Exploring time series retrieved from cardiac implantable devices for optimizing patient follow-up.

Marie Guéguin, Emmanuel Roux, Alfredo I Hernández, Fabienne Porée, Philippe Mabo, Laurence Graindorge, Guy Carrault

Exploring time series retrieved from cardiac implantable devices for optimizing patient follow-up

Guéguin Marie [1 *], Roux Emmanuel [2], Hernández Alfredo I [1], Porée Fabienne [1], Mabo Philippe [3], Graindorge Laurence [4], Carrault Guy [1]

[1] *LTSI, Laboratoire Traitement du Signal et de l'Image INSERM : U642, Université Rennes I, Campus de Beaulieu, 263 Avenue du Général Leclerc - CS 74205 - 35042 Rennes Cedex,FR*

[2] *LMTG, Laboratoire des Mécanismes et Transfert en Géologie CNRS : UMR5563, IRD : UR154, Observatoire Midi-Pyrénées, Université Paul Sabatier - Toulouse III, 14 avenue Edouard Belin 31400 Toulouse,FR*

[3] *Département de cardiologie et maladies vasculaires CHU Rennes, Hôpital Pontchaillou, Université Rennes I, 2 rue Henri Le Guilloux 35033 RENNES cedex 9,FR*

[4] *ELA Medical company Sorin Group, FR*

* Correspondence should be adressed to: Marie Guéguin <gueguinm@yahoo.fr>

Abstract

Current cardiac implantable devices (ID) are equipped with a set of sensors that can provide useful information to improve patient follow-up and to prevent health deterioration in the postoperative period. In this paper, data obtained from an ID with two such sensors (a transthoracic impedance sensor and an accelerometer) are analyzed in order to evaluate their potential application for the follow-up of patients treated with a cardiac resynchronization therapy (CRT). A methodology combining spatio-temporal fuzzy coding and multiple correspondence analysis (MCA) is applied in order to: i) reduce the dimensionality of the data and provide new synthetic indices based on the "factorial axes" obtained from MCA, ii) interpret these factorial axes in physiological terms and iii) analyze the evolution of the patient's status by projecting the acquired data into the plane formed by the first two factorial axes named "factorial plane". In order to classify the different evolution patterns, a new similarity measure is proposed and validated on simulated datasets, and then used to cluster observed data from 41 CRT patients. The obtained clusters are compared with the annotations on each patient's medical record. Two areas on the factorial plane are identified, one being correlated with a health degradation of patients and the other with a stable clinical state.

time-series ; trajectories ; monitoring ; cardiac implantable devices ; data mining

Introduction

Cardiac resynchronization therapy (CRT) is indicated for patients suffering from drug-refractory congestive heart failure (CHF) associated with intraventricular dyssynchrony [1]. CRT improves hemodynamic parameters, ejection fraction or distance covered in the 6 minutes walking test [2]. Furthermore, CRT has shown to decrease hospitalizations for patients treated with the implantable devices (ID). Although the efficiency of this treatment has been proven, 20 to 30% of patients show either no improvement or worsening of their symptoms [3].

Individual follow-up of implanted patients is a key to understand the difference between responders and non-responders, and to prevent severe health degradation. Besides regular follow-up visits, during the post-operative period, an everyday follow-up is possible with the new IDs recently developed for CRT. They offer an increased storage capability of data acquired by the ID, providing information on the ID itself (e.g. event counters of pacing and sensing activities) or on the state of the patient (e.g. arrhythmias, electrograms) and on its activity [4]. Recorded data are very promising towards the home monitoring of patients, the prediction of adverse events or the reduction of hospitalizations. However, this source of information is under-exploited because data are large, multivariate, time-dependent and heterogeneous, and consequently difficult to interpret for caregivers.

The objective of the present study is to propose a methodology to process this amount of multivariate data, in order to i) evaluate and extract the information content of the time-dependent data downloaded from the pacemaker memory, ii) define synthetic indices which are easy to interpret and iii) characterize and compare different populations of patients. Given the dimensionality of the recorded data, methods of data reduction are investigated. The interest of the multidimensional analysis of the data recorded in the ID memory to objectively assess the patients' response to the therapy and the validity of the exploratory techniques to process these data have been shown in two previous studies, using principal component analysis (PCA) [5] and multiple correspondence analysis (MCA) associated with a spatio-temporal fuzzy coding of the time-series [6]. The former method has been successfully used to differentiate a test population (patients with rate-responsive pacemakers)

from a population of patients suffering from CHF by jointly exploiting a number of physiological variables and using a simple representation for the temporal dimension. Providing an appropriate adaptation of its table of analysis, MCA has been successfully applied to the analysis of the evolution of time-series across time, and is then used here as well. MCA performs a reduction of the dimensionality of the data and provides synthetic indices called “factorial axes”. A plane formed by two factorial axes is called “factorial plane”. Each patient is finally represented by trajectories on the factorial plane.

From a methodological point of view, several questions are raised: i) how to link the factorial axes with the variables acquired from the ID? ii) do patients with similar trajectories on the factorial plane have a similar clinical state, and if yes, how to cluster patients according to their evolution in the factorial plane? and iii) are the obtained clusters consistent with the clinical data available from the patients? The study addresses the clustering of trajectories with different numbers of points in the factorial plane, which implies the choice of appropriate distance measure and clustering method. This problem is related to temporal clustering (i.e. the clustering of time-series) [7], [8].

The paper is organized as follows. In section II, the clinical protocol and the ID data are presented. Then, in section III, the proposed methodology is described and the issues arising from the clustering of trajectories are presented and solutions are proposed. The proposed methodology is tested on simulated datasets and applied on the real recorded data in Section VI. The validity of the obtained clusters is evaluated in comparison with the medical records of patients participating in the protocol.

Data recorded by cardiac implantable devices

Patients

Forty-one patients (34 males, 7 females) participated in the present study. The mean age was 64 (minimum 38 and maximum 87). They suffered from refractory heart failure (RHF) associated with intraventricular dyssynchrony and present a thin QRS complex (< 120 ms), a NYHA class from III to IV and a left ventricular ejection fraction (LVEF) = 25% (± 7). They were candidate for cardiac resynchronization therapy and were then implanted with cardiac implantable devices. The patients were informed about the research protocol and gave their fully informed consent for participating in this study.

Description of the follow-up time-series

Data stored in the ID memory are retrieved in individual records at the end of the third, the sixth and the twelfth postoperative months. Each record covers a three-month length period. These data result from two sensors [9]: a transthoracic impedance sensor which reflects the respiratory activity of the patient and its intensity of effort, and an 1-D accelerometer which is linked to the intensity of the physical activity of the patient.

By combining information from the two sensors, the activity level of the patients is classified automatically by the device into two states: exercise and rest. For each state, 24-hour cumulative values of a number of variables are computed and recorded in the ID memory over 30-day follow-up periods. More details concerning the two sensors can be found in [10], [11]. The final set of thirteen physiological variables is listed in Table I and is constituted of seven recorded variables and of six additional variables deducted from the seven recorded ones.

A methodology for clustering multivariate time-series

An overview of the proposed methodology is provided in Figure 1. The analysis is first performed on a reference population to determine significant factorial axes and clusters of similar evolutions in the factorial planes. In a follow-up situation, new data would regularly be retrieved from patients’ implantable cardiac devices. The results of the analysis (i.e. factorial axes and clusters) would then be used by projecting the new data on the factorial planes and assigning the evolution of the patient to an existing cluster towards the diagnosis (i.e. the patient is improving or degrading). In this paper, only the analysis is presented.

The analysis consists of 2 successive steps, namely fuzzy coding of the data and multidimensional analysis with smoothed multiple correspondence analysis (SMCA), described in the following subsections.

Space-time fuzzy coding

A coding of the recorded data is required, as MCA has been at first conceived for categorical variables. MCA exploits disjunctive tables $Z = (z_{ij})_{(i,j)\in[1,R]\times[1,C]}$ where z_{ij} is the membership value of the i^{th} object to the j^{th} modality. As a coding of the multivariate time-series, a fuzzy space-time windowing, defined by Loslever and Bouilland for characterizing and coding biomechanical temporal data [12], is proposed. Instead of an indicator matrix Z (i.e. $z_{ij} \in \{0,1\}$), the MCA analyses a fuzzy version of Z where $z_{ij} \in [0,1]$ with the condition $\Sigma_{j \in J_v} z_{ij} = 1$, J_v being the set of modalities of the v^{th} attribute (variable). In statistics, one modality of a variable is one possible level of this variable. In

classical crisp coding, continuous variables are divided into several modalities (whose number depends on the distribution of the variable). For example, the variable “age” of a population can be split into 3 modalities “age ≤ 30”, “30 < age < 60” and “age ≥ 60”.

As depicted in Figure 2, the fuzzy space-time windowing divides the time domain of each variable into N_T overlapping windows. The membership value of the q^{th} time sample t_q to the time window T_j is denoted $\mu_{T_j}(t_q)$ and falls between [0,1]. The membership values of each data point in the N_T time windows meet the condition $\sum_{j=1}^{N_T} \mu_{T_j}(t_q) = 1$. The amplitudes of each variable (referred to as spatial domain) are coded in a similar manner with N_A spatial fuzzy windows verifying the same properties.

From the membership values in the time and the spatial (amplitude) domains, the membership value of the space-time window $W_{i,j}^n$, for a given time-series (signal) TS_k and the variable V_n, is denned as:

$$\mu_{W_{i,j}^n}(TS_k) = \frac{\sum_{q=1}^{Q} \mu_{T_j}(t_q) \cdot \mu_{A_{i,n}}(V_n(t_q))}{\sum_{q=1}^{Q} \mu_{T_j}(t_q)}$$

where $V_n(t_q)$ is the value taken by the n^{th} variable at time unit t_q, $\mu_{T_j}(t_q)$ is the membership value of the time window T_j for the time unit t_q, $\mu_{A_{i,n}}(V_n(t_q))$ is the membership value of the i^{th} space window $A_{i,n}$ for the $V_n(t_q)$ value, and Q is the number of time units in TS_k. With this definition, $\mu_{W_{i,j}^n}(TS_k)$ is the weighted average of the space membership values with the time membership values as weights and verifies $\sum_{i=1}^{N_A} \mu_{W_{i,j}^n}(TS_k) = 1$. This property is required to maintain the statistical context and to allow $\mu_{W_{i,j}^n}(TS_k)$ to be interpreted as the frequency of appearance of the signal in the space-time window $W_{i,j}^n$.

Multiple correspondence analysis (MCA)

MCA is of great interest to explore the data recorded by ID, as it handles both quantitative and qualitative data and captures nonlinear relationships between variables. It deals with a two-way cross-table with the observations (also called statistical individuals) as rows and the variables (or attributes) characterizing these observations as columns in the table. In MCA, rows and columns of the cross-table play a symmetric role and can be represented on the same plot. Another advantage of this method is the possibility of displaying supplementary variables and individuals jointly with the variables and individuals of analysis. They are not involved in the MCA but their projection on the factorial plane: i) refines and enriches the interpretation of the MCA factorial axes by relating them to meaningful variables (e.g. age, sex, etc.) and ii) enables the characterization of supplementary individuals according to their location with respect to individuals of analysis. Consequently in MCA, data acquired from other patients' ID can be represented on the factorial plane jointly with the reference population: it is possible to study their evolution with respect to the evolution of the patients of analysis.

Being in a follow-up frame, the temporal information contained in the recorded data is primordial and has to be taken into account. Inherently MCA does not exploit time, but the temporal dimension can be introduced artificially. A simple way consists in representing each time sample (or time window) of a time-series by one statistical individual (a row of the cross-table of analysis) and in applying MCA to the resulting table [12]. Data can then be organized in a table such as Table II, where a time-series is represented by as many rows as it has time samples or time windows. This method is simple and leads to a rather easy interpretation of the results. However, the temporal dimension is not explicitly exploited by the subjacent model of data representation (e.g. the same factorial axes would be obtained by introducing the lines on the analysis table in any particular order). Details and examples on the computation and interpretation of MCA can be found in [13]. With this convention, each time-series (i.e., in this study, each three-month length period of a given patient) is represented by a trajectory onto the factorial plane.

One of our objectives is to cluster the patients according to the evolution of their trajectories in the factorial plane defined by MCA. To facilitate this clustering, it is possible to smooth the trajectories in the factorial plane by applying a weighted and smoothed temporal average on the table Z analyzed by MCA. This method is named smoothed multiple correspondence analysis (SMCA) and has been introduced by Benali and Escofier [14]. The final table of analysis is $S = P \cdot Z$, where $P = (p_{ij})_{(i,j) \in [1,L]^2}$ is a proximity matrix denning the weighted and smoothed temporal average and is such as $\sum_{j=1}^{L} p_{ij} = 1$.

Application of the fuzzy coding and multidimensional analysis on the recorded database

The first two steps of the proposed methodology, namely the fuzzy spatio-temporal coding and the SMCA, are applied to the time-series available in the recorded database. In this study, a fuzzy window set $T = \{T_1, \cdots, T_J, \cdots, T_{N_T}\}$, where each T_j is 7-day long, is considered. The length of the fuzzy time windows has been chosen by considering the patients' behaviors, being quite similar from one week to another. Each trajectory, representing a three-month length period (i.e. around 13 7-day long), links then around 13 points.

For each variable V_n , $N_A = 3$ modalities are considered: "Low" corresponding to the spatial fuzzy window $A_{1,n} = [-\infty, \text{median}(V_n)]$, " Medium" corresponding to $A_{2,n} = [\text{prctile}(V_n, 2.5), \text{prctile}(V_n, 97.5)]$ and "High" corresponding to $A_{3,n} = [\text{median}(V_n), +\infty]$, where $\text{prctile}(V_n, p)$ is the p^{th} percentile of the variable V_n . The 3 spatial fuzzy windows are denoted with the suffixes "−H" for the higher level (modality), "−M" for the medium level and "−L" for the lower level.

The elements of the proximity matrix P are defined as:

$$p_{ij} = p_{ji} = \begin{cases} 0.2 & \text{if } j = i+1 \\ 0.1 & \text{if } j = i+2 \\ 0.05 & \text{if } j = i+3 \\ 1 - \sum_{k=1,\, k \neq j}^{L} p_{ik} & \text{if } j = i \\ 0 & \text{otherwise} \end{cases}$$

performing a temporal average of the statistical individuals and respecting the condition $\sum_{j=1}^{L} p_{ij} = 1$.

The protocol provided 58 records for 41 patients. SMCA is thus applied to 793 statistical individuals related to these 41 patients, i.e. to an array of 793 rows and 39 columns (13 variables with $N_A = 3$ spatial fuzzy windows).

The 1 st and 2 nd factorial axes represent respectively 68.0% and 15.7% of the total variance, showing that the majority (83.7%) of the information is contained in the first factorial plane. So a great part of the variance of the initial data can be represented by only two factorial axes when 13 variables were initially considered. These first two factors define synthetic indices for patient follow-up but their interpretation, from physiological and functional points of view, has to be performed. Consequently, this study will focus on the clustering of trajectories on this factorial plane.

The projections of the variables and individuals of analysis on the first factorial plane of the SMCA are provided in Figures 3 and 4, respectively. The first axis is mainly defined by the lower (−L) and higher (−H) modalities of Imp_E , Dur_E , NbCardCyc_E , NbBreaths_E , ACC_E and ImpRate, which reflect the time spent in exercise and the intensity of the efforts made by the patient. Consequently, in Figure 4, the more the individuals are located to the right of the plane, the lower is the time spent in exercise and the less important are the efforts they make. The second axis is mainly defined by the medium (−M) and extreme (−H, −L) modalities of ImpMinVent_E , ImpOverAcc, Imp_R and HeartRate_E , which define the ventilation activity in terms of amplitude, frequency and flow rate, especially in rest. This axis can be interpreted as an "axis of cardiovascular efficiency". In Figure 4 the more the individuals are located to the lower part of the plane, the less important is their ventilation in rest (i.e. their cardiovascular system is more "efficient"), independently of the daily activity duration and intensity (i.e. of the position along the first axis).

As it can be seen on Figure 4, trajectories present different locations and evolutions on the factorial plane, with a high overlapping.

Clustering trajectories on the factorial plane

The aim of the present study is to cluster patients according to their clinical state during the follow-up period, which, in terms of methodology, corresponds to the clustering of trajectories on the factorial plane according to their location and evolution. Considering the characteristics of the trajectories, the clustering methodology has to address the following issues:

- Unsupervised: no a priori knowledge on the clusters is required.

- Similarity measure: relevance in comparing trajectories on the factorial plane having different numbers of points and possibly subjects to nonlinear spatio-temporal deformations.

- Location and evolution: both locations and evolutions of the trajectories on the factorial plane have to be taken into account, as they are both informative.

Each of the previous three points is described in the following sections.

An appropriate clustering algorithm

Among the unsupervised clustering algorithms, k-means and agglomerative hierarchical clustering (AHC) are two classical methods.

The k-means algorithm implies the definition of centroids for each cluster, which in the present study is not trivial as the considered objects are trajectories possibly constituted of different numbers of points within the same cluster. The k-means method seems then not relevant for this particular problem.

Agglomerative hierarchical clustering only requires the definition of the dissimilarity matrix between the objects (i.e. the trajectories) and of the aggregation link. It is chosen as the clustering technique is the present study with the complete link as an aggregation link.

A relevant similarity measure between trajectories

The main difficulty is the definition of a similarity measure corresponding to the trajectories on factorial planes. As mentioned above, these trajectories can potentially be subject to deformations and are constituted of different numbers of points. Consequently, the Euclidian distance is not suitable. Among the measures of dissimilarity, dynamic time warping (DTW) and longest common subsequence (LCSS) both allow stretching in time and comparing time-series of different lengths. DTW has been widely used as a measure of dissimilarity in time-series clustering, indexing and retrieving [15], in speech or handwriting recognition [16]. LCSS has also been studied as a similarity measure for heterogeneous multivariate time-series or for multidimensional trajectories [17]. DTW presents the advantage over LCSS to be non-parametric and seems thus more appropriate for unsupervised clustering.

The computation of the DTW vector can be adapted in two dimensions to deal with two trajectories instead of two time-series. Given one trajectory $Traj_k = \{(Traj_k(x_i), Traj_k(y_i))\}_{i \in [1, m]}$ constituted of m data points and one trajectory $Traj_l = \{(Traj_l(x_j), Traj_l(y_j))\}_{j \in [1, n]}$ constituted of n data points, the DTW vector is denoted $DTW_{k,l}(i, j)$. It is defined according to the equation:

$$DTW_{k,l}(i, j) = D_{k,l}(i, j) + \min [DTW_{k,l}(i-1, j-1), DTW_{k,l}(i, j-1), DTW_{k,l}(i-1, j)]$$

where $D_{k,l}(i, j)$ is the Euclidian distance between the coordinates $(Traj_k(x_i), Traj_k(y_i))$ and $(Traj_l(x_j), Traj_l(y_j))$. Then, the distance DTW between the two trajectories is $DTW_{k,l} = DTW_{k,l}(m, n)$.

A similarity measure considering both locations and evolutions of the trajectories in the factorial plane

The positions of the modalities on the factorial plane in Figure 3 indicate that two trajectories with similar shapes but located at different positions on the factorial plane are related to different modalities. This observation underlines the fact that both their location and their evolution, i.e. both their coordinates and their derivatives, in the factorial plane are informative to cluster similar trajectories and to compute the dissimilarity matrix. The DTW can then be used with the following modifications.

Given the k^{th} trajectory $Traj_k$, its derivative is:

$$dTraj_k = \{(Traj_k(x_i) - Traj_k(x_{i-1}), Traj_k(y_i) - Traj_k(y_{i-1}))\}_{i \in [1, m]}.$$

The DTW vector $dDTW_{k,l}(i, j)$ between the two derivatives $dTraj_k$ and $dTraj_l$ is defined according to equation 3 and the distance DTW between the two derivatives is $dDTW_{k,l} = dDTW_{k,l}(m,n)$.

Thus in this study, given $\overline{Deuclid_{k,l}} = euclid(mean(Traj_k), mean(Traj_l))$ the Euclidian distance between the means of the coordinates of $Traj_k$ and $Traj_l$ in the factorial plane, the dissimilarity measure between the two trajectories $Traj_k$ and $Traj_l$ is defined as:

$$DM_{k,l} = dDTW_{k,l} + \overline{Deuclid}_{k,l},$$

where $dDTW_{k,l}$ takes into account the derivatives of the trajectories and $\overline{Deuclid}_{k,l}$ their locations.

The $N \times N$ dissimilarity matrix used for the AHC is then $DM = \{DM_{k,l}\}_{k,l \in [1,N]^2}$, N being the number of trajectories to be clustered. After computation of the AHC with complete linkage and the dissimilarity matrix DM, a dendrogram is obtained.

Cluster validity criterion

In AHC, in order to choose the threshold of cut in the dendrogram (i.e. the number of clusters) and to verify the validity of the clustering, cluster validity indices are used (for a review, see [18]). As no information on the data is available (unsupervised clustering), only internal validation indices are suitable. They are based on computing the properties of the resulting clusters, as the intra- and inter-cluster distances. The internal validation index used in the study is the mean silhouette value, denoted S and described in [19]. S is in the interval [−1, +1], where values close to −1 indicate a wrong clustering and values close to +1 indicate a correct clustering.

Results obtained with the clustering

Before applying the proposed clustering method, namely agglomerative hierarchical clustering with complete linkage and the dissimilarity matrix based on dynamic time warping, it is tested on two simulated datasets, described in the following sections.

Tests on simulated datasets

The aim of the first test is to explore the capability of the proposed similarity measure to take into account both locations and evolutions of the trajectories on the factorial plane. The second test interests in the complete methodology described in Figure 1, from the fuzzy space-time coding to the clustering.

1. Test on the similarity measure

From two trajectories selected among the trajectories obtained with the real dataset (cf. Figure 4), three prototypes of trajectories are computed, the third being obtained by inverting the time samples of the first one.

Nine trajectories are simulated from the first prototype at different locations in the factorial plane, 6 from the second prototype and 12 from the third prototype. For each trajectory, a white noise is added to obtain slightly different trajectories for the same prototype. The simulated dataset of 27 artificial trajectories is represented in Figure 5. Figure 5 shows that the simulation reproduces the characteristics of the real dataset, namely the overlapping of several trajectories of different shapes at different locations on the factorial plane.

The clustering method proposed above is applied to the simulated dataset with three dissimilarity measures based on the DTW which is alternately computed on: i) the coordinates of the trajectories, ii) the derivatives of the trajectories, iii) the derivatives with addition of the Euclidian distance between the means of the coordinates as proposed in equation 5.

Figure 6 illustrates the resulting dendrogram for each dissimilarity measure and provides the mean silhouette value against the number of clusters. A dendrogram is a tree-like plot where each step of hierarchical clustering is represented as a fusion of two branches of the tree into a single one. The branches represent clusters obtained on each step of hierarchical clustering. This representation eases the choice of the number of clusters. The dendrogram is cut at the threshold (horizontal dashed line) producing the number of clusters corresponding to the maximum mean silhouette value.

For each of the three measures, the mean silhouette value S presents one global maximum, indicating a value for the number of clusters K to be chosen, and leads to a different clustering. For the DTW on coordinates, $\max_K S$ is obtained for $K = 3$ and it tends to group the trajectories close in the sense of their location on the factorial plane. For the DTW on derivatives, $\max_K S$ is obtained for $K = 3$, but it regroups trajectories only according to their evolution. For the DTW on derivatives with the addition of the Euclidian distance between coordinates, $\max_K S$ is obtained for $K = 9$ and each obtained cluster is composed of trajectories with similar evolutions and close locations. Despite the overlapping of several trajectories with different evolutions at the same location, the proposed dissimilarity measure is able to group similar trajectories. Moreover, it is suitable to distinguish between shapes (independent of any notion of time) and evolutions (with start- and end-points), as trajectories created with the first prototype and with the third prototype are assigned to two different clusters, even if they have similar shapes.

Test of the complete method

In this section, the noise robustness of the proposed methodology, described in Figure 1, is tested. Among the trajectories obtained with the real database (cf. Figure 4), seven trajectories with different evolutions and locations are selected. The corresponding time-series (i.e. the 13 variables of analysis) are retrieved. For each selected trajectory, a white noise is added independently to each of its 13 variables with a given signal-to-noise ratio (SNR). The original time-series and 10 realizations of the noisy time-series are coded by fuzzy space-time coding and constitute the individuals of analysis for the SMCA. The 77 resulting trajectories are then clustered by AHC with the DTW on derivatives with the addition of the Euclidian distance between coordinates as a dissimilarity measure (cf. equation 5). Figure 7 presents the mean silhouette value against the number of clusters for SNR = 3 dB, and the clusters obtained for the maximum value of the silhouette value being K = 7. Despite the noise, the correct number of clusters is determined by the internal validation indices, the trajectories being grouped according to their evolution and location.

It appears that above SNR = 0 dB the clustering is correct and is not disturbed by the noise added to the time-series and that under SNR = 0 dB, S is unable to provide the correct value of K. The method seems then robust to noise on the variables of analysis, which can be explained by two of its steps: the time fuzzy coding and the smoothing performed during the SMCA. The mean silhouette value increases when the length of the time fuzzy windows increases and is higher with the SMCA than with the MCA. The averaging performed by both the fuzzy coding and the SMCA smoothes the time-series and the trajectories, respectively, improving the performance of the clustering.

The previous two tests, performed on datasets close to the real database, have proven the validity of the proposed approach in terms of clustering methodology and dissimilarity measure relevant for the processed trajectories, and of noise robustness.

Performance of the clustering on the recorded database

In this section, the recorded database, constituted of 58 trajectories on the first factorial plane, is clustered with the proposed approach. Figure 8 provides the mean silhouette value S against the number of clusters K. The number of clusters is chosen at K = 10, after the maximum value of S. The resulting clusters are provided in Figure 8, where for each cluster individuals of analysis are in gray and individuals of the given cluster are in black. One can notice that trajectories within a given cluster have visually similar evolutions and close locations.

To evaluate the methodology, the resulting clusters have to be compared with the appreciation of the physicians on the evolution of the patients during each of the three-month length periods. For the present protocol, information is available on the global state of each patient, updated at the end of each 3-month period of the follow-up, and each “adverse event” is reported (date and type). In the present study, the records with adverse events other than cardiac events are discarded, as a non cardiac event like a fall or a bronchitis can have very different effects on patients with RHF. As no everyday report of patients’ clinical state is available, the difficulty resides in defining the actual evolution of a given patient: has a patient undergoing a cardiac adverse event at the beginning of his/her three-month length period and recovering rapidly after hospitalization a favorable or an unfavorable evolution? In this context, indices like the specificity or the sensitivity are difficult to compute. Consequently, the present study can only focus on the relation between the different areas of the factorial plane defined by SMCA and the health of the patients as reported at the end of each follow-up period. More data would be necessary to study the evolutions of patients on this factorial plane.

According to the analysis performed on the position of the individuals of analysis relatively to the position of the modalities of analysis (cf. Figures 3 and 4), the evolution of a patient (during one of his/her three-month length period) is a priori i) favorable if the corresponding trajectory is located on the left of the plane or evolves from the right to the left of the plane, and ii) unfavorable if the corresponding trajectory is located on the right part of the plane or evolves from the left to the right of the plane. The study of each cluster is necessary to confirm this division of the factorial plane into several areas with different meanings in terms of patient's health.

On the left of the plane, clusters 2, 4 and 5 contain 13 trajectories in total. The medical reports indicate that all the corresponding patients were, during the given period, in a favorable state. On the right of the plane, clusters 8 and 9 contain 5 trajectories in total, associated with patients all undergoing a health deterioration during the given period. These five clusters enable the definition of two distinct areas in the factorial plane, the left one associated with a favorable state and the right one with a health deterioration.

The other five clusters are interesting as their trajectories show transitions from one area to the other. Cluster 1 comprises 27 trajectories, with small evolutions on the upper half-plane, overlapping both right and left half-planes. During the given period, patients whose trajectories belong to cluster 1 were all in good health, except one patient. According to the medical records on this patient, he was tired during the period of interest (3 months post-op) and died 1 month after the end of this period, the date of his/her adverse event is not reported. Figure 9 shows that the corresponding trajectory (solid line) is evolving from the upper-plane to the right of the plane, with a loop around the 7 th post-op week. The trajectory in cluster 3 is very specific, evolving bottom-to-top firstly and right-to-left secondly. This patient underwent an adverse event (heart failure) 28 days after the beginning of the period of interest, corresponding to the changing of direction in his/her trajectory, and

recovered rapidly which explains why the trajectory evolves from right to left. Cluster 6 groups 6 trajectories evolving in the same direction (bottom-right to top), associated with patients all, except one, presenting a favorable state during the given period. This patient (bold line in Figure 9) underwent a cardiac adverse event around 2 months after the beginning of the period of interest. He was hospitalized for 1 week and recovered rapidly, as can be seen from his/her trajectory finally evolving right to left. In Figure 9, the arrow corresponds to the reported date of the adverse event, and one can notice that a changing in the direction of the trajectory has occurred around 5 weeks before the adverse event. Cluster 7 comprises 2 trajectories evolving from top-right to top-left of the plane and corresponding to one patient with a favorable evolution and one patient with an unfavorable evolution. This patient underwent a cardiac adverse event around 2 weeks before the end of the period of interest, as can be seen with the abrupt U-turn in his/her trajectory (dashed line in Figure 9). Cluster 10 groups 4 trajectories with two different evolutions, two trajectories have very small dynamics, associated with patients in good health, and the two others have large evolutions from the top to the right of the plane, associated with patients undergoing adverse events during the given period.

The clustering performed on the real dataset provides 10 clusters, grouping trajectories with similar evolutions and close locations, as expected. The obtained clusters seem consistent with the medical records of the patients, but no quantitative evaluation is available. Two areas in the factorial plane have been identified, the bottom-right quarter-plane is related with a health degradation and the bottom-left one with a stable clinical state. The other two areas, although being distinguished by the clustering algorithm, are not so easily identifiable.

Discussion and conclusion

This study was designed to show the informative potential of acceleration and impedance data recorded in implantable devices and to evaluate the appropriateness of a multiple correspondence analysis in this particular context. A clinical protocol has been designed to provide data on patients that suffer from heart failure.

MCA has been chosen as it is a multivariate method that exhibits linear and non-linear relationships between variables. However, MCA demands to transform continuous variables into nominal ones, i.e. to code amplitude and time domains of the variables by means of crisp or fuzzy modalities. MCA does not explicitly exploit time, but the temporal dimension is introduced in this study with a basic solution that consists in representing each time sample (or time window) of a time-series by one statistical individual (a row of the table of analysis) and to perform MCA. Statistical individuals are represented by trajectories onto the factorial plane and their temporal evolution can then be exploited. Consequently, the proposed method can be useful for graphically follow the evolution of a given patient's state by means of a synthetic representation that takes into account the most pertinent information in the data.

In the present study, it has been possible to evaluate and discuss the synthetic indices provided by the first two factorial axes of the MCA and to explain them according to functional and physiological points of view. The 58 records provided by the clinical protocol have been represented by trajectories on the first factorial plane of the MCA. The definition of an appropriate similarity measure has been discussed and has enabled the clustering of the trajectories into groups of trajectories with similar locations and evolutions on the factorial plane. The proposed distance has been validated on simulated datasets and enables the clustering of trajectories in the factorial plane according to their location and/or their evolution depending on the problem to be solved. Experiments on simulated datasets regarding the noise robustness also show the relevance of the association of fuzzy spatio-temporal coding and smoothed MCA. The proposed method is robust to white noise with SNR as low as 3 dB.

A database constituted of clinical observations from 41 patients has also been analyzed by using a data mining approach so as to characterize the data. Two areas in the factorial plane, corresponding to two large groups of patients, have been identified, the bottom-right quarter-plane being related with a health degradation and the bottom-left one with a stable clinical state. Most of the trajectories projected on this first factorial plane were correctly clustered according to their location and shape. Discussed individually, these clusters were efficient in grouping trajectories corresponding to similar patients' clinical state. The present study has shown that patients undergoing an adverse event often present trajectories with abrupt variations. A detection of such phenomena would permit to identify patients with a critical evolution.

In the future, additional data would enable the identification and detection of typical evolutions related to health deterioration. Rules may be extracted from the location of modalities on the factorial plane, and enable the definition of thresholds on time-series to generate alarms associated with adverse events or health deterioration. These alarms could be sent from the pacemaker, via a data communication device (such as a mobile phone, PDA, etc.), to a telemonitoring center. These results are encouraging and may be useful for the definition of new selection criteria of candidate patients for CRT. Finally, the proposed methodology can be generalized to other monitoring problems.

References:

- 1. Cazeau S , Alonso C , Jauvert G , Lazarus A , Ritter P Cardiac resynchronization therapy. Europace. 5: (Suppl 1) S42- S48 2004;
- 2. Cazeau S , Leclercq C Effects of multisite biventricular pacing in patients with heart failure and intraventricular conduction delay. N Engl J Med. 344: 873- 880 2001;
- 3. Leclercq C , Kass DA Retiming the failing heart: principles and current clinical status of cardiac resynchronization. J Am Coll Cardiol. 39: 194- 201 2002;

- 4. Germany R , Murray C Use of device diagnostics in the outpatient management of heart failure. Am J Cardiol. 99: 11G- 16G 2007;
- 5. Roux E , Hernandez A , Graindorge L , Carrault G , Mabo P Multivariate analysis of follow-up physiological data recorded by cardiac implantable devices. Computers in Cardiology. 33: 2006; 765- 768
- 6. Guéguin M , Roux E Clustering follow-up time-series recorded by cardiac implantable devices. 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society 2007; 3848- 3851
- 7. Antunes C , Oliveira A Temporal data mining: an overview. Workshop on Temporal Data Mining at the 7th International Conference on Knowledge Discovery and Data Mining (KDD) 2001;
- 8. Srivatsan L , Sastry PS A survey of temporal data mining. Sadhana. 31: (2) 173- 198 2006;
- 9. Simon R , Ni Q Comparison of impedance minute ventilation and direct measured minute ventilation in a rate adaptative pacemaker. PACE. 26: 2127- 2133 2003;
- 10. Bonnet J , Geroux L Active implantable medical device having a control function responsive to at least one physiological parameter. Patent US 5 722 996 Al 1998;
- 11. Bonnet J Implantable active medical device enslaved to at least one physiological parameter. Patent US 6 336 048 Bl 2002;
- 12. Loslever P , Bouilland S Marriage of fuzzy sets and multiple correspondence analysis: Examples with subjective interval data and biomedical signals. Fuzzy Sets and Systems. 107: 255- 275 1999;
- 13. Abdi H , Valentin D Editor: NS Multiple correspondence analysis. Encyclopedia of Measurement and Statistics. Thousand Oaks (CA) Sage; 2007; 651- 657
- 14. Benali H , Escofier B Editor: Coppi R , Bolasco S Smooth factorial analysis and factorial analysis of local differences. Multiway Data Analysis. North-Holland 1989; 327- 339
- 15. Liao TW Clustering of time series data - a survey. Pattern Recognition. 38: 1857- 1874 2005;
- 16. Myers C , Rabiner L A comparative study of several dynamic time-warping algorithms for connected word recognition. The Bell System TechnicalJournal. 60: (7) 1389- 1409 1981;
- 17. Vlachos M , Kollios G , Gunopulos D Discovering similar multidimensional trajectories. 18th International Conference on Data Engineering (ICDE) San Jose, CA, USA 2002; 673- 684
- 18. Brun M , Sima C Model-based evaluation of clustering validation measures. Pattern Recognition. 40: 807- 824 2007;
- 19. Rousseeuw P Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics. 20: 53- 65 1987;

Fig. 1

Overview of the proposed methodology. The analysis is performed on the reference population and leads to the determination of clusters of similar evolutions – according to an appropriately chosen dissimilarity measure – on the factorial plane defined by the smoothed multiple correspondence analysis (SMCA). A practical application of this methodology would be the projection of subsequent follow-up data as supplementary individuals on the factorial axes defined during the analysis and the assignation of the obtained trajectories to the “closest” cluster, in the sense of the chosen distance.

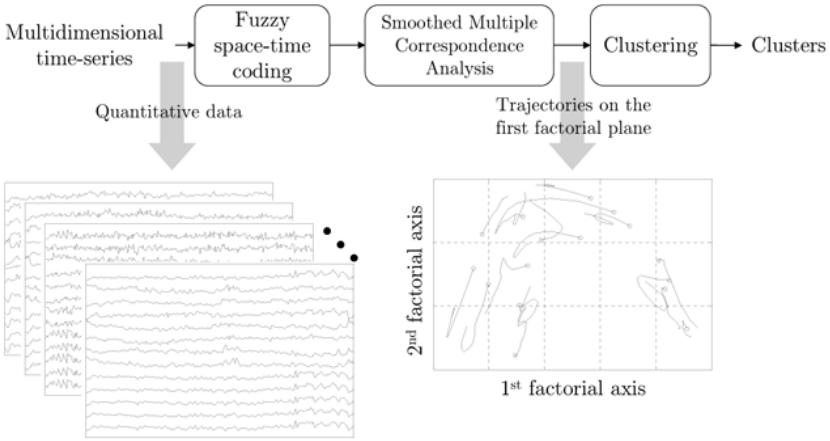

Fig. 2

Temporal and spatial (amplitude) fuzzy coding of a continuous signal.

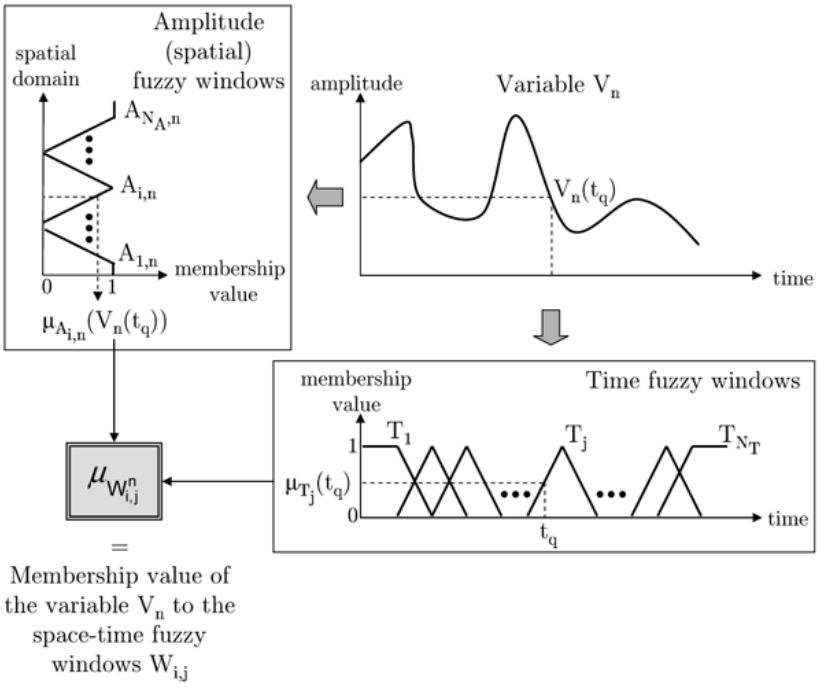

Fig. 3

Variables of analysis represented on the first plane of the Smoothed Multiple Correspondence Analysis (SMCA). For each variable, only the first (Low, −L) and the third (High, −H) levels are labelled, unlabelled squares correspond to the second levels (Medium, −M).

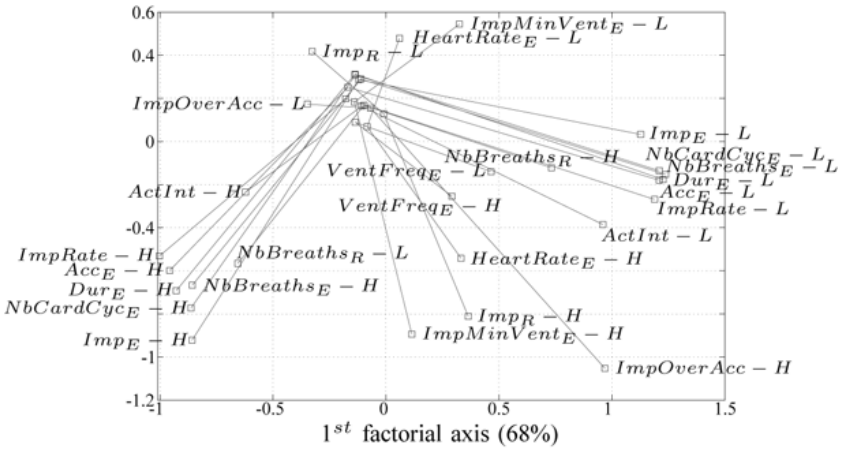

Fig. 4

Individuals of analysis represented on the first plane of the Smoothed Multiple Correspondence Analysis (SMCA). Each record of each patient (corresponding a three-month period) is represented by one trajectory. The first point in time for each trajectory is marked up by a circle.

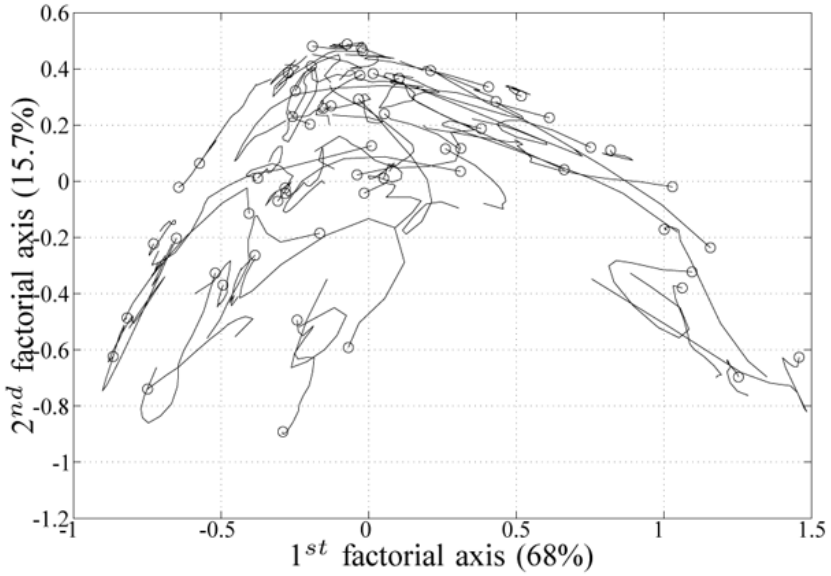

Fig. 5

Simulated dataset of 27 trajectories obtained from three real trajectories and represented on the first factorial plane. The first point in time for each trajectory is marked up by a circle.

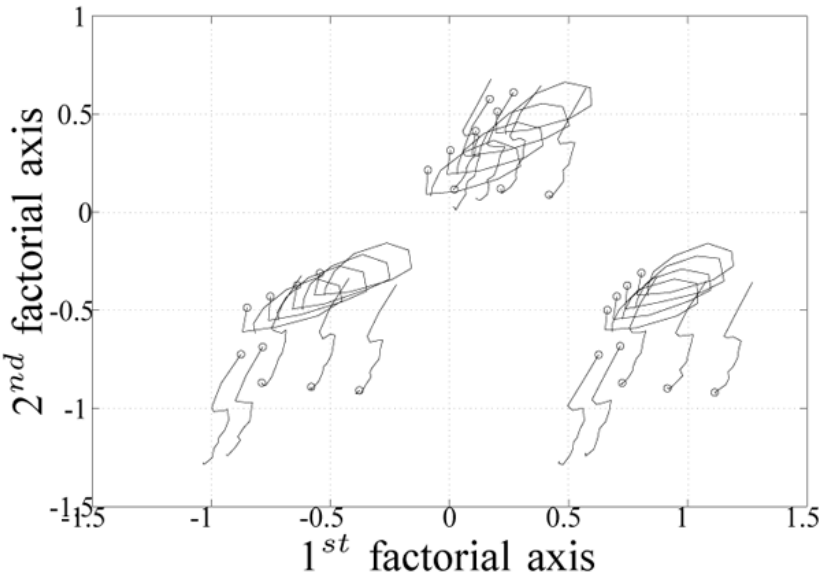

Fig. 6

Dendrograms and mean silhouette value vs. the number of clusters, for three dissimilarity measures based on the DTW. The clustering is applied to the simulated dataset of 27 trajectories. The dendrogram is cut at the threshold whose value is obtained after the maximum of the mean silhouette.

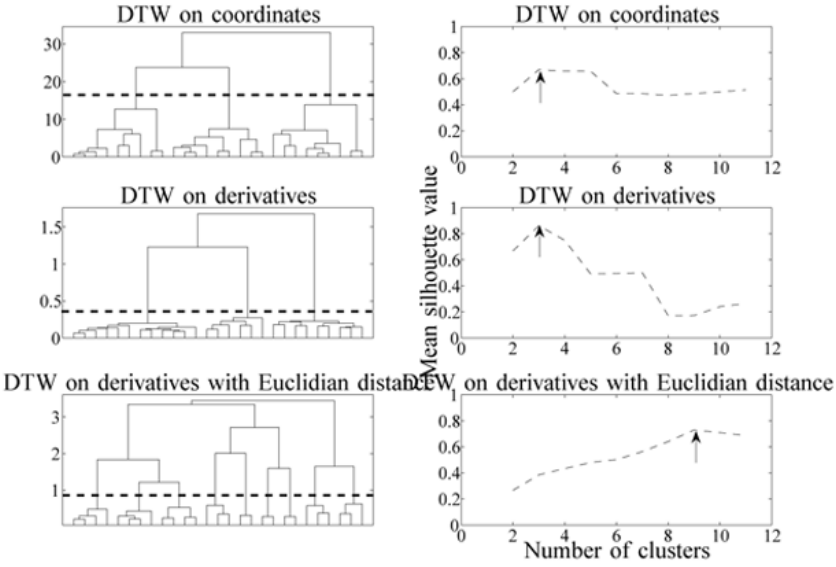

Fig. 7

Test of the complete method on 77 trajectories computed from 7 selected trajectories of the real dataset. For each selected trajectory, a white noise is added independently to each of its 13 variables with a given signal-to-noise ratio (SNR = 3 dB), and the resulting 77 trajectories are clustered. For each cluster, individuals of analysis (in gray) and individuals of the given cluster (in black) are represented in the first plane of the SMCA. The first point in time for each trajectory is marked up by a circle.

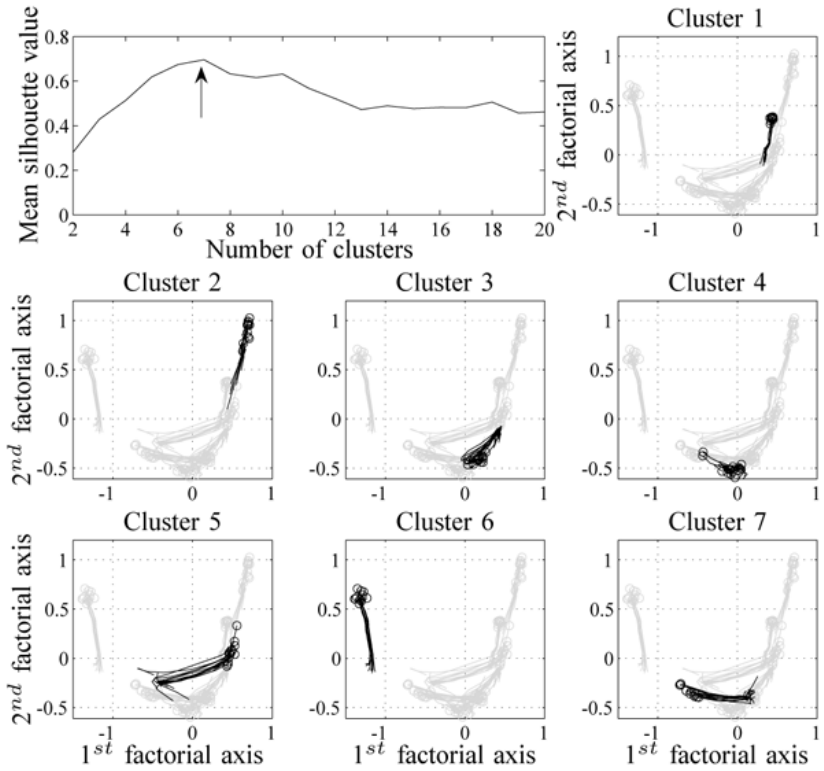

Fig. 8

Mean silhouette value S against the number of clusters K. Ten clusters are determined with the proposed methodology for the recorded database ($\max_K S$ for K = 10). For each cluster, individuals of analysis (in gray) and individuals of the given cluster (in black) are represented in the first plane of the SMCA. Each record of each patient (corresponding to a three-month period) is represented by one trajectory. The first point in time for each trajectory is marked up by a circle.

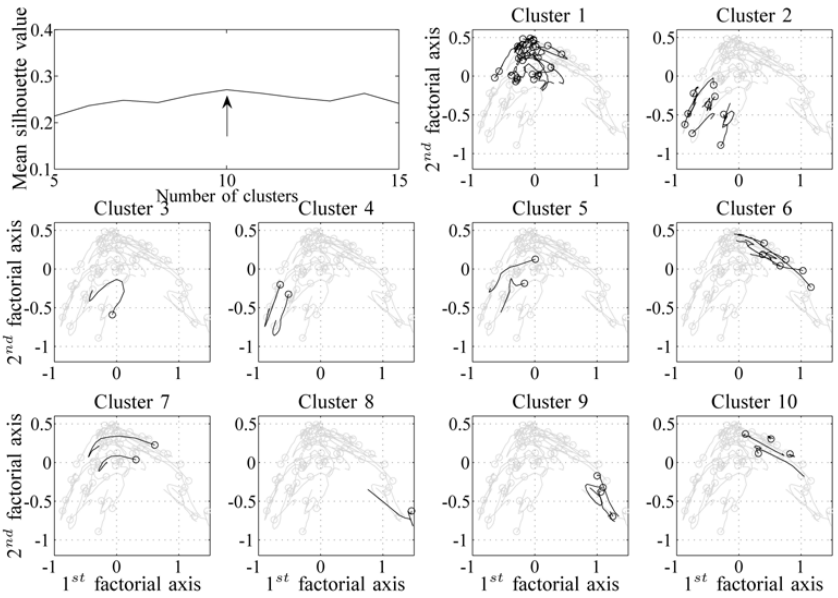

Fig. 9

Three specific trajectories of patients undergoing an adverse event during the period of interest. Solid line: trajectory in cluster 1, bold line: trajectory in cluster 6, dashed line: trajectory in cluster 7. Reported dates of adverse events are indicated by an arrow. The first point in time for each trajectory is marked up by a circle.

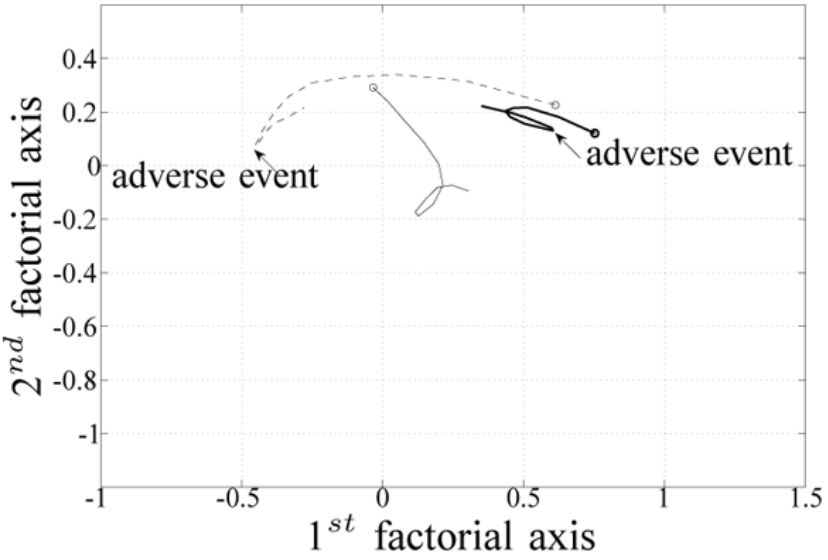

TABLE I

List of 7 physiological variables recorded in the cardiac implantable devices and 6 variables computed from these recorded variables.

Description	Names	Units
Total duration within the activity level	Dur_E 1	s
Cumulative values of acceleration	ACC_E	$M{\cdot}s^{-2}(g)$
Cumulative values of impedance	Imp E Imp R	millivolts (mV)
Cumulative number of ventilation cycles	NbBreaths E NbBreaths R	NbVC
Cumulative number of cardiac cycles	$NbCardCyc_E$	NbCC
“Mean”2 activity intensity	ActInt	$g{\cdot}s^{-1}$
Imp_E over ACC_E	ImpOverAcc	$mV{\cdot}g^{-1}$
“Mean” heart rate	$HeartRate_E$	Beats per minute (bpm)
“Mean” impedance minute ventilation	$ImpMinVent_E$	$mV{\cdot}min^{-1}$
“Mean” ventilation frequency	$VentFreq_E$	$NbVC{\cdot}min^{-1}$
Imp_E over Imp_R	ImpRate	none

1 Subscripts E and R are for Exercise and Rest, respectively.

2 The duration of each Exercise and Rest period that occurs within 24 hours is unknown. Only the cumulative duration is known. Consequently, this “mean” is not the average of the variable value hours, excepted if all the periods are of the same duration.

TABLE II

Construction of the table Z for multiple correspondence analysis (MCA) applied to fuzzy coded data and implicitly exploiting the temporal dimension.

Variables	…	$\mathbf{V_n}$	…
Fuzzy space windows	…	$A_{i\,,\,n}$	…
TS_{1} at fuzzy time window 1	…	$\mu_{W_i}(TS_1^n)$	…
TS_{1} at fuzzy time window 2	…	$\mu_{W_i}(TS_1^n)$	…
…	…	…	…
TS_k at fuzzy time window j	…	$\mu_{W_i}(TS_k^n)$	…
…	…	…	…

$A_{i,n}$ is the fuzzy space window corresponding to the i^{th} modality of the n^{th} variable, TS_k is the k^{th} time-series and (TS_k) is the membership value of the n^{th} variable to the fuzzy time-space window f time-series.