Towards the determination of the optimal recording duration for Heart Rate Variability
Applications. A bootstrap study on the “Fantasia” HRV database

José Luis Hernández Cáceres,* Ambrose Vincent Namugowa,* Jehu Iputo,* Rolando Hong,** Luis Garcia** and Miguel Sautié.**

*Department of Physiology, Health Sciences Faculty, University of Transkei, Private Bag X1 5117, Umtata, South Africa.
**Grupo de Bioinformática, Centro de Cibernética Aplicada a la Medicina, Instituto Superior de Ciencias Médicas de La Habana, Ciudad de La Habana, Cuba.


Palabras clave: variabilidad de la frecuencia cardíaca, bootstrap, electrocardiograma.
Key words: heart rate variability, bootstrap, electrocardiogram.

RESUMEN. Para determinar la mínima duración (L) de un cardiotacograma (CTG) capaz de garantizar un análisis de Variabilidad de la Frecuencia Cardíaca (VFC) estadísticamente confiable se analizaron CTG correspondientes a cinco sujetos jóvenes (Y) y cinco ancianos (O), todos saludables. Los datos fueron obtenidos de la base de datos “Fantasia”, disponible en el sitio www.physionet.org. Condiciones de registro: 2 h de registro continuo en posición supina. Se aplicó una técnica de “bootstrap” mediante la cual se generaba un gran número de experimentos replicativos a partir de los cuales, se obtenían distribuciones empíricas correspondientes a tres diferentes índices de evaluación de la VFC en el dominio del tiempo. Se seleccionaron los índices siguientes: PNN₅₀, SDNN, según la definición aceptada en la literatura, así como el “índice de ley de potencia (PLI)”, una medida de autoafinidad fractal. Los índices seleccionados permitieron una separación entre grupos si se comparaban segmentos de larga duración (L = 4 000 intervalos). La capacidad discriminativa disminuyó al reducir el valor de L. La aplicación del método “Bootstrap” mostró diferencias significativas cuando comparaban segmentos de registros de seguimiento de 75 a 150 intervalos. Duraciones tan cortas como 15 intervalos son suficientes para hacer comparaciones basadas en SDNN, y PNN₅₀, lo que se corresponde con estudios recientes donde otros autores utilizan trazos convencionales de ECG para estudios de VFC. La probabilidad de solapamiento entre grupos, estimada mediante el método ‘bootstrap’ fue elevada para duraciones cortas. Esa probabilidad disminuyó al incrementar L, y dependía del tipo de índice seleccionado. Los valores óptimos de L correspondientes a un solapamiento mínimo fueron cercanos a 1 000 intervalos para SDNN, 2 000 intervalos para PNN₅₀ y cercanos a 4 000 para PLI. La aplicación de un método estadístico computacionalmente intensivo permitió extraer información relevante acerca de los valores óptimos de L para la VFC; es de esperar que estos resultados sean útiles para otros estudios de VFC, en particular, para el diseño óptimo de grupos de experimentación, y para estudios individuales de registros de VFC. El estudio de la VFC a corto plazo es importante para la investigación relacionada con la discriminación entre grupos experimentales. Para la determinación de índices clínicamente relevantes de VFC en un paciente individual, parecen ser necesarios registros de al menos 1 h de duración.

ABSTRACT. In order to determine the shortest duration (L) of a cardiotachogram (CTG) that allowed statistically reliable analysis of heart rate variability (HRV) the CTGs corresponding to five young (Y) and to five elderly (O) healthy subjects were downloaded from the “Fantasia” database; data from “Fantasia” is freely available at www.physionet.org. Recording conditions: 2 h of continuous recording in supine position. Bootstrap analysis was applied for generating a large amount of replicate experiments to obtain empirical distributions corresponding to different time-domain HRV indices, PNN₅₀, and SDNN are generally accepted definition of HRV indices; the Power Law Index (PLI), a measure of fractal self-similarity was also used. The selected indices accounted for group separation when large duration segments were used (L = 4 000 heartbeats). The discriminative capability, however, was reduced as L was shortened. The application of the bootstrap method revealed that significant differences (P₁ < 0.05) may be found for all the three indices when cardiotachograms as short as 70-150 heartbeats are compared. Duration as short as 15 heartbeats are sufficient for obtaining faithful discriminations based on SDNN, and PNN₅₀ which is in agreement with recent studies by other authors using conventional ECG traces for HRV studies. The probability of group overlapping assessed by the bootstrap method was high for short duration segments. This probability decreased as L increased, in an HRV index-dependent way. Optimal L values corresponding to minimal group overlapping were around 1 000 heartbeats for SDNN, 2 000 heartbeats for PNN₅₀ and close to 4 000 for PLI. With the application of a computer intensive statistical method it was able to extract relevant information.
about optimal L values for HRV analysis. Hopefully, these results may be used for other HRV studies. In particular, for optimally designing group as well as individual HRV studies. Short-term HRV is supported for research aimed at finding differences between groups. For clinically relevant assessment of the HRV indices of an individual patient, at least 1 h of recording seems to be required.

INTRODUCTION

Heart rate variability (HRV) is emerging as a noninvasive and reliable approach to global autonomic nervous system (ANS) activity assessing.1 The primary information in HRV analysis comes from the cardio-tachogram (CTG), a plot of the time interval between the peaks of two successive Electrocardiogram (ECG) R-waves as a function of the heart-beat number (Fig. 1). In HRV literature, the term “I-RR signal” is frequently used as a synonym for CTG.

Current knowledge about ANS correlates of HRV emerges mainly from empirical evidences relating quantitative indices of the I-RR signal to known interventions upon ANS function.2,5 Methods and concepts from Nonlinear Dynamical Systems theory have been regarded convenient for the analysis of CTG data. Nonlinear methods are especially suitable for dealing with signals generated by complex nonlinear systems whose detailed mathematical description cannot be formulated directly due to incomplete information about the object.6-10,18

Unfortunately, the HRV phenomenon is difficult to enunciate even using the relatively loose framework of nonlinear science. Thus, according to different authors, HRV may either be regarded as a random process, a chaotic system, a fractal, a critical process, or a signal with intermittency.6,11,12,22

Despite all the theoretical difficulties for its interpretation, HRV has been, up to this time, the most reliable predictor of sudden cardiac death in post infarction patients.12,13 However, there are still several limitations to the widespread introduction of HRV. Besides theoretical constraints, there are also implementation drawbacks. Particularly, ECG machines are not suited for HRV signal acquisition. Usually 24-hour Holter recordings are recommended, but this option is both costly and computer demanding. Thus, the alternative concept of “Short-term HRV” (S-HRV) has recently appeared.14

The I-RR signal is nonstationary from the viewpoint of time series statistics.15 As a consequence, simple concepts emerging from the “Large Numbers Law” cannot be applied to HRV analysis. In HRV data the standard deviation of the R-R intervals (known in literature as the SDNN index) obtained from a 24-hour I-RR recording is larger than that obtained from a 5-min recording. As a consequence, it is not possible to compare short-term and long-term signals in a direct way.

In this paper, a bootstrap-type computer intensive statistical method16-17 has been applied in an attempt to shed light into the question of how short may an I-RR signal be for providing clinically relevant information.

The analysis was carried on a clinically well-characterized database of HRV recordings, whereas two widely used indices (SDNN and PNNn) as well as the Power Law Index (PLI), a theoretically correct self similarity index for fractal time series, were used for quantitative HRV evaluation.

Our results suggest that all of the three indices accounted for statistically significant differences between groups even when the I-RR signal duration was as short as 70-150 heartbeats (corresponding to 1-2 min of ECG recording). When SDNN and PNNn were used, even duration corresponding to routine 10 s ECG recordings were adequate for detecting differences between groups. If the restriction is posed of no individual recordings overlapping among the two groups, in that instance the optimal segment duration was from one to several thousand beats long. This points to the inadequacy of S-HRV for research purposes when comparisons between groups are undertaken, though for individual diagnostic purposes long-duration HRV is required.

MATERIALS AND METHODS

The data

The database included 10 I-RR recordings, corresponding to 5 young subjects aged 21-34 years old (Y), and 5 elderly, 68-81 year old subjects (O). All subjects were rigorously screened, and healthy condition was certified. ASCII files with individual recordings (O1.txt, O2.txt, ..., O5.txt, Y1.txt, ..., Y5.txt) were downloaded from the “Fanta-sia” database, freely available at “www.physionet.org”. Details about the Physionet website, as well as about the possibility to use these data for research purposes was described.19

Each trace corresponded to an I-RR signal obtained from 2 h of continuous ECG recording in supine position, and contained at least 4,000 heartbeat counts. The authors of (20) provide a further description of the data.

Statistical analysis

Rationale of the bootstrap method

A I-RR recording of length T may be conceived as composed from T/L segments of duration L (provided that T is a multiple of L). Since the database contains 10 recordings, at least 10 · T/L combinations of virtual experiments can be carried out. By analyzing all the possible combinations, an empirical distribution for a given quantitative index (e.g. PNNn) may be generated. From this empirical distribution probability estimations could be carried out. Since the groups O and Y respectively seem to belong to different populations according to different HRV indices, some of these parameters may be statistically different when comparing both groups. This is a way of generating large empirical distributions from relatively small data sets taking advantage of the great numerical capability of modern personal computers. The term “Computer Intensive Methods” (CIM) was proposed by Efron,16-17 for this type of analysis. Differences between group means according to a putatively sensitive parameter are very useful for empirical distribution generation. The authors also used another more restrictive distance criterion, defined as the difference between closest neighbors belonging to two groups. Thus, for example, the difference between the lowest value in SDNN in the Y group, and the highest SDNN value on the O group is defined as the overlapping distance for SDNN:

\[ OD_{SDNN} = \min(\text{SDNN}_Y) - \max(\text{SDNN}_O) \]

A negative value of OD_{SDNN} implies group overlapping according to the SDNN parameter. The proportion of negative distances among all the values in the distribution gives an estimate of the probability of type I error (P).
Ideally $P_1$ will depend upon $L$. The value of $L$ corresponding to a previously taken significance level (e.g. $P_1 = 0.05$) is the optimal $L$ value looked for in our study.

The bootstrap design

A bootstrap method\textsuperscript{16-17} was used for generating small duration segments from the original full-length recordings. Randomly chosen segments were analyzed to obtain the corresponding empirical distributions.

The first 4,000 data points of each I-RR signal were chosen. Numbers between 10 and $4,000 - L$, were randomly chosen. A set of ten different random numbers indicated, for each replication experiment, the starting position at each of the original CTG recording from which an L-beat-duration segment will be obtained. The segment duration $L$ was fixed for each of the $N$ replications corresponding to an empirical distribution generation process. The value of $L$ ranged from 2 to 2,000. After testing with different values of $N$, we found that the character of the empirical distribution was stable beyond $N = 200$ trials. Thus no less than 200 iterative experiments were used for each empirical distribution assessing.

HRV Indices

The following parameters were selected for analysis:

- **SDNN.** Is defined as the standard deviation of R-R intervals over a segment of duration $L$.\textsuperscript{15} If the I-RR signal were a Gaussian process, SDNN would converge into a constant value as $L$ increases. Due to HRV nonstationarity the dependence of SDNN upon $L$ is not easily predictable. SDNN also carries indirect information about other HRV measures such as spectral indices or fractal dimension.

- **PNN$_L$.** Is defined as the proportion of interval couples where the succeeding R-R interval has 50 ms more than the preceding one (15). This value is expressed in number of counts per hour.

- **Power Law Index (PLI).** This index was introduced by Higuchi\textsuperscript{21} as a time-domain-theoretically-correct index for fractal time series evaluation. PLI describes the exponent of the divergence of the signal’s variance as $L$ increases:

$$SD = C \cdot L^{(PLI)}$$

Where $SD$ is the standard deviation and $C$ is a proportionality factor. This parameter is exactly related to the fractal dimension of an ideal fractal.\textsuperscript{21} In HRV literature there are different indices for evaluating HRV, as the slope in log-log plots of power spectra,\textsuperscript{15} or the fractal index introduced by Yamasato.\textsuperscript{22} The PLI was chosen for its straightforward character, theoretical correctness, as well as its relation to SDNN [for details about a previous application of the PLI index to physiological time series, see (23)].

Dissimilarity measures

In this study, two dissimilarity measures were chosen:

- a) The difference between the mean value corresponding to each group during each replication.

- b) The difference between the lower value of the $Y$ group and the highest value of the $O$ group during each replication; The authors call this measure overlapping distance. A negative value of this difference indicates overlapping between groups.

Computational resources

The necessary algorithms for this work were separately programmed in C language using the GNU C compiler (GCC) on a PC (Pentium III, 866 Mz and 256 RAM memory) running Red Hat Linux V7.1 as operative system. Algorithms used in this study are freely available upon request from authors.

RESULTS

Long term recording comparisons

The statistical analyses were carried out on database recordings (Fig. 1). The mean values, standard deviations, and standard errors for each index in both groups of sub-

---

**Fig. 1.** Compose plot of all the recordings from the “Fantasia” Database, corresponding to the $Y$ group (left column) and the $O$ group (right column). An expert, based solely on recordings’ visual appreciation, classified the traces into three clusters. One cluster contained only traces from the $Y$ group, other cluster contained traces only from the $O$ group, whereas a third cluster contained traces from the both groups ($Y_3$, $Y_5$ and $O_4$).
jectors are shown for the case when 4 000 data points were taken from each recording (Fig. 2). Conventional one-sided significance tests revealed significant differences between the Y and the O group according to all three indices ($P_I = 0.007$ for SDNN, $P_I = 0.001$ 3 for $P_{NN_{50}}$, and $P_I = 0.003$ 6 for PLI).

**Short term data segments comparisons**

If instead of the whole recording only 150 data points segments were selected, the results from the previous section would change not only in numerical values, but also in statistical significance (Fig. 3). Estimated differences between the Y and the O group according to all three indices ($P_I = 0.126$ 5 for SDNN; $P_I = 0.008$ 3 for $P_{NN_{50}}$, and $P_I = 0.003$ 1 for PLI). When 300 data points duration were analyzed, the results were $P_I = 0.010$ 3 for SDNN, $P_I = 0.004$ 8 for $P_{NN_{50}}$, and $P_I = 0.010$ 3 for PLI.

These results illustrate the obvious condition that increasing $L$ leads to a better separation between groups. However, for relatively small data samples ($N = 5$) the applicability of statistical tests based on normal distribution assumptions is not theoretically justified. In the next section, empirical distributions from the data are obtained using the bootstrap method.

**Bootstrap method**

a) **Differences between groups**

If 150-point segments are taken and as much as 10 000 experiments are repeated, then an empirical distribution can be generated for the difference of the mean value corresponding to the O and the Y groups. The probability for Type I error (the probability of classifying recordings of different groups as belonging to the same one)$^{16-17}$ was less than 0.01 for all the three indices. Even when 150 heartbeats corresponding to about 2 min of recording, seem to be a “too-short” recording for HRV analysis, this result may be interpreted as a safety factor allowing to use 5-min recordings for HRV.

Recent literature reports have made claims about the utility of carrying out heart rate variability analysis from data obtained in a conventional 10 s ECG recording.$^{24}$

Assuming that 10 s roughly correspond to 15 heartbeats it was tested whether some HRV indices preserve their predictive value when very short segments are used. It was found that both $P_{NN_{50}}$ and SDNN preserved their predictive value under these conditions.

The bootstrap method was applied for generating an empirical distribution when only one segment 8 beats long was recorded from each of the 10 subjects. This corresponds to a null hypothesis for no difference between groups. The distribution of the mean was generated, and the probability for the measured difference between Y and O to belong to the homogeneous population was estimated. Since a value of $P_I = 0.005$ was obtained, the null hypothesis is not accepted. Thus the theoretical conclusion based on the study of many segments combinations is supported by the result from the analysis of only one measurement.

Not surprisingly, PLI did not show significant differences for 15-heartbeats segments ($P_I = 0.10$). At the same time, for $L = 70$ heartbeats, $P_I$ was less than 0.05. This result points to usefulness of PLI as a reliable index for data corresponding to about 1 min of recording.

The dependence of $P_I$ for differences between groups with respect to the segment length is clear for all the three indices (Fig. 4).
Thus statistical support has been provided for the use of short term HRV in studies where finding out differences between groups is the objective.

**An example with conventional ECG traces**

According to the results from figure 4a, it would be possible to find statistically significant differences between groups, when the SDNN is computed from I-RR signals as short as 3 heartbeats. In this section we describe an attempt to test that prediction using conventional paper ECG traces.

ECG data corresponded to a sample of 27 putatively healthy male students in the age range from 8 to 19 years. Data were collected at Umtata Primary School, and at Umtata High School, in an urban area of Umtata, Eastern Cape. Portable ECG machines (Dr. Lee, South Korea) were used for paper traces acquisition. A significant negative correlation with age was obtained (r = 0.44, PI < 0.05).

Since the decline of SDNN with age is a well-documented fact, our ECG traces provided additional support to the predictions obtained from the statistical analysis of the “Fantasia” database.

**b) Individual overlapping probability**

To account for statistically significant difference between groups under different risk factors is a frequent purpose in HRV research. In most of the clinical applications, however, we need to decide whether or not a given patient shows abnormal HRV. In this case the presence of overlapping in the distribution of abnormal and normal subjects may create difficulties, even if the differences between two groups are statistically sound.

For those purpose the overlapping distance was used, according to the previously introduced definition.

\[
OD_k = \min_{k \leq Y} - \max_{k \geq Y},
\]

where \( k \) corresponds to an HRV index (e.g. SDNN), and \( \min_{k \leq Y} \) is the minimal value of the index \( k \) among young subjects. Thus if each individual SDNN value at the group Y is higher than any individual value at the O group, the distance OD will be positive, and no group overlapping is present.

Compared to the mean difference obtained when overlapping difference was used for 150 data-length segments the probabilities for type I errors increased notably for all indices (\( P_I = 0.95 \) for SDNN; \( P_I = 0.43 \) for PNN\(_{50}\), and \( P_I = 0.89 \) for PLI). When 300 data points were taken, the situation improved, but not to a satisfactory extent (\( P_I = 0.97 \) for SDNN; \( P_I = 0.369 \) for PNN\(_{50}\), and \( P_I = 0.85 \) for PLI).

Increasing \( L \), and using the bootstrap method, we observe that the probability for Type I error is reduced in a different way for each of the indices analyzed (Fig. 5).

As is apparent, \( P_I \) for SDNN falls from practically 1 to 0 in an almost stepwise way, being the border point close to \( L = 1 \ 000 \) (Fig. 5a). \( P_I \) for PNN\(_{50}\) declines in a seemingly smooth way (Fig. 5b), whereas an apparently biphasic and slow decrease is the pattern corresponding to the PLI index (Fig. 5c).

\( L = 2000 \) is the optimal value for PNN\(_{50}\), whereas no value of \( L \) below 4 000 seem to be optimal for PLI.

**DISCUSSION**

The results of this study may be summarized as follows.

Using the “Fantasia” HRV data base, statistically significant differences between HRV recordings from a group of young and of elderly healthy subjects may be observed according to the following indices: SDNN, PNN\(_{50}\), and PLI. Differences between groups are significant even when 70-150 data points are used as recording length. Overlapping between groups will be absent in a statistical sense only if much longer data segments are taken for analysis (1 000 for SDNN, 2 000 for PNN\(_{50}\) and about 4 000 for PLI).

Related to a plausible explanation for the results obtained in figures 5A-5C, as is apparent, the three observed indices differ, not only in the optimal \( L \) value for no overlapping, but also in the way the probability of type I error depend upon \( L \).

**PNN\(_{50}\)** When using 4 000 data points, the standard deviations of PNN\(_{50}\) in both groups almost overlap (Fig. 2). A detailed look at individual recordings reveals that a subject in the Y group had a PNN\(_{50}\) = 238 whereas another individual in the O group had a PNN\(_{50}\) = 195. These values are expressed as counts over 1 h recording time. When these values are expressed as probability (or frequency) values, we obtain a probability of 0.063 for the recordings from the Y group and 0.056 for the recording from the O group. Assuming Gaussianity, a calculation of Type I error probability can bring a result similar to that obtained for the dependence of PNN\(_{50}\) upon \( L \). This in case, the dependence seems to emanate

---

**Fig. 4.** Type I error probability for group differences as a function of segment length. Plots correspond to the following indices: PNN50 and SDNN (4A), and PLI (4B).

Notice the difference between plots as to the optimal \( L \) value as well as to the type of the functional dependence.
from the close probability values among individual recordings.

**SDNN.** The non-stationary nature of HRV makes the analysis of SDNN complicated. If we assume that HRV is a fractal process following a power law dependence of SDNN upon $L$.

$$SDNN = k \cdot L^{PLI}$$

If there were some overlapping at low $L$ values, for identical PLI values, it would not be possible to reduce the probability of overlapping in SDNN at the expense of increasing $L$. If PLI were lower in the $O$ group, then a stepwise function as that observed for SDNN in figure 5b would be theoretically expected. Thus this result is an indirect evidence of differences in PLI among the two groups. Since HRV is not a pure fractal process, these conclusions cannot be applied in a straightforward way. Nevertheless they provide qualitative information about the interplay between the variance and the fractal properties of HRV in both groups.

**PLI.** This index is a measure of the long-term behavior of the I-RR signal. In the case of a pure fractal, this measure should ideally be constant over all the time windows, in accordance with the self-similarity properties of fractal processes. However, as stressed by Yamamoto, the I-RR signal is not a pure fractal, since short-term components, appearing at the high frequency region of the power spectrum, are far from linearity in the log-log spectral plot. Thus the need for longer $L$ values providing statistically significant differences with the PLI index is a corollary from the non-fractal processes contribution at short-term range. These results are also in agreement with those of Hernandez et al., who found that the I-RR signal includes several types of dynamical attractors, whose behavior cannot be described as that of a pure fractal, though fractal-like components may also be detected.

The fact that only long range data are useful for statistical discrimination with PLI are in agreement with those from Enzmann et al. They found that complexity measures for discriminating between hemodynamically stable and unstable patient’s HRV data provide a complete separation of groups if several hour duration recordings are taken.

It is paradoxical that, in spite of the great interest in the related literature, there are very few conditions where it is possible to definitively assert that HRV may be a diagnostic tool. From that viewpoint it seems that the situation with most of the possible applications of HRV would be similar to the data analyzed in this study.

The fact that the theoretical prediction about the usefulness of traces as long as three heartbeats for statistical group stratification was supported by the analysis of conventional paper ECG recordings may be illustrative.

The authors suggest that a preliminary survey, with a relatively small data sample, and using methods similar to those described in this work, could be of great help for the final design of a through HRV study.

**CONCLUSIONS**

A computer-intensive statistical method applied to a data base of five recordings from young healthy male subjects ($Y$) and five old male healthy subjects ($O$) was used to...
approach the problem of the minimal data length necessary for achieving enough group separation using commonly accepted indices for HRV analysis. The main result from the study is that, for answering questions related to statistical difference between groups, 70-150 data points, corresponding to 1-2 min of recording, are enough. At the same time, for avoiding overlapping of individual recordings pertaining to different groups, much larger recordings are needed. Since intergroup differences are relevant in research with large groups of individuals, whereas overlapping safety is important for individual diagnosis purposes, these results suggest that short term HRV is useful for research purposes whereas long term HRV is necessary for focused diagnostic studies.

Finally, the authors recommend carrying out a study with a similar design as a preliminary stage of any HRV research where new categories of subjects are submitted to analysis.

BIBLIOGRAPHY