50 000 stays (roughly 1/20 and 1/400 of the number of stays in 2008, respectively).

We monitored the computation time for all the embeddings. The embeddings were computed using the gensim library [41] on Python version 3.7.6. The machine was a Dell PowerEdge C6220 II, with a biprocessor Intel Xeon E5-2620v2 (2,1GHz, 6C, Cache 15Mo, 7,2GT/s QPI, 80W, Turbo, HT) 32 Go RAM.

2.3 Embeddings' description

In this descriptive step, we illustrated the ability of the main embeddings to regroup similar diagnoses together. Firstly, we retrieved the 5 closest neighbors of the 5 most frequent diagnoses, according to the embeddings. The distance between 2 diagnostic codes was defined as the cosine similarity between the vector representations of these codes. Secondly, graphical representations were built using the t-distributed stochastic neighbor embedding (t-SNE) algorithm [42] to depict the embeddings in two dimensional plots. To visually compare on this representation the groups obtained with the main embeddings and those of the ICD10 classification, the diagnoses were

colored according to ICD10 chapters. The t-SNE representation was computed using the tsne R package [43].

2.4 Embeddings' evaluation

We evaluated the performance of the main embeddings in several epidemiological case-control studies. Then, we conducted several secondary analyses for all the secondary embeddings.

In each of these studies, predictive models were built from the main embeddings, comparatively with the use of bare diagnostic codes. To obtain a vector representation of one stay, we used the mean of the diagnoses vectors of the corresponding stay. Each dimension of the stay vector was then used as a variable in the predictive model. In the model built from bare diagnosis, each diagnosis was considered as a binary variable.

Several case-control studies were built with the following characteristics:

- Two types of cases were considered: (i) The cases were patients with readmission within 30 days (in 2008 or 2009) following an hospitalization for chronic obstructive pulmonary disease (COPD) in 2008, or (ii) the cases were death during such an hospitalization for COPD in 2008. For each case, we randomly picked an inpatient stay which did not present a rehospitalization (which did not end with the death of the patient, respectively) to be "control" with a 1:1 ratio. The COPD hospitalizations were identified thanks to the algorithm suggested by the French Institute for Public Health Surveillance [44]. The deaths were identified thanks to the discharge code in the PMSI. A readmission was defined as a second hospitalization within 1 to 30 days after a discharge to home. Hospitalizations for organ transplant, chemotherapy, dialysis, irradiation, curietherapy or for which the patient did not come from home were not included.
- 80% of these stays were included in the learning sample, and 20% were kept for the test sample.
- As we wanted to work in the context of a classical epidemiological study, the learning sample for the predictive models had to be relatively small: so, after checking the occurrences of each case, we chose to keep 1/25th of the initial sample. Secondly, to assess the effect of the learning sample size, we also built predictive models keeping 1/5th of the initial sample, then keeping the whole initial sample. The controls were still randomly picked with a 1:1 ratio. Whatever the size of the learning sample, the test sample considered was always the same and corresponded to the initial 20%.

Then, a supervised model was conducted in order to predict the outcome (i.e. cases). We used a classification random forest. Five hundred trees were computed and the number of candidate variables at each split was the square root of the total number of variables.

2.5 Evaluation of the model

The performance of the predictive model was estimated using the receiver operating characteristic area under the curve (ROC AUC) of the predicted value and the corresponding 95% confidence interval (CI). The ROC AUC and 95%CI were computed. All the analyses were conducted on R version 3.6.3 [45], with packages randomForest [40] and pROC [46].

3 Results

3.1 Embeddings' computing

In France in 2008, we found 16,254 distinct diagnostic codes. We found 11,904 diagnostic codes in the 1 million stays sample, and 6,370 in the 50,000 stays sample. For each stay, there were between 1 and 63 diagnoses (1 and 54 diagnoses and 1 and 30 diagnoses for the 1 million and the 50,000 stays samples respectively), with a median of 2 (Q1 :1, Q3 : 4). The CBOW main embedding took 7 minutes 22 seconds to compute and the Skip-Gram main embedding took 11 minutes 30 seconds to compute. The computations times for the secondary embeddings can be found in Supplementary Table 1.

3.2 Embeddings' description

The most frequent diagnostic code was I10 (Essential – primary – hypertension). Its closest neighbors according to the CBOW embedding were I11.9 (Hypertensive heart disease without – congestive - heart failure), I11.0 (Hypertensive heart disease with – congestive - heart failure), I15.0 (Renovascular hypertension), I20.9 (Angina pectoris, unspecified) and I15.9 (Secondary hypertension, unspecified). There were some differences with the results of the Skip-gram embedding, where the closest neighbors of I10: I11.9, I20.9 and I11.0 are still in the list, but I15.0 and I15.9 were replaced by H40.9 (Glaucoma, unspecified) and I24.8 (Other forms of acute ischemic heart disease). Table 1 presents the most similar diagnoses to the 5 most frequent diagnoses, according to CBOW and skip-gram main embeddings.

Figure 6 shows a graphical representation of the main skip-gram embedding in two dimensions. There is a good pooling of diagnoses for some chapters, like chapter 15 "Pregnancy, childbirth and the puerperium", but still a lot of scattering for other chapters, like chapter 18 "Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified". The main CBOW embedding is represented in Supplementary figure 1.

Table 1 Most similar diagnoses to the 5 most frequent diagnoses, according to CBOW and skip-gram embeddings from 21.7 million inpatient stays (100 dimensions)

Diagnosis	Most similar diagnoses	
	CBOW embedding	**Skip-gram embedding**
I10 Essential (primary) hypertension	**I11.9** Hypertensive heart disease without (congestive) heart failure **I11.0** Hypertensive heart disease with (congestive) heart failure **I15.0** Renovascular hypertension **I20.9** Angina pectoris, unspecified **I15.9** Secondary hypertension, unspecified	**I11.9** Hypertensive heart disease without (congestive) heart failure **I20.9** Angina pectoris, unspecified **H40.9** Glaucoma, unspecified **I11.0** Hypertensive heart disease with (congestive) heart failure **I24.8** Other forms of acute ischaemic heart disease
Z51.1 Chemotherapy session for neoplasm	**Z08.2** Follow-up examination after chemotherapy for malignant neoplasm **Z51.0.0** Radiotherapy session – Preparatory care **Z51.0.1** Radiotherapy session - Irradiation **C78.1** Secondary malignant neoplasm of mediastinum **C79.8** Secondary malignant neoplasm of other specified sites	**Z08.2** Follow-up examination after chemotherapy for malignant neoplasm **Z08.7** Follow-up examination after combined treatment for malignant neoplasm **C79.8** Secondary malignant neoplasm of other specified sites **C78.8** Secondary malignant neoplasm of other and unspecified digestive organs **C79.2** Secondary malignant neoplasm of skin
Z51.0.1 Radiotherapy session - Irradiation	**Z51.0.0** Radiotherapy session – Preparatory care **C78.1** Secondary malignant neoplasm of mediastinum **Z51.1** Chemotherapy session for neoplasm **C79.8** Secondary malignant neoplasm of other specified sites **C76.0** Malignant neoplasms - Head, face and neck	**Z51.0.0** Radiotherapy session – Preparatory care **C78.1** Secondary malignant neoplasm of mediastinum **Z51.1** Chemotherapy session for neoplasm **C79.8** Secondary malignant neoplasm of other specified sites **Z08.2** Follow-up examination after chemotherapy for malignant neoplasm
Z49.1 Extracorporeal dialysis	**Z49.0** Preparatory care for dialysis **Z99.2** Dependence on renal dialysis **I77.0** Arteriovenous fistula, acquired **I12.0** Hypertensive renal disease with renal failure **Z49.2.1** Other dialysis - Continuous ambulatory peritoneal dialysis	**I12.0** Hypertensive renal disease with renal failure **Z49.0** Preparatory care for dialysis **Z99.2+0** Dependence on renal dialysis - Hemodialysis **Z99.2** Dependence on renal dialysis **T82.4** Mechanical complication of vascular dialysis catheter
N18.0 End-stage renal disease	**N18.8** Other chronic renal failure **N18.9** Unspecified renal failure **I77.0** Arteriovenous fistula, acquired **Z49.0** Preparatory care for dialysis **Z49.1** Extracorporeal dialysis	**I12.0** Hypertensive renal disease with renal failure **Z49.0** Preparatory care for dialysis **Z99.2** Dependence on renal dialysis **N18.9** Unspecified renal failure **Z99.2+0** Dependence on renal dialysis - Hemodialysis

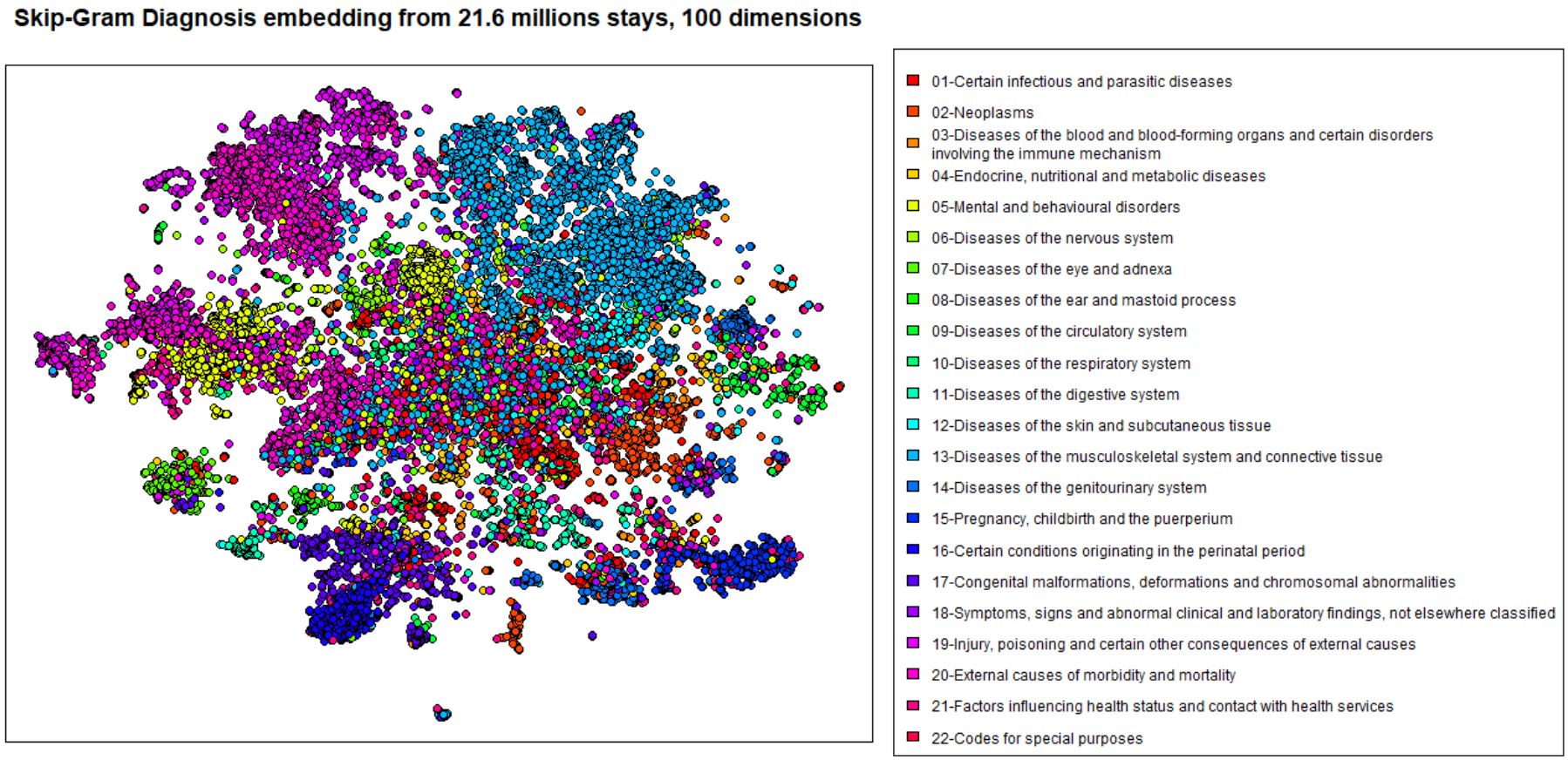


Figure 6 Representation of the skip-gram diagnosis embedding from 21.6 million stays. The reduction from 100 dimensions to 2 was obtained by t-SNE. The diagnoses are colored according to ICD10 chapters

3.3 Embeddings evaluation through predictive models

In 2008 in France, there were 66,758 COPD hospitalizations that ended with a discharge to home (60,845, 91%) or with death (5,913, 9%). Among the stays that ended with a discharge to home, 12,068 (20%) were followed by a readmission within 1 to 30 days.

For the readmission predictive models, our main learning sample comprised 770 stays (385 cases and 385 controls). Our test sample comprised 4,874 stays (2,437 cases). The ROC AUC of the CBOW based random forest was 0.548 (95% CI: [0.532; 0.564]), the one of the Skip-Gram based random forest was 0.568 [0.552; 0.584] and the one of the bare diagnosis based random forest was 0.547 [0.531; 0.563]. The secondary learning samples comprised 3,852 and 19,262 stays (including 1,926 and 9,631 cases, respectively). The ROC AUC of the CBOW based random forest were 0.561 [0.545; 0.577] and 0.582 [0.566; 0.598] respectively. The ones of the Skip-Gram based random forest were 0.567 [0.551; 0.583] and 0.590 [0.574; 0.606] respectively. The ones of the bare diagnosis based random forest were 0.565 [0.549; 0.581] and 0.589 [0.573; 0.605] respectively. These results are illustrated by Figure 7.

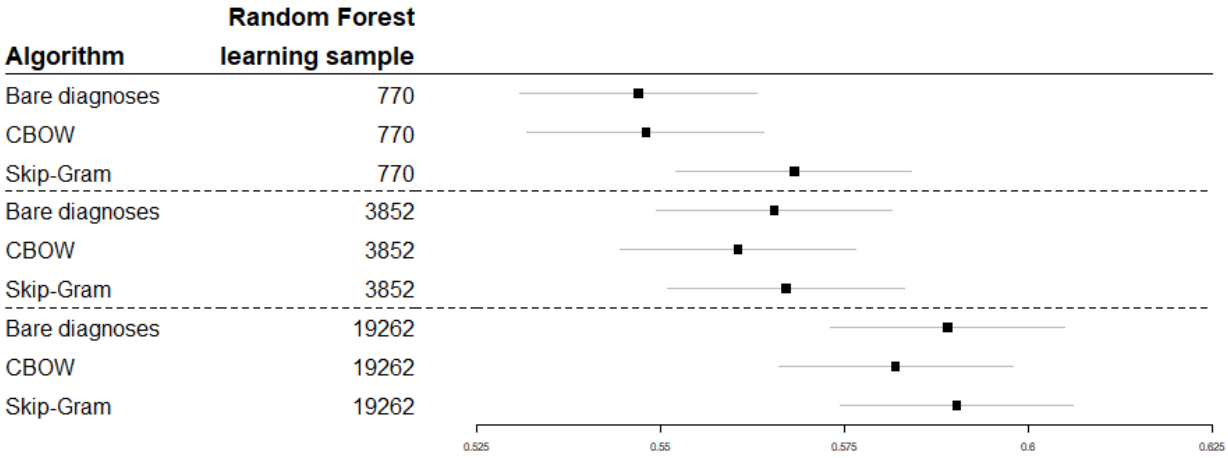


Figure 7 Area under the ROC curve of the readmission prediction random forest. The CBOW and skip-gram models were computed from 21.6 million stays and had 100 dimensions. The stay embedding was the mean of the diagnosis embedding

For the death predictive models, our main learning sample comprised 380 stays (190 cases). Our test sample comprised 2,324 stays (1,162 cases). The ROC AUC of the CBOW based random forest was 0.862 (95% CI: [0.848; 0.877], the one of the Skip-Gram based random forest was 0.868 [0.854; 0.883] and the one of the bare diagnosis based random forest was 0.840 [0.824; 0.856]. The secondary learning samples comprised 9,502 and 1,900 stays (including 4,751 and 950 cases, respectively). The ROC AUC of the CBOW based random forest were 0.874 [0.860; 0.888] and 0.874 [0.859; 0.888] respectively. The ones of the Skip-Gram based random forest were 0.878 [0.865; 0.892] and 0.877 [0.863; 0.891], respectively. The ones of the bare diagnosis based random forest were 0.871 [0.857; 0.885] and 0.882 [0.869; 0.896], respectively. These results are illustrated on Figure 8.

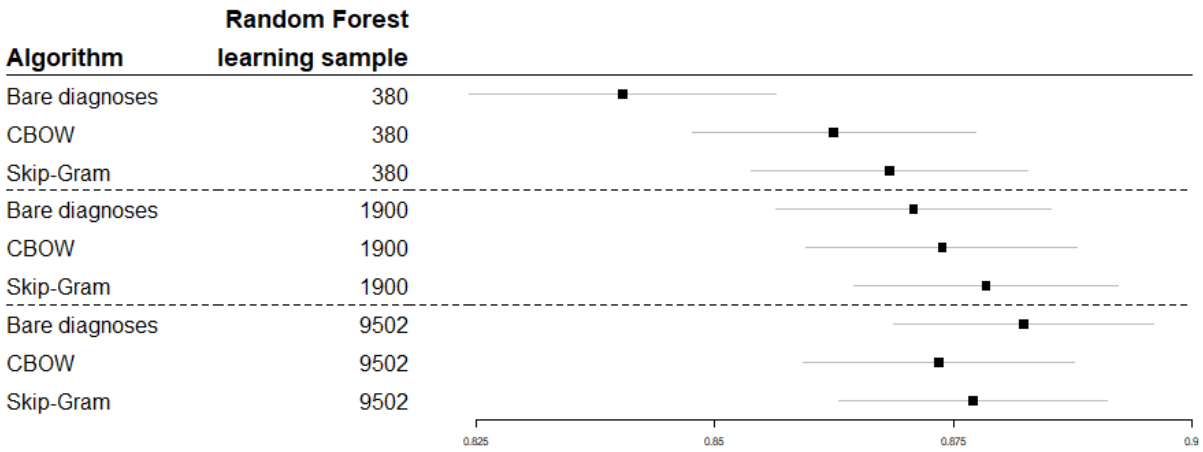


Figure 8 Area under the ROC curve of the death prediction random forest. The CBOW and skip-gram models were computed from 21.6 million stays and had 100 dimensions. The stay embedding was the mean of the diagnosis embedding

As detailed in the appendix, the construction of predictive models from embeddings obtained from the smaller sample size (n=50,000) did not yield similar results. Indeed, in the case of skip-gram with an embedding dimension of 100, the AUC [95% CI] was 0.538 [0.522; 0.555] and 0.793 [0.775; 0.811] for, respectively, the death (sample size n=700) and the readmission (sample size n=380).

All the results of the predictive models based on all embeddings can be found in Supplementary Table 2 for readmission prediction models and in Supplementary Table 3 for death prediction models.

4 Discussion

We built a diagnosis embedding on a national claims database and then transferred this representation to a classical epidemiological study context. Considering a prediction task learned on small samples, for both readmission and death prediction, the Skip-gram representation seems to perform better than the bare diagnoses one. In addition, we obtained a prediction quality close to that obtained from a larger sample size but without using this embedding. The diagnosis embedding built in this work is not dedicated to one type of medical specialty, it is generalist so that it can be transferred to multiple contexts.

However, this difference between models built with or without diagnosis embedding seems to disappear as the learning sample grows. In addition, the Skip-gram models seem to outperform the CBOW models, which was expected as each diagnosis was rather infrequent.

To the best of our knowledge, no transfer learning work using epidemiological databases (be it from data reuse or from studies with data collection by investigators) had been performed, so it is difficult to compare this result specifically.

Some authors have proposed embeddings based on structured medical data which were then used on the same dataset that they were built on. Farhan et al built an embedding inspired by the Word2Vec method to predict 80 diagnoses from a temporal sequence of events [26]. They used a selection of laboratory results, drug prescriptions, diagnostic codes, medical conditions and symptoms from 5,195 critical

care patients of the MIMIC-III database[25].Other works focused on predicting the future of a patient from deep learning representations of EHR: Miotto et al. built a “deep patient” representation that allowed them to predict future diseases of a patient [47]. Choi et al. built a recurrent neural network model to predict the diagnostic codes and drug prescriptions of the next admission of a patient [48].

In comparison, some experimental steps of our work use an embedding built from a sample size close to the learning sample size (e.g. embedding on 50,000 and learning on 9,500): nevertheless, in this case, whatever the model considered, the result obtained from the raw diagnoses seems higher than the result obtained from the diagnosis embedding (see appendices). In our opinion, this result validates the interest of the transfer task from a large database.

Our work takes advantage of the availability of the very large number of data produced routinely, such as those from EHRs or claims databases, and makes it possible to transfer representations to contexts where the data are actively collected with high effort, as well as epidemiological studies carried out on medico-administrative databases but with a lack of statistical power (e.g. rare events in pharmacoepidemiology). It is reasonable to think that these representations could be used in particular for epidemiological studies where predictive models need to be developed, such as for example studies where a propensity score (from a model predicting the presence of the drug of interest) is used. Nevertheless, at this stage, the usefulness of this representation has not been evaluated on an external validation dataset.

More generally, the use of complex representations raises the question of the explicability of models constructed from these representations; indeed, the use of embedding-based models may appear to be a relative contradiction with the models classically carried out in the field of epidemiology, for which the candidate explanatory variables are carefully selected by experts. In our opinion, these two approaches could be combined, for example by transforming with an embedding the adjustment variables previously chosen by an expert.

Moreover, the dissemination of such a generic embedding raises several questions. Firstly, in terms of confidentiality, it is not possible to re-identify patients with the embedding; and secondly, the sharing of our diagnosis embedding raises the question of its interoperability, particularly if it would be reused in the context of classic cohorts in epidemiology; the use of the ICD-10 nevertheless ensures its interoperability with most medico-administrative databases. Following this perspective, it seems useful to be able, in a next step, to interface these embeddings with some frameworks such as the one proposed by OMOP [49].

Regarding the method used to build the diagnosis embedding, the Skip-gram architecture of Word2Vec takes the order of words into account, when the order in which the diagnoses are recorded in a stay has no meaning. This could have impaired the precision of the model by artificially increase the relation between some diagnoses. We randomized the order of the diagnoses for each stay to address this issue. In addition, in the present study, we did not consider the entire care pathway of a same patient (i.e. the sequence of hospitalizations), we considered hospital stays independently of each other, which is a simplification: still by analogy with the text, the use of more recent methods such as BERT [50] could probably be envisaged in order to model the whole of this sequence. Moreover, it would be of great interest, in our

opinion, to develop multimodal strategies to build generic embeddings combining other types of structured data such as procedure codes, lab results or drug prescriptions.

Finally, in our opinion, if the interest of such medical embeddings was confirmed, we formulate the hypothesis that an embedding built on international large databases could be shared by researchers in the biomedical field, similarly to the features developed on ImageNet [51] for images or text embeddings proposed with the development of BERT.

5 Conclusion

A diagnosis embedding built from a large medico-administrative database has been proposed. Using this embedding in a new epidemiological context, allowed to improve a prediction task. In addition, we obtained a prediction quality close to that obtained from a larger sample size but without using this embedding. This suggests a gain in statistical power for studies with a limited number of patients where this embedding can be applied. This work demonstrates, in our opinion, the first application case of transferring a representation built from massive data to an epidemiological study context.

Discussion en français

1 Résultats principaux

Nous avons construit un *embedding* des diagnostics à partir de la base de données nationale du PMSI, puis nous avons transféré cette représentation dans un contexte d'étude épidémiologique classique. Lorsque l'on considère une tâche de prédiction apprise sur de petits échantillons, à la fois pour la prédiction de la réhospitalisation et du décès, la représentation Skip-gram semble mieux fonctionner que l'utilisation des codes diagnostiques bruts. De plus, nous avons obtenu une qualité de prédiction proche de celle obtenue lorsque l'on utilise les diagnostics bruts sur un échantillon de plus grande taille. *L'embedding* des diagnostics étant construit dans ce travail à partir de l'ensemble des séjours hospitaliers sur une année, il n'est pas réduit à une spécialité médicale, il est généraliste et peut être transposé dans de multiples contextes.

Cependant, la différence entre les modèles construits avec ou sans *embedding* des diagnostics semble disparaître à mesure que la taille de l'échantillon d'apprentissage augmente. D'autre part, les modèles Skip-gram semblent surpasser les modèles CBOW, ce qui était attendu car chaque diagnostic était plutôt rare.

2 Travaux précédents

À notre connaissance, aucun travail d'apprentissage par transfert utilisant des bases de données épidémiologiques (qu'il s'agisse de réutilisation des données ou d'études pour lesquelles les données ont été recueillies expressément) n'avait été effectué, il est donc difficile de comparer ce résultat de manière spécifique.

Certains auteurs ont proposé des *embeddings* construits à partir de données médicales structurées qui ont ensuite été utilisés sur les données ayant servi à les construire. Farhan et al. ont construit un *embedding* inspiré de Word2Vec pour prédire 80 diagnostics à partir d'une séquence temporelle d'événements [26]. Ils ont utilisé une sélection de résultats de laboratoire, de prescriptions de médicaments, de codes diagnostics, de pathologies et de symptômes de 5195 patients suivis en soins intensifs à partir de la base de données MIMIC-III [25]. D'autres travaux ont porté sur la prévision du devenir d'un patient à partir de représentations d'apprentissage profond des dossiers patients informatisés : Miotto et al. ont construit une représentation « *deep patient* » qu'ils ont ensuite utilisée pour prédire la probabilité pour les patients de développer certaines pathologies [47]. Choi et al. ont construit un réseau de neurones récurrent pour prédire les codes diagnostiques et les prescriptions de médicaments de la prochaine hospitalisation d'un patient [48].

Certaines étapes de notre travail utilisent aussi un *embedding* construit à partir d'un échantillon dont la taille est proche de celle de l'échantillon d'apprentissage (par exemple lorsque *l'embedding* est construit sur 50 000 séjours et le modèle prédictif est construit sur 9500 séjours). Néanmoins, dans ce cas, quel que soit le modèle considéré, le résultat obtenu à partir des diagnostics bruts semble supérieur aux résultats obtenus à partir des *embeddings* de diagnostics (voir les Supplementary

Table 2 et Supplementary Table 3 en annexe). À notre avis, ce résultat souligne l'importance de construire les *embeddings* à partir d'une grande base de données et l'intérêt de la tâche de transfert vers des échantillons plus petits.

Notre travail tire parti de la disponibilité du très grand nombre de données produites en routine, telles que celles des dossiers patients informatisés ou des données de facturation, et permet construire puis transférer des représentations vers des contextes dans lesquels les données sont activement collectées au coût d'un investissement important d'une part, et vers des études épidémiologiques réalisée sur des bases de données médico-administratives pâtissant d'un manque de puissance statistique (ex: événements rares en pharmaco-épidémiologie) d'autre part. Il est raisonnable de penser que ces représentations pourraient être utiles notamment dans le cas des études épidémiologiques où des modèles prédictifs doivent être développés, comme par exemple des études où un score de propension (prédisant la présence de la drogue d'intérêt ou de l'intervention d'intérêt) est utilisé. Néanmoins, à ce stade, l'utilité de ce type de représentation n'a pas été évaluée sur un ensemble de données de validation externe.

3 Validité externe de ce travail

Plus généralement, l'utilisation de représentations complexes pose la question de l'explicabilité des modèles construits à partir de ces représentations. En effet, l'utilisation de modèles fondés sur les *embeddings* peut paraître en relative contradiction avec les modèles classiquement réalisés dans le domaine de l'épidémiologie, pour lesquels les variables explicatives candidates sont soigneusement sélectionnées par des experts. À notre avis, ces deux approches pourraient être combinées, par exemple en utilisant un *embedding* pour représenter les variables d'ajustement préalablement choisies par un expert.

La possibilité de partage des *embeddings* une fois construits soulève plusieurs questions. Premièrement, s'agissant de données de santé, se pose la question de la confidentialité. Les *embeddings* n'étant constitués que de valeurs numériques associés à des objets (des diagnostics dans le cadre de notre travail), il n'est pas possible de réidentifier les patients dont les données ont été utilisées. Deuxièmement, le partage de nos *embeddings* de diagnostics pose la question de leur interopérabilité, notamment s'ils étaient réutilisés dans le cadre de cohortes classiques en épidémiologie pour lesquelles les informations recueillies ne sont pas toujours structurées selon des classifications existantes. L'utilisation de la CIM-10 garantit néanmoins leur interopérabilité avec la plupart des bases de données médico-administratives. Dans cette perspective, il semble utile de pouvoir, dans une prochaine étape, interfacer ces *embeddings* avec certains *frameworks* existants, comme celui proposé par OMOP [49].

4 Validité interne de ce travail

En ce qui concerne la méthode utilisée pour construire les *embeddings* de diagnostics, l'architecture Skip-gram de Word2Vec prend en compte l'ordre des mots, alors que l'ordre dans lequel les diagnostics sont enregistrés dans un séjour n'a pas de signification. Cela aurait pu nuire à la précision du modèle en augmentant

artificiellement la relation entre certains diagnostics. Nous avons randomisé l'ordre dans lequel apparaissaient les diagnostics pour chaque séjour pour résoudre ce problème.

Dans la présente étude, nous n'avons pas considéré l'intégralité du parcours de soins d'un même patient (i.e. la séquence des hospitalisations), mais nous avons considéré les séjours à l'hôpital indépendamment les uns des autres, ce qui est une simplification: Toujours par analogie avec le texte, le l'utilisation de méthodes plus récentes telles que BERT [50] pourrait probablement être envisagée afin de modéliser l'ensemble de cette séquence.

Par ailleurs, il serait d'un grand intérêt, à notre avis, de développer des stratégies multimodales pour construire des *embeddings* génériques combinant d'autres types de données structurées telles que les codes d’actes médicaux, les résultats de laboratoire ou les prescriptions de médicaments.

Enfin, à notre avis, si l'intérêt de tels *embeddings* médicaux se confirmait, un *embedding* construit sur de grandes bases de données internationales pourrait être partagé par les chercheurs du domaine biomédical, à l'instar des fonctionnalités développées sur ImageNet pour les images [51] ou les intégrations de texte proposé avec le développement du BERT.

Conclusion

Nous proposons un *embedding* de diagnostics construit à partir d'une grande base de données médico-administrative. L'utilisation de cet *embedding* dans un contexte épidémiologique distinct a permis d'améliorer une tâche de prédiction par rapport aux méthodes classiques. De plus, nous avons obtenu une qualité de prédiction proche de celle obtenue sans *embedding* sur d'un échantillon de plus grande taille. Cela suggère un gain de puissance statistique potentiel si cet *embedding* est utilisé pour les études avec un nombre limité de patients. Ce travail démontre, à notre avis, le premier cas d'application du transfert d'une représentation construite à partir de données massives vers un contexte d'étude épidémiologique.

Liste des tables et figures

Références bibliographiques

[1] Meystre SM, Lovis C, Bürkle T, Tognola G, Budrionis A, Lehmann CU. Clinical Data Reuse or Secondary Use: Current Status and Potential Future Progress. Yearb Med Inform 2017;26:38–52. https://doi.org/10.15265/IY-2017-007.

[2] Schlegel DR, Ficheur G. Secondary Use of Patient Data: Review of the Literature Published in 2016. Yearb Med Inform 2017;26:68–71.

[3] Safran C. Reuse of clinical data. Yearb Med Inform 2014;9:52–4. https://doi.org/10.15265/IY-2014-0013.

[4] Chazard E, Ficheur G, Caron A, Lamer A, Labreuche J, Cuggia M, et al. Secondary Use of Healthcare Structured Data: The Challenge of Domain-Knowledge Based Extraction of Features. Stud Health Technol Inform 2018;255:15–9.

[5] Baro E, Degoul S, Beuscart R, Chazard E. Toward a Literature-Driven Definition of Big Data in Healthcare. BioMed Res Int 2015;2015. https://doi.org/10.1155/2015/639021.

[6] Laney D. 3D Data Management: Controlling Data Volume, Velocity, and Variety. META Group 2001.

[7] Chazard E. Réutilisation et fouille de données massives de santé produites en routine au cours du soin. Habilitation à Diriger des Recherches. Lille, 2017.

[8] Mitchell T. Machine Learning. McGrow Hill. 1997.

[9] Pennington J, Socher R, Manning C. Glove: Global Vectors for Word Representation. Proc. 2014 Conf. Empir. Methods Nat. Lang. Process. EMNLP, Doha, Qatar: Association for Computational Linguistics; 2014, p. 1532–43. https://doi.org/10.3115/v1/D14-1162.

[10] Mihálta M. mmihaltz/word2vec-GoogleNews-vectors. 2020.

[11] Mikolov T, Le QV, Sutskever I. Exploiting Similarities among Languages for Machine Translation. ArXiv13094168 Cs 2013.

[12] Bengio S, Heigold G. Word embeddings for speech recognition. Interspeech 2014.

[13] Wang Y, Liu S, Afzal N, Rastegar-Mojarad M, Wang L, Shen F, et al. A comparison of word embeddings for the biomedical natural language processing. J Biomed Inform 2018;87:12–20. https://doi.org/10.1016/j.jbi.2018.09.008.

[14] Bonina DCM. New business models and the value of open data: definitions, challenges and opportunities n.d.:30.

[15] Cuggia M, Polton D, Wainrib B, Combes S. Rapport health data hub 2018.

[16] Tang B, Cao H, Wang X, Chen Q, Xu H. Evaluating Word Representation Features in Biomedical Named Entity Recognition Tasks. BioMed Res Int 2014;2014:1–6. https://doi.org/10.1155/2014/240403.

[17] Liu S, Tang B, Chen Q, Wang X. Effects of Semantic Features on Machine Learning-Based Drug Name Recognition Systems: Word Embeddings vs.

Manually Constructed Dictionaries. Information 2015;6:848–65. https://doi.org/10.3390/info6040848.

[18] Jagannatha A, Chen J, Yu H. Mining and Ranking Biomedical Synonym Candidates from Wikipedia. Proc. Sixth Int. Workshop Health Text Min. Inf. Anal., Lisbon, Portugal: Association for Computational Linguistics; 2015, p. 142–51. https://doi.org/10.18653/v1/W15-2619.

[19] Jiang Z, Jin L, Li L, Qin M, Qu C, Zheng J, et al. A CRD-WEL System for Chemical-disease Relations Extraction 2017:10.

[20] Wang Y, Rastegar-Mojarad M, Komandur-Elayavilli R, Liu S, Liu H. An Ensemble Model of Clinical Information Extraction and Information Retrieval for Clinical Decision Support n.d.:10.

[21] Wu Y, Xu J, Zhang Y, Xu H. Clinical Abbreviation Disambiguation Using Neural Word Embeddings. Proc. BioNLP 15, Beijing, China: Association for Computational Linguistics; 2015, p. 171–6. https://doi.org/10.18653/v1/W15-3822.

[22] Asgari E, Mofrad MRK. Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics. PLOS ONE 2015;10:e0141287. https://doi.org/10.1371/journal.pone.0141287.

[23] Ng P. dna2vec: Consistent vector representations of variable-length k-mers. ArXiv170106279 Cs Q-Bio Stat 2017.

[24] Choi Y, Chiu CY-I, Sontag D. Learning Low-Dimensional Representations of Medical Concepts. AMIA Summits Transl Sci Proc 2016;2016:41–50.

[25] Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016;3:160035. https://doi.org/10.1038/sdata.2016.35.

[26] Farhan W, Wang Z, Huang Y, Wang S, Wang F, Jiang X. A Predictive Model for Medical Events Based on Contextual Embedding of Temporal Sequences. JMIR Med Inform 2016;4. https://doi.org/10.2196/medinform.5977.

[27] Nielsen MA. Neural Networks and Deep Learning 2015.

[28] Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. ArXiv13013781 Cs 2013.

[29] Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed Representations of Words and Phrases and their Compositionality. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, editors. Adv. Neural Inf. Process. Syst. 26, Curran Associates, Inc.; 2013, p. 3111–3119.

[30] Firth JR. A synopsis of linguistic theory 1930-1955. Stud. Linguist. Anal., Oxford: Philological Society: 1957, p. 1–32.

[31] de-obfuscated Python + question - Google Groupes n.d.

[32] Google Code Archive - Long-term storage for Google Code Project Hosting. n.d. https://code.google.com/archive/p/word2vec/ (accessed May 22, 2020).

[33] McCormick C. Word2Vec Tutorial - The Skip-Gram Model 2016. http://www.mccormickml.com.

[34] Chia D. An implementation guide to Word2Vec using NumPy and Google Sheets 2018. https://derekchia.com/an-implementation-guide-to-word2vec-using-numpy-and-google-sheets/ (accessed May 29, 2020).

[35] Morin F, Bengio Y. Hierarchical Probabilistic Neural Network Language Model. AISTATS'05, 2005, p. 246–52.

[36] Kass GV. An exploratory technique for investigating large quantities of categorical data. J R Stat Soc Ser C Appl Stat 1980;29:119–127.

[37] Chazard E, Ficheur G, Merlin B, Serrot E, PSIP Consortium, Beuscart R. Adverse drug events prevention rules: multi-site evaluation of rules from various sources. Stud Health Technol Inform 2009;148:102–11.

[38] Chazard E, Ficheur G, Bernonville S, Luyckx M, Beuscart R. Data mining to generate adverse drug events detection rules. IEEE Trans Inf Technol Biomed Publ IEEE Eng Med Biol Soc 2011;15:823–30. https://doi.org/10.1109/TITB.2011.2165727.

[39] Breiman L. Heuristics of instability and stabilization in model selection. Ann Stat 1996.

[40] Liaw A, Wiener M. Classification and Regression by randomForest. R News 2002;2:18–22.

[41] Řehůřek R, Sojka P. Software Framework for Topic Modelling with Large Corpora. Proc. LREC 2010 Workshop New Chall. NLP Framew., Valletta, Malta: ELRA; 2010, p. 45–50.

[42] van der Maaten L, Hinton G. Visualizing Data using t-SNE. J Mach Learn Res 2008;9.

[43] Donaldson J. jdonaldson/rtsne. 2019.

[44] Fuhrman C, Delmas M-C. Hospitalisations pour exacerbations de BPCO : comment les identifier à partir des données du programme de médicalisation des systèmes d'information (PMSI) ? Saint-Mauricce (Fra): Institut de Veilles Sanitaire; 2009.

[45] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2020.

[46] Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011;12:77. https://doi.org/10.1186/1471-2105-12-77.

[47] Miotto R, Li L, Kidd BA, Dudley JT. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. Sci Rep 2016;6. https://doi.org/10.1038/srep26094.

[48] Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. ArXiv151105942 Cs 2016.

[49] Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc 2012;19:54–60. https://doi.org/10.1136/amiajnl-2011-000376.

[50] Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv181004805 Cs 2019.

[51] Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. ArXiv160207261 Cs 2016.

Annexes

1 Tables supplémentaires de l'article en anglais

Supplementary Table 1 : Computation time of diagnoses embeddings according to the learning sample size and the number of dimensions

Algorithm	Embedding learning sample size	Dimensions	Computation time
CBOW	21.7 millions	200	7min 31s
		100	7min 22s
	1 million	200	24s
		100	24s
	50,000	200	3s
		100	3s
Skip-gram	21.7 millions	200	14min 28s
		100	11min 30s
	1 million	200	44s
		100	34s
	50,000	200	4s
		100	3s

CBOW : Continuous bag of words

Supplementary Table 2 : Performances of readmission predicitive models

Algorithm	Emb. learning sample size	Diagnosis-to-stay aggregation	Emb. size	RF learning sample size	AUC [CI95]	RF Time
Bare diagnoses	_	_	_	19262	0.589 [0.573; 0.605]	21h 30min
				3852	0.565 [0.549; 0.581]	1h 42min
				770	0.547 [0.531; 0.563]	5min 35s
CBOW	21.7 millions	Mean	200	19262	0.582 [0.566; 0.598]	3min 7s
				3852	0.561 [0.545; 0.570]	22s
				770	0.540 [0.524; 0.557]	3s
			100	19262	0.582 [0.566; 0.598]	1min 32s
				3852	0.561 [0.545; 0.577]	12s
				770	0.548 [0.532; 0.564]	2s
		Maximum absolute value	200	19262	0.579 [0.563; 0.595]	3min 18s
				3852	0.562 [0.545; 0.578]	22s
				770	0.535 [0.518; 0.551]	3s
			100	19262	0.574 [0.558; 0.590]	1min 35s
				3852	0.558 [0.542; 0.574]	12s
				770	0.534 [0.518; 0.550]	2s
CBOW	1 million	Mean	200	19262	0.580 [0.564; 0.596]	3min 9s
				3852	0.564 [0.547; 0.580]	22s
				770	0.548 [0.532; 0.564]	3s
			100	19262	0.576 [0.560; 0.592]	1min 34s
				3852	0.561 [0.545; 0.577]	12s
				770	0.540 [0.524; 0.556]	2s
		Maximum absolute value	200	19262	0.570 [0.554; 0.586]	3min 21s
				3852	0.559 [0.543; 0.575]	22s
				770	0.536 [0.520; 0.553]	3s
			100	19262	0.569 [0.553; 0.585]	1min 37s
				3852	0.565 [0.549; 0.581]	12s
				770	0.538 [0.521; 0.554]	2s
CBOW	50,000	Mean	200	19262	0.552 [0.536; 0.568]	3min 28s
				3852	0.534 [0.518; 0.550]	23s
				770	0.521 [0.505; 0.537]	3s
			100	19262	0.552 [0.535; 0.568]	1min 41s
				3852	0.534 [0.518; 0.550]	12s
				770	0.525 [0.509; 0.541]	2s
		Maximum absolute value	200	19262	0.551 [0.535; 0.567]	4min 35s
				3852	0.541 [0.525; 0.558]	22s
				770	0.540 [0.524; 0.556]	3s
			100	19262	0.557 [0.541; 0.573]	1min 34s
				3852	0.540 [0.524; 0.556]	12s
				770	0.542 [0.526; 0.558]	2s

Algorithm	Emb. learning sample size	Diagnosis-to-stay aggregation	Emb. size	RF learning sample size	AUC [CI95]	RF Time
Skip-gram	21.7 millions	Mean	200	19262	0.590 [0.574; 0.606]	4min 1s
				3852	0.573 [0.557; 0.589]	22s
				770	0.564 [0.548; 0.580]	3s
			100	19262	0.590 [0.574; 0.606]	1min 39s
				3852	0.567 [0.551; 0.583]	16s
				770	0.568 [0.552; 0.584]	2s
		Maximum absolute value	200	19262	0.586 [0.571; 0.602]	3min 9s
				3852	0.568 [0.552; 0.584]	22s
				770	0.552 [0.536; 0.568]	3s
			100	19262	0.586 [0.570; 0.602]	1min 39s
				3852	0.564 [0.548; 0.580]	14s
				770	0.551 [0.535; 0.567]	2s
Skip-gram	1 million	Mean	200	19262	0.586 [0.570; 0.602]	3min 6s
				3852	0.564 [0.548; 0.580]	22s
				770	0.554 [0.538; 0.570]	3s
			100	19262	0.585 [0.569; 0.601]	1min 39s
				3852	0.571 [0.555; 0.587]	13s
				770	0.560 [0.544; 0.576]	2s
		Maximum absolute value	200	19262	0.584 [0.568; 0.600]	3min 14s
				3852	0.568 [0.552; 0.584]	22s
				770	0.544 [0.528; 0.560]	3s
			100	19262	0.586 [0.570; 0.602]	1min 38s
				3852	0.563 [0.547; 0.579]	13s
				770	0.547 [0.531; 0.563]	2s
Skip-gram	50,000	Mean	200	19262	0.564 [0.548; 0.580]	3min 24s
				3852	0.540 [0.523; 0.556]	22s
				770	0.532 [0.516; 0.548]	3s
			100	19262	0.564 [0.548; 0.580]	1min 45s
				3852	0.539 [0.523; 0.555]	13s
				770	0.538 [0.522; 0.555]	2s
		Maximum absolute value	200	19262	0.567 [0.551; 0.583]	3min 15s
				3852	0.562 [0.546; 0.578]	22s
				770	0.543 [0.527; 0.559]	3s
			100	19262	0.565 [0.549; 0.581]	1min 38s
				3852	0.558 [0.542; 0.574]	13s
				770	0.541 [0.525; 0.557]	2

CBOW : Continuous Bag of Words, Emb. : embedding, RF : Random Forest, CI95 : 95% confidence interval

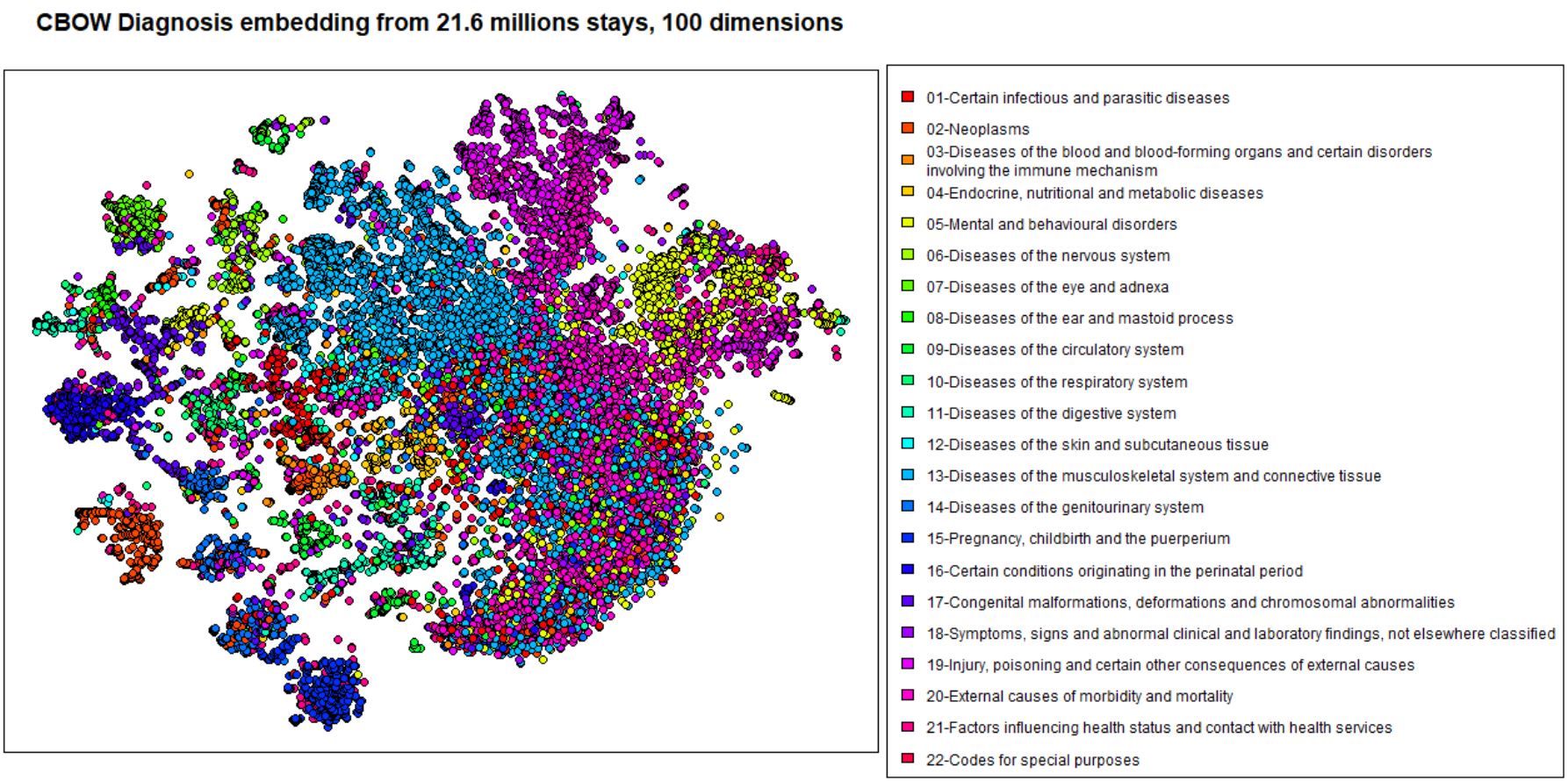
Supplementary Table 3 : Performances of death predicitve models

Algorithm	Emb. learning sample size	Diagnosis-to-stay aggregation	Emb. size	RF learning sample size	AUC [CI95]	RF Time
Bare diagnoses	_	_	_	9502	0.882 [0.869; 0.896]	3h 52min
				1900	0.871 [0.857; 0.885]	16min 23s
				380	0.840 [0.824; 0.856]	1min 9s
CBOW	21.7 millions	Mean	200	9502	0.874 [0.860; 0.888]	1min 8s
				1900	0.872 [0.857; 0.886]	8s
				380	0.861 [0.846; 0.876]	1s
			100	9502	0.874 [0.859; 0.888]	32s
				1900	0.874 [0.860; 0.888]	4s
				380	0.862 [0.848; 0.877]	1s
		Maximum absolute value	200	9502	0.872 [0.858; 0.887]	1min 5s
				1900	0.870 [0.856; 0.884]	8s
				380	0.853 [0.838; 0.868]	1s
			100	9502	0.871 [0.857; 0.885]	31s
				1900	0.870 [0.855; 0.884]	4s
				380	0.850 [0.835; 0.865]	1s
CBOW	1 million	Mean	200	9502	0.871 [0.857; 0.885]	1min 8s
				1900	0.871 [0.857; 0.886]	8s
				380	0.854 [0.838; 0.869]	1s
			100	9502	0.869 [0.855; 0.884]	32s
				1900	0.869 [0.854; 0.883]	4s
				380	0.850 [0.835; 0.865]	1s
		Maximum absolute value	200	9502	0.868 [0.853; 0.882]	1min 4s
				1900	0.868 [0.853; 0.882]	8s
				380	0.847 [0.831; 0.863]	1s
			100	9502	0.863 [0.848; 0.877]	32s
				1900	0.864 [0.850; 0.879]	4s
				380	0.845 [0.830; 0.861]	1s
CBOW	50,000	Mean	200	9502	0.792 [0.774; 0.810]	1min 5s
				1900	0.787 [0.768; 0.805]	8s
				380	0.748 [0.729; 0.768]	1s
			100	9502	0.787 [0.768; 0.805]	32s
				1900	0.780 [0.761; 0.799]	4s
				380	0.740 [0.720; 0.760]	1s
		Maximum absolute value	200	9502	0.827 [0.810; 0.843]	1min 37s
				1900	0.821 [0.804; 0.838]	8s
				380	0.800 [0.782; 0.818]	1s
			100	9502	0.825 [0.808; 0.841]	30s
				1900	0.816 [0.799; 0.833]	4s
				380	0.795 [0.777; 0.813]	1s

Algorithm	Emb. learning sample size	Diagnosis-to-stay aggregation	Emb. size	RF learning sample size	AUC [CI95]	RF Time
Skip-gram	21.7 millions	Mean	200	9502	0.874 [0.860; 0.888]	1min 6s
				1900	0.876 [0.861; 0.890]	8s
				380	0.868 [0.853; 0.882]	1s
			100	9502	0.877 [0.863; 0.891]	33s
				1900	0.878 [0.865; 0.892]	6s
				380	0.868 [0.854; 0.883]	1s
		Maximum absolute value	200	9502	0.876 [0.862; 0.890]	1min 6s
				1900	0.875 [0.861; 0.889]	8s
				380	0.858 [0.843; 0.874]	1s
			100	9502	0.877 [0.863; 0.892]	33s
				1900	0.875 [0.861; 0.890]	5s
				380	0.861 [0.846; 0.876]	1s
Skip-gram	1 million	Mean	200	9502	0.872 [0.858; 0.886]	1min 7s
				1900	0.874 [0.860; 0.888]	8s
				380	0.868 [0.853; 0.882]	1s
			100	9502	0.873 [0.859; 0.887]	33s
				1900	0.876 [0.861; 0.890]	5s
				380	0.863 [0.848; 0.878]	1s
		Maximum absolute value	200	9502	0.876 [0.862; 0.890]	1min 5s
				1900	0.874 [0.860; 0.888]	8s
				380	0.861 [0.846; 0.876]	1s
			100	9502	0.875 [0.861; 0.890]	33s
				1900	0.873 [0.859; 0.887]	5s
				380	0.859 [0.844; 0.874]	1s
Skip-gram	50,000	Mean	200	9502	0.818 [0.801; 0.835]	1min 7s
				1900	0.809 [0.792; 0.827]	8s
				380	0.786 [0.768; 0.805]	1s
			100	9502	0.820 [0.803; 0.837]	34s
				1900	0.814 [0.797; 0.831]	5s
				380	0.793 [0.775; 0.811]	1s
		Maximum absolute value	200	9502	0.850 [0.835; 0.865]	1min 5s
				1900	0.845 [0.829; 0.861]	8s
				380	0.828 [0.812; 0.845]	1s
			100	9502	0.847 [0.831; 0.862]	32s
				1900	0.843 [0.828; 0.859]	4s
				380	0.826 [0.809; 0.842]	1s

CBOW : Continuous Bag of Words, Emb. : embedding, RF : Random Forest, CI95 : 95% confidence interval

2 Figure supplémentaire de l'article en anglais



Supplementary figure 1 : Representation of the CBOW diagnosis embedding from 21.6 million stays. The reduction from 100 dimensions to 2 was obtained with t-SNE. The diagnoses are coloured according to ICD10 chapters

AUTEUR : Nom : Riant **Prénom :** Margaux

Date de soutenance : 30 juin 2020

Titre de la thèse : Apprentissage par transfert en épidémiologie : construction et évaluation d'un « *diagnosis embedding* » à partir de la base nationale médico-administrative du PMSI

Thèse - Médecine - Lille 2020

Cadre de classement : Médecine

DES + spécialité : Santé publique et médecine sociale

Mots-clés : Réutilisation de données ; apprentissage par transfert ; données massives ; épidémiologie

Résumé :

Introduction : Dans le domaine biomédical, la réutilisation des données fournit des informations sur un nombre limité de variables pour de nombreux patients, alors que les études épidémiologiques classiques permettent de décrire de nombreuses variables pour un nombre de patients limité. Le transfert de représentations construites à partir de la réutilisation de données vers des études épidémiologiques pourrait améliorer la qualité de la modélisation dans ces études. Les méthodes de *word embedding* permettent de construire des représentations à faible dimensionalité et peuvent être appliquées à des données médicales structurées. Notre objectif était de construire un *embedding* de diagnostics à partir d'une base de données médico-administrative nationale, puis de la transférer à une étude épidémiologique avec un nombre limité de patients.

Méthodes : Nous avons construit des *embeddings* de diagnostics, à partir des versions Continuous Bag Of Words (CBOW) et Skip-gram de Word2Vec, à partir de 21,7 millions de séjours de 2008 de la base nationale du PMSI. Nous avons transféré ces *embeddings* à deux études épidémiologiques et comparé leurs performances avec celles obtenues à partir des codes diagnostiques bruts. Les événements à prédire étaient la réhospitalisation à 30 jours ou le décès intra-hospitalier consécutifs à une hospitalisation pour une broncho-pneumopathie chronique obstructive. Les modèles étaient évalués par la valeur de l'aire sous la courbe ROC (AUC)

Résultats : Les modèles prédictifs de réhospitalisation ont été construits sur 770 séjours. L'AUC des modèles basés sur CBOW, Skip-gram et sur les diagnostics bruts étaient 0,548 (IC95%: [0,532 ; 0,564], 0,568 [0,552 ; 0,584] et 0,547 [0,531 ; 0,563], respectivement. Les modèles prédictifs de décès ont été construits sur 380 séjours. L'AUC des modèles basés sur CBOW, Skip-gram et sur les diagnostics bruts étaient de 0,862 (IC95%: [0,848 ; 0,877], 0,868 [0,854 ; 0,883] et 0,840 [0,824 ; 0,856], respectivement.

Conclusion : L'utilisation d'un *embedding*, construit à partir d'une grande base de données médico-administrative puis transféré dans un contexte épidémiologique, a permis d'améliorer les tâches de prédiction, démontrant la possibilité de transférer une représentation construite à partir de données massives.

<u>Composition du Jury</u> :

Président : Professeur P. AMOUYEL

Assesseurs : Professeur E. CHAZARD, Docteur A. LAMER, Docteur P. BALAYE

Directeur de thèse : Docteur G. FICHEUR