BMC Bioinformatics

Selection of entropy-measure parameters for knowledge discovery in heart rate variability data

Christopher C Mayer[1][*][†], Martin Bachler[1,2][†], Matthias Hörtenhuber[2], Christof Stocker[3], Andreas Holzinger[3], Siegfried Wassertheurer[1]

Abstract

Background: Heart rate variability is the variation of the time interval between consecutive heartbeats. Entropy is a commonly used tool to describe the regularity of data sets. Entropy functions are defined using multiple parameters, the selection of which is controversial and depends on the intended purpose. This study describes the results of tests conducted to support parameter selection, towards the goal of enabling further biomarker discovery.

Methods: This study deals with approximate, sample, fuzzy, and fuzzy measure entropies. All data were obtained from PhysioNet, a free-access, on-line archive of physiological signals, and represent various medical conditions. Five tests were defined and conducted to examine the influence of: varying the threshold value r (as multiples of the sample standard deviation σ, or the entropy-maximizing r_{Chon}), the data length N, the weighting factors n for fuzzy and fuzzy measure entropies, and the thresholds r_F and r_L for fuzzy measure entropy. The results were tested for normality using Lilliefors' composite goodness-of-fit test. Consequently, the p-value was calculated with either a two sample t-test or a Wilcoxon rank sum test.

Results: The first test shows a cross-over of entropy values with regard to a change of r. Thus, a clear statement that a higher entropy corresponds to a high irregularity is not possible, but is rather an indicator of differences in regularity. N should be at least 200 data points for $r = 0.2\ \sigma$ and should even exceed a length of 1000 for $r = r_{Chon}$. The results for the weighting parameters n for the fuzzy membership function show different behavior when coupled with different r values, therefore the weighting parameters have been chosen independently for the different threshold values. The tests concerning r_F and r_L showed that there is no optimal choice, but $r = r_F = r_L$ is reasonable with $r = r_{Chon}$ or $r = 0.2\sigma$.

Conclusions: Some of the tests showed a dependency of the test significance on the data at hand. Nevertheless, as the medical conditions are unknown beforehand, compromises had to be made. Optimal parameter combinations are suggested for the methods considered. Yet, due to the high number of potential parameter combinations, further investigations of entropy for heart rate variability data will be necessary.

Background

Heart rate variability (HRV) is the variation of the time interval between consecutive heartbeats. It highly depends on the extrinsic regulation of the heart rate (HR) and reflects the balance between the sympathetic and the parasympathetic nervous system [1]. Batchinsky et al. [2] develop a collection of methods for using HRV to describe regular periodic oscillations in the heart rate, attributed to the vagal and/or sympathetic branches of the autonomic nervous system.

Research on HRV has attracted considerable attention in the fields of psychology and behavioral medicine. It has its origin in the search for non-invasive correlates of injury severity which can be extracted from available signals in order to discover new cardiac biomarkers [2]. These signals are usually ones that are routinely measured, and include sources like a photoplethysmogram or the electrocardiogram (ECG) [3]. Electrocardiography is an interpretation of the electrical activity of the heart over some

---

* Correspondence: christopher.mayer@ait.ac.at
† Contributed equally
[1]AIT Austrian Institute of Technology, Health & Environment Department, Biomedical Systems, Donau-City-Str. 1, 1220 Vienna, Austria
Full list of author information is available at the end of the article

period of time. It is a non-invasive procedure, using electrodes attached to the surface of the skin, and is commonly used to measure the heart rate, the regularity of the beats, and characterize properties or injuries in the heart chamber. An *R-to-R interval (RRI)* describes the latency between two consecutive R peaks in the ECG. The RRI time series are used as input for the determination of HRV parameters.

In studies of HRV, both time- and frequency-domain measures are typically used by practitioners and researchers [1,4]. Additionally, further knowledge about the subject's status can be discovered by the evaluation of certain patterns and shifts in an "apparent ensemble amount of randomness" of a stochastic process [5]. This randomness, as well as the predictability of this process, can be measured by entropy [6]. Thus, it is a commonly used tool to describe the regularity of large biomedical data sets. The more regulatory inputs a system has, the higher its irregularity is, due to interference of those regulatory systems. This assumption is also true for many biomedical systems such as HRV [7]. Therefore, it is a reasonable hypothesis that a more regular heart rate variability is connected to a defect in a regulatory system. To measure this irregularity, Pincus et al. proposed approximate entropy (ApEn) [8]. Further types of entropy were developed based on this method to improve it [9]. Because of its ability to measure regularity, entropy is widely used as a diagnostic tool in medicine to derive and discover biomarkers in large biomedical data. Its applications range from sudden infant death syndrome [7], to complexity analysis of intracranial pressure dynamics during periods of severe intracranial hypertension [10], to quantification of amplitude variations in mechanomyographic signals [11], to analysis of short gait data sets [12], to automatic detection of normal, pre-ictal, and ictal conditions from recorded electroencephalography signals [13], to the postural sway in stroke patients [14].

The approach to quantify the structural complexity (or, inversely, regularity) of the HRV is called heart rate complexity (HRC) and utilizes methods derived from nonlinear dynamics. Note that complexity and variability are not necessarily the same [15]. A periodical signal, such as a sine wave, is variable but not complex. This property allows complexity measures to ignore the complicated periodic oscillations to at least some extent [16].

To date, numerous entropy types are widely accepted as measures of the HRC. However, since their function is controlled by three, four or six parameters, there are many possible combinations to choose from. The selection of criteria for these parameters is controversial and heavily depends on the intended purpose and the data at hand [17]. The variation of only one parameter often results in highly non-linear behaviour [12]. Therefore, results of calculations with different parameters cannot be easily extrapolated from existing data, but have to be computed individually. Hence, the variation of several parameters in order to optimize the reliability of the results is a very time consuming process. To be used in daily routine, reasonable parameters have to be selected prior to the evaluation of HRC.

The determination of appropriate parameters for entropy in general [11,12,18-20] and for HRV applications in particular [16,17,21,22] is the subject of ongoing research. A summary of the current knowledge regarding parameter selection is given in the subsection "Parameter selection". In order to verify the results of previous publications and to extend them by the usage of more entropy measures, this work focuses on the parameter selection for approximate, sample, fuzzy and fuzzy measure entropy and their implications for HRV data. The objective of this paper is to provide a reference for choosing parameters for HRV applications based on their influence on the different entropies' ability to distinguish significantly between pathological and non-pathological recordings. The main research questions addressed in this paper are: (1) what does the choice of the threshold value(s) mean for the entropies to be a direct measure of regularity, (2) how does the data length influence significance for different data sets, (3) how should the weighting factor(s) be chosen for fuzzy and fuzzy measure entropy, and finally, (4) what are the constraints for choosing the threshold values?.

Methods
In the following sections approximate (ApEn), sample (SampEn), fuzzy (FuzzyEn) and fuzzy measure entropy (FuzzyMEn) are described in detail. They are built on each other and ordered by increasing complexity. The number of parameters increases from three for ApEn to six for FuzzyMEn, with the basic parameters staying the same and being extended by new ones in each step. Afterwards, the challenge of parameter selection and the tests conducted for the parameters to the different entropy types are described. Finally, the data used for all performed tests are briefly described.

Approximate entropy
The main idea behind approximate entropy is that a sequence is regular if a subsequence and an expansion of the subsequence are similar. It was developed by Pincus [8] and is calculated the following way.

Given a sample sequence $\{u_1, \ldots, u_N\}$, a template length m and a threshold value r, the sequence is first split into overlapping sequences $\{X_1^m, \ldots, X_{N-m+1}^m\}$ of length m, with $X_i^m := \{u_i, \ldots, u_{i+m-1}\}$. For example, for the input sequence $\{1, 2, 3, 4, 5\}$ the overlapping sequences are $\{\{1, 2, 3\}, \{2, 3, 4\}, \{3, 4, 5\}\}$ for $m = 3$.

Next, define C_i^m as the number of $j = 1, \ldots, N - m + 1$, for which $d(X_i^m, X_j^m) < r$, where $d(X_i^m, X_j^m)$ is the Chebyshev distance, i.e., the maximum distance between two elements of X_i^m and X_j^m. Constructing the same sequences, but with template length $m + 1$, yields C_i^{m+1}. Then, ϕ^m and ϕ^{m+1} are defined as:

$$\phi^m := \frac{1}{N - m + 1} \sum_{i=1}^{N-m+1} \ln \left(\frac{C_i^m}{N - m + 1} \right), \text{ and} \quad (1)$$

$$\phi^{m+1} := \frac{1}{N - m} \sum_{i=1}^{N-m} \ln \left(\frac{C_i^{m+1}}{N - m} \right). \quad (2)$$

The approximate entropy is then defined as ApEn $(m, 4) := \lim_{N \to \infty} (\phi^m - \phi^{m+1})$, which can be estimated by ApEn $(m, r, N) := \phi^m - \phi^{m+1}$.

Sample entropy
Richman and Moorman showed in [23] that approximate entropy is biased towards regularity. Thus, they modified it to sample entropy. The main difference between the two is that sample entropy does not count self-matches, and only compares the first $N - m$ subsequences instead of all $N - m + 1$, so the same amount of subsequences are used in φ^m and φ^{m+1} [23]. It is calculated in the following way.

Given a sample sequence $\{u_1, \ldots, u_N\}$, a template length m, and a threshold value r, first the overlapping sequences, $\{X_1^m, \ldots, X_{N-m}^m\}$ are constructed as for ApEn. As opposed to ApEn, C_i^m is now defined as the number of $j = 1, \ldots, N - m$, for which $d(X_i^m, X_j^m) < r$ where $i \neq j$, where $d(X_i^m, X_j^m)$ is again the Chebyshev distance. Applying the same for template length $m + 1$ results in:

$$\phi^m := \frac{1}{N - m} \sum_{i=1}^{N-m} \frac{C_i^m}{N - m - 1}, \text{ and} \quad (3)$$

$$\phi^{m+1} := \frac{1}{N - m} \sum_{i=1}^{N-m} \frac{C_i^{m+1}}{N - m - 1}. \quad (4)$$

Sample entropy is then defined as SampEn$(m, r) := \lim_{N \to \infty} (\ln \phi^m - \ln \phi^{m+1})$, which can be estimated by SampEn$(m, r, N) := \ln \phi^m - \ln \phi^{m+1}$.

Fuzzy entropy
ApEn and SampEn are very sensitive with respect to the threshold parameter r. They show, on one hand, a very abrupt behavior, and on the other hand less significance for small r, as was discussed in [9] and can be seen in figure 1. To soften these effects, Chen et al. developed fuzzy entropy, which uses a fuzzy membership function

instead of the Heaviside function [9]. FuzzyEn is calculated as follows.

Let m and r again be the template length and the threshold value and n a weight for the fuzzy membership function. Sequences $\{X_1^m, \ldots, X_{N-m+1}^m\}$ are defined as for ApEn and SampEn from an input sequence $\{u_1, \ldots, u_N\}$, with $X_i^m := \{u_i, \ldots, u_{i+m-1}\}$. Next, these sequences are transformed into sequences, $\{\overline{X}_1^m, \ldots, \overline{X}_{N-m+1}^m\}$, where $\overline{X}_i^m := \{ui - u0_i, \ldots, u_{i+m-1} - u0_i\}$ and $u0_i$ is the mean value of X_i^m, i.e.,

$$u0_i := \sum_{j=0}^{m-1} \frac{u_{i+j}}{m}. \quad (5)$$

Next, using a fuzzy membership function $x \to \mu(x, n, r)$ a membership matrix D^m is calculated, where each element is defined as $D_{i,j}^m := \mu(d(\overline{X}_i^m, \overline{X}_j^m), n, r)$. According to Chen et al. [9], fuzzy membership functions must be continuous and convex. The first property guarantees only slow similarity changes, and by the second condition self-similarity is a maximum of the function. For all tests conducted in this study, $x \to e^{-(x/r)^n}$ was chosen as the fuzzy membership function as in [9]. For $n \to \infty$ this fuzzy membership function converges to the Heaviside function. These steps are repeated for template length $m + 1$ to get D^{m+1}. Consequently, ϕ^m and ϕ^{m+1} are calculated the following way:

$$\phi^m := \frac{1}{N - m} \sum_{i=1}^{N-m} \sum_{j=1, j \neq i}^{N-m} \frac{D_{i,j}^m}{N - m - 1}, \text{ and} \quad (6)$$
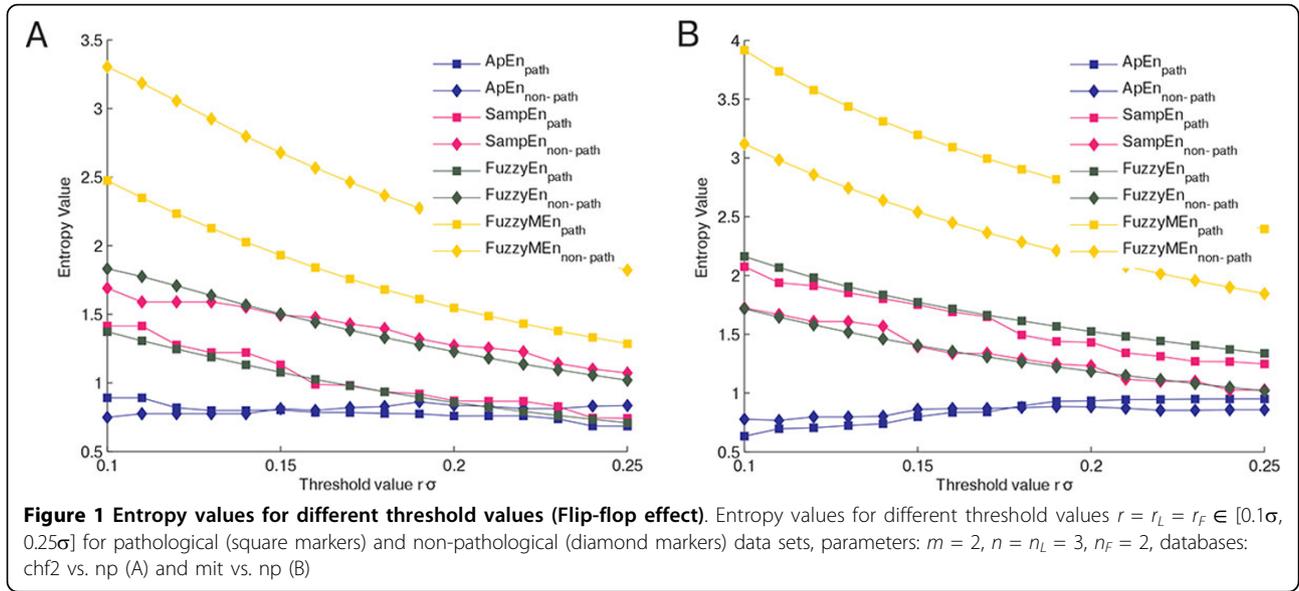
$$\phi^{m+1} := \frac{1}{N - m} \sum_{i=1}^{N-m} \sum_{j=1, j \neq i}^{N-m} \frac{D_{i,j}^{m+1}}{N - m - 1}. \quad (7)$$

Fuzzy entropy is defined as FuzzyEn$(m, r, n) := \lim_{N \to \infty} (\ln \phi^m - \ln \phi^{m+1})$ and can be estimated by FuzzyEn$(m, r, n, N) := \ln \phi^m - \ln \phi^{m+1}$.

Fuzzy measure entropy
Liu et al. proposed adding a function for global similarity to the fuzzy entropy and called the combination fuzzy measure entropy [24]. It can be calculated as follows.

Given a data sequence $\{u_1, \ldots, u_N\}$, a template length m, two threshold values r_L and r_F, and two weighting parameters n_L and n_F (r_L and n_L correspond to the local term and r_F and n_F to the global one), a sequence $\{X_1^m, \ldots, X_{N-m+1}^m\}$ is constructed like before. In the next step, it is transformed into a local sequence $\{XL_1^m, \ldots, XL_{N-m+1}^m\}$ and a global sequence, $\{XF_1^m, \ldots, XF_{N-m+1}^m\}$, with $XL_i^m := X_i^m - u0_i$ where $u0_i$ is defined as in (5) and $XF_i^m := X_i^m - u_{\text{mean}}$, and u_{mean} is the mean value of the complete data sequence $\{u_1, \ldots, u_N\}$.

Figure 1 Entropy values for different threshold values (Flip-flop effect). Entropy values for different threshold values $r = r_L = r_F \in [0.1\sigma,$ $0.25\sigma]$ for pathological (square markers) and non-pathological (diamond markers) data sets, parameters: $m = 2$, $n = n_L = 3$, $n_F = 2$, databases: chf2 vs. np (A) and mit vs. np (B)

Using the local and the global parameters for the fuzzy membership functions, the matrices DL^m and DF^M are defined as $DL_{i,j}^m := \mu(d(XL_i^m, XL_j^m), n_L, r_L)$ and $DF_{i,j}^m := \mu(d(XF_i^m, XF_j^m), n_F, r_F)$. Afterwards, all these steps are repeated for template length $m + 1$ to get DL^{m+1} and DF^{m+1}. Finally, $\phi_L^m, \phi_F^m, \phi_L^{m+1}$ and ϕ_F^{m+1} are defined as:

$$\phi_L^m := \frac{1}{N-m} \sum_{i=1}^{N-m} \sum_{j=1,j\neq i}^{N-m} \frac{DL_{i,j}^m}{N-m-1}, \tag{8}$$

$$\phi_F^m : \frac{1}{N-m} \sum_{i=1}^{N-m} \sum_{j=1,j\neq i}^{N-m} \frac{DF_{i,j}^m}{N-m-1}, \tag{9}$$

$$\phi_L^{m+1} := \frac{1}{N-m} \sum_{i=1}^{N-m} \sum_{j=1,j\neq i}^{N-m} \frac{DL_{i,j}^{m+1}}{N-m-1}, \text{ and} \tag{10}$$

$$\phi_F^{m+1} := \frac{1}{N-m} \sum_{i=1}^{N-m} \sum_{j=1,j\neq i}^{N-m} \frac{DF_{i,j}^{m+1}}{N-m-1}. \tag{11}$$

Fuzzy measure entropy is then defined as FuzzyMEn $(m, r_L, r_F, n_L, n_F) := \lim_{N\to\infty} (\ln \phi_L^m - \ln \phi_L^{m+1} + \ln \phi_F^m - \ln \phi_F^{m+1})$, which can be estimated by FuzzyMEn$(m, r_L, r_F, n_L, n_F, N) := (\ln \phi_L^m - \ln \phi_L^{m+1} + \ln \phi_F^m - \ln \phi_F^{m+1})$.

Parameter selection

As one can see in the description of each entropy, the various entropy types have three, four or six parameters. Thus, there are many possible combinations to choose from. The parameters which are varied in the test cases, their ranges and values mentioned in the literature, and the choice of certain fixed parameters are described here.

The most common choice for the template length is $m = 2$, as it was recommended by Pincus and Goldberger for ApEn [7], by Yentes et. al for SampEn [12], and confirmed by other studies, e.g. [20]. A false nearest neighbor method is also sometimes used, but according to Chon et al. the standard choice leads to the statistically best solutions for ApEn for human heart rate variability data [25]. In [9,24], the template length is set to $m = 2$ for other entropy types as well. For comparability, this value was used for all tests.

Some publications describe a sensitivity of the entropies to the data length N [1,6,8]. Therefore, the significance of the entropies calculated of parts of the heart rate variability data has been tested with increasing data set size.

For the threshold parameter r, Pincus suggested in [8] to choose a value between 0.1 σ and 0.25 σ, where σ is the sample standard deviation of the data sequence. This is also the standard range in most publications, with the most common choice of $r = 0.2 \sigma$ [17,21,23,24,26]. To examine threshold parameters in the standard range, they were tested with heart rate variability data.

In [19,27], the so called flip-flop effect is described, where for some values of the threshold value r a signal has a higher entropy compared to another and for other choices of r a lower one. This is also shown in [21] for heart rate variability data. To test for this effect, various entropies of signals from one database were calculated with parameters inside the standard range. Lu et al. also showed this effect in [26]. Therefore, they proposed choosing $r \in [0.1 \cdot \sigma, 1.0 \cdot \sigma]$ for ApEn in such a way as to

maximize the entropy value. Since finding a maximum is computationally very expensive, Chon et al. created in [25] an empirical formula to calculate an r, hereinafter called r_{Chon}, which approximates a maximizing r. For $m = 2$, it can be formulated as:

$$r_{\text{Chon}} := (-0.036 + 0.26\sqrt{\sigma_1/\sigma_2})/\sqrt[4]{N/1000}, \qquad (12)$$

where σ_1 is the standard deviation of the distances in the data sequence, i.e., the standard deviation of $\{(u_1 - u_2), \dots, (u_{N-1} - u_N)\}$, and σ_2 the standard deviation of the complete data sequence. This formula was derived from non-physiological data and Liu et al. showed in [17] that it is not always a good approximation of the maximizing r, but actually leads to more significant results than the maximizing r, when applied to heart rate variability data.

Regarding fuzzy function parameters, Chen et al. [9] used $n = 2$ and $r = 0.2\ \sigma$ for test signals. They described in [9] that for a larger n, the closer data points are weighted more strongly. Liu et al. [24] used the weighting factors $n_L = 3$, $n_F = 2$ and $r_L = r_F = 0.2\sigma$ for heart rate variability analysis. Their choices for n_L and n_F were given without any motivation.

Test cases
To get further knowledge of the parameters, heart rate variability data were used to compare different choices of r_L, r_F and n, n_L and n_F with respect to the resulting significance of the statistical tests comparing pathological and normal cases.

The following test cases using the data described in the Data section have been conducted to answer the research questions stated in the Background section and to support an "optimal" parameter selection for all entropy types for heart rate variability data. For each research question, one test case has been designed (except for the third, which is covered in two test cases: one for each entropy type under consideration). All tests have been conducted consecutively. Fixed parameters have been taken from literature or based on the outcomes of preceding tests.

Test case 1
Variation of the threshold values $r = r_L = r_F$ within the standard interval $[0.1\ \sigma, 0.25\ \sigma]$ [8] to show its influence on the entropy values and the aforementioned flip-flop effect [21,27]. Fixed parameters were $m = 2$, $n = n_L = 3$ and $n_F = 2$ [7,24]. For data length N, the data length of the shortest RR interval sequence of the available data $N = 1126$ sets was used.

Test case 2
Variation of the data length $N = x \cdot 110$ with $x \in [1, 10]$ to show its influence on the significance. Maximum N was 1100 due to the length of the available test data. Fixed parameters were $m = 2$, $n = n_L = 3$, $n_F = 2$ and $r = r_{\text{Chon}}$ or $r = 0.2\ \sigma$ [7,17,21,23-26].

Test case 3
Variation of the weighting factor n for FuzzyEn in the interval $[1, 6]$ to show its influence on the significance. Fixed parameters were $m = 2$ and $r = r_{\text{Chon}}$ or $r = 0.2\ \sigma$ [7,17,21,23-26]. $N = 1000$ was chosen based on the results of test case 2.

Test case 4
Variation of the weighting factors n_L and n_F for Fuzzy-MEn in the interval $[1, 6]$ to show their influence on the significance. Fixed parameters were $m = 2$ and $r_L = r_F = r_{\text{Chon}}$ or $r_L = r_F = 0.2\ \sigma$ [7,17,21,23-26]. $N = 1000$ was chosen based on the results of test case 2.

Test case 5
Variation of the threshold values r_L and r_F for Fuzzy-MEn in an interval of $[0.25 \cdot r_{\text{Chon}}, 6 \cdot r_{\text{Chon}}]$ and in an interval of $[0.1\ \sigma, 0.25\ \sigma]$[8] to show their influence on the significance. The parameter $m = 2$ was fixed [7]. $N = 1000$, n_L and n_F were chosen based on the results of the previous tests.

In order to test for their statistical significance, the calculated entropies were first tested for normality using Lilliefors' composite goodness-of-fit test [28]. If this test was positive for all results within a subset of a test case (i.e., certain database and/or choice for r), the p-value was calculated with a two sample t-test, and otherwise a Wilcoxon rank sum test was performed. To ensure comparability, the same statistical test was used for each subset of a test case.

Since statistical tests used in this work are based on the same null hypothesis, the same subject groups, the same endpoints and only slight variations of the analysis method, interaction of the observed results is not only possible, but highly probable. On the other hand, p-value adjustments such as the commonly used Bonferroni correction assume uncorrelated endpoints and are therefore considered inappropriate for the tasks in this work [29]. Besides, the aim of this work is not to test whether there is a difference between groups, but to investigate the ability of varied methods to detect those differences.

Due to the high computational complexity of the tests conducted (e.g., the calculation of the variation of the weighting factors n_L and n_F for FuzzyMEn takes several hours for only one r and one comparison of databases), the authors had to refrain from more robust randomized testing strategies (i.e., the permutation of the original data), since the computation time would multiply by at least a thousand times.

Data
All data used for the described tests have been taken from http://Physionet.org[30], a free-access, on-line archive of physiological signals. They are described in detail in this section.

To create a control group all databases described as non-pathological were combined into one database,

afterwards called *np*. This database contains the Normal Sinus Rhythm RR Interval Database, which consists of the beat annotations of 54 long-term ECG recordings, digitized at 128 samples per second, of subjects with normal sinus rhythm (30 men, aged 28.5 to 76, and 24 women, aged 58 to 73). Furthermore, it includes the MIT-BIH Normal Sinus Rhythm Database, which consists of 18 long-term recordings (5 men, aged 26 to 45, and 13 women, aged 20 to 50) digitized at 128 samples per second. The Fantasia Database of 120-minute recordings of twenty young (10 men and 10 women; 21 - 34 years old) and twenty elderly (10 men and 10 women; 68 - 85 years old) healthy subjects with ECG digitized at 250 Hz was also used [31]. The record "fantasia/f1o09" had to be excluded due to its high number of supraventricular premature beats. This results in a total database size of 111 recordings. RR intervals greater than 2.5 seconds were excluded to ignore artifacts.

Two databases were used to search for pathological effects. One was the Congestive Heart Failure RR Interval Database, afterwards referred to as *chf2*, which includes 29 long-term ECG recordings, with a sampling frequency of 128 Hz, of subjects aged 34 to 79 (8 men and 2 women; gender not known for the remaining subjects) with congestive heart failure (NYHA classes I, II, and III) [32].

The second one was the MIT-BIH Arrhythmia Database, afterwards called *mit*, which contains 48 half-hour recordings, sampled with a frequency of 360 Hz, from 47 subjects (25 men aged 32 to 89 years and 22 women aged 23 to 89 years) [33]. It contains a set of randomly chosen signals and 25 signals especially chosen to include examples of uncommon but clinically important arrhythmias recorded at the BIH Arrhythmia Laboratory [33].

The two databases *chf2* and *mit* were always evaluated separately to keep the results as homogeneous as possible and to avoid the mutual neutralization of ab- normalities. To ensure comparability, the data length N was equal for all recordings. Longer recordings were cropped at the beginning and the end in equal shares.

In a way, this work can be considered as a pilot study, as only previously recorded data are used. Furthermore, the findings of this work will be incorporated in follow-up studies.

Results

The following sections show the results of our tests, which are described in the Test cases section above. An overview of these results is given in Table 1.
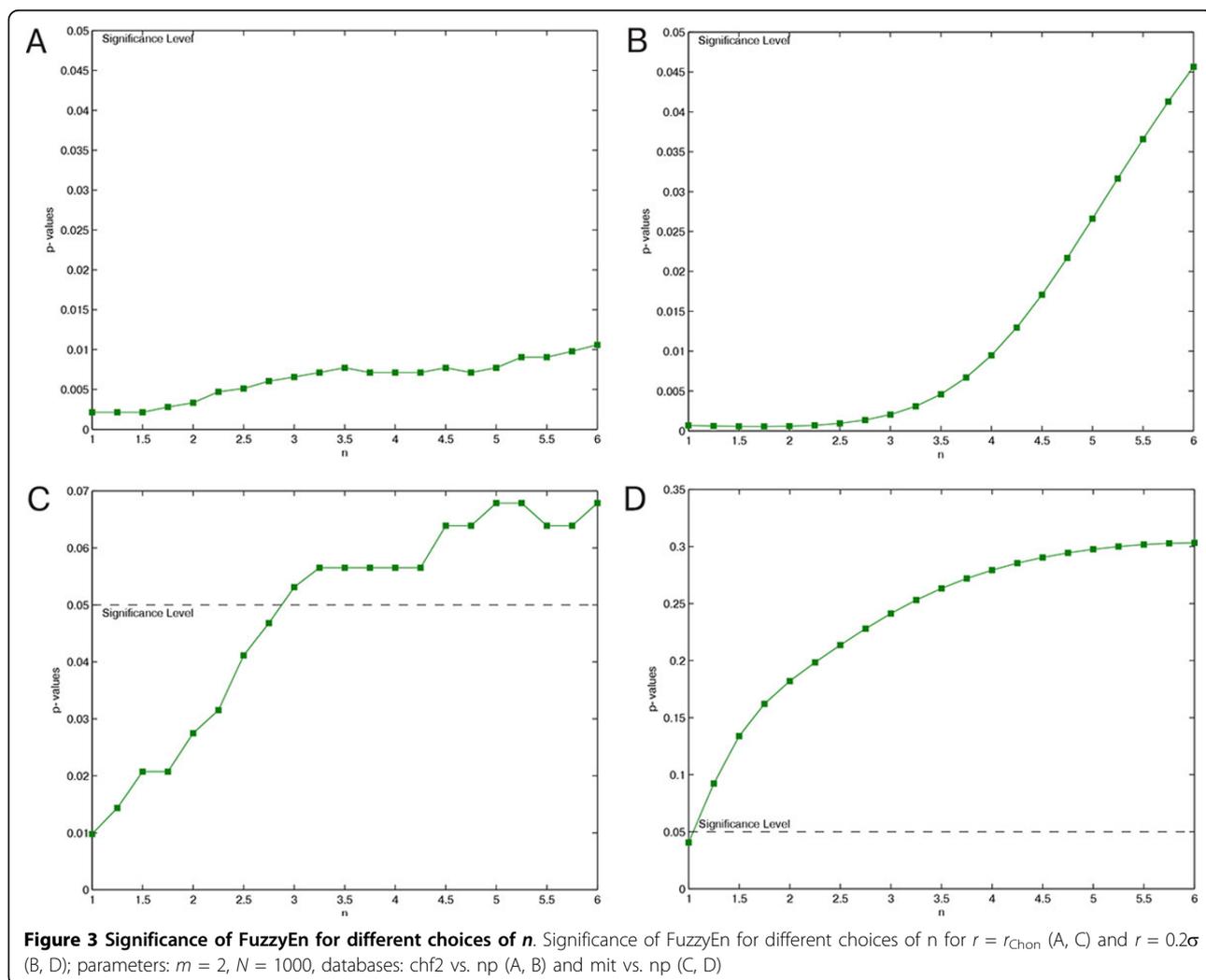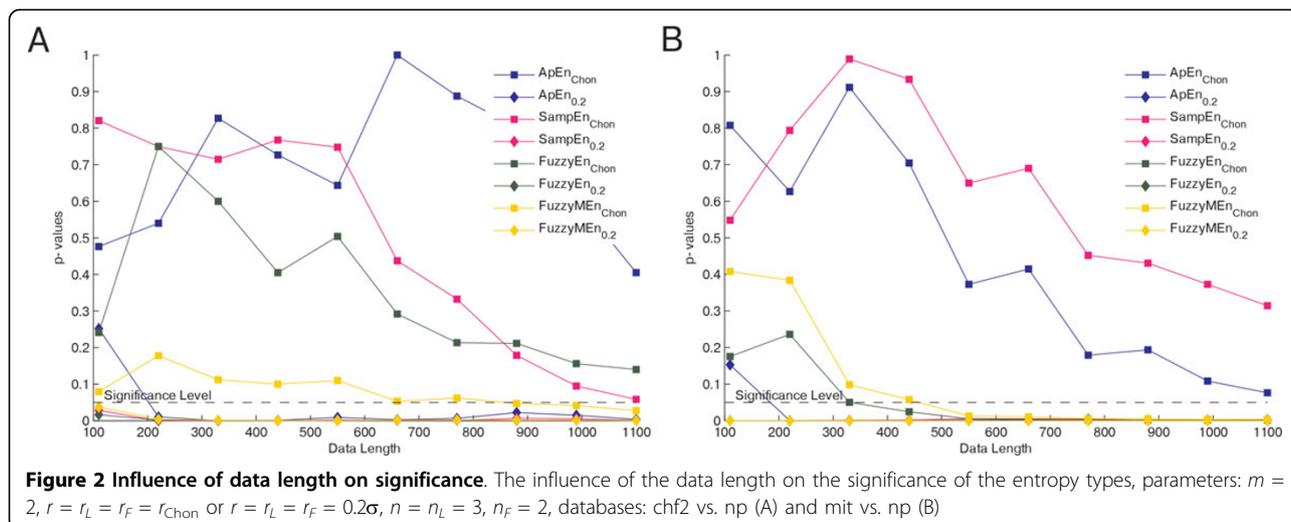
The results were tested for normality using Lilliefors' composite goodness-of-fit test. Throughout the whole test case 2, which combines results of two different ways to determine r (see figure 2), data were either normally or not normally distributed. Hence, entropy values were compared using the Wilcoxon rank sum test. Entropies calculated with r as multiple of σ were distributed normally in our test scenarios and could therefore be compared using the two sample t-test (as in figures 3 (B, D), 4 (B, D) and 5 (B, D)). Entropy values derived using r as multiple of r_{Chon}, on the other hand, were not normally distributed and hence compared using Wilcoxon's rank sum test (as in figures 3 (A, C), 4 (A, C) and 5 (A, C)). Only findings of the same statistical test are combined in the following sections.
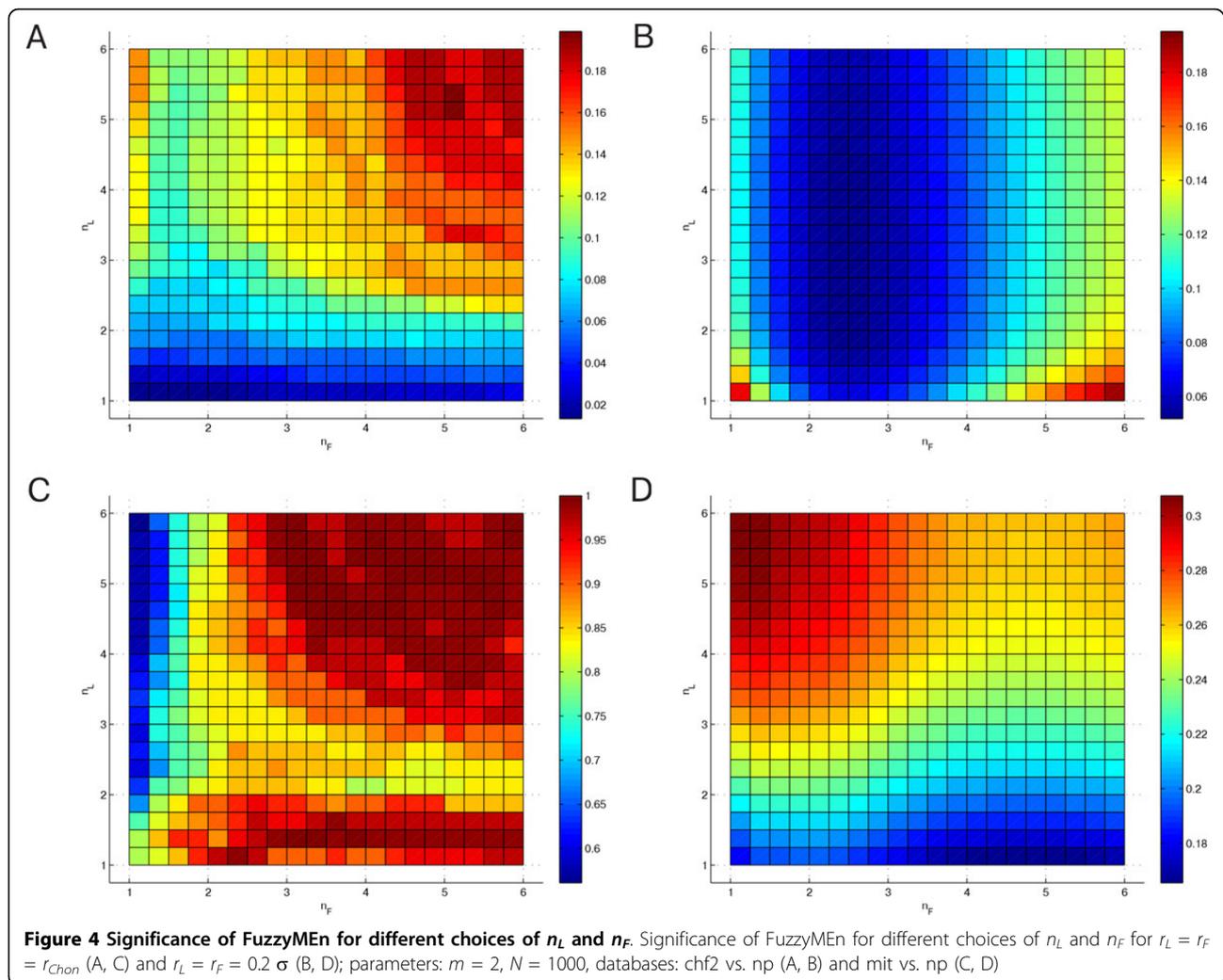
Figure 1 corresponds to test case 1 and shows the effect of different threshold values on the entropy values. One of the defining characteristics of FuzzyEn and FuzzyMEn, their relative insensitivity to changes in r, is clearly visible. The flip-flop effect can be observed for ApEn. For the *chf2* database, entropy values are higher for pathological data where $r < 0.15\ \sigma$, whereas they are lower for $r > 0.15\ \sigma$. This behavior is reversed for the *mit* database, with lower entropy values for pathological data for $r < 0.18\ \sigma$ and higher entropy values for pathological data for $r > 0.18\ \sigma$.

The results for test case 2, analyzing the sensitivity of the entropy types to the data length are represented in figure 2. The more data points evaluated, the higher the

Table 1 Summary of the best results achieved for each test case.

Test case number	Test case description	Results with	
		$r = r_{\text{Chon}}$	$r = 0.2\ \sigma$
1	Variation of the threshold values $r = r_L = r_F$ within the standard interval [0.1 σ, 0.25 σ]	Not applicable	Flip-flop effect verified for ApEn but not for other entropies
2	Variation of the data length $N = x \cdot 110$ with $x \in [1,10]$	Significant with $N > 1000$	Significant with $N > 200$
3	Variation of the weighting factor n for FuzzyEn in the interval [1,6]	Significant with $n < 3$	Significant with $n = 1$
4	Variation of the weighting factors n_L and n_F for FuzzyMEn in the interval [1,6]	Best results with $n_L = 2$, $n_F = 1$	Best results with $n_L = 1$, $n_F = 3$
5	Variation of r_L and r_F for FuzzyMEn in an interval of [0.25 · r_{Chon}, 6 · r_{Chon}] and in an interval of [0.1σ, 0.25σ]		inconclusive

Figure 2 Influence of data length on significance. The influence of the data length on the significance of the entropy types, parameters: $m = 2$, $r = r_L = r_F = r_{Chon}$ or $r = r_L = r_F = 0.2\sigma$, $n = n_L = 3$, $n_F = 2$, databases: chf2 vs. np (A) and mit vs. np (B)

Figure 3 Significance of FuzzyEn for different choices of *n*. Significance of FuzzyEn for different choices of n for $r = r_{Chon}$ (A, C) and $r = 0.2\sigma$ (B, D); parameters: $m = 2$, $N = 1000$, databases: chf2 vs. np (A, B) and mit vs. np (C, D)

Figure 4 Significance of FuzzyMEn for different choices of n_L and n_F. Significance of FuzzyMEn for different choices of n_L and n_F for $r_L = r_F = r_{Chon}$ (A, C) and $r_L = r_F = 0.2\ \sigma$ (B, D); parameters: $m = 2$, $N = 1000$, databases: chf2 vs. np (A, B) and mit vs. np (C, D)

separation between pathological and non-pathological data. When using the threshold value $r = 0.2\ \sigma$, significance is already reached with $N \geq 200$ data points, whereas with $r = r_{Chon}$ more data ($N \geq 1000$) are needed before significance is reached. Comparing the different methods when using $r = 0.2\ \sigma$, one can see that they only differ when N is very small. With higher N, their behavior converges.

Figure 3 presents the results of test case 3, showing that the results become insignificant for $n > 3$ with $r = r_{Chon}$, and for $n > 1$ with $r = 0.2\ \sigma$ in the *mit* database (C, D). The increasing p-values are below the significance level ($p < 0.05$) for the *chf2* database within the observed range of n (A, B).

Figure 4, corresponding to test case 4, shows a higher significance for $n_L \leq 2$ and all values of n_F (A), or in the case of $r = 0.2\ \sigma$ for $n_F \in [2, 3.5]$ (B). In figure 4 (C and D) the situation is different: the best results are achieved with $n_F < 1.5$ and $n_L \geq 2$ for $r = r_{Chon}$ in case C, whereas for $r = 0.2\ \sigma$ (D) the best performance is reached with $n_L < 1.5$ and any n_F.

The results for test case 5 are shown in figure 5. They represent the ability of FuzzyMEn to differentiate between *chf2* and *np* (A and B) and *mit* and *np* (C and D) for different choices of the threshold values r_F and r_L, when chosen as multiples of r_{Chon} in the range of $[0.25 \cdot r_{Chon}, 6 \cdot r_{Chon}]$ (A, C), and calculated according to the results of test case 4 with $n_L = 2$ and $n_F = 1$, or inside the standard range of $[0.1\ \sigma, 0.25\ \sigma]$ (B, D), calculated with $n_L = 1$ and $n_F = 3$. In (A), the best results are achieved with $r_L \lesssim 1 \cdot r_{Chon}$ or $r_F \lesssim 1 \cdot r_{Chon}$, whereas they have to be greater than or equal to $1 \cdot r_{Chon}$ in (C) for good performance. In (B), $r_L < 0.12\ \sigma$ or $r_L \gtrsim 0.2\sigma$ yield significant results, however rF does not matter that much. In contrast, the best performance in (D) is achieved with $r_F \gtrsim 0.18\sigma$ and $r_L \in [0.14\ \sigma, 0.22\ \sigma]$.

Discussion

As a summary of the following discussion of the previously presented results, the combinations of parameters which yielded the best results are listed in Table 2.

Figure 5 Significance of FuzzyMEn for different choices of r_L and r_F. Significance of FuzzyMEn for different choices of r_L and r_F in an interval of $[0.25 \cdot r_{\text{Chon}}; 6 \cdot r_{\text{Chon}}]$ with $n_L = 2$, $n_F = 1$ (A, C) and in an interval of $[0.1\sigma, 0.25\sigma]$ with $n_L = 1$, $n_F = 3$ (B, D); parameters: $m = 2$, $N = 1000$, databases: chf2 vs. np (A, B) and mit vs. np (C, D)

Some of the tests concerning the parameter selection showed no clear results. A couple of evaluations revealed contradicting results depending on the database (*chf2* vs. *mit*) or on the chosen parameters (e.g., $r = r_L = r_F = r_{\text{Chon}}$ vs. $r = r_L = r_F = 0.2\sigma$). The latter is easy to overcome by choosing a different set of parameters based on the choice of the threshold value r. However, a compromise must be found in order to find parameters suitable for both databases.

One of the hardest difficulties lies in the choice of the threshold value r due to the flip-flop effect, i.e., for some parameters one data set has a higher entropy compared to another, but this order is reversed for different parameter choices [17,27]. This can occur for simple signals, but also when analyzing heart rate variability data, as in figure 1. This leads to difficulties with the interpretation of the entropy, i.e., the direct assignment of entropy values to pathological or non-pathological data without a given r. In our tests, the effect occurred for ApEn, but it

is also reported to occur for SampEn and FuzzyEn as well by Boskovic et al. [27]. The flip-flop effect does not allow us to make a clear statement, as in [7-9,17], that a higher entropy corresponds to a higher irregularity. Since this effect can happen for all different entropy values [27], they should not be seen as a direct measure of regularity, but rather as an indicator of differences in regularity with regard to certain time periods.

Compared to the threshold value, the data length N seems to have a smaller effect on the ability of the entropy measures to differentiate pathological from non-pathological data sets. Generally, a larger N leads to a higher probability of significance. If possible, the data length N should be longer than 200 data points when using $r = 0.2\,\sigma$. This finding is consistent with previous studies, e.g. [12]. Surprisingly, it should even exceed a length of 1000 data points when using $r = r_{\text{Chon}}$, assuming a continuing trend in figure 2. Due to the length of the available test data, the range had to be restricted for this test. In case of HRV

Table 2 Recommendations for different combinations of parameters.

$r = r_L = r_F$	m	N	$n = n_L$	n_F
r_{Chon}	2	>1000	2	1
$0.2\,\sigma$	2	>200	1	3

data, N is the number of recorded heartbeats and therefore proportional to the duration of the recording. Thus, to increase N, longer measurements are necessary. The Task Force of The European Society of Cardiology and The North American Society of Pacing and Electrophysiology [4] recommends a duration of five minutes for short time recordings, which would result in 300 data points at an average heart rate of 60 beats per second. This would be sufficient when using $r = 0.2\,\sigma$, but not for $r = r_{Chon}$. These considerations should be kept in mind when dealing with HRV data.

Due to the different behavior when varying n, n_L and n_F given different threshold parameters $r = r_L = r_F = r_{Chon}$ and $r = r_L = r_F = 0.2\,\sigma$, the parameters n, n_F and n_L have been chosen independently for the different threshold values. This is no constraint to the method, since the choice of r is known beforehand anyway (and is not based on the potentially unknown medical condition of the subject). The values $n_L = 2$ and $n_F = 1$ for $r_L = r_F = r_{Chon}$, and $n_L = 1$ and $n_F = 3$ for $r_L = r_F = 0.2\,\sigma$ showed better results than $n_L = 3$, $n_F = 2$ as proposed by Liu et al. in [24]. Unsurprisingly, similar values were found for n, as n for FuzzyEn equals n_L for FuzzyMEn. For consistency, $n = n_L = 2$ for $r = r_{Chon}$ and $n = n_L = 1$ for $r = 0.2\,\sigma$ are recommended.

The tests concerning r_F and r_L showed that there is no optimal choice, since the results in figure 5 (A) and (C) as well as (B) and (D) contradict each other. Nevertheless, both $r_F = r_L = r_{Chon}$ and $r_F = r_L = 0.2\,\sigma$, as described in the literature [9,24,25], are the most reasonable compromise between figure 5 (A) and (C), and (B) and (D), respectively.

Finally, a number of important limitations of this study need to be considered. First, this study is limited to previously recorded signals of only two different cardiac diseases due to the availability of data. As already mentioned in the Data section, a separate evaluation is warranted, to avoid the mutual neutralization of abnormalities. Furthermore, the template length was fixed to $m = 2$ for all calculations, as the number of possible variations of parameters would get too high otherwise and $m = 2$ seems to be a reasonable choice in all investigated literature, e.g., [8]. In [34] it was reported that spikes due to recording errors in heart rate variability data can disturb ApEn and SampEn. No tests were done to examine this behavior for the included entropies. Finally, the study did not evaluate the dependency of the entropies on age and gender as reported in literature [35,36].

Conclusions

The results of the presented study clearly stress the need for further investigations of signal entropy for heart rate variability data. Given the wide range of different medical conditions of subjects, the assortment of available methods to calculate entropy, and their customizability with up to six degrees of freedom, it is almost impossible to cover all combinations in a single study. Future work will therefore be focused on overcoming the limitations of the presented work, i.e., extending the evaluations to other cardiac diseases, the variation of the template length m, investigating the robustness with respect to recording errors, and the relation of the parameter choice to age and gender.

Authors' contributions
CM and MB have equally contributed to this work. CM participated in the design and coordination of the study, the interpretation of the results and drafted the manuscript. MB carried out the calculations and interpretation of the results and participated in drafting the manuscript. MH carried out the implementation of methods, calculations and interpretations. CS participated in the implementation. AH helped to draft the manuscript. SW participated in the design of the study and drafting the manuscript. All authors read and approved the final manuscript.

Authors' details
[1]AIT Austrian Institute of Technology, Health & Environment Department, Biomedical Systems, Donau-City-Str. 1, 1220 Vienna, Austria. [2]Vienna University of Technology, Institute for Analysis and Scientific Computing, Wiedner Hauptstr. 8-10, 1040 Vienna, Austria. [3]Medical University Graz, Institute of Medical Informatics, Auenbruggerplatz 2/9/V, 8036 Graz, Austria.

References
1. Rajendra Acharya U, Paul Joseph K, Kannathal N, Lim C, Suri J: **Heart rate variability: a review.** *Med Biol Eng Comput* 2006, **44**:1031-1051, doi:10.1007/s11517-006-0119-0.
2. Batchinsky AI, Salinas J, Cancio LC: **Assessment of the Need to Perform Life-Saving Interventions Using Comprehensive Analysis of the Electrocardiogram and Artificial Neural Networks.** *RTO-MP-HFM-182: Use of Advanced Technologies and New Procedures in Medical Field Operations* 2010, 39-116.
3. Bachler M, Mayer C, Hametner B, Wassertheurer S, Holzinger A: **Online and offline determination of qt and pr interval and qrs duration in electrocardiography.** In *Pervasive Computing and the Networked World Lecture Notes in Computer Science. Volume 7719.* Springer, Berlin Heidelberg; Zu, Q., Hu, B., El¸ci, A 2013:1-15.
4. American Heart Association Inc, European Society of Cardiology: **Guidelines - heart rate variability.** *Eur Heart J* 1996, **17**:354-381.
5. Pincus SM: **Assessing serial irregularity and its implications for health.** *Annals of the New York Academy of Sciences* 2001, **954**:245-267.
6. Hornero R, Aboy M, Abasolo D, McNames J, Goldstein B: **Interpretation of approximate entropy: analysis of intracranial pressure approximate**

entropy during acute intracranial hypertension. *IEEE Trans Biomed Eng* 2005, **52**(10):1671-1680.

7. Pincus SM, Goldberger AL: **Physiological time-series analysis: what does regularity quantify?** *Am J Physiol* 1994, **266**(4 Pt 2):1643-1656.

8. Pincus SM: **Approximate entropy as a measure of system complexity.** *Proc Natl Acad Sci USA* 1991, **88**(6):2297-2301.

9. Chen W, Zhuang J, Yu W, Wang Z: **Measuring complexity using FuzzyEn, ApEn, and SampEn.** *Med Eng Phys* 2009, **31**(1):61-68.

10. Hornero R, Aboy M, Abasolo D, McNames J, Wakeland W, Goldstein B: **Complex analysis of intracranial hypertension using approximate entropy.** *Crit Care Med* 2006, **34**(1):87-95.

11. Sarlabous L, Torres A, Fiz JA, Gea J, Martínez-Llorens JM, Morera J, Jané R: **Interpretation of the approximate entropy using fixed tolerance values as a measure of amplitude variations in biomedical signals.** *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE* 2010, 5967-5970, doi:10.1109/IEMBS.2010.5627570.

12. Yentes J, Hunt N, Schmid K, Kaipust J, McGrath D, Stergiou N: **The appropriate use of approximate entropy and sample entropy with short data sets.** *Annals of Biomedical Engineering* 2013, **41**(2):349-365, doi:10.1007/s10439-012-0668-3.

13. Acharya UR, Molinari F, Sree SV, Chattopadhyay S, Ng KH, Suri JS: **Automated diagnosis of epileptic eeg using entropies.** *Biomedical Signal Processing and Control* 2012, **7**(4):401-408.

14. Roerdink M, De Haart M, Daffertshofer A, Donker SF, Geurts AC, Beek PJ: **Dynamical structure of center-of-pressure trajectories in patients recovering from stroke.** *Exp Brain Res* 2006, **174**(2):256-269.

15. Kaplan DT, Furman MI, Pincus SM, Ryan SM, Lipsitz LA, Goldberger AL: **Aging and the complexity of cardiovascular dynamics.** *Biophysical Journal* 1991, **59**(4):945-949, doi:10.1016/S-0006-3495(91)82309-8.

16. Holzinger A, Stocker C, Bruschi M, Auinger A, Silva H, Gamboa H, Fred ALN, Huang R, Ghorbani AA, Pasi G, Yamaguchi T, Yen NY, Jin B: **On applying approximate entropy to ecg signals for knowledge discovery on the example of big sensor data.** In *AMT Lecture Notes in Computer Science. Volume 7669.* Springer, Berlin Heidelberg; 2012:646-657.

17. Liu C, Liu C, Shao P, Li L, Sun X, Wang X, Liu F: **Comparison of different threshold values r for approximate entropy: application to investigate the heart rate variability between heart failure and healthy control groups.** *Physiol Meas* 2011, **32**(2):167-180.

18. Ramdani S, Bouchara F, Lagarde J: **Influence of noise on the sample entropy algorithm.** *Chaos: An Interdisciplinary Journal of Nonlinear Science* 2009, **19**(1), doi:10.1063/1.3081406.

19. Chen W, Wang Z, Xie H, Yu W: **Characterization of surface emg signal based on fuzzy entropy.** *Neural Systems and Rehabilitation Engineering, IEEE Transactions on* 2010, **15**(2):266-272, doi:10.1109/TNSRE.2007.897025.

20. Alcaraz R, Abásolo D, Hornero R, Rieta JJ: **Optimal parameters study for sample entropy-based atrial fibrillation organization analysis.** *Computer Methods and Programs in Biomedicine* 2010, **99**(1):124-132, doi:10.1016/j.cmpb.2010.02.009.

21. Castiglioni P, Di Rienzo M: **How the threshold "r" influences approximate entropy analysis of heart-rate variability.** *IEEE* 2008, 561-564.

22. Lake DE, Richman JS, Griffin MP, Moorman JR: **Sample entropy analysis of neonatal heart rate variability.** *American Journal of Physiology - Regulatory, Integrative and Comparative Physiology* 2002, **283**(3):789-797[http://ajpregu.physiology.org/content/283/3/R789.full.pdf], doi:10.1152/ajpregu.00069.2002..

23. Richman JS, Moorman JR: **Physiological time-series analysis using approximate entropy and sample entropy.** *Am J Physiol Heart Circ Physiol* 2000, **278**(6):2039-2049.

24. Liu C, Li K, Zhao L, Liu F, Zheng D, Liu C, Liu S: **Analysis of heart rate variability using fuzzy measure entropy.** *Comput Biol Med* 2013, **43**(2):100-108.

25. Chon K, Scully CG, Lu S: **Approximate entropy for all signals.** *IEEE Eng Med Biol Mag* 2009, **28**(6):18-23.

26. Lu S, Chen X, Kanters JK, Solomon IC, Chon KH: **Automatic selection of the threshold value R for approximate entropy.** *IEEE Trans Biomed Eng* 2008, **55**(8):1966-1972.

27. Boskovic A, Loncar-Turukalo T, Japundzic-Zigon N, Bajic D: **The flip-flop effect in entropy estimation.** *IEEE* 2011, 227-230.

28. Lilliefors HW: **On the kolmogorov-smirnov test for normality with mean and variance unknown.** *Journal of the American Statistical Association* 1967, **62**(318):399-402[http://amstat.tandfonline.com/doi/pdf/10.1080/01621459.1967.10482916], doi:10.1080/01621459.1967.10482916.

29. Altman DG, Machin D, Bryant TN, Gardner MJ, Gardner S, Bird S, Campbell M, Daly LE, Morris J: **Statistics with confidence: confidence intervals and statistical guidelines.** 2000.

30. Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE: **PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals.** *Circulation* 2000, **101**(23):215-220.

31. Iyengar N, Peng C, Morin R, Goldberger A, Lipsitz LA: **Age-related alterations in the fractal scaling of cardiac interbeat interval dynamics.** *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology* 1996, **271**(4):1078-1084.

32. Baim DS, Colucci WS, Monrad ES, Smith HS, Wright RF, Lanoue A, Gauthier DF, Ransil BJ, Grossman W, Braunwald E: **Survival of patients with severe congestive heart failure treated with oral milrinone.** *J Am Coll Cardiol* 1986, **7**(3):661-670.

33. Moody GB, Mark RG: **The impact of the MIT-BIH arrhythmia database.** *IEEE Eng Med Biol Mag* 2001, **20**(3):45-50.

34. Molina-Picó A, Cuesta-Frau D, Aboy M, Crespo C, Miró-Martínez P, Oltra-Crespo S: **Comparative study of approximate entropy and sample entropy robustness to spikes.** *Artif Intell Med* 2011, **53**(2):97-106.

35. Beckers F, Verheyden B, Aubert AE: **Aging and nonlinear heart rate control in a healthy population.** *American Journal of Physiology-Heart and Circulatory Physiology* 2006, **290**(6):2560-2570.

36. Ryan SM, Goldberger AL, Pincus SM, Mietus J, Lipsitz LA: **Gender- and age-related differences in heart rate dynamics.** *Are women more complex than men? Journal of the American College of Cardiology* 1994, **24**(7):1700-1707, doi:10.1016/0735-1097(94)90177-5.