A methodology to quantify the differences between alternative methods of heart rate variability measurement

2016 Physiol. Meas. 37 128

(http://iopscience.iop.org/0967-3334/37/1/128)

OPEN ACCESS
IOP Publishing | Institute of Physics and Engineering in Medicine Physiological Measurement
Physiol. Meas. **37** (2016) 128–144 doi:10.1088/0967-3334/37/1/128

A methodology to quantify the differences between alternative methods of heart rate variability measurement

M A García-González[1], M Fernández-Chimeno[1], F Guede-Fernández[1], V Ferrer-Mileo[1], A Argelagós-Palau[1], L Álvarez-Gómez[1], E Parrado[2], J Moreno[2], L Capdevila[2] and J Ramos-Castro[1]

[1] Group of Biomedical and Electronic Instrumentation at the Department of Electronic Engineering of the Universitat Politècnica de Catalunya BARCELONATECH (UPC), C/ Jordi Girona 1–3, Edifici C-4, 08034 Barcelona, Spain
[2] Laboratory of Sport Psychology, Universitat Autonoma de Barcelona, Edifici B, 08193, Bellaterra, Spain

E-mail: miquel.angel.garcia@upc.edu

Received 27 July 2015, revised 9 October 2015
Accepted for publication 3 November 2015
Published 14 December 2015

Abstract

This work proposes a systematic procedure to report the differences between heart rate variability time series obtained from alternative measurements reporting the spread and mean of the differences as well as the agreement between measuring procedures and quantifying how stationary, random and normal the differences between alternative measurements are. A description of the complete automatic procedure to obtain a differences time series (DTS) from two alternative methods, a proposal of a battery of statistical tests, and a set of statistical indicators to better describe the differences in *RR* interval estimation are also provided. Results show that the spread and agreement depend on the choice of alternative measurements and that the DTS cannot be considered generally as a white or as a normally distributed process. Nevertheless, in controlled measurements the DTS can be considered as a stationary process.

Keywords: heart rate variability, surrogate measurements, agreement, stationarity, normal distribution, randomness

(Some figures may appear in colour only in the online journal)

1. Introduction

Heart rate variability (HRV) has become a popular research tool because it reflects changes in the cardiac autonomic regulation (Billman *et al* 2015). Hence, HRV is considered as a window to the autonomic nervous system (ANS). There are several approaches to HRV analysis but most start with the estimation of the *RR* time series that is the series obtained from the successive time intervals between consecutive QRS complexes using a properly acquired electrocardiogram (ECG) (Task Force 1996) or some sort of surrogate fiducial point such as the distal pulse wave (Lu *et al* 2008) or the vibrocardiogram (Ramos-Castro *et al* 2012).

The advent of smartphones, smartwatches and wearable sensors has created a demand for biomedical measurement solutions that ideally must be non-obtrusive, robust to movement and accurate enough (Milošević *et al* 2011). Such measurement methods often are a surrogate measure or a poor-quality estimate of the ECG so, necessarily, there are differences between the time series obtained using the ECG or the new measurement method. The problem is that there are no guidelines or consensus on how to document or quantify the differences between these time series. This problem becomes of paramount importance taking into account that the number of approaches to quantify HRV grows every year as well as the demands of high precision and resolution of the time series for certain methods does too (Garcia-Gonzalez *et al* 2004, Garcia-Gonzalez *et al* 2009). Moreover, the sensitivity of most methods to differences in the estimation of the time series depends on the dynamic changes of these differences and furthermore, some sources of errors between estimated HRV time series cannot be modelled as white normal processes. As an example, the error in a *RR* time series obtained from an ECG sampled at a sampling frequency f_s using a perfect QRS detector can be modelled as a random process with power spectrum density (Merri *et al* 1990)

$$S_{ee}(f) = \frac{1}{6 \cdot f_s^2} \cdot [1 - \cos(2 \cdot \pi \cdot f)] \quad |f| \leqslant 0.5 \tag{1}$$

where f is in cycles/beat as defined in (Lisenby and Richardson 1977). This sampling error has a higher spectral density on higher frequencies and accordingly has a greater effect on indexes that reflect fast changes in HRV than in indexes that reflect slow changes.

Currently there are two ways (that in some studies intersect) to document differences in the estimation of HRV parameters using different methods to obtain the *RR* time series. The first approach measures the difference for each obtained heart period and summarizes it by its mean, standard deviation and/or level of agreement (Parrado *et al* 2010, Gamelin *et al* 2006). This approach does not provide useful information as how the differences are distributed or if they can be considered as white noise so the effect on different HRV indexes is difficult to judge. Nevertheless, a more in depth analysis of differences between two *RR* time series measured with two different methods can provide useful information on validating new HRV measurement methods. The second approach quantifies the HRV using certain (very often standardized) indexes and compares the differences in the indexes (once again, mean, standard deviation and/or level of agreement) when using the two measurement methods (Nunan *et al* 2008, Gil *et al* 2010) This approach does not provide information on the effect of choosing one method instead of the other for any not analyzed indexes so it is difficult to extrapolate the results to novel indexes. So, more detailed information to document the differences between *RR* time series obtained simultaneously using different measurement methods is needed. This information can be very useful not only to validate alternative methods but also to simulate the effect of these differences when proposing novel HRV indexes.

The aim of this work is to propose a methodology to report not only the spread and mean of the differences in *RR* time series when using different heartbeat detection methods but if these

differences can be considered stationary, white and/or normal random processes. The paper focuses on the validation of alternative or surrogate *RR* time series measurement procedures when a trusted measurement procedure is available.

2. Materials and methods

2.1. Proposal of a methodology for differences documentation

We will say that two HRV measurement methods are interchangeable when the bias and standard deviation of the differences are low enough to not affect the interpretation of results. Moreover, it is often desirable that the differences can be reasonably modelled as a stationary white normal process to simplify the analysis of discrepancies between measurement methods. Nevertheless, any additional knowledge on the properties of these differences may add to a better understanding of their effect on any HRV index. With this rationale in mind, we propose a set of procedures and tests for differences among measurement methods characterization and documentation.

Let us suppose two different methods (a and b) are employed to measure the *RR* time series for the same individual at approximately the same time. We define as $RR_a(n)$ the *RR* time series obtained by method a (where n ranges from 1 to N) while $RR_b(m)$ is the corresponding time series using a different method b (m ranges from 1 to M). We deliberately express each series indicating the beats n and m because both time series may not start and/or not end at the same time so the number of beats in each time series may not be equal. Figure 1 shows our recommended proposed procedure to document the measurable differences between both time series. This systematic procedure can be applied to any experiment intended to validate new different *RR* time series measurement methods when a trusted one is also simultaneously employed.

Because misdetections on heart period may happen with any of the measurement methods, the first step of the procedure is to recognize outliers in both *RR* time series and correct them. The correction must be done in a way that after proper synchronization of time series and removal of periods of no simultaneous *RR* detections (it may happen that a method starts prior or ends after the other) there would be a consecutive correspondence of estimated *RR* time intervals so the differences can be properly computed by simply applying:

$$\mathrm{DTS}(i) = RRc_a(n_s + i) - RRc_b(m_s + i) \qquad \forall\, i \in [1, I] \tag{2}$$

In (2), n_s and m_s are constants determined by the synchronization procedure and the differences time series (DTS) is evaluated for I consecutive *RR* intervals. Figure 2 depicts the need for the correction and synchronization steps. Because the correction of artifacts in each *RR* time series is not always perfect (this is true for ectopic beats and missing beats), the differences involving *RR* intervals may be very high. In order to avoid this, the next step is the correction of outliers in the DTS. Once an acceptable DTS is available, its characterization starts. The stationarity tests aim to ascertain if the DTS is stable, at least, in mean and variance so the bias and dispersion have any meaning. The randomness tests try to measure the degree of whiteness of the DTS. The normality test is necessary when in further studies confidence intervals of the differences must be estimated from the standard deviation or vice versa. Preferably, a sample distribution for the DTS should be provided. Finally, the DTS quantification, as stated in the introduction, is a very often reported measure of differences in HRV applications and consists in the quantification of the bias and dispersion of the DTS. In some studies, this quantification reports levels of agreement or confidence intervals.

There are several tests and procedures to complete all the steps of figure 1. In the next section we recommend some of them and describe thoroughly our employed algorithm for DTS

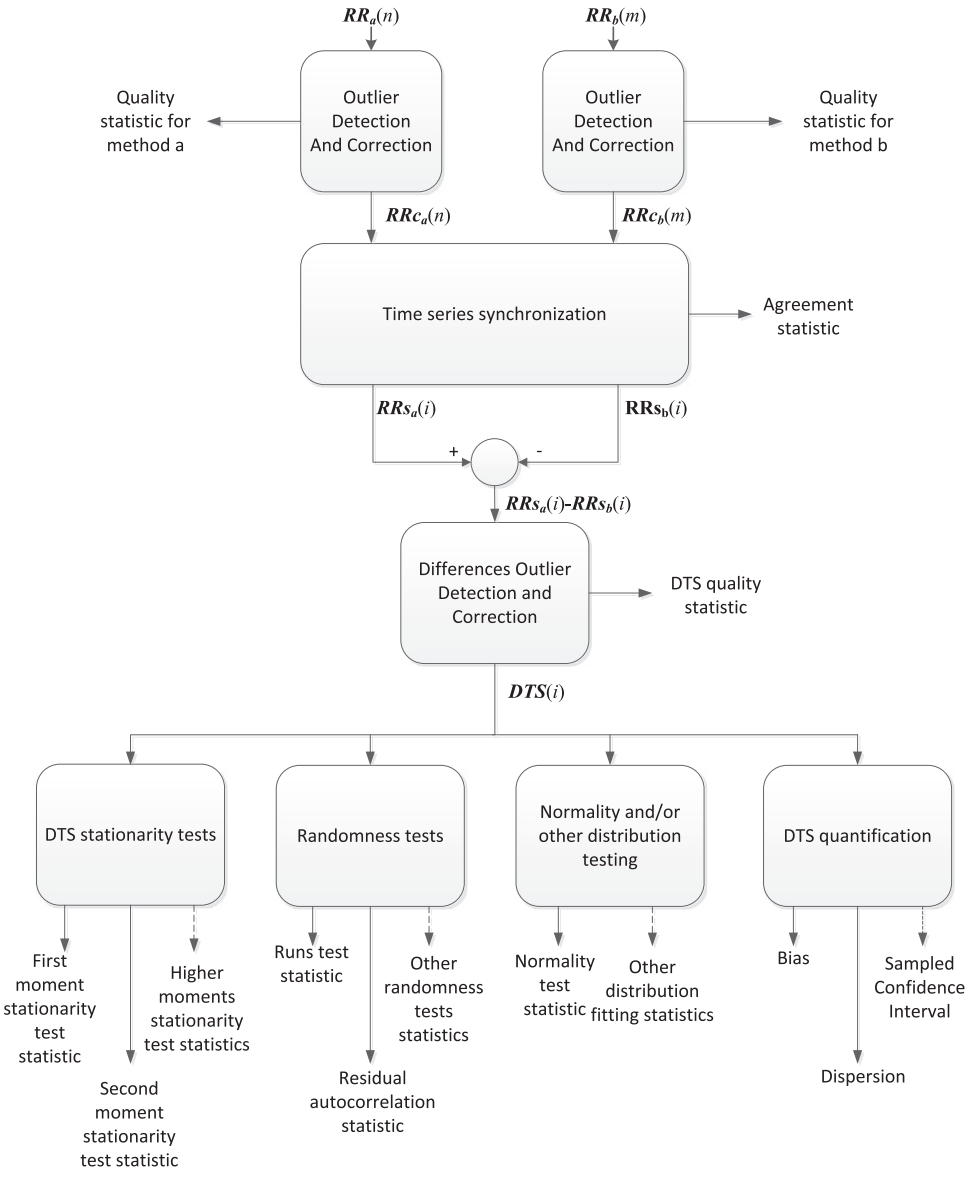

Figure 1. Recommended procedure to obtain and document the differences in *RR* time series obtained from alternative measurement methods.

documentation. Nevertheless, we do not intend to claim this procedure as the optimal approach but to straighten the importance to better characterize the differences among alternative methods and show how the results of the tests can be interpreted in some actual recordings.

2.2. A suggestion of tests and procedures

For outlier detection in both time series and the DTS we have employed a slight modification of the Grubbs test (Grubbs 1969) To decide if an outlier is present in the time series $x(i)$, the modified Z-score time series (Iglewicz and Hoaglin 1993) is computed as:

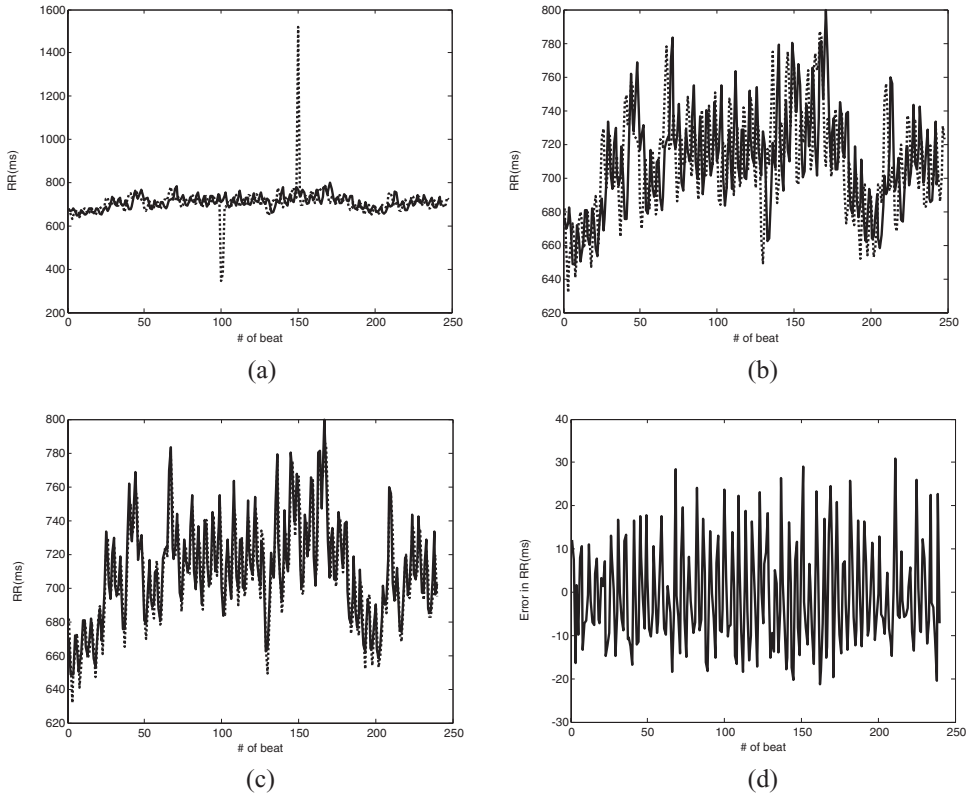

Figure 2. Example of the need of outlier correction and synchronization. The upper panel on the left (a) shows two unsynchronized time series where there is a measurement method (dotted line) that has two outliers. The upper panel onthe right (b) shows the two unsynchronized time series once the outliers have been detected and corrected. The lower panel on the left (c) shows the two synchronized time series while the lower panel on the right (d) shows the difference time series between the two synchronized time series. Note that the scales of figures (a) and (d) are, for the sake of visualization, different from those of figures (b) and (c).

$$G(i) = \frac{|0.6745 \cdot (x(i) - \text{median}(x))|}{\text{median}(|x(i) - \text{median}(x)|)} \tag{3}$$

and its maximum is found ($G_{\max}$ at sample $i_{\max}$). It is decided that the sample at $i_{\max}$ is an outlier if:

$$G(i_{\max}) = G_{\max} > \frac{N-1}{\sqrt{N}} \cdot \sqrt{\frac{t^2}{N-2+t^2}} \tag{4}$$

where N is the number of samples in the time series and t is the critical value of the t Student distribution with N-2 degrees of freedom and a risk $\alpha/(2 \cdot N)$. In this work we have considered $\alpha = 0.001$ so the probability that a normal sample is treated as an outlier is kept very low. This procedure is repeated by eliminating the outlier from the time series (the new time series has N-1 samples and do not contains the sample $i_{\max}$) until the condition in (4) does not hold. The chosen quality statistic of the time series in figure 1 is the percentage of outliers found by this described test.

After detection and for *RR* time series, the outliers were classified as false positives (at least two consecutive *RR* intervals with a value significantly lower than the median of the previous 10 *RR* intervals), false negatives (at least one consecutive *RR* interval with a value significantly higher than the median of the previous 10 *RR* intervals) or ectopic beats (two consecutive *RR* intervals, one with a significantly lower value and the other with a significantly higher value than the median of the previous 10 *RR* intervals). For a false positive, as many consecutive *RR* intervals were added as needed to obtain a corrected *RR* beat with a value near the mean of the previous ten *RR* intervals. A false negative was split in as many *RR* intervals with equal value as needed to be near to the mean of the previous ten *RR* intervals. Ectopic beats were substituted with two equal *RR* intervals corresponding to the mean of the two *RR* intervals involved in the artifact.

After artifact correction, the two time series (namely, RRc_a and RRc_b) look as in figure 2(b) To synchronize them and obtain two aligned time series with the same number of samples as in figure 2(c) the procedure delays or advances a time series respect to the other a finite number of samples until the Fisher intraclass correlation coefficient (ICC) is maximized (Fisher 1934) The ICC is estimated as:

$$\mathrm{ICC}(n_s, m_s) = \frac{1}{I \cdot s^2} \cdot \sum_{n=1}^{I} (RRc_a(n_s + n) - \overline{RR}) \cdot (RRc_b(m_s + n) - \overline{RR}) \tag{5}$$

where:

$$\overline{RR} = \frac{1}{2 \cdot I} \cdot \sum_{n=1}^{I} (RRc_a(n_s + n) + RRc_b(m_s + n)) \tag{6}$$

and:

$$s^2 = \frac{1}{2 \cdot I} \left\{ \sum_{n=1}^{I} (RRc_a(n_s + n) - \overline{RR})^2 + \sum_{n=1}^{I} (RRc_b(m_s + n) - \overline{RR})^2 \right\} \tag{7}$$

Finally, if L_a is the number of samples in RRc_a and L_b is the number of samples in RRc_b then I is determined as:

$$I = \min\{L_a - n_s, L_b - m_s\} \tag{8}$$

In this work and because the *RR* time series to compare are measured nearly simultaneously, we have searched the maximum of ICC by advancing or delaying a *RR* time series respect to the other by no more than 10 beats.

Time series are considered as synchronized when ICC is maximum. The agreement statistic on figure 1 is, precisely, the maximum ICC that corresponds to use the delay parameters $n_{s_{max}}$ and $m_{s_{max}}$. On the other hand, the DTS is finally obtained as:

$$\mathrm{DTS}(i) = RRc_a(n_{s_{max}} + i) - RRc_b(m_{s_{max}} + i) \tag{9}$$

Because outlier correction on the *RR* time series can generate outliers in the DTS, the previously described procedure for the identification of outliers has been employed to label artifacts in this time series. The chosen DTS quality statistic on figure 1 is, as for the original time series, the ratio between the number of detected outliers and the number of available DTS samples. After outlier identification and depending on the next steps of analysis, two corrected DTS have been derived: If the characterization tests and statistics does not depend on the ordering of the time series (i.e. histogram based tests, standard deviation, etc) a time series is created in which all segments without outliers are merged. When the dynamics of the time series is of interest (as is the case of tests that look for residual correlation), any sample

labelled as an outlier is set to zero. From now on, we will name $\mathrm{DTS}_m(i)$ the DTS rejecting outliers and merging the other data and $\mathrm{DTS}_z(i)$ the DTS where the outliers have been replaced by zeros.

The differences quantification and distribution testing have been performed using the $\mathrm{DTS}_m(i)$ time series. Bias and dispersion have been estimated using the arithmetic mean ($\overline{\mathrm{DTS}_m}$) and the sample standard deviation (s_{DTS_m}) while the sample confidence interval has been estimated using the difference between the 97.5% and 2.5% percentiles (U_{DTS_m}). Normal distribution testing has been assessed using the Anderson-Darling goodness-of-fit hypothesis test (Anderson and Darling 1954) and the reported normality test statistic of figure 1 has been their proposed statistic:

$$W_T^2 = -T - \frac{1}{T}\cdot\sum_{j=1}^{T}(2\cdot j-1)\cdot[\log(u_j)+\log(1-u_{T-j+1})] \tag{10}$$

where T is the number of samples in the error time series (lower than I if there are outliers in the DTS of equation (9)) and u_j is the value of the cumulative distribution function of the normal distribution evaluated at the j sample of the standardized and sorted error time series. The standardization procedure consists of removing the mean and normalizing the resulting values by the sample standard deviation of the time series. The higher the statistic, the lower the probability that the time series can be considered as normally distributed. The asymptotic distribution of W_T^2 is empirical and the critical values to accept or reject the null hypothesis of samples normally distributed can be found in statistical packages, tables or estimated using Monte Carlo simulations.

Moreover, to reinforce the agreement (or disagreement) of the normal hypothesis, the sample coverage factor has been computed as:

$$k = \frac{U_{e_m}}{2\cdot s_{e_m}} \tag{11}$$

If the distribution is approximately normal, this coverage factor should be (in average) around 1.96

To quantify the stationarity of the mean (first order moment) we have used the Kwiatkowski, Phillips, Schmidt and Shin (KPSS) test (Kwiatkowski *et al* 1992) with the null hypothesis that a time series is stationary around a constant value against the alternative that it is a nonstationary unit-root process. The reported first moment stationarity test statistic on figure 1 is the KPSS statistic that is computed as:

$$\mathrm{KPSS} = \frac{\sum_{i=1}^{I}\left(\sum_{j=1}^{i}\left(\mathrm{DTS}_z(j)-\frac{1}{I}\cdot\sum_{k=1}^{I}\mathrm{DTS}_z(k)\right)\right)^2}{s^2\cdot I^2} \tag{12}$$

where:

$$s^2 = \frac{\sum_{i=1}^{I}\left(\mathrm{DTS}_z(j)-\frac{1}{I}\cdot\sum_{k=1}^{I}\mathrm{DTS}_z(k)\right)^2}{I} + \frac{2}{I}\cdot\sum_{i=1}^{L}\left(1-\frac{i}{L+1}\right)\cdot\sum_{j=i+1}^{I}\left[\left(\mathrm{DTS}_z(j)-\frac{1}{I}\cdot\sum_{k=1}^{I}\mathrm{DTS}_z(k)\right)\right.$$
$$\left.\times\left(\mathrm{DTS}_z(j-i)-\frac{1}{I}\cdot\sum_{k=1}^{I}\mathrm{DTS}_z(k)\right)\right], \tag{13}$$

$$L = \text{round}\left(4 \cdot \sqrt[4]{\frac{I}{100}}\right) \tag{14}$$

and I is the number of error samples in the $\text{DTS}_z(i)$ time series. The greater the KPSS statistic, the higher the nonstationarity of the mean is. The asymptotic distribution has a quite complicated formulation and it is related with the integration of a mix of Brownian motions. Generally the critical values are obtained by simulation or are included in statistical software packages.

To assess the stationarity of the variance (second order moment) we have used the test suggested by (Inclan and Tiao 1994) that is based on the M statistic estimated as:

$$M = \max\left\{\sqrt{\frac{I}{2}} \cdot |D(i)|\right\} \qquad i \in N[1, I] \tag{15}$$

where the distance time series is obtained from the $\text{DTS}_z(i)$ time series as:

$$D(i) = \frac{\sum_{j=1}^{i}\left(\text{DTS}_z(j) - \frac{1}{I}\cdot\sum_{l=1}^{I}\text{DTS}_z(l)\right)^2}{\sum_{j=1}^{I}\left(\text{DTS}_z(j) - \frac{1}{I}\cdot\sum_{l=1}^{I}\text{DTS}_z(l)\right)^2} - \frac{i}{I} \tag{16}$$

M is the second moment stationarity test statistic on figure 1 and the greater the M statistic, the higher the nonstationarity of the variance is. The asymptotic distribution of M is empirical and the critical values to accept or reject the null hypothesis can be found in statistical packages, tables or estimated using Monte Carlo simulations.

The whiteness of the time series has been assessed using two statistics from two complementary tests: the normalized Q statistic from the Ljung–Box Q-test (Ljung and Box 1978) has been chosen as the residual autocorrelation statistic of figure 1 while the U statistic of the nonparametric Wald–Wolfowitz runs test for randomness on the sign of the DTS (with the residual mean previously removed) (Wald and Wolfowitz 1940) has been chosen as the statistic for the runs test in figure 1.

The Q statistic of the Ljung–Box test is estimated as:

$$Q = I \cdot (I+2) \cdot \sum_{k=1}^{h} \frac{\widehat{\rho}_k^2}{I-k} \tag{17}$$

where

$$\widehat{\rho}_k = \frac{\sum_{i=1}^{I-k}\left(\text{DTS}_z(i) - \frac{1}{I}\cdot\sum_{l=1}^{I}\text{DTS}_z(l)\right)\cdot\left(\text{DTS}_z(i+k) - \frac{1}{I}\cdot\sum_{l=1}^{I}\text{DTS}_z(l)\right)}{\sum_{i=1}^{I}\left(\text{DTS}_z(i) - \frac{1}{I}\cdot\sum_{l=1}^{I}\text{DTS}_z(l)\right)^2} \tag{18}$$

is the estimation of the autocorrelation of the DTS at lag k and h is the number of analyzed lags that are chosen, as suggested by (Tsay 2005) as:

$$h = \text{round}(\ln(I)) \tag{19}$$

Finally, because Q asymptotically behaves as a χ^2 random variable with h degrees of freedom and in order to provide a randomness indicator with no dependence on the number of samples of the time series, we report the normalized Q statistic defined as:

$$Q_n = \frac{Q}{h} \tag{20}$$

To compute the U statistic of the Wald–Wolfowitz runs test, first the sign time series is obtained as:

$$se_z(i) = \text{sign}\left(\text{DTS}_z(i) - \frac{1}{I}\cdot\sum_{l=1}^{I}\text{DTS}_z(l)\right) \tag{21}$$

Next, the number of runs (R) is computed as the number of sign changes in the time series plus one. Moreover, the number of positive (np) and negative (nn) samples in $\text{se}_z(i)$ are counted.

If the time series was a random white process, the mean number of runs would be:

$$\overline{R}_r = \frac{2\cdot\text{np}\cdot\text{nn}}{I} + 1 = \frac{2\cdot\text{np}\cdot\text{nn}}{\text{np}+\text{nn}} + 1 \tag{22}$$

while its standard deviation would be:

$$\sigma_{R_r} = \sqrt{\frac{(\overline{R}_r - 1)\cdot(\overline{R}_r - 2)}{I-1}} \tag{23}$$

Finally a continuity correction (c) that depends on R and the expected number of runs for a random white process is applied to obtain the U statistic as

$$U = \left|\frac{R - c - \overline{R}_r}{\sigma_{R_r}}\right| \tag{24}$$

where

$$c = \begin{cases} 0.5 & \text{if } R > \overline{R}_r \\ 0 & \text{if } R = \overline{R}_r \\ -0.5 & \text{if } R < \overline{R}_r \end{cases} \tag{25}$$

The runs test statistic asymptotically behaves as a standardized normal distribution so the critical values can be found elsewhere. Table 1 summarizes the proposed tests and statistics as well as their intended use.

2.3. Assessing the differences in estimating actual RR time series

In order to show the potentiality of the previously described battery of tests, we have employed three different databases: the first one (ECG database) consists of two high quality ECG channels simultaneously sampled, the second database (*RR* database) consists on *RR* time series recorded simultaneously by two different devices and the last one (PP database) has a high quality ECG signal and a finger photoplethysmography signal simultaneously sampled.

The ECG database has the purpose to assess how big the differences in *RR* time series estimation are when choosing a certain ECG lead instead of another. The database contains measurements from 20 young healthy subjects. Data were acquired using a Biopac MP36 data acquisition system (Santa Barbara, CA, USA). Channels 1 and 2 of the system were devoted to measure conventional ECG with a bandwidth from 0.05 Hz to 150 Hz while channel 3 was employed to measure the respiratory signal obtained from a thoracic piezoresistive band (SS5LB sensor by Biopac, Santa Barbara, CA, USA) with a bandwidth from 0.05 Hz to 10 Hz. Channel 1 measured the ECG standard lead I while channel 2 measured the ECG

Table 1. A summary of the proposed tests and statistics and their correspondence with the procedure described in figure 1.

Test/procedure	Involved equations	Statistic	Use
Repeated modified Grubbs test	(3) and (4)	%oa, %ob, %oerr	Quality statistic for method a, b and error
Fisher intraclass correlation coefficient	(5)–(7)	ICC	Time series synchronization Agreement statistic
Anderson-darling goodness-of-fit hypothesis test	(10)	W_T^2	Normality test statistic
Sample coverage factor	(11)	k	Sample level of agreement estimation
Kwiatkowski, Phillips, Schmidt and Shin (KPSS) test	(12)–(14)	KPSS	First moment stationarity test statistic
Inclan and Tiao heteroscedasticity test	(15) and (16)	M	Second moment stationarity test statistic
Ljung-Box test of residual autocorrelation	(17)–(20)	Q_n	Residual autocorrelation statistic
Wald–Wolfowitz runs test	(21)–(25)	U	Runs test statistic

standard lead II. For the ECG measurement monitoring electrodes with foam tape and adhesive gel (3M Red Dot 2560) were employed. Each channel was sampled at 5 kHz. During the measurement, the subjects were asked to be very still in supine position on a comfortable conventional single bed and awake. After attachment of sensors, we recorded a total of 60 min. From minute 5 to minute 55, the subjects were listening to a playlist of classical music. For analysis purposes, the 60 min measurement was segmented in 12 recordings of 5 min. For each of these recordings, the QRS complexes were detected for both leads. A first rough fiducial point was obtained using the Pan-Tompkins QRS detector (Pan and Tompkins 1985) but was further refined by maximizing the correlation between any detected QRS complex and the first detected QRS complex using templates of 200 ms duration centred on the rough fiducial point. After this detection and for each recording, two *RR* time series were obtained computing the successive time differences between consecutive R-wave fiducial points of each lead. The raw ECG among other signals can be downloaded from www.physionet.org/physiobank/database/cebsdb/ (Garcia-Gonzalez *et al* 2013a, Garcia-Gonzalez *et al* 2013b, Goldberger *et al* 2000). For each five minute epoch and each subject, 12 indicators were computed from the two *RR* and DTS after synchronization (DTS was defined as the difference of the *RR* from the ECG standard lead I minus the *RR* from the ECG standard lead II). The indicators were the percentage of outliers in the *RR* time series obtained from the standard lead I using the Grubbs test and $\alpha = 0.001$ (%oa), the same percentage for the *RR* time series obtained from the standard lead II (%ob), the percentage of outliers using the Grubbs test and $\alpha = 0.001$ for the DTS once they have been synchronized (%oDTS), the ICC of the synchronized time series (ICC), the mean of the DTS (MDTS) expressed in milliseconds, the standard deviation of the DTS (SDDTS) also expressed in milliseconds, the sample coverage factor of the DTS (k), the Anderson–Darling statistic for normal distribution testing (W_T^2), the stationarity of the mean test statistic (KPSS), the stationarity of the variance test statistic (M), the normalized residual autocorrelation statistic (Q_n) and the nonparametric Wald–Wolfowitz runs test for randomness statistic (U). The mean, median, standard deviation and 5% and 95% percentiles of each

Table 2. Critical values for the different evaluated statistics (both asymptotic and for 300 samples). If the estimated statistic is higher than the reported critical value then the null hypothesis can be rejected with risk p.

N	p	W_T^2	KPSS	M	Q_n	U
300	<0.05	0.750	0.454	1.314	2.099	1.968
$\rightarrow\infty$	<0.05	0.752	0.462	1.358	≈1.600	1.960
300	<0.001	1.438	1.053	1.899	3.743	3.323
$\rightarrow\infty$	<0.001	1.441	1.170	1.951	≈2.400	3.291

indicator have been estimated and the raw values of W_T^2, KPSS, M, Q_n and U have been compared to critical values. Moreover, the Friedman repeated measures analysis of variance on ranks has been used to ascertain if there are significant differences in any indicator associated with the subject under measurement or the time epoch. The whole processing procedure and statistical analyses (for this and the other databases) have been performed using MATLAB® and its statistical toolbox.

Table 2 shows the critical values of W_T^2, *KPSS*, M, Q_n and U when analyzing time series with $N = 300$ samples (around the number of beats analyzed in a 5 min epoch) as well as the asymptotic value ($N \rightarrow \infty$) for $p < 0.05$ and $p < 0.001$. Note that except for Q_n, the dependence of the critical value with the number of samples is negligible. The critical values for W_T^2 were obtained using the statistical toolbox of MATLAB®. The critical values for the KPSS in the statistical toolbox are provided only for probabilities between 0.01 and 0.10 so two Monte Carlo simulations with random normal noise, the first with 300 samples and 10^7 realizations and the second with 100 000 samples (assumed as a high number of samples enough to characterize the asymptotic value) and 10^5 iterations were performed. The critical values of KPSS were estimated by computing the percentiles 95% ($p = 0.05$) and 99.9% ($p = 0.001$). Because M is not implemented in the statistical toolbox, the critical values were obtained using the same procedures as for the KPSS statistic except for the asymptotic valued for $p = 0.05$ that was obtained directly from (Inclan and Tiao 1994). The critical values for Q_n and for 300 samples were obtained by dividing the critical value reported by the statistical toolbox by 6 (the rounding of the natural logarithm of 300). The asymptotic critical values were obtained by extrapolating towards infinity the curve of the critical value reported by the statistical toolbox divided by the rounding of the natural logarithm of the number of samples. Finally, the critical values for the U statistic were obtained using the critical values of the t Student distribution as provided by the statistical toolbox. Because the critical values experience little change when the number of samples is high enough (are quite similar with 150 or 450 samples that are a minimum and maximum number of heartbeats in 5 min), all the computed statistics have been compared with the critical values for $N = 300$. The critical value with most variability is Q_n that changes from 2.21 for 150 samples to 2.10 for 450 samples.

The *RR* database (Parrado *et al* 2010) measures a sample of 90 healthy subjects. *RR* intervals were recorded simultaneously with a Polar Heart Rate Monitor (Polar S810) with a resolution of 1 ms, and the Omegawave Sport Technology System (Eugene, OR) with the same resolution. Subjects rested comfortably and adequately dressed during the recording in a supine position in a quiet room maintained at the temperature of 19–23 °C. After 3 min of rest lying down, subjects were asked to remain supine quietly during 10 min without speaking or making any movements. HRV data was registered continuously for 5 min of free breathing rhythm and 5 min of paced breathing at the frequency of 0.20 Hz (12 breaths/min) using

the two systems. For each recording, the same 12 indicators as in the ECG database were obtained. The DTS was defined after synchronization as the difference of the *RR* time series obtained from the Omegawave system minus the *RR* time series obtained from the Polar system. The mean, median, standard deviation and 5% and 95% percentiles of each indicator have been estimated and the raw values of W_T^2, KPSS, M, Q_n and U have been compared to critical values. Moreover, a Friedman repeated measures analysis of variance on ranks has been used to assess if there are significant differences associated to the subject under measurement while a Wilcoxon signed rank test has been used to detect if there are changes associated to the breathing pattern.

The PP database is quite similar to the ECG database except that it also acquires a channel of finger pulse photoplethismography (PPG). The database measured 22 healthy subjects acquiring the signal at 5 kHz during 60 min that were divided in 12 consecutive epochs each with duration of 5 min. Data were acquired using a Biopac MP36 data acquisition system (Santa Barbara, CA, USA). During the measurement, the subjects lay in a bed while listening to the same playlist of pop-rock music. QRS complexes were obtained from the standard II lead using a Pan–Tompkins detector and a further fiducial point refinement based on template matching. After detection, we looked for the pulse arrival in the PPG channel during 600 ms after each QRS complex. To detect the pulse we used three fiducial points after bandpass filtering the PPG between 0.05 Hz and 10 Hz using a 4th order bidirectional Butterworth filter. The peak (P) fiducial point corresponds to the maximum of the filtered PPG inside the 600 ms window, the maximum derivative (MD) fiducial point corresponds to the maximum of the differentiated filtered PPG inside the same window, and finally, the tangent intersection (T) fiducial point corresponds to the intersection of the straight line (linear approximation) of the filtered PPG at the maximum derivative with the minimum value of the filtered PPG in the analysis window (Peng *et al* 2015) For each recording, the *RR* time series obtained from the QRS fiducial points were compared with each of the three pulse to pulse (PP) time series that can be obtained using the P, MD or T fiducial points. For each pulse wave fiducial point, the DTS was defined as the difference between the *RR* time series obtained from the ECG minus the time series obtained from successive intervals between pulse arrivals. The same methodology as in the previous databases has been employed to characterize the time series. The raw values of W_T^2, KPSS, M, Q_n and U have been also compared to critical values and a Friedman repeated measures analysis of variance on ranks has been employed to assess the effect of the subject under measurement, the time epoch and, of course, the employed fiducial point.

3. Results and discussion

Table 3 shows the summary of the results for the computed indicators in the ECG database while table 4 shows the percentage of recordings where the statistics are higher than the critical values (considering $N = 300$ in table 2). As seen in the tables, the standard deviation of the DTS (SDDTS) is around 0.5 ms in mean but the error cannot be considered as normal distributed and rarely as a white process (Q_n or U statistics are higher than the critical values for most of the 5 min epochs). On the other hand, the mean and variance of the DTS are highly stable indicating that the measurement conditions have not changed throughout the 5 min. The Friedman repeated measures analysis of variance on ranks shows that %oa, %ob, %oDTS, SDDTS, k, W_T^2, KPSS, Q_n and U change very significantly among subjects ($p < 0.001$) while M changes significantly among subjects ($p < 0.05$) On the other hand, the same test shows that no indicator changes significantly through time epoch.

Table 3. Summary of results for each indicator considering the 240 available recordings (20 individuals and 12 time series of 5 min duration for individual) of the ECG database. The indicators have been summarized by the mean, standard deviation, median and percentiles.

Indicator	Mean	Standard deviation	5% percentile	Median	95% percentile
%oa	0.29	0.60	0.00	0.00	1.66
%ob	0.26	0.53	0.00	0.00	1.61
%oDTS	0.25	1.21	0.00	0.00	1.15
ICC	1.00	0.00	1.00	1.00	1.00
MDTS (ms)	0.00	0.00	−0.01	0.00	0.01
SDDTS (ms)	0.52	0.27	0.28	0.47	1.06
k	1.90	0.14	1.68	1.90	2.12
W_T^2	4.67	2.41	1.89	3.92	9.58
KPSS	0.02	0.02	0.01	0.01	0.05
M	0.98	0.40	0.51	0.88	1.85
Q_n	24.8	21.6	4.01	16.3	78.4
U	3.27	1.94	0.44	3.08	6.27

Table 4. Percentage of 5 min recordings of the ECG database with statistics higher than the critical values in table 2. With $Q_n \cup U$ we denote when either Q_n or U (or both) statistics are higher that the critical value and with $Q_n \cap U$ when both statistics are simultaneously higher than the critical value.

P	W_T^2	KPSS	M	Q_n	U	$Q_n \cup U$	$Q_n \cap U$
0.05	99.6%	0.00%	17.9%	100%	71.7%	100%	70.0%
0.001	97.5%	0.00%	4.58%	95.4%	42.9%	95.8%	42.5%

Table 5. Summary of results for each indicator considering the 180 available recordings (90 individuals and two time series of 5 min duration for individual (periodic and spontaneous breathing) of the *RR* database.

Indicator	Mean	Standard deviation	5% percentile	Median	95% percentile
%oa	0.11	0.56	0.00	0.00	0.36
%ob	0.11	0.56	0.00	0.00	0.41
%oDTS	1.06	2.05	0.00	0.54	4.30
ICC	1.00	0.00	1.00	1.00	1.00
MDTS (ms)	0.01	0.02	−0.03	0.01	0.05
SDDTS (ms)	1.28	0.27	1.05	1.20	1.77
k	1.82	0.15	1.59	1.82	2.09
W_T^2	9.78	2.87	4.55	9.74	14.4
KPSS	0.07	0.06	0.02	0.05	0.19
M	0.98	0.37	0.55	0.90	1.69
Q_n	11.4	4.24	6.76	10.4	18.9
U	4.59	2.82	1.10	4.16	12.2

Table 6. Percentage of 5 min recordings of the *RR* database with statistics higher than the critical values in table 2.

p	W_T^2	KPSS	M	Q_n	U	$Q_n \cup U$	$Q_n \cap U$
0.05	100%	0.00%	15.6%	100%	90.0%	100%	89.4%
0.001	100%	0.00%	2.78%	99.4%	73.9%	100%	73.3%

Tables 5 and 6 show the results for the *RR* database. In this case, the standard deviation of the DTS is (as expected) greater than in the ECG database (around 1.3 ms versus 0.5 ms) and can be justified in part by the lower resolution of the *RR* time series (1 ms in the *RR* database versus 0.2 ms for the ECG database). Once again, the DTS is stable in mean and variance and cannot be considered, in general, nor as a normal process neither as a white process. The Friedman repeated measures analysis of variance on ranks shows that SDDTS and W_T^2 change very significantly among subjects ($p < 0.001$) while *%oa*, *%ob*, *%oDTS* and U change significantly among subjects ($p < 0.05$) On the other hand, the Wilcoxon signed rank test shows that there are significant differences ($p < 0.05$) in SDDTS associated with the way of subjects are breathing (median: 1.19 ms when breathing at will versus 1.23 ms when breathing periodically) and in U (median: 4.32 when breathing at will versus 3.88 when breathing periodically).

Tables 7 and 8 show the results for the PP database separated according to the three different fiducial points. The main difference in results is that the ICC is not 1.00 as in the previous databases although the agreement between the *RR* time series and the PP time series is still very high in accordance with previous studies (Gil *et al* 2010, Schäfer *et al* 2013) The Friedman repeated measures analysis of variance on ranks shows that the employed fiducial point changes very significantly ($p < 0.001$) %oDTS, ICC, MDTS, SDDTS, k, W_T^2, KPSS, M and Q_n. There were no significant differences in indicators when comparing those obtained with the MD or the T fiducial points. This result indicates that the PP time series obtained by the MD or the T fiducial points are quite similar. On the other hand, ICC is higher with these fiducial points indicating that they are better suited to be used as surrogate measures of the QRS complex instead of the peak of the pulse wave. This is confirmed by the lower MDTS and SDDTS for these fiducial points. Finally, the DTS using either MD or T are more normally distributed, more stable in mean and variance than using the P fiducial point but cannot be considered as white random process. All the indicators show very significant differences ($p < 0.001$) among subjects and the time epoch affects significantly ($p < 0.05$) to %oa, %oDTS and k and very significantly ($p < 0.001$) to %ob and SDDTS. Taking into account that the time between the QRS complex and the pulse arrival fiducial point is the pulse arrival time we can relate the SDDTS with the standard deviation of the pulse arrival time variability as follows:

$$\begin{gathered}\mathrm{PAT}(n) = P(n) - R(n)\\ PP(n) = P(n+1) - P(n)\\ RR(n) = R(n+1) - R(n)\\ \mathrm{DTS}(n) = PP(n) - RR(n) = P(n+1) - P(n) - R(n+1) - R(n) = \mathrm{PAT}(n+1) - PAT(n)\end{gathered} \tag{26}$$

So the very significant changes in SDDTS can be attributed to changes in the pulse arrival time variability that may be caused with changes in the relaxation of the subjects while listening to music.

As a summary of the results for the three databases, generally the observed differences are not normal neither white (although they are reasonably stable in mean and variance). A high number of studies dealing with error sources in the HRV measurement study the effect of

Table 7. Summary of results for each indicator considering the 264 available recordings (22 individuals and 12 time series of 5 min duration for individual of the PP database.

Indicator	Fiducial point	Mean	Standard deviation	5% percentile	Median	95% percentile
%oa	all	0.32	1.05	0.00	0.00	1.70
%ob	P	0.48	1.15	0.00	0.00	1.98
	MD	0.37	0.97	0.00	0.00	1.84
	T	0.34	0.98	0.00	0.00	1.71
%oDTS	P	3.48	4.89	0.00	1.38	15.1
	MD	0.43	0.91	0.00	0.00	2.25
	T	0.58	1.16	0.00	0.00	3.28
ICC	P	0.97	0.05	0.88	0.99	1.00
	MD	0.99	0.01	0.97	1.00	1.00
	T	0.99	0.01	0.98	1.00	1.00
MDTS (ms)	P	0.22	0.66	−0.34	0.05	1.48
	MD	0.00	0.08	−0.13	0.00	0.13
	T	−0.02	0.10	−0.18	0.00	0.11
SDDTS (ms)	P	11.6	8.75	3.79	9.48	24.5
	MD	6.97	2.46	3.81	6.42	12.0
	T	6.83	2.27	3.98	6.37	11.6
k	P	2.11	0.15	1.86	2.10	2.36
	MD	1.94	0.13	1.75	1.93	2.16
	T	1.93	0.12	1.73	1.92	2.13
W_T^2	P	1.77	1.19	0.33	1.49	4.11
	MD	1.00	0.91	0.25	0.70	2.18
	T	1.10	0.93	0.26	0.86	2.90
KPSS	P	0.07	0.08	0.01	0.04	0.23
	MD	0.03	0.03	0.01	0.02	0.08
	T	0.03	0.03	0.01	0.02	0.09
M	P	1.63	0.70	0.79	1.41	3.05
	MD	1.22	0.56	0.54	1.11	2.25
	T	1.25	0.60	0.49	1.13	2.28
Q_n	P	7.10	6.43	1.21	5.42	19.3
	MD	16.5	12.5	4.42	12.8	37.5
	T	18.2	12.3	6.76	14.3	40.7
U	P	2.39	1.51	0.13	2.34	4.98
	MD	2.60	1.60	0.19	2.37	5.47
	T	2.73	1.62	0.23	2.57	5.50

Table 8. Percentage of 5 min recordings of the PP database with statistics higher than the critical values in table 2.

p	Fiducial point	W_T^2	KPSS	M	Q_n	U	$Q_n \cup U$	$Q_n \cap U$
0.05	P	77.3%	0.38%	61.0%	86.4%	57.6%	90.2%	53.8%
	MD	47.0%	0.00%	34.8%	99.2%	60.2%	99.2%	60.2%
	T	58.7%	0.00%	39.4%	99.6%	63.6%	99.6%	63.6%
0.001	P	52.3%	0.00%	26.9%	68.9%	27.3%	70.1%	26.1%
	MD	21.6%	0.00%	13.3%	96.2%	31.8%	96.2%	31.8%
	T	20.5%	0.00%	13.6%	98.9%	37.5%	98.9%	37.5%

errors in HRV indexes by adding white noise (or differentiated white noise) to a gold standard time series and evaluating the effect of the noise on the indexes. The obtained results in those studies may be compromised by the fact that actual differences between time series do not obey such simple models. Previous results (Garcia-Gonzalez *et al* 2013b) have shown that some part of the differences associated with the QRS detection is synchronous with breathing at least in the ECG database. Further work will study in more depth the role of the breathing in the dynamics of the DTS.

4. Conclusions

This work has described the complete automatic procedure to obtain the differences time series from two alternative measurement methods of *RR* time series and has proposed a battery of statistical tests and a set of statistical indicators to better describe these. Results show that the differences in practical measurements cannot be considered as normal nor as random processes. On the other hand, the differences are reasonably stable in mean and variance. Simulations of the effect of error sources in the *RR* time series using white normal processes are discouraged because they are not realistic.

Acknowledgments

This work has been supported by the Recercaixa 2013 project 'Desenvolupament de marcadors d'estils de vida saludable per a gent gran basats en Smartphones' and by MINECO project PSI2011-29807

References

Anderson T W and Darling D A 1954 A test of goodness of fit *J. Am. Stat. Assoc.* **49** 765–9

Billman G E, Huikuri H V, Sacha J and Trimmel K 2015 An introduction to heart rate variability: methodological considerations and clinical applications *Front. Physiol.* **6** 1–3

Fisher R A 1934 *Statistical Methods for Research Workers No. 5* ed F A E Crew and D Ward-Cutler (Edinburgh, UK: Oliver and Boyd) pp 198–235

Gamelin F X, Berthoin S and Bosquet, L 2006 Validity of the polar S810 heart rate monitor to measure *RR* intervals at rest *Med. Sci. Sports Exerc.* **38** 887–93

Garcia-Gonzalez M A, Fernandez-Chimeno M and Ramos-Castro J 2004 Bias and uncertainty in heart rate variability spectral indices due to the finite ECG sampling frequency *Physiol. Meas.* **25** 489–504

Garcia-Gonzalez M A, Fernandez-Chimeno M and Ramos-Castro J 2009 Errors in the estimation of approximate entropy and other recurrence-plot-derived indices due to the finite resolution of *RR* time series *IEEE Trans. Biomed. Eng.* **56** 345–51

Garcia-Gonzalez M A, Argelagos-Palau A, Fernandez-Chimeno M and Ramos-Castro J 2013a A comparison of heartbeat detectors for the seismocardiogram *Proc. of the Computing in Cardiology Conference* pp 461–4

Garcia-Gonzalez M A, Argelagos-Palau A, Fernandez-Chimeno M and Ramos-Castro J 2013b Differences in QRS locations due to ECG lead: relationship with breathing *Proc. of the XIII Mediterranean Conf. on Medical and Biological Engineering and Computing* pp 962–4

Gil E, Orini M, Bailón R, Vergara J M, Mainardi L and Laguna P 2010 Photoplethysmography pulse rate variability as a surrogate measurement of heart rate variability during non-stationary conditions *Physiol. Meas.* **31** 1271–90

Goldberger A L, Amaral L A N, Glass L, Hausdorff J M, Ivanov P Ch, Mark R G, Mietus J E, Moody G B, Peng C-K, Stanley H E 2000 Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals *Circulation* **101** e215–20

Grubbs F 1969 Procedures for detecting outlying observations in samples *Technometrics* **11** 1–21
Iglewicz B and Hoaglin D 1993 *How to Detect and Handle Outliers* ed E F Mykytka vol 16 (Milwaukee, WI: American Society for Quality Control) pp 1–87
Inclan C and Tiao G C 1994 Use of cumulative sums of squares for retrospective detection of changes of variance *J. Am. Stat. Assoc.* **89** 913–23
Kwiatkowski D, Phillips P C B, Schmidt P and Shin Y 1992 Testing the null hypothesis of stationarity against the alternative of a unit root: how sure are we that economic time series have a unit root? *J. Econom.* **54** 159–78
Lisenby M J and Richardson P C 1977 The beatquency domain: an unusual application of the fast fourier transform *IEEE Trans. Biomed. Eng.* **24** 405–8
Ljung G M and Box G E 1978 On a measure of lack of fit in time series models *Biometrika* **65** 297–303
Lu S, Zhao H, Ju K, Shin K, Lee M, Shelley K and Chon K H 2008 Can photoplethysmography variability serve as an alternative approach to obtain heart rate variability information? *J. Clin. Monit. Comput.* **22** 23–9
Merri M, Farden D C, Mottley J G and Titlebaum E L 1990 Sampling frequency of the electrocardiogram for spectral analysis of the heart rate variability *IEEE Trans. Biomed. Eng.* **37** 99–106
Milošević M, Shrove M T and Jovanov E 2011 Applications of smartphones for ubiquitous health monitoring and wellbeing management *J. Inf. Technol. Appl.* **1** 5–13
Nunan D, Jakovljevic D G, Donovan G, Hodges L D, Sandercock G R and Brodie D A 2008 Levels of agreement for *RR* intervals and short-term heart rate variability obtained from the polar S810 and an alternative system *Eur. J. Appl. Physiol.* **103** 529–37
Pan J and Tompkins W J 1985 A real-time QRS detection algorithm *IEEE Trans. Biomed. Eng.* **32** 230–6
Parrado E, Garcia M A, Ramos J, Cervantes J C, Rodas G and Capdevila L 2010 Comparison of omega wave system and polar S810i to detect *RR* intervals at rest *Int. J. Sports Med.* **31** 336–41
Peng R C, Zhou X L, Lin W H and Zhang Y T 2015 Extraction of heart rate variability from smartphone photoplethysmograms *Comput. Math. Methods Med.* **2015** 1–11
Ramos-Castro J, Moreno J, Miranda-Vidal H, Garcia-Gonzalez M A, Fernandez-Chimeno M, Rodas G and Capdevila L 2012 Heart rate variability analysis using a seismocardiogram signal *Proc. of the 34th Ann. Int. Conf. of the IEEE Engineering in Medicine and Biology Society* pp 5642–5
Schäfer A and Vagedes J 2013 How accurate is pulse rate variability as an estimate of heart rate variability?: a review on studies comparing photoplethysmographic technology with an electrocardiogram *Int. J. Cardiol.* **166** 15–29
Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology 1996 Heart rate variability: standards of measurement, physiological interpretation and clinical use *Circulation* **93** 1043–65
Tsay R S 2005 *Analysis of Financial Time Series* (Hoboken, NJ: Wiley)
Wald A and Wolfowitz J 1940 On a test whether two samples are from the same population *Ann. Math. Stat.* **11** 147–62