Blood Pressure Pattern Identification in AKI Patients using Linguistic Summaries

Ana Luísa Martins
ana.luisa.martins@tecnico.ulisboa.pt

Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal

June 2018

Abstract

The high development of Information Systems that has been witnessed in the past few years has led to a new problem of data analysis and processing. In the past few years, the access and production of information have considerably eased. Hardier, though, is it to filter such information so that relevant conclusions can be withdrawn. Linguistic summaries, a data mining and knowledge discovery approach to extract patterns and sum up large volume of data into simple sentences, are one interesting solution to overcome this challenge. This dissertation aims to study the evolution of two time-varying variables, Mean Arterial Blood Pressure (MAP) and Urine Output (UO), from patients who stayed more than 24 hours at an intensive care unite (ICU) resorting to linguistic summarization. For a better and comparative understanding of their evolutions, patients were initially split into two different groups: those who had a critical drop in UO before a certain MAP level and those who had it above the same value. In order to treat the problem different summarization methods were applied: categorical and multivariable linguistic summaries, linguistic summaries obtained from clusters of datasets, linguistic summaries concerning demographic variables and temporal linguistic summaries.
Keywords: Linguistic Summary, MAP, Urine Output, Health Care, Similarity Analysis

1. Introduction

The high development of Information Systems that has been witnessed in the past few years has led to a new problem of data analysis and processing. In the past few years, the access and production of information has considerably eased. Hardier, though, is it to filter such information so that relevant conclusions can be withdrawn.

This problem reaches dramatic levels when considering medical databases. For each patient, there is a big number of variables that may be described through time. Even though the amount of information is not necessarily higher than for other systems, the inherent complexity of each variable combined with huge databases with several individuals, makes its concise analysis almost a herculean task.

In hospitals, huge amounts of data are recorded concerning the diagnosis and treatments of patients in what is typically called event-logs [6]. Event logs show occurrences of events at specific moments in time, where each event refers to a specific process and case. Usually event logs are large text files with some standardized format, sorted according to the time they were stored [9]. This information, however, may be of difficult read and interpretation and, sometimes, not easy to translate into a visual environment. Thus, a need for a methodology that suitably represents this data in a understandable and easy way prevails.

Linguistic summarization is a data mining and knowledge discovery approach to extract patterns and sum up large volume of data into simple sentences [1].

A linguistic summary of data is then a concise description of this same data in terms of natural human language. Its main aim is to capture the essence of the behaviour of a large data set and translate it into small and human comprehensible statements.

Distinguished and several times complementary to data visualization, which is used to increase the understanding of data by emphasizing or discovering visual cues, linguistic summarization aims to show the most characteristic and relevant aspects of data by natural language and help, this way, humans discovering relations in data that would have, otherwise, remained unknown or difficult to notice. However and despite its shown advantages in translating complex data events into simple sentences, linguistic summarization is still sparsely present in field literature [9].

At the intensive care unit (ICU), established pro-

tocols recommend that patients with sepsis should be treated with vasopressors or fluids as soon as their mean arterial blood pressure (MAP) drops below 65 mmHg. However, older patients with atherosclerosis or previous hypertension, for example, may benefit from being treated at a higher MAP than younger patients without any cardiovascular conditions. There are currently no guidelines indicating a target MAP for groups of patients with distinct characteristics.

A critical urine output (UO) entails an impending organ failure, advocating the initiation of treatment at a specific point in time - being it below or above the established protocol.

From a clinical perspective, it is very important that treatment starts before the onset of a critical UO, to avoid severe kidney complications. Previous studies have tried to establish a relationship between MAP and UO [7]. However, the relation is very complex and not possible to describe through linear systems.

In order to understand the relation between MAP and UO two groups are considered: group 1, the ones who had a critical event (UO dropping below 30 $ml\backslash h$) before the established protocol (MAP below 65 mmHg) or group 2, the ones after (MAP above 65 mmHg).

The idea is to understand the differences between these two groups through the study of linguistic summaries that characterize their time series and this way understanding which variables are affecting the most the occurrence of the critical event.

If the correlation between variables and the critical event is well defined, the vasopressor can always be administered before the occurrence of the critical event before the onset of a critical UO, avoiding serious kidney complications.

2. Linguistic Summaries

A linguistic summary of data is a concise description of this same data in terms of natural human language. Its main aim is to capture the essence of the behavior of a large data set and translate it into small and human comprehensible statements. The concept of linguistic summaries has been developed by Yager [12] and further developed by Kacprzyk and Yager [4] and Kacprzyk, Yager and Zadrozny [5].

Here the notation as proposed by Yager [12] is followed:

1. $Y = \{y_1, ..., y_n\}$ is a set of objects in a database, for instance a set of employees.
2. $A = \{A_1, ..., A_n\}$ is a set of attributes characterizing y_i from Y, e.g., *salary* and $A_j(y_i)$ is a value attribute A_j of y_i.

A linguistic summary of a data set consists of:

1. A summarizer P, i.e., an attribute with linguistic value (fuzzy predicate) defined on the domain of A_j (e.g. *high* for the attribute *salary*);
2. A quantity of agreement Q, i.e., a linguistic quantifier (e.g. *most*);
3. The Truth (validity), $\mathcal{T}$, of the summary, i.e. a number from the interval [0,1], assessing the truth (validity) of the summary as, e.g., "$\mathcal{T}$ (*most* of employees have *high* salaries) = 0.7";
4. A qualifier R (optional), i.e., another attribute with linguistic value (also fuzzy predicate) defined on the domain of A_k determining a fuzzy subset of Y (e.g. *young* for the attribute *age*).

Thus, the core of a linguistic summary is a linguistically quantified proposition in the sense of Zadeh which can be written in its simple form as:

$$Qy\text{'s are}P \tag{1}$$

e.g.: "*Most* employees have *high* salaries."
And in its extended form as:

$$QRy\text{'s are}P \tag{2}$$

e.g.: "*Most young* employees have *high* salaries."
The truth value ($\mathcal{T}$) of them can be calculated by using either original Zadehs calculus of linguistically quantified propositions (cf. [14]), or other interpretations of linguistic quantifiers, i.e., respectively:

$$\mathcal{T}\ (Qy\text{'s are } P) = \mu_Q\left(\frac{1}{n}\sum_{i=1}^{n}\mu_P(y_i)\right) \tag{3}$$

$$\mathcal{T}\ (QRy\text{'s are } P) = \mu_Q\left(\frac{\sum_{i=1}^{n}\mu_R(y_i)\cap\mu_P(y_i)}{\sum_{i=1}^{n}\mu_R(y_i)}\right) \tag{4}$$

The computation of a temporal linguistic summary is very similar to the previous ones, it will have a temporal expression, E_T, as an additional qualifier (R), as the temporal expression, similar to the qualifier, limits the universe of interest to those segments that only occur on the time axis described by a fuzzy set modelling the expression E_T. That is the proportion of trends that are P and happen in E_T to those that occur in E_T. Then the quantity of agreement of this proportion is computed.

Thus the truth value of a temporal protoform is computed by:

$$\mathcal{T}(E_T \text{ among all } y\text{'s}, Q \text{ are } P) = \mu_Q\left(\frac{\sum_{i=1}^{n}\mu_{E_T}(y_i)\cap\mu_P(y_i)}{\sum_{i=1}^{n}\mu_{E_T}(y_i)}\right) \tag{5}$$

where μ_{E_T} is the degree to which a trend occurs during the time span described by E_T.
Similarly we compute the truth of an extended temporal protoform as:

$$\mathcal{T}(E_T \text{ among all R}y\text{'s, } Q \text{ are } P) = \mu_Q\left(\frac{\sum_{i=1}^{n} \mu_{E_T}(y_i) \cap \mu_P(y_i) \cap \mu_R(y_i)}{\sum_{i=1}^{n} \mu_{E_T}(y_i) \cap \mu_R(y_i)}\right) \quad (6)$$

3. Time Series Segmentation and Linguistic Summarization

In many studies, while resorting to time series data, the aim is to look at the relation between certain data points at a certain time, instead of just considering their absolute value. The time series may be very long and take too much space, for which a more compact version is needed in which a time series data set is split into more compact subsets (i.e. time segments or trends). This procedure is called segmentation.

The trends obtained from the segmentation will then be the basis for linguistic summarization as it is explained is the following subsection.

All segments are here considered linear. The following approach to study time segments and compute linguistic summaries based on them was developed by Kacprzyk, Wilbik and Zadrozny [3] and is largely followed. It considers three aspects of the time series: (1) dynamic of change, (2) duration, (3) and variability.

1. **Dynamic of Change**
   The dynamic of change refers to the speed of changes that can be described by the slope of a line representing a trend. Thus, to quantify dynamics of change it is used the interval of possible angles $\alpha \in \langle -90, 90 \rangle$.

2. **Duration**
   The duration describes linguistically the length of a single trend that is related to a fuzzy set defined over a time span of time series.

3. **Variability**
   Variability refers to how the data is spread out and varies through time.

3.1. Quality Measures for linguistic summaries

In order to evaluate the quality of the generated linguistic summaries, quality measures have been used. Except for the truth value, the quality measures here introduced have been proposed in Yager Ford and Canas [13]. Another quality criteria, the degree of focus, found in [2] and widely used will be also here introduced. This way, the linguistic summaries are commonly evaluated through the following criteria:

- truth value (validity),
- degree of imprecision,
- degree of specificity,
- degree of fuzziness,
- degree of covering,
- degree of focus,
- degree of appropriateness,
- measure of informativeness,
- length of the summary,

4. Background

4.1. Data Set Characterization and Treatment

This study used data from the Multi-parameter Intelligent Monitoring for Intensive Care (MIMIC III) database. This is a large database of ICU patients admitted to the Beth Israel Deaconess Medical Center, collected from 2001 to 2012, that has been deidentified by removal of all Protected Health Information.

Clinical records store various measurements for each patient. These measurements can include patient measurements, fluid output, and laboratory measurements

The used data contains information about all the patients who stayed longer than 24 hours and without chronic renal disease.

At this present study, from all the variables available at MIMIC III database, only age, gender, history of hypertension, MAP, the UO rate from previous hour, if vasopressor where administrated at current hour, magnesium, potassium, OASIS scores, OASIS respiratory rate score and Elixhause for fluid and electrolyte were used. The selection of this particular variables was made taken into consideration the work done in [7].

A critical event, that will be the main concern of this study, is characterized by a drop at the UO below $30ml/h$ in the next hour. Depending on MAP measures during the first critical event, the dataset was split into two different groups as stated on the first chapter: first group with the patients who had a critical event before the established protocol (MAP below 65 mmHg) and the second group with the patients who had a critical event after and even though the established protocol was applied (MAP above 65 mmHg). From now on, all data will be seen and characterized through these two groups, named group 1 and group 2.

The dataset was initially processed to remove outliers, patients with missing data or with no critical event, and then was divided into the two initial groups as said above. From this initial cleanse, a first group with 103 patients and a second group with 601 patients were obtained.

As previously stated, MAP and UO are measured hourly. However a linear regression was made in order to better show the temporal evolution of both variables for all patients. This option took also into consideration the approach described in Section 1.2 to deal with time-varying Linguistic Summaries. For these reasons, all graphics at this report concerning MAP and UO will be linear instead of discrete.

The following figures show examples of MAP and UO for all patients linearised conformed what was said above.

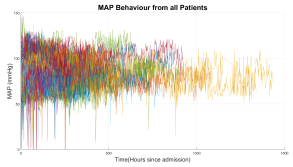

Figure 1: MAP for all patients

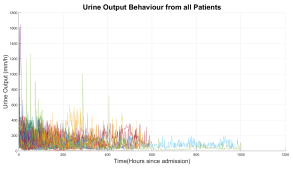

Figure 2: UO for all Patients

4.2. Clustering Techniques

As it will be further observed in the following chapter, some complementary techniques were used both previous and for a possible linguistic summaries creation. In this chapter these methods will be better explained.

Clustering is a technique that partitions a data set into small groups of similar objects (data points), producing, this way, a concise system's representation. It's applied in several fields such as data mining, machine learning, pattern recognition data compression, etc.

The techniques here described were used in both data partition and linguistic summaries grouping.

- **K-Means**
  This algorithm divides the data from a group of n vectors $x_j j = 1, , n$ in c groups $G_i i = 1, , c$ and finds the cluster centres or group profile so that the distance is minimized. The Euclidean distance was the metric chosen and the cost function is:

$$J = \sum_{i=1}^{c} \left(\sum_{k, x_k \in G_i} \|x_k - c_i\|^2 \right) \tag{7}$$

  The function J_i is the cost function inside group i. The group partition is typically defined as a binary matrix with a membership function $U_{(c \times n)}$ where $u_{i,j}$ is 1 if x_j belongs to group I and 0 if it doesn't.

- **C-Means**
  Fuzzy C-Means is a data clustering technique wherein each data point belongs to a cluster to some degree that is specified by a membership grade. This technique was originally introduced by Jim Bezdek in 1981 as an improvement on earlier clustering methods. The FCM algorithm aims to partition a finite collection of n elements $X = x_1, ..., x_n$ into a collection of clusters with respect to some given criterion. The algorithm returns a list of centres $C = c_1, ..., c_c$ and a partition matrix $W = w_{i,j} \in [0, 1], i = 1, , n; j = 1, ..., c$ where each element $w_{i,j}$ represents the degree to which element x_i belongs to cluster c_j.
  Fuzzy C-Means aims to minimize the objective function

$$\underset{C}{\operatorname{argmin}} \sum_{i=1}^{n} \sum_{j=1}^{c} w_{i,j}^{m} \|x_i - c_j\|^2$$

  where,

$$w_{i,j} = \frac{1}{sum_{k=1}^{c} \frac{\|x_i - c_j\|}{\|x_j - c_k\|}^{\frac{2}{m-1}}}$$

- **Single-linkage Clustering**
  Single-linkage clustering is one method of hierarchical clustering. It is based on grouping clusters in bottom-up fashion, at each step combining two clusters that contain the closest pair of elements that don't belong to the same cluster.

  At the beginning, each element is in a cluster of its own. The clusters are then sequentially combined into larger clusters, until all similar elements end up being in the same cluster. At each step, the two clusters separated by the shortest distance are combined.

  The distance between two clusters is determined by a single element pair, more precisely two elements (one in each cluster) that are closest to one another. The shortest of these links which remains after all steps, causes the fusion of the two clusters whose elements are tangled.

The method is also known as nearest neighbour clustering for this reason.

The linkage function is computed as

$$D(X,Y) = \min_{x \in X,\ y \in Y} d(x,y) \tag{8}$$

- **NERF C-Means**
  NERF C-Means as defined in [10] is an extension of relational FCM, which is dual to FCM when D is a Euclidean matrix. The non-Euclidean added to RFCM yields NERFCM, shortened to NERF. The extension is based on the b spread transformation, which is used to adjust D when distances in D are non-Euclidean. It is necessary to specify the number of clusters for each run of NERF.

- **Correlation Cluster Validity**
  The Correlation Cluster Validity (CCV) indices compare a given U to the structure in the data matrix D by first transforming U into a matrix of dissimilarities using the formula

$$D(U) = 1_n - \left[\frac{U^T U}{\max_{i,j}(U^T U)_{i,j}} \right] \tag{9}$$

  Pearson and Spearman's CCV were used in the present work.

5. Implementation

As stated in section 2, every summary is composed by a quantifier (in this case almost all, most, many, about a half and few) and a summarizer. Here, two types of summaries were considered: categorical with simple ("Qy's are P) and extended ("QRy's are P") forms and extended temporal summaries ("E_T among all Ry's, Q are P").

For all procedures, five linguistic quantifiers, Q, were considered *almost all*, *most*, *many*, *about a half*, and *few*. There were also considered five summarizers,P: *quickly decreasing*, *decreasing*, *constant*,*increasing* and *quickly increasing*.

For the extended categorical linguistic summaries and extended temporal summaries, the R qualifier was concerned with duration and had three linguistic values: *short*, *medium* and *long*. Finally, for the last, E_T related to the time span the summary occurs had also three linguistic value: *Far from critical event*, *Moderately close to critical event* and *Close to critical event*. The quantifiers, Q and qualifier, P, were defined using trapezoidal membership functions.

For this study, categorical univariable summaries, categorical multivariable summaries and temporal linguistic summaries were chosen. Also, for the univariable summaries an analysis based on C-means clustering and demographic variables (history of hypertension) was performed.

5.1. Categorical linguistic summaries

Categorical linguistic summaries analyse different sets of time series at the same hour interval and look for similarities or differences between the two patients' groups (for the same time intervals). Additionally, extended categorical summaries were generated considering R as the trend's duration. These last summaries aimed to compare the similarities and differences between patients from the two different groups considering their longest trends.

- **Categorical multivariable summaries**
  It is an extension of categorical summaries that analyse unions of univariable linguistic summaries by computing the minimum, product and Hamacher product between individual degrees of membership.

  Multivariable Linguistic Summaries were here only applied to the most representative summaries found.

5.2. Temporal linguistic summaries

Temporal linguistic looked at data from a combination of both length, position in the time span and dynamic of change aiming to give a different perspective and focus that may not be found in conventional data analysis. The duration of each trend is a discrete value that has 0 and 5 as its range values. The time span is also given by a discrete function with also 0 and 5 has its range values.

5.3. Clustering of Linguistic Summaries

Additionally, in order to achieve greater conciseness, clustering based on similarity was used to obtain sets of linguistic summaries, each represented by their medoid. The methods used to evaluate similarity were SL and NERF with both Pearson and Spearman correlations to determine the ideal number of clusters. This technique was found to be a substitute of other quality criteria as a means not to evaluate the quality of the summaries in absolute value but to determine the most representative ones.

6. Results

In this section, the summaries obtained and their respective quality measures will be presented. A soft analysis will be made concerning their practical meaning and how this final sets can summarize entire agglomerates of patients' time series.

Here the most representative linguistic summaries for each analysis will be shown. This was made resorting to the validity criteria.

As already stated, the summaries must meet a minimal value of validity (and degree of focus when extended). However, a "good" linguistic summary should have other good values for the criteria presented in 2.4. The criteria can be ranked by importance in the following way [8] (the degree of impre-

cision, degree of fuzziness, degree of specificity and length of the summary weren't here considered for being concerned only to the form of summary):

1. truth value, $\mathcal{T}$,
2. measure of informativeness, $\mathcal{I}$,
3. degree of covering (support), d_c, degree of appropriateness, d_a,
4. degree of focus, d_f.

6.1. Categorical linguistic summaries

Considering the criteria ranking presented before, the following rank of summaries is obtained for each group and variable (MAP and UO). For means of simplification, only the three more relevant summaries will be shown.

1. $\mathcal{S}_{11}$: About a half patients of group 1 have MAP three to four hours before CE quickly decreasing.
2. $\mathcal{S}_{12}$:Few patients of group 1 have MAP two to three hours before CE quickly decreasing.
3. $\mathcal{S}_{13}$: Many patients of group 1 have MAP one hour before CE until CE quickly decreasing.

For group 2 the most representative summaries found for this analysis were:

1. $\mathcal{S}_{21}$: Few patients of group 2 have MAP one hour before CE until CE quickly increasing.
2. $\mathcal{S}_{22}$: Few patients of group 2 have MAP three to four hours before CE quickly decreasing.
3. $\mathcal{S}_{23}$: Few patients of group 2 have MAP four to five hours before CE quickly decreasing.

As the metrics for this group are very similar for every summary, the summaries concerning the same time interval has the ones for group 1 were sought so that a analyses of the similarity of the two groups through their set of times series was possible. In what UO is concerned, the most relevant summary for both groups is evidently the one referring to the last hour. For the rest of the summaries, the metrics are very similar and between groups no significant differences between summaries were found, for which no ranking was made.

The linguistic summaries obtained above, for concerning very specific time periods came in a small amount which turns the inter-group analysis hard and inconclusive. For this reason, a new analysis was made looking to the time series in its all and describing their behaviour through the six hours span.

Similar to what was done for the summaries for time intervals, the summaries for the two groups were ranked according to their criteria:

1. $\mathcal{S}_{31}$: Few medium MAP trends from patients of group 1 are decreasing.
2. $\mathcal{S}_{32}$: About a half medium MAP trends of group 1 are increasing.
3. $\mathcal{S}_{33}$: Few medium MAP trends from group 1 are quickly increasing.

For group 2 the most representative summaries found for this analysis are:

1. $\mathcal{S}_{41}$: Few medium MAP trends from patients of group 2 are decreasing.
2. $\mathcal{S}_{42}$: Few short MAP trends from patients of group 2 are quickly decreasing.
3. $\mathcal{S}_{43}$: Few medium MAP trends from group 2 are increasing.

Multivariable linguistic summaries

The multivariable categorical linguistic summaries were generated from the most representative ones for MAP (once that for the UO the summaries were equal in both groups). This procedure aimed to see how the conjunction between groups of two from the most representative would influence the summary, studying this way what could be the influence of two MAP intervals on a critical event.

For group 1 five summaries were created while for group 2 only one was generated. None of the summaries had a more restrictive quantifier than "few". However, their generation shows a minimum of validity (more than 0.65) and this way show that this time periods may have a relative significance.

Nonetheless no relevant conclusions could be withdrawn from this procedure.

6.2. Temporal linguistic summaries

This way, the most representative temporal extended linguistic summaries for MAP trends from patients of group 1 are:

1. $\mathcal{S}_{11}$: Moderately close to critical event among all short MAP trends, almost all are increasing.
2. $\mathcal{S}_{12}$:Close to critical event among all long MAP trends, many are increasing.
3. $\mathcal{S}_{13}$: Far from critical event among all short MAP trends , few are quickly decreasing.

From the summaries above, $\mathcal{S}_{12}$ is the one that draws more attention. It reports a clearly distinct (and contradictory) behaviour to the one reported by one of the categorical summaries in the first analysis (remembering: "Many patients have MAP one hour before CE until CE quickly decreasing"). Nonetheless, while the second reports only to the

last hour and to **all** MAP trends, the first considers only **long** trends that finish at a time span "close to the critical event".

Concerning patients from group 2, the summaries generated were:

1. $\mathcal{S}_{21}$: Far from critical event among all short MAP trends , few are quickly decreasing.

2. $\mathcal{S}_{22}$: Moderately close to critical event among all short MAP trends , few are constant.

3. $\mathcal{S}_{23}$: Moderately close to critical event among all medium MAP trends , few are quickly decreasing.

Summary $\mathcal{S}_{21}$ is similar to $\mathcal{S}_{13}$ and ranks very high among all criteria. This may show that it is very representative of this two groups.

Finally, for the UO trends for both group 1 and 2 the one with highest criteria is "Close to critical event among all long UO trends, most are increasing", being for group 2 by far the most representative according to the quality criteria. This summary is very surprising once again seems to contradict what the categorical summaries said and most importantly what a critical event is. However, again these summaries is only related to **long** trends and ends one or two hours before the critical event.

6.3. Clustering of linguistic summaries

Following [10], the termination norm for the NERF was $e = 0.0001$, initialization of U was random and the degree of fuzziness was $m = 1.5$.

The following table lists values of the CCV validity indices for terminal partitions of SL and NERF. Recall that the CCV indices must be maximized.

c	**SL**		**NERF**	
	ccvp	**ccvs**	**ccvp**	**ccvs**
2	0.5328	0.8122	0.5390	0.6966
3	0.6279	0.9572	0.6712	0.4995
4	0.7845	0.8248	0.7827	0.6812
5	0.8514	0.8071	-	-
6	0.9260	0.7490	-	-
7	0.9742	0.6768	0.9742	0.6768

Table 1: Validation indices for different clustering methods

From Table 1, it is clear that for both SL with Pearson correlation and NERF with Pearson correlation the ideal number of clusters is seven, that is all linguistic summaries will be represented by themselves once they are in singleton clusters. On the other hand, for Spearman correlation, only three clusters will be needed for SL and two for NERF.

As $c = 7$ is equivalent to the all set of linguistic summaries, only the medoids computed for 2 and 3 clusters will be shown for both SL and NERF.

Being N_i the number of elements in cluster i and SL_{ij} or $NERF_{ij}$ the cluster j from SL or NERF respectively from a total of i clusters.

From an examination of Table 2 and the linguistic summaries chosen for each set, it is quickly evident that both NERF and SL aggregate the linguistic summaries according to their dynamic of change. However SL also distinguishes the trend duration, separating a long trend from the other "about a half long MAP trends from patients of group 1 are decreasing.", all other groups only have "short" or "medium" trends and both cluster method don't distinguish between quantifiers "few" and "about a half". NERF only looks to the dynamic of change in this particular case. Moreover, it also becomes clear from the table above that this method differentiates fast changing trends from moderately changing trends when they are increasing, but the same doesn't happen when they are decreasing. This may be related to how the definition of "quickly increasing" and "increasing", embodied by their trapezoidal membership functions, were done, that is, the difference in increasing trends by definition might be bigger than in decreasing trends.

Finally is it interesting to notice that the linguistic summary "about a half short MAP trends from patients of group 1 are quickly decreasing." is present in both clustering methods. This might mean that this particular linguistic summary is a good representative of the description of the MAP behaviour.

With two clusters, as shown in Table 3, both methods aggregated the linguistic summaries in an identical way, clearly reporting to the dynamic of change. Again the summary "about a half short MAP trends from patients of group 1 are quickly decreasing." which is a good indicator of robustness, it "survives" as a medoid to different cluster procedures. The second medoid varies from SL and NERF: the first considers a moderately increasing trend and the second a fast increasing trend.

From this short example it was possible to understand that the dynamic of change play a key rule when defining clustering. Additionally, in this short model exists at least a linguistic summary that may be representative of what is happening in the system. However, more analyses with a bigger set of linguistic summaries are needed for better conclusions.

It is important to take into consideration that, as shown above, the linguistic medoids will vary with the chosen clustering procedure. This may lead to loss of important information when seeking for conciseness. This way clustering with different meth-

	SL medoids for c = 3	$\mathcal{T}$	N_i
SL_{31}	about a half long MAP trends from patients of group 1 are decreasing.	1	1
SL_{32}	about a half short MAP trends from patients of group 1 are quickly decreasing.	0.808	2
SL_{33}	about a half medium MAP trends from patients of group 1 are increasing.	1	3
	NERF medoids for c = 3	$\mathcal{T}$	N_i
$NERF_{31}$	about a half short MAP trends from patients of group 1 are quickly decreasing.	0.808	2
$NERF_{32}$	few short MAP trends from patients of group 1 are increasing.	1	2
$NERF_{33}$	few medium MAP trends from patients of group 1 are quickly increasing.	1	2

Table 2: Linguistic medoid for SL and NERF with 3 clusters

	SL medoids for c = 2	$\mathcal{T}$	N_i
SL_{21}	about a half short MAP trends from patients of group 1 are quickly decreasing.	0.808	3
SL_{22}	few short MAP trends from patients of group 1 are increasing.	1	4
	NERF medoids for c = 2	$\mathcal{T}$	N_i
$NERF_{21}$	about a half short MAP trends from patients of group 1 are quickly decreasing.	0.808	3
$NERF_{22}$	few short MAP trends from patients of group 1 are quickly increasing	1	4

Table 3: Linguistic medoid for SL and NERF with 2 clusters

ods should always be done for a further understanding of which summaries may be more representative and thus better medoids.

7. Conclusions

Following what has been done in linguistic summarization, this work aimed to study the Medium Arterial Blood Pressure and the Urine Output from two groups of patients, understanding, this way, what could be affecting the occurrence of a critical event.

For this to be accomplished and pursuing what has been done in target blood pressure, this work considered the six hour period previous to the first critical event and studied the demographical variables from both groups that could help to contribute to their distinction. As this isolated variables weren't found enough to differentiate both groups, a statistical analysis for the behaviour of MAP and UO for patients from both groups was made.

Considering the results obtained in this analysis and acknowledging linguistic summarization to be a procedure that condensates information contained in big data sets into concise sentences, approaches were made, according to expertise literature, to derive a comprehensive and global characterization of the behaviours of MAP and UO from both groups. The followed methodology used elements of fuzzy logic in order to transpose and handle the characteristic imprecision of natural language.

Two major analysis were done to better study the differences between the two groups: one considered simple protoforms and was easily corroborated by the previous statistical analysis of trends. The second considered extended temporal protoforms that looked into data with a new perspective, not having the present study found a way to validate them other than the quality criteria found in expertise literature.

The similarities between the two sets was computed or described through the similarity between their most representative linguistic summaries, as purposed in expertise literature.

At the end of this work and for means of exemplification, clustering of linguistic summaries was performed to further condensate data to its essential. This procedure was done as an alternative to the representative summaries selection through quality criteria and followed the premiss that a set of similar summaries could be represented by its medoid, being it enough to describe data.

From the categorical linguistic summaries, concise descriptions of the two groups were found that seemed to be concordant with what statistical analysis showed. The generated categorical summaries found what may be MAP tendencies in group 1 while reporting in group 2 what seems to be a well distributed MAP behaviour. The UO was similar for both groups and sub-groups what may indicate the changes in this variable are only a consequence of a critical event (UO drop below $30ml/h$ and not an indicator by themselves. However, the extended temporal summaries showed an UO occurrence close to the critical event in both groups that was deemed as unexpected and should be looked with further attention and care.

The performed methodology was subjected to several simplifications that may have compromised the final conclusions and results. For this reasons, it is considered that for future explorations of the target blood pressure with linguistic summarisation some considerations should be made. They will be reported in the following lines.

- A study of high-frequency data may lead to different and interesting results. The major difference to the procedure present in this work is that the time series will be continue instead of discrete. This way, a similar linearisation process similar to the ones done in expertise literature (piece-wise linearisation, for instance) could be done and a new variable, i.e. the variability, would be obtained.
- A study based on which granulations are more suitable for the studied data and which will generate more reliable summaries for this specific data set could be done.
- Introduction of causal protoforms such as if-then clauses introduced by [11] could be interesting to study the possibility of causality in this data.
- A study on how to deal with bi-dimentional summaries also should be carried out. A possible course to deal with the problem of trends discarding would be to perform summaries considering only shorts when finding short, medium and long trends and considering only medium when finding both medium and long trends. This way a comparative analysis could be done to the final summaries set that would include summaries that may have discard short and medium trends, summaries that may have discard medium and long trends and, finally, summaries that may have discard long and short trends.
- A more insightful analysis to the medoids obtained with clustering of linguistic summaries could be performed followed by a comparative analysis between them and the most representative summaries obtained with quality criteria.

Linguistic summarization offers a new and intuitive manner of looking and studying data: it gives natural language to the users, speaks to them with simple sentences and tries to embody the natural imprecision and incongruence that they are so used to. However, linguistic summarization is not the source of all answers. It is though a possible guideline for future and more insightful studies, showing hidden ways, possible lights that couldn't be found through the vast density of big data and conventional methodologies.

Acknowledgements
I would like to thank Prof. Susana Vieira and Eng. Cátia Salgado for their guidance throughout this work.

References

[1] R. J. Almeida, M.-J. Lesot, B. Bouchon-Meunier, U. Kaymak, and G. Moyse. Linguistic summaries of categorical time series for septic shock patient data. *IEEE International Conference on Fuzzy Systems*, 2013.

[2] J. Kacprzyk and A. Wilbik. Towards an efficient generation of linguistic summaries of time series using a degree of focus. *Proceedings of the 28th North American Fuzzy Information Processing Society Annual Conference*, 2009.

[3] J. Kacprzyk, A. Wilbik, and S. Zadrony. Linguistic summarization of trends: a fuzzy logic based approach. *Proceedings of the 11th International Conference Information Processing and Management of Uncertainty in Knowledge-based Systems*, 2006.

[4] J. Kacprzyk and R. R. Yager. Linguistic summaries of data using fuzzy logic. *International Journal of General Systems*, 2001.

[5] J. Kacprzyk, R. R. Yager, and S. Zadrony. A fuzzy logic based approach to linguistic summaries of databases. *International Journal of Applied Mathematics and Computer Science*, 2000.

[6] R. S. Mans, W. Aalst, R. J. B. Vanwersch, and A. J. Moleman. Process mining in healthcare: Data challenges when answering frequently posed questions. *Process Support and Knowledge Representation in Health Care*, 2013.

[7] R. Pacheco, C. Salgado, R. O. Deliberato, and S. Vieira. Short-term prediction of low kidney function in icu patients. *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2017.

[8] A. Wilbik. *Linguistic summaries of time series using fuzzy sets and their application for performance analysis of mutual funds*. PhD thesis, Systems Research Institute, Polish Academy of Sciences, 2010.

[9] A. Wilbik and U. Kaymak. Linguistic summarization of processes a research agenda. *16th World Congress of the International Fuzzy Systems Association (IFSA)*, 2015.

[10] A. Wilbik, J. Keller, and G. Alexander. Linguistic summarization of sensor data for eldercare. *Systems, Man, and Cybernetics (SMC),2011 IEEE International Conference*, 2011.

[11] D. Wu and J. M. Mendel. Linguistic summarization using ifthen rules and interval type-2 fuzzy sets. *Fuzzy Systems, IEEE Transactions*, 2011.

[12] R. R. Yager. A new approach to the summarization of data. *Information Sciences*, 1982.

[13] R. R. Yager, K. M. Ford, and A. J. Canas. An approach to the linguistic summarization of data. *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, 1990.

[14] L. A. Zadeh. A computational approach to fuzzy quantifiers in natural languages. *Computers and Mathematics with Applications*, 1983.