Heart rate variability measures: a fresh look at reliability

Gian Domenico PINNA*, Roberto MAESTRI*, Antoni TORUNSKI†, Ludmila DANILOWICZ-SZYMANOWICZ†, Małgorzata SZWOCH†, Maria Teresa LA ROVERE* and Grzegorz RACZAK†

*Department of Cardiology and Biomedical Engineering, S. Maugeri Foundation-IRCCS, Scientific Institute of Montescano, Montescano (PV), Italy, and †II Department of Cardiology, Medical University, Gdansk, Poland

ABSTRACT

Despite its extensive use in physiological and clinical research, the analysis of HRV (heart rate variability) is still poorly supported by rigorous reliability studies. The main aim of the present study was to perform 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 min of supine ECG during spontaneous and paced (15 breaths/min) breathing. The test was repeated on the next day under the same conditions. From the RR intervals we computed standard indexes of HRV: SDNN (S.D. of RR interval values), RMSSD (root-mean-square of successive RR interval differences), LF (low frequency) and HF (high frequency) power (absolute and normalized units) and LF/HF. Absolute reliability was assessed by 95% limits of random variation; relative reliability was assessed by the ICC (intraclass correlation coefficient). The sample size needed to detect a mean difference ⩾30% of the between-subject S.D. was also estimated. Although there was no significant mean change between the two tests, we found that in individual subjects the second measurement was 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 24–98 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 part of the between-subject variability; therefore observed differences between individuals mostly reflect 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], HRV (heart rate variability) is still a matter of great interest in clinical and physiological research. According to the PubMed database (www.pubmed.com), from 2000 to 2006 the number of yearly publications related to HRV has increased steadily from 391 to 584, and this trend is going to be continued 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.

Key words: ECG, heart rate variability, high-frequency power, low-frequency power, paced breathing, RR interval.

Abbreviations: BMS, between-subject mean square; HF, high frequency; HF_AR, HF power according to the autoregressive method; HF_BT, HF power according to the Blackman–Tukey method; HRV, heart rate variability; ICC, intraclass correlation coefficient; LF, low frequency; LF_AR, LF power according to the autoregressive method; LF_BT, LF power according to the Blackman–Tukey method; LFnu, LF power in normalized units; RMSSD, root-mean-square of successive RR interval differences; SDNN, S.D. of RR interval values; SEM, standard error of measurement; WMS, within-subject mean square.

Correspondence: Dr Gian Domenico Pinna (email gdpinna@fsm.it).
which are those most commonly used in non-invasive investigations of autonomic cardiovascular control. Major methodological limitations of published reliability studies include: (i) inadequate protocol (e.g. replicate measurements were taken too far from each other); (ii) insufficient sample size; (iii) too short or too long recordings according to current guidelines [3]; (iv) limited selection of studied HRV parameters; (v) inadequate assessment of reliability due to the use of inappropriate reliability indexes or to misapplication/incomplete inclusion of appropriate indexes; and (vi) 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 papers published previously [2,4,5].

In the present 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.

MATERIALS AND METHODS

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

Study protocol
All tests were performed between 08.00 and 13.00 hours, with the subjects in the supine position in a quiet and dimmed room at a comfortable temperature. Before the study all subjects fasted for more than 2 h and refrained from smoking, alcohol or coffee for 24 h.

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. Stabilization of the signals was then performed for 15 min, following which subjects breathed spontaneously for 8 min, and then breathed at the paced breathing frequency for another 8 min. During these two sessions ECG recordings were performed. An identical session was repeated on the next day at the same time of day.

Signal analysis and measurements
Beat-by-beat RR interval values (resolution 1 ms) were obtained from the ECG using a software package developed in-house [7]. The RR interval time series was 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 in short-term HRV [3].

After detrending via least-square second-order polynomial fitting, the power spectral density of the RR interval 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 LF (low frequency; 0.04–0.15 Hz) and HF (high frequency; 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). The 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 (termed LF_AR and HF_AR respectively) were computed by 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 LFnu (LF power in normalized units) was computed as LF_AR/(LF_AR + HF_AR). To avoid redundancy, the HF power in normalized units, being equal to 1 – LFnu, was not analysed.

We also computed time-domain HRV parameters suitable for short-term analysis, including SDNN (S.D. of RR interval values) and RMSSD (root-mean-square of successive RR interval differences) [3]. Although the 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
A short introduction to the statistical basis for the assessment of reliability is described in the Appendix. The key point is that observed measurements were 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 the 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 +− 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
Reliability of heart rate variability measures

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

Values are expressed as means (S.D.). Skewed variables are also reported after log transformation. One subject had an outlier in RMSSD, HF_BT and HF_AR, whereas another one had an outlier only in RMSSD. These measurements and those derived from them (i.e. LF_nu and LF/HF) were ignored. * P value for the test of no difference between the two measurements (paired Student’s t test). For skewed variables the test was carried out after log-transformation.

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>Measurement 1 (X₁)</th>
<th>Measurement 2 (X₂)</th>
<th>Difference (X₂ − X₁)</th>
<th>* P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean RR (ms)</td>
<td>39</td>
<td>910 (115)</td>
<td>927 (128)</td>
<td>17 (80)</td>
<td>0.19</td>
</tr>
<tr>
<td>SDNN (ms)</td>
<td>39</td>
<td>43 (17)</td>
<td>46 (22)</td>
<td>3 (12)</td>
<td></td>
</tr>
<tr>
<td>Ln SDNN (ln ms)</td>
<td>39</td>
<td>3.68 (0.39)</td>
<td>3.71 (0.40)</td>
<td>0.03 (0.26)</td>
<td>0.44</td>
</tr>
<tr>
<td>RMSSD (ms)</td>
<td>37</td>
<td>34 (23)</td>
<td>39 (24)</td>
<td>5 (21)</td>
<td></td>
</tr>
<tr>
<td>Ln RMSSD (ln ms)</td>
<td>37</td>
<td>3.35 (0.58)</td>
<td>3.48 (0.64)</td>
<td>0.13 (0.41)</td>
<td>0.07</td>
</tr>
<tr>
<td>LF_BT (ms²)</td>
<td>39</td>
<td>687 (583)</td>
<td>814 (759)</td>
<td>127 (604)</td>
<td></td>
</tr>
<tr>
<td>Ln LF_BT (ln ms²)</td>
<td>39</td>
<td>6.18 (0.88)</td>
<td>6.26 (1.01)</td>
<td>0.08 (0.62)</td>
<td>0.43</td>
</tr>
<tr>
<td>HF_BT (ms²)</td>
<td>38</td>
<td>468 (542)</td>
<td>522 (653)</td>
<td>− 54 (423)</td>
<td></td>
</tr>
<tr>
<td>Ln HF_BT (ln ms²)</td>
<td>38</td>
<td>5.60 (1.11)</td>
<td>5.63 (1.22)</td>
<td>0.03 (0.63)</td>
<td>0.74</td>
</tr>
<tr>
<td>LF_AR (ms²)</td>
<td>39</td>
<td>684 (596)</td>
<td>804 (771)</td>
<td>120 (632)</td>
<td></td>
</tr>
<tr>
<td>Ln LF_AR (ln ms²)</td>
<td>39</td>
<td>6.16 (0.89)</td>
<td>6.23 (1.03)</td>
<td>0.07 (0.63)</td>
<td>0.51</td>
</tr>
<tr>
<td>HF_AR (ms²)</td>
<td>38</td>
<td>464 (540)</td>
<td>518 (645)</td>
<td>54 (418)</td>
<td></td>
</tr>
<tr>
<td>Ln HF_AR (ln ms²)</td>
<td>38</td>
<td>5.58 (1.12)</td>
<td>5.62 (1.22)</td>
<td>0.04 (0.64)</td>
<td>0.71</td>
</tr>
<tr>
<td>LFnu (%)</td>
<td>38</td>
<td>62 (18)</td>
<td>63 (17)</td>
<td>1 (15)</td>
<td>0.66</td>
</tr>
<tr>
<td>LF/HF</td>
<td>38</td>
<td>2.5 (2.4)</td>
<td>2.5 (2.3)</td>
<td>0.0 (1.9)</td>
<td></td>
</tr>
<tr>
<td>Ln LF/HF (ln)</td>
<td>38</td>
<td>0.56 (0.86)</td>
<td>0.59 (0.82)</td>
<td>0.03 (0.69)</td>
<td>0.76</td>
</tr>
</tbody>
</table>

method allows for any systematic change between the tests to be observed 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 Student’s 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].

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

One-way random effects ANOVA was carried out on all variables to estimate the SEM (standard error of measurement), the major index of absolute reliability. This was obtained as the square root of the WMS (within-subject mean square) from the ANOVA table [11,12] as:

\[
\text{ICC} = \frac{\text{BMS} - \text{WMS}}{\text{BMS} + \text{WMS}}
\]

where BMS is the between-subject mean square. 95% 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 a change of ≥ 30% of between-subject S.D as ‘relevant’. 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 version 8.02 (SAS Institute).

RESULTS

Spontaneous breathing

All variables with the exception of mean RR interval and LFnu had a marked right-skewed distribution (P < 0.01, as determined by the Shapiro–Wilk test). One subject had an outlier in RMSSD, HF_BT and HF_AR, whereas another one had an outlier only in RMSSD. These measurements and those derived from them (i.e. LFnu and LF/HF) were ignored in the subsequent analyses. Descriptive statistics of parameters measured in the two tests are shown in Table 1.
For all HRV parameters the difference between the two tests was negligible and non-significant, indicating the absence of a systematic change.

Reliability indexes for homoscedastic variables (mean RR interval and LFnu) are shown in Table 2. If we consider LFnu, the limits of random variation indicated that, in order to be 95% confident that a real change has occurred in an individual, the observed difference between two measurements had to be $>0.29$ or $<-0.29$. For the same parameter, the ICC indicated that 65% of the variability in LFnu measurements across the population studied was due to variability in the true value of the subjects, whereas the remaining 35% was due to random error. The last column of Table 2 shows that 98 subjects were needed in a test–retest experiment to detect a change in mean LFnu of $\geq 0.04$ (30% of between-subject S.D.), with a significance level of 5% and a power of 80%.

Reliability indexes for heteroscedastic HRV parameters are shown in Table 3. 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 were expressed in log units. The limits of random variation indicated that, in order to be 95% confident that a real change had occurred in an individual, the ratio between two measurements $X_2/X_1$ had to lie outside the indicated interval. The asymmetry of this interval was 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 approx. $3.5/0.3$ times the first measurement due to pure random variation. To relate these inferential figures to the data of the present 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 the raw data.

The ICC values in Table 3 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 shown in Table 4. 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 heteroscedastical behaviour and were successfully log-transformed. The difference between transformed measurements was largely non-significant.

Reliability indexes for homoscedastic and heteroscedastic variables are shown in Tables 2 and 3 respectively. Compared with spontaneous breathing we found: (i) an improvement of all reliability indexes (SEM, limits of
Table 2  Reliability indexes for homoscedastic HRV parameters
SEM is the S.D. 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 is the proportion of measurement variability that is due to variability in the subjects’ ‘true value’, the remaining fraction being due to random variability. 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 S.D. In the computation of the sample size we assumed a two-tailed test with a significance level of 5% and a power of 80%. CI, confidence interval.

<table>
<thead>
<tr>
<th>HRV parameter</th>
<th>SEM</th>
<th>95% limits of random variation of the difference between two measurements (X2 − X1)</th>
<th>ICC (95% CI)</th>
<th>Change in the mean to be detected and required sample size (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean RR interval</td>
<td>57 ms</td>
<td>−159 ms, +159 ms</td>
<td>0.78 (0.62–0.88)</td>
<td>32 ms, (n = 52)</td>
</tr>
<tr>
<td>LFnu</td>
<td>0.10</td>
<td>−0.29, +0.29</td>
<td>0.65 (0.43–0.80)</td>
<td>0.04, (n = 98)</td>
</tr>
</tbody>
</table>

Paced breathing

| 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) |

Table 3  Reliability of heteroscedastic HRV parameters
SEM is the S.D. of the random component of the measurement 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 is the proportion of measurement variability that is due to variability in the subjects’ ‘true value’, the remaining fraction being due to random variability. 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 S.D. In the computation of the sample size we assumed a two-tailed test with a significance level of 5% and a power of 80%.

<table>
<thead>
<tr>
<th>HRV parameter</th>
<th>SEM</th>
<th>95% limits of random variation of the ratio between two measurements (X2/X1)</th>
<th>ICC (95% CI)</th>
<th>Change in the mean to be detected and required sample size (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDNN</td>
<td>0.18 ln ms</td>
<td>0.60, 1.67</td>
<td>0.82 (0.68–0.90)</td>
<td>0.12 ln ms, (n = 40)</td>
</tr>
<tr>
<td>RMSSD</td>
<td>0.30 ln ms</td>
<td>0.44, 2.29</td>
<td>0.76 (0.59–0.87)</td>
<td>0.16 ln ms, (n = 57)</td>
</tr>
<tr>
<td>LF_BT</td>
<td>0.43 ln ms²</td>
<td>0.30, 3.32</td>
<td>0.79 (0.64–0.88)</td>
<td>0.25 ln ms², (n = 48)</td>
</tr>
<tr>
<td>HF_BT</td>
<td>0.44 ln ms²</td>
<td>0.30, 3.36</td>
<td>0.86 (0.75–0.92)</td>
<td>0.32 ln ms², (n = 30)</td>
</tr>
<tr>
<td>LF_AR</td>
<td>0.44 ln ms²</td>
<td>0.29, 3.42</td>
<td>0.79 (0.64–0.88)</td>
<td>0.26 ln ms², (n = 49)</td>
</tr>
<tr>
<td>HF_AR</td>
<td>0.49 ln ms²</td>
<td>0.29, 3.44</td>
<td>0.86 (0.74–0.92)</td>
<td>0.33 ln ms², (n = 31)</td>
</tr>
<tr>
<td>LF/HF</td>
<td>0.45 ln</td>
<td>0.28, 3.53</td>
<td>0.70 (0.50–0.84)</td>
<td>0.21 ln, (n = 77)</td>
</tr>
</tbody>
</table>

Paced breathing

| 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 ms² | 0.33, 3.00 | 0.86 (0.75–0.92) | 0.30 ln ms², (n = 29) |
| HF_BT | 0.45 ln ms² | 0.28, 3.51 | 0.88 (0.78–0.94) | 0.37 ln ms², (n = 25) |
| LF_AR | 0.40 ln ms² | 0.33, 3.04 | 0.86 (0.76–0.93) | 0.30 ln ms², (n = 29) |
| HF_AR | 0.45 ln ms² | 0.29, 3.50 | 0.88 (0.79–0.94) | 0.38 ln ms², (n = 24) |
| LF/HF | 0.39 ln | 0.34, 2.94 | 0.87 (0.76–0.93) | 0.30 ln, (n = 28) |

random variation, ICC and sample size) for LF_BT, LF_AR and LF/HF; (ii) an increased ICC and reduced sample size for LFnu; and (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 have recently been advocated [2]. In the present study we performed 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**

The present 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 approx. 3.5 times and decreases of as much as approx. 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 approx. 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, except 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 S.D.), 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 0.65–0.88. Although the definition of a categorical rating of relative reliability based on ICC is still controversial, these values can reasonably be considered to indicate 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 (see 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 the 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].

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 et 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 the 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 S.D. 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 vary largely 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 compared with paced breathing**

Paced breathing substantially improved the reliability of LFnu and LF/HF, with a consequent dramatic reduction (>50%) in the estimated sample size. This result is in agreement with the findings of Pitzalis et al. [16]. Moreover, voluntary control of breathing moderately improved the reliability of LF power, while leaving substantially unchanged that of 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 compared with autoregressive spectral estimation**

Our present study suggests that the reliability of HRV parameters is not affected by the method used for spectral estimation (classical Blackman–Tukey or autoregressive). A similar result has been 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 the present study may yield different reliability figures.

**Conclusions**

In healthy subjects, short-term HRV parameters are subject to large day-to-day random variations which would make the detection of treatment effects in individual subjects difficult. For most indexes, however, random variation represents a limited portion of the between-subject variability; therefore observed differences between individuals mostly reflect differences in the subjects’ true value rather than random error. The sample size for an experiment depends markedly 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.

**APPENDIX**

**Statistical basis for the assessment of reliability**

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

\[ X = \tau + \varepsilon \]  

(1)

In this equation, the measurement \( X \) made on a given subject is assumed to be the sum of a fixed quantity \( \tau \), the ‘true value’ or ‘error-free value’ of the characteristic being measured in that subject, and a random quantity \( \varepsilon \), commonly referred to as random error of measurement, which accounts for intra-subject variability.

There are some basic assumptions underlying eqn (1): the random error \( \varepsilon \) is normally distributed, has zero mean, is uncorrelated within and between subjects, and has a fixed S.D. independent of \( \tau \), the latter property being commonly referred to as homoscedasticity [5]. Moreover, for a population of subjects, \( \tau \) is assumed to be normally distributed. Before analysing 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 eqn (1) we have that the difference \( \Delta X \) between them will be:

\[ \Delta X = X_2 - X_1 = (\tau + \varepsilon_2) - (\tau + \varepsilon_1) = \varepsilon_2 - \varepsilon_1 = \delta \]  

(2)

where the suffixes 2 and 1 indicate the measurements at the two occasions. Eqn (2) clearly shows that the repeatability of the measurement depends on the random quantity \( \delta \): the lower the \( \delta \), the closer the two measurements will be to each other. The magnitude of \( \delta \), as expressed by its S.D., is \( \sqrt{2 \cdot \sigma_{\delta}} \), where \( \sigma_{\delta} \), the so-called SEM, is the S.D. of \( \varepsilon \). 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 \( -1.96 \cdot \sqrt{2} \cdot \text{SEM} + 1.96 \cdot \sqrt{2} \cdot \text{SEM} \), or \( -2.77 \cdot \text{SEM} + 2.77 \cdot \text{SEM} \) [5,10]. Therefore the extremes of this interval can be viewed as the 95% limits of random variation. The value \( 2.77 \cdot \text{SEM} \) is called repeatability coefficient [10]. Secondly, 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 ICC, which is also commonly referred to as the 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]:

\[
\text{ICC} = \frac{\sigma^2_x}{\sigma^2_x + \sigma^2_{\epsilon}} = \frac{1}{1 + \tau^2/\sigma^2_{\epsilon}} \tag{3}
\]

where \(\sigma^2_x\) is the total variance of observed measurements among the subjects of the considered population, \(\sigma^2_{\epsilon}\) is the part of total variance due to differences in the patients’ true value and \(\sigma^2_{\epsilon}\) 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 probable differences between individuals will be detected. 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, whereas an ICC between 0.6 and 0.8 may be taken to represent substantial reliability. From eqn (3) it clearly appears that the ICC, besides being dependent on the random variability of the measurement, depends on the variability (i.e. the heterogeneity) of the population being studied. Therefore results obtained in one population cannot be extrapolated to a new and 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 for 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 and ICC) and related parameters (e.g. 95% limits of random variation and sample size) are finally estimated.

### REFERENCES


Received 13 February 2007/19 March 2007; accepted 23 March 2007
Published as Immediate Publication 23 March 2007, doi:10.1042/CS20070055