Predicting Survival in Critical Patients by Use of Body Temperature Regularity Measurement Based on Approximate Entropy

D. Cuesta [a], M. Varela [b], P. Miró [c], P. Galdós [b], D. Abásolo [d], R. Hornero [d], M. Aboy [e]

[a] *Technological Institute of Informatics, Polytechnic University of Valencia, Alcoi Campus, Alcoi (Spain)*

[b] *Hospital de Móstoles, Madrid (Spain)*

[c] *Department of Applied Statistics and Operational Research and Quality, Polytechnic University of Valencia, Alcoi Campus, Alcoi (Spain)*

[d] *Biomedical Engineering Group, ETSI Telecomunicación, University of Valladolid, Valladolid (Spain)*

[e] *Electronics Engineering, Oregon Institute of Technology, Oregon (USA)*

Abstract

Body temperature is a classical diagnostic tool for a number of diseases. However, it is usually employed as a plain binary classification function (febrile or not febrile), and therefore its diagnostic power has not been fully developed. In this paper we describe how body temperature regularity can be used for diagnosis. Our proposed methodology is based on obtaining accurate long–term temperature recordings at high sampling frequencies and analyzing the temperature signal using a regularity metric (approximate entropy). In this study we assessed our methodology using temperature registers acquired from patients with multiple organ failure admitted to an intensive care unit. Our results indicate there is a correlation between the patient's condition and the regularity of the body temperature. This finding enabled us to design a classifier for two outcomes (survival or death) and test it on a dataset including 36 subjects. The classifier achieved an accuracy of 72%.

Key words: Body temperature, Approximate entropy, Temperature regularity, ROC analysis, Biomedical signal processing
PACS:

Email address: dcuesta@iti.upv.es (D. Cuesta).
URL: www.iti.upv.es (D. Cuesta).

1 Introduction

Body temperature is an important diagnostic tool since its changes accompany many diseases (Dinarello and Gelfand, 2001; Dale, 2004; Mackowiack, 2000). It can also be used to monitor the course of the disease or the efficiency of treatment. Body temperature is often measured discontinuously at time intervals of up to hours and is used to establish if the patient is febrile or not. Recently, other diagnostic techniques based on continuous body temperature monitoring have been proposed (Varela et al., 2005, 2003). The rationale behind these methods is that there may be a correlation between body temperature and patient's condition that a measurement every several hours can not show. The methodology in this cases consists on registering body temperature for a long time (days or even weeks) at higher sampling rates (a few minutes instead of hours), and analyzing not only the absolute values of the temperature but also its evolution, differences, changes, and patterns.

Based on this last approach, we propose a method to map the regularity of a body temperature register into the clinical outcome using a classifier. In this paper we estimated the regularity of the body temperature register using approximate entropy (ApEn) averaged along overlapping epochs of the registers (Pincus, 1991). Previous studies have demonstrated there are clinical implications of body temperature curve complexity (Varela et al., 2005, 2003) and ApEn has been successfully used to estimate the regularity of other biomedical signals (Pincus and Keefe, 1992; Kaplan et al., 1991). ApEn changes have often been seen to be predictive of subsequent clinical changes. For instance, ApEn has been applied to studies to discriminate atypical EEGs (Bruhn et al., 2000) and respiratory patterns (Engoren, 1998) from normative counterparts, it has been used to quantify the differences in apparent regularity between the heart rate interval time series of aborted SIDS and healthy infants (Pincus et al., 1993) and to characterize postoperative ventricular dysfunction (Fleischer et al., 1993). Preliminary evidence suggests that ApEn of EEGs is predictive of epileptic seizures (Radhakrishnan and Gangadhar, 1998). It has also been applied to extract features from EEG and respiratory recordings of a patient during Cheyne-Stokes respiration (Rezek and Roberts, 1998) and to quantify the depth of anesthesia (Zhang and Roy, 2001). Within endocrinology, it has been used in multifaceted ways; for instance, in the analysis of endocrine hormone release pulsatility (Pincus, 1996), and the impact of pulsatility on the ensemble orderliness of neurohormone secretion (Veldhuis et al., 2001). ApEn has also been used to analyze intracranial pressure (ICP) signals from patients with traumatic brain injury (TBI) during episodes of abrupt intracranial hypertension (ICH). (Hornero et al., 2005, 2006) studied episodes of acute ICH in pediatric patients with severe TBI and found that the ApEn of ICP decreases during acute elevations. This suggests that the complex regulatory mechanisms that govern intracranial pressure are disrupted during

acute rises in ICP. Additionally, this study carried out a series of experiments where ApEn was used to analyze synthetic signals of different characteristics with the objective of gaining a better understanding of ApEn itself, specially with regards to its interpretability in the context of biomedical signal analysis and TBI. The results of this simulation study enable researchers to interpret the ApEn metric in terms of classical signal processing concepts such as frequency, number of harmonics, frequency variability of harmonics, and signal bandwidth.These results showed that 1) ApEn increases as the frequency and the number of harmonics of a sinusoidal signal increases, 2) ApEn is correlated with noise bandwidth, increasing as the noise bandwidth increases (ApEn is lower in the case of colored noise than for white noise), 3 typical values of ApEn for sinusoidal signals 0.001 to 0.007 ($m = 1, r = 0.25s$), and 4) the ApEn of biomedical pressure signals increases as the variability of the cardiac component increases and decreases as the the pulse morphology becomes more rounded.

2 Materials and Methods

2.1 Subjects and Study Overview

Body temperature registers were recorded for 36 subjects with multiple organ failure admitted to the Intensive Care Unit (ICU) of Mostoles Hospital, Madrid(Spain) using a portable temperature data logger (Veriteq, 2005). An example of such register is shown in Fig. 1. The subjects were assigned to one of two classes: survivors A and non-survivors B.

All patients defined as"non-survivors" died in the ICU, before discharge. All 36 studied patients were adults, with age ranging from 37 to 83 years. Mean age was 63.0 (s=11.7). There was no significant difference between survivors and non-survivors regarding age (60.2 versus 65.5). Patients suffered from either medical (pneumonia, myocardial infarction, etc) or surgical conditions (trauma, abdominal surgery, etc). There were 10 postoperative multiorgan failure due mostly to oncologic interventions, 9 sepsis of respiratory origin, 4 of abdominal origin, 3 from an urologic source, 2 secondary to politraumatism and 10 multiorgan failures of diverse etiologies.

Temperature was measured all along the admission, until the patient was discharged or considered dead and all monitoring devices were retired. Nevertheless, to avoid the influence of pre-mortem or peri-mortem conditions, the last hour was not included in the analysis. The patients were monitored for a median of 210 hours, with a (statistically non-significant) trend towards more hours of monitoring in dying patients (median 323 versus 195).

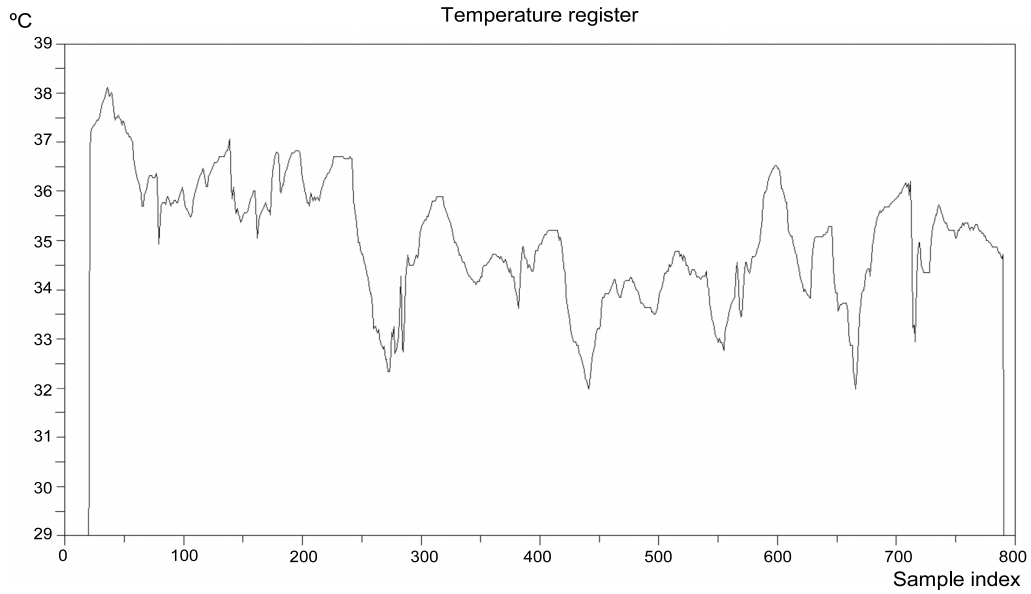

Fig. 1. Example of temperature register. Measurements are taken every 10 minutes. There are low temperature measurements at the beginning and end of the register because of the sensor hysteresis and disconnection from the patient, respectively.

The temperature was measured with a thermistor sensor attached on the right hypochondrium (right upper abdomen quadrant). When this was not possible (because of surgical wounds, etc), the left hypochondrium was used.

Patient's body temperatures were recorded using a data logger Spectrum 1000 (Veriteq, 2005) with a precision NTC 100K thermistor remote probe EPT-010. This device is able to record temperature during 10 years with an accuracy of 0.05 ℃. Measures were taken every 10 minutes, and epoch length L was 180 (30 hours, a period long enough to assess clinical evolution).

The base precision of the sensor used is 0.021℃, at 25℃. The software introduces "smoothing"' (averaging), which reduces quantization noise in the data. In order to display the maximum possible precision at a given temperature, the temperature was displayed to 0.01℃. As for the accuracy, it is difficult to answer, and even the idea of a gold standard is debatable. Temperatures are known to differ in different points of the body at the same time. Nevertheless, we believe that what the real temperature is is not so relevant. What we are studying is rather how the temperature fluctuates, and the absolute values may not be so important.

The presence of edema was not considered. Although any part of the skin may be edematous, patients were almost all the time laying on their back, and in this position the upper abdomen is not specially prone to edema. We did not analyze the relation between temperature and blood pressure. Indeed, blood pressure regulation frequently involves peripheral vasoconstriction or

vasodilatation and thus may influence temperature readings (specially in peripheral locations). Nevertheless, as stated before, we believe the issue is not what the real temperature is and how our measures correlate with it, but how the organism thermoregulates and what are the physiologic and clinical consequences of it. Rather than proposing a new thermometry technique, we try to explore the temperature variability. Whatever the mechanism or the absolute temperature is, severely ill patients seem to thermoregulate poorly (or, at the very least, their peripheral temperature variability is blunted), and this may be a marker of bad prognosis.

The resulting discrete time signals obtained from the body temperature registers, $y[n]$, were preprocessed to remove invalid values (present, for example, when a sensor disconnection takes place). Next, regularity was estimated using ApEn for both classes, and a statistical test was conducted to determine whether the ApEn means in the different classes were statistically different (Daya, 2003; Yue and Wang, 2001). Finally, a ROC analysis was performed to design a classifier. The major methodological steps are shown in Fig. 2. These stages are described in detail in the following subsections and in the appendix.

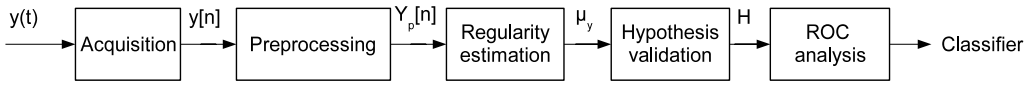

Fig. 2. Block diagram summarizing the steps followed in this study, from temperature recording to the classifier design.

2.2 Preprocessing

Prior to analysis we removed artifacts from the temperature registers. The input registers $y[n]$ may contain incorrect temperature values due to:

- Sensor disconnection. Temperature registers often contain artifacts or invalid measurements due to sensor disconnection. This disconnection is sometimes intentional (for example, when the patient is taken to surgery, long duration) or accidental (patient movement, short duration). Such measurements should be removed or, if there are only a few, interpolated, if possible. If the duration of the disconnection exceeds a threshold, register finishes at the last valid measurement, and a new register starts when correct measurements are resumed (for instance, when the patient returns from surgery).
- Border effects. There are also border effects since at the beginning of the recording there is an adaptation period for the sensor to provide a correct temperature measurement because of hysteresis, and at the end, when patient is supposed to be about to be discharged or in a pre-mortem condition,

measurements are not significant.

To deal with these invalid signal measures, the following procedure was established:

- A minimum threshold t_v for valid temperature measurements was set. Measures below that threshold were considered to be due to sensor disconnection (ambient temperature is measured instead). There is no upper threshold since disconnection only implies lower temperature measurements.
- Measurements at the beginning and at the end of the recording, below the minimum valid temperature threshold, were discarded. This is to eliminate border effects.
- Missing samples, namely, measurements below the minimum valid temperature threshold, within the recording, were linearly interpolated provided there are at most t_m consecutive missing samples. Otherwise, register finishes at the last valid measurement, and a new one starts when new measurements above t_v is found.

Since the regularity measurement was computed for fixed duration intervals, the temperature recording was split into valid overlapping epochs of length L, provided no missing sample was found. Thus, analysis was performed on uninterrupted signal epochs of fixed length.

The output of this stage is a set Y_p of valid $y[n]$ epochs. For simplicity, we will refer to any of these epochs as $x[n]$.

3 Results

In this study, 18 registers were available for each class. Long term temperature registers were preprocessed prior to the computation of the ApEn as described in section 2. Thresholds were $t_v = 30$℃(minimum valid temperature), $t_m = 2$ (maximum number of missing samples to interpolate), and $t_l = 6$ (number of overlapping samples between consecutive epochs). The percentage of invalid samples removed according to these thresholds was 0.46%.

Values for the ApEn parameters were $m = 1$ and $r = 0.2$ times the standard deviation of y. As r is normalized by the standard deviation, ApEn is amplitude scale independent. These values have proven to perform well in many cases (Abásolo et al., 2005; Pincus and Keefe, 1992; Kaplan et al., 1991). Results for each register and each class are shown in Table 1. The overall results for class A were $\mu_A = 0.7646$ and $s_A = 0.1595$ and for class B, $\mu_B = 0.5832$ and $s_B = 0.1275$.

Table 1
ApEn average values for each register. Second column corresponds to patients that survived (A), and fourth column to those who died (B).

Register	ApEn(A)	Register	ApEn(B)
01	0.8224	19	0.5996
02	0.7543	20	0.4849
03	0.6886	21	0.6629
04	0.7828	22	0.5641
05	0.5449	23	0.8788
06	0.7917	24	0.4850
07	0.9154	25	0.4197
08	0.5305	26	0.6568
09	0.8706	27	0.7176
10	0.9054	28	0.7312
11	1.0639	29	0.5733
12	0.7774	30	0.6329
13	1.0172	31	0.6466
14	0.6135	32	0.4634
15	0.6177	33	0.4520
16	0.6967	34	0.5661
17	0.5370	35	0.3560
18	0.6526	36	0.6049
μ_A	0.7546	μ_B	0.5832
s_A	0.1595	s_B	0.1275

The results of the statistical tests confirmed the existence of statistically significant differences between the classes ($\alpha = 0.01$). The parametric Student t-test yielded $t = -3.5635$ with $p-$value$= 0.0011$, rejecting H_0, and confirming $\mu_A \neq \mu_B$. This test was based on additional assumptions that were also confirmed. The normality test of Shapiro-Wilks provided the following results: $\omega_A = 0.9601$, $p-$value$= 0.5961$, $\omega_B = 0.9776$, $p-$value$= 0.9035$, and therefore normality was accepted. The homoscedasticity test yielded $F = 0.6392$ with $p-$value$= 0.3657$, namely, standard deviations were considered to be equal. Finally, while uncorrelation doesn't imply independence, the two classes could potentially be assumed to be independent based on the results of the correlation test ($\rho = 0.0653$, $p-$value$= 0.7967$).

For the non-parametric test, MW statistic was $U = 257$, with a $p-$value$=$ 0.0028. Therefore it rejected H_0 and confirmed there were significative differences between the medians of the two classes compared.

Fig.3 shows the corresponding ROC and accuracy curves. From the accuracy curve, optimal threshold was found to be 0.69, providing the best possible accuracy of 72% and AUC of 0.73. An area of 0.73 means that a randomly selected individual from the class A has an ApEn value larger than that of a randomly chosen individual from the class B in 73% of the time (Zweig and Campbell, 1993).

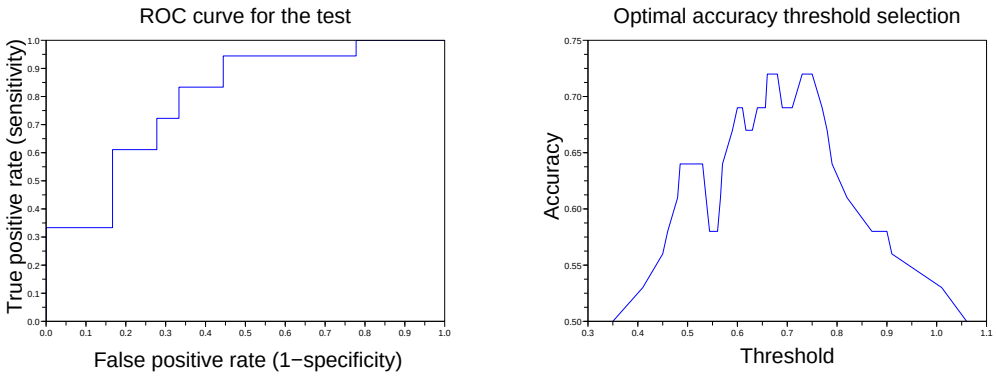

Fig. 3. ROC and accuracy curves. Area below ROC curve was 0.73, which means the classifier performance is reasonable. Optimal accuracy threshold was 0.69 for which accuracy was 72%.

4 Discussion

The results of our experiments with real body temperature registers confirmed there is a statistically significant difference between the regularity of signals from patients that survived and patients that did not. In order to use our proposed methodology the sampling frequency must be relatively high compared to the classical sampling frequency in these cases, minutes instead of hours, and acquisition has to last days or even weeks.

The importance of body temperature and its correlation with certain physiopathologically relevant parameters in healthy subjects has been shown in previous studies (Varela et al., 2003). The relationship between loss of complexity of temperature time series and clinical status of patients measured with the Sequential Organ Failure Assesment (SOFA) score has also been demonstrated (Varela et al., 2005). Our findings are consistent with those of previous studies. A distinguishing characteristic of our study with respect to previous works includes the design of a classifier to be used as a prognostic tool and having conducted a complete statistical analysis to confirm all the assumptions.

The physiological basis of these findings can only be hypothesized. Nevertheless, they are by no means unexpected. Complex biological systems have frequently been shown to display a loss of complexity when injured, and temperature may well be just another example. Thermoregulation is a vital homeostatic function, with several regulatory loops. In situation of severe illness (i.e. multiple organ failure) there may be a loss of afferent or processing functions, which would produce a "decomplexification" of its output. Furthermore, it has been proposed that one of the earliest sign of dysfunction of complex systems is an"uncoupling"' of its regulatory loops, which would determine a loss of "fine-regulation" and consequently a loss of complexity in its output.

ApEn has also been used in other similar medical studies with good results. It has been used to analyse information provided by the electroencephalogram related to the Alzheimer's disease (Abásolo et al., 2005), or study implications of heart rate dynamics (Ryan et al., 1994), among others. For temperature registers, our results confirm the usefulness of this measure for distinguishing the two classes (t-test, $p = 0.0011$, and fMW test, $p = 0.0028$).

Although ApEn also has limitations (Richman and Moorman, 2000), they do not affect the results in this case. The lack of relative consistency is avoided by means of averaging several epochs of each temperature register, and dependency on record duration is eliminated by using the same epoch length for all the registers.

ROC and accuracy curves provide information about the sensitivity and false positive rate. Since ROC curves does not depend on the scale of the test results offsets in the measures do not change it. The practical lower bound for the AUC is 0.5. Diagnostic tests with AUCs greater than 0.5 have some ability to discriminate between two classes (Obuchowski, 2003). Since our AUC is $0.73 > 0.50$, we can state discrimination ability of the classifier is adequate to have clinical utility. From the accuracy curve, optimal threshold is 0.69, for a maximum accuracy of 72%, greater than the random guess of 50%.

5 Conclusion

This work demonstrates body temperature registers provide important clinical information when considered as continuous signals using a regularity measurement based on ApEn. It applies a subpattern similarity analysis of these temperature registers to assess their regularity. Our results indicate that the worse is the condition of a patient, the more regular his body temperature register is.

We designed and assessed a classifier to map the temperature regulatity into

the two classes. The objective of this work is to have a diagnostic tool that may help forecast patient outcome, helping physicians in certain decision (i.e. intensifying or curtailing therapeutic efforts or monitoring). Furthermore, our proposed methodology is less invasive and less labor-consuming than conventional scores, such as SOFA, or APACHE (Acute Physiology and Chronic Health Evaluation), while retaining a similar predictive power (Varela et al., 2005).

References

Abásolo, D., Hornero, R., Espino, P., Poza, J., Sánchez, C., la Rosa, d., 2005. Analysis of regularity in the eeg background activity of alzheimer's disease patients with approximate entropy. Clinical Neurophysiology 8 (116), 1826–1834.

Altman, D., 1991. Practical statistics for medical research. Chapman & Hall.

Bruhn, J., Ropcke, H., Rehberg, B., Bouillon, T., Hoeft, A., 2000. Electroencephalogram approximate entropy correctly classifies the occurrence of burst suppression pattern as increasing anesthetic drug effect. Anesthesiology 93, 981–985.

Dale, D., 2004. The febrile patient, 22nd Edition. Cecil textbook of medicine. Saunders (Elsevier).

Daya, S., 2003. The t-test for comparing means of two groups of equal size. Evidence-based obstetrics and gynecology (5), 4–5.

Dinarello, C., Gelfand, J., 2001. Alteration in body temperature, 15th Edition. Harrison's Principles of Internal Medicine. McGraw Hill.

Engoren, M., 1998. Approximate entropy of respiratory rate and tidal volume during weaning from mechanical ventillation. Crit Care Med 26, 1817–1823.

Fawcett, T., jun 2006. An introduction to roc analysis. Pattern Recognition Letters 27 (8), 861–874.

Fleischer, L., Pincus, S., Rosenbaum, S., 1993. Approximate entropy of heart rate as a correlate of postoperative ventricular dysfunction. Anesthesiology 78, 683–692.

Hines, W. W., Montgomery, D. C., 1990. Probability and Statistics in Engineering and Management Science, 3rd Edition. Jon Wiley & Sons.

Ho, K., Moody, G., Peng, C., Mietus, J., Larson, M., Levy, D., Goldberger, A., aug 1997. Predicting survival in heart failure case and control subjects by use of fully automated methods for deriving nonlinear and conventional indices of heart rate dynamics. Circulation 3 (96), 842–848.

Hornero, R., Aboy, M., Abasolo, J., McNames, J., Goldstein, B., 2005. Interpretation of approximate entropy. analysis of intracranial pressure approximate entropy during acute intracranial hypertension. IEEE Trans Biomed Eng 52, 1671–1680.

Hornero, R., Aboy, M., Abasolo, J., McNames, J., Wakeland, W., Goldstein,

B., 2006. Complex analysis of intracranial hypertension using approximate entropy. Crit Care Med 34, 87–95.

Kaplan, D., Furman, M., Pincus, S., Ryan, S., Goldberger, A., 1991. Aging and the complexity of cardiovascular dynamics. Byophys. J. (59), 945–949.

Lim, T., Loh, W., 1996. A comparison of tests of equality of variances. Computational statistics & Data Analysis (22), 287–301.

Mackowiack, P., 2000. Temperature regulation and the pathogenesis of fever, 5th Edition. Mandell, Douglas, and Bennett's Principles and Practice of Infectious Diseases. Churchill Livingstone.

Obuchowski, N. A., 2003. Receiver operating characteristic curves and their use in radiology. Radiology (229), 3–8.

Pincus, S., 1996. Older males secrete luteinizing hormone and testosterone more irregularly and joint more asynchronously, than younger males. Proc Natl Acad Sci USA 93, 14100–14105.

Pincus, S., Cummings, T., Haddad, G., 1993. Heart rate control in normal and aborted SIDS infants. Am J Physiol (Regulatory Integrative Comp Physiol) 264, R638–R646.

Pincus, S. M., mar 1991. Approximate entropy as a measure of system complexity. Proc. Natl. Acad. Sci. USA 88, 2297–2301.

Pincus, S. M., Keefe, D. L., 1992. Quantification of hormone pulsatility via an approximate entropy algorithm. Am. J. Physiol. (Endocrinol. Metab.) (262), 741–754.

Radhakrishnan, N., Gangadhar, B., 1998. Estimating regularity in epileptic seizure time series data. a complexity-measure approach. IEEE Eng Med Biol Mag 17, 89–94.

Rezek, I., Roberts, S., 1998. Stochastic complexity measures for physiological signal analysis. IEEE Trans Biomed Eng 45, 1186–1191.

Richman, J. S., Moorman, J. R., 2000. Physiological time-series analysis using approximate entropy and sample entropy. Am. J. Physiol. Heart Circ. Physiol. (278), 2039–2049.

Ryan, S. M., Goldberger, A. L., Pincus, S. M., Mietus, J., Lipsitz, L. A., 1994. Gender and age-related differences in heart rate dynamics: Are women more complex than men? J. Am. Coll. Cardiol. (24), 1700–1707.

Shapiro, S. S., Wilk, M. B., 1965. An analysis of variance test for normality. Biometrika (52), 591–611.

Varela, M., Calvo, M., Chana, M., Gómez-Mestre, I., Asensio, R., Galdós, P., dec 2005. Clinical implications of temperature curve complexity in critically ill patients. Critical care medicine 33 (12), 2764–71.

Varela, M., Jiménez, L., Faria, R., 2003. Complexity analysis of the temperature curve: new information from body temperature. Eur. J. Appl. Physiol. (89), 230–237.

Veldhuis, J., Johnson, M., Veldhuis, O., Straume, M., Pincus, S., 2001. Impact of pulsatility on the ensemble orderliness (approximate entropy) of neurohormone secretion. Am J Physiol (Regulatory Integrative Comp Physiol) 281, R1975–R1985.

Veriteq, 2005. Data loggers. Veriteq, www.veriteq.com.
Yue, S., Wang, C. Y., 2001. The influence of serial correlation on the mann-whitney test for detecting a shift in median. Advances in water resources (25), 325–333.
Zhang, J., Wu, Y., 2005. Likelihood-ratio tests for normality. Computational statistics & data analysis (49), 709–721.
Zhang, X.-S., Roy, R., 2001. Derived fuzzy knowledge model for estimating the depth of anesthesia. IEEE Trans Biomed Eng 48, 312–323.
Zweig, M., Campbell, G., 1993. Receiver-operating characteristic (roc) plots: a fundamental evaluation tool in clinical medicine. Clin. Chem. (39), 561–577.

A Appendix

A.1 Regularity estimation

Biomedical signal regularity measurement has proven to be an effective way to obtain new information from these signals that correlates well with clinical condition (Varela et al., 2005). One of the most used mathematical tools to compute regularity in signals is ApEn (Abásolo et al., 2005). This is a measure aimed at obtaining the regularity of a data series because it reflects the probability that patterns within the series are not followed by similar ones. Therefore, a data series containing many repetitive patters will have a low ApEn, whereas a less predictable one will have a higher ApEn (Ho et al., 1997).

The algorithm for computing ApEn is as follows. Given an input data series $x[n]$ of length N, an epoch of valid temperature recordings, two input parameters must be chosen in order to compute its ApEn, the length of the pattern m, and the distance threshold r.

A data series pattern of length m is given by:

$$x_m(i) = \{x[i], x[i+1], \ldots, x[i+m-1]\},$$

that is, m refers to the number of consecutive temperature measures assumed to form a possible repetitive pattern within $x[n]$, and starting at sample $x[i]$.

The distance between two generic patterns $x_m(i)$ and $x_m(j)$ is given by:

$$d\left(x_m\left(i\right), x_m(j)\right) = \max\left(\left|x\left[i+k\right] - x\left[j+k\right]\right|\right), 1 \leq k \leq m \qquad \text{(A.1)}$$

The distance threshold r determines if $x_m(i)$ and $x_m(j)$ can be considered

similar when $d\left(x_m\left(i\right), x_m(j)\right) \leq r$. Given the set of all possible patterns of length $m, (x_m(1), x_m(2), \ldots, x_m(N-m+1))$, we define:

$$C_{r,m}(i) = \frac{k_{i,m}(r)}{N-m+1} \tag{A.2}$$

where $k_{r,m}(i)$ is the number of patterns $x_m(j)$ that are similar to $x_m(i)$ according to the distance threshold r. Hence, $C_{r,m}(i)$ is the fraction of patterns of length m starting at $j, 1 \leq j \leq N-m+1$ whose distance to pattern starting at i, is below the threshold r, that is, they are considered to be similar to pattern $x_m(i)$. This fraction is computed for each pattern, and then another quantity can be defined as:

$$\phi^m(r) = \frac{1}{N-m+1} \sum_{i=1}^{N-m+1} \log C_{r,m}(i).$$

Finally, the computation of the ApEn of a temperature epoch $x[n]$, $\mathrm{ApEn}(m, r)$ is given by:

$$\mathrm{ApEn}(m, r) = \left[\phi^m(r) - \phi^{m+1}(r)\right] \tag{A.3}$$

Namely, ApEn quantifies the relative prevalence of repetitive patterns of length m compared with patterns of length $m+1$ (Ho et al., 1997). ApEn is computed for all the epochs in the temperature register, and then the mean $\mu_y = \mathrm{mean}(ApEn(x_i[n])), \forall x_i[n] \in Y_p$, is obtained.

A.2 Hypothesis validation

ApEn was calculated for every temperature register in classes A and B as described in previous section, and the mean for both classes was obtained, $\mu_A = \mathrm{mean}(\mu_{y_i}), \forall y_i \in A$, and $\mu_B = \mathrm{mean}(\mu_{y_j}), \forall y_j \in B$. The objective of the hypothesis validation was aimed at assessing if μ_A and μ_B differences were statistically significant. There are several statistic tests for this validation but in order to consider all the possible scenarios, we chose two complementary tests (Hines and Montgomery, 1990). The first one, the classical parametric Student's t-test (Daya, 2003), based on the assumptions of data normality and homoscedasticity, difficult to make when not many input instances are available, and the second one, the Mann-Whitney test (Yue and Wang, 2001), a non-parametric method that does not require the normality assumption.

For the Student's t-test, the null hypothesis H_0 is that the two ApEn means for classes A and B are considered to be equal, and then the objective is to decide whether to accept or reject such hypothesis. In order to be able

to carry out this test, data normality, homoscedasticity, and independence must apply. Normality can be assured using the Shapiro-Wilks test (Shapiro and Wilk, 1965; Zhang and Wu, 2005). Homoscedasticity can be confirmed by means of the Bartlett test (Lim and Loh, 1996), and independency by a sample correlation study (Hines and Montgomery, 1990).

Taking a distribution as normal when not many observations are available may lead to incorrect conclusions. The Mann-Whitney U test (MW) (Altman, 1991), a non parametric test, can be carried out instead in order not to make such assumptions, and be able to assess if there are significative differences between the two populations with respect to their medians. Again, null hypothesis H_0 states the two populations from which samples have been drawn have equal medians, and the alternative hypothesis H_1 states medians are different.

To carry out the test, both groups are put together and observations are rank-ordered from lowest to highest. Then rankings are returned to the class, A or B, to which they belong. The test statistic U is given by (Yue and Wang, 2001):

$$U = min\{U_1, U_2\}$$

with:

$$\begin{aligned} U_1 &= n_A n_B + \tfrac{n_A(n_A+1)}{2} - W_A \\ U_2 &= n_A n_B + \tfrac{n_B(n_B+1)}{2} - W_B \end{aligned}$$

and where U_1 is the total number of class A observations preceding class B observations, and the other way round for U_2. W_A and W_B are the rank sums for each class.

Finally, additional tests were carried out to accept or reject the assumptions of normality, homoscedasticity and independence for the data (Shapiro and Wilk, 1965; Zhang and Wu, 2005).

A.3 ROC analysis

ROC analysis is a very useful tool to select a classifier and visualize its performance and behaviour (Fawcett, 2006). It has been used in many medical diagnosis applications. If the previous statistical tests determine that both classes have different means, a classifier can be designed with this method. The input to the classifier is the regularity measure obtained with ApEn, and the output is a mapping to a predicted class.

Our classification problem consists of mapping an input instance (mean ApEn of a temperature epoch) to one of the classes in the discrete set $\{A, B\}$. If we call A the positive class, and B the negative class, we can define the following performance metrics for the classifier:

- True positive (TP): instance is A and it is classified as A.
- False positive (FP): instance is B but it is incorrectly classified as A.
- True negative (TN): instance is B and it is classified as B.
- False negative (FN): instance is A and it is incorrectly classified as B.
- Sensitivity: Correctly classified instances of A divided by the total number of A instances.
- Specificity: Correctly classified instances of B divided by the total number of B instances.
- Accuracy: Ratio of correctly classified instances: $\frac{(TP+TN)}{(P+N)}$, where P and N are the total number of positives and negatives, respectively.

A threshold is used to obtain a crisp classifier, that is, instances can only belong to a single class. If the score is greater than that threshold we map instance into class A, otherwise into class B. The objective is to find an optimal threshold that maximizes accuracy.

The ROC curve is plotted considering each possible threshold as a different classifier, obtaining a set of points in the ROC space that form the resulting curve, a step function. Only one of the possible classifiers is finally chosen, that considered optimal from the accuracy point of view. Finally, the area under the ROC curve (AUC) is computed in order to assess performance.