International Journal of Healthcare Information Systems and Informatics
Volume 16 • Issue 4

A Data Representation Model for Personalized Medicine

Hafid Kadi, Department of Computer Science, University of Mustapha Stambouli, Mascara, Algeria & Normandie University, UNICAEN, ENSICAEN, CNRS, GREYC, Caen, France*

Mohammed Rebbah, Department of Computer Science, University of Mustapha Stambouli, Mascara, Algeria

Boudjelal Meftah, Department of Computer Science, University of Mustapha Stambouli, Mascara, Algeria

Olivier Lézoray, Normandie University, UNICAEN, ENSICAEN, CNRS, GREYC, Caen, France

ABSTRACT

Personalized medicine exploits the patient data, for example, genetic compositions and key biomarkers. During the data mining process, the key challenges are the information loss, the data types heterogeneity, and the time series representation. In this paper, a novel data representation model for personalized medicine is proposed in light of these challenges. The proposed model will account for the structured, temporal, and non-temporal data and their types, namely numeric, nominal, date, and Boolean. After the "date and Boolean" data transformation, the nominal data are treated by dispersion while several clustering techniques are deployed to control the numeric data distribution. Ultimately, the transformation process results in three homogeneous representations with these representations having only two dimensions to ease the exploration of the represented dataset. Compared to the symbolic aggregate approximation technique, the proposed model preserves the time-series information, conserves as much data as possible, and offers multiple simple representations to be explored.

KEYWORDS

Clustering, Data Representation, Electronic Health Records (EHRs), Medical Event, Personalized Medicine (PM), Time-series Data

1. INTRODUCTION

Personalized medicine (PM) refers to the individualization of medical treatments based on the unique dataset of each patient. It generates and exploits stored patient data, which are often captured digitally in an "Electronic Health Record (EHR)" comprising profiles of many different patients. Essentially, an EHR refers to a longstanding, comprehensive health database resource that stores and manages all patient data files digitally under the custody of a licensed health entity. More specifically, it provides a digitalized view of the patient's demographics, data associated with the patient's clinical and medication history, diagnostic trajectory, social and economic environmental conditioning, geographical relocation, if any, as well as the patient's genetic data, if these exist (Jensen et al., 2012).

Together, this massive data resource available via the EHR often includes not only homogeneous, heterogeneous, structured, unstructured and/or semi-structured data, but also the temporal and non-

DOI: 10.4018/IJHISI.295822 *Corresponding Author

temporal data. Mixed in this huge bag of patient data are many captured medical events of different individual patients such as their body temperature measurements, blood pressure recordings and other time-series information, with different sorts and forms of data. As Ghazi (2015) noted, we consider time-series data to include all the observational sequences of a patient being captured vis-à-vis a medical event. Moreover, the EHR data resource contains a lot of hidden information and knowledge waiting to be mined and/or discovered. The process of reporting, evaluation, and medical decision-making based on the EHR data involves the extraction of relevant information and knowledge via specialized methods known as data mining techniques. The quality of information processing and knowledge discovery are thus directly linked to the availability, accessibility, type and form of the data to be extracted and aggregated for analysis. The objective of our work is to produce a high fidelity model for the representation of PM structured data. This is a challenging problem and our proposed model addresses several important scientific gaps: data heterogeneity, loss of data during data transformation, and interpretability of the representation over the course of a data mining process. To accomplish this non-trivial task, we represent the data by two parts. The first is dedicated to the representation of numeric data with clustering techniques, whereas the second part considers the representation of nominal data with respect to its dispersion. These two bodies of information are then joined into a single global representation table. Thanks to the simplicity of the obtained representation, healthcare specialists will be able to identify in the dataset both the key patient events, as well as the variations in the information conveyed by the data series. However, it is intended for the obtained representation to be used within automated medical decision-making processes such as disease prevention and/or adverse drug events prediction. Importantly, this paper emphasizes the need to explore the EHR data mining process that informs and challenges PM, which will ultimately enhance the ability of physicians and other care professionals to personalize high quality care to the inflicted individuals.

The rest of the paper is organized as follows. Section 2 explains the time-series representation process limitations. A novel data representation model proposed for PM is then detailed in Section 3 with Section 4 continuing on the discussion about the experimentation and the evaluation of the proposed model and the results analysis. The final section, Section 5, will provide concluding remarks and offer insights into practical implications and potential future works.

2. EHR INFORMATION & DATA LOSS

Most EHR data exploration models require a lot of the stored data, including temporal data, to be represented and transformed into an appropriate, meaningful and interoperable form. Research along these lines constitutes a common point among several past efforts. Bagattini et al. (2019) proposed an approach that belongs to one of the medical branches advocating PM: the prediction of adverse drug events. The authors have used three phases: symbolic data representation, subsequences generation and classification. Their approach exploits sparse time series features, that is, it processes only the numeric data and ignores the other types of data that may appear in the EHR data. In the representation phase, they applied SAX technique in order to produce symbolic representation sequences for all the time-series. Regardless of the limits of the SAX technique and of the processed data qualities (different types), this approach provides a timely example on the use of PM data and puts forward the need for new techniques for EHR data representation. Among the emerging data mining methodologies, for example, the Deep Integrated Prediction (DIP) project (Milad et al., 2018) uses deep learning to represent numeric type attributes and the GloVe algorithm for discriminating nominal attributes in EHRs of patient records to showcase the cardiac disease. Graph-based Attention Model (GRAM) is an approach based on marking each medical visit events on a directed acyclic graph (Edward et al., 2017). The work of Mallick et al. (2018) presents an approach to study the genes interdependence in cancer cases. These researchers apply fuzzy logic to compute the information gain to represent the genes data vis-à-vis a graph. In Anima et al. (2015), windows fixed with marking according to the

time factor during diagnosis ('time-Windows') have been proposed. The time period was initially fixed to six months, thereby causing possible information loss concerning changes in the event's behavior. Years later, Jing et al. (2017) argue for a different approach to capture data temporality treatment as applied to the clinical events and representation extraction. For the transformation process, they used the Symbolic Aggregate Approximation (SAX) method on all observations to generate a representation as character chains, a process that may also cause the loss of information.

Treatment unification on the numerical data of patient records sometimes implies the need for data normalization. From a data transformation perspective, normalization assists in unifying the treatment scales so as to make the results comparable. Yet, this process can also generate information loss as encapsulated in the data series. Representation techniques applied on the results of these transformations can therefore lead to loss in the series as constraints such as the minimum length of the series to be processed may not be accommodated. In best practices, especially with time-series data, only one or two data types may generally be treated at the same time with the information and data loss problem omitted. To overcome challenges faced with the representation problems of time-series, data types homogeneity, and the information loss, an alternative data representation model is proposed. The proposed data model can represent the structured, temporal and/or non-temporal data and their differing types, including numeric, nominal, date, and Boolean. The temporal observations exist in a three-dimensional form (3-D: Patients, Events, Time). As an endeavor to unify the treatment process for the non-temporal data that are actually found in two-dimensions (2D: Patient, Data), we consider them in a three-axis form (Patient, Data, Time-1) such that "Time-1" takes only one value, i.e., "1". Such a model conserves as much data as is possible even for short time-series. Applying several techniques with a different number of clusters reallocate the data distribution. Accordingly, this work will result in a simplified representation with only two dimensions, making it easy to explore. With such results, it is clear that attention has now been given to the annotated challenges, including data type heterogeneity, the maximum coverage of data resource during treatment, the minimization of loss information and data during the transformation process.

For large-scale data transformation, imagine having a growing number of multiple data types to be extracted simultaneously. This would surely weigh down on the data exploration process, which is needed for the investigator(s) to obtain useful insights from the entire data set; more precisely, with satisfying the objective of having the emergence of irrelevant data to be minimized. Sometimes, novel data exploration approaches are used to represent data and extract dataset features, and may be enhanced with a combination of classification techniques and optimization tools such as the artificial bee colony (ABC) algorithm (Razmjooy & Khalilpour, 2015; Khalilpour et al., 2013), an optimization technique based on variance reduction of the Gaussian distribution (Namadchian et al., 2016), an invasive weed optimization approach (Razmjooy & Ramezani, 2014), an imperialist competitive algorithm (Razmjooy et al., 2013), a particle swarm optimization (PSO) methodology (Kolekar & Pawar, 2014) and/or a genetic algorithm (GA) (Nguyen et al., 2014) so as to train a classification and decision-making model.

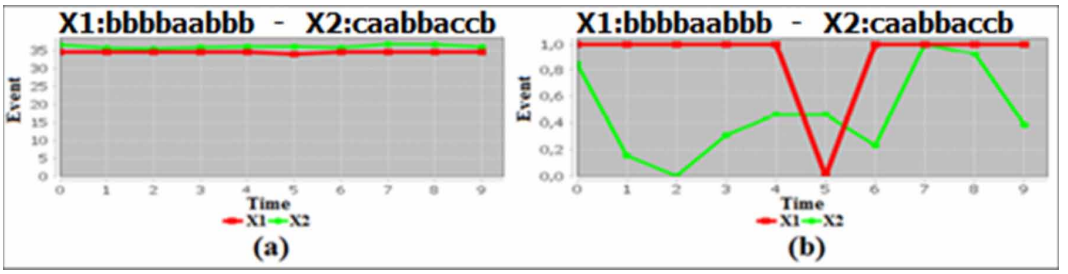
Yeh et al. (2009) and Nguyen et al. (2014), for example, are two medical data classification methodologies useful for EHR data exploration. The first separates patients with data foreshowing a high risk of breast cancer(Yeh et al., 2009). These researchers apply the PSO technique on features selection via statistical methods. The second uses the breast cancer and heart disease datasets for its exploration(Nguyen et al., 2014). The proposed model incorporates a feature extraction and transformation phase via the "Wavelet transformation" technique, followed by a training phase via the fuzzy standard additive model and the GA technique. The EHR data exploration can also involve the image-processing task stored in the data source. Kavya et al. (2020) and Guo & Razmjooy (2019) are examples of medical image processing and breast cancer detection approaches.

Symbolic Aggregate Approximation (SAX) is a representation and dimension reduction technique for time-series(e.g., Lin et al., 2007; Park & Jung, 2020). It has the capacity to transform the numeric data series into sequences of consecutive symbols. Its first phase divides the time-series of length N

into W equal segments where (W < N), and for each segment, the average of its values is computed. The second phase defines a corresponding strategy between a proposed alphabet with Z symbols and the new computed averages series.

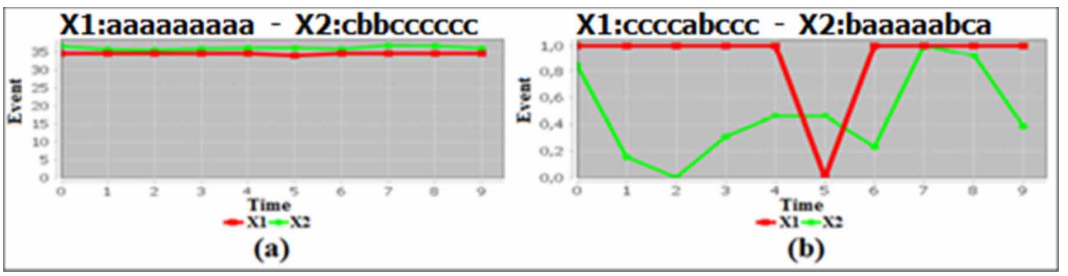
Jing et al. (2017) use the SAX technique to normalize time-series values. The normalization step generates two problems. To better understand these problems, we use two (2) time-series that represent the real temperature measures of two patients: X1={34.6, 34.6, 34.6, 34.6, 34.6, 34.0, 34.6, 34.6, 34.6, 34.6} and X2={36.6, 35.7, 35.5, 35.9, 36.1, 36.1, 35, 8, 36.8, 36.7, 36.0}. They will be considered to generate a string of nine characters at the base of three representation symbols. The *first* problem appears when applying the SAX method on X1 and X2 separately. Indeed, the generated representation is the same for both normalized v. non-normalized data. However the same symbol is used in the two representations that may correspond to two different intervals, for example, without normalization the technique SAX generates a symbol 'a'∈]-∞;34,462] for X1 and 'a'∈]-∞;35,939] for X2as shown in Figure 1.a. With normalization, it generates 'a'∈]-∞;0,771] for X1 and 'a'∈] -∞; 0,338] for X2as shown in Figure 1.b. The other symbols have the same behavior as the first symbol 'a'.

Figure 1. SAX representation of: (a) non-normalized X1 and X2, (b) normalized X1 and X2.



The *second* problem is observed after applying the multi-series SAX method, that is, processing over all time-series, based on the mean, variance and standard deviation or SD (Razmjooy et al., 2016). The generated representations are not identical for normalized v. non-normalized series, but the behavioral observation of these series shows that there is a loss of meaning after the normalization process. Without normalization (Figure 2.a), the curve that presents the series X1 appeared totally under the curve of X2, in which all data values of X1 are inferior to X2 values. This apparently shows a meaningful behavior with the non-normalized dataset. Comparatively, with normalization (Figure 2.b), most, but not all, of the parties of the X2 curve are presented below the portions of X1, which implies the loss of data meaning and the behavioral direction of the two series.

Figure 2. Multi-series SAX representation of: (a) non-normalized X1 and X2, (b) normalized X1 and X2.



In addition to the information loss during normalization, the SAX technique requires an input parameter to specify the resulting symbol series length, which constitutes a *third* problem. This parameter must be less than or equal to the length of series to be represented; otherwise, all series that have a minimum length will be wasted. The impact of this criterion can lead to the loss of all data should and if the data contain only short series.

To eliminate these limitations, including the information and data loss, we propose a novel model representation that would be more protective on both information and data.

3. OUR PROPOSED MODEL

3.1 Problem Formulation

To formalize our proposed model, we use D to notify the EHR dataset that includes a patient set $P=\{P_1, P_2, ..., P_n\}$ and an event set $E=\{E_1, E_2, ..., E_m\}$ captured on time T as represented by the observational chronologies $\{T_1, T_2, ..., T_q\}$ for each event.

Let e_{ijr} represents the observation value on the event E_i for the patient P_j and the chronology T_r. We then use $E_i.T$ to present the longest series chronology in an event E_i. Let *NuE* represents the numeric events set and *NoE* represents the nominal events set:

where,

$$\forall E_i \subset NuE \wedge \forall e_{ijr} \in E_i \Rightarrow e_{ijr} \in \mathbb{R} \tag{1}$$

$$\forall E_i \subset NoE \wedge \forall e_{ijr} \in E_i \Rightarrow e_{ijr} \text{ is a String Type} \tag{2}$$

Equation 1 means that all observations of numerical events are real type values whereas Equation 2 means that all observations of nominal events are string type values.

We use IWk_i for intra-class inertia and IBk_i for inter-class inertia in relation to the event E_i and the clusters number k_i, with the ITk_i to be their total inertia(Choukri et al., 2019).

3.2 Description of the Model

Our approach treats structured, temporal and non-temporal data vis-à-vis their differing data types. As depicted in Figure 3,with data types to be inclusive of numeric, date, binary, and nominal, our Data Representation model per Region and Dispersion (DRRD) may be divided vertically into two main parts. The first part treats the numeric and date type data while the second part treats the nominal and Boolean type. Each part comprises several steps, including data representation and/or transformation, event marking, and result linearization.

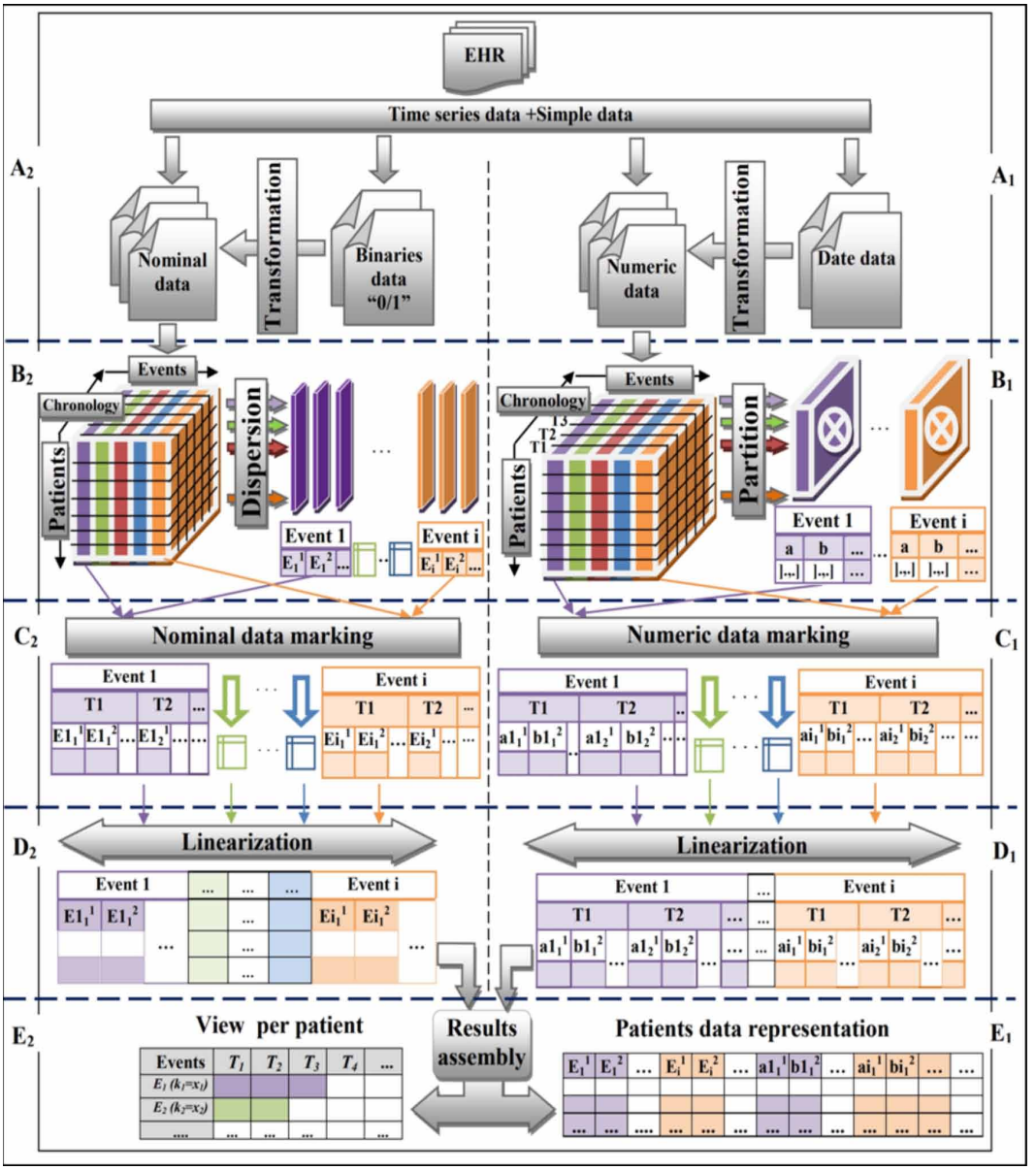
3.2.1 Numeric & Date Data Representation

This step selects the numerical type of data and transforms all date data types into a numeric (Part A1).

For all values e_{bj1} a transformation of the birth date event E_b is performed into the patient age. The other dates e_{ijr} are transformed by computing the observation appearance year according to the following formula:

$$DateToNumeric\left(e_{ijr} \,/\, e_{bjl}\right) = \; NbYears\left(e_{ijr} - e_{bjl}\right) \tag{3}$$

Figure 3. Proposed personalized medicine data representation model.



The function $NbYears(e_{ijr} - e_{bj1})$ computes the difference between the date of e_{ijr} observation and the birth date e_{bj1} in terms of the years number, that is, it calculates the e_{ijr} observation age.

All simple or temporal data are presented as numeric temporal events, with this task to be presented as a time-series representation task. We consider event chronologies, and the result is a three-dimensional data reorganization (3D: P,NuE,T).

3.2.2 Partitioning Numeric Events

Clustering (or, Partitioning) is the task of grouping objects into subsets known as clusters based on a given similarity criteria between the object's properties to be examined (Hancer, 2020). Each

clustering technique has a specific strategy for defining the clusters while associating an object to each cluster with respect to the cluster centers.

The clustering intends forming similar groups (clusters or classes) based on the ordered data of the same event. These clusters will be the belonging regions and the data comparison units of each patient on this event.

In this step (Part B_1), for each event E_i in the reorganized data cube with time factor being neglected, we apply a clustering technique to produce k_i clusters and their ordered centers list (C_1, C_2,...,C_p,..., C_u,...,C_{ki}) as given in Equation 4. Figure 4details this step.

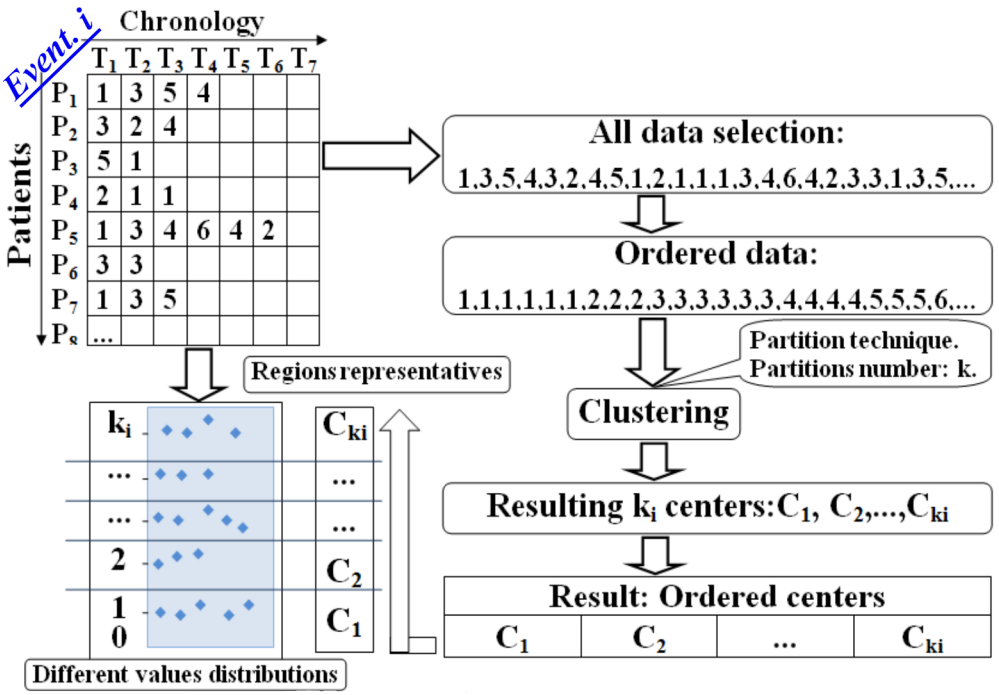
$$\forall p < u \wedge \forall u \leq k_i \Rightarrow C_p < C_u \tag{4}$$

The centers of clusters are ordered for their appropriate use with the symbols of an alphabet S to be defined later. Treatments sequentially include data selection, data sorting, data clustering, and finally the sorting of the resulting cluster centers as applied to each event E_i. The association of relevant data to the region of the nearest center will inform the data distribution in the regions on the ordered centers.

By distribution, we mean the manner of data dissemination, organization and display in the data presentation space. Four (4) clustering techniques have been evaluated: PAA or Piecewise Aggregate Approximation (e.g., Seunghye, 2017; Vineetha & Heggere, 2014), K-MEANS (e.g., Hartigan & Wong, 1979), EM or Expectation-Maximization (e.g., Dempster et al., 1977), MDBC or Make Density Based Clusterer, based on hierarchical clustering techniques (e.g., Witten & Frank, 2005).

Briefly, the principle of each of the referenced techniques in the sequel is as follows:

Figure 4. An example of the clustering process of a numeric Event E_i.



- **PAA:** Partitions a series of length n into N segments of length n/N, and the average data belonging to each segment is the representation of the latter.
- **K-MEANS:** Involves partitioning a dataset into a given number of clusters, each cluster being associated with a center point known as the centroid. Each point is assigned to the cluster with the closest centroid.
- **EM:** A probabilistic extension of the K-means algorithm, this iterative technique has been developed for incomplete data cases based on two steps; the first evaluates the expectation (E) while the second maximizes (M) the conceivable expectation.
- **MDBC:** A meta-clustering technique that relies on the results of a basic clustering approach to generate a probability distribution and a corresponding density for all observations. Specifically, the basic technique we use herein is the hierarchical clustering.

Each of these techniques will be applied to form k_i clusters on each event E_i. Several k_i values will be tested and the best result will be considered. Each event E_i must have different list values (L_i) length with $length(L_i)$ greater than or equal to k_i:

$$\forall e_{ijr},\ e_{ils} \in L_i \Rightarrow e_{ijr} \neq e_{ils} \wedge length\left(L_i\right) \geq k_i \tag{5}$$

This is the task (Part C_1) of notifying the values of each numeric reorganization cube event in a separate table. Each table generated for an event E_i has a number of columns, $ColumnsNumber(E_i)$ being equal to:

$$ColumnsNumber\left(E_i\right) = k_i * length\left(E_i.\ T\right) \tag{6}$$

where,

The function $length(E_i.\ T)$ means the longest series T size in the event E_i.

The notification for each event E_i consists in indicating on its notification table $MarkingTable_i$ one cell for each value of each series. This indication must be pointed in front of the cell corresponding to the nearest cluster center,

where,

$$\forall e_{ijr} \in E_i,\ \forall C_{NC} \in \left\{C_1,\ldots,C_{k_i}\right\} /\ Distance\left(C_{NC},e_{ijr}\right) = \min_{p=1,\ldots,k_i}\left(Distance\left(C_p,e_{ijr}\right)\right) \Rightarrow MarkingTable_i\left[j\right]\left[\left((r)*k_i\right)+NC\right] = \mathrm{X} \tag{7}$$

$Distance(C_{NC},e_{ijr})$ is the subtraction function between the C_{NC} center and the e_{ijr} observation.

3.2.3 Numeric Data Marking

The marking process (see *Algorithm 1*) fills the cells indicated in the notifications table. We propose three (3) marking types.

The first type is marking via real value as illustrated in Table 1, comprising the rewriting of the data value in the indicated cell ($X=e_{ijr}$). The second type, the binary marking (as illustrated in Table 2), is a presence signaling form. It notifies by "1" for the corresponding cells (X =1) and for the others by "0" (X=0). The last type is the marking via a symbol (as illustrated in Table 3). As centers are being ordered, we can associate a symbol from a predefined and ordered alphabet to each one. The marking task places the symbol that carries the same order with the center used during the notification (X = *S*[NC]).

```
Algorithm 1: Numeric Event Ei Marking.
EventMarking(P, E_i, Centers_i, MarkingType){
MarkingTable_i=new Table[n][k_i* Length(E_i. T)];
if(MarkingType= Symbol)// Define an ordered alphabet S.
Create an alphabet S of ordered symbols S_0,S_1,...,S_k;
for all P_j in P do{
int h=0;
for all T_r in E_i.T do{
NC=NearestCenter(Centers_i, e_ijr); // Nearest center index
if(MarkingType= Real Value)// Marking per real value.
MarkingTable_i[j][(h* k_i)+NC]=e_ijr;
if(MarkingType= Binary) // Binary marking.
 MarkingTable_i[j][(h* k_i)+NC]= 1;
if(MarkingType= Symbol) // Marking per symbol.
 MarkingTable_i[j][(h* k_i)+NC]= S_NC;
h++;
}
}
return MarkingTable_i;
}
```

Table 1. Marking by Real value.

$\mathbf{E_i}$						
T_1			T_2			…
C_1	...	C_{ki}	C_1	...	C_{ki}	...
					5.4	...
3						
…						…

Table 2. Binary marking.

E_i						
T_1			T_2			…
C_1	...	C_{ki}	C_1	...	C_{ki}	...
0		0	0		1	...
1		0	0		0	
…						…

Table 3. Marking by Symbol.

E_i						
T_1			T_2			…
C_1	...	C_{ki}	C_1	...	C_{ki}	...
					c	...
a						
…						…

3.2.4 The Numeric Data Linearization Task

Part D1 entails arranging the marking tables into a single table. Two (2) such arrangements are proposed:

- The first is the numeric event linearization (as illustrated in Table 4). In a table that collects all the events, we arrange(denoted by $\prod$) the marking tables one after the other,

  where,

$$EventLinearization = \prod_{i=1,..,m} (MarkingTable_i) \tag{8}$$

Table 4. Event linearization.

E_1							E_2							...	E_m							
T_1			T_2			...	T_1			T_2			...	…	T_1			T_2			…

- The second is the chronological linearization (as illustrated in Table 5). It joins (denoted by $\sum$) all columns of T_1, the first times in a marketing table with T_2, the columns of the second times; then with T_3 and so on until the last time is reached:

$$ChronologicalLinearization = \sum_{j=1,..,SizeOf(E_1.\ T)}^{i=1,..,m} (MarkingTable_{iT_j}) \tag{9}$$

Table 5. Chronological linearization.

E_1			E_2			...	E_m			E_1			E_2			... ...	E_m			...
T_1			T_1			...	T_1			T_2			T_2			...	T_2			...

3.2.5 The Nominal & Boolean Data Representation

Part A2 considers the data of nominal and Boolean type.

The Boolean data transformation task converts temporal and non-temporal Boolean data into symbolic data. It entails replacing all values containing "0" by "F" and the value containing "1" by "Y":

$$\begin{cases} BooleanToNominal(e_{ijr}) = \text{"F"}\ if\ (e_{ijr} = 0) \\ \qquad\qquad and \\ BooleanToNominal(e_{ijr}) = \text{"Y"}\ if\ (e_{ijr} = 1) \end{cases} \tag{10}$$

The result will be a dataset that contains temporal as well as non-temporal nominal data. In this new nominal reorganization of events and following their chronologies, the data are represented according to three axes (P, NoE, T).

3.2.6 Dispersion "Scattering" of Nominal Events

Nominal event dispersion process (Part B2) selects the different values set L_i of each event E_i and creates a new corresponding empty table $DispersionTable_i$.

We associate $L_i.length$ cells for each value in the longest series $E_i.T$ of this event E_i. The columns number in each table must be equal:

$$\forall P_j \in P,\ DispersionTable_i[P_j].length = L_i.length * length(E_i.\ T) \tag{11}$$

where,

$DispersionTable_i[P_j].length$ is the columns number in the table $DispersionTable_i$ with all patients P_j.

Algorithm 2 describes this step:

```
Algorithm 2: Nominal Event E_i Dispersion
EventDispersion(E_i){
Set L_i = DistinctDataSelection(E_i); //Distinct data values
selection.
DispersionTable_i=new Table[n][L_i. length * length(E_i. T)];
```

```
Y=0;
For all z in (0,..., length(E_i. T)-1) do {
For each value v in L_i do { //Labeling columns.
DispersionTable_i. Column[(z* L_i. indexOf(v))+Y]. Name = E_i.
ID+"_"+z+"_"+v;
Y++;
}
}
return DispersionTable_i and L_i;
}
```

3.2.7 Nominal Data Marking

The notification (Part C2) uses the result of the previous task. We use the chronology value r of each e_{ijr} observation as an index to notify all the nominal data on the dispersion tables:

$$\forall e_{ijr} \in E_i,\ DispersionTable_i\left[j\right]\left[\left(r^{*}L_i.\ length\right)+L_i.\ indexOf\left(e_{ijr}\right)\right]=\ \mathrm{X} \tag{12}$$

The function $L_i.\ indexOf(e_{ijr})$ returns the e_{ijr} observation index in the set L_i of different values to the event E_i.

The three (3) types of markings described above are reused.

Algorithm 3 describes these steps.

```
Algorithm 3: Nominal Event E_i Marking.
EventMarking(P, E_i, L_i, DispersionTable_i) {
MarkingTable_i= DispersionTable_i;
if(MarkingType_i= Symbol) {// Define an ordered alphabet S.
w= L_i. length-1;
Create S of an ordered symbols S_0,S_1,...,S_w ;
}
for all P_j in P do {
for all T_r in E_i.T do {
SI= L_i. indexOf(e_ijr);
RI=(r* L_i.length)+SI;
if(MarkingType= Real Value)  // Marking per real value.
 MarkingTable_i[j][RI]= e_ijr;
if(MarkingType= Binary)  // Binary marking.
 MarkingTable_i[j][RI]= 1;
if(MarkingType= Symbol)  // Marking per symbol.
 MarkingTable_i[j][RI]= S_SI;
}
}
return MarkingTable_i;
}
```

3.2.8 The Nominal Data Linearization Task

Using the same numeric data linearization logic to collect the nominal events marking results into a single table (Part D2). There are always two (2) types of rearrangement, that via the nominal event v. that via the chronology of nominal observations.

3.2.9 Results Assembly

The two (2) linearization parts of numeric and nominal data require the assembly of their results. Two (2) proposals are advanced.

The first (part E1) joins the results of the numeric linearization and the nominal linearization in a single global representation for all patients. The consequences of this operation achieve a representation via real values, symbols, or binary representation. These results are two-dimensional tables, which highlights the simplicity of our final representation.

The second (part E2) is the view per patient (as illustrated in Table 6). Here, the proposal is to provide assistance to the health professionals, giving a clear idea of the variations in each event observations for a given patient. The view per patient comprises associating a data table to the patient P_j. The rows correspond to the captured events vis-à-vis the patient P_j representation. For each cell k_i at each time T_r, we take only the non-null cells that contain the representation symbol. The created table must contain q columns so that T_q is the longest series $E_i.T$ chronology for the event E_i. To provide the longest series $E_i.T$ chronology information, we will color only the cells whose position is less or equal to the $E_i.T$ length.

Table 6. View per patient prototype.

Events	T_1	T_2	T_3	T_4	...
E_1(On k_1 Clusters)					
E_2(On k_2 Clusters)					
E_3(On k_3 Clusters)					
... .	...	...	...		...

4. EXPERIMENTATION

For the model evaluation, we apply it to a real EHR dataset and compare the obtained representations with the results of the state-of-the-art SAX technique.

4.1 Dataset Description

We use the free data from OpenMRS Wiki (2018), system "Open Medical Record System". The OpenMRS platform in OpenMRS Project (2004) is an application for personalized EHRs. Based on 2,528 medical concepts, this dataset stores 476,973 observations for 5,284 patients. Two (2) diseases are observed in this dataset, namely, the "HIV Program" v. the "TB Program".

4.2 Data Selection/Transformation/Codification

For our experiments, we use only observations on the disease "TB Program". Only concepts that have more than one observation are selected. We obtained 21 numerical events, two (2) nominal events and four (4) date type events require a transformation operation. However, this dataset does not include observations of temporal nominal and Boolean type.

In the first step and after the transformation, the data have been re-arranged in a three-dimensional (3D) form, thereby repositioning twenty-five (25) numeric events.

Table 7 presents the statistics on ordered numerical events with each frequency.

For the second step, the three-dimensional (3D) form of nominal events uses only the "GENDER" observations data and the "TRIBE" observations that are only available of this type.

Table 7. Statistics of numeric events.

N°	Concept ID / Concept Name	Different instances number	Frequency
1	21 / HEMOGLOBIN	82	282
2	654 / SERUM GLUTAMIC-PYRUVIC TRANSAMINASE	207	278
3	678 / WHITE BLOOD CELLS	81	282
4	729 / PLATELETS	204	282
5	730 / CD4%	52	476
6	790 / SERUM CREATININE	154	160
7	851 / MEAN CORPUSCULAR VOLUME	58	282
8	853 / CD8 COUNT	396	470
9	952 / ABSOLUTE LYMPHOCYTE COUNT	190	275
10	980 / BODY SURFACE AREA	3	3
11	1113 / TUBERCULOSIS DRUG TREATMENT START DATE	7	9
12	1279 / NUMBER OF WEEKS PREGNANT	19	38
13	5085 / SYSTOLIC BLOOD PRESSURE	31	2. 144
14	5086 / DIASTOLIC BLOOD PRESSURE	21	2. 143
15	5087 / PULSE	113	2. 269
16	5088 / TEMPERATURE (C)	64	2. 274
17	5089 / WEIGHT (KG)	216	2. 262
18	5090 / HEIGHT (CM)	79	135
19	5092 / BLOOD OXYGEN SATURATION	25	2. 267
20	5096 / RETURN VISIT DATE	2. 049	2. 265
21	5242 / RESPIRATORY RATE	15	105
22	5314 / HEAD CIRCUMFERENCE	31	122
23	5497 / CD4 COUNT	330	478
24	5599 / DATE OF CONFINEMENT	6	7
25	5919 / BIRTH YEAR	803	824

Table 8 shows the statistics.

Table 8. Statistics of nominal events.

N°	Concept ID / Concept Name	Different instances number	Frequency
1	992843 / GENDER	2	824
2	992844 / TRIBE	3	824

4.3 Results & Discussion

In line with the minimum length constraint (see Equation 5) vis-à-vis the list of different values applied on the numeric events to be partitioned, we have found that a single case generated by the event "980" has three (3) different values so that the clustering process is executed only for k= 2 and k= 3. The other cases of k = 4, 5 and 6 are not applicable.

Table 9. Statistics of series maximums lengths and the columns produced number.

N°	Concept ID	Max Length of the longer series	Produced columns number				
			k=2	k=3	k=4	k=5	k=6
1	21	3	6	9	12	15	18
2	654	2	4	6	8	10	12
3	678	3	6	9	12	15	18
4	729	3	6	9	12	15	18
5	730	3	6	9	12	15	18
6	790	2	4	6	8	10	12
7	851	3	6	9	12	15	18
8	853	3	6	9	12	15	18
9	952	3	6	9	12	15	18
10	980	1	2	3			
11	1113	3	6	9	12	15	18
12	1279	4	8	12	16	20	24
13	5085	10	20	30	40	50	60
14	5086	10	20	30	40	50	60
15	5087	10	20	30	40	50	60
16	5088	10	20	30	40	50	60
17	5089	10	20	30	40	50	60
18	5090	4	8	12	16	20	24
19	5092	10	20	30	40	50	60
20	5096	9	18	27	36	45	54
21	5242	4	8	12	16	20	24
22	5314	4	8	12	16	20	24
23	5497	3	6	9	12	15	18
24	5599	2	4	6	8	10	12
25	5919	1	2	3	4	5	6

The columns are named in relation to the event identifier, the cluster number and the observation number in the series, for example, if we take k =2 to partition the event "21 / HEMOGLOBIN," the maximum series of this event is observed on the patient number "3618" with three (3) values, which produces a representation with six (6) columns; in this case, "21_C0_S0, 21_C1_S0, 21_C0_S1, 21_C1_S1, 21_C0_S2, 21_C1_S2".

Table 9 summarizes the statistics of the lengths of series maximums and the produced columns number (i.e., the length of the empty notifications tables to be created) during the partitioning of each event.

Data dispersion affects only the two (2) selected nominal concepts. The first "992843/GENDER" generates a dispersion table with two (2) columns"992843_0_M and 992843_0_F". The second concept "992844/TRIBE" generates a dispersion table with three (3) columns "992844_0_Luo, 992844_0_Luhya, 992844_0_Unknown".

Table 10 shows the data notification values for the numeric event "790/SERUM CREATININE" corresponding to the patient identified by "5126". We use the three (3) markings types, and the K-MEANS technique with k=2. The alphabet S={a,b} of marking symbols has only two (2) symbols vis-à-vis the number of centers.

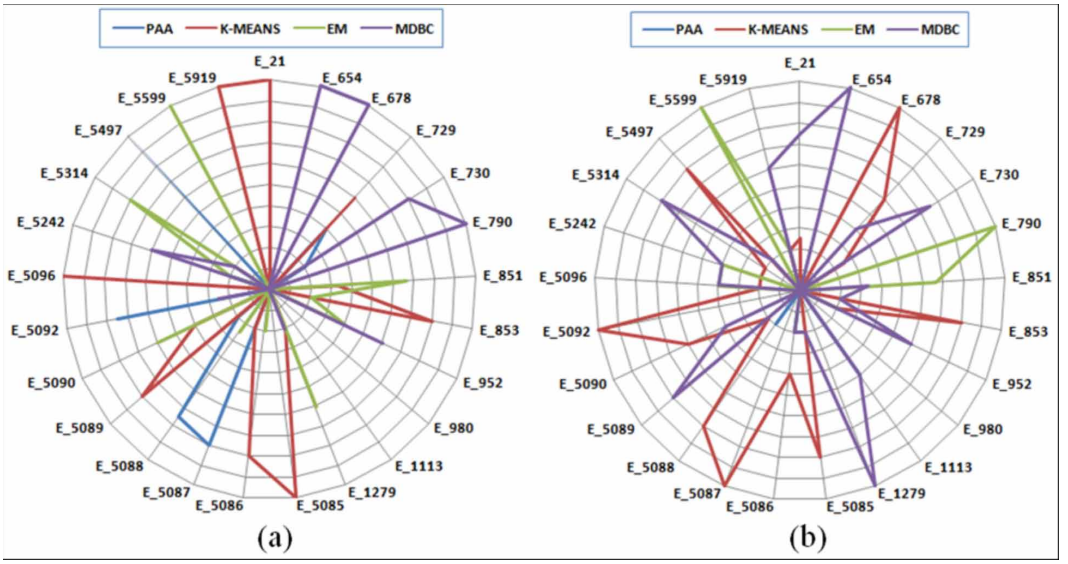
Table 10. Numeric event notification.

Marking types	790_C_0–S_0	790_C_1–S_0	790_C_0–S_1	790_C_1–S_1
Real value		27. 000,0	51,9	
Binary		1	1	
Symbol		b	a	

Table 11. Nominal event notification.

Marking types	992843_0_F	992843_0_M
Real value	F	
Binary	1	
Symbol	a	

Figure 5. Appropriate clustering techniques corresponding to statistics of: (a) intra-class inertia, (b) inter-class inertia.



Also, Table 11 shows the nominal events "992843/GENDER" notification for the same patient.

As argued by Chenguang et al. (2018), a better partition has either the lowest intra-class inertia, or the highest inter-class inertia. Hence, in order to compare the used classification techniques, we evaluate the results of our numerical events representation, at the base of intra and inter-class inertia. Radar graphs shown in Figure 5.a and Figure 5.b have been generated based on the results of intra-class and inter-class inertia respectively. These graphs visually highlight the statistics of the

Table 12. Global statistics of intra and inter class inertia.

Inertia	PAA	K-MEANS	EM	MDBC
Intra class	4/23	8/23	5/23	6/23
Inter class	0/21	10/21	3/21	8/21

Table 13. Chosen techniques and clusters numbers results.

N°	Concepts ID	Chosen technique	Chosen k_i
1	21	MDBC	4
2	654	PAA	5
3	678	EM	2
4	729	MDBC	3
5	730	EM	2
6	790	PAA	2
7	851	PAA	2
8	853	MDBC	2
9	952	EM	2
10	980	PAA	2
11	1113	PAA	4
12	1279	MDBC	4
13	5085	MDBC	4
14	5086	MDBC	3
15	5087	MDBC	4
16	5088	MDBC	3
17	5089	MDBC	4
18	5090	MDBC	2
19	5092	K-MEANS	6
20	5096	MDBC	5
21	5242	MDBC	2
22	5314	K-MEANS	4
23	5497	MDBC	2
24	5599	PAA	3
25	5919	MDBC	2

used clustering techniques of PAA, K-MEANS, EM, and MDBC and illustrate the most appropriate clustering technique according to the event considered.

For each technique, we associate a best cases counter. This counter will be initialized to zero for each event; for each event and for each k_i (k_i=2,..., k_i =6), we incrementally adjust the counter of the technique corresponding to the best inertia. If several techniques have the same best inertia in a given partition k_i, we eliminate these results.

Altogether, Table 12 summarizes the global situation for the techniques used. For twenty-three (23) observed cases on twenty-five (25) events in the intra-class inertia evaluation, the K-MEANS technique is found to be superior to all the other approaches, followed by the MDBC, EM, and PAA techniques. The same result is obtained when evaluating the inter-class inertia, and for twenty-one (21) observed cases on twenty-five (25) events, the K-MEANS technique always remains the best.

The K-MEANS technique is superior among all others, and it can be used with all events. Yet, the qualities of each event, especially the distribution of the values, push us to choose the most appropriate classification technique for each event. To choose the technique and the number of clusters k_i for each event E_i, we compute for all k_i and all techniques the total inertia ITk_i:

$$IT_{k_i} = IW_{k_i} + IB_{k_i} \tag{13}$$

The technique and the k_i corresponding to the maximum ITk_i will be used for the event E_i representation.

Table 13 displays the evaluation results.

For the nominal data dispersion statistics, no loss of information has occurred. The instances sum in each dispersion table of event "GENDER" (Figure 6.a) and event "TRIBE" (Figure 6.b) equals the total number of patients (824).

Finally, the global representation has been generated on 431 columns. Of these, 426 columns associate with numeric events, several between them containing time-series in its initial form. The remaining 5 columns are associated with nominal events. These results prove the capacity of time-series treatments via our model.

Figure 6. Dispersion of nominal events: (a) "GENDER", (b) "TRIBE".

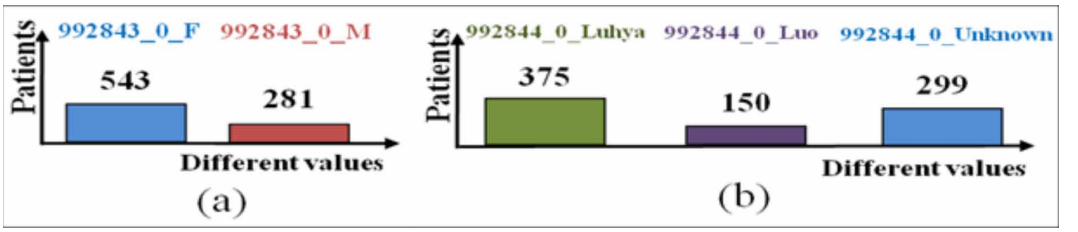


Figure 7. Part of a representation by real value.

E_5088_0	E_5088_1	E_5088_2	E_5088_3	E_5088_4	E_5088_5	E_5088_6	E_5088_7	E_5088_8
	36.40			36.30				
	35.80							
		39.50			38.40			39.70
	36.10			35.80				37.30
	36.00			36.10				36.80
	36.50				36.90			36.90

Depending on the type of marking used, we obtain a global representation via real value (as shown in Figure 7), a binary global representation (as shown in Figure 8) or a global representation via symbol (as shown in Figure 9).

The binary and symbolic global representations use only one type of data, which provides homogeneous representations and eliminates heterogeneity in the initial data types of events "Numeric, Nominal, Date, Boolean".

Figure 8. Binary representation part.

E_5088_0	E_5088_1	E_5088_2	E_5088_3	E_5088_4	E_5088_5	E_5088_6	E_5088_7	E_5088_8
	1			1				
	1							
		1			1			1
	1			1				1
	1			1				1
	1				1			1

Figure 9. Part of a representation by symbol.

E_5088_0	E_5088_1	E_5088_2	E_5088_3	E_5088_4	E_5088_5	E_5088_6	E_5088_7	E_5088_8
	b			b				
	b							
		c			c			c
	b			b				c
	b			b				c
	b				c			c

4.4 Representation Example & Evaluation

To evaluate our approach, we compute three (3) data symbolic representations of the patient identified with the id =75. Table 13 results have been enlisted for this purpose. The first representation demonstrates our DRRD approach whereby each event representation is taken separately from the others. The strategy entails concatenating the representation symbols of each event vis-à-vis their chronological order. The second and third representations apply the SAX technique with two (2) values different of the segments number W. W indicates the resulting representation length, which must be less than or equal to the series length in treatment. $W=1$ is used here as the minimum length acceptable by SAX in the second representation named SAX1, with $W=10$ as the maximum length corresponding to the longest series for the third representation named SAX10. The SAX technique requires an input symbol number Z comparable to the number of clusters. We then apply SAX1 and SAX10 on each event with $Z_i=k_i$.

Table 14 shows three (3) events symbolic representations to showcase the essential points in the comparison. Empty cells allude to the unobserved events (e.g., event 790) on this patient. Cells containing "/" are presented only for the SAX technique, and are informed by the minimum lengths conditions not respected for the series to be presented (e.g., event 21). For the nineteen (19) marked events, DRRD and the SAX1 generate new representations, whereas for the SAX10, new

representations are generated for 8 events only. Additionally, our model represents each observation, whatever the length of the series. Even so, the SAX technique generates only the results with the same length if the series have passed the minimum length condition *W*. More generally, the SAX technique eliminates all series having a length less than *W*, which constitutes an information loss that directly affects the quality of the obtained representation. For instance, patients representation having only series lengths less than or equal to one will fail directly with the SAX technique and the parameter $W \geq 2$.

This example demonstrates the capacity of the proposed novel model to preserve time-series information and to conserve as much data as possible throughout the process of representation and transformation.

Table 14. Representation by symbol of the patient identified by id = 75.

Event Id	Z=k_i	DRRD	SAX1	SAX10
21	4	c	d	/
654	5	a	c	/
678	2	a	b	/
729	3	a	a	/
730	2	a	a	/
790	2			
851	2	a	a	/
853	2	a	b	/
952	2	a	b	/
980	2			
1113	4			
1279	4			
5085	4	dcdccccdd	c	dbdaadcbdd
5086	3	ccccccccc	b	cacaabbacc
5087	4	ccccccccc	c	bcabdcbbbd
5088	3	bbbbbbbbb	c	bccccccccc
5089	4	ccccccccc	c	cbccccccc
5090	2			
5092	6	adbbbbbabb	a	adaaaaaaaa
5096	5	ccccccccc	e	/
5242	2			
5314	4			
5497	2	a	a	/
5599	3			
5919	2	b	b	/
992843	2	b	b	b
992844	3	a	a	a

4.5 View Per Patient Example

This solution consists in presenting a window on the data and its variations for each patient taken individually. Figure 10 shows an instance of this window for the patient identified by Id = 75. This data visualization can help practitioners in analyzing the observations. In this example, the patient has only one appearance of the "HEMOGLOBIN" event. The 'c' character encodes this event as it corresponds to the third level of this event among 4 levels (the number of clusters). The practitioner can also know the maximum length of the time-series of this event according to the number of colored cells. For the "HEMOGLOBIN" event, the time-series include a maximum of three samples. According to this information, a practitioner can easily decide if the patient has a strong need for other specimens or diagnostics. After certain experiments on the patients, the practitioners can notice then the key events for a given disease. This allows them to directly examine the desired event on this dashboard for the suspicious cases. Overall, this solution presents a detailed view of the data of a given patient in view of all the patient data.

Figure 10. View per patient example "Patient Id=75".

N	Concept Name	T_0	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9
1	HEMOGLOBIN (On 4 Clusters)	c									
2	SERUM GLUTAMIC-PYRUVIC TRANSAMINASE (On 5 Clusters)	a									
3	WHITE BLOOD CELLS (On 2 Clusters)	a									
4	PLATELETS (On 3 Clusters)	a									
5	CD4% (On 2 Clusters)	a									
6	SERUM CREATININE (On 2 Clusters)										
7	MEAN CORPUSCULAR VOLUME (On 2 Clusters)	a									
8	CD8 COUNT (On 2 Clusters)	a									
9	ABSOLUTE LYMPHOCYTE COUNT (On 2 Clusters)	a									
10	BODY SURFACE AREA (On 2 Clusters)										
11	TUBERCULOSIS DRUG TREATMENT START DATE (On 4 Clusters)										
12	NUMBER OF WEEKS PREGNANT (On 4 Clusters)										
13	SYSTOLIC BLOOD PRESSURE (On 4 Clusters)	d	c	d	c	c	c	c	c	d	d
14	DIASTOLIC BLOOD PRESSURE (On 3 Clusters)	c	c	c	c	c	c	c	c	c	c
15	PULSE (On 4 Clusters)	c	c	c	c	c	c	c	c	c	c
16	TEMPERATURE (C) (On 3 Clusters)	b	b	b	b	b	b	b	b	b	b
17	WEIGHT (KG) (On 4 Clusters)	c	c	c	c	c	c	c	c	c	c
18	HEIGHT (CM) (On 2 Clusters)										
19	BLOOD OXYGEN SATURATION (On 6 Clusters)	a	d	b	b	b	b	b	a	b	b
20	RETURN VISIT DATE (On 5 Clusters)	c	c	c	c	c	c	c	c	c	
21	RESPIRATORY RATE (On 2 Clusters)										
22	HEAD CIRCUMFERENCE (On 4 Clusters)										
23	CD4 COUNT (On 2 Clusters)	a									
24	DATE OF CONFINEMENT (On 3 Clusters)										
25	BIRTH YEAR (On 2 Clusters)	b									
26	GENDER (On 2 Clusters)	b									
27	TRIBE (On 3 Clusters)	a									

5. CONCLUSION

In this paper, we first discuss PM, and its associated data, citing some types of patient profile data. The need to consider the composition of these data and to extract the hidden knowledge underlying the data have become a necessity for various critical data-intensive PM tasks, including reporting, evaluation, analysis, and medical decision-making.

Eventually, we argue that the patient data preparation and its representation can provide a staging basis for these different PM tasks. To date, the exploration of data for realizing PM has attracted a plethora of research. Notwithstanding, most of these works carry certain limits in their methodologies. Inevitably, the heterogeneity and number of data types, the information and data loss are the key challenges, and in order to showcase these difficulties, we cite some of the prominent works and describe certain limits in their results.

Subsequently, we propose a novel representation model working on numeric, nominal, date and Boolean data types to ensure maximum data resource coverage. By means of transformations and treatments on the nominal and numeric data, our proposed model overcomes the data types heterogeneity issues. The nominal data are processed by dispersion, and the numeric data is processed by data belonging region (Clustering and nearest centers). Such a resolution has a strong contribution on the time-series representation. This last is treated as a part of numeric data. Following the experiments, a single global representation is obtained for all patients. This representation includes the basic data detail even if the data have only one observation. It conserves the information embedded within the data especially for time-series as it represents each observation in relation to its original region.

On the individual patient scale, the results of the view per patient shows other details at the level of each patient, including the absence and presence of medical events, the number of tests and the specimens from the patient and their levels in relation to a global scale (number of clusters) of the event examined. The comparison of our approach vis-à-vis the SAX technique and its parameters shows the high efficiency and the capacity of our proposed model to minimize the information and time-series data loss (as compared to SAX10, where the segments number W is large) during the transformation and to provide strong descriptiveness of patient data (as compared to SAX1, where W is small). Put simply, our approach has the capability to process several data types at the same time whereas the SAX technique processes only the numeric data.

Our work simplifies the understanding of personalized medicine data by healthcare practitioners. It also provides a tool for the analysis and comparison of data between patients, and facilitates the tracking of key events and important observations by practitioners. Both diagnosis and treatment planning can benefit of such a tool. In particular, we believe that our proposed representation makes EHR data exploration easier and can help practitioners to turn towards personalized medicine as well as using computer-aided decision systems that mine EHR data.

The other essential point of our model is the three marking types proposal "By real value, Binary, By symbol", and the production of a simple, homogeneous and unified representation in the form of a table for each marking type. Achieving this solution will open the way for our future work, such as medical decision-making, classification and data exploration. Our representation richness in terms of data types, and data mining richness in terms of classification algorithms availability and their capacities for processing different data types, will give us the possibility to do other work that takes this representation as a basis for initial and forward learning. Currently, we are actively working to apply our representation model to other similar data resources, even if they are not medical. We believe the step ahead is to tackle the challenges in transforming and representing massive datasets containing time-series data.

ACKNOWLEDGMENT

This work was realized as part of the Hubert Curien partnership (PHC) TASSILI cooperation program, between France and Algeria under the project code 19MDU212.

REFERENCES

Anima, S., Girish, N., Omri, G., Stephen, B. E., Erwin, P. B., & John, V. G. (2015). Incorporating temporal EHR data in predictive models for risk stratification of renal function deterioration. *Journal of Biomedical Informatics*, *53*, 220–228. doi:10.1016/j.jbi.2014.11.005 PMID:25460205

Bagattini, F., Karlsson, I., Rebane, J., & Papapetrou, P. (2019). A classification framework for exploiting sparse multi-variate temporal features with application to adverse drug event detection in medical records. *BMC Medical Informatics and Decision Making*, *19*(1), 7. doi:10.1186/s12911-018-0717-4 PMID:30630486

Chenguang, Y., Yuhang, Y., Xinyang, L., & Ruowei, W. (2018). Development of a neuro-feedback game based on motor imagery EEG. *Multimedia Tools and Applications*, *77*(12), 15929–15949. doi:10.1007/s11042-017-5168-x

Choukri, A., Hamzaoui, Y., Amnai, M., & Fakhri, Y. (2019). Classification Algorithm Based on Nodes Similarity for MANETs. *International Journal of Online and Biomedical Engineering*, *15*(05), 86–100. doi:10.3991/ijoe.v15i05.9742

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series A (General)*, *39*(1), 1–38.

Edward, C., Mohammad, T. B., Le, S., & Walter, F. S. (2017). GRAM: graph-based attention model for healthcare representation learning. *International Conference on Knowledge Discovery and Data Mining*, *23*, 787–795.

Ghazi, A. (2015). Mitigating the influence of the curse of dimensionality on time series similarity measures. *International Journal of Computer Applications in Technology*, *52*(1), 94–105. doi:10.1504/IJCAT.2015.071424

Guo, G., & Razmjooy, N. (2019). A new interval differential equation for edge detection and determining breast cancer regions in mammography images. *Systems Science & Control Engineering*, *7*(1), 346–356. doi:10.1080/21642583.2019.1681033

Hancer, E. (2020). A new multi-objective differential evolution approach for simultaneous clustering and feature selection. *Engineering Applications of Artificial, 87*.

Hartigan, J. A., & Wong, M. A. (1979). A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series A (General)*, *28*(1), 100–108.

Jensen, P., Jensen, L., & Brunak, S. (2012). Mining electronic health records: Towards better research applications and clinical care. *Nature Reviews. Genetics*, *13*(6), 395–405. doi:10.1038/nrg3208 PMID:22549152

Jing, Z., Panagiotis, P., Lars, A., & Henrik, B. (2017). Learning from heterogeneous temporal data in electronic health records. *Journal of Biomedical Informatics*, *65*, 105–119. doi:10.1016/j.jbi.2016.11.006 PMID:27919732

Kavya, N., Sriraam, N., Usha, N., Hiremath, B., Suresh, A., Sharath, D., Venkatraman, B., & Menaka, M. (2020). Breast Cancer Lesion Detection From Cranial-Caudal View of Mammogram Images Using Statistical and Texture Features Extraction. *International Journal of Biomedical and Clinical Engineering*, *9*(1), 16–32. doi:10.4018/IJBCE.2020010102

Khalilpour, M., Valipour, K., Shayeghi, H., & Razmjooy, N. (2013). Designing a Robust and Adaptive PID Controller for Gas Turbine Connected to the Generator. *Research Journal of Applied Sciences, Engineering and Technology*, *5*(5), 1543–1551. doi:10.19026/rjaset.5.4902

Kolekar, J. S., & Pawar, C. (2014). Clinical decision making using artificial neural network with particle swarm optimization algorithm. *International Journal of Research in Advent Technology*, *2*(1), 311–315.

Lin, J., Keogh, E., Wei, L., & Lonardi, S. (2007). Experiencing SAX: A novel symbolic representation of time series. *Data Mining and Knowledge Discovery*, *15*(2), 107–144. doi:10.1007/s10618-007-0064-z

Mallick, P., Seth, P., & Ghosh, A. (2018). Entropy-based fuzzy hybrid framework for gene prediction network – an application to identify and rank the biomarkers for human lung adenocarcinoma. *International Journal of Computers and Applications*, *41*(1), 62–77. doi:10.1080/1206212X.2018.1508865

Milad, Z. N., Dongxiao, Z., Najibesadat, S., & Kai, Y. (2018). *A Predictive Approach Using Deep Feature Learning for Electronic Medical Records: A Comparative Study*. Academic Press.

Namadchian, A., Ramezani, M., & Razmjooy, N. (2016). A New Meta-Heuristic Algorithm for Optimization based on Variance Reduction of Gaussian Distribution. *Majlesi Journal of Electrical Engineering*, *10*, 49–56.

Nguyen, T., Khosravi, A., Creighton, D., & Nahavandi, S. (2014). Classification of Healthcare Data using Genetic Fuzzy Logic System and Wavelets. *Expert Systems with Applications*, *42*(4), 2184–2197. doi:10.1016/j.eswa.2014.10.027

OpenMRS Project. (2004). https://openmrs.org

OpenMRS Wiki. (2018). *Demo Data* [Data file]. https://wiki.openmrs.org/

Park, H., & Jung, J. (2020). SAX-ARM: Deviant event pattern discovery from multivariate time series using symbolic aggregate approximation and association rule mining. *Expert Systems with Applications*, *141*, 141. doi:10.1016/j.eswa.2019.112950

Razmjooy, N., & Khalilpour, M. (2015). A new design for PID controller by considering the operating points changes in Hydro-Turbine Connected to the equivalent network by using Invasive Weed Optimization (IWO) Algorithm. *International Journal of Information Security and Systems Management*, *4*(2), 468–475.

Razmjooy, N., Khalilpour, M., & Ramezani, M. (2016). A New Meta-Heuristic Optimization Algorithm Inspired by FIFA World Cup Competitions: Theory and Its Application in PID Designing for AVR System. *Journal of Control. Automation and Electrical Systems*, *27*(4), 419–440. doi:10.1007/s40313-016-0242-6

Razmjooy, N., Mousavi, B. S., & Soleymani, F. (2013). A hybrid neural network Imperialist Competitive Algorithm for skin color segmentation. *Mathematical and Computer Modelling*, *57*(3-4), 848–856. doi:10.1016/j.mcm.2012.09.013

Razmjooy, N., & Ramezani, M. (2014). An improved quantum evolutionary algorithm based on invasive weed optimization. *Indian Journal of Scientific Research*, *4*(2), 413–422.

Seunghye, J. W. (2017). Data representation for time series data mining: time domain approaches. *WIREs Comput Stat, 9*(1), 1–6.

Vineetha, B., & Heggere, S. R. (2014). An Analysis of Time Series Representation Methods Data Mining Applications Perspective. *Proceedings of the 2014 ACM Southeast Regional Conference*.

Witten, H., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.

Yeh, W. C., Chang, W. W., & Chung, Y. Y. (2009). A new hybrid approach for mining breast cancer pattern using discrete particle swarm optimization and statistical method. *Expert Systems with Applications*, *36*(4), 8204–8211. doi:10.1016/j.eswa.2008.10.004

Kadi Hafid is a PhD student in computer science, he is preparing his doctoral thesis between the university of Mustapha Stambouli, Mascara (Algeria) and university of Caen Normandy (France). Data mining and medical data exploration are the main axis of his research.

Mohammed Rebbah obtained his MSC in computer science from University of Sciences and Technology of Oran (USTO) Algeria. After he received his PhD in computer science in 2015 USTO, Algeria. Actually, he is an Assistant professor in Computer Sciences Department at the University of Mascara, Algeria. His research interest Grid computing, cloud computing and distributed data mining.

Meftah Boudjelal is an associate professor at the computer science department of Mascara University, Algeria. He obtained his engineering diploma in computer science in 1997 from Sidi Belabes University, his MSc in computer science option pattern recognition and artificial intelligence in 2005 from the University of USTOran, Algeria, and the Ph.D. degree in computer science from the University of USTO, Oran, Algeria in 2011. From September 2009 to August 2011, he joined GREYC Laboratory of the University of Caen Normandie in France as a member in Image Group. Actually, he is the head of research team (Medical Images Processing by New Bio-inspired Approaches) in University of Mascara. His research interests include pattern recognition application, neural network, spiking neural networks and also image processing. He has publications in many journals and participated in many international conferences. He served as a reviewer of known journals and joined some international conferences as organizing, a scientific program committee member.

Olivier Lézoray received the M.Sc. and doctoral degrees in computer science from University of Normandie, France, in 1996 and 2000, respectively. From 1999 to 2000, he was an assistant professor in the Computer Science Department at the same university. From 2000 to 2009, he was an associate professor at the Cherbourg Institute of Technology in the Communication Networks and Services Department. Since 2010, he has been a full professor at the Cherbourg Institute of Technology. His research is focused on color image segmentation and filtering (graph-based variational and morphological methods) and machine-learning techniques for image mining (neural networks and support vector machines).