Temporal phenotyping for characterisation of hospital care pathways of COVID19 patients*

Mathieu Chambard[1], Thomas Guyet[2]
Yên-Lan NGuyen[3], and Etienne Audureau[4]

[1] ENS Rennes/IRISA, Rennes, France
[2] Institut Agro/IRISA, Rennes, France
[3] AP-HP, Hôpital Cochin, Sorbonne Université, INSERM UMR S 1138, Pierre Louis Institute of Epidemiology and Public Health
[4] AP-HP, Henri Mondor Hospital, University Paris Est Créteil

Abstract. During the COVID19 crisis, Intensive Care Units admitted many patients with breathing disorders up to respiratory insufficiency. The care strategy of such patients was difficult to find and preventing patients to drift away toward a critical situation was one of the first challenge of physicians. In this study, we would like to characterize care pathways of patients that required a mechanical ventilation. The mechanical ventilation is an invasive treatment for the most critical respiratory insufficiencies. Through the analysis of the sequence of cares, we aim at supporting physicians to better understand patients evolution and let them propose new medical procedures to prevent some patients to be ventilated. This article proposes a method which combines a tensor factorization and sequence clustering. The tensor factorization enables to represent the care sequences as a sequence of daily phenotypes. Then, the sequences of phenotypes is clustered to extract typical care trajectories. This method is experimented on real data from Greater Paris university Hospital and is compared to a direct clustering of the sequences. The results show that the outputs are more easily interpretable with the proposed method.

Keywords: Tensor factorization · sequence clustering · phenotypes · care pathways.

1 Introduction

The advent of the COVID19 crisis show us the need to support physicians to identify as early as possible people who may have medical complications. This illustrates the need for predictive analytic tools that may support stakeholders

* This project is partly founded by Fondation APHP through the Chair AI-RACLES and received the agreement from the AP-HP CDW Scientific and Ethics Committee (CSE-20-11-COVIPREDS). Data used in preparation of this article were obtained from the AP-HP Covid CDW Initiative database. A complete listing of the members can be found at: (https://eds.aphp.fr/covid-19)

to better manage crises in the future: better individual patient management, better patient flows organization, etc.

In our case study, we would like to characterize the pathways of patients that required mechanical ventilation. Mechanical ventilation is an invasive treatment indicated when the patient's spontaneaous breathing is inadequate to maintain effective gaz exchange. It is a heavy treatment that physicians try to avoid for their patients. Their characterization may help physicians to improve their medical management procedures in these cases [5].

The characterization of a patient suffering from a disease is often called a phenotype. A phenotype may be a collection of conditions (smoker status, comorbidities, BMI, treatments), but the notion of phenotype may be extended to recent procedures and drugs that have been delivered to the patient. The information of such procedure becomes a proxy for the patient condition.

In this work, we use Electronic Health Record (EHR) data from AP-HP (Greater Paris university hospital) to build phenotypes of patients. The data collected by information systems provide access to rich information on hospital stays and for a very large population of hospitalized patients. Then, the care trajectory of a patient is described as a matrix $\boldsymbol{X}$ with features (procedures or drugs) in columns and days in rows. The value $X_{i,j}$ is 1 when patient p received the procedure/drug i the day j. AP-HP has identified more than $20,000$ patients hospitalized due to COVID19 from the beginning of the French crisis in March 2020 until March 2021.

There are potential flaws in the data but their volumetry and their sanitization make them valuable for extracting meaningful phenotypes. During the COVID19 crisis, physicians lack of time to code accurately the procedures or being exhaustive in their report. A sanitization of the database has been conducted all along the crisis to spur their use for research and operational purposes. These massive data should help to identify typical patient pathways, so called temporal phenotypes. A phenotype is a list of clinical features occurring in the same day for a subgroup of patients. For instance, the phenotype of patients suffering from a disease may be a combination of diagnosis codes, drugs or procedures he/she received, etc. A temporal phenotype describes a patient profile by the evolution of its "features" during his/her hospitalization. It groups together medications and procedures to best describe some visits.

This article proposes to use tensor factorization to identify automatically temporal phenotypes, so called typical care trajectories, from EHR data. More specifically, we investigate a simplified version of the CNTF model [14] which proposes to apply machine learning techniques in order to efficiently address tensor factorization (see Section 3). Our hypothesis is that depending on patients and procedures, their health status evolves in different ways. Discovering a temporal phenotype means to discover both what and when procedures occur during a patient stay and, if possible, to correlate the temporal phenotypes to patient outcomes such as mechanical ventilation. The dataset is presented in Section 4. Finally, Section 5 presents and analyses the first results of our approach on COVID19 care pathways and is compared to KMeans clustering.

2 Related works

Our goal is to discover phenotypes from an EHR database. Discovering phenotypes is an unsupervised task that aims at both describing phenotypes as typical sets of diseases and cares; and at identifying typical groups of patients having different types of phenotype.

There are several types of approach to address this problem. UPhenome [12] is a probabilistic approach based on Latent Dirichlet Allocation (LDA). It describes a patient by a set of cares without considering the temporal dimension. In our case study, we are interested in describing the longitudinal care trajectory of patients to characterize the dynamic of their disease. This dynamic of cares is characterized by careflow mining [3] using pattern mining techniques. In this approach, a careflow is a sequence of cares. But in case of sparse events, temporal patterns mining are more meaningful than sequential pattern mining. For instance, Dauxais et *al.* [4] proposed to discover patterns describing both the structural sequence of cares and the delay between. This problem has also been addressed in the statistical machine learning community. Many works have been proposed to discover structures in EHR data in supervised fashion. For instance, MedGraph [6] proposes a supervised EMR embedding method that captures the visit-code associations, and the temporal sequencing of visits through a point process.

In this article, we propose to explore an unsupervised statistical machine learning technique called non-negative tensor factorization (NNTF). NNTF has been studied extensively and many models have been proposed to tackle it [8]. The seminal methods are PARAFAC and Canonical Polyadic (CP) decompositions [7] which are the decomposition of a tensor in a finite collection of unidimensional vectors of rank R. The main limitation with this method is that it considers a tensor with fixed dimensions. In practice, it enforces all patients to have the same length of stay. Therefore, PARAFAC2 [9] extends the CP decomposition for a collection of matrices with different sizes (along one dimension). Both CP and PARAFAC2 are statistical approaches with good formal properties (*e.g.*, uniqueness of the CP decomposition). Nonetheless, these approaches are not computationally tractable on large datasets. Recently, SPARTan [10] proposed an algorithmic reformulation of PARAFAC2 to be faster and more memory-efficient. Another way to solve the tensor factorization consists in using machine learning techniques that provide efficient approximated solving processes. Since the last years, several machine learning solutions for tensor factorization have been proposed with additional features, for instance temporal regularization [14], handling missing values [13] or optimized for sparse data [1,2].

CNTF [14] (Collective Non-Negative Tensor Factorization) made two contributions: on the one hand, it is a flexible model which includes initial condition modeling, temporal regularization and classification regularization. Thus, CNTF is suitable for a wide range of care trajectories analysis. On the other hand, it proposes a description of a phenotype by a 2 dimensional matrices: one dimension for drugs and procedures and one dimension for lab tests. This matrix representation aims at capturing correlations between the two dimensions. Nonetheless,

CNTF only enables to extract daily phenotypes, but not groups of entire care trajectories.

3 Care pathway characterization through tensor factorization

In this section, we present a method for characterizing care pathways based on tensor factorization. The proposed method has two steps:

1. A tensor factorization identifies the daily phenotypes from patient care pathways,
2. The sequences of phenotypes are clustered to create groups of similar care pathways. The representative of each group is a *typical care trajectories*.

3.1 Tensor factorization

In this section, we propose a factorization model inspired by CNTF [14]. Our model borrowed from CNTF the principle of tensor factorization through function minimization and the temporal regularization. We simplified the model by discarding the other constraints (including correlation modeling between lab tests and cares).

Tensor factorization is a data analysis technique that consists in decomposing a multidimensional tensor $\mathcal{X}$ into a collection of lower dimensional tensors $\mathcal{Y}_1, \ldots, \mathcal{Y}_k$ such that $\mathcal{X} \approx \mathcal{Y}_1 \otimes \cdots \otimes \mathcal{Y}_k$ where $\otimes$ is a matrix product. A non-negative tensor factorization enforces $\mathcal{Y}_.$ matrices to contain only positive values.

In the context of EHR data analysis, $\mathcal{X}$ is seen as a three-dimensional tensor whose dimensions are the patient id (p patients), the time (d time units) and the medical events (N types of event). The length of stay of each patient visit are not all the same. Then, PARAFAC2 proposes a sparse representation of $\mathcal{X}$ as a collection of p two-dimensional matrices. I_k denotes the length of stay of the k-th patient such that its matrix $\boldsymbol{X}_k$ is of size $I_k \times N$.

Given $R \in \mathbb{N}$ the number of phenotypes, the matrix factorization problem consists in finding the matrices $\boldsymbol{U}$ of size $R \times N$ and the collections of p matrices $\boldsymbol{W}_k$ of size $I_k \times R$ such that for all $k \in \mathbb{N}_p$:

$$\boldsymbol{X}_k \approx \boldsymbol{W}_k \otimes \boldsymbol{U}$$

where $\boldsymbol{U}$ is the non-negative matrix describing the R phenotypes, and $\boldsymbol{W}_k$ is a non-negative matrix that describes the patient stay by the occurrence of the phenotypes each day. w_{krt} describes how likely the r-th phenotype exists at the particular time point t of patient k.

Inspired by CNTF, the problem is to minimize the following function:

$$f_{\boldsymbol{U},\boldsymbol{W}_{1..p}} = \sum_{k=1}^{p} \frac{1}{I_k} \sum_{i,j} \hat{x}_{kij} - x_{kij} \log(\hat{x}_{kij})$$

where $\hat{\boldsymbol{X}}_k = \boldsymbol{W}_k \otimes \boldsymbol{U}$ for all $k = 1..p$ is the tensor reconstruction from the phenotypes. In this problem formulation, the matrix reconstruction error is divided by the number of days. It aims at balancing contribution of patients who stayed for a long time or not.

At the moment, the temporal relationship is not taken into account in the model. However, for the course of a disease, we cannot look at the days independently of each other. The technique proposed in CNTF is to penalize a reconstruction model that does not allow to accurately predict the next sequences events or the stay outcomes. In both cases, we use a LTSM to model sequential dependencies between $w_{k \cdot t}$ vectors. The LTSM predicts the next state of the patient or the patient outcomes. In the first case, we want to minimize the mean square error between the real and predicted values, *i.e.*:

$$R(\boldsymbol{W}_k) = \frac{1}{I_k} \sum_{t=2}^{I_k} ||g_k(w_{k \cdot (t-1)}) - w_{k \cdot t}||^2$$

where g denotes the prediction function of the LSTM trained on the sequence.

In the second case, we want to minimize the prediction error. In our practical case, the outcome of the stay is whether a patient has been mechanically ventilated or not. In such case, the error may be evaluated through the cross-entropy between the predicted and real outcomes.

Finally, the tensor factorization from EHR data is formalized by the following optimization problem:

$$\begin{aligned} \underset{\boldsymbol{U}, \boldsymbol{W}_{1..p}}{\arg\min} \;\; & \mathcal{L} = \sum_{k=1}^{p} \frac{1}{I_k} \left(\sum_{i,j} \hat{x}_{kij} - x_{kij} \log(\hat{x}_{kij}) + \alpha \times \sum_{t=2}^{I_k} ||g_k(W_{t-1}) - W_t||^2 \right) \\ \text{subject to} \;\; & \hat{\boldsymbol{X}}_k = \boldsymbol{W}_k \otimes \boldsymbol{U} \\ & \boldsymbol{U} \geq 0 \\ & \boldsymbol{W}_k \geq 0, \; \forall k = 1..p \end{aligned}$$

where $\alpha \in \mathbb{R}^+$ is a parameter to balance the contribution of the two terms of the function.

To minimize $\mathcal{L}$, whatever optimization technique may be used. We use an alternating minimization strategy, illustrated in Algorithm 1. For each mini-batch B, the $\boldsymbol{U}$ is optimized given $\boldsymbol{W}_{1..p}$ values, then $\boldsymbol{W}_{1..p}$ is optimized using the $\boldsymbol{U}$ values. As the $\boldsymbol{U}$ is optimized several times per epoch while $\boldsymbol{W}_{1..p}$ is optimized only once (for each batch, only one part of the matrix actually changes), then we used different learning rates for each optimizer. In addition, the learning rate is decreased along the epochs to prevent from algorithm instability.

It is worth noting that we actually extract the phenotypes for the p patients. This means that the loss function $\mathcal{L}$ is evaluated on the p patients and splitting the datasets in train/test is not required.

Algorithm 1: Alternating minimization strategy (n epochs)

Data: $\boldsymbol{X}_{1..p}$: patient stays, R: the number of phenotypes
Result: $\boldsymbol{U}$: phenotypes, $\boldsymbol{W}_{1..p}$, phenotype occurrences in patient stays

$\boldsymbol{U} \leftarrow random$, $\boldsymbol{W}_{1..p} \leftarrow random$;
for $e = 1..n$ **do**
  for *Patient batch* B **do**
    $\boldsymbol{U}^* \leftarrow \boldsymbol{U} + \frac{\partial f_{\boldsymbol{U}, \boldsymbol{W}_{k\in B}}}{\partial \boldsymbol{U}}$;
    $\boldsymbol{W}^*_{k\in B} \leftarrow \boldsymbol{W}_{k\in B} + \frac{\partial f_{\boldsymbol{U}^*, \boldsymbol{W}_{k\in B}} + \sum_{k\in B} R(\boldsymbol{W}_k)}{\partial \boldsymbol{W}_{k\in B}}$;
  $\boldsymbol{U} \leftarrow \boldsymbol{U}^*$, $\boldsymbol{W}_{1..p} \leftarrow \boldsymbol{W}^*_{1..p}$;

3.2 Typical care trajectories

The tensor factorization enables us to change the representation of patient care pathways from sequences of cares $\boldsymbol{X}_{1..p}$ to sequences of phenotypes $\boldsymbol{W}_{1..p}$. In these two cases, the clustering of patients' matrices built typical care trajectories. It gathers similar pathways in clusters, and the representative of each cluster is a typical care trajectory.

In the general case, the patients' matrices do not have the same size due to different length of stay. Then, the classical clustering algorithms may not be applied. Our proposal is to use the Dynamic Barycentre Averaging (BDA) clustering approach [11]. DBA is a clustering algorithm for time series. It adapts the KMeans algorithm to the DTW distance. Thanks to the use of the DTW, it can cluster time series with different lengths. In our typical case, the sequence of phenotypes occurrences of a patient k, *i.e.* $\boldsymbol{W}_k$, is seen as a multidimensional time series of length I_k and R dimensions. The centroid of a cluster computed by DBA is then a typical care trajectory.

In our experiments, all patients' stays have the same length. In this case, a simple KMeans algorithm applies for clustering the $\boldsymbol{W}_{1..p}$ matrices.

4 Dataset of ventilated COVID19 patients

The objective of this study is to characterize the stays that have been admitted for COVID19. This disease affects the respiratory functions and may lead patients to critical respiratory insufficiency. In this case, patients have to be mechanically ventilated. This critical care procedure saves lives, but may lead to longer stays and to medical complications. For these reasons, physicians do their best to prevent patients from being mechanically ventilated.

In this section, we present the dataset that has been constructed to address the problem of the characterization of care pathways of patient who were ventilated.

Data were obtained from the AP-HP clinical data warehouse. It contains information for $27,370$ ICU admissions with at least one positive PCR[5] test

[5] PCR (Polymerase chain reaction) denotes here a test for COVID19 infection.

in one of the hospitals in the Greater Paris region between March 2020 and March 2021. It represents 17,400 unique patients. The database includes dates of admission to the intensive care unit, gender, age of each patient and possibly date of death.

For this study, patients were selected from people in the AP-HP database over 18 years old at ICU admission with a positive PCR test. We discarded patients having short visits (less than a day). In the original database there are 3.5 times more visits (27,370 visits) without ventilation procedure than visits leading to at least a mechanical ventilation procedure (6,066 visits). In order to balance the dataset, we subsample the patients without ventilation procedures. Indeed, the goal is to compare ventilated and non-ventilated patients. So the cohort must have roughly the same number of people and a similar age distribution. We adopted a stratified subsampling of the ventilated patients to have similar populations in age. More precisely, patients were drawn randomly to have for each age group (18-20, 20-40, 40-60, 60-80, 80-100, 100-120) as many ventilated as non-ventilated people. Figure 1, on the left, displays the age distributions of ventilated and non-ventilated patients. This figure also details the distributions of lengths of stay and of ages of death (for COVID19 or another reason). Note that in this study, we are not interested in the patient death but only on whether their stay leads to a mechanical ventilation or not.

The database contains medical and administrative information about each visit: clinical observations, lab test results, care performed or also textual medical reports. We decided to focus on medical procedures and prescription drugs, and to discard lab tests and medical reports. This information is collected with a suitable quality due to their administrative purpose (patient reimbursement). On the contrary, laboratory tests are too sparsely available and it is difficult to extract reliable information from medical reports.

All drugs and procedures delivered are timestamped and coded using standard taxonomies. Drugs are coded with ATC[6] codes and procedures are coded with CCAM[7] codes. CCAM is a French codification for medical procedures. Each code is a type of medical event in the $\mathcal{X}$ tensor. Drugs and procedure deliveries are timestamped with dates and times. We keep only the dates. For some procedures performed along several days (*e.g.*, mechanical ventilation), the procedure is accurately recorded daily. Contrary to procedures, drugs are tagged with start and end dates, but the ends of drug exposures is not reliable. This is currently a potential weakness in our data.

The next step was to select a subset of potentially meaningful drugs/procedure among all possible codes. Indeed, the temporal and spatial complexity is exponential with the number of features. Considering the limited computational resources available on hospital servers, a selection of features was required. In addition, less medical features eases the interpretation of the results. The out-

[6] ATC: Anatomical, Therapeutical and Chemical

[7] CCAM: Classification commune des actes médicaux/Common classification of medical procedures.

Table 1. Statistical characteristics of the cohorts/datasets. Raw database denotes the database of 21,901 patients with positive PCR tests, and final database denotes the stratified patients, with medical feature selection.

	Raw database	Final database
Number of patients	21,901	7,358
Number of visits	37,312	8,937
Average age	69 years	64 years
Gender distribution	M:56%, F:44%	M:62%, F:38%
Average length of stay	10 days	10 days
Number of different drugs	1,120	166
Number of different procedures	2,635	44
Death rate	23%	28 %

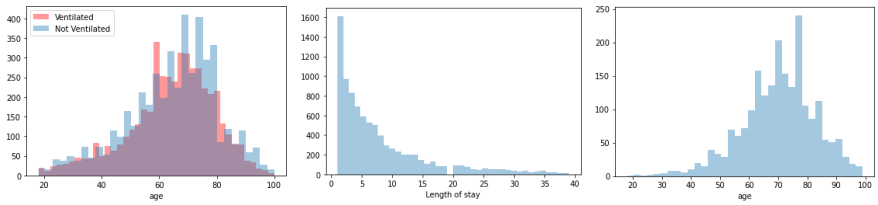

Fig. 1. Population characteristics. From left to right: age distribution, length of stay, age distribution of deceased people

putted phenotypes are more concise and there are less potential correlations to analyze for physicians.

In the case of the COVID19 study, the patients are very heterogeneous and have very different pathologies. The total number of medical events is very large, 1,120 different drugs and 2,635 different procedures. The selection of the medical features have been done in two steps. Firstly, the 500 most frequent drugs and 200 most frequent procedures were selected. Secondly, physicians selected 166 types of drugs and 44 types of procedure from the frequency-based selection. They selected the potentially most interesting medical features in the context of COVID19.

Table 1 sums up some characteristics of the cohort. Figure 1 on the right shows the distribution of the age of death. The distribution matches the known indicators: the people most affected are people over 60 years old.

Finally, for each visit, we select the events that occurs d days after the entry in an ICU. In case the patient visit started in another service, it is not taken into account. In this study, the pathway starts the first day in an ICU service. The entry date is used as an index date that is valid for patients who were ventilated or not. In addition, in the perspective of having a decision support tool, it is interesting to observe the care trajectory of a patient since its entry to decide as soon as possible the action to take to prevent a ventilation.

5 Experiments and results on COVID19 care pathways

In this section, our method is applied to the database presented in the previous section. We remind that our objective is to investigate the care pathways of patients who have been mechanically ventilated or not. We set $d = 6$ meaning that 6 days were kept per patient from their arrival in ICU.

The tensor factorization model has been adapted from the CNTF implementation. It is implemented within the PyTorch framework. An initial study of the algorithm convergence shown that the algorithm does not significantly improve the results after 100 epochs. Then, we set the number of epochs to 100 and batch size of 100 patients. The running time on the dataset detailed in the next section is from 5 to 15 minutes on a server dedicated to AP-HP data analysis. This reasonable time makes the approach practical on real data. For the clustering algorithm, we use a K-means algorithm that suits our particular dataset which contains sequences of the same length. We used the K-Means *sklearn* library with a smart initialization of the centers.

In the remaining of this section, we start by studying the daily phenotypes extracted from care pathways of the whole dataset (ventilated and non-ventilated). Then, we investigate the results of the clustering phase of our method (typical care trajectories). Finally, we propose to compare the obtained results with the direct clustering of care trajectories.

5.1 Phenotypes of COVID19 patients

The main parameters of our method are R, the number of phenotypes, and ρ, initial random state. Due to the stochastic nature of the optimization process, the results also depends on the initial random state (ρ). The method was tested with different $R \in [6, 12]$ and ρ in order to find which value to give to R and to ρ to have insightful and robust results.

In the following, we illustrate two cases: $R = 8$ and $R = 10$. The outputted phenotypes are illustrated in Figure 2.

The detailed phenotypes are presented in Tables 2 and 3. After a physician's expertise, several pieces of information emerged from these phenotypes. First of all, we recognize phenotypes that characterize the pathway of patients in a intensive care unit. These are phenotypes with a prescription of thromboprophylaxis like *Enoxaparin* and also those who received antibiotics (*cefotaxime*, *amoxicillin* and inhibitor). This corresponds to a large part of the results: phenotypes 1.0, 1.2, 1.6, 1.7 and phenotypes 2.1, 2.2, 2.7, 2.8 and 2.9.

We also find deliveries of analgesics such as morphine, tramadol, nefopam or paracetamol in phenotypes 1.0, 1.2, 1.5, 1.7 and also in phenotype 2.7 and 2.8. These drugs treat muscle pain or fever caused by COVID19.

After some deaths from pulmonary embolism, a link has been discovered between a severe form of COVID19 and a risk of venous thrombosis. Patients gradually benefited from a preventive treatment for thrombosis such as *enoxaparin*. It appears in phenotypes 1.0, 1.2, 1.6, 2.1, 2.7 and 2.8.

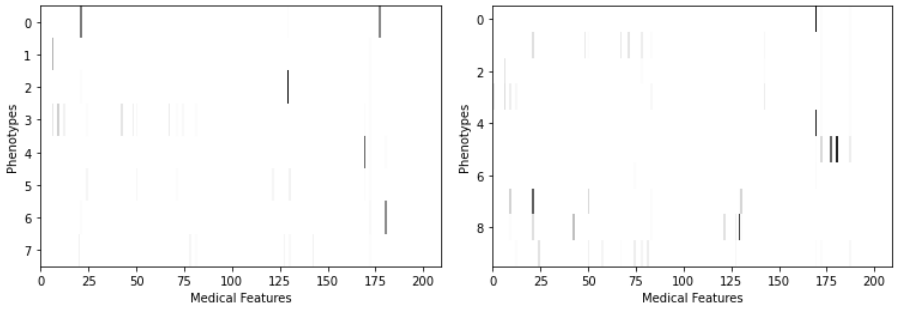

Fig. 2. Drugs phenotype result for $R = 8$ (on the left) and $R = 10$ (on the right). Each row corresponds to a phenotype, the columns correspond to drugs identifiers. A dark square means that the drug in column is part of the phenotype in row. The darker the square, the more likely the drug in this phenotype. (see Tables 2 and 3 for detailed values).

The cohort has a high average age. This explains the appearance of *furosemide* in phenotype 1.3 and phenotype 2.8. This drug is an anti-hypertensive agent prescribed for elderly.

In addition, we observe common diseases in patients suffering from COVID19. First, some patients suffer from diabetes. Some have been intubated (phenotype 2.2), others have hypertension (phenotype 2.7: *amlodipine*). Second, some patients have cholesterol and cardiovascular problems. They are found in phenotypes 1.3 and 2.1. Finally, phenotypes 1.5 and 2.9 correspond to patients suffering from hypertension (*amlodipine*, *ramipril*) with also cardiovascular problems (*acetylsalicylic*).

Interestingly, the procedures are gathered in two or three phenotypes (1.4, 1.6, 2.0 and 2.4). Such phenotypes describes the standard monitoring procedures in a ICU service (*e.g.* electrocardiogram, intra arterial pressure). Thus, the stay of a patient being monitored in a ICU service is described with a combination of one of such phenotypes and phenotypes for drugs deliveries. It also highlight intubation procedures and the injection of *dobutamine / dopamine* present in phenotype 2.5.

Finally, it is worth noting that ρ parameter has a low impact on the results of the system. By repeating the experiment several times with different values, we observe similarities between the results of phenotypes. This robustness makes us confident in the significance of the results. However, these are not exactly the same phenotypes. Sometimes a phenotype of an experiment is the mixture of two phenotypes of an experiment with a different value of ρ. This may be disturbing for physicians.

Table 2. Phenotypes with R = 8. Numbers indicate the likelihood of the occurrence of a drug for a phenotype. Drugs names correspond to the French official denomination.

$Ph_{1.0}$	Enoxaparine: 1.2, Injection dobutamine/dopamine: 1.11, Paracetamol: 0.03, Dexamethasone: 0.0, Amlodipine: 0.0
$Ph_{1.1}$	Insuline aspart: 0.73, Monitoring of intra-arterial pressure: 0.03, Continuous monitoring of electrocardiogram: 0.0
$Ph_{1.2}$	Paracetamol: 2.33, Monitoring of intra-arterial pressure: 0.03, Enoxaparine: 0.03, Tramadol: 0.0, Acetylsalicylique acide: 0.0
$Ph_{1.3}$	Insuline glargine: 0.44, Insuline aspart: 0.24, Furosemide: 0.24, Atorvastatine: 0.18, Bisoprolol: 0.15
$Ph_{1.4}$	Continuous monitoring of electrocardiogram: 1.44, Central intra-arterial or intravenous pressure monitoring : 0.03, Monitoring of intra-arterial pressure: 0.03
$Ph_{1.5}$	Nefopam: 0.1, Acetylsalicylique acide: 0.09, Morphine: 0.09, Amlodipine: 0.08, Monitoring of intra-arterial pressure: 0.03
$Ph_{1.6}$	Central intra-arterial or intravenous pressure monitoring : 1.05, Monitoring of intra-arterial pressure: 0.06, Enoxaparine: 0.03, Acetylsalicylique acide: 0.0, Amlodipine: 0.0
$Ph_{1.7}$	Heparine: 0.12, Zopiclone: 0.12, Amoxicilline et inhibiteur d'enzyme: 0.09, Tramadol: 0.09, Nefopam: 0.06

5.2 Care trajectories

In this section, we describe the different pathways that lead to use mechanical ventilation or not. Then, we investigate the typical patient trajectories.

In the previous section, we analyzed the phenotypes, $\boldsymbol{U}$. This section analyzes the information contained in $\boldsymbol{W}_{1..p}$ matrices. These matrices represent the sequence of cares during the first 6 days of the ICU stay.

A cluster is a group of patients having the same kind of sequences during the first days of its stay. In our particular case, the clustering can be done with the DBA algorithm (see section 3.2) or with a regular KMeans using the Froebenius distance between matrices having the same dimensions. For computational reasons, we applied this second alternative and set up the algorithm with $k = 6$.

Figure 3 illustrates the six cluster centers. For a better clarity, values lower than the half of the maximum of a matrix have been set to 0. A dark cell means that the phenotype is significantly present in average at a given day before starting ventilation for the group of patients.

The clusters could be split into three types of clusters. The clusters CT_0, CT_1 and CT_2 are mostly present in unventilated people. They are 2 to 3 times more present in non-ventilated patients than in ventilated patients. Then the clusters CT_4 and CT_5 are especially present in ventilated people. Finally, cluster CT_3 lies in both visits from ventilated and non-ventilated people.

We remind that the hospital stay of patients is aligned with the first days of hospitalization. Therefore, we can have a shift in phenotypes between patients depending on their health status at arrival. This shift is observed with cluster

Table 3. Phenotypes with R = 10. (see legend of Table 2)

$Ph_{2.0}$	Continuous monitoring of electrocardiogram: 2.17, Monitoring of intra-arterial pressure: 0.03, Intubation trachéale: 0.03
$Ph_{2.1}$	Enoxaparine: 0.27, Dexamethasone: 0.22, Atorvastatine: 0.21, Bisoprolol: 0.18, Amoxicilline et inhibiteur d'enzyme: 0.15
$Ph_{2.2}$	Insuline aspart: 0.24, Zopiclone: 0.03, Monitoring of intra-arterial pressure: 0.03, Intubation trachéale: 0.03, Amoxicilline et inhibiteur d'enzyme: 0.03
$Ph_{2.3}$	Insuline aspart: 0.3, Phloroglucinol: 0.18, Insuline glargine: 0.15, Zopiclone: 0.15, Metformine: 0.06
$Ph_{2.4}$	Continuous monitoring of electrocardiogram: 1.96, Monitoring of intra-arterial pressure: 0.03, Intubation trachéale: 0.03
$Ph_{2.5}$	Central intra-arterial or intravenous pressure monitoring : 1.87, Injection dobutamine/dopamine: 1.29, Monitoring of intra-arterial pressure: 0.33, Intubation trachéale: 0.15, Continuous monitoring of electrocardiogram: 0.03
$Ph_{2.6}$	Continuous monitoring of electrocardiogram: 0.04, Prednisone: 0.03
$Ph_{2.7}$	Enoxaparine: 1.36, Insuline glargine: 0.38, Nefopam: 0.35, Amlodipine: 0.34, Ceftriaxone: 0.03
$Ph_{2.8}$	Paracetamol: 1.6, Furosemide: 0.55, Enoxaparine: 0.27, Morphine: 0.26, Tramadol: 0.12
$Ph_{2.9}$	Acetylsalicylique acide: 0.24, Cefotaxime: 0.15, Prednisone: 0.15, Amlodipine: 0.12, Ramipril: 0.09

Table 4. Repartitions of ventilated/unventilated patients.

Care trajectories	Patients	Unventilated	Ventilated
CT_0	362	257	105
CT_1	3093	2110	983
CT_2	2889	1376	1513
CT_3	883	640	243
CT_4	767	30	737
CT_5	884	1	883

CT_1 and CT_3. These two clusters have almost the same phenotypes: 2.1, 2.2, 2.3, 2.7, 2.8, 2.9. One is filled on the first day of hospitalization, the other on the second day. In addition, the clusters have the same proportion of ventilated and unventilated, which supports the fact that these clusters represent the same kinds of patients.

The phenotypes 2.0, 2.4, 2.5 and 2.6 are mostly present in ventilated patients. They correspond to cluster CT_4 and CT_5 for which 98% of the patients have been ventilated. Indeed, the phenotypes contained in the clusters are phenotypes linked to classical resuscitation procedures and are very similar to the ones in the previous section.

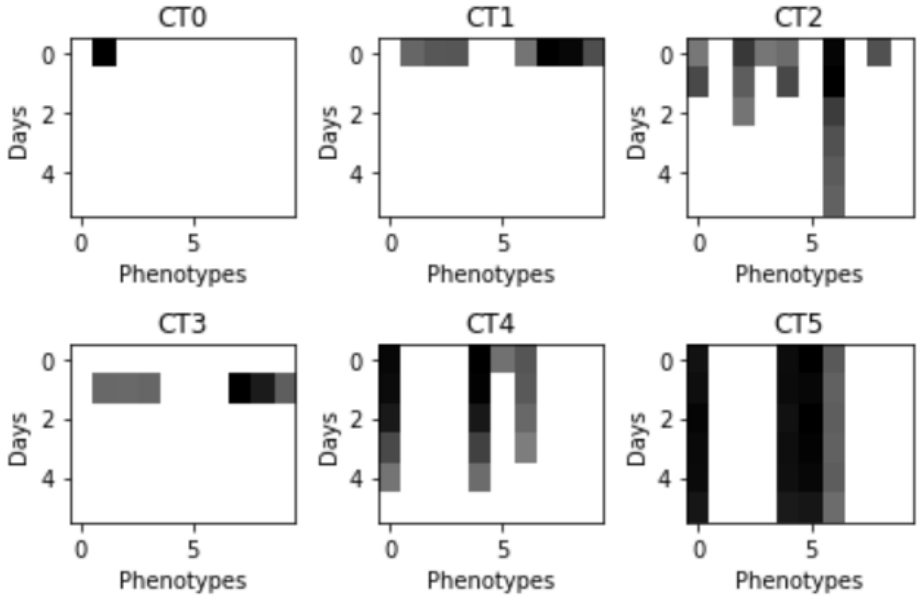

Fig. 3. Typical care trajectories (CT_i for $i = 1..6$) of patient during the first 6 days of hospitalization. A typical care trajectory gives how likely a phenotype appears at a given day of the stay. This figure uses the phenotypes extracted with $R = 10$ (see Figure 2, on the right).

Finally, cluster CT_2 represents as many ventilated and non-ventilated patients. This cluster appears in almost a third of patients. It is made up of phenotypes 2.0, 2.4 and 2.6 which are mainly made of ICU procedures. The other phenotypes present are phenotypes 2.2 and 2.6 which are mainly prescriptions.

5.3 Comparison with direct clustering

In this section, the goal is to compare the trajectories extracted with our method and the ones extracted with a direct clustering (K-Means). Figure 5 shows the KMeans cluster centers. It illustrates the medical events' occurrences wrt days. The clusters of this figure are compared to the results of our method presented in Figure 4. The matrices of this later figure are computed by multiplying the clusters matrices with the phenotype matrix $\boldsymbol{U}$ (see Figure 2).

Table 5 provides the number of ventilated and non-ventilated patients in each cluster.

We can observe common clusters between the two method results. For instance, there is a strong similarity between CT_1 and KM_3. Additionally, CT_4 and CT_5 clusters look like KM_0, KM_4 and KM_5 clusters. Then, we can conclude that the approaches extract almost the same care trajectories.

Table 5. Repartitions of ventilated/unventilated patients.

Care trajectories	Patients	Unventilated	Ventilated
KM_0	541	1	540
KM_1	2422	1778	644
KM_2	3839	1963	1876
KM_3	1005	644	361
KM_4	450	0	450
KM_5	621	28	593

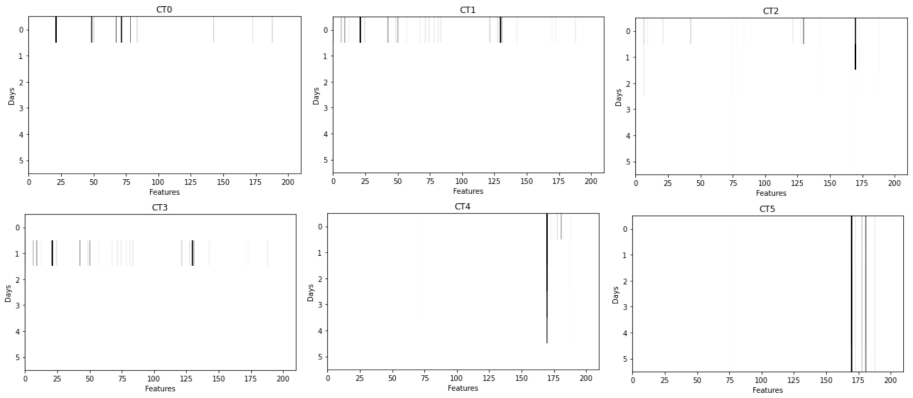

Fig. 4. Typical care trajectories: medical events along the first 6 days of hospitalization (alternative view of the result presented in Figure 3)

Nonetheless, KM_3 and KM_4 are quite similar while there is more diversity in the phenotypes extracted by our methods. A possible explanation is that clustering the sequence of few phenotypes is easier than clustering the sequence of all the medical events.

The second advantage of our method is in the ease to interpret the results and get insight from them. We have seen that the daily phenotypes can be interpreted by physicians. This intermediary interpretation enables physicians also to get insights from the typical care trajectories of Figure 2. We believe a direct clustering providing the care trajectories without intermediary phenotype is harder to interpret.

6 Conclusion

We presented a method to extract typical care trajectories from EHR care pathways. Our method combines a tensor factorization to extract daily phenotypes and a clustering of phenotype sequences. This method has been applied to the analysis of COVID19 patients admitted in ICU to investigate the use of mechanical ventilation.

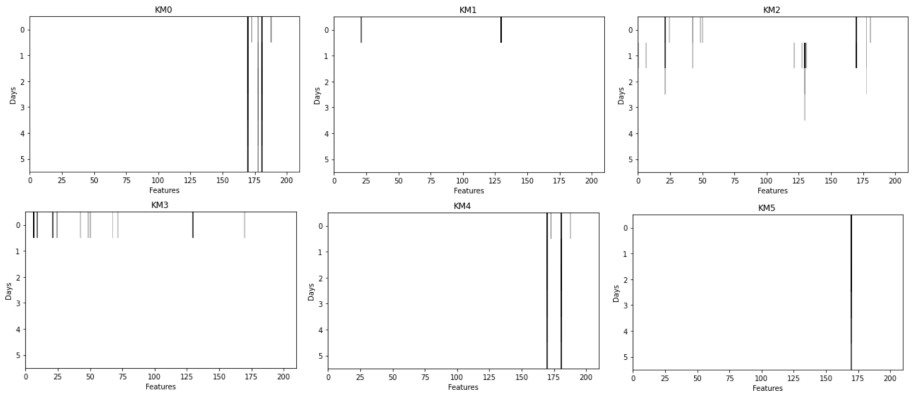

Fig. 5. Typical care trajectories (K-Means clustering): medical events along the first 6 days of hospitalization.

The first results with this method are interesting. First, the use of an approximate tensor factorization inspired by CNTF enables to process a large number of patient sequences. Phenotypes have been easily interpreted by physicians as their evolution over days. Compare to the direct clustering of the sequences, we argue that the use of phenotype is more insightful and easier to interpret. Finally, these results are promising. It is important to continue to look at the evolution of phenotypes in patients to compare the course of the disease in different subgroups of the population. For the future, the goal will also be to compare the evolution in ventilated and non-ventilated people using supervised tensor factorization techniques.

References

1. Afshar, A., Perros, I., Papalexakis, E.E., Searles, E., Ho, J., Sun, J.: COPA: Constrained PARAFAC2 for sparse & large datasets. p. 793–802 (2018)
2. Afshar, A., Yin, K., Yan, S., Qian, C., Ho, J.C., Park, H., Sun, J.: SWIFT: Scalable wasserstein factorization for sparse nonnegative tensors. In: Proceedings of the AAAI conference (2021)
3. Dagliati, A., Sacchi, L., Zambelli, A., Tibollo, V., Pavesi, L., Holmes, J., Bellazzi, R.: Temporal electronic phenotyping by mining careflows of breast cancer patients. Journal of Biomedical Informatics **66**, 136–147 (2017)
4. Dauxais, Y., Guyet, T.: Generalized chronicles for temporal sequence classification. In: Workshop on Advanced Analytics and Learning on Temporal Data (AALTD). pp. 30–45 (2020)
5. Ferté, T., Cossin, S., Schaeverbeke, T., Barnetche, T., Jouhet, V., Hejblum, B.P.: Automatic phenotyping of electronical health record: Phevis algorithm. Journal of Biomedical Informatics **117**, 103746 (2021)
6. Hettige, B., Wang, W., Li, Y., Le, S., Buntine, W.L.: Medgraph: Structural and temporal representation learning of electronic medical records. In: Proceedings of the European Conference on Artificial Intelligence (ECAI). vol. 325, pp. 1810–1817 (2020)

7. Hitchcock, F.L.: The expression of a tensor or a polyadic as a sum of products. Journal of Mathematics and Physics **6**(1-4), 164–189 (1927)
8. Hong, D., Kolda, T.G., Duersch, J.A.: Generalized canonical polyadic tensor decomposition. SIAM Review **62**(1), 133–163 (2020)
9. Kiers, H.A., Ten Berge, J.M., Bro, R.: PARAFAC2–part i. A direct fitting algorithm for the PARAFAC2 model. Journal of Chemometrics: A Journal of the Chemometrics Society **13**(3-4), 275–294 (1999)
10. Perros, I., Papalexakis, E.E., Wang, F., Vuduc, R., Searles, E., Thompson, M., Sun, J.: Spartan: Scalable parafac2 for large & sparse data. In: Proceedings of the International Conference on Knowledge Discovery and Data Mining -ACM SIGKDD). pp. 375–384 (2017)
11. Petitjean, F., Ketterlin, A., Gançarski, P.: A global averaging method for dynamic time warping, with applications to clustering. Pattern recognition **44**(3), 678–693 (2011)
12. Pivovarov, R., Perotte, A.J., Grave, E., Angiolillo, J., Wiggins, C.H., Elhadad, N.: Learning probabilistic phenotypes from heterogeneous ehr data. Journal of Biomedical Informatics **58**, 156–165 (2015)
13. Yin, K., Afshar, A., Ho, J.C., Cheung, W.K., Zhang, C., Sun, J.: LogPar: Logistic PARAFAC2 factorization for temporal binary data with missing values. In: Proceedings of the International Conference on Knowledge Discovery & Data Mining (ACM SIGKDD). pp. 1625–1635 (2020)
14. Yin, K., Qian, D., Cheung, W.K., Fung, B.C.M., Poon, J.: Learning phenotypes and dynamic patient representations via rnn regularized collective non-negative tensor factorization. Proceedings of the AAAI Conference on Artificial Intelligence **33**(01), 1246–1253 (2019)