Compression entropy contributes to risk stratification in patients with cardiomyopathy

Kompressionsentropie zur verbesserten Risikostratifizierung bei Patienten mit DCM

Sandra Truebner¹, Iwona Cygankiewicz², Rico Schroeder³, Mathias Baumert³,Montserrat Vallverdú⁴, Pere Caminal⁴, Rafael Vazquez⁵, Antoni Bayés de Luna⁶ and Andreas Voss¹*,

¹ Department of Medical Engineering, University of Applied Sciences Jena, Jena, Germany
² MUSIC Trial, Institute of Cardiology, Medical University of Lodz, Lodz, Poland
³ Department of Medical Engineering, University of Applied Sciences Jena, Jena, Germany
⁴ Biomedical Engineering Research Centre, Technical University of Catalonia, Barcelona, Spain
⁵ MUSIC Trial, University of Sevilla, Seville, Spain
⁶ MUSIC Trial, Hospital de la Santa Creu i Sant Pau, Barcelona, Spain

Abstract

Sudden cardiac death (SCD) is a leading cause of mortality with an incidence of 3 million cases per year worldwide. Therapies for patients who have survived an SCD episode or have a high risk of developing lethal ventricular arrhythmia are well established and depend mainly on risk stratification. In this study we investigated the suitability of the non-linear measure compression entropy (H_C) for improved risk prediction in cardiac patients. We recorded 24-h Holter ECG for 300 patients with congestive heart failure (CHF). During a mean follow-up period of 12 months, 32 patients died due to a cardiac event. H_C depends on the compression parameters window length w and buffer length b, which were optimised by analysing a subgroup of patients. Compression entropies based on the beat-to-beat interval (BBI) were subsequently calculated and compared with standard heart-rate variability parameters. Statistical analysis revealed significant differences between high- and low-risk CHF patients in standard HRV measures, as well as compression entropy based on the BBI (cardiac death, p=0.005; SCD, p=0.02). In conclusion, the implementation of non-linear compression entropy analysis in multivariate analysis seems to be useful for enhanced risk stratification of cardiac death, especially SCD, in ischaemic cardiomyopathy patients.

Keywords: cardiomyopathy; compression entropy; non-linear analysis; risk stratification; sudden cardiac death.

Zusammenfassung

Der plötzliche Herztod (SCD) ist die Haupttodesursache weltweit (3 Millionen Fälle/Jahr). Moderne Methoden zur Therapie und Prävention des SCD sind abhängig von der Erkennung der Hochrisikopatienten. Das Ziel dieser Studie war die Untersuchung der Eignung des nichtlinearen Parameters der Kompressionsentropie (H_C) zur Risikostratifizierung bei ischämischer Herzensuffizienz (CHF). Von 300 CHF-Patienten wurden 24-h Holter-EKGs im Rahmen einer spanischen Multicenter-Studie (MUSIC) aufgezeichnet. Innerhalb der anschließenden Follow-up-Phase (12 Monate) verstarben 32 Patienten aufgrund eines kardialen Ereignisses (Hochrisikogruppe). Mittels einer Patientenuntergruppe wurden die in die H_C-Analyse eingehenden Parameter Fenster- und Bufferlänge optimiert. Zusätzlich zur Berechnung von H_C wurden die Standardparameter der Herzfrequenzvariabilität (HRV) bestimmt. Die statistische Analyse zeigte signifikante Unterschiede zwischen CHF-Patienten mit hohem und niedrigem Risiko in den Standardparametern der HRV (kardialer Tod: p=0,02; SCD: p=0,04) sowie Parametern der H_C (kardialer Tod: p=0,005; SCD: p=0,02). Diese Ergebnisse zeigen die prinzipielle Eignung der H_C für die Risikoanalyse des kardialen Todes insbesondere des plötzlichen Herztodes bei Patienten mit ischämischer Kardiomyopathie. Durch eine anschließende multivariate Analyse dieses nichtlinearen Parameters soll die Verbesserung der Ergebnisse bezüglich Sensitivität und Spezifität bestätigt werden.

Schlüsselwörter: Kardiomyopathie; Kompressionsentropie; nichtlineare Analyse; plötzlicher Herztod; Risikostratifizierung.

Introduction

Heart failure is recognised as a major and escalating public health problem in industrialised countries with ageing populations. The overall prevalence of clinically identified heart failure is estimated at up to 20 cases/1000 population, but rises to >100 cases/1000 population in those aged >65 years. The prevalence of confirmed left-ventricular systolic dysfunction also increases with age and is more common in men [7]. Sudden cardiac death (SCD) is a leading cause of mortality with an incidence of 3 million cases per year worldwide and may occur at any stage [9].
Therapies for patients who have survived an SCD episode or have a high risk of developing lethal ventricular arrhythmia are now well established, but strategies such as the application of an implantable cardioverter defibrillator (ICD) depend mainly on risk stratification. Various methods based on non-linear systems theory have been proposed to obtain information about physiological and pathological cardiovascular regulation. Several studies were performed to investigate the usefulness of clinical application of linear and non-linear heart-rate variability analysis for physiological interpretation and risk stratification in cardiac diseases [1, 3–5, 10, 13, 16]. However, the problem of SCD risk stratification in patients still remains.

In this study we investigated the suitability of the non-linear compression entropy method based on beat-to-beat interval (BBI) time series, BBI after mean subtraction, and differences between successive BBIs for non-invasive risk stratification in congestive heart failure (CHF) patients.

Materials and methods

Data and pre-processing

A total of 300 CHF patients with sinus rhythm were enrolled in the Spanish MUSIC (Muerta Subita en Insuficiencia Cardiaca, sudden death in heart failure) study, a multicentre study designed to assess risk predictors of sudden cardiac death in patients with heart failure characterised by class II–III of the New York Heart Association (NYHA) classification system. The majority of these patients were in NYHA class II (216 patients, 80%). Further criteria for patient enrolment were a time-lag of more than 3 months after the last hospitalisation due to heart failure decompensation and left ventricular ejection fraction (LVEF) <40%, left ventricular diastolic diameter (LVDD) >60 mm or abnormal relaxation patterns characteristics for diastolic dysfunction. Patients with sustained ventricular tachycardia, atrial fibrillation, or ICD, as well as ICD implantation or heart transplantation during the follow-up period, were excluded from the study. For all 300 patients, 24-h Holter ECG with a sampling frequency of 200 Hz (Syneflash-MMC; ELA Medical, Barcelona, Spain) was recorded.

Ventricular premature beats and artefacts were interpolated in the BBI series to obtain normal-to-normal BBI. Filtering was performed by applying an adaptive variance estimation algorithm, considering the variance within the time series just before and directly after an ectopic beat [15]. In this regard, patients with a percentage of artefacts and ectopic beats greater than 10% were excluded from the analysis (n=30) to reduce the influence of filtering on the time series and the analysis results.

Of the patients, 59 suffered from idiopathic cardiomyopathy (CM), 146 had ischaemic CM caused by coronary artery disease, and 65 patients suffered mainly from alcoholic, hypertrophic and hypertensive CM. In 119 patients, diabetes mellitus had been diagnosed. All patients received optimal medical treatment with ACE inhibitors (74%), beta-blockers (66%) and diuretics (72%). During the median follow up-period of 24 months (range 11–27 months), 121 ischaemic patients remained in stable physiological conditions (low-risk group, LR) and 32 patients died due to a cardiac event (ischaemic CM; n=25, high-risk group, HR), whereas 16 patients suffered from SCD (ischaemic CM; n=14, high-risk group, HRS).

Methodology

In 1977 Ziv and Lempel developed a universal algorithm for lossless data compression [17]. Lossless compression ensures that the original information can be exactly reproduced from the compressed data. This algorithm is widely used in compression utilities such as gzip, GIF image compression and the V.42 modem standard.

LZ77 algorithm

The Ziv and Lempel algorithm LZ77 employs a principle called a sliding-window when a sequence of symbols $x=x_1,x_2,x_3,...,x_n$ has to be compressed.

![Figure 1](https://via.placeholder.com/150)

**Figure 1** Scheme of the compression algorithm.

The window ($w=3$) and look-ahead buffer ($b=3$) are shifted through a data stream. The parameters pos, len and next are stored in a three-column matrix.
The algorithm keeps the most recently encoded symbols in a sliding window of size \( w \). The not-yet-encoded sequence of symbols is stored in a look-ahead buffer of size \( b \) (Figure 1). The encoder \( p \) positioned at the first string in look-ahead buffer \( x_p \) looks for the longest string match of length \( len \) between the not-yet-encoded \( n \)-string \( x_n \) and the strings in the recently encoded window. If \( x_p \) matches with strings in the sliding window, the position \( pos \) of the beginning string match in the sliding window, the length \( len \) of the string match, and the first string after the string-matching sequence (next) in the look-ahead buffer are stored in a three-column matrix. The sliding window and look-ahead buffer are displaced by \( len + 1 \) samples. If the strings of the look-ahead buffer do not match with the string sequence in the sliding window with \( pos = 0 \), \( len = 0 \) and next \( = x_p \), the sliding window and buffer are displaced by 1 sample. At the end of the procedure, an original data series \( x \) of length \( L \) is compressed into a three-column matrix of length \( M \). In terms of information theory, the smallest algorithm that produces a string is the entropy of that string, the Chain-Kolmogorov entropy [2, 6]. Although it is theoretically impossible to develop such an algorithm, data compression techniques might represent a good approximation. Assuming that the source is an ergodic process, the entropy \( H_c \) of the compressed string is determined as the length \( M \) of the compressed string divided by the length \( L \) of the original data series:

\[
H_c = \frac{M}{L}. \tag{1}
\]

**Compression entropy for heart rate**

The modified LZ77 algorithm was used for the analysis of BBI long-term heart-rate time series. In addition, the differences between successive BBIs, as well as the beat-to-beat time series after subtraction of the mean, were analysed. Thus, the algorithm was applied to three different variations of one BBI time series \( (x, x_{mv}, x_{ab}) \):

- **BBI without any modification:**
  \[
x = \text{BBI}_1, \text{BBI}_2, \text{BBI}_3, \ldots, \text{BBI}_L
\]

- **BBI after mean subtraction:**
  \[
x_{mv} = \text{BBI}_1 - \text{BBI}_1, \ldots, \text{BBI}_L - \text{BBI}_L
\]

- **Differences of consecutive BBIs:**
  \[
x = \text{BBI}_2 - \text{BBI}_1, \ldots, \text{BBI}_L - \text{BBI}_{L-1}
\]

where \( \text{BBI} \) is the mean BBI value.

The alphabet analysed for each data stream is affected by the sampling rate. For this reason, the implementation of compression entropy \( H_c \) for heart-rate time series has to consider integer numbers. Furthermore, \( H_c \) is mainly influenced by window length \( w \) and look-ahead buffer length \( b \). Considering the parameters \( w \) and \( b \) that affect it, the compression entropy \( H_c \) is denoted as:

\[
H_c^{w,b} \text{ for time series } x; \quad H_c^{mv} \text{ for BBI time series after mean subtraction; and } H_c^{ab} \text{ for times series based on differences.}
\]

**Surrogate analysis**

To verify that compression entropy is a non-linear method, a surrogate data test was carried out. Assuming the hypothesis that the time series includes only linear components, 50 linear realisations from two different original BBI time series (I, a healthy subject; II, a high-risk CHF patient) – so-called surrogates – were synthesised [12]. Compression entropy \( H_c^{w,b} \) in different window-buffer combinations and the mean value and standard deviation of the original time series of the healthy subject (X) and the CHF patient (X) and of the respective surrogate sets (S, S) were quantified and analysed using discriminant statistics. The statistical deviation between the original time series and the surrogates indicates the presence of a non-linear structure in the original time series.

**Optimisation of \( w \) and \( b \)**

According to the influence of the window and buffer length, the compression parameters \( w \) and \( b \) were incremented stepwise and compression entropies \( H_c^{w,b} \), \( H_c^{mv} \) and \( H_c^{ab} \) were calculated for a subset of 10 high-risk and 10 low-risk patients. The ranges investigated for window and buffer length were \( 2 \leq w \leq 200 \) and \( 2 \leq b \leq 100 \), respectively.

**Standard HRV analysis**

Previous studies have shown that some standard measures of heart rate variability (HRV) provide independent prognostic information for CHF patients [8, 14]. A set of time-domain and frequency-domain standard HRV parameters [11] was calculated (Table 1) to compare the efficiency of compression entropy with standard HRV measures. The frequency analysis was performed with linear interpolated time series (resolution 100 ms) after mean subtraction using fast Fourier transformation with Blackman Harris windowing.

**Statistical analysis**

The Mann-Whitney U-test for skewed variables was applied for surrogate analysis, as well as optimisation of window and buffer length for best discrimination between high-risk (cardiac death, SCD) and low-risk CHF patients. Univariate differences between ischaemic and idiopathic patients were evaluated. Pearson’s correlation coefficient and significance were computed for all clinical measures.
and HRV parameters, as well as compression entropies. Owing to the limited number of high-risk idiopathic patients (cardiac death $n = 3$; SCD $n = 1$), the suitability of compression entropy parameters for risk stratification (cardiac death and SCD) was investigated only in ischaemic patients. For the variable follow-up duration, a Cox regression model was applied for multivariate analysis and to illustrate the univariate effectiveness of compression entropy in risk stratification.

**Results and discussion**

**Surrogate analysis**

Comparison of meanNN and sdNN for the surrogate sets (consisting of 50 surrogates from one original time series) with the values for the original time series showed no significant differences between surrogate sets $S_i$, $S_s$ and the original BBI $X_i$, $X_s$ (Table 2).

**Optimisation of $w$ and $b$**

Figure 2 displays the mean compression entropy $H_c$ within the low-risk subgroup as a function of the window and buffer length. The compression entropy decreases with increasing window length $w$. The influence of the buffer size on $H_c$ is only marginal.

The compression entropy of the BBI and BBI after mean subtraction revealed similar significances for discriminating HR and LR patients. Possible subtraction of BBI led to the elimination of only linear structures in the BBI time series. Therefore, both time series $x$ and $x_{mv}$ differed only in their linear properties, and all non-linear structures were preserved.

According to the optimisation procedure, compression entropies $H_c^{7,3}$, $H_{c,mv}^{7,3}$ and $H_{c,diff}^{25,3}$ were selected for application to patient data sets.

**HRV and compression entropy**

All the groups compared (ischaemic and idiopathic, LR, HR, HR-SCD) were age- and gender-matched and could not be discriminated by the clinical parameter LVDD (Table 3). Comparison of the parameters for idiopathic and ischaemic CM patients showed only significant differences for sdNN and medication (diuretics, ACE inhibitors).

However, the clinical measures NYHA and EF were different in the ischaemic groups (NYHA $p < 0.005$; EF $p = 0.024$), and the parameters sdNN, LF and compression entropy revealed significant differences between LR and HR, and LR and HR-SCD.

After adjustment for multiple comparisons (Bonferroni correction), the compression parameter $H_c^{7,3}$ remained significantly different, especially between LR and HR. The parameters meanNN and sdNN decreased with increasing risk of cardiac death. Furthermore, the parameters LF/HF and LFn revealed significant differences for discrimination between low-risk patients and cardiac deaths, but could not contribute to the identification of ischaemic patients at risk of SCD.

The application of compression entropies offered a promising result: the compression entropy of BBIs was significantly lower in patients suffering from cardiac death, especially SCD. Figure 3 shows the receiver operator characteristic curves for compression entropy $H_c^{7,3}$ (sensitivity 78.6%; specificity 61.2%; area under curve 74.4%) in comparison to the traditional parameter sdNN (sensitivity 64.3%; specificity 62.0%; area under curve 66.6%). Multivariate combination of the parameters sdNN and $H_c^{7,3}$ showed marginal enhancement of risk stratification (sensitivity 71.4%; specificity 68.6%; area under curve 74.6%). Assuming that the compressibility of a time series is a measure of its non-linear complexity, the complexity of heart rate in high-risk patients is reduced and, therefore, compression entropy decreases with increasing risk. Compression entropy based on the differences between successive BBI $x_{diff}$ revealed no significant differences for discrimination between HR and LR patients after adjustment for multiple comparisons. Compression entropy based on BBI-Hc seems to be

**Table 2** Results of surrogate analysis: selection of parameters.

<table>
<thead>
<tr>
<th>$H_c^{7,3}$</th>
<th>$H_{c,mv}^{7,3}$</th>
<th>$H_{c,diff}^{25,3}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.759</td>
<td>0.759</td>
<td>0.774</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$H_c^{7,3}$</th>
<th>$H_{c,mv}^{7,3}$</th>
<th>$H_{c,diff}^{25,3}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.753</td>
<td>0.753</td>
<td>0.751</td>
</tr>
<tr>
<td>&lt;0.05</td>
<td>&lt;0.05</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>0.464</td>
<td>0.464</td>
<td>0.457</td>
</tr>
<tr>
<td>0.44</td>
<td>0.44</td>
<td>0.429</td>
</tr>
<tr>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

$S_i$, surrogate set (average group value); $X_i$, original BBI; n.s., not significant; $p$, univariate significance.
Cross-correlations between the clinical parameters, measures of HRV and Cdiff were confirmed. Measures of HRV and variability. In this case, the pressibility of a time series is mainly influenced by its meanNN as established measures of variability can be explained. Diuretics as a specific risk marker of arrhythmic death. The NYHA measure has some potential limitations as a specific risk marker of arrhythmic death. The results ---  

**Table 3** Results of univariate statistical analysis presented as median and interquartile ranges.

<table>
<thead>
<tr>
<th></th>
<th>Ischaemic</th>
<th>Idiopathic</th>
<th>LR</th>
<th>HR</th>
<th>HRSCD</th>
</tr>
</thead>
<tbody>
<tr>
<td>NYHA class II</td>
<td>113 (77.4%)</td>
<td>49 (83.1%)</td>
<td>103 (85.1%)</td>
<td>10 (40%)</td>
<td>7 (50%)</td>
</tr>
<tr>
<td>Beta-blockers [n]</td>
<td>107 (73.3%)</td>
<td>35 (59.3%)</td>
<td>91 (75.2%)</td>
<td>16 (64%)</td>
<td>9 (64.3%)</td>
</tr>
<tr>
<td>Diuretics [n]</td>
<td>91 (62.3%)</td>
<td>48 (81.4%)</td>
<td>71 (58.7%)</td>
<td>20 (80%)</td>
<td>12 (85.7%)</td>
</tr>
<tr>
<td>ACE [n]</td>
<td>101 (69.2%)</td>
<td>52 (88.1%)</td>
<td>86 (71.1%)</td>
<td>15 (60%)</td>
<td>11 (78.6%)</td>
</tr>
<tr>
<td>LF/HF</td>
<td>2.3 [1.5–4.0]</td>
<td>2.5 [1.3–3.9]</td>
<td>2.4 [1.6–4.1]</td>
<td>1.8 [1.3–2.9]</td>
<td>1.9 [1.3–3.3]</td>
</tr>
<tr>
<td>LFn</td>
<td>0.71</td>
<td>0.64</td>
<td>0.64</td>
<td>0.65</td>
<td>n.s.</td>
</tr>
<tr>
<td>$H_c^{7,3}$</td>
<td>0.7 [0.6–0.8]</td>
<td>0.57–0.79</td>
<td>0.7 [0.62–0.81]</td>
<td>0.56–0.74</td>
<td>0.57–0.77</td>
</tr>
<tr>
<td>$H_{car}^{7,3}$</td>
<td>0.05 [0.43–0.57]</td>
<td>0.46–0.60</td>
<td>0.5 [0.46–0.58]</td>
<td>0.38–0.51</td>
<td>0.37–0.53</td>
</tr>
<tr>
<td>$H_{car}^{7,3}$</td>
<td>0.33</td>
<td>0.34</td>
<td>0.33</td>
<td>0.31</td>
<td>n.s.</td>
</tr>
<tr>
<td>Cdiff</td>
<td>0.31–0.35</td>
<td>0.31–0.36</td>
<td>0.32–0.35</td>
<td>0.30–0.34</td>
<td>0.32 [0.3–0.34]</td>
</tr>
<tr>
<td>Cdiff</td>
<td>0.32</td>
<td>0.33</td>
<td>0.33</td>
<td>0.31</td>
<td>n.s.</td>
</tr>
</tbody>
</table>

ACE, angiotensin-converting enzyme inhibitors. *p < 0.05; **p < 0.005 by univariate analysis; n.s., not significant.

more useful than $H_{car}$ for risk stratification. The results of window and buffer optimisation were confirmed.

Cross-correlations between the clinical parameters, measures of HRV and $H_c$ were analysed using Pearson’s correlation coefficient ($r$). $H_c^{7,3}$ was slightly correlated with meanNN ($r_p = 0.5$) and sdNN ($r_p = 0.67$). The compressibility of a time series is mainly influenced by its variability. In this case, the $H_c$ correlation with sdNN and meanNN as established measures of variability can be explained. $H_c$ shows nearly no dependence on traditional clinical parameters ($r_p = H_c/EF = 0.06$; $r_p = H_c/NYHA = 0.18$; $r_p = H_c/LVDD = 0.07$).

Despite the clinical applicability of assessment of functional class, the NYHA measure has some potential limitations as a specific risk marker of arrhythmic death. The degree of functional impairment, the degree of left ventricular dysfunction, and the prevalence of fatal arrhythmias are not linearly related, as the proportion of sudden death from total cardiac mortality is higher among the patients with mild to moderate heart failure than in those with severe heart failure [4].

One of the limitations of this study was that the clinical parameters were diagnosed by very experienced specialists, which increases the positive predictive value of NYHA classification.

Because of the limited number of patients with diabetes mellitus in the HR groups, the influence of diabetes on the efficiency of compression entropy and HRV could not be analysed. Analysis of ECG data in combination with the BBI time series may lead to further enhancement of the risk stratification.

For further investigations it would be useful to implement the compression entropy of heart-rate time series into a multivariate analysis for risk stratification of SCD in idiopathic cardiomyopathy patients.

**Conclusion**

In this study we investigated the suitability of compression entropy based on different variations of BBI time series for the analyses of non-linear structures in heart-rate time series, and especially for risk stratification in ischaemic cardiomyopathy patients. Surrogate analysis proved that compression entropy can detect non-linear structures in BBI time series. Using this new method, significant discrimination between low-risk patients and patients with a high risk of cardiac death was achieved. Hence, the implementation of compression entropy analysis based on BBI seems to be useful for multivariate analysis for enhanced risk stratification of SCD in ischaemic cardiomyopathy patients.
Acknowledgements

This study was supported by the MUSIC trial (I. Carlos III – Spain network group) and grants from the Federal Ministry of Education, Science, Research and Technology BMBF (13N7720/4) and the Ministry of Culture and Education of Thuringia (HWP project 25713723).

References