Heart rate variability measures: a fresh look at reliability

Gian Domenico Pinna, Roberto Maestri, Antoni Torunski, Ludmila Danilowicz-Szymanowicz, Malgorzata Szwoch, Maria Teresa La Rovere, Grzegorz Raczak

Clinical Science Immediate Publication. Published on 23 Mar 2007 as manuscript CS20070055

HEART RATE VARIABILITY MEASURES:

A FRESH LOOK AT RELIABILITY

Gian Domenico Pinna[1], Roberto Maestri[1], Antoni Torunski[2], Ludmila Danilowicz-Szymanowicz[2], Malgorzata Szwoch[2], Maria Teresa La Rovere[1] and Grzegorz Raczak[2]

[1] Department of Cardiology and Biomedical Engineering, S. Maugeri Foundation-IRCCS, Scientific Institute of Montescano, Montescano (PV), Italy

[2]II Department of Cardiology, Medical University, Gdansk, Poland

Keywords: heart rate variability, reproducibility, repeatability, reliability, paced breathing

Running title: Reliability of Heart Rate Variability Measures

Address for correspondence:
Gian Domenico Pinna,
Servizio di Bioingegneria,
Fondazione S. Maugeri, IRCCS
Istituto Scientifico di Montescano,
27040 Montescano (PV)- ITALY
Tel. ++39-0385 247256
Fax ++39-0385 61386
E-Mail: gdpinna@fsm.it

Clinical Science Immediate Publication. Published on 23 Mar 2007 as manuscript CS20070055

Abstract

Despite its extensive use in physiological and clinical research, the analysis of heart rate variability (HRV) is still poorly supported by sound reliability studies. The main aim of this study was to make an in-depth assessment of absolute and relative reliability of standard indexes of HRV from short-term laboratory recordings.

In 39 healthy subjects (mean age (min-max): 38 (26-56) years, 18 men and 21 women) we recorded 5 + 5 min of supine ECG during spontaneous and paced (15 breaths/min) breathing. The test was repeated on the next day in the same conditions. From RR intervals we computed standard indexes of HRV: SDNN, RMSSD, LF and HF power (absolute and normalized units) and LF/HF. Absolute reliability was assessed by the 95% limits of random variation (RV); relative reliability was assessed by the intraclass correlation coefficient (ICC). The sample size needed to detect a mean difference ≥30% of the between-subject SD was also estimated.

Although there was no significant mean change between the two tests, we found that in individual subjects the second measurement can be as high/low as 1.9/0.5 times (SDNN, best case) and 3.5/0.3 times (LF/HF, worst case) the first measurement, due to pure random variation. For most parameters the ICC was > 0.8 (range: 0.65-0.88). The estimated sample size ranged from 24 to 98 subjects. Reliability indexes tended to improve during paced breathing.

We conclude that short-term HRV parameters are subject to large day-to-day random variations. Random error, however, represents a limited portion of between-subject variability; therefore observed differences between individuals reflect mostly differences in the subjects' error-free value rather than random error. Overall, paced breathing improves reliability.

INTRODUCTION

After almost 25 years since the publication of the pioneering paper from Akselrod and co-workers [1], heart rate variability (HRV) is still a matter of great interest in clinical and physiological research. According to Pubmed database, from 2000 to 2006 the number of yearly publications related to HRV has steadily increased from 391 to 584, and this trend is going to be confirmed in 2007. Although such a broad use of this methodology would implicitly assume that the reliability of HRV measurements has been thoroughly evaluated previously, an unprejudiced look at available reliability studies clearly shows that this evaluation has often been inadequate [2]. This is particularly true for HRV indexes derived from short-term laboratory experiments, which are those most commonly used in non-invasive investigations of autonomic cardiovascular control. Major methodological limitations of published reliability studies include: 1) inadequate protocol (e.g.: replicate measurements were taken too far from each other), 2) insufficient sample size, 3) too short or too long recordings according to current guidelines [3], 4) limited selection of studied HRV parameters, 5) inadequate assessment of reliability due to the use of inappropriate reliability indexes or to misapplication/incomplete inclusion of appropriate indexes, and 6) lack of indications on the practical implications of computed reliability indexes (e.g., for the assessment of individual responses or for sample size estimation). A detailed review of these issues can be found in recently published papers [2, 4, 5].

In this study we accurately and comprehensively assessed the reliability of short-term HRV measurements in healthy individuals, in order to overcome major methodological limitations of previous investigations. Both absolute and relative reliability have been considered and implications for sample size calculation and assessment of individual changes are presented. As laboratory recordings are mostly carried out during spontaneous and/or paced breathing, both experimental conditions have been considered.

METHODS

Subjects

We studied 39 healthy volunteers (mean age (min–max): 38 (26-56) years, 18 men and 21 women. None of them was on medication or suffered from chronic or acute disease. The study was approved by the local Ethical Committee and all subjects gave their written informed consent before participation.

Study protocol

All tests were performed between 8.00 a.m. and 1.00 p.m., with the subjects in the supine position in a quiet and dimmed room at a comfortable temperature. All subjects fasted for more than 2 hours and refrained from smoking, alcohol or coffee for the 24 hours preceding the study.

After instrumentation, subjects carried out a session of familiarization with the paced breathing protocol [6]. They followed a recorded voice instruction to breathe in and out at a frequency of 0.25 Hz. After 15 minutes for signal stabilization, subjects breathed spontaneously for 8 min, and then breathed at the paced breathing frequency for another 8 min. During these two sessions ECG was recorded. An identical session was repeated on the next day at the same hour.

Signal analysis and measurements

Beat-by-beat RR interval values (resolution 1 ms) were obtained from the ECG signals using a software package developed in-house [7]. The RR interval time series were then re-sampled at 2 Hz by cubic spline interpolation. The analysis was carried out on the central 5-min window of each recording. This analysis window has become the standard one in short-term HRV [3].

After detrending via least-square second-order polynomial fitting, the power spectral density of RR time series was estimated by the Blackman-Tukey method using a Parzen window with a spectral bandwidth of 0.015 Hz [8]. The power in the low frequency (LF, 0.04-0.15 Hz) and high frequency (HF, 0.15-0.45 Hz) bands were obtained by numerical integration. These spectral indexes will be referred to as LF_BT and HF_BT. Spectral analysis was also performed using the autoregressive method (Burg algorithm) with spectral decomposition (Johnsen and Andersen algorithm). Autoregressive model order was set at 26, but was interactively increased when negative components appeared in the spectral decomposition table [8]. Spectral powers of the LF and HF bands (LF_AR, HF_AR) were computed summing the respective spectral components. Components showing < 10% of the overall power in the band were ignored as they probably represented pure noise contributions. The LF power in normalized units (LFnu) was computed as LF_AR/(LF_AR+HF_AR). The HF power in normalized units (=1-LFnu) was not analyzed, to avoid redundancy.

We also computed time-domain HRV parameters suitable for short-term analysis: the standard deviation of normal-to-normal beats (SDNN) and the root mean square of successive differences (RMSSD) [3]. Although mean RR interval is not, strictly speaking, an HRV parameter, it was included in the analysis as a major index of cardiac autonomic control.

Statistical analysis

The reader is referred to the appendix for a short introduction to the statistical basis for the assessment of reliability. The key point is that observed measurements are assumed to be the sum of i) a fixed quantity, which represents the true or error-free value of the characteristic being measured in each subject, and ii) a random quantity, commonly referred to as random error of measurement, which accounts for within-subject variability. According to these concepts, we first examined the distribution of measurements obtained in the two tests, as well as of their difference, to detect and discard possible outliers. An observed value was deemed to be an outlier if it was greater/less than the upper/lower quartile plus/minus 1.5 times the interquartile range [9]. This method required a preliminary log-transformation of skewed variables (assessed by the Shapiro-Wilk test for normality).

Following the recommendations of Bland and Altman [10], we plotted the difference between the two measurements ($X_2 - X_1$) against their average. This graphical method allows to look for any systematic change between the tests and also to check whether the random error is related to the size of the characteristic being measured (this is referred to as heteroscedasticity). This qualitative investigation was followed by a formal verification of the assumptions underlying the assessment of reliability (see appendix). Specifically, we tested the hypothesis of normality and zero mean of the difference between the two tests (two-sided paired t-test at the 0.05 significance level); we also verified the homoscedasticity assumption by regressing the absolute differences between the two measurements against their average [5, 10].

Since many variables showed non-normality and heteroscedasticity, this problem was solved by log-transformation (natural logarithm, ln) [10].

One-way random effects ANOVA was carried out on all variables to estimate the standard error of measurement (SEM), the major index of *absolute reliability*. This was obtained as the square root of the within-subject mean square (WMS) from the ANOVA table [11, 12]. From the SEM value we derived the 95% limits of random variation, i.e. the range of values within which 95% of the *differences* between two measurements ($X_2 - X_1$) are expected to lie due to pure random variation. For log-transformed variables, these limits were back-transformed (antilogarithm), giving the range of values within which 95% of the *ratios* between the two measurements (X_2 / X_1) are expected to lie due to pure random variation [5, 10].

From mean square values of the ANOVA table we also computed the intraclass correlation coefficient (ICC) as [11, 12] :

$$ICC = \frac{BMS - WMS}{BMS + WMS}$$

where *BMS* is the between-subject mean square. Ninety-five percent confidence intervals for ICC were calculated [13].

As the magnitude of the random error affects the sample size of an experiment [14], we estimated the sample size needed to detect a relevant change in the mean of HRV parameters after a treatment (test-retest experiment). Conventionally, we considered as "relevant" a change of ≥ 30% of between-subject standard deviation. We also assumed a two-sided test with a significance level of 5% and a power of 80%.

All analyses were carried out using the SAS/STAT statistical package, release 8.02 (SAS Institute Inc., Cary, NC, USA).

RESULTS

Spontaneous breathing

All variables except mean RR interval and LFnu showed a marked right-skewed distribution (Shapiro-Wilk test: $p < 0.01$). One subject showed an outlier in RMSSD, HF_BT and HF_AR, while another one showed an outlier only in RMSSD. These measurements and those derived from them (i.e., LFnu and LF/HF) were ignored in the subsequent analysis. Descriptive statistics of measured parameters in the two tests are reported in table 1.

Bland-Altman plots of the difference between the two measurements against their average displayed a symmetrical distribution of points around the zero line in all parameters, indicating absence of a systematic change. The width of the scatter around the same line, however, showed a clear increasing trend in SDNN, RMSSD, LF_BT, LF_AR, HF_BT, HF_AR and LF/HF, indicating heteroscedasticity. Two representative examples of respectively an homoscedastic variable (mean RR) and an heteroscedastic variable (LF_BT power) are displayed in figure 1 a and b respectively. Visual findings were confirmed by regression analysis**.**

Heteroscedastic variables were successfully log-transformed, obtaining at the same time homoscedasticity and normality. A representative Bland-Altman plot is shown in fig. 1c.

For all HRV parameters the difference between the two tests was negligible and non significant, indicating absence of systematic change.

Reliability indexes for homoscedastic variables (mean RR interval and LFnu) are reported in table

3a. If we consider LFnu, the limits of random variation indicate that, in order to be 95% confident that a real change has occurred in an individual, the observed difference between two measurements has to be > 0.29 or < -0.29. For the same parameter, the ICC indicates that 65% of the variability of LFnu measurements across the studied population is due to variability in the true value of the subjects, while the remaining 35% is due to random error. The last column of table 3a shows that 98 subjects are needed in a test-retest experiment to detect a change in mean LFnu of ≥ 0.04 (30% of between-subject standard deviation), with a significance level of 5% and a power of 80%.

Reliability indexes for heteroscedastic HRV parameters are reported in table 4a. Since the statistical analysis of these variables was carried out after log-transformation, the SEM and the change in the mean for the estimation of the sample size, are expressed in log units. The limits of random variation indicate that in order to be 95% confident that a real change has occurred in a given parameter, the ratio between two measurements (X_2/X_1) has to lie outside the indicated interval. The asymmetry of this interval is simply the result of the antilogarithmic transformation. It can be seen that for all spectral parameters the second measurement can be as large/small as about 3.5/0.3 times the first measurement due to pure random variation. To relate these inferential figures to the data of the study, figure 2 shows the plots of the ratio X_2/X_1 against the mean of the two measurements. Note the close agreement between theoretical 95% limits of variation and raw data.

ICC values reported in table 4a show that the proportion of total measurement variability of heteroscedastic HRV parameters explained by the variability of the subjects' error-free value ranged from 70% (LF_HF) to 86% (HF_BT and HF_AR); therefore random error accounted for 30% to 14% of total measurement variability.

Paced breathing

Descriptive statistics of HRV parameters derived from paced breathing recordings are reported in table 2. There was only one outlier in RMSSD. As for spontaneous breathing, SDNN, RMSSD, LF_BT, LF_AR, HF_BT, HF_AR and LF/HF exhibited an heteroscedasticity behavior and were successfully log-transformed. The difference between transformed measurements was largely non significant.

Reliability indexes for homoscedastic and heteroscedastic variables are reported in table 3b and 4b respectively. Compared to spontaneous breathing we found: i) an improvement of all reliability indexes (SEM, limits of random variation, ICC and sample size) for LF_BT, LF_AR and LF/HF, ii) an increased ICC and reduced sample size for LFnu; iii) a slight worsening of all reliability indexes for

SDNN. Analysis indexes of the other HRV parameters were almost unchanged.

DISCUSSION

Despite its extensive use in physiological and clinical research, the analysis of HRV is still poorly supported by sound reliability studies, and further investigations in this field has recently been advocated [2]. In this study we carried out an in-depth assessment of absolute and relative reliability of standard indexes of HRV from short-term laboratory recordings in healthy subjects during spontaneous and paced breathing. The experimental protocol and the analysis of collected data were performed according to state-of-the art methodology and best-practice criteria. We found that HRV parameters were characterized by large random variations within individuals, thus showing low absolute reliability. Random error, however, in most parameters represented a limited portion of total measurement variability across individuals, thus indicating good relative reliability. Overall, paced breathing improved the reliability of spectral parameters, particularly those derived from ratios of raw quantities (LFnu and LF/HF).

Reliability of HRV parameters: absolute reliability

Our study reveals the presence of a large random error (SEM) in all HRV parameters, particularly in those computed in the frequency domain. Indeed increases as great as about 3.5 times and decreases of as much as about 30% may occur from one measurement to the next due to pure random variation. Limits of random variation are lower for time-domain indexes, being about 1.9-2.4 times and 50-60% for SDNN and RMSSD respectively. These results question the use of HRV indexes in assessing treatment effects in individual subjects. Of note, in all HRV parameters but LFnu random error increases as the magnitude of the parameter increases, which is the hallmark of heteroscedasticity.

Random error of HRV parameters is in part due to sampling variability of the estimated parameters, as they are nothing but statistics computed on a finite number of RR intervals. Therefore, they are subjected to random changes from one sample to another [15]. Part of intra-subject variability is also due to an intrinsic lability of HRV parameters - probably because they are under the influence of such factors as mood, alertness and mental activity which are very difficult to control for in any study. Changes associated with frequency and depth of respiration also play an important role [6].

In the two previous reliability studies that provide estimates of the SEM (or equivalently of the

within-subject standard deviation), slightly greater values of this index were found [16, 17]. This might be due to differences in the experimental protocol.

Reliability of HRV parameters: relative reliability

The ICC of HRV parameters ranged between 0.65 and 0.88. Although the definition of a categorical rating of relative reliability based on ICC is still controversial, these values can reasonably be considered as indicating substantial to good reliability [5, 11, 18]. For most measurements the ICC was > 0.80, indicating that they reflect mostly the true value of HRV parameters relative to random error. From the mathematical definition of ICC (se the appendix), it clearly appears that such high values of relative reliability are the consequence of a large between-subject variability, a fact well-known to investigators involved in HRV analysis. The lowest values of ICC were found in the LFnu and LF/HF parameters during spontaneous breathing (0.65 and 0.70 respectively). This result probably derives from a relatively greater random error, as LFnu and LF/HF "carry" the error of both the LF and HF power. A further insight into the practical meaning of observed ICCs can be gained by remembering that in the context of our test-retest reliability study the ICC equals the correlation coefficient between paired measurements [19].

The estimated ICCs for time-domain HRV parameters are very close to those reported by Sinnreich et al [20] and similar to those found by Schroeder et al. [4]. Lower ICCs were obtained in the same parameters by Pitzalis et al. [16] and Gerritsen at al. [21]; their estimates, however, were based on raw (i.e., not transformed) data. The ICCs of spectral parameters were similar to or higher than those found by others [4, 16, 20]. A comparatively reduced LFnu during spontaneous breathing was also observed by Sandercock et al [22].

Implications of reliability in sample size estimation

A major implication of measurement reliability is the size of the sample needed to test a scientific hypothesis with a preset significance level and power. We explored this point by simulating a simple test-retest study to investigate the effect of a treatment on the mean value of HRV parameters. Since reference values as to what change in HRV parameters would be clinically relevant are lacking, we conventionally adopted the criterion of 30% of between-subject standard deviation. The rationale is that the more the subjects have dispersed values, the larger the shift in mean value should be to be clinically relevant. We found that the sample size can largely vary from parameter to parameter and

that it is inversely proportional to the ICC. The latter result also applies to the estimation of the sample size in group comparison studies [11].

Spontaneous versus paced breathing

Paced breathing substantially improved the reliability of LFnu and LF/HF, with a consequent dramatic reduction (> 50%) of the estimated sample size. This result is in agreement with the findings of Pitzalis and coworkers [16]. Moreover, voluntary control of breathing moderately improved the reliability of the LF power, while leaving substantially unchanged that of the HF power and RMSSD, and slightly decreasing the reliability of SDNN. We argue that the improvements observed during paced breathing might be due to a better stabilization of LF oscillations brought about by the virtual abolition of respiratory-related frequency components within the LF band [6].

Classical versus autoregressive spectral estimation

Our study suggests that the reliability of HRV parameters is not affected by the method used for spectral estimation (classical Blackman-Tukey or autoregressive method). A similar result was found by other investigators [16]. It should be stressed, however, that spectral estimates depend to a certain extent on the design criteria adopted in the analysis. For instance, autoregressive measurements depend on the criterion used for model order selection [8]. Therefore, the use of algorithms markedly different from those adopted in this study might yield different reliability figures.

CONCLUSION

In healthy subjects, short-term HRV parameters are subjected to large day-to-day random variations which would make difficult the detection of treatment effects in individual subjects. For most indexes, however, random variation represents a limited portion of the between-subject variability; therefore observed differences between individuals reflect mostly differences in the subjects' true value rather than random error. The sample size for an experiment markedly depends on the reliability of the HRV index considered; this implies that the design of experiments based on the measurement of a set of HRV parameters should be tuned to the indexes with lowest reliability. Paced breathing appears to provide more reliable HRV measurements, particularly those related to the spectral content of the LF band.

Clinical Science Immediate Publication. Published on 23 Mar 2007 as manuscript CS20070055

Appendix

Statistical basis for the assessment of reliability

A measurement is said to be to reliable when the values obtained under identical conditions on the same individuals at different times agree closely each other. Therefore, reliability is synonymous with *intrasubject reproducibility, repeatability or consistency.* Reliability has classically been investigated assuming the following statistical model [11, 12]:

$$X = \tau + \varepsilon \qquad (1)$$

In this model, the measurement X made on a given subject is assumed to be the sum of a fixed quantity τ, the "true value" or "error-free value" of the characteristic being measured in that subject, and a random quantity ε, commonly referred to as random error of measurement, which accounts for *intra-subject variability*.

There are some basic *assumptions* underlying model (1): the random error ε is normality distributed, has zero mean, is uncorrelated within and between subjects, and has a fixed standard deviation independent of τ, the latter property being commonly referred to as *homoscedasticity* [5]. Moreover, for a population of subjects, τ is assumed to be normally distributed. Before analyzing reliability, all these assumptions should be carefully verified or confidently assumed on the basis of a properly conducted experiment. Failure to satisfy homoscedasticity and normality assumptions is commonly dealt with by variable transformation [10].

If we take two replicate measurements on the same individual under identical conditions, from equation (1) we have that the difference ΔX between them will be:

$$\Delta X = X_2 - X_1 = (\tau + \varepsilon_2) - (\tau + \varepsilon_1) = \varepsilon_2 - \varepsilon_2 = \delta \qquad (2)$$

where the suffixes 2 and 1 indicate the measurements at the two occasions. Equation (2) clearly shows that the repeatability of the measurement depends on the random quantity δ: the lower δ, the closer the two measurements will be with each other. The magnitude of δ, as expressed by its standard deviation, is $\sqrt{2} \cdot \sigma_\varepsilon$, where σ_ε, the so-called *standard error of measurement* (SEM), is the standard deviation of ε. Therefore, the repeatability of a measurement ultimately depends on the magnitude of the SEM. Accordingly, the SEM is considered the key statistical indicator of *absolute reliability* [5, 11, 19].

The SEM has the following two major uses. First, if we take two measurements on the same

individual before and after a treatment and want to be 95% confident that a real change has occurred, the observed difference has to lie outside the interval $-1.96 \cdot \sqrt{2} \cdot SEM$, $+1.96 \cdot \sqrt{2} \cdot SEM$, or $-2.77 \cdot SEM$, $+2.77 \cdot SEM$ [5, 10]. Therefore, the extremes of this interval can be viewed as the *95% limits of random variation*. The value $2.77 \cdot SEM$ is called *repeatability coefficient* [10]. Second, the SEM is a crucial parameter in determining the sample size for an experiment [5, 14]. This is because the higher the random error of a measurement, the greater the "noise" that will tend to obscure a possible treatment effect.

Another classical way of assessing reliability is through the intraclass correlation coefficient (ICC), which is also commonly referred to as *reliability coefficient*. It expresses the proportion of the variability of observed measurements that is explained by the variability of the subjects' error-free value. The ICC is then defined as [11, 12]:

$$ICC = \frac{\sigma_\tau^2}{\sigma_X^2} = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_\varepsilon^2} = \frac{1}{1 + \sigma_\varepsilon^2 / \sigma_\tau^2} \qquad (3)$$

where σ_x^2 total variance of observed measurements among the subjects of the considered population, σ_τ^2 is the part of total variance due to differences in the patients' true value and σ_ε^2 is the part due to random error (i.e., the square of the SEM).

The ICC ranges from 0 to 1: the lower the random error relative to subject-to-subject variability, the closer the ICC will be to 1; conversely, the greater the random error relative to subject-to-subject variability, the closer the ICC will be to 0. Therefore, the higher the ICC, the more a measurements will reflect the true value and the more probably we will be able to detect differences between individuals. For these reasons the ICC is considered the key statistical indicator of *relative reliability* [5]. When reliability is assessed taking two replicate measurements per subject, the ICC turns out to be mathematically equivalent to the correlation coefficient between paired measurements [19].

Various categories of reliability based on ICC have been proposed so far [5, 11, 18]. Although these criteria do not fully agree with each other, an ICC >0.8 is usually regarded as indicating good to excellent reliability, while an ICC between 0.6 and 0.8 may be taken to represent substantial reliability. From equation (3) it clearly appears that the ICC, besides being dependent on the random variability of the measurement, depends on the variability (i.e.: heterogeneity) of the population being studied. Therefore results obtained in one population can not be extrapolated to a new an possibly more/less homogeneous population [5, 19].

Assessment of reliability

To assess reliability, the measurement of interest is taken on a sample of individuals during two or more replicated experiments under as close to uniform conditions as possible. To fulfill the statistical assumptions for analysis, replicated experiments should be performed neither to far apart in time, to ensure constancy of the characteristic being measured, nor too close, to avoid potential carry-over or learning effects. Collected data should be carefully checked to detect outliers and to verify the required statistical assumptions (see above) [12]. Graphical methods such as Bland-Altman plots as well as formal hypothesis testing are used at this purpose [5, 10, 12]. Whenever the data do not satisfy distributional assumptions and/or the homoscedasticity requirement, variable transformation is applied [5, 10]. Indexes of absolute and relative reliability (e.g., SEM, ICC) and related parameters (e.g., 95% limits of random variation, sample size) are finally estimated.

References

[1] Akselrod S, Gordon D, Ubel FA, Shannon DC, Berger AC, Cohen RJ. Power spectrum analysis of heart rate fluctuation: a quantitative probe of beat-to-beat cardiovascular control. *Science.* 1981; 213(4504): 220-222.

[2] Sandercock GR, Bromley PD, Brodie DA. The reliability of short-term measurements of heart rate variability. *Int J Cardiol.*. 2005; 103 (3): 238-247.

[3] Task force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. Heart rate variability. Standards of measurement, physiological interpretation and clinical use. *Circulation.* 1996; 93: 1043-1065.

[4] Schroeder EB, Whitsel EA, Evans GW, Prineas RJ, Chambless LE, Heiss G. Repeatability of heart rate variability measures. *J Electrocardiol.* 2004; 37 (3): 163-172.

[5] Atkinson G, Nevill AM. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Med.* 1998; 26 (4): 217-238.

[6] Pinna GD, Maestri R, La Rovere MT, Gobbi E, Fanfulla F. Effect of paced breathing on ventilatory and cardiovascular variability parameters during short-term investigations of autonomic function. *Am J Physiol Heart Circ Physiol.* 2006; 290 (1): H424-433.

[7] Maestri R, Pinna GD. POLYAN: a computer program for polyparametric analysis of cardio-respiratory variability signals. *Comput Methods Programs Biomed.* 1998; 56 (1): 37-48.

[8] Pinna GD, Maestri R, Di Cesare A. Application of time series spectral analysis theory: analysis of cardiovascular variability signals. *Med Biol Eng Comput.* 1996; 34 (2): 142-148.

[9] Wilcox RR. Fundamentals of Modern Statistical Methods. *Springer-Verlag New York.* 2001; Chapter 3, 33-36.

[10] Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res.* 1999; 8 (2): 135-160.

[11] Fleiss JL. The design and analysis of clinical experiments. *In: Wiley series in probability and mathematical statistics, John Wiley & Sons, New York.* 1986.

[12] Dunn G. Design and analysis of reliability studies. *Stat Methods Med Res.* 1992; 1: 123-157.

[13] Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psycological Bulletin.* 1979; 86 (2): 420-428.

[14] Desu MM, Raghavarao D. Sample size methodology. *In: Statistical modeling and decision science, Academic Press, Inc, San Diego.* 1990.

[15] Pinna GD, Maestri R, Sanarico M. Effects of record length selection on the accuracy of spectral estimates of heart rate variability: a simulation study. *IEEE Trans Biomed Eng.* 1996; 43 (7): 754-757.

[16] Pitzalis MV, Mastropasqua F, Massari F, et al. Short-and long-term reproducibility of time and frequency domain heart rate variability measurements in normal subjects. *Cardiovascular Research.* 1996; 32: 226-233.

[17] Lord SW, Senior RR, Das M, Whittam AM, Murray A, McComb JM. Low-frequency heart rate variability: reproducibility in cardiac transplant recipients and normal subjects. *Clin Sci (Lond).* 2001; 100 (1): 43-46.

[18] Donner A, Eliasziw M. Sample size requirements for reliability studies. *Stat Med.* 1987; 6: 441-448.

[19] Dunn G. Design and analysis of reliability studies. The statistical evaluation of measurement errors. *Oxford University Press, New York.* 1989.

[20] Sinnreich R, Kark JD, Friedlander Y, Sapoznikov D, Luria MH. Five minute recordings of heart rate variability for population studies: repeatability and age-sex characteristics. *Heart.* 1998; 80 (2): 156-162.

[21] Gerritsen J, TenVoorde BJ, Dekker JM, et al. Measures of cardiovascular autonomic nervous function: agreement, reproducibility, and reference values in middle age and elderly subjects. *Diabetologia.* 2003; 46 (3): 330-338.

[22] Sandercock GR, Bromley P, Brodie DA. Reliability of three commercially available heart rate variability instruments using short-term (5-min) recordings. *Clin Physiol Funct Imaging*. 2004; 24: 359-367

Clinical Science Immediate Publication. Published on 23 Mar 2007 as manuscript CS20070055

Table 1. Descriptive results for HRV parameters during spontaneous breathing in the two tests taken one day apart from each other.

	N	Measurement 1	Measurement 2	Difference (2-1)	p*
Mean RR (ms)	39	910 (115)	927 (128)	17 (80)	0.19
SDNN (ms)	39	43 (17)	46 (22)	3 (12)	
Ln SDNN (ln ms)	39	3.68 (0.39)	3.71 (0.48)	0.03 (0.26)	0.44
RMSSD (ms)	37	34 (23)	39 (24)	5 (21)	
Ln RMSSD (ln ms)	37	3.35 (0.58)	3.48 (0.64)	0.13 (0.41)	0.07
LF_BT (ms^2)	39	687 (583)	814 (759)	127 (604)	
Ln LF_BT (ln ms^2)	39	6.18 (0.88)	6.26 (1.01)	0.08 (0.62)	0.43
HF_BT (ms^2)	38	468 (542)	522 (653)	-54 (423)	
Ln HF_BT (ln ms^2)	38	5.60 (1.11)	5.63 (1.22)	0.03 (0.63)	0.74
LF_AR (ms^2)	39	684 (596)	804 (771)	120 (632)	
Ln LF_AR (ln ms^2)	39	6.16 (0.89)	6.23 (1.03)	0.07 (0.63)	0.51
HF_AR (ms^2)	38	464 (540)	518 (645)	54 (418)	
Ln HF_AR (ln ms^2)	38	5.58 (1.12)	5.62 (1.22)	0.04 (0.64)	0.71
LFnu (%)	38	62 (18)	63 (17)	1 (15)	0.66
LF/HF	38	2.5 (2.4)	2.5 (2.3)	0.0 (1.9)	
Ln LF/HF (ln)	38	0.56 (0.86)	0.59 (0.82)	0.03 (0.69)	0.76

Data are expressed as mean (SD). Skewed variables are also reported after log transformation.
SDNN=standard deviation of normal-to-normal beats; RMSSD=root mean square of successive differences; LF_BT, LF_AR= power in the low frequency band (0.04-0.15 Hz) according to the Blackman-Tukey (BT) or autoregressive (AR) method; HF_BT, HF_AR= power in the high frequency band (0.15-0.45 Hz) according to the Blackman-Tukey (BT) or autoregressive (AR) method; LF/HF= ratio between low frequency and high frequency power (AR method); Ln= natural logarithm.
One subject showed an outlier in RMSSD, HF_BT and HF_AR, while another one showed an outlier only in RMSSD. These measurements and those derived form them (i.e., LF_nu and LF/HF) were ignored.
*) p value for the test of no difference between the two measurements (one-sample t-test). For skewed variables the test was carried out after log-transformation.

Table 2. Descriptive results for HRV parameters during paced breathing in the two tests taken one day apart from each other.

	N	Measurement 1	Measurement 2	Difference (2-1)	p*
Mean RR (ms)	39	933 (130)	928 (124)	-4 (89)	0.76
SDNN (ms)	39	45 (23)	46 (21)	1 (15)	
Ln SDNN (ln ms)	39	3.68 (0.48)	3.74 (0.45)	0.05 (0.32)	0.30
RMSSD (ms)	38	41 (34)	42 (27)	1 (25)	
Ln RMSSD (ln ms)	38	3.49 (0.67)	3.56 (0.64)	0.07 (0.44)	0.30
LF_BT (ms^2)	39	602 (780)	641 (697)	39 (638)	
Ln LF_BT (ln ms^2)	39	5.88 (1.05)	5.94 (1.07)	0.06 (0.56)	0.48
HF_BT (ms^2)	39	1047 (1909)	944 (1429)	-103 (1028)	
Ln HF_BT (ln ms^2)	39	5.88 (1.05)	5.94 (1.07)	0.08 (0.70)	0.49
LF_AR (ms^2)	39	599 (787)	637 (707)	38 (653)	
Ln LF_AR (ln ms^2)	39	5.86 (1.09)	5.92 (1.10)	0.06 (0.57)	0.51
HF_AR (ms^2)	39	1044 (1885)	977 (1593)	-67 (987)	
Ln HF_AR (ln ms^2)	39	6.05 (1.36)	6.13 (1.29)	0.08 (0.70)	0.45
LFnu (%)	39	46 (22)	46 (22)	0 (14)	0.27
LF/HF	39	1.4 (1.4)	1.3 (1.2)	-0.1 (1.1)	
Ln LF/HF (ln)	39	-0.19 (1.08)	-0.22 (1.09)	-0.03 (0.62)	0.62

Data are expressed as mean (SD). Skewed variables are also reported after log transformation.
SDNN=standard deviation of normal-to-normal beats; RMSSD=root mean square of successive differences; LF_BT, LF_AR= power in the low frequency band (0.04-0.15 Hz) according to the Blackman-Tukey (BT) or autoregressive (AR) method; HF_BT, HF_AR= power in the high frequency band (0.15-0.45 Hz) according to the Blackman-Tukey (BT) or autoregressive (AR) method; LF/HF= ratio between low frequency and high frequency power (AR method); Ln= natural logarithm.
*) p value for the test of no difference between the two measurements (one-sample t-test). For skewed variables the test was carried out after log-transformation.

Table 3. Reliability indexes for homoscedastic heart rate variability (HRV) parameters.

a) Spontaneous breathing

HRV parameter	SEM	95% limits of random variation for the difference between 2 measurements ($X_2 - X_1$)	ICC (95% CI)	Change in the mean to be detected and required sample size (N)
Mean RR interval	57 ms	-159 ms, +159 ms	0.78 (0.62-0.88)	32 ms, N=52
LFnu	0.10	-0.29, +0.29	0.65 (0.43-0.80)	0.04, N=98

b) Paced breathing

HRV parameter	SEM	95% limits of random variation for the difference between 2 measurements ($X_2 - X_1$)	ICC (95% CI)	Change in the mean to be detected and required sample size (N)
Mean RR interval	62 ms	-173 ms, +173 ms	0.76 (0.59-0.87)	33 ms, N=58
LFnu	0.10	-0.26, +0.26	0.81 (0.67-0.90)	0.06, N=42

LFnu= low frequency power in normalized units. SEM= standard error of measurement; is the standard deviation of the random component of the measurement. The limits of random variation give the range of values within which 95% of the differences between the two measurements are expected to lie due to pure random variability. Therefore, when the observed difference lies outside this interval we can be 95% confident that a real change has occurred. ICC= intraclass correlation coefficient; is the proportion of measurement variability that is due to variability in the subjects' "true value", the remaining fraction being due to random variability. CI= confidence interval. The last column gives the estimate of the sample size needed to detect in a test-retest experiment a change in the mean of the parameter ≥30% of the between-subject standard deviation. In the computation of the sample size we assumed a two-tail test with a significance level of 5% and a power of 80%.

Table 4. Reliability of heteroscedastic heart rate variability (HRV) parameters.

a) Spontaneous breathing

HRV parameter	SEM	95% limits of random variation of the ratio between 2 measurements (X_2/X_1)	ICC (95% CI)	Change in the mean to be detected and required sample size (N)
SDNN	0.18 ln ms	0.60, 1.67	0.82 (0.68-0.90)	0.12 ln ms, N=40
RMSSD	0.30 ln ms	0.44, 2.29	0.76 (0.59-0.87)	0.16 ln ms, N=57
LF_BT	0.43 ln ms^2	0.30, 3.32	0.79 (0.64-0.88)	0.25 ln ms^2, N=48
HF_BT	0.44 ln ms^2	0.30, 3.36	0.86 (0.75-0.92)	0.32 ln ms^2, N=30
LF_AR	0.44 ln ms^2	0.29, 3.42	0.79 (0.64-0.88)	0.26 ln ms^2, N=49
HF_AR	0.45 ln ms^2	0.29, 3.44	0.86 (0.74-0.92)	0.33 ln ms^2, N=31
LF/HF	0.45 ln	0.28, 3.53	0.70 (0.50-0.84)	0.21 ln, N=77

b) Paced breathing

HRV parameter	SEM	95% limits of random variation of the ratio between 2 measurements (X_2/X_1)	ICC (95% CI)	Change in the mean to be detected and required sample size (N)
SDNN	0.23 ln ms	0.54, 1.87	0.77 (0.60-0.87)	0.12 ln ms, N=55
RMSSD	0.31 ln ms	0.42, 2.37	0.77 (0.61-0.87)	0.17 ln ms, N=54
LF_BT	0.40 ln ms^2	0.33, 3.00	0.86 (0.75-0.92)	0.30 ln ms^2, N=29
HF_BT	0.45 ln ms^2	0.28, 3.51	0.88 (0.78-0.94)	0.37 ln ms^2, N=25
LF_AR	0.40 ln ms^2	0.33, 3.04	0.86 (0.76-0.93)	0.30 ln ms^2, N=29
HF_AR	0.45 ln ms^2	0.29, 3.50	0.88 (0.79-0.94)	0.38 ln ms^2, N=24
LF/HF	0.39 ln	0.34, 2.94	0.87 (0.76-0.93)	0.30 ln, N=28

The legend of HRV parameters is given in table 1. SEM= standard error of measurement; is the

standard deviation of the random component of the measurement (see appendix) and was computed on log-transformed data. The limits of random variation give the range of values within which 95% of the ratios between the two measurements are expected to lie due to pure random variability. Therefore, when the observed ratio lies outside this interval, we can be 95% confident that a real change has occurred. ICC= intraclass correlation coefficient; is the proportion of measurement variability that is due to variability in the subjects' true value, the remaining fraction being due to random variability. CI= confidence intervals. The last column gives the estimate of the sample size needed to detect in a test-retest experiment a change in the mean of the parameter ≥30% of the between-subject standard deviation. In the computation of the sample size we assumed a two-tail test with a significance level of 5% and a power of 80%.

Figure legends

Fig. 1. Representative Bland-Altman plots of the difference between the two measurements ($X_2 - X_1$) against their average: (a) mean RR, (b) LF power (LF_BT, Blackman-Tukey method). The scatter of points is symmetrical around the zero line in both parameters, indicating absence of systematic change between the two tests. The magnitude of the scatter around the same line, however, is pretty homogeneous for mean RR, indicating homoscedasticity, while it steadily increases for LF power, indicating heteroscedasticity. Figure 1 (c) shows how log-transformation (natural logarithm, ln) of the LF power was successful in producing homoscedasticity. All plotted data are from recordings during spontaneous breathing.

Fig. 2. Plots of the ratio between the two measurements $\left(X_2 / X_1\right)$ against their mean for heteroscedastic heart rate variability parameters. SDNN=standard deviation of normal-to-normal beats; RMSSD=root mean square of successive differences; LF_BT, LF_AR= power in the low frequency band (0.04-0.15 Hz) according to the Blackman-Tukey (BT) and autoregressive (AR) method; HF_BT, HF_AR= power in the high frequency band (0.15-0.45 Hz); LF/HF= ratio between low frequency and high frequency power (AR method). All plotted data are from recordings during spontaneous breathing.

Clinical Science Immediate Publication. Published on 23 Mar 2007 as manuscript CS20070055

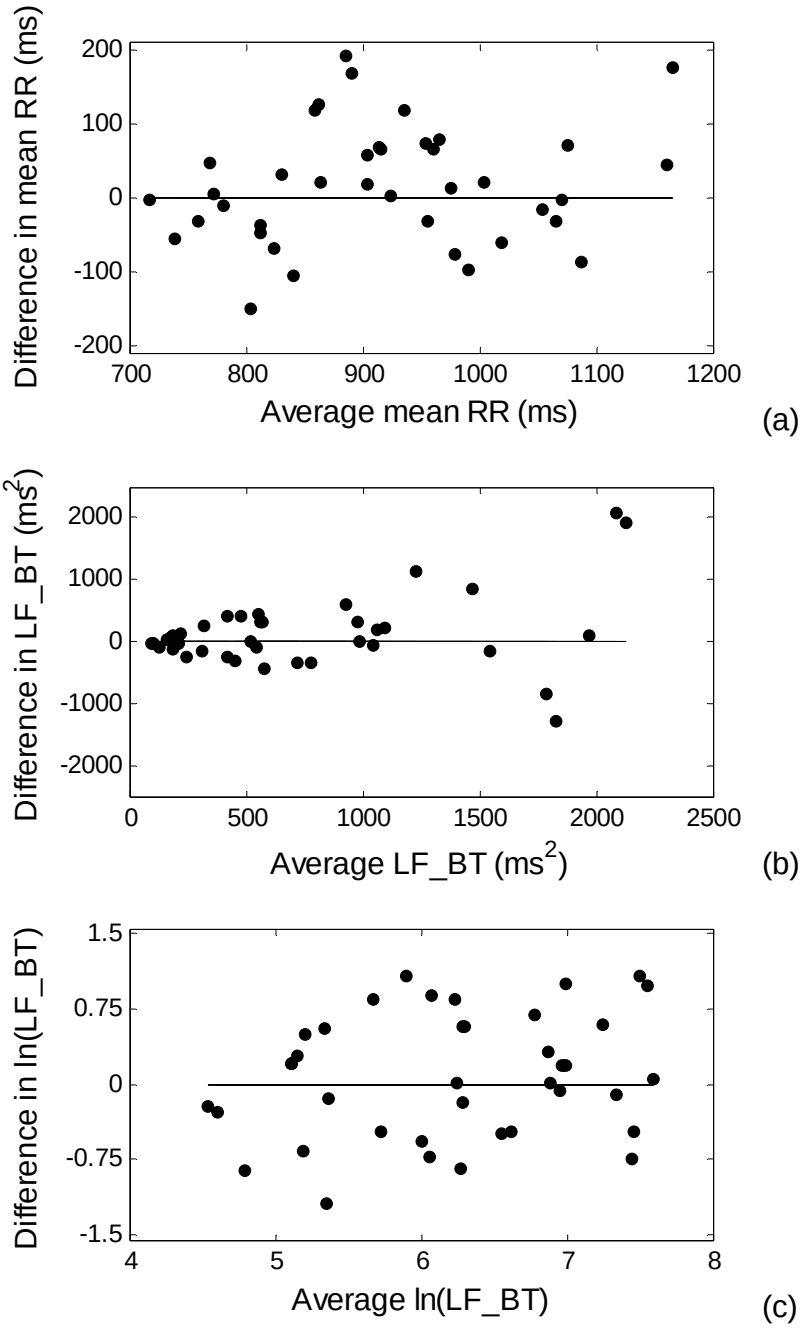

Figure 1

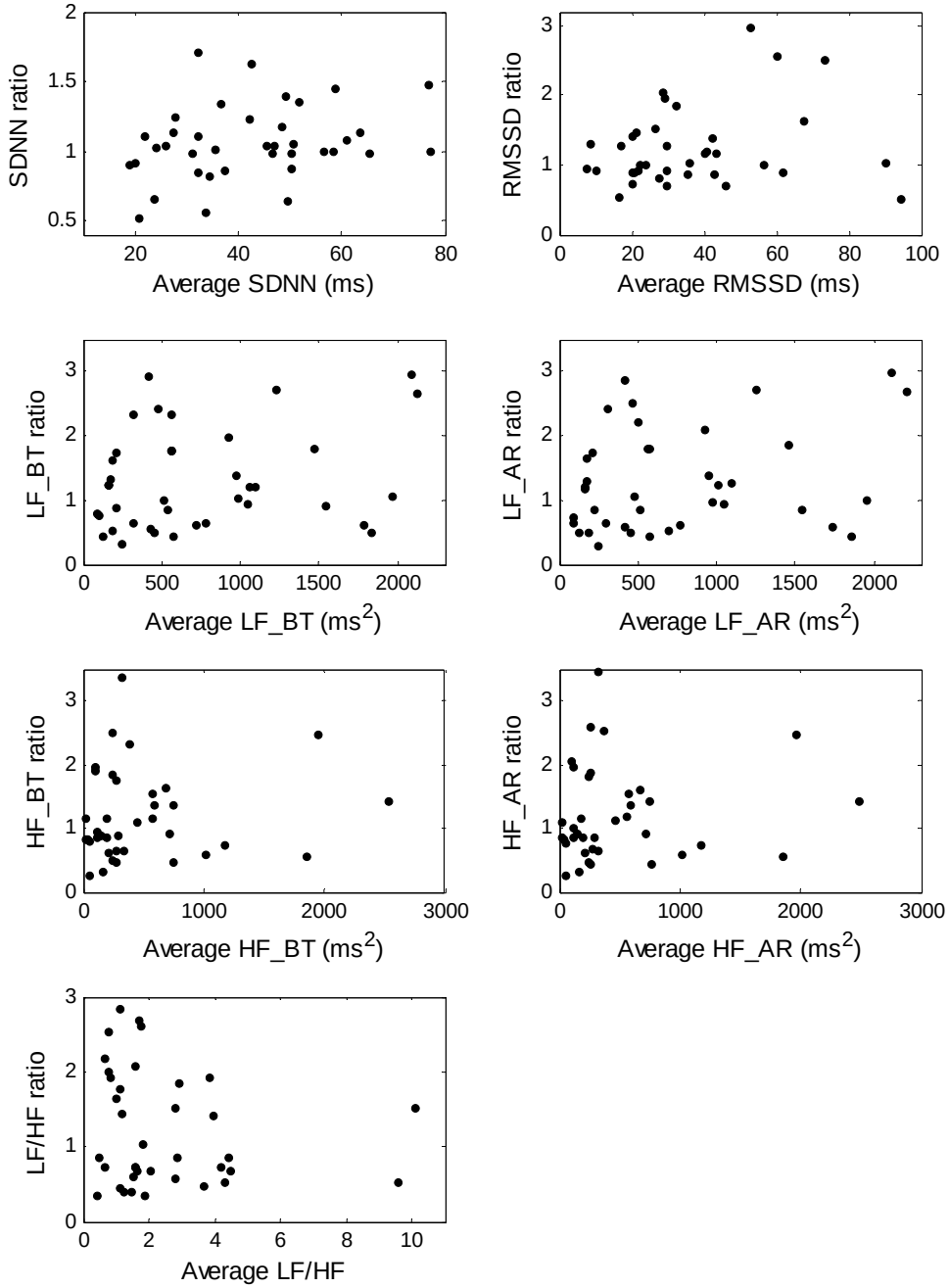

Figure 2