Actes de la journée d'étude sur la Similarité entre Patients, SimPa 2023

Adrien Coulet, Christel Gérardin, Aurélie Névéol, Xavier Tannier

Actes de la journée d'étude sur la similarité entre patients SimPa 2023

Avec le soutien de l'ATALA, de PariSanté Campus et du laboratoire iPLesp

Éditeurs

Adrien Coulet
Inria, Inserm, Université Paris Cité, HeKA

Christel Gérardin
iPLesp, AP-HP

Aurélie Névéol
Université Paris-Saclay, CNRS, LISN

Xavier Tannier
Sorbonne Université, Insemr, LIMICS

13 mars 2023
PariSanté Campus, 2 - 10 Rue d'Oradour-sur-Glane, 75015 Paris

Préface

La notion de similarité est importante pour représenter les connaissances médicales et traiter les documents véhiculant des informations de santé. Les développements méthodologiques récents en traitement automatique de la langue ont permis d'explorer différentes facettes de la similarité en santé en s'appuyant sur la représentation des contenus structurés ou non structurés typiquement présents dans les entrepôts de données de santé.

Dans ce contexte, il nous a semblé intéressant d'examiner les approches proposées dans les communautés de l'informatique médicale, de la gestion de connaissances et du traitement automatique de la langue sur des questions telles que la définition et l'identification des phénomènes de similarité textuelle en santé, explorant un continuum allant de l'identité des chaînes de caractères à la similarité sémantique et thématique entre documents ; l'identification de « cas similaires » (par exemple pour la définition de cohortes, l'inclusion dans des essais cliniques ou la détection de dossiers patients dupliqués) ; la représentation de patients ; la création de jumeaux numériques par génération de données patient synthétiques.

L'objectif de la journée SimPa était double :

— documenter les cas pratiques en santé dans lesquels les méthodes de calcul de similarité entre patients sont utiles ;
— documenter les solutions existantes, par exemple, mais sans y être limité, pour les systèmes fondées sur des méthodes d'apprentissage automatique.

En bref, nous souhaitions que cette journée permette une rencontre des communautés d'informatique médicale, santé publique et traitement automatique de la langue et des échanges sur la recherche de contenus textuels « similaires » en TAL. Elle a bien rempli cet objectif en permettant de rassembler 45 personnes (30 en présentiel, 15 à distance) autour de la thématique de similarité entre patients.

Le comité d'organisation.

Comité d'organisation

Adrien Coulet, Christel Gérardin, Aurélie Névéol, Xavier Tannier.

Comité de programme

François Antonini, Sandra Bringay, Vincent Claveau, Natalia Grabar, Thomas Guyet, Vianney Jouhet, Lina Soualmia, ainsi que les membres du comité d'organisation.

Remerciements

Nous remercions l'ATALA et l'Institut Pierre Louis d'Épidémiologie et de Santé Publique (iPLesp) pour leur soutien financier. Nous remercions Inria Paris et PariSanté Campus pour leur accueil en leurs locaux. Nous remercions Caio Corro et Gaël Lejeune pour avoir répondu à nos nombreuses questions concernant l'organisation des journées d'étude et l'édition des actes. Nous remercions les membres du comité de programme pour avoir soigneusement relu les résumés soumis. Enfin, nous remercions les conférenciers invités et les auteur·ices des résumés qui ont fait de cette journée d'étude un succès.

Programme de la journée

	Accueil à partir de 9 :30
9:50 - 10:00	**Mot d'accueil** Comité d'organisation
10:00 - 11:00	Présentation invitée **Recherche d'information dans un domaine spécifique et intelligence artificielle** Nicolas Fiorini
11:00 - 11:30	**Exploration de la représentation latente d'un modèle profond de patient sous forme de graphes.** Hugo Le Baher, Jérôme Azé, Sandra Bringay, Caroline Dunoyer, Pascal Poncelet & Nancy Rodriguez
11:30 - 12:00	**Event2vec, democratizing medical concept embeddings at scale.** Matthieu Doutreligne, Antoine Neuraz & Gaël Varoquaux
12:00 - 12:15	**Similarité des documents médicales en se basant sur des embeddings appris pour le codage médical.** Leonardo Moros, Jérôme Azé, Maximilien Servajean, Sandra Bringay, Pascal Poncelet & Caroline Dunoyer
	Repas (offert par l'ATALA et l'iPLesp)
14:00 - 14:30	**Détection de zones dupliquées dans des comptes rendus médicaux.** Thibault Fabacher, Olivier Birot, Camila Arias-Villamil, Kim-Tâm Huynh, Antoine Neuraz & Bastien Rance
14:30 - 14:45	**Patient similarity study to identify hospital units with the highest rate of unplanned readmissions.** Nzamba Bignoumba, Sadok Ben Yahia & Nedra Mellouli
14:45 - 15:15	**Exploring Similarities and Dissimilarities in Patient Representations for Analogical Reasoning.** Safa Alsaidi, Miguel Couceiro, Nicolas Garcelon & Adrien Coulet
15:15 - 14:45	**Similarité surfacique et similarité sémantique dans des cas cliniques générés.** Nicolas Hiebel, Olivier Ferret, Karën Fort & Aurélie Névéol
	Pause café
16:30 - 17:00	**Patient-patient similarity-based screening of a clinical data warehouse to support rare disease diagnosis.** Xiaoyi Chen, Carole Faviez, Anita Burgun & Nicolas Garcelon
16:50 - 17:20	Présentation invitée **Adventures in using real-world evidence at the bed-side** Nigam H. Shah

Table des matières

Recherche d'information dans un domaine spéci que et intelligence arti cielle

Nicolas Fiorini
R&D, Algolia, Paris

RÉSUMÉ

La recherche d'information en domaine de spécialité est un problème majeur, aujourd'hui toujours partiellement résolu. Les problématiques sont multiples : les requêtes di èrent signi cativement des contenus, la pertinence est complexe à caractériser, le vocabulaire est lui aussi spéci que, et souvent nous faisons face à un manque d'annotations ou de jeux de données pour améliorer les systèmes. Cette complexité a motivé plusieurs initiatives consistant à se reposer sur de l'intelligence arti cielle pour trouver des solutions, là où un humain serait dépassé par les possibilités combinatoires. Dans cet exposé, nous présenterons un ensemble de méthodes qui ont pu faire leurs preuves dans le contexte de la recherche d'information en domaine de spécialité, ainsi que leur potentielle adaptation d'un domaine à un autre. Les problématiques nouvellement créées lors de l'utilisation de ces approches (notamment éthiques, de con dentialité, ou d'explicabilité) seront aussi couvertes, a n de donner une intuition globale de la mise en œuvre de telles approches sur des cas d'applications concrets.

BIOGRAPHIE

Nicolas Fiorini a réalisé une thèse en informatique à l'Ecole des Mines d'Alès avec un accent sur l'intelligence arti cielle, la classi cation de données et l'utilisation de bases de connaissances. Il a étendu ses recherches au National Center for Biotechnology Information (NCBI, Washington, DC) en appliquant des méthodes d'apprentissage automatique sur la recherche de littérature biomédicale via l'outil PubMed. De retour en France, il a travaillé chez Doctrine, une startup proposant un moteur de recherche de contenu juridique français (décision de justice, lois, etc.). Il a récemment rejoint Algolia en tant que Director of Engineering, où il supervise une équipe construisant des moteurs de recherche et de recommandation avec une forte composante IA.

Exploration de la représentation latente d'un modèle profond de patient sous forme de graphes

Hugo Le Baher[1, 2, 3], Jérôme Azé[1], Sandra Bringay[1, 4], Pascal Poncelet[1], Nancy Rodriguez[1], Caroline Dunoyer[2, 5]

(1) LIRMM, UMR 5506, Université de Montpellier, CNRS, Montpellier, France
(2) Département d'Information Médicale, CHU Montpellier, Montpellier, France
(3) 5 DEGRÉS, Paris, France
(4) AMIS, Université Paul-Valéry, Montpellier, France
(5) IDESP, UMR UA11, INSERM - Université de Montpellier, Montpellier, France

RÉSUMÉ

L'article étudie la représentation latente d'un modèle profond représentant un patient sous forme de graphe et sa capacité à révéler des caractéristiques communes chez les patients, pour l'utilisation du modèle dans de nouvelles tâches, comme la construction de cohortes.

ABSTRACT

Exploration of the Latent Representation of a Deep Graph Patient Model

The article studies the latent representation of a deep graph patient model and whether it can reveal common characteristics of patients, to see if the model can be used for new tasks (e.g. cohort building).

MOTS-CLÉS : Représentation de patients, GCN, MIMIC-III, Dossier patient informatisé.

KEYWORDS: Patient representation, GCN, MIMIC-III, Electronic health record.

1 Introduction

Dans (Le Baher *et al.*, 2023), nous proposons de modéliser un parcours patient sous la forme d'un graphe où la temporalité est représentée via les arêtes. De manière à valider notre représentation, nous proposons un modèle basé sur des GCN (*Graph Convolutional Network*) que nous expérimentons sur des données issues de MIMIC-III (Johnson *et al.*, 2016) afin de prédire le décès à 24 heures. Les résultats obtenus (0.897 AUC) sont similaires à ceux de l'état de l'art qui utilisent des modèles basés principalement sur des LSTM (Harutyunyan *et al.*, 2019). Dans cette nouvelle communication, nous nous intéressons plus particulièrement à la représentation latente apprise au cours de l'apprentissage. Nous souhaitons étudier si cette représentation pourrait mettre en évidence des patients aux caractéristiques communes. L'objectif est d'étudier si le processus mis en place pour une tâche spécifique peut être utilisé sur de nouvelles tâches, avec des applications comme le sous-typage de patients (Baytas *et al.*, 2017) ou de maladies (Kaneko *et al.*, 2021), la construction de cohortes (Wirbka *et al.*, 2020) ou la définition de modèles prédictifs personnalisés (Hatton *et al.*, 2019). En d'autres termes, est-ce que le modèle appris pour prédire des décès à 24h, peut être utilisé pour prédire d'autres événements médicaux ?

2 La représentation latente

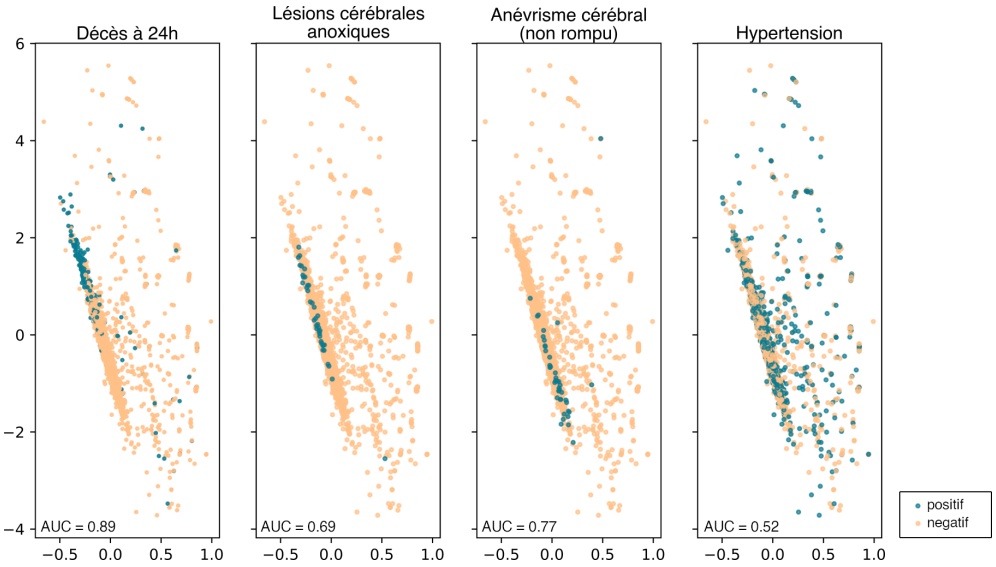

FIGURE 1 – Représentation extraite du modèle GCN.

Notre modèle étant basé sur une architecture à base de GCN, nous pouvons extraire directement les informations issues des sorties des convolutions. La Figure 1 illustre la représentation extraite d'une couche intermédiaire sur laquelle nous appliquons une ACP pour réduire les dimensions et les visualiser. Par la suite, nous sélectionnons 5 760 patients pour lesquels nous identifions la présence ou l'absence d'un diagnostic en particulier. Nous considérons quatre diagnostics : décès à 24h, lésions cérébrales anoxiques, anévrisme cérébral non rompu et hypertension (représentés sur la figure de gauche à droite). Pour la prédiction du décès à 24h, nous pouvons constater un cluster très localisé dans l'espace (points bleus) ce que confirme la forte valeur de l'AUC. De même, les lésions cérébrales anoxiques et l'anévrisme cérébral sont aussi localisés. Au contraire, l'hypertension ne semble pas posséder de localité dans cette représentation latente.

Les lésions cérébrales anoxiques et l'anévrisme cérébral concernent respectivement la perte d'oxygène dans le cerveau et la dilatation de la paroi d'une artère. La détection de ces diagnostics est fortement liée à la mortalité du patient et la représentation latente apprise reste adaptée. L'hypertension est un diagnostic identifié chez des patients à gravité très variable d'où le fait qu'il n'y ait pas de localité bien définie par rapport à la représentation latente apprise. Ces expérimentations montrent que pour une tâche particulière, la représentation latente est intéressante et met en évidence des caractéristiques communes. Par contre pour des diagnostics où le décès n'intervient que rarement, on peut constater qu'elle n'est pas adaptée. Pour cette raison, nos travaux actuels s'intéressent à l'apprentissage de la représentation des nœuds en prédisant l'existence des arcs qui les relient.

Remerciements

Ce projet est financé par une bourse CIFRE, financée par 5 DEGRÉS, établie en collaboration avec le LIRMM et le CHU de Montpellier.

Références

BAYTAS I. M., XIAO C., ZHANG X., WANG F., JAIN A. K. & ZHOU J. (2017). Patient subtyping via time-aware LSTM networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* : ACM. DOI : 10.1145/3097983.3097997.

HARUTYUNYAN H., KHACHATRIAN H., KALE D. C., VER STEEG G. & GALSTYAN A. (2019). Multitask learning and benchmarking with clinical time series data. *Scientific Data*, **6**(1), 96. DOI : 10.1038/s41597-019-0103-9.

HATTON C. M., PATON L. W., MCMILLAN D., CUSSENS J., GILBODY S. & TIFFIN P. A. (2019). Predicting persistent depressive symptoms in older adults : A machine learning approach to personalised mental healthcare. *Journal of Affective Disorders*, **246**, 857–860. DOI : 10.1016/j.jad.2018.12.095.

JOHNSON A. E. W., POLLARD T. J., SHEN L., LEHMAN L.-W. H., FENG M., GHASSEMI M., MOODY B., SZOLOVITS P., CELI L. A. & MARK R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, **3**, 160035. DOI : 10.1038/sdata.2016.35.

KANEKO H., UMAKOSHI H., OGATA M., WADA N., IWAHASHI N., FUKUMOTO T., YOKOMOTO-UMAKOSHI M., NAKANO Y., MATSUDA Y., MIYAZAWA T., SAKAMOTO R. & OGAWA Y. (2021). Machine learning based models for prediction of subtype diagnosis of primary aldosteronism using blood test. *Scientific Reports*, **11**(1). DOI : 10.1038/s41598-021-88712-8.

LE BAHER H., AZÉ J., BRINGAY S., PONCELET P., RODRIGUEZ N. & DUNOYER C. (2023). Modélisation de parcours patients : graphes temporels pour la supervision médicale. *Extraction et Gestion des Connaissances*.

WIRBKA L., HAEFELI W. E. & MEID A. D. (2020). A framework to build similarity-based cohorts for personalized treatment advice – a standardized, but flexible workflow with the r package SimBaCo. *PLOS ONE*, **15**(5), e0233686. DOI : 10.1371/journal.pone.0233686.

Event2vec, democratizing medical concept embeddings at scale

Matthieu Doutreligne, Antoine Neuraz, Gaël Varoquaux

1 Motivation and context

Large observational databases –clinical health records, or claims– contain multiple tables, geographical disparities and heterogeneous data types and temporalities. Complex, time-consuming and costly work in medical and computer skills is required to clean and transform these sources into data tables suited for statistical analysis (Bacry *et al.*, 2020; Hripcsak *et al.*, 2015).

Simplifying the link from the collection of raw data, heterogeneous between sites to the ideal medical concepts driving study questions, is a key step to better utilize large observational databases.

The formatting of big observational databases as a collection of events has emerged in numerous machine learning application in healthcare (Rajkomar *et al.*, 2018; Beam *et al.*, 2019; Bacry *et al.*, 2020; Chazard *et al.*, 2022). A patient record can be described as a sequence of events e, each one of them being represented by a person identifier, a datetime t and a medical code c, $e = (i, t, c)$. This collection of ordered tokens is analogous to language. Building on a formulation of word2vec (Mikolov *et al.*, 2013) as the factorization of the Positive Pointwise Information Matrix (PPMI), Beam *et al.* (2019) create medical concept embeddings from both text and structured data. However they do not provide a package implementing the computational heavy part of the algorithm: the construction of the event cooccurrence matrix.

Sharable concept embeddings. The underlying information at the basis of the vectors is a sharable aggregated data: the cooccurrence matrix. The easy creation and exchanges of medical concept embeddings open different applications that efficiently pulls data from multiple sites, overcoming both the heterogeneity of the data, and the administrative barriers to access individual patient data.

2 Qualitative evaluation on two french healthcare databases

We propose event2vec [1], a python package democratizing medical concept embeddings that leverages the spark parallel computing framework (Salloum *et al.*, 2016) to perform the SVD-PPMI algorithm (Beam *et al.*, 2019; Levy & Goldberg, 2014). We apply our package on a medium and a large medical datasets from different nature. For each dataset, the events have been grouped by individual and sorted by date of care, forming sequences of medical codes.

[1] https://gitlab.com/strayMat/event2vec/

- **Claims:** 950 million of care events in the pathways of a random sample of 3,112,565 patients with a vocabulary of 15, 968 unique concepts have been extracted from the System National de Données de Santé (SNDS) (Tuppin *et al.*, 2017). This work has already presented in (Doutreligne *et al.*, 2020).
- **Clinical data warehouse:** 15 million of care events with a vocabulary of 10583 unique concepts, have been extracted from a random sample of 200, 000 patients from the clinical warehouse of the Greater Hospital of Paris (APHP).

We used the Scalpel preprocessing pipeline (Bacry *et al.*, 2020) for the SNDS and the eds-scikit package (Petit-Jean *et al.*, 2023) for the APHP.

Tsne projections (Van der Maaten & Hinton, 2008) of the embeddings –Figure 1– display clear groups of pathologies recovered by the algorithm, a reflection of the local usage of the codes. Thanks to the aggregated nature of the embeddings, we can share interactive tsne plots and full embeddings (currently SNDS only) [2].

(a) APHP

(b) SNDS

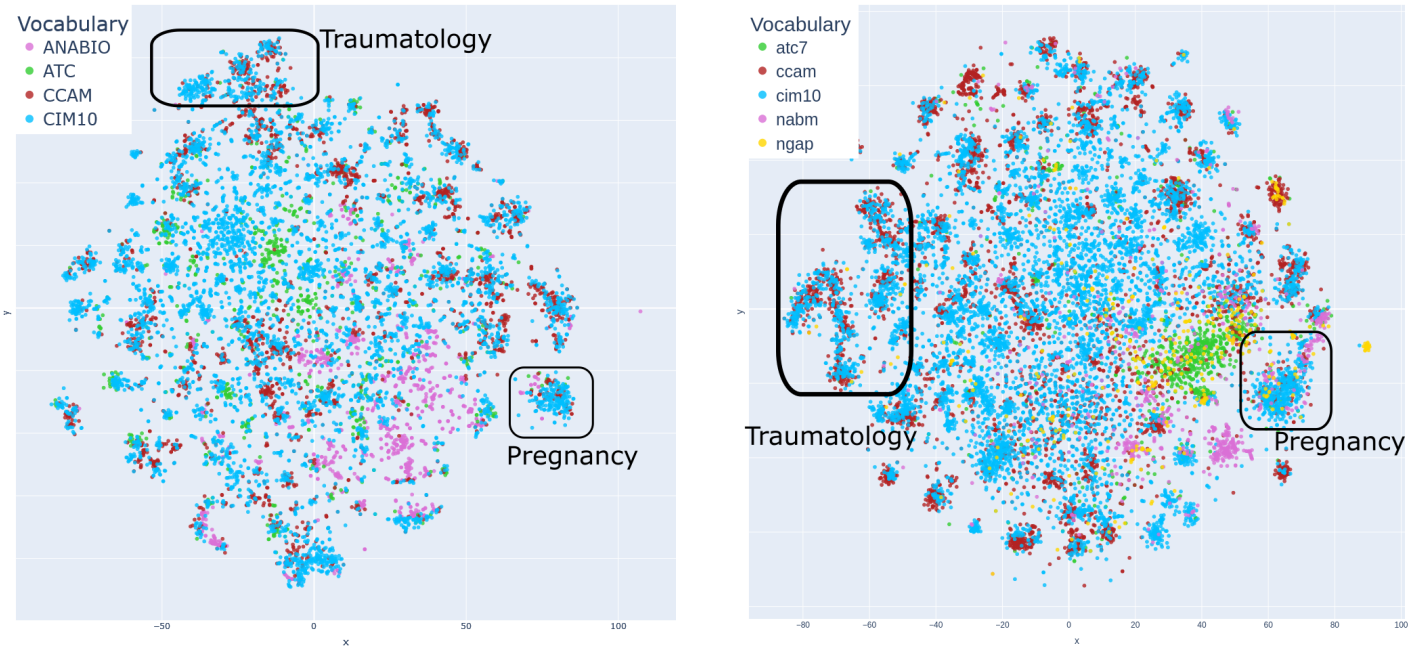

Figure 1: TSNE projection of event embeddings a) in the APHP Clinical Data Warehouse, b) in the French Medical Claims (SNDS).

3 Empirical evaluation

It is delicate to assess the *general value* of these embeddings (Hong *et al.*, 2021). Instead, their *utility* should be measured for specific downstream tasks, some of which we begin to explore.

[2] https://straymat.gitlab.io/event2vec/visualizations.html

3.1 Downstream task: Rehospitalization at 30 days

Rehospitalization is a common benchmarking task for EHR data (Beaulieu-Jones *et al.*, 2021). Good rehospitalization models would be useful for patient risk stratification, patient followup or hospital planning. It is also a key outcome for epidemiological studies that should be well modeled and included in causal outcome modelizations for treatment effect estimation (Wendling *et al.*, 2018; Dorie *et al.*, 2019).

Data We use a random sample of 200, 000 patients from the clinical warehouse of the Greater Hospital of Paris (APHP).

Cohort creation The study period is restricted to [2017-01-01, 2022-06-01]. Before this period the data is not in a stable regime because of the ramping up of the Information System. After this period, we don't have the followup to evaluate rehospitalization.

We used the eds-scikit package (Petit-Jean *et al.*, 2023) for preprocessing. Included patients were those with at least one hospitalization in the study period, aged over 18 at admission, with a minimum of 30 days of followup after the admission, strictly less than 7 hospitalizations, not deceased before the end of their inclusion visit. A detail of the inclusion criteria is given in Figure 5.

Among this cohort, the prevalence of reshospitalization at 30 days is 10%.

3.2 Methods

Pre-processing The followup starts at the end of the first hospitalization, used for inclusion. We select all events before the start of followup in the following tables: CIM10 billing codes, CCAM billing codes, drug administrations. Only three information are kept for each event: the timestamp, the code and the person_id.

Evaluation protocol The population is split into a train $\mathcal{T}$ and a test set $\mathcal{S}$, with a ratio of 0.7/0.3. To evaluate the advantage of embeddings with respect to the sample size, the effective train set size is restricted to increasing ratio of its full size: $N_{train,r} = r \cdot N_{\mathcal{T}} \quad with \quad r \in [0.1, 0.25, 0.5, 0.9, 1]$. The test set is left unchanged. Every model is learned on a sub-train set and evaluated on the test set.

Models compared We study different models for this predictive task. A model is the combination of a featurizer and a classifier.

For the representations, we consider three different featurizers **a) Count vectorizer** $[C, C_{decay}]$ a one hot encoded representation of the codes, ie. a very sparse matrix of size $N \times V$ where the entry (i, j) is the number of times the code j occurs for patient i. To take account of the temporality, we concatenate to the simple count matrix, a weighted decayed version, where each count has been multiplied by a exponentially decreasing weight: $C_{decay}(i,j) = C_{i,j} \cdot exp(-\delta(t_{i,j})/T)$. $\delta(t_{i,j})$ is the time delta between the start of followup for patient i and the event j. The caracteristic time T is set to 30 days. **b) In-domain embeddings** $[C \cdot \Phi_{train}, C_{decay} \cdot \Phi_{train}]$, an application of the the svd-ppmi, trained

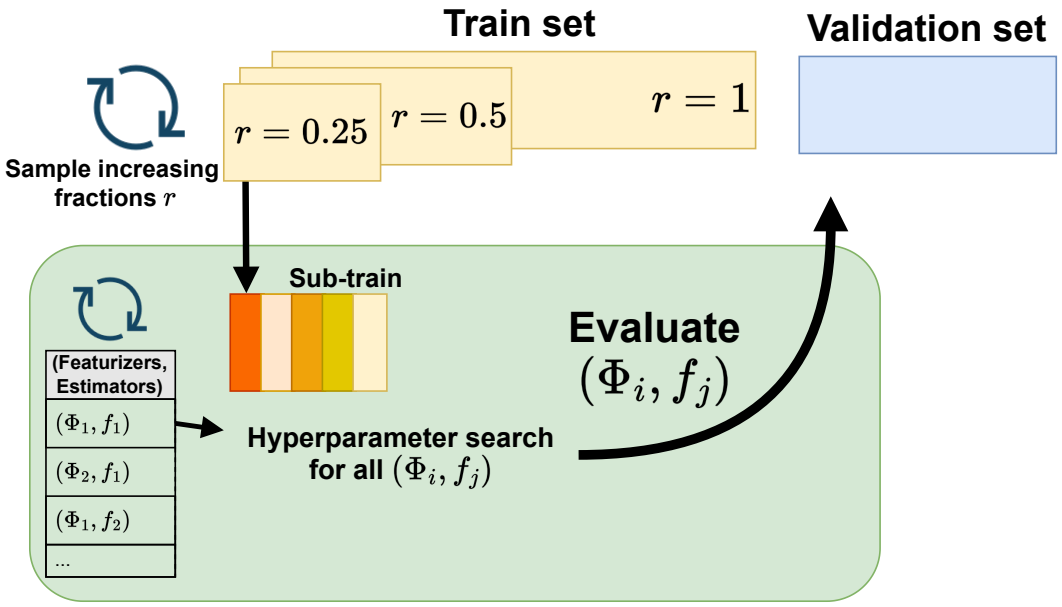

Figure 2: Growing train set selection procedure

on the restricted train set, **c) SNDS embeddings** $[C \cdot \Phi_{SNDS}, C_{decay} \cdot \Phi_{SNDS}]$, a transfer from the SVD-PPMI embeddings learned on the SNDS on the studied population.

We also compare adding SVD compression to the count vectorizer and the SNDS embeddings to reduce them to a small dimension (D=30).

For th estimators, we compare penalized logistic regression, random forests. For each of this estimator, the hyper-parameter have been cross-validated on the sub-train set.

3.3 Results

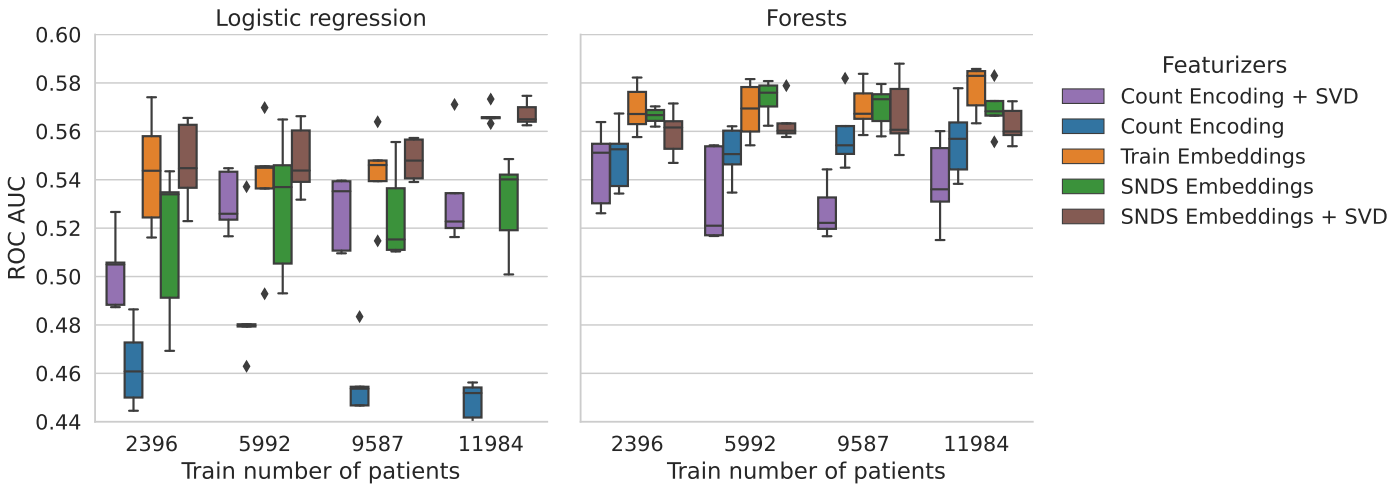

Figure 3: With a ridge estimator, **in domain embeddings are more performant than out of domain embeddings**. The event count featurizer has the worst performances, even when followed by SVD. Random forests smoothes this differences, certainly because it leverages the missing value mask of the count vectorizer. Overall, performances are low, underlying the difficulty of the task.

4 Perspectives

The easy creation and exchange of medical concept embeddings open a wide range of applications that efficiently pulls data from multiple sites, overcoming both the heterogeneity of the data, and the administrative barriers to access individual patient sequences. More work is needed to study these opportunities.

Another unexplored usage of concept embeddings is to conciliate raw heterogeneous medical codes into clinically meaningfull groups that can be used to characterize a population. The embeddings define a distance between medical concepts than can be used to group together codes used in similar contexts.

References

BACRY E., GAIFFAS S., LEROY F., MOREL M., NGUYEN D.-P., SEBIAT Y. & SUN D. (2020). Scalpel3: a scalable open-source library for healthcare claims databases. *International Journal of Medical Informatics*, **141**, 104203.

BEAM A. L., KOMPA B., SCHMALTZ A., FRIED I., WEBER G., PALMER N., SHI X., CAI T. & KOHANE I. S. (2019). Clinical concept embeddings learned from massive sources of multimodal medical data. In *Pacific symposium on biocomputing 2020*, p. 295–306: World Scientific.

BEAULIEU-JONES B. K., YUAN W., BRAT G. A., BEAM A. L., WEBER G., RUFFIN M. & KOHANE I. S. (2021). Machine learning for patient risk stratification: standing on, or looking over, the shoulders of clinicians? *NPJ digital medicine*, **4**(1), 62.

CHAZARD E., BALAYE P., BALCAEN T., GENIN M., CUGGIA M., BOUZILLÉ G. & LAMER A. (2022). Book music representation for temporal data, as a part of the feature extraction process: A novel approach to improve the handling of time-dependent data in secondary use of healthcare structured data. *Studies in Health Technology and Informatics*, **290**, 567–571.

DORIE V., HILL J., SHALIT U., SCOTT M. & CERVONE D. (2019). Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition.

DOUTRELIGNE M., LEDUC A., NGUYEN D.-P. & VUAGNAT A. (2020). Snds2vec, représentations continues pour les concepts médicaux du système national des données de santé. *Revue d'Épidémiologie et de Santé Publique*, **68**, S35.

HONG C., RUSH E., LIU M., ZHOU D., SUN J., SONABEND A., CASTRO V. M., SCHUBERT P., PANICKAN V. A., CAI T. *et al.* (2021). Clinical knowledge extraction via sparse embedding regression (keser) with multi-center large scale electronic health record data. *NPJ digital medicine*, **4**(1), 1–11.

HRIPCSAK G., DUKE J. D., SHAH N. H., REICH C. G., HUSER V., SCHUEMIE M. J., SUCHARD M. A., PARK R. W., WONG I. C. K., RIJNBEEK P. R. *et al.* (2015). Observational health data sciences and informatics (ohdsi): opportunities for observational researchers. In *MEDINFO 2015: eHealth-enabled Health*, p. 574–578. IOS Press.

LEVY O. & GOLDBERG Y. (2014). Neural word embedding as implicit matrix factorization. *Advances in neural information processing systems*, **27**.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. S. & DEAN J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, **26**.

PETIT-JEAN T., REMAKI A., MALADIÈRE V., VAROQUAUX G. & BEY R. (2023). eds-scikit: data analysis on omop databases. DOI : 10.5281/zenodo.7401549.

RAJKOMAR A., OREN E., CHEN K., DAI A. M., HAJAJ N., HARDT M., LIU P. J., LIU X., MARCUS J., SUN M. *et al.* (2018). Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, **1**(1), 18.

SALLOUM S., DAUTOV R., CHEN X., PENG P. X. & HUANG J. Z. (2016). Big data analytics on apache spark. *International Journal of Data Science and Analytics*, **1**(3), 145–164.

TUPPIN P., RUDANT J., CONSTANTINOU P., GASTALDI-MÉNAGER C., RACHAS A., DE ROQUEFEUIL L., MAURA G., CAILLOL H., TAJAHMADY A., COSTE J. *et al.* (2017). Value of a national administrative database to guide public decisions: From the système national d'information inter-régimes de l'assurance maladie (sniiram) to the système national des données de santé (snds) in france. *Revue d'epidemiologie et de sante publique*, **65**, S149–S167.

VAN DER MAATEN L. & HINTON G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, **9**(11).

WENDLING T., JUNG K., CALLAHAN A., SCHULER A., SHAH N. H. & GALLEGO B. (2018). Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. *Statistics in medicine*, **37**(23), 3309–3324.

A SVD-PPMI algorithm

We recall the SVD-PPMI algorithm developped by (Beam *et al.*, 2019) and used for transfering phenotyping in (Hong *et al.*, 2021).

The algorithm takes a sequence of coded events as input and outputs vector representations. As shown in Figure 4, it builds a context window around every event e, then update the cooccurrence matrix for the corresponding medical code $P(c, c_j) \forall\, c_j \in Vocabulary$.

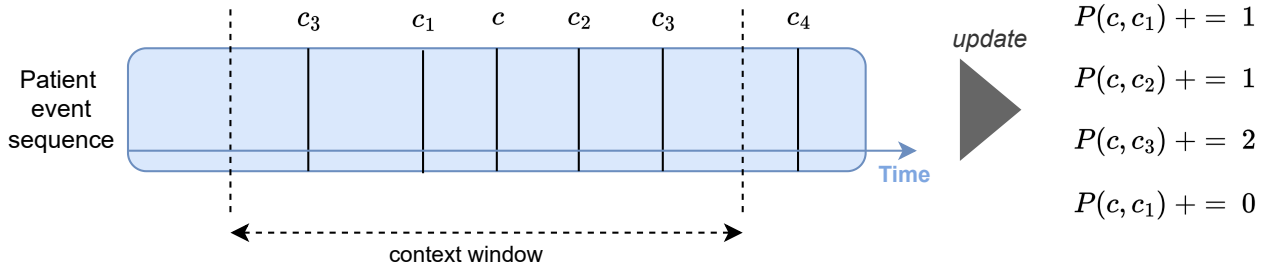

Figure 4: The cooccurrence matrix is updated when two events occur in a specified time window.

The PPMI matrix is then computed as the logged-shifted version of the cooccurrence matrix.

$$PMI(c_i, c_j) = log \frac{P(c_i, c_j)}{P(c_i)P(c_j)} \tag{1}$$

where $P(c_i)$ is the total count of the event c_i.

$$PPMI = max(0, PMI - log(k)) \tag{2}$$

Finally, concept embeddings are recovered by SVD factorisation and the mean .

$$PPMI = U \cdot \Sigma \cdot V \tag{3}$$

$$embeddings = U_d \cdot \sqrt{\Sigma_d} + V_d \cdot \sqrt{\Sigma_d} \tag{4}$$

Where the subscript d denotates the restriction to the first d components of the matrices.

B Empirical Study

B.1 Selection of the population

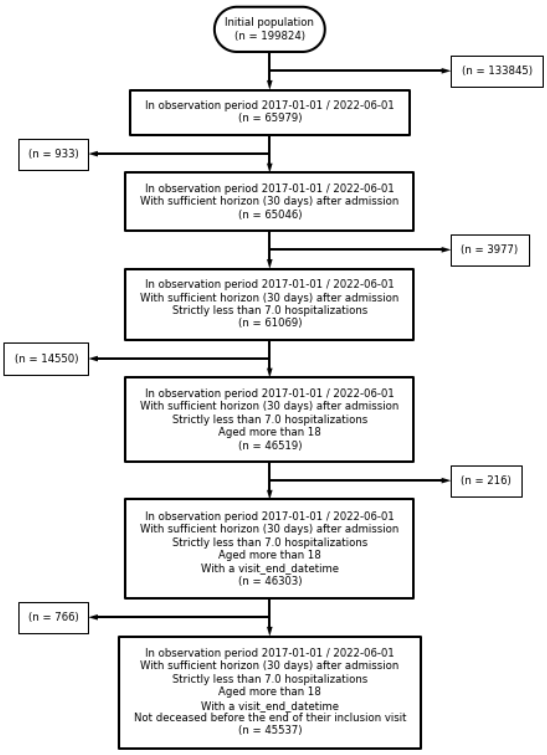

Figure 5: Cohort creation flowchart

Similarité des documents médicales en se basant sur des embeddings appris pour le codage médical

Leonardo Moros[1, 2, 3], Jérôme Azé[1], Sandra Bringay[1, 4], Pascal Poncelet[1], Maximilien Servajean[1, 4], Caroline Dunoyer[2, 5]

(1) LIRMM, UMR 5506, Université de Montpellier, CNRS, Montpellier, France

(2) Département d'Information Médicale, CHU Montpellier, Montpellier, France

(3) 5 DEGRÉS, Paris, France

(4) AMIS, Université Paul-Valéry, Montpellier, France

(5) IDESP, UMR UA11, INSERM - Université de Montpellier, Montpellier, France

RÉSUMÉ

Dans ce travail, nous utilisons la représentation apprise par un modèle qui fait du codage médical pour retrouver des patients (ou séjours) similaires.

ABSTRACT

Clinical document similarity based on embeddings learned for a clinical coding task

In this work we use the embeddings learned by a clinical coding model similar patients (or hospital visits).

MOTS-CLÉS : Codage médical, apprentissage automatique, similarité.

KEYWORDS: Clinical coding, machine learning, similarity.

1 Introduction

Plusieurs travaux ont été effectués pour la tâche du codage médical qui consiste à assigner des codes alphanumériques aux documents médicaux correspondant aux diagnostiques et traitements administrés à un patient lors d'un séjour à l'hôpital. Actuellement, l'état de l'art est un modèle appelé LAAT (Vu *et al.*, 2020) qui est basé sur un LSTM et des mécanismes d'attention qui génèrent une représentation du document par rapport à chaque code sur lequel le modèle a été entraîné. Dans ce travail, nous étudions la possibilité d'utiliser ces représentations pour trouver des séjours similaires.

2 Méthode

Nous avons entraîné LAAT en utilisant un split du MIMIC-III contentant les 1 000 codes les plus fréquents tel que décrit dans (Moros *et al.*, 2023). La représentation apprise contient 1 000 vecteurs (un par code) de 1 000 dimensions. Nous concaténons toutes les représentations dans un vecteur de 1 000 000 dimensions. Ensuite, nous appliquons l'algorithme PCA pour réduire le vecteur à 100 dimensions et obtenir la représentation finale du document. Pour évaluer notre méthode, nous appliquons l'algorithme UMAP (McInnes & Healy, 2018) pour réduire en 2 dimensions et dessinons

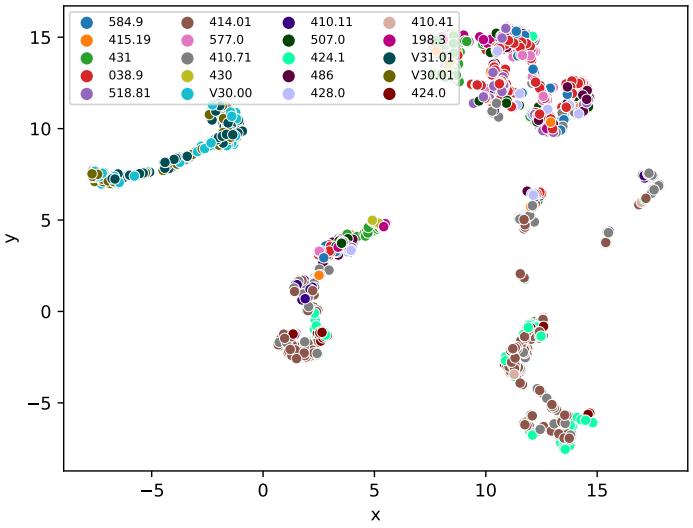

FIGURE 1 – UMAP sur nos représentations

	584.9	518.81	038.9	486
584.9	1	0.21	0.32	0.16
518.81	0.21	1	0.24	0.28
038.9	0.32	0.24	1	0.19
486	0.16	0.28	0.19	1

TABLE 1 – Matrice de correlation

un nuage de points où chaque point correspond à un compte rendu de sortie avec un code couleur montrant le diagnostic principal associé à ce séjour. Les résultats sont présentés dans la figure 1.

Nous pouvons remarquer la présence de clusters dans les données. Par exemple, à gauche, nous trouvons un cluster avec les codes V30.00 (*Single liveborn, born in hospital, delivered without mention of cesarean section*), V30.01 (*Single liveborn, born in hospital, delivered by cesarean section*) et V31.01 (*Twin birth, mate liveborn, born in hospital, delivered by cesarean section*) qui semble intéressant car ces 3 codes sont sémantiquement proches en raison du lien avec l'accouchement.

Nous trouvons aussi des codes regroupés car correlés comme, les codes 584.9 (*Acute kidney failure*), 518.81 (*Acute respiratory failure*), 038.9 (*Sepsis, unspecified organism*) et 486 (*Pneumonia*). Dans la table 1, nous voyons la matrice de corrélation pour ces 4 codes.

3 Conclusion

Ces résultats préliminaires montrent un potentiel intéressant dans la recherche des séjours similaires (e.g. construction de cohortes (Qian *et al.*, 2015), sous-typage de maladies (Li *et al.*, 2015), médecine personnalisée (Zhang *et al.*, 2014)) en se basant sur une représentation apprise pour une tâche de codage médical.

Remerciements

Ce projet a été soutenu par le LabEx NUMEV (ANR-10-LABX-0020) intégré à l'I-Site MUSE (ANR-16-IDEX-0006) et le CHU de Montpellier.

Références

LI L., CHENG W.-Y., GLICKSBERG B. S., GOTTESMAN O., TAMLER R., CHEN R., BOTTINGER E. P. & DUDLEY J. T. (2015). Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Science Translational Medicine*, **7**, 311ra174 – 311ra174.

MCINNES L. & HEALY J. (2018). Umap : Uniform manifold approximation and projection for dimension reduction. *ArXiv*, **abs/1802.03426**.

MOROS L., AZÉ J., BRINGAY S., PONCELET P., SERVAJEAN M. & DUNOYER C. (2023). Apprendre à classer des textes hospitaliers rédigés en anglais selon la classification cim-9 avec une approche par budget. *Extraction et Gestion des Connaissances*.

QIAN B., WANG X., CAO N., LI H. & JIANG Y.-G. (2015). A relative similarity based method for interactive patient risk prediction. *Data Mining and Knowledge Discovery*, **29**, 1070–1093.

VU T., NGUYEN D. Q. & NGUYEN A. N. (2020). A label attention model for icd coding from clinical text. In *IJCAI*.

ZHANG P., WANG F., HU J. & SORRENTINO R. (2014). Towards personalized medicine : Leveraging patient similarity and drug similarity analytics. *AMIA Summits on Translational Science Proceedings*, **2014**, 132 – 136.

Détection de zones dupliquées dans des comptes rendus médicaux

Thibaut Fabacher[1,5] Olivier Birot[1] Camila Arias-Villamil[1] Kim-Tâm Huynh[1]
Antoine Neuraz[1, 2, 3, 4] Bastien Rance[1, 2, 3, 4]

(1) Inria, HeKA, PariSantéCampus, Paris, France
(2) Université Paris Cité, Paris, France
(3) Centre de Recherche des Cordeliers, UMRS 1138, Inserm, Paris, France
(4) Assistance Publique - Hôpitaux de Paris, Paris, France
(5) CHU de Strasbourg, Strasbourg, France
bastien.rance@aphp.fr, kim-tam.huynh@inria.fr

RÉSUMÉ·

Des duplications peuvent être présentes dans les comptes rendus médicaux. Ces duplications sont souvent dues à des copier/coller de portions de textes anciens vers de nouveaux comptes rendus. Elles peuvent poser des problèmes lors de l'utilisation secondaire des textes et en particulier pour la gestion de la temporalité. Nous présentons une méthode de détection des duplications basée sur l'utilisation de n-gram de mots et son implémentation dans le framework de traitement de données de santé medkit.

ABSTRACT·

Duplications can be found in medical reports. Often due to old sections copy/pasted into new documents. These duplications can present signi cant challenges for the secondary use of the text data, especially for the detection and management of temporality. We present here a method to detect duplications using word n-grams, and its implementation with the framework of the biomedical data treatment library medkit.

MOTS-CLÉS : Traitement Automatique des Langues, Dossier Patient Informatisé, Duplications
KEYWORDS: Natural Language Processing, Electronic Health Record, Duplications

1. Duplications dans les dossiers patients

Les comptes rendus médicaux sont d'importants moyens d'échange entre les professionnels de santé. Les comptes rendus permettent de transmettre des informations factuelles sur l'état des patients, les décisions cliniques, mais également de discuter des hypothèses ou des remarques

diverses. Ces comptes rendus textuels sont cruciaux dans les échanges, diverses études ont montré qu'une grande partie de l'information médicale n'est présente que dans les textes (Neuraz et al., 2020).

Le processus d'écriture des comptes rendus fait souvent appel à des copier / coller : une portion d'un document ancien est copiée dans le document courant a n de conserver dans un document unique des informations sur l'historique de prise en charge de la patiente ou du patient ou simplement de reporter des antécédents médicaux. Dans une étude réalisée en 2018 sur un corpus de 650 000 documents médicaux (Digan et al., 2019), nous avions identi é un volume de duplication correspondant à environ 30% de la taille globale des documents.

Les duplications peuvent poser des problèmes lors de l'utilisation secondaire des données, et en particulier pour les tâches d'extraction d'information par méthodes de traitement automatique des langues (Cohen et al., 2013; Liu et al., 2022). Les duplications peuvent spéci quement perturber l'extraction et le travail sur les notions de temporalité. En e et, les sections dupliquées ne sont le plus souvent pas identi ées comme telles. La présence dans ces zones de marqueurs de temporalité peut être biaisée avec les références temporelles faussées. Un exemple simple est la présence de marqueurs de temporalité relative (e.g. il y a deux jours, dans trois mois).

Nous présentons ici une optimisation d'une méthode pour détecter e cacement les zones dupliquées dans des documents (Digan et al., 2019), ainsi que son implémentation dans le Framework medkit (https://heka.gitlabpages.inria.fr/medkit/).

2. Méthode

La section ci-dessous décrit la méthode de détection des duplications. En résumé, les textes sont découpés en n-grams et les duplications sont détectées en cherchant les éléments communs entre deux documents.

1. Pré-traitements

Nous supprimons les en-têtes et pieds de page des documents. Cette étape n'est pas indispensable mais réduit le temps de traitement des documents. Il faut noter que cette étape est très dépendante de la provenance des documents et doit être optimisée localement.
Deux étapes sont ensuite réalisées :

- Transformation de l'ensemble du texte en caractères minuscules
- Normalisation des espaces. Toute occurrence de plusieurs espaces successifs est normalisée en un espace simple.

1. Identi cation des zones dupliquées

L'objectif de la méthode est de découvrir les zones dupliquées et en particulier celles créées par mécanique de copier/coller. Nous travaillons donc en considérant indépendamment les documents de chaque patient, et considérons la liste de documents d'un patient classés chronologiquement par ordre de création.

Conditions de comparaisons. Nous comparons ensuite les documents deux à deux. La recherche est orientée, la paire de documents (i,j) est considérée seulement si la date de création de j est strictement supérieure à celle du document i. Le nombre de comparaisons est $n(n\text{-}1)/2$ avec n le nombre de documents.

Empreintes et fusion. Un document source et un document cible sont respectivement découpés en séquences de mots de taille n (des n-grams) qui vont constituer des empreintes. Chaque empreinte est associée à sa position dans le document et l'origine du document (source ou cible). En cherchant l'intersection des empreintes entre deux documents, on identi e des zones communes entre les deux documents, potentiellement dupliquées. Pour reconstruire les zones dupliquées on « fusionne » les empreintes contigües dans le document cible à l'aide d'un arbre intervalle.

2. Implémentation

Medkit est une bibliothèque Python ayant pour objectif de faciliter le traitement des données biomédicales (https://heka.gitlabpages.inria.fr/medkit/). Medkit est non-destructive : les traitements appliqués aux données ne changent pas les données sources, medkit embarque également une gestion ne de la provenance. Nous avons implémenté notre méthode de détection des duplications sous la forme d'un module medkit.

Références

Cohen, R., Elhadad, M., & Elhadad, N. (2013). Redundancy in electronic health record corpora: Analysis, impact on text mining performance and mitigation strategies. *BMC Bioinformatics*, *14*(1), 10. https://doi.org/10.1186/1471-2105-14-10

Digan, W., Wack, M., Looten, V., Neuraz, A., Burgun, A., & Rance, B. (2019). Evaluating the Impact of Text Duplications on a Corpus of More than 600,000 Clinical Narratives in a French Hospital. *Studies in Health Technology and Informatics*, *264*, 103–107. https://doi.org/10.3233/SHTI190192

Liu, J., Capurro, D., Nguyen, A., & Verspoor, K. (2022). "Note Bloat" impacts deep learning-based NLP models for clinical prediction tasks. *Journal of Biomedical Informatics*, *133*, 104149. https://doi.org/10.1016/j.jbi.2022.104149

Neuraz, A., Lerner, I., Digan, W., Paris, N., Tsopra, R., Rogier, A., Baudoin, D., Cohen, K. B., Burgun, A., Garcelon, N., & others. (2020). Natural language processing for rapid response to emergent diseases: Case study of calcium channel blockers and hypertension in the COVID-19 pandemic. *Journal of Medical Internet Research*, *22*(8), e20773.

Étude de similarité des patients pour identifier les unités hospitalières ayant le taux le plus élevé de réadmissions non planifiées

Nzamba Bignoumba[1] Sadok Ben Yahia[1] Nedra Mellouli[2]
(1) Tallinn University of Technology, Akadeemia tee 15a, 12618 Tallinn, Estonie
(2) LIASD University of Paris8, DVRC Devinci Group
Paris, France
nzamba.bignoumba@taltech.ee, sadok.ben@taltech.ee,
n.mellouli@iut.univ-paris8.fr

RÉSUMÉ

La réadmission non planifiée est un fardeau financier tant pour les patients que pour les prestataires de soins de santé. Si elle a un impact négatif sur le mode de vie du patient, elle donne également une mauvaise réputation aux prestataires de soins de santé. Pour alléger la charge des patients et des prestataires de soins, nous proposons un modèle qui prédit le risque de réadmission à la sortie de l'hôpital en se basant sur un encodage profond de l'état de santé du patient. Comme l'utilisation d'un seul patient n'est pas suffisamment cohérente pour identifier la ou les unités hospitalières (où services) responsables des admissions non planifiées, nous utilisons la représentation encodée des états de santé des patients pour calculer la similarité entre ces derniers et obtenir un meilleur modèle qui nous permet d'identifier ces unités hospitalières. Nous utilisons MIMIC-4 comme base de données pour l'étude expérimentale.

ABSTRACT

Patient similarity study to identify hospital units with the highest rate of unplanned readmissions.

Unplanned readmission is a financial burden for both patients and healthcare providers. While it has a negative impact on the patient's lifestyle, it also gives healthcare providers a bad reputation. To alleviate the burden on both patients and healthcare providers, we propose a model that predicts the risk of readmission at hospital discharge based on a deep encoding of the patient's health status. As the use of a single patient is not consistent enough to identify the hospital unit(s) (or services) responsible for unplanned admissions, we use the encoding representation to calculate the similarity of patients and obtain a better pattern that allows us to identify these hospital units. We use MIMIC-4 as the database for the experimental study.

MOTS-CLÉS : réadmission non planifiée, similarité entre les patients, modèle prédictif.

KEYWORDS: unplanned readmission, patients' similarities, predictive model.

1 Introduction

Advances in machine and learning (including deep learning) and the growing adoption of electronic records as a means of storing patient visit data have undoubtedly contributed to the explosion of varied and powerful models designed for medical problems such as identification of chronic cough patients (Luo *et al.*, 2021), coronavirus detection (Ghoshal & Tucker, 2020), cancer prediction (Xiao *et al.*, 2018b), mortality prediction (Narayan Shukla & Marlin, 2021), unplanned readmission prediction (Pishgar *et al.*, 2022) to name a few. Among the wide variety of medical problems addressed with machine learning algorithms, unplanned readmission prediction is probably one of the most discussed problems in the literature. Indeed, the growing interest in solving this problem is due to the fact that unplanned readmissions are a financial burden for both patients and healthcare providers. While they have a negative impact on the lifestyle and health of patients, they also give healthcare providers a bad name.

As the prediction of unplanned readmissions often involves processing clinical codes (those associated with diagnoses, medications, procedures and laboratory events) and/or clinical notes, the latter is usually approached, in whole or in part, as a natural language processing (NLP) task. Indeed, visits can be viewed as a set of ordered documents and clinical codes recorded during those visits as words. Based on this modeling, several model have been proposed. For example, (Xiao *et al.*, 2018a) and (Ashfaq *et al.*, 2019) proposed RNN-based model to predict the readmission of heart failure patients. As consecutive words (phrases) can be modeled by Recurrent Neural Network (RNN) which is one of the state-of-the-art models for sequence modelling, they modeled consecutive visits in the same fashion. Rather than using the clinical code as input features, other studies such as (Liu *et al.*, 2019; Huang *et al.*, 2019; Thapa *et al.*, 2022) relied on clinical notes. While (Liu *et al.*, 2019) used Convolutional Neural Network (CNN) for sequence modeling, (Huang *et al.*, 2019; Thapa *et al.*, 2022) used Transformer-based model (Vaswani *et al.*, 2017) which is considered as the state-of-the-art model for NLP tasks. What makes Transformer a powerful model is its attention mechanism that encodes the hidden relationship of the input features into real numbers. In addition, it incorporates a residual block (He *et al.*, 2016) that allows it to prevent gradient problems. Other works, such as (Low *et al.*, 2018; Doryab *et al.*, 2019; Qian *et al.*, 2021), used data collected from patients' wearable devices to predict unplanned readmissions.

Although the models listed above are functional, they did not model an important aspect, namely the irregular elapsed time between visits. Without taking this aspect into account, the models will give the same importance to medical features recorded in the past as to those recorded recently. Furthermore, these models are often built for a specific cohort of patients. (Bai *et al.*, 2018) proposed a model that takes into account the irregular elapsed time between visits. However, their model was intended for the prediction of diagnoses. Moreover, the approach they used to model irregular elapsed time between visits differs from ours.

Aware of the impact of the above-mentioned limitations on the task of predicting unplanned readmissions, we proposed a model composed of a Transformer-encoder block for better encoding of heterogeneous patients and a magnitude management block to weight current and historical medical events. As our main objective is to provide a more detailed report of the cause of unplanned readmission, we used the latent representation of patients computed through the proposed model for :

- Clustering of patients likely to be readmitted into two sub-cohorts of patients ;
- Identifying recurrent diseases in each sub-cohorts of patients ;
- Listing the different hospital units where the patients in each sub-cohorts were examined ;

— Conducting an analysis to check whether the medical profile of patients in each sub-cohort matches the hospital units in which they were examined ;
— Identification of hospital units responsible for unplanned readmissions.

These steps are carried out using Principal Component Analysis (PCA) and the Gaussian Mixture Model (GMM). Figure 1 summarizes the different components of the proposed model and their role.

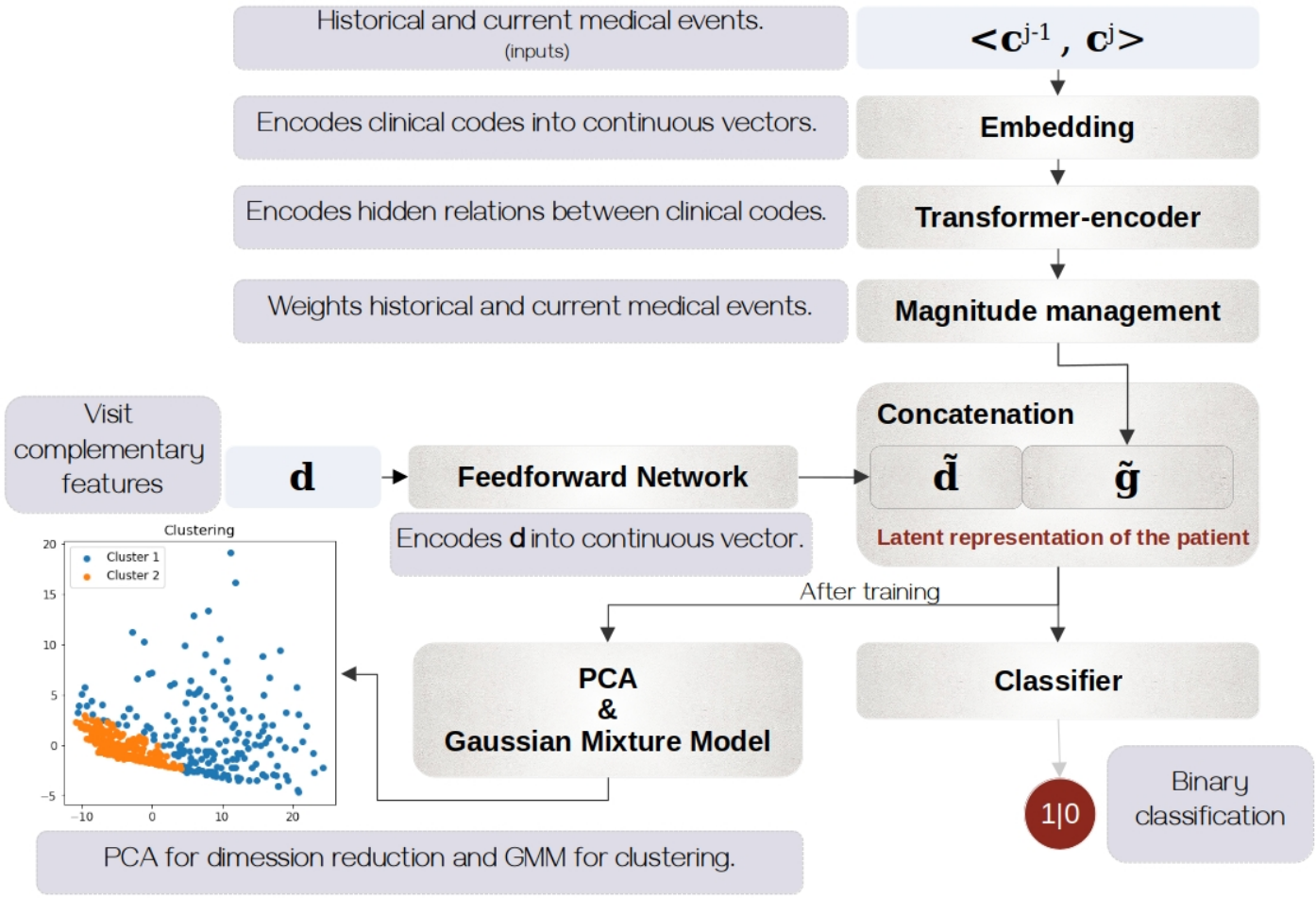

FIGURE 1 – Model's architecture. $< . >$ is the concatenation. $\mathbf{c}^{j-1}$ and $\mathbf{c}^{j}$ are the clinical codes of the previous visit (historical medical event) and the current visit (current medical event), respectively. As patient demographics (age, gender) and visit information (visit duration, type of visit) carry relevant information for prediction, we also consider them as additional input features. We call the set of these features visit complementary features, and represent it by a vector $\mathbf{d}$. $\tilde{\mathbf{d}}$ is the latent representation of the visit complementary features and $\tilde{\mathbf{g}}$ the patient's overall latent health status.

2 Preliminary results

In this section, we present the results obtained on the task of predicting unplanned readmissions and on the identification of hospital units causing unplanned readmissions. Before presenting the results, we describe the dataset used for the empirical evaluation.

2.1 Dataset

For the empirical evaluation, we used the publicly available MIMIC-IV database. MIMIC-IV is an updated version of MIMIC-III. It contains the anonymized health-related data of more than forty thousand patients who stayed in Intensive Care Units of Beth Israel Deaconess Medical Center between 2001 and 2012. From this database, we initially extract data from 180747 patients. After filtering, 65904 admissions were used for the study. If the admission that follows the current admission occurs before 30 days, the label associated with the current admission is 1, i.e. an admission that will lead to an unplanned readmission. Otherwise, the label is 0, i.e. an admission that will not lead to an unplanned readmission. On the basis of this criterion, we obtained 46979 planned readmissions and $18925(40.28\%)$ unplanned readmissions.

2.2 Model performance

As we are dealing with an unbalanced dataset, we compare the average Area Under the ROC Curve (AUC) and the Area Under the Precision-Recall Curve (AUPRC) scores over the 5-fold cross validation that we obtained against those of our competitors. Long Short-Term Memory (LSTM), CNN-LSTM (CNN stands for Convolutional Neural Network), GRU-Decay (GRU stands for Gated Recurrent Unit) and Feedforward Neural Network (FNN) are the competing models that we also evaluate to benchmark our model. All the models aforementioned incorporate an embedding layer and have the same classifier as ours. From Table 1, we can see that our model outperforms all competitors.

Models	AUC	AUPRC
GRU-Decay	0.678 ± 0.005	0.487 ± 0.006
FNN	0.667 ± 0.005	0.476 ± 0.004
CNN-LSTM	0.675 ± 0.007	0.487 ± 0.007
LSTM	0.670 ± 0.006	0.479 ± 0.005
Ours wo\mm	0.663 ± 0.009	0.475 ± 0.008
Ours	$\mathbf{0.681 \pm 0.005}$	$\mathbf{0.490 \pm 0.007}$

TABLE 1 – AUC and AUPRC scores on the unplanned readmission prediction task. "wo\mm" stands for without the magnitude management block.

These results confirm our assertion on the importance of using a component capable of encoding the hidden relations between the clinical codes and another component in charge of weighting the historical events according to the dates on which they were recorded.

2.3 Identification of hospital units causing unplanned readmissions

In this subsection, we used the latent representation of patients calculated by the model to identify the hospital units responsible for unplanned readmissions. In order to obtain reliable results, we only consider patients whose probability of unplanned readmission is equal to or greater than 0.8. After this filtering, we obtained 577 patients. Using the Gaussian Mixture Model, we cluster these patients into two sub-cohorts of patients, namely, cluster 1 and cluster 2. Table 2 shows the 10 most frequent diagnoses in each cluster. With the help of a practitioner, it was found that the cluster 1 may involve patients suffering from diseases caused by poor nutrition, while the cluster 2 may involve patients

suffering from depression. It is then suggested that patients in the cluster 1 be examined in the hospital unit responsible for diseases related to poor nutrition, while those of the cluster 2 be examined in the hospital unit dedicated to psychiatric diseases. Figure 2 summarizes the number of examinations per hospital unit for both clusters.

Although it is normal for patients to consult a general practitioner first before going to a specialist (if necessary), we find that the number of examinations in the Medical-general service for internal medicine (MED) is far too large compared to the other department. Based on our analysis above, we would have expected to have a similar number of examinations in the psychiatry department (PSYCH) as in MED, for cluster 2, but this is not the case. Concerning cluster 1, in the absence of specific hospital units specifically dedicated to patients suffering from nutrition-related diseases, we cannot draw any conclusions at this stage. A more detailed analysis is required.

From the analysis conducted on cluster 2, we assume that patients with depression are prone to unplanned readmissions due to misdiagnoses by general practitioners. We therefore recommend a restructuring of the Medical-general service (MED) so that patients with depression are better identified and referred to the psychiatric department. It should be emphasized that this is only a recommendation. The aim of our study is to help health staff to strengthen or possibly restructure hospital units in order to reduce the rate of unplanned admissions, not to replace them in the decision-making process.

Top ten diagnoses

Cluster 1	Cluster 2
Other disorders of fluid, electrolyte and acid-base balance ; Place of occurrence of the external cause ; Personal history of other diseases and conditions ; Disorders of lipoprotein metabolism and other lipidaemias ; Long term (current) drug therapy ; Essential (primary) hypertension ; Pedal cycle rider injured in collision with railway train or railway vehicle ; Panic disorder [episodic paroxysmal anxiety] ; Type 2 diabetes mellitus	Major depressive disorder, single episode ; Symptoms and signs involving emotional state ; Occupant of heavy transport vehicle injured in collision with two- or three-wheeled motor vehicle ; Depressive disorder, not elsewhere classified ; Depressive disorder, not elsewhere classified ; Other anxiety disorders ; Cannabis related disorders ; Reaction to severe stress, and adjustment disorder ; Nicotine dependence

TABLE 2 – The ten most frequent diagnoses in each cluster.

3 Conclusion

In this study, our main concern was to predict unplanned readmissions and to identify the hospital units that may be the source of these unplanned readmissions. We show that better encoding of the hidden

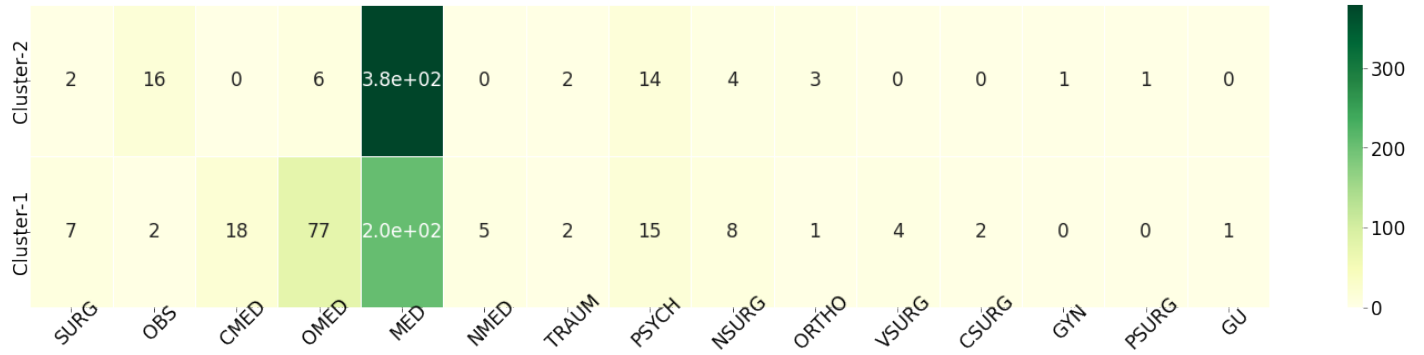

FIGURE 2 – Distribution of the number of examinations in the different hospital units. SURG : Surgical - general surgical service not classified elsewhere ; OBS : Obstetrics - conerned with childbirth and the care of women giving birth ; CMED : Cardiac Medical - for non-surgical cardiac related admissions ; OMED : Orthopaedic medicine - non-surgical, relating to musculoskeletal system ; MED : Medical - general service for internal medicine ; NMED : Neurologic Medical-non-surgical, relating to the brain ; TRAUM : Trauma-injury or damage caused by physical harm from an external source ; PSYCH : Psychiatric-mental disorders relating to mood, behaviour, cognition, or perceptions ; NSURG : Neurologic Surgical - surgical, relating to the brain ; ORTHO : Orthopaedic-surgical, relating to the musculoskeletal system ; VSURG : Vascular Surgical - surgery relating to the circulatory system ; CSURG : Cardiac Surgery-for surgical cardiac admissions ; GYN : Gynecological-female reproductive systems and breasts ; PSURG : Plastic-restortation/reconstruction of the human body (including cosmetic or aesthetic) ; GU : Genitourinary-reproductive organs/urinary system.

relation of clinical codes and weighting of historical medical events recorded long ago improves the accuracy of the model. We have also shown that it is possible to leverage the latent representation of patients to provide a more detailed report of possible causes of unplanned readmission.

Although we obtained satisfactory results, the AUC and AUPRC scores are not high enough for the model to be deployed in real situations. At this stage, it is simply an experimental model. We admit that, although we have involved a practitioner in this study, it is imperative to involve more in order to obtain a more thorough analysis and, therefore, more meaningful conclusions. Another limitation of this work is that we only considered structured data. Unstructured data such as clinical notes that are very informative should also be considered as input features. We believe that a slight modification of our model, namely the integration of the position encoding component, to process clinical notes, will result in better performance.

Acknowledgment

Our work, has been partially conducted in the project "ICT program" which was supported by the European Union through the European Social Fund.

Références

ASHFAQ A., SANT'ANNA A., LINGMAN M. & NOWACZYK S. (2019). Readmission prediction

using deep learning on electronic health records. *Journal of biomedical informatics*, **97**, 103256.

BAI T., ZHANG S., EGLESTON B. L. & VUCETIC S. (2018). Interpretable representation learning for healthcare via capturing disease progression through time. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, p. 43–51.

BENAMARA F., HATOUT N., MULLER P. & OZDOWSKA S., Éds. (2007). *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, Toulouse. ATALA, IRIT.

DIAS G., Éd. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

DORYAB A., DEY A. K., KAO G. & LOW C. (2019). Modeling biobehavioral rhythms with passive sensing in the wild : a case study to predict readmission risk after pancreatic surgery. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, **3**(1), 1–21.

GHOSHAL B. & TUCKER A. (2020). Estimating uncertainty and interpretability in deep learning for coronavirus (covid-19) detection. *arXiv preprint arXiv :2003.10769*.

HE K., ZHANG X., REN S. & SUN J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 770–778.

HUANG K., ALTOSAAR J. & RANGANATH R. (2019). Clinicalbert : Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv :1904.05342*.

LAIGNELET M. & RIOULT F. (2009). Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs. In A. NAZARENKO & T. POIBEAU, Éds., *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis : ATALA LIPN.

LANGLAIS P. & PATRY A. (2007). Enrichissement d'un lexique bilingue par analogie. In (Benamara *et al.*, 2007), p. 101–110.

LIU X., CHEN Y., BAE J., LI H., JOHNSTON J. & SANGER T. (2019). Predicting heart failure readmission from clinical notes using deep learning. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, p. 2642–2648 : IEEE.

LOW C. A., BOVBJERG D. H., AHRENDT S., CHOUDRY M. H., HOLTZMAN M., JONES H. L., PINGPANK JR J. F., RAMALINGAM L., ZEH III H. J., ZUREIKAT A. H. *et al.* (2018). Fitbit step counts during inpatient recovery from cancer surgery as a predictor of readmission. *Annals of Behavioral Medicine*, **52**(1), 88–92.

LUO X., GANDHI P., ZHANG Z., SHAO W., HAN Z., CHANDRASEKARAN V., TURZHITSKY V., BALI V., ROBERTS A. R., METZGER M. *et al.* (2021). Applying interpretable deep learning models to identify chronic cough patients using ehr data. *Computer Methods and Programs in Biomedicine*, **210**, 106395.

NARAYAN SHUKLA S. & MARLIN B. M. (2021). Multi-time attention networks for irregularly sampled time series. *arXiv e-prints*, p. arXiv–2101.

PISHGAR M., THEIS J., DEL RIOS M., ARDATI A., ANAHIDEH H. & DARABI H. (2022). Prediction of unplanned 30-day readmission for icu patients with heart failure. *BMC Medical Informatics and Decision Making*, **22**(1), 1–12.

QIAN C., LEELAPRACHAKUL P., LANDERS M., LOW C., DEY A. K. & DORYAB A. (2021). Prediction of hospital readmission from longitudinal mobile data streams. *Sensors*, **21**(22), 7510.

SERETAN V. & WEHRLI E. (2007). Collocation translation based on sentence alignment and parsing. In (Benamara *et al.*, 2007), p. 401–410.

THAPA N. B., SEIFOLLAHI S. & TAHERI S. (2022). Hospital readmission prediction using clinical admission notes. In *Australasian Computer Science Week 2022*, p. 193–199.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł. & POLOSUKHIN I. (2017). Attention is all you need. *Advances in neural information processing systems*, **30**.

XIAO C., MA T., DIENG A. B., BLEI D. M. & WANG F. (2018a). Readmission prediction via deep contextual embedding of clinical concepts. *PloS one*, **13**(4), e0195024.

XIAO Y., WU J., LIN Z. & ZHAO X. (2018b). A deep learning-based multi-model ensemble method for cancer prediction. *Computer methods and programs in biomedicine*, **153**, 1–9.

Exploring Similarities and Dissimilarities in Patient Representations for Analogical Reasoning

Safa Alsaidi [1, 2], Miguel Couceiro [3], Nicolas Garcelon [1,2,4,5] and Adrien Coulet [1, 2]
(1) Inria Paris, Paris; (2) CRC, Inserm, Université Paris Cité, Sorbonne Université, Paris ; (3) LORIA, CNRS, Universite de Lorraine, Nancy ; (4) Imagine Institute, Paris ; (5) Service d'Informatique Biomédicale, Hôpital Necker-Enfants Malades, Assistance Publique - Hôpitaux de Paris, Paris, France

Background : An analogical proportion, or an analogy, is a relation between four objects A, B, C, and D that reads as "A is to B as C is to D". Two typical tasks associated with Analogy Reasoning (AR) are *analogy detection* and *analogy solving*. Analogy detection corresponds to the task of deciding whether a quadruple $\langle A, B, C, D\rangle$ is a valid analogy. Analogy solving corresponds to the task of finding an x for a given triple A, B, C such that $A : B :: C : x$ is a valid analogy. Analogies are classified and grouped based on the type of relation that exists between the pairs of objects and can have various meanings depending on what the four objects or concepts resemble. Therefore, analogies have been applied to a variety of reasoning and classification tasks in and out NLP such as mining paradigm tables in linguistics or image generation (Fam & Lepage, 2016; Reed *et al.*, 2015) and have achieved promising results. Because analogical reasoning is a cognitive process used in clinical practice, we think its automation may find numerous applications in the healthcare domain.

Method : We explored the potential of adapting the analogy framework to solve two biomedical tasks. In the first one, we addressed the classical patient matching (or record linkage) task by answering the question "does a hospital stay belong to a patient ?" In the second one, we addressed the task of disease prognosis, *i.e.* "will a certain disease develop in the same way for two distinct patients." To investigate these tasks, we defined two settings of the analogical framework that we named *(i)* Identity ; *(ii)* Identity + Sequent settings. For each setting, we introduced sets of constraints over objects A, B, C, D that need to be fulfilled to establish valid analogies.

To investigate these settings, we proposed to leverage the availability of large sets of Electronic Health Records (EHRs) data. EHRs are complex data that combine structured data (*e.g.*, diagnostic codes, drug prescriptions, laboratory results) and unstructured data (*e.g.*, clinical texts, images) in two dimensions : patient and time. From these data, we built sets of analogies, where objects are patient-stay representations *i.e.*, a numeric vector representation of EHRs data that belong to a single hospital stay. Representation learning has enabled notable progress in natural language processing as it focuses on obtaining compact vector representation of complex data. Similar representations can also be acquired from patient data to solve tasks such as predicting readmission, diagnosis, or length of stays.

We presented our patient-stay analogies as quadruples of four stays $(s^{i_1}_{t_1}, s^{i_1}_{t_2}, s^{i_2}_{t_3}, s^{i_2}_{t_4})$. For the Identity setting, we defined only one constraint that forces each pair of two stays to belong to a single patient i_j. Accordingly, analogies of this setting took the form of *Alice's visit of last January : Alice's visit of last June : : Bob's last visit : Bob's visit of 3/25/2025* . In the Identity + Sequent setting, we added a temporal constraint, $t_1 < t_2 \wedge t_3 < t_4$ that forces, in each pair, the first visit to occur before the second. We also added a diagnosis constraint that forces $s^{i_1}_{t_1}$ and $s^{i_2}_{t_3}$ to be associated with a same (or similar) diagnosis. Here analogies took the form of *Alice's Jan. 22 visit, with diagnostic A : Alice's*

Mar. 22 visit with diagnostic B : : Bob's Oct. 12 visit with diagnostic A : Bob's visit of Dec. 13 with diagnostic D. Accordingly, we built for each setting a set of valid and invalid analogies that either fulfil or violate these constraints. To augment the number our analogies, we used analogical properties (*e.g.*, reflexivity, symmetry) to generate more valid analogies.

In this preliminary study, we only addressed the analogy detection task to investigate healthcare applications of the analogy based framework with patient representations learned from EHRs. Our approach chains two neural network models as depicted in Figure 1. First, patient-stay representations are learned with the Fusion CNN embedding model developed by (Zhang *et al.*, 2020), which enables combining both structured and unstructured EHRs data. Second, analogy detection is based on a CNN classification model inspired by (Lim *et al.*, 2019; Alsaidi *et al.*, 2021), which we adapted to patient-stay representations. This last model determines whether a given quadruple (A, B, C, D) constitutes a valid analogy. For the first setting, we used the analogy detection to determine if a particular stay belongs to a certain patient. For the second setting, we used the analogy detection to determine if it is plausible for a patient's condition to evolve a certain way.

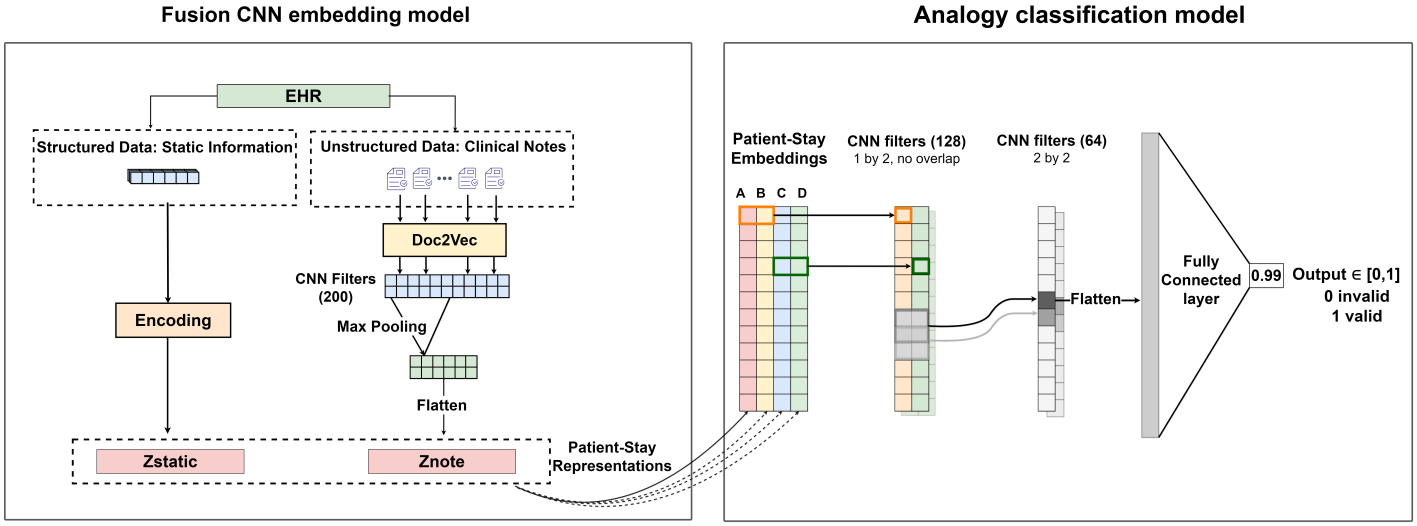

FIGURE 1 – The chaining of two models for analogy detection from patient-stay representations.

Results : We experimented with EHRs from the MIMIC-III database (Johnson *et al.*, 2016) as a source of patient data and considered both structured and unstructured data to define our analogies. The dataset contains data associated with around 40,000 patients admitted to the ICU (Intensive Care Unit) of Beth Israel Deaconess Medical Center between 2001 and 2012.

For each setting we compared the contribution of each type of data (structured only, unstructured only, or structured and unstructured). Results showed that the accuracy is the highest for valid analogies when the patient-stay representations are made of the concatenation of both types of data. This indicates that adding or using static information when learning patient-stay representations improves the performance of our model : it allows the model to better distinguish the stays and match them to the patient they belong to. Detailed experimental design and results are available in (Alsaidi *et al.*, 2022a,b).

Conclusion : In this abstract we present experimental settings to study how the analogy framework can contribute to solving two tasks relevant to the healthcare domain : patient matching and disease prognosis. Although preliminary, we believe that the analogy-based settings we defined will serve as building blocks to further investigate the applications of analogical reasoning in this domain.

Références

ALSAIDI S., COUCEIRO M., MARQUER E., QUENNELLE S., BURGUN A., GARCELON N. & COULET A. (2022a). An analogy based framework for patient-stay identification in healthcare. In *Proceedings of the ICCBR Workshop on Analogies : from Theory to Applications ATA 2022, Nancy, France, September 2022*.

ALSAIDI S., COUCEIRO M., QUENNELLE S., BURGUN A., GARCELON N. & COULET A. (2022b). Exploring analogical inference in healthcare. In M. COUCEIRO & P. MURENA, Éds., *Proceedings of the IJCAI-ECAI Workshop on the Interactions between Analogical Reasoning and Machine Learning, IARML 2022, Vienna, Austria, July 23, 2022*, volume 3174 de *CEUR Workshop Proceedings*, p. 40–50 : CEUR-WS.org.

ALSAIDI S., DECKER A., LAY P., MARQUER E., MURENA P.-A. & COUCEIRO M. (2021). A neural approach for detecting morphological analogies. In *Proceedings of the 8th IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, p. 1–10.

FAM R. & LEPAGE Y. (2016). Morphological predictability of unseen words using computational analogy. In *Workshops Proceedings for the Twenty-fourth International Conference on Case-Based Reasoning (ICCBR)*, volume 1815, p. 51–60.

JOHNSON A. E. W., POLLARD T. J., SHEN L., WEI H. LEHMAN L., FENG M., GHASSEMI M. M., MOODY B., SZOLOVITS P., CELI L. A. & MARK R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific Data*, **3**.

LIM S., PRADE H. & RICHARD G. (2019). Solving word analogies : A machine learning perspective. In *Proceedings of the Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU)*, volume 11726, p. 238–250.

REED S. E., ZHANG Y., ZHANG Y. & LEE H. (2015). Deep visual analogy-making. In C. CORTES, N. D. LAWRENCE, D. D. LEE, M. SUGIYAMA & R. GARNETT, Éds., *Advances in Neural Information Processing Systems 28 : Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, p. 1252–1260.

ZHANG D., YIN C., ZENG J., YUAN X. & ZHANG P. (2020). Combining structured and unstructured data for predictive models : a deep learning approach. *BMC Medical Informatics Decision Making*, **20**(1), 280. DOI : 10.1186/s12911-020-01297-6.

Similarité surfacique et similarité sémantique dans des cas cliniques générés

Nicolas Hiebel[1] Olivier Ferret[2] Karën Fort[3] Aurélie Névéol[1]
(1) Université Paris-Saclay, CNRS, LISN, 91400, Orsay, France
(2) Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
(3) Sorbonne Université / Université de Lorraine, CNRS, Inria, LORIA, 54506, Vandœuvre-lès-Nancy, France
[1]prenom.nom@lisn.upsaclay.fr, [2]olivier.ferret@cea.fr,
[3]karen.fort@loria.fr

MOTS-CLÉS : Génération, Similarité, Texte clinique, Texte synthétique, Français.

KEYWORDS: Generation, Similarity, Clinical Text, Synthetic Text, French.

Contexte La disponibilité restreinte des documents cliniques est un frein à la recherche en traitement automatique de la langue dans le domaine médical. Les corpus cliniques dont l'accès est relativement facile en français (E3C (Magnini *et al.*, 2020), CAS (Grabar *et al.*, 2018)) ne sont pas tout à fait représentatifs des documents confidentiels présents dans les hôpitaux. Le partage des connaissances au sein de la communauté scientifique est compliqué. Aucune reproductibilité n'est possible, tout comme les comparaisons avec d'autres méthodes / données. Une piste de création de ressource partageable en substitut des données confidentielles est la génération de données similaires à ces données privées. Cela pourrait permettre à des personnes ayant accès à un corpus privé de générer un corpus librement distribué à partir du premier. En partageant la méthode de génération, il serait également possible de reproduire l'expérience sur d'autres données confidentielles. La mise à disposition des données générées donnerait alors à la communauté scientifique un terrain de test, de comparaison, de discussion et d'entraide dans la recherche en TAL biomédical.

Les métriques existantes d'évaluation automatique de la génération de texte reposent majoritairement sur des mesures de similarité avec des références : une réponse idéale dans un contexte donné. Cette référence est comparée avec la ou les hypothèses du système (BLEU (Papineni *et al.*, 2002), ROUGE (Lin, 2004), BERTScore (Zhang *et al.*, 2020)). Nous présentons ici des alternatives à ces mesures pour évaluer une génération ouverte dans le domaine médical.

Objectifs L'évaluation de la génération est multidimensionnelle. Il est nécessaire d'observer la qualité de la langue dans les données générées, ainsi que son adéquation avec la langue des données sources. Dans cette étude, nous nous concentrons sur l'évaluation de la similarité entre des données médicales réelles et des données synthétiques générées à partir de ces données réelles selon deux critères complémentaires : la similarité surfacique et la similarité sémantique. Ces similarités sont importantes à estimer : une trop grande proximité pourrait signifier une copie d'information dans le corpus synthétique, représentant un potentiel risque de fuite d'information confidentielle. Inversement, il ne faut pas que les données synthétiques s'éloignent trop des données réelles pour que leur étude soit pertinente. Il y aura donc forcément des ressemblances. Nous cherchons par conséquent à repérer les éléments identiques ou ressemblants dans les deux jeux de données à l'aide de mesures de similarité,

en essayant de différencier les similarités traduisant un risque et les similarités acceptables.

Méthodologie Pour cette expérience, nous utilisons le corpus E3C, un corpus multilingue de documents biomédicaux librement disponible dans lequel nous sélectionnons les cas cliniques en français. Nous le désignerons par $\texttt{E3C}_{FR}$. Nous générons quatre corpus de données synthétiques de taille similaire au corpus $\texttt{E3C}_{FR}$ à l'aide de quatre configurations de modèles de langue génératifs pré-entraînés et adaptés (*fine-tune*) aux données cliniques de $\texttt{E3C}_{FR}$. Les modèles sont entraînés à générer des documents entiers en encadrant le texte généré par des balises de début et de fin du document. Nous étudions la similarité entre les données synthétiques générées et les données réelles à deux niveaux. Au niveau lexical, en observant les recouvrement de ngrammes entre les deux corpus, ce qui nous permet d'estimer la quantité de ngrammes que le modèle génératif a observé à l'entraînement et généré à l'identique à l'inférence. En complément, une comparaison avec un autre corpus de cas cliniques en français est ajoutée avec le corpus CAS. Deuxièmement, nous étudions la proximité sémantique entre les phrases des corpus. Pour cela, nous calculons les plongements des phrases des corpus à l'aide de l'outil SENTENCE-BERT (Reimers & Gurevych, 2019). Le modèle SENTENCE-BERT est ajusté sur le corpus de paires de phrases cliniques annotées en similarité CLISTER (Hiebel *et al.*, 2022) pour s'adapter au domaine clinique. Nous utilisons ensuite les plongements de phrases des deux corpus pour calculer une matrice de similarité entre les phrases des corpus, ce qui permet de récupérer pour chaque phrase du corpus généré les phrases les plus similaires dans le corpus réel avec les scores de similarité associés. C'est à partir de ces scores que nous essayons de déterminer les phrases et documents potentiellement problématiques du point de vue de la confidentialité.

Résultats Le tableau 1 présente les taux de recouvrement de ngrammes entre des corpus générés dans des configurations différentes et $\texttt{E3C}_{FR}$. Nous avons ajouté la comparaison avec le corpus CAS. On remarque que le corpus CAS présente plus d'unigrammes en commun avec le corpus $\texttt{E3C}_{FR}$. En revanche, il présente moins de séquences longues (8grammes) en commun avec ce même $\texttt{E3C}_{FR}$ que plusieurs corpus générés. On remarque également des différentes importantes entre les différents corpus générés. Le tableau 1 donne par ailleurs la similarité sémantique moyenne des phrases des corpus générés (et de CAS), calculée comme décrite à la section précédente. On observe ainsi que le corpus $\texttt{Bloom}_{E3C+T}$ présente une similarité moyenne plus élevée que les autres et le corpus $\texttt{LLF}_{E3C}$ la moins élevée, proche même de la proximité obtenue sur le corpus réel CAS.

Corpus	1gramme	4gramme	8gramme	Sim Sém
$\texttt{Bloom}_{E3C}$	0,16419	0,00368	0,00011	0,561
$\texttt{Bloom}_{E3C+T}$	0,13887	0,00531	0,00020	0,585
$\texttt{LLF}_{E3C}$	0,11740	0,00447	0,00023	0,531
$\texttt{LLF}_{E3C+T}$	0,11935	0,00505	0,00013	0,557
CAS	0,20373	0,00899	0,00013	0,530

TABLEAU 1 – Comparaison entre les différents corpus générés et $\texttt{E3C}_{FR}$. Nous avons ajouté la comparaison entre $\texttt{E3C}_{FR}$ et CAS. Sim Sém = Similarité Sémantique moyenne.

Conclusion Ce travail donne un premier aperçu de la similarité surfacique et sémantique de corpus cliniques générés et naturels, montrant notamment que ces deux plans ne sont pas strictement corrélés.

Références

GRABAR N., CLAVEAU V. & DALLOUX C. (2018). CAS : French corpus with clinical cases. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, p. 122–128, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/W18-5614.

HIEBEL N., FERRET O., FORT K. & NÉVÉOL A. (2022). CLISTER : A corpus for semantic textual similarity in French clinical narratives. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 4306–4315, Marseille, France : European Language Resources Association.

LIN C.-Y. (2004). ROUGE : A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, p. 74–81, Barcelona, Spain : Association for Computational Linguistics.

MAGNINI B., ALTUNA B., LAVELLI A., SPERANZA M. & ZANOLI R. (2020). The E3C Project : Collection and Annotation of a Multilingual Corpus of Clinical Cases. In J. MONTI, F. DELL'ORLETTA & F. TAMBURINI, Éds., *Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020, Bologna, Italy, March 1-3, 2021*, volume 2769 de *CEUR Workshop Proceedings* : CEUR-WS.org.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). BLEU : a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, p. 311–318, Philadelphia, Pennsylvania, USA : Association for Computational Linguistics. DOI : 10.3115/1073083.1073135.

REIMERS N. & GUREVYCH I. (2019). Sentence-BERT : Sentence Embeddings using Siamese BERT-Networks. In K. INUI, J. JIANG, V. NG & X. WAN, Éds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, p. 3980–3990 : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1410.

ZHANG T., KISHORE V., WU F., WEINBERGER K. Q. & ARTZI Y. (2020). Bertscore : Evaluating text generation with BERT. In *International Conference on Learning Representations (ICLR)*, online.

Patient-patient similarity-based screening of a clinical data warehouse to support rare disease diagnosis

Xiaoyi Chen[1, 2, 3] Carole Faviez[2, 3] Anita Burgun[1, 2, 3, 4] Nicolas Garcelon[1, 2, 3]
(1) Data Science Platform, Imagine Institute, Université de Paris, Inserm UMR 1163, 24 Boulevard du Montparnasse, 75015 Paris, France
(2) Inserm, Centre de Recherche des Cordeliers, Sorbonne Université, Université de Paris Cité, 15 Rue de l'école de médecine, 75006 Paris, France
(3) HeKA, Inria Paris, 2-10 Rue d'Oradour-sur-Glane, 75015 Paris, France
(4) Department of medical informatics, Hôpital Necker-Enfant Malades, AP-HP, 149 Rue de Sèvres, Paris, France
xiaoyi.chen@institutimagine.org, carole.faviez@inserm.fr, anita.burgun@aphp.fr, nicolas.garcelon@institutimagine.org

ABSTRACT

Background: A timely diagnosis is a key challenge for many rare diseases (Faviez *et al.*, 2020). As an expanding group of rare and severe monogenic disorders with a broad spectrum of clinical manifestations, ciliopathies, notably renal ciliopathies, suffer from important underdiagnosis issues. The major obstacles include the phenotypic and genetic heterogeneity of ciliopathies, and the growing pace at which new clinical phenotypes are being described. In such situation, patient-patient similarity measures may be useful to search for potential ciliopathy patients in clinical data warehouses (Chen *et al.*, 2019). The wide adoption of electronic health records (EHR) systems in hospitals enables the reuse and mining of patient data collected during care to support diagnosis (Garcelon *et al.*, 2017, 2020). To do so, the similarity model is required to be able to handle the semantic relations between medical concepts extracted from EHRs and their different levels of relevance, which is important due the characteristics of EHR data: incompleteness, inaccurate phenotyping, noisy phenotypes related to multiple comorbidities and medical histories. In the context of C'IL-LICO program, our objective is to develop an approach for screening large-scale clinical data warehouses and detecting patients with similar clinical manifestations to those from diagnosed ciliopathy patients. We expect that the top-ranked similar patients will benefit from genetic testing for an early diagnosis.
Method: We combined natural language (NLP) processing techniques and statistical modeling to build a similarity method. The dependence and relatedness between phenotypes were taken into account through medical concept embedding, derived from a collection of 2.5 million French clinical narratives from the clinical data warehouse of Necker Hospital (Dr. Warehouse). We further assessed the adequacy of other existing embeddings, including cui2vec (Beam *et al.*, 2019) and HPO2Vec+ (Shen *et al.*, 2019). To aggregate concept similarity into patient similarity, we considered the adjusted average best-match method (Chen *et al.*, 2021) to take into account the relevance of each phenotype

to each patient. A ranking model based on the best-subtype-average similarity was proposed to address the phenotypic overlapping and heterogeneity of ciliopathies.

Results: Our results showed that using less than one-tenth of learning sources, our language and center specific embedding provided comparable or better performances than other existing medical concept embeddings. Combined with the best-subtype-average ranking model, our patient-patient similarity-based screening approach was demonstrated effective in two large scale unbalanced datasets containing approximately 10,000 and 60,000 controls with kidney manifestations in the clinical data warehouse (about 2% and 0.4% of prevalence respectively) (Chen *et al.*, 2022).

Conclusion: Our approach will offer the opportunity to identify candidate patients who could go through genetic testing for ciliopathy. Earlier diagnosis, before irreversible end-stage kidney disease, will enable these patients to benefit from appropriate follow-up and novel treatments that could alleviate kidney dysfunction.

Références

BEAM, A. L., KOMPA, B., SCHMALTZ, A., FRIED, I., WEBER, G., PALMER, N. P., … KOHANE, I. S., (2019), Clinical Concept Embeddings Learned from Massive Sources of Multimodal Medical Data, *arXiv:1804.01486 [cs, stat]*.

CHEN, X., FAVIEZ, C., VINCENT, M., BRISEÑO-ROA, L., FAOUR, H., ANNEREAU, J.-P., … BURGUN, A., (2022), Patient-Patient Similarity-Based Screening of a Clinical Data Warehouse to Support Ciliopathy Diagnosis, *Frontiers in Pharmacology*, vol. 13, p. 786710.

CHEN, X., FAVIEZ, C., VINCENT, M., GARCELON, N., SAUNIER, S. & BURGUN, A., (2021), Identification of Similar Patients Through Medical Concept Embedding from Electronic Health Records: A Feasibility Study for Rare Disease Diagnosis, *Studies in Health Technology and Informatics*, vol. 281, p. 600-604.

CHEN, X., GARCELON, N., NEURAZ, A., BILLOT, K., LELARGE, M., BONALD, T., … BURGUN, A., (2019), Phenotypic similarity for rare disease: Ciliopathy diagnoses and subtyping, *Journal of Biomedical Informatics*, vol. 100, p. 103308.

FAVIEZ, C., CHEN, X., GARCELON, N., NEURAZ, A., KNEBELMANN, B., SALOMON, R., … BURGUN, A., (2020), Diagnosis support systems for rare diseases: a scoping review, *Orphanet Journal of Rare Diseases*, vol. 15, n°1, p. 94.

GARCELON, N., BURGUN, A., SALOMON, R. & NEURAZ, A., (2020), Electronic health records for the diagnosis of rare diseases, *Kidney International*.

GARCELON, N., NEURAZ, A., BENOIT, V., SALOMON, R., KRACKER, S., SUAREZ, F., … BURGUN, A., (2017), Finding patients using similarity measures in a rare diseases-oriented clinical data warehouse: Dr. Warehouse and the needle in the needle stack, *Journal of Biomedical Informatics*, vol. 73, p. 51-61.

SHEN, F., PENG, S., FAN, Y., WEN, A., LIU, S., WANG, Y., … LIU, H., (2019), HPO2Vec+: Leveraging heterogeneous knowledge resources to enrich node embeddings for the Human Phenotype Ontology, *Journal of Biomedical Informatics*, vol. 96, p. 103246.

Adventures in using real-world evidence at the bed-side

Nigam H. Shah
Stanford University, USA

ABSTRACT

Using evidence derived from previously collected medical records to guide patient care has been a long-standing vision of clinicians and informaticians and one with the potential to transform medical practice. We will review research at Stanford University that developed an on-demand consultation service to derive evidence from patient data to answer clinicians' questions and support their bedside decision-making. We will describe the design and implementation of the service as well as a summary of their experience in responding to the first 100 requests. Consultation results informed individual patient care, resulted in changes to institutional practices, and motivated further clinical research.

BIOGRAPHY

Dr. Nigam Shah is Professor of Medicine at Stanford University, and Chief Data Scientist for Stanford Health Care. His research group analyzes multiple types of health data (EHR, Claims, Wearables, Weblogs, and Patient blogs), to answer clinical questions, generate insights, and build predictive models for the learning health system. At Stanford Healthcare, he leads artificial intelligence and data science efforts for advancing the scientific understanding of disease, improving the practice of clinical medicine and orchestrating the delivery of health care. Dr. Shah is an inventor on eight patents and patent applications, has authored over 200 scientific publications and has co-founded three companies. Dr. Shah was elected into the American College of Medical Informatics (ACMI) in 2015 and was inducted into the American Society for Clinical Investigation (ASCI) in 2016. He holds an MBBS from Baroda Medical College, India, a PhD from Penn State University and completed postdoctoral training at Stanford University.