

# A careful look at ECG sampling frequency and R-peak interpolation on short-term measures of heart rate variability

Robert J Ellis<sup>1</sup>, Bilei Zhu<sup>2</sup>, Julian Koenig<sup>3</sup>, Julian F Thayer<sup>3</sup> and Ye Wang<sup>1</sup>

<sup>1</sup> School of Computing, National University of Singapore, 13 Computing Drive, Singapore 117417

<sup>2</sup> School of Computer Science, Fudan University, Shanghai, People's Republic of China

<sup>3</sup> Department of Psychology, The Ohio State University, Columbus, Ohio

E-mail: [wangye@comp.nus.edu.sg](mailto:wangye@comp.nus.edu.sg)

Received 13 January 2015, revised 13 May 2015

Accepted for publication 29 May 2015

Published 3 August 2015




CrossMark

## Abstract

As the literature on heart rate variability (HRV) continues to burgeon, so too do the challenges faced with comparing results across studies conducted under different recording conditions and analysis options. Two important methodological considerations are (1) what sampling frequency (SF) to use when digitizing the electrocardiogram (ECG), and (2) whether to interpolate an ECG to enhance the accuracy of R-peak detection. Although specific recommendations have been offered on both points, the evidence used to support them can be seen to possess a number of methodological limitations. The present study takes a new and careful look at how SF influences 24 widely used time- and frequency-domain measures of HRV through the use of a Monte Carlo-based analysis of false positive rates (FPRs) associated with two-sample tests on independent sets of healthy subjects. HRV values from the first sample were calculated at 1000 Hz, and HRV values from the second sample were calculated at progressively lower SFs (and either with or without R-peak interpolation). When R-peak interpolation was applied prior to HRV calculation, FPRs for all HRV measures remained very close to 0.05 (i.e. the theoretically expected value), even when the second sample had an SF well below 100 Hz. Without R-peak interpolation, all HRV measures held their expected FPR down to 125 Hz (and far lower, in the case of some measures). These results provide concrete insights into the statistical validity of comparing datasets obtained at (potentially) very different SFs; comparisons which are particularly relevant for the domains

of meta-analysis and mobile health.

Keywords: heart rate variability, electrocardiogram sampling frequency, R-peak interpolation, Monte Carlo simulation, false positive rate

 Online supplementary data available from [stacks.iop.org/PM/36/091827/mmedia](http://stacks.iop.org/PM/36/091827/mmedia)

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

## 1. Introduction

Quantitative analysis of heart rate variability (HRV)—fluctuations in the time interval between successive heart beats, associated with the intimate interplay of the sympathetic and parasympathetic branches of the autonomic nervous system—remains a growing area of interest within psychophysiology. The combined advantages of easy and robust electrocardiogram (ECG) recording, affordable and valid ambulatory monitoring devices (e.g. Vanderlei *et al* 2008), and intuitive and comprehensive open-source analysis software (e.g. Tarvainen *et al* 2009) suggest that this growth will continue well into the future.

Two of the many important choices faced in designing and analyzing HRV are (1) the sampling frequency (SF) at which the ECG is digitized during recording, and (2) whether mathematical interpolation of the digitized signal is performed to enhance or ‘refine’ the R-wave fiducial point. The importance of these choices was acknowledged in a seminal consensus paper published in 1996 by the Task Force of the European Society of Cardiology and the North American Society of Pacing Electrophysiology, published simultaneously in the flagship journals of the American Heart Association (Task Force 1996a) and the European Society of Cardiology (Task Force 1996b). Specifically (Task Force 1996a, p 1047–1048):

‘The sampling rate must be properly chosen. A low sampling rate may produce a jitter in the estimation of the R-wave fiducial point, which alters the [HRV power] spectrum considerably. The optimal range is 250 to 500 Hz or perhaps even higher ... while a lower sampling rate (in any case  $\geq 100$  Hz) may behave satisfactorily only if an algorithm of interpolation ([e.g.] parabolic) is used to refine the R-wave fiducial point.’

Another contemporary review paper (Berntson *et al* 1997, p 630) suggests a somewhat more conservative approach, advising that ‘an optimal and generally applicable digitization rate would be 500–1000 Hz’ and that ‘some type of template matching or interpolation algorithm’ should be used ‘especially with digitization rates below 250 Hz’.

Several sources of experimental evidence were used to support these statements, and a number of subsequent investigations exploring how SF and R-peak interpolation influence HRV measures have also been performed; all are summarized in table 1. In reviewing all this evidence, however, a number of methodological issues become apparent.

### 1.1. SF, R-peak interpolation, and HRV: prior work

A first issue pertains to sample size. Several studies investigated samples of fewer than five (Merri *et al* 1990, Pinna *et al* 1994, Riniolo and Porges 1997, Hejjel and Roth 2004, Bragge

**Table 1.** Prior investigations of the influence of SF (with optional R-peak refinement) on R-peak timing or HRV measures, highlighting their downsampling and interpolation methods, as well as how differences (relative to the original SF) were assessed.

Lead author	Year	Data (HC / PT)	SF <sub>Or</sub> (Hz)	Downsampling		Interpolation		Assessment	
				Method	SF <sub>Ex</sub> (Hz)	Method	SF <sub>Re</sub> (Hz)	Method	Measure
Merri	1990	1 / 0	300	Trigonometric	256, 128, 64	— <sup>a</sup>	—	Descriptive	TD, FD
Bianchi	1993	0 / 13	300	Diff. devices	100	Parab. spline	300	Descriptive	FD
Pinna	1994	1 <sup>b</sup> / 1 <sup>b</sup>	2000	Diff. devices	128, 125 <sup>c</sup>	—	—	ANOVA	FD
Aboud	1995	5 / 10	1000	Diff. subjects	500, 333, 200, 142 <sup>d</sup>	—	—	% error	FD
Daskalov	1997	0 / 30	1000 <sup>e</sup>	ECG decim.	250, 100	Cubic spline	1000	% error	RR interval
Riniolo	1997	3 / 0	10000	Rounding	1000, 500, 360, 333, 200, 100	—	—	Descriptive	RSA
Dinh	2001	0 / 25	360	—	—	Cubic spline wavelet	[Wavelet peak]	% error	R-peak
Castiglioni	2003	21 / 0	500	ECG decim.	250, 166, 125, 100, ... 50	Inverse DFT	[Inverse DFT peak]	RMS error	R-peak, A <sub>QRS</sub>
García-González	2004	54	128 <sup>f</sup>	R-peak shift	1000, 500, 250, 125	—	—	NB, NU	FD
Hejfel	2004	1 / 1	1000	R-peak shift	500, 333, 250, 200, ... 100	—	—	% error	TD
Ward	2004	1 / 0	1000	ECG decim.	512, 256, 128	—	—	Descriptive	FD <sup>g</sup>
Chellakumar	2005	7 / 0	1000	Diff. device	100	—	—	ANOVA	TD, FD, NL
Bragge	2005	5 / 0	20000	ECG decim.	100	Double exponential	[Fitted peak]	% error	R-peak
McSharry	2006	100 <sup>h</sup> / 0	2048	ECG decim.	1024, 512, 256, 128, 64	—	—	Signed error	FD

(Continued)

Table 1. (Continued)

Lead author	Year	Data (HC / PT)	SF <sub>Or</sub> (Hz)	Downsampling		Interpolation		Assessment	
				Method	SF <sub>Ex</sub> (Hz)	Method	SF <sub>Re</sub> (Hz)	Method	Measure
García-González	2009	94/0	128, 250	R-peak shift	Inf. <sup>k</sup>	—	—	Signed error	NL
Bhatia	2010	[31 rats]	1000	ECG decim.	500, 333, 250, 200, 100, 50	Cubic spline	1000	ANOVA	TD, FD

Abbreviations: ANOVA: analysis of variance; Aqrs: area under each QRS complex; decim.: decimation; DFT: discrete Fourier transform; diff: different; FD: frequency-domain HRV measures; HC: healthy control subjects; NB: normalized bias; NL: nonlinear HRV measures; NU: nonlinear HRV measures; PT: patients; RSA: respiratory sinus arrhythmia; SF<sub>Re</sub>: resampled SF; TD: time-domain HRV measures. 'R-peak shift' refers to the addition of stochastic error to the R-peaks at detected at SF<sub>Or</sub>.  
 Note: <sup>a</sup> Parabolic interpolation used to improve the R-peak fiducial point only for the 300-Hz original signal, not the downsampled signals; <sup>b</sup> A single reference PQRST complex was created for each subject, and the time between successive T and P waves artificially shortened or lengthened to create a 10-min ECG waveform; <sup>c</sup> Across different devices; <sup>d</sup> Across different subjects; <sup>e</sup> 'The quality of the records was not high ...' (p 376); <sup>f</sup> ... in order to obtain a 'true' RR time series, a 20th-order autoregressive model was fitted ...' (p. 494); <sup>g</sup> PR interval (not RR interval) statistics were examined; <sup>h</sup> simulated ECGs derived from Clifford and McSharry (2003); <sup>i</sup> thousands of 5 min excerpts were extracted from the original records; <sup>k</sup> 'infinite-resolution' RR interval series were generated by adding uniform noise to observed R-peaks.

*et al* 2005) or ten (Abboud and Barnea 1995, Chellakumar *et al* 2005) subjects from the same population (i.e. either healthy control or a patient type), making it challenging to infer whether the effects of the experimental manipulation (i.e. ECG decimation and/or R-peak interpolation) hold in larger, more representative samples of subjects.

A second issue pertains to the SF of the original records (hereafter, 'SF<sub>Or</sub>') relative to the SF values induced experimentally (hereafter, 'SF<sub>Ex</sub>'). In some cases, SF<sub>Or</sub> itself was itself relatively low (e.g. 300 Hz in Merri *et al* 1990 and Bianchi *et al* 1993) and was compared to further-downsampled SF<sub>Ex</sub> values. If systematic errors in HRV estimation were *already* be present at 300 Hz, however, it may confound the ability to accurately quantify additional error that is accrued at lower SF<sub>Ex</sub> values (or accurately quantify the magnitude of error reduction as a result of R-peak refinement).

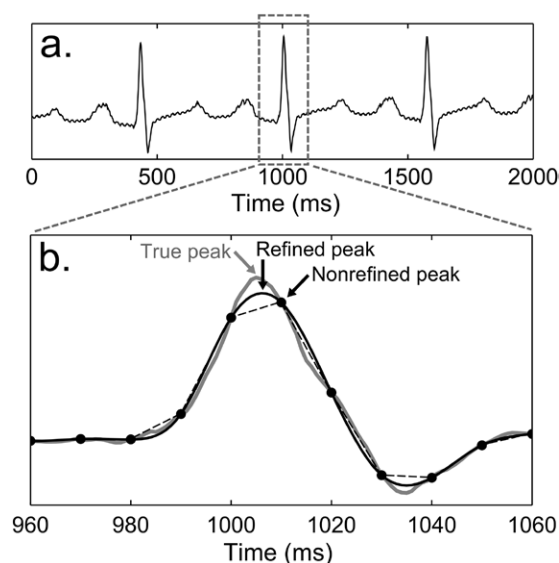
A third issue pertains to the range of SF<sub>Ex</sub> values examined. Several studies in table 1 examined numerous SF<sub>Ex</sub> (e.g. Castiglioni *et al* 2003, Ward *et al* 2004, McSharry and Clifford 2006). Other studies, however, examined only one (Bragge *et al* 2005, Chellakumar *et al* 2005) or two (Pinna *et al* 1994, Daskalov and Christov 1997) SF<sub>Ex</sub> values, often at a great distance from SF<sub>Or</sub>. For example, Pinna *et al* (1994) compared HRV measures derived from Holter-recorded ECGs at 128 Hz and 125 Hz with HRV measures derived from an ECG at 2000 Hz—with no intermediate SF values for reference.

A fourth issue pertains to the downsampling procedure used to achieve SF<sub>Ex</sub>. Only a few studies have performed a true decimation of the ECG digitized at SF<sub>Or</sub> (Daskalov and Christov 1997, Castiglioni *et al* 2003, Bragge *et al* 2005, Bhatia *et al* 2010). Other studies induced different SF<sub>Ex</sub> values using various techniques. Some studies added stochastic temporal error to the location of R-peaks detected at SF<sub>Or</sub> (Garcia-Gonzalez *et al* 2004, Hejjeel and Roth 2004). In a few further cases, a separate device was used to record the ECG at one or more SF<sub>Ex</sub> (Bianchi *et al* 1993, Pinna *et al* 1994, Chellakumar *et al* 2005), introducing a possible confound (i.e. if device-induced measurement artifacts were present). In one final case (García-González *et al* 2009), the addition of temporal noise to detected R-peak locations was used to simulate an *infinite*-resolution SF (i.e. the added noise was used as a surrogate for increased temporal precision for R-peaks).

A fifth issue pertains to the presumed benefits of R-wave interpolation. Figure 1 illustrates the rationale behind *cubic spline* interpolation, widely used in the literature and in several studies cited in table 1. In figure 1(a), two seconds of an ECG at a high (1000 Hz) sampling rate. If, however, the ECG were sampled only at 100 Hz, the resultant waveform would be coarser (figure 1(b); black points and dashed line) than the true ECG (figure 1(b); gray line). By interpolating through points and resampling at 1000 Hz (figure 1(b); solid black line), the resultant R-wave fiducial point is presumed to be closer in time to the true fiducial point than that of the noninterpolated waveform. Although this operation is mathematically straightforward, it nevertheless raises questions. For example, at what SFs does R-wave interpolation yield a meaningful improvement in HRV measure accuracy (i.e. compared to measures calculated at a high SF), and at what SFs is it unnecessary?

A sixth issue pertains to the degree in which statements about minimum SFs and the benefits of R-peak interpolation generalize across different HRV measures. Most prior work on this topic has focused on a particular class of measure (e.g. time-domain only or frequency-domain only). No prior investigation has performed a systematic evaluation of whether SF might affect different HRV measures in unique ways.

Taken together, these methodological issues suggest that widely cited recommendations concerning minimum SFs and the benefits of R-peak interpolation (e.g. Task Force 1996a, Berntson *et al* 1997) may be due for a careful re-evaluation. A minimum SF of 250–500 Hz might be unnecessarily high—or it might not be high enough.



**Figure 1.** R-peak refinement using cubic spline interpolation. From an original ECG trace at 1000 Hz (a), a single R-wave (b) is isolated to better illustrate the poorer temporal resolution (with respect to the accuracy of the R-wave fiducial point) at 100 Hz (black points and dashed black line), and the improved temporal resolution after spline interpolation is applied and the waveform is resampled to 1000 Hz (solid black line).

### 1.2. Implications for the HRV literature

The above issues are further compounded when turning to the broader HRV literature, in which a highly heterogeneous set of SF values may be observed—a decade or more after the recommendations of Task Force (1996a, 1996b) and Berntson *et al* (1997) were published (table 2). With so much SF diversity, a question emerges: Can HRV measures derived from different SFs (e.g. across different studies) be directly compared without introducing systematic errors? The answer to this question has implications for at least two research methodologies.

First, for *meta-analysis*. In the face of heightened attention paid towards the prevalence of publication bias (e.g. Ioannidis 2005), meta-analysis—a set of statistical methods designed to illuminate patterns of similarity across studies—has now reached a favored position in many disciplines (e.g. Cumming 2012). One key assumption of meta-analysis, however, is that the conditions under which data from the individual studies was obtained is equivalent; or, perhaps more accurately, treated as a random source of error. As such, meta-analysis cannot directly compensate for differences in SF among the constituent studies (as it can for differences in sample size). Whether differences in SF among studies may impact the results of the meta-analysis is thus an open question, but one that has not been addressed by any of the several recent meta-analyses examining HRV (Maser *et al* 2003, Sandercock *et al* 2005, Tak *et al* 2009, Kemp *et al* 2010, Nunan *et al* 2010, Lotufo *et al* 2012, Thayer *et al* 2012, Chalmers *et al* 2014).

Second, for the telemedicine domain of *mobile health* (or *mHealth*). With an increasing focus on ‘pervasive monitoring’ applications through the use of components

**Table 2.** A tally of SFs used in the HRV literature, divided into four-year segments. For each SF value  $h$  (Hz), a Google Scholar search ([scholar.google.com](http://scholar.google.com)) was performed (31 December 2014) for works containing the phrase ‘heart rate variability’ in the main text (not the cited works list), and at least one of the following phrases: ‘sampled at  $h$  Hz’, ‘sampling frequency of  $h$  Hz’, ‘sampling rate of  $h$  Hz’, ‘ $h$  samples per second’, ‘ $h$  samples/sec’, or ‘ $h$  samples/s’.

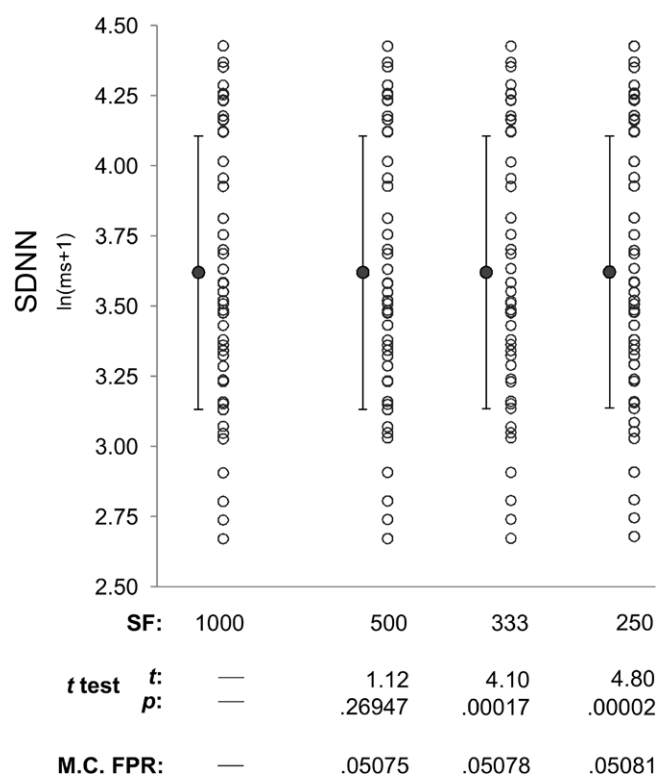
SF	Tally			
	1999–2002	2003–2006	2007–2010	2011–2014
64	3	8	12	22
100	46	75	116	169
125	10	21	26	63
128	57	78	110	163
200	75	116	181	222
250	69	84	133	281
256	43	69	125	177
360	14	29	79	97
500	117	179	220	359
512	11	34	48	142
1000	136	242	374	658
1024	9	31	56	110
2000	17	22	37	76

(e.g. low-profile external sensors; Patel *et al* 2012) that interface directly with the smart-phone *already* in a patient’s pocket (e.g. Boulos *et al* 2011, Kay 2011, Free *et al* 2013) comes important considerations about system resources (e.g. Tarkoma *et al* 2014). Specifically, lower sampling frequencies can translate into lower power consumption of the processor, thus prolonging battery life for continuous physiological monitoring applications (e.g. 24 h recordings). However, although halving the sampling frequency (clock rate) of a transistor cuts its power consumption in half (e.g. Dieter *et al* 2005), it may also decrease the quality recorded physiological signal, thereby adding a source of error during HRV outcome measure calculation.

### 1.3. Quantifying the influence of SF on HRV: within- versus between-subjects analyses

To determine whether HRV measures obtained at different SFs can be meaningfully compared, the definition of ‘meaningfully compared’ must first be established. Most previous investigations into this topic have used various forms of within-subjects analysis (see table 1): comparing HRV values obtained at SF<sub>Or</sub> and one or more SF<sub>Ex</sub> in the same set of subjects. Examples of within-subjects analysis include repeated-measures analysis of variance (ANOVA; Pinna *et al* 1994, Chellakumar *et al* 2005, Bhatia *et al* 2010); Bland–Altman plots (Ziemssen *et al* 2008); and signed error (either in original units or as a percentage; Abboud and Barnea 1995, Daskalov and Christov 1997, Dinh *et al* 2001, Hejjel and Roth 2004, García-González *et al* 2009) or normalized bias (García-Gonzalez *et al* 2004).

Although a within-subjects analysis has increased statistical power to detect a difference between experimental conditions, it may yield a test that is *too* sensitive with respect to quantifying systematic error induced at different SFs. To illustrate this point, consider the following example.<sup>4</sup> A 2 min ECG, digitized at SF<sub>Or</sub> = 1000 Hz, was recorded from 47



**Figure 2.** Illustrating a within-subjects approach to SF-induced error in SDNN values. A conventional paired-samples *t*-test reveals statistically significant inflation of SDNN values 250 Hz, 100 Hz, and 50 Hz relative to 1000 Hz. By contrast, the proposed Monte Carlo (M.C.) false positive rate (FPR) analysis (see section 2.6.2) suggests that only the 50 Hz SF resulted in a meaningful inflation of the expected FPR of the *t*-test.

healthy adults. For each subject, R-peaks were extracted (using the classic algorithm by Pan and Tompkins 1985), and the standard deviation of the RR interval series (SDNN) was calculated. Next, each ECG was decimated to four different experimental SFs:  $SF_{Ex} = \{500, 333, 250\}$  by taking every second, third, or fourth data point from the original signal. Again, R-peaks were extracted (without any R-peak interpolation) and SDNN calculated. Figure 2 plots natural log transformed SDNN data points, group means, and group standard deviations. A Shapiro–Wilk test on each set of values confirmed the absence of non-normality (all  $p_s > 0.05$ ). Next, a paired *t*-test (i.e. a within-subjects analysis) was performed between SDNN values at each  $SF_{Ex}$  relative to  $SF_{Or}$ . Compared to  $SF_{Or}$ , SDNN values at 500 Hz were not statistically different (two-tailed  $p = 0.269$ ), but were statistically different at 333 Hz ( $p = 0.0002$ ) and 250 Hz ( $p = 0.00002$ ). Such findings would seem to support the hypothesis that SF can influence SDNN.

Upon closer inspection, however, these results reveal an illustration of the difference between *statistical* significance and *practical* significance (a distinction which has been

<sup>4</sup>The data for this example comes directly from the set of 47 subjects from the PTBDB database used for the main set of experiments in this paper; thus, all processing steps (SF downsampling, R-peak extraction, HRV measure calculation) are identical to the steps outlined in section 2.2–2.5.



		Decision	
		"0"	"1"
Truth	0	True Negative (TN) — Correct rejection	False Positive (FP) Type I error False alarm
	1	False Negative (FN) Type II error Miss	True Positive (TP) — Hit

**Figure 3.** The possible outcomes of a binary classification test as a function of the true state of events and the decision based on the test outcome.

advocated across numerous scientific domains; e.g. Kirk 1996, Wilkinson 1999). For example, when derived from ECGs at 333 Hz versus 1000 Hz, SDNN values were slightly larger in 34 out of 47 subjects: a group average increase of 0.001 in log units. Although this consistency yielded a 'highly' significant result ( $p = 0.0002$ ), the actual amount of inflation is negligible in the context of the wide range of SDNN values observed across subjects. Within-subjects tests, however, either ignore inter-subject differences entirely (as in a paired  $t$ -test), or treat it as a separate error term (as in ANOVA); in both cases, inter-subject differences do not impact the significance of the actual within-subject test statistic. In other words, a within-subjects analysis to quantify how SF influences HRV may yield a result that is *too* sensitive: a result having a 'bark' that is worse than its 'bite'.

As an alternative, we propose that a more informative perspective is gained from a *between*-subjects analysis of the relationship between SF and the *false positive rate* (FPR). The FPR (also termed the *Type I error rate*, *false alarm rate*, and *fall-out rate*) is a fundamental component of any binary classification test, illustrated in figure 3, and is relevant in numerous domains: for example, disease diagnosis in clinical medicine (e.g. 'test result positive' versus 'test result negative'); detection theory in psychology (e.g. 'stimulus was perceived' versus 'stimulus was not perceived'); machine learning in computer science (e.g. 'pattern is type A' versus 'pattern is type B'); and frequentist statistical inference, used widely in the behavioral and biological sciences (e.g. 'significant,  $p < 0.05$ ' versus 'not significant,  $p \geq 0.05$ '). Mathematically, FPR is defined as the number of false positives divided by the sum of false positives and true negatives:  $FP / (FP + TN)$ . It is the expected probability associated with falsely declaring a test result as significant/positive/successful when in fact no 'true' effect is present (e.g. because both samples, signals, or patterns come from the same underlying population or category). In frequentist statistical inference, the critical value of statistical tests (e.g.  $t$ -test,  $F$ -test, correlation, regression) is set *a priori* so as to control the expected rate at which false positives should occur; by convention, 5% (i.e.  $\alpha = 0.05$ ); for some context, (see Lehmann 1993, Berger 2003).

In this vein, a new question can be stated: 'For a given population of individuals, if a particular HRV measure was obtained from an ECG recorded at a high SF (e.g. 1000 Hz) in one sample of subjects, and obtained from an ECG recorded at a lower SF ( $h$  Hz) in a different sample of subjects, and this process were repeated many times, how often would a two-sample test (e.g. a  $t$ -test) be significant (using the conventional two-tailed  $\alpha = 0.05$ )?' The proportion significant test results is the observed FPR. If the observed FPR (comparing samples at 1000 Hz versus  $h$  Hz) were found to be substantially higher than the expected FPR (i.e.  $\alpha = 0.05$ ), concern is warranted: it indicates that HRV values at  $h$  Hz appear to come from a

*different population* of individuals than HRV values at 1000 Hz. (This method will be illustrated in section 2.6.2.)

We believe an FPR emphasizes the practical significance of how SF affects a given HRV measure, as it is more in line with a plausible methodological concern—whether HRV values measured in independent samples of subjects from a common population (and which were measured at different SFs across the different studies) can be compared without systematic errors (i.e. as indexed by an inflated FPR) being introduced. Such concern is relevant for the domains of meta-analysis and mHealth, as argued in section 1.2. By contrast, a significant *p*-value from a paired-samples test would be a cause for concern if a researcher were attempting to compare HRV values from the *same subjects* using different SFs at different measurement occasions. A decision which prompted a change in SF over the course of a longitudinal study would likely reflect some larger methodological issue with its own set of experimental caveats.

Quantifying the relationship between SF and observed FPRs across commonly used time- and frequency-domain measures of HRV in nominally healthy individuals (and how that relationship is influenced by cubic spline R-peak refinement) will be the central focus of the present paper.

## 2. Methods

A summary of all steps described below is provided in figure 4.

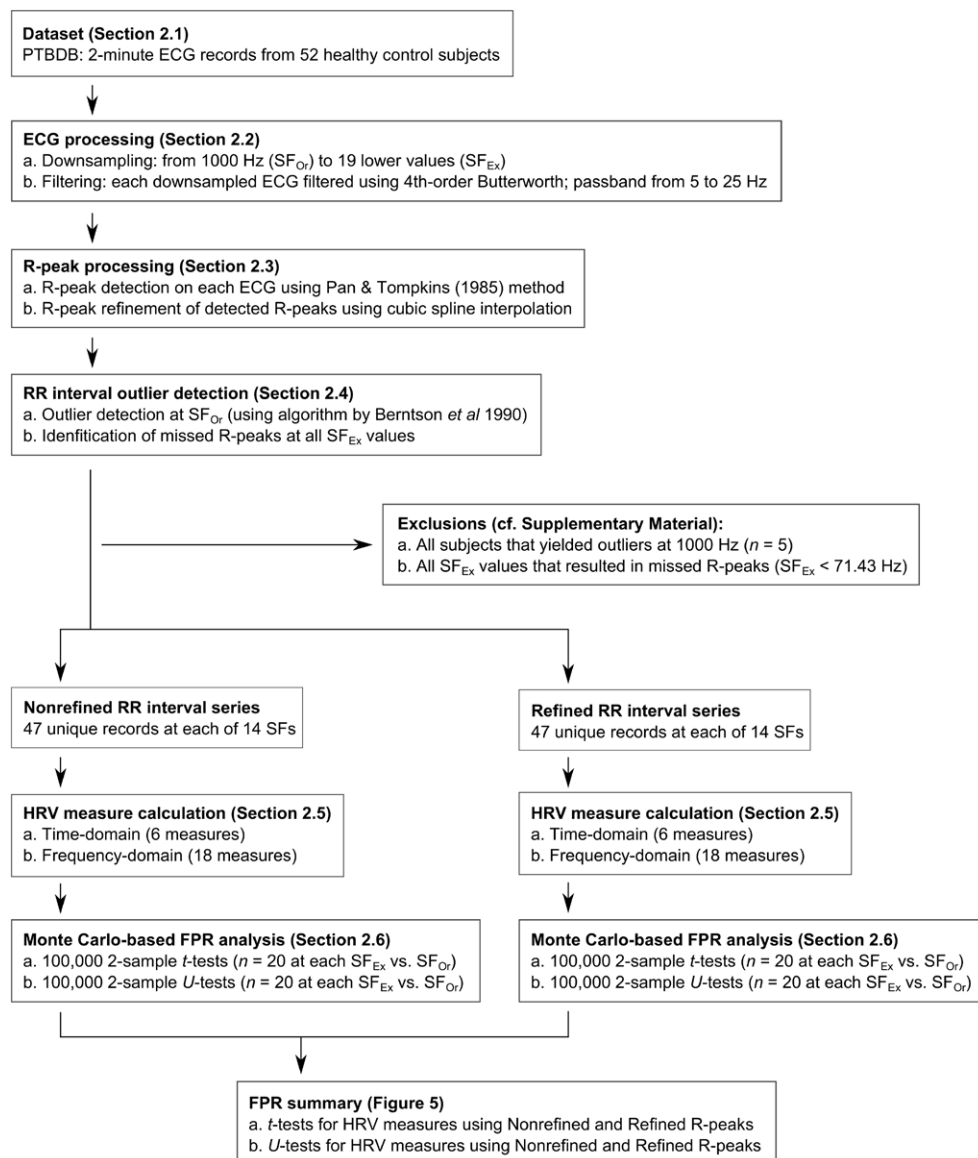
### 2.1. Dataset

The PhysioNet archive ([www.physionet.org/physiobank/database/ptbdb](http://www.physionet.org/physiobank/database/ptbdb); Goldberger *et al* 2000) contains numerous ECG data sets, but only a few at a high ( $\geq 1000$  Hz) sampling rate. For the present study, the PTB Diagnostic ECG Database (PTBDB; Bousseljot *et al* 1995) was selected. The PTBDB includes records from 52 healthy subjects aged 17 to 81. Each record comprises 15 simultaneous ECGs (12 conventional leads and the 3 Frank leads) of approximately 120 s duration, digitized at 1000 Hz with 16-bit resolution. Because ECG lead configuration can induce systematic effects on HRV (García-González *et al* 2011), only lead II ECGs were examined.

### 2.2. ECG processing

**2.2.1. ECG downsampling.** Each ECG was downsampled (using Matlab's `downsample.m`) from the original SF ( $SF_{Or}$ ) to one of the 19 'experimental' sampling frequencies ( $SF_{Ex}$ ) by taking every  $q$ th data point (where  $q$  is the set of integers from 2 to 19), yielding the relationship  $SF_n = 1000 / q_n$ . (A  $q > 19$  would result in  $SF_{Ex}$  values with a Nyquist frequency that falls within the passband of the subsequent ECG filtering stage.) This downsampling method results in a *true* decimation of the original signal. (Note: not all of these  $SF_{Ex}$  values were retained in the final FPR analysis, for reasons detailed in section 2.4.)

**2.2.2. ECG filtering.** Bandpass filtering of an ECG is typically performed prior to R-peak detection (e.g. Kohler *et al* 2002). Each ECG (at each distinct SF) was filtered using a zero-lag 4th-order Butterworth filter (constructed using Matlab's `butter.m` and `filtfilt.m`) with a passband from 5 to 25 Hz.



**Figure 4.** Methods pipeline for the current study.

### 2.3. R-peak processing

**2.3.1. R-peak detection.** R-peak detection was accomplished using `rpeakdetect.m` (Clifford 2003), an instantiation of the classic digital filter-based algorithm by Pan and Tompkins (1985). (Any detected peaks within the first and last 1 s of data were ignored to exclude possible false positive ‘partial’ R-peaks.) Although more advanced R-peak detection methods have been proposed (e.g. wavelet transforms, filter banks, neural networks, hidden Markov models, and genetic algorithms; for a comprehensive review, see Kohler *et al* 2002), the classic method was used here so as to align the present methodology with the widest possible set of past and current experimental literature.

**2.3.2. R-peak refinement.** Next, the timestamps of detected R-peaks (i.e. R-wave fiducial points) were optionally refined (independently for each  $SF_{Ex}$ ) using a two-step procedure. First, a piecewise cubic spline (e.g. Daskalov and Christov 1997) was fit to each downsampled ECG (using Matlab's `spline.m`), and resampled at 1000 Hz (as in Daskalov and Christov 1997, Bhatia *et al* 2010).

Second, for each R-peak detected in the ECG at  $SF_{Ex}$ , the local maximum of the interpolated and resampled ECG was located (i.e. within a window of  $\pm 100$  ms of the detected R-peak), yielding the refined R-peak time series. It is worth noting that the procedure of (1) downsampling an ECG, (2) detecting R-peaks, and (3) refining R-peak timing is *not* necessarily equivalent to the procedure of (1) downsampling an ECG, (2) interpolating the ECG with a cubic spline, and (3) detecting R-peaks. The first order of operations, used in the present paper, will not change the accuracy of R-peak detection. That is, any R-peaks that are 'missed' at a low  $SF_{Ex}$  will also be missed even after R-peak refinement is applied. Performing R-peak refinement in this manner is a cleaner experimental manipulation, however, as it guarantees the same number of detected R-peaks between nonrefined and refined versions of a particular R-peak series at a given  $SF_{Ex}$ .

#### 2.4. RR interval series outlier detection, PTBDB record exclusion, and $SF_{Ex}$ exclusion

The first-order difference of each R-peak series yields an *RR interval* series. Artifacts (outliers) in an RR interval series can arise from several sources (e.g. Friesen *et al* 1990). A disturbance in the cardiac electrical rhythm (ectopic beats), electrical noise during ECG recording, or errors during QRS detection itself can compromise the accuracy of HRV measures (e.g. Kim *et al* 2007, 2009). An additional source of outliers—one of particular relevance to the current project—is SF: if an ECG is digitized too sparsely, R-spike events may be partially or entirely missed. In order to present the 'cleanest' experimental investigation of the role of SF on HRV measures, only those ECG records which (1) yielded an RR series at  $SF_{Or}$  (1000 Hz) that was free from any outliers (as detected by the widely used algorithm developed by Berntson *et al* 1990 and implemented by Kaufmann *et al* 2011), and (2) did not accrue any additional spurious or missed R-peaks relative to the R-peaks detected at  $SF_{Or}$  at any  $SF_{Ex}$ .

This two-step procedure is detailed in section 2.4 of the supplementary material ([stacks.iop.org/PM/36/091827/mmedia](http://stacks.iop.org/PM/36/091827/mmedia)). To summarize, of the original set of 52 healthy controls, 47 possessed an ECG record that was free of RR outliers at  $SF_{Or}$ . When examining the presence of false positive and false negative R-peaks (with respect to R-peaks at  $SF_{Or}$ ) at each  $SF_{Ex}$ , all R-peaks extracted from all 47 ECG records were present down to  $SF_{Ex} = 71.43$  Hz (i.e. downsampling the 1000 Hz ECG by taking every 14th data point). Thus, the final data set used in the FPR analysis was a total of 658 RR interval series (47 subjects  $\times$  14 SFs), all of which were free of RR interval outliers.

#### 2.5. HRV measure calculation

A total of 24 HRV measures (6 time-domain and 18 frequency-domain) were calculated on each outlier-free RR interval series at each SF. All calculations were derived using publicly available Matlab toolboxes, as detailed briefly below. (Thorough explanations of these HRV measures are available elsewhere; e.g. TF96; Rajendra Acharya *et al* 2006, Bravi *et al* 2011, Tarvainen and Niskanen 2012, Smith *et al* 2013a).

**2.5.1. Time-domain.** Six time-domain measures were calculated on each RR interval series. (Formulas for all these time-domain measures are provided in section 2.5.1 of the

supplementary material) ([stacks.iop.org/PM/36/091827/mmedia](http://stacks.iop.org/PM/36/091827/mmedia)). Four are defined directly using simple transformations of the RR interval series: the average RR interval ('AVNN'); the standard deviation of RR intervals ('SDNN'); the standard deviation of successive RR interval differences ('SDSD'); the square root of the mean of squared successive RR interval differences ('RMSSD'). To this list are added the two most common measures derived from a Poincaré plot: the length of the short axis ('SD1') and long axis ('SD2') of the fitted ellipse. Although SD1 and SD2 are sometimes classed as *nonlinear* measures of HRV, their formulas can, in fact, be expressed in terms of SDNN and SDSD (Brennan *et al* 2001, equations (8) and (12)). This intimate mathematical relationship distinguishes SD1 and SD2 from other nonlinear measures (noted briefly in section 2.5.3), and motivated their inclusion here.

**2.5.2. Frequency-domain.** A total of 18 frequency-domain measures derived from various features of three distinct power spectrum density (PSD) estimates of the periodicities present in an RR interval series were calculated. Prior to PSD estimation, each RR interval series was detrended using the widely used smoothness priors method proposed by Tarvainen *et al* (2002) and implemented in Matlab, with the regularization parameter ( $\lambda$ ) set to 500 (per Tarvainen and Niskanen 2012). All three PSD estimates were performed using the freqDomain.m Matlab script from the HRVAS package (Ramshur 2014), modified slightly to accommodate the PTBDB data and parameter values suggested by Tarvainen and Niskanen (2012).

The first PSD estimate was obtained from a fast Fourier transform (FFT) using a Welch periodogram (calling Matlab's pwelch.m), with the RR interval series resampled at 4 Hz, an FFT window length of 256 samples (i.e. 64 s), and a window overlap of 50%.

The second PSD estimate was obtained from an autoregressive (AR) model using a Burg periodogram (calling Matlab's pburg.m), with the RR interval series resampled at 4 Hz and a model order of 16. Spectral factorization of the AR output was not performed, per the recommendation of Tarvainen and Niskanen (2012, p 28).

The third PSD estimate was obtained from least-squares spectral analysis via a Lomb – Scargle (LS) periodogram (calling the HRVAS script lomb2.m). Less well-known than FFT and AR approaches, the LS method of PSD estimation is considered superior to FFT and AR methods (for a detailed discussion, see Clifford and Tarassenko 2005), as it operates directly on the observed RR interval values and does not require their projection onto a regular time axis (via cubic spline interpolation and resampling; here, at 4 Hz), which attenuates higher-frequency periodicities present in the original signal.

All three periodograms were calculated with 512 points-per-Hz resolution. For each analysis method, the PSD itself was quantified from the power spectrum using trapezoidal integration (calling Matlab's trapz.m) within the low- (0.04 to 0.15 Hz) and high- (0.15 to 0.40 Hz) frequency bands, and converted to  $\text{ms}^2/\text{Hz}$  units. The final set of frequency-domain measures from each PSD estimation method were: low- and high-frequency power in absolute units ('LF<sub>au</sub>', 'HF<sub>au</sub>'), as well as their ratio ('LF<sub>au</sub>/HF<sub>au</sub>') and their sum ('LF<sub>au</sub> + HF<sub>au</sub>'); and LF<sub>au</sub> and HF<sub>au</sub> expressed in normalized units: 'LF<sub>nu</sub>' = LF<sub>au</sub> / (LF<sub>au</sub> + HF<sub>au</sub>), and 'HF<sub>nu</sub>' = HF<sub>au</sub> / (LF<sub>au</sub> + HF<sub>au</sub>). (This 'simplified' definition excludes very low frequency power from the denominator, which cannot be fully resolved in 2 min recordings; for a more detailed explanation, see section 2.5.2 of the supplementary material) ([stacks.iop.org/PM/36/091827/mmedia](http://stacks.iop.org/PM/36/091827/mmedia)).

**2.5.3. HRV measures not examined.** A number of widely used measures of HRV were, after careful consideration, excluded from the present analysis. For some measures, the decision was motivated by the 'limitations' of imposed by the available set of 2 min ECGs. Although TF96 notes that 5 min ECG recordings are 'preferred' for estimating HRV, it also states that

recordings of ‘approximately 2 min’ accurately quantify the low-frequency component of HRV. Demonstrations of the diagnostic sensitivity of two-minute analysis of HRV in both the time-domain (e.g. Dekker *et al* 2000) and frequency-domain (e.g. Singh *et al* 1998, Hall *et al* 2004) are readily found in the literature.

Additional empirical evidence supports the validity of 2 min HRV, at least for some measures. Schroeder *et al* (2004) examined several HRV statistics (AVNN, SDNN, RMSSD, HF<sub>au</sub>, LF<sub>au</sub>, HF<sub>nu</sub>, and LF<sub>nu</sub>) derived from 2 min versus 6 min ECGs at 1000 Hz are highly-correlated (their table 5), and yield a similar breakdown of variance components (e.g. between-person, between-visit, within-visit) when analyzed using mixed linear models (their table 4). SD1 and SD2 were not examined in Schroeder *et al* (2004), but would be expected to show similar performance given their close mathematical relationship to SDSA and SDNN (Brennan *et al* 2001).

By contrast, the diagnostic sensitivity of a number of other measures of HRV is predicated on longer recordings of HRV. Geometric measures (e.g. the HRV triangular index and the triangular interpolation of RR interval histogram) require at least 20 min (and ‘preferably’ 24 h, per TF96), as do entropy measures (e.g. approximate entropy and sample entropy; Hogue *et al* 1998, Vikman *et al* 1999). Another popular nonlinear HRV measure, detrended fluctuation analysis, is typically analyzed using windows of at least 8000 intervals, or roughly 2 h of continuous data (e.g. Huikuri *et al* 2003, Voss *et al* 2009).

A final exclusion was the ‘pNN<sub>x</sub> family’ of statistics (Mietus *et al* 2002): the percentage of RR intervals greater than  $x$  ms (typically, with  $x = 50$ ). Although pNN<sub>x</sub> is commonly encountered in the literature and often reported as showing high correlations with other measures which index rapid RR interval changes (such as RMSSD and HF<sub>au</sub>; e.g. Task Force 1996a, Kleiger *et al* 2005, Thayer and Fischer 2009, Smith *et al* 2013b), its statistical properties imply some important caveats. The value of  $x$  plays a key role in its diagnostic sensitivity; Mietus *et al* (2002) argue that  $x = 50$  is less sensitive than lower values of  $x$ . An additional concern with pNN<sub>x</sub> is the potential for ‘0’-valued statistics, either as a consequence of a too-short recording duration or a too-high value of  $x$ , which could mask subtle but important differences among subjects with low HRV. By contrast, RMSSD is a continuously valued statistic; a ‘0’ value would only emerge if all RR intervals were identical (a physiologically implausible scenario).

## 2.6. Quantifying the influence of SF and R-peak refinement

**2.6.1. Quantifying differences between samples.** The choice of the two-sample test statistic used to quantify differences in HRV values at SF<sub>Or</sub> versus SF<sub>Ex</sub> is important. Because distributions of HRV values vary widely in terms of their (non)normality (e.g. positively skewed SDSA, RMSSD, LF<sub>au</sub>, and HF<sub>au</sub>), assessment of between-sample differences typically follow one of two analysis options (see Riniolo and Porges 2000, Ellis *et al* 2008). One option submits the original data values to a natural log (ln) transform ( $\ln(\text{value} + 1)$ ) prior to analysis with parametric statistical tests (e.g. a two-sample  $t$ -test). Another option utilizes a rank-based (nonparametric) test (for an overview, see Conover and Iman 1981), eliminating the need for any data transformations; for example, the Mann–Whitney–Wilcoxon  $U$ -test. Unlike a two-sample  $t$ -test (i.e. a test of group mean differences), a  $U$ -test is sensitive not only to differences in distribution location (i.e. the median rank of each sample), but also to differences in distribution shape (Hart 2001); it is also more powerful than a  $t$ -test under a variety of non-normal distributional assumptions (e.g. Blair and Higgins 1980, Fay and Proschan 2010).

For sake of comparison and completeness, both a  $t$ -test (Matlab’s `ttest2.m` applying the Welch–Satterthwaite correction for unequal variances; see Ruxton 2006) and a  $U$ -test (Matlab’s `ranksum.m`, on the original values) were performed. All time-domain measures (other than AVNN) and all power spectral density values in absolute units were ln-transformed prior to two-sample tests.



**2.6.2. Quantifying FPR.** A Monte Carlo subsampling paradigm was implemented to quantify the observed FPR of the two-sample tests. (For a theoretical and practical overview of Monte Carlo-based evaluations of FPR, see Serlin 2000). For each of 100 000 iterations (a number empirically determined to yield a stable estimate of FPR values, as detailed in section 2.6.2 of the supplementary material) ([stacks.iop.org/PM/36/091827/mmedia](http://stacks.iop.org/PM/36/091827/mmedia)), a random permutation of the integers 1 to 47 was taken. The first 20 values defined the set of subjects for sample 1, and the next 20 values defined the set of subjects for sample 2 (thus leaving seven subjects out per iteration). The *t*- and *U*-tests were then performed across the full set of HRV measures (separately for values calculated using nonrefined and refined R-peaks), with sample 1 at SF<sub>Or</sub> and sample 2 at successive levels of SF<sub>Ex</sub>. The binary significance of each test (i.e. either  $p < 0.05$  or  $p \geq 0.05$ ) was recorded for each iteration. (For the *U*-test, significance was determined using the exact critical value rather than its Normal approximation; see Bergmann *et al* 2000). The observed FPR was then taken as the proportion of iterations that yielded a significant test result.

The utility of this approach can be illustrated by analyzing the SDNN data presented in figure 2, using 100 000 Monte Carlo iterations 20 data points per sample. A separate FPR was calculated for each of the SF<sub>Ex</sub> values versus SF<sub>Or</sub>, and are presented in sequence as ‘M.C. FPR’*m*, below the figure. These results present a rather different picture than the paired-sample *t*-tests performed on the same data: in fact, two-sample tests between SDNN at 1000 Hz and SDNN at 500 Hz, 333 Hz, and 250 Hz had nearly *identical* FPRs ( $\approx 0.0508$ ) just slightly higher than the expected  $\alpha = 0.05$ . This indicates that independent samples of SDNN values obtained at any of these SFs could be compared without risk of inflating the rate of false positive test results. As argued in section 1.3, we believe that such an inference is of greater practical significance than the *p*-value associated with a paired-samples test.

### 3. Results

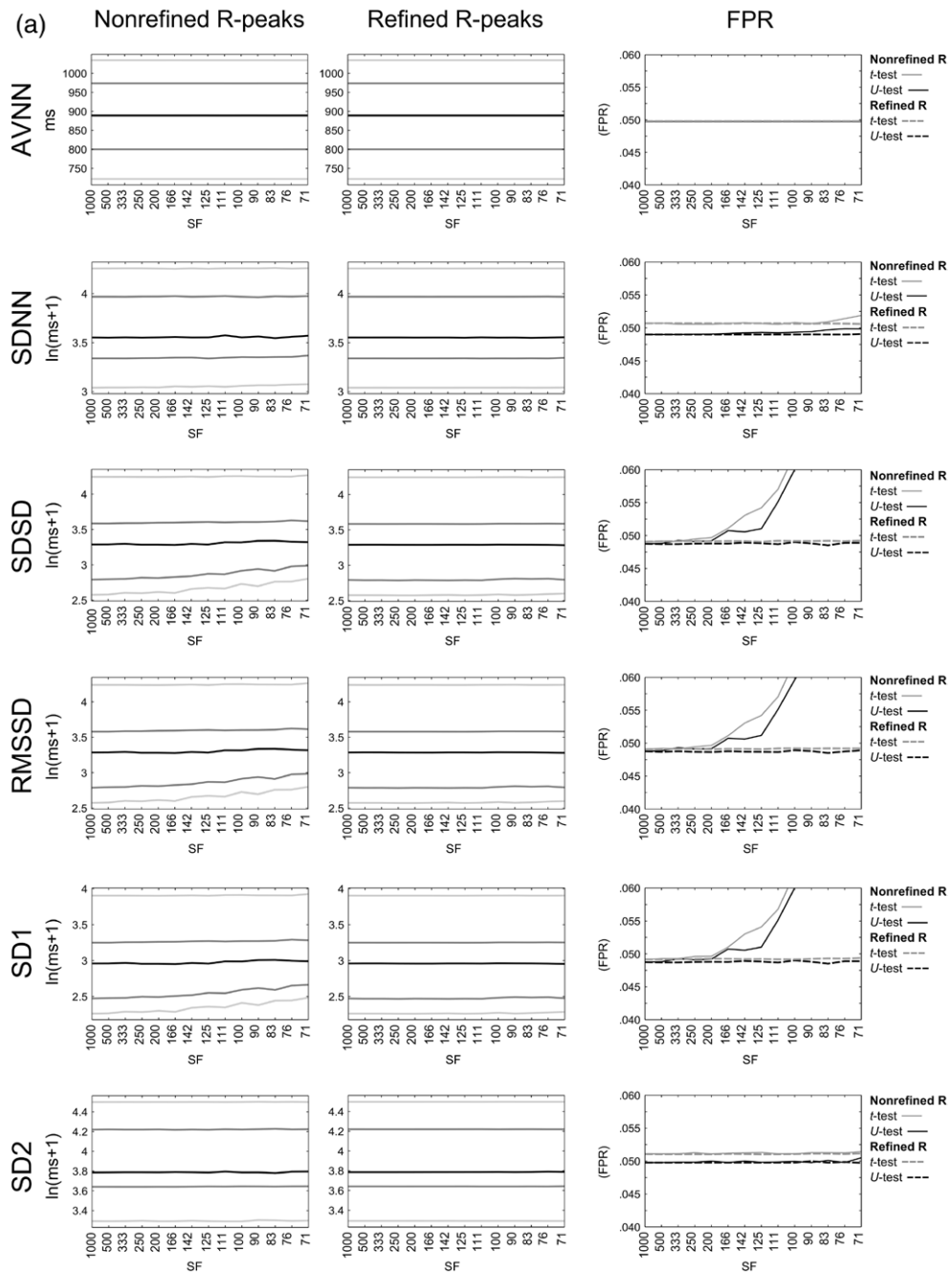
Figure 5 presents a summary of each HRV measure as a function of SF, for SF<sub>Or</sub> (1000 Hz) down to SF<sub>Ex</sub> = 71.42 Hz, arranged into four sets: time-domain (figure 5(a)), FFT model (figure 5(b)), autoregressive model (figure 5(c)), and Lomb–Scargle model (figure 5(d)). Three subplots are shown for each measure. The first two are percentile-based summaries of observed HRV values derived from R-peaks that were nonrefined (left plot) versus refined (middle plot). (Measures which were natural-log transformed are labeled explicitly.) The results of the Monte Carlo-based FPR analysis for each HRV measure (and using either nonrefined or refined R-peaks) are presented in the right plot.

Results are presented in three parts. First, some remarks about SF-induced changes in different HRV measures. Second, some remarks about general differences between *t*-test and *U*-test FPRs. Finally, a summary of how FPR is influenced by SF and R-peak interpolation.

#### 3.1. SF, R-peak interpolation, and HRV measures

In the absence of R-peak interpolation, several measures showed clear changes as SF<sub>Ex</sub> decreased: an *inflation* of low-valued statistics (e.g. the 10th percentiles of S<sub>SDS</sub>, R<sub>MSD</sub>, S<sub>D1</sub>, and all three HF<sub>nu</sub> values), or a *deflation* of high-valued statistics (e.g. the 90th percentiles of all three LF<sub>nu</sub> and all three LF<sub>au</sub>/HF<sub>au</sub> values) relative to SF<sub>Or</sub>.

The explanation for this finding is relatively straightforward: as SF decreases, the magnitude of stochastic measurement error at each R-peak (i.e. with respect to the ‘true’ R-peak location at SF<sub>Or</sub>) increases; for an illustration of this, see figure S3 in the supplementary material ([stacks.iop.org/PM/36/091827/mmedia](http://stacks.iop.org/PM/36/091827/mmedia)). This increased error at the level of R-peaks translates into greater stochastic error in successive RR interval differences, thereby increasing the



**Figure 5.** Statistics for the 24 HRV measures: time-domain (a); and frequency-domain using fast Fourier transform (b) auto-regressive (c) and Lomb – Scargle (d) algorithms. The distribution of observed values at each SF is summarized by five percentiles: 10th (light gray), 30th (dark gray), 50th (black), 70th (dark gray), and 90th (light gray), and plotted separately for HRV measures calculated using nonrefined R-peaks (left) and refined R-peaks (middle). Observed false positive rates (FPRs) for each measure are plotted as a function of SF in the right panel.



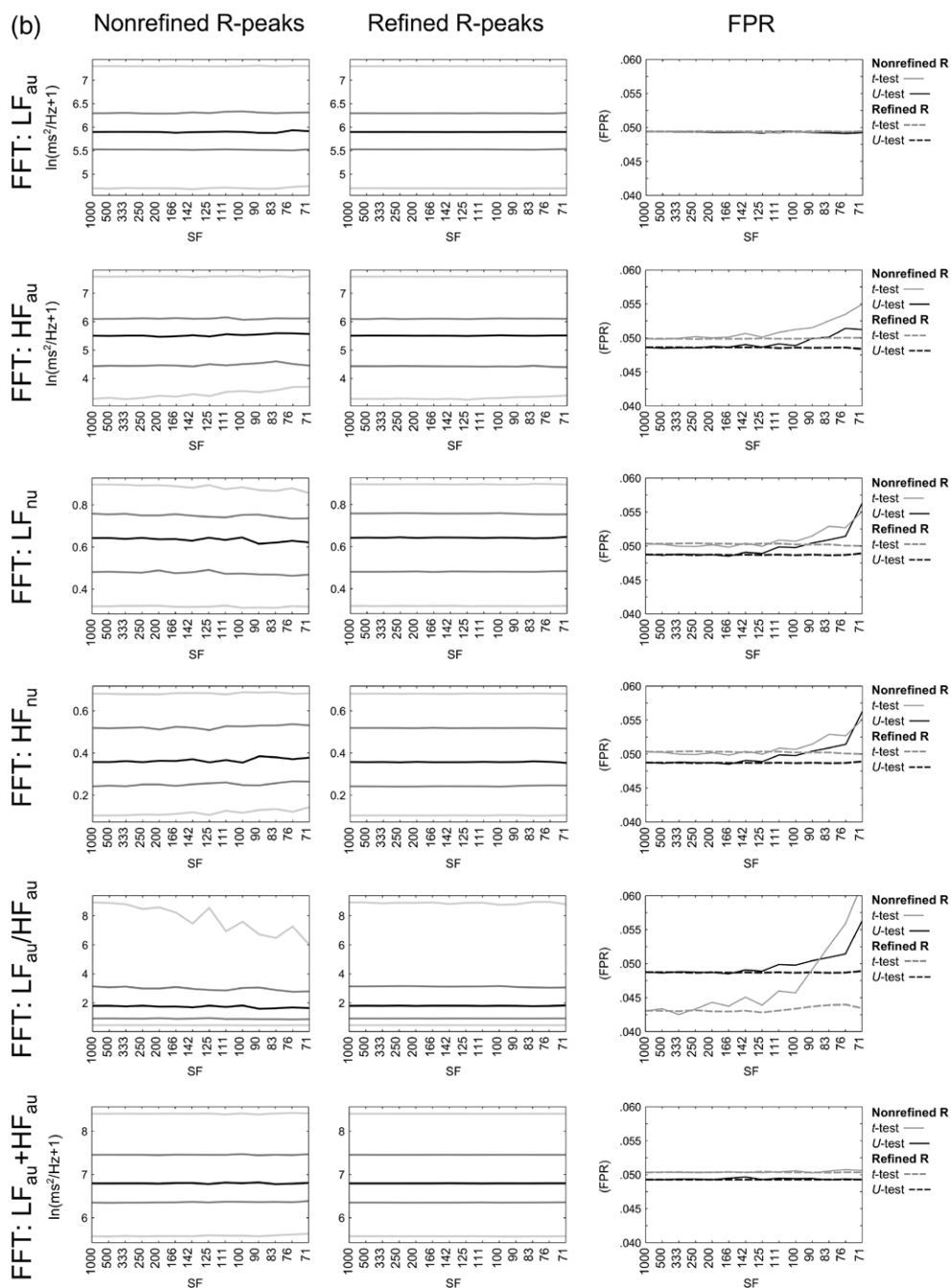


Figure 5. (Continued)

amount of high-frequency ‘noise’ in the RR interval series. In subjects with relatively low beat-to-beat changes in RR interval at SF<sub>Or</sub>, as SF decreases, this increasing amount of noise leads to increasing *inflation* among HRV measures that are sensitive to rapid beat-to-beat fluctuations (SDSD, RMSSD, SD1, and HF<sub>au</sub>). Even though LF<sub>au</sub> does not exhibit any systematic

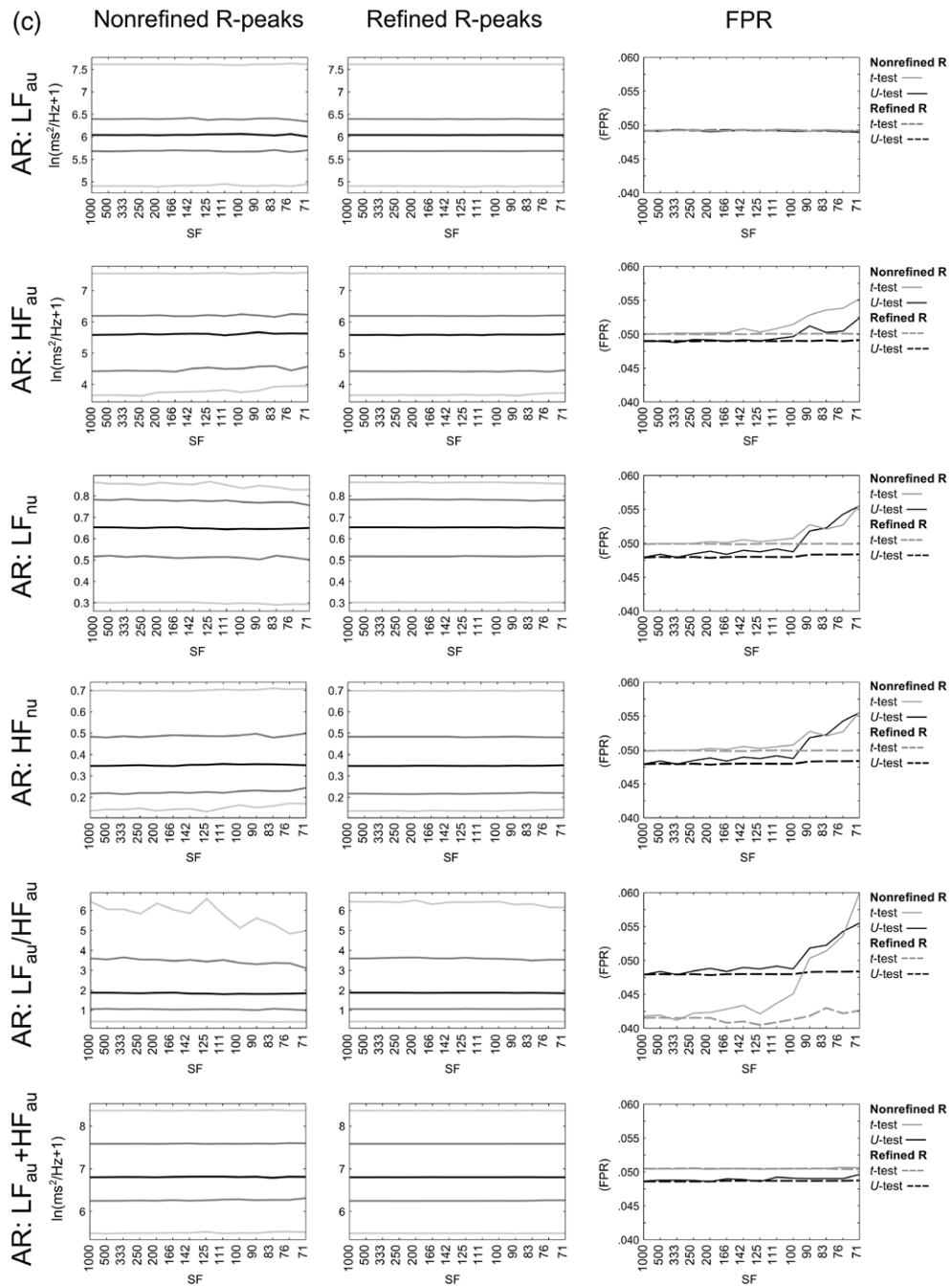


Figure 5. (Continued)

SF-induced inflation or deflation as SF decreases, the inflation of HF<sub>au</sub> as SF decreases leads to a corresponding *deflation* of ratio-based measures with LF<sub>au</sub> in the numerator (LF<sub>nu</sub> and LF<sub>au</sub>/HF<sub>au</sub>).

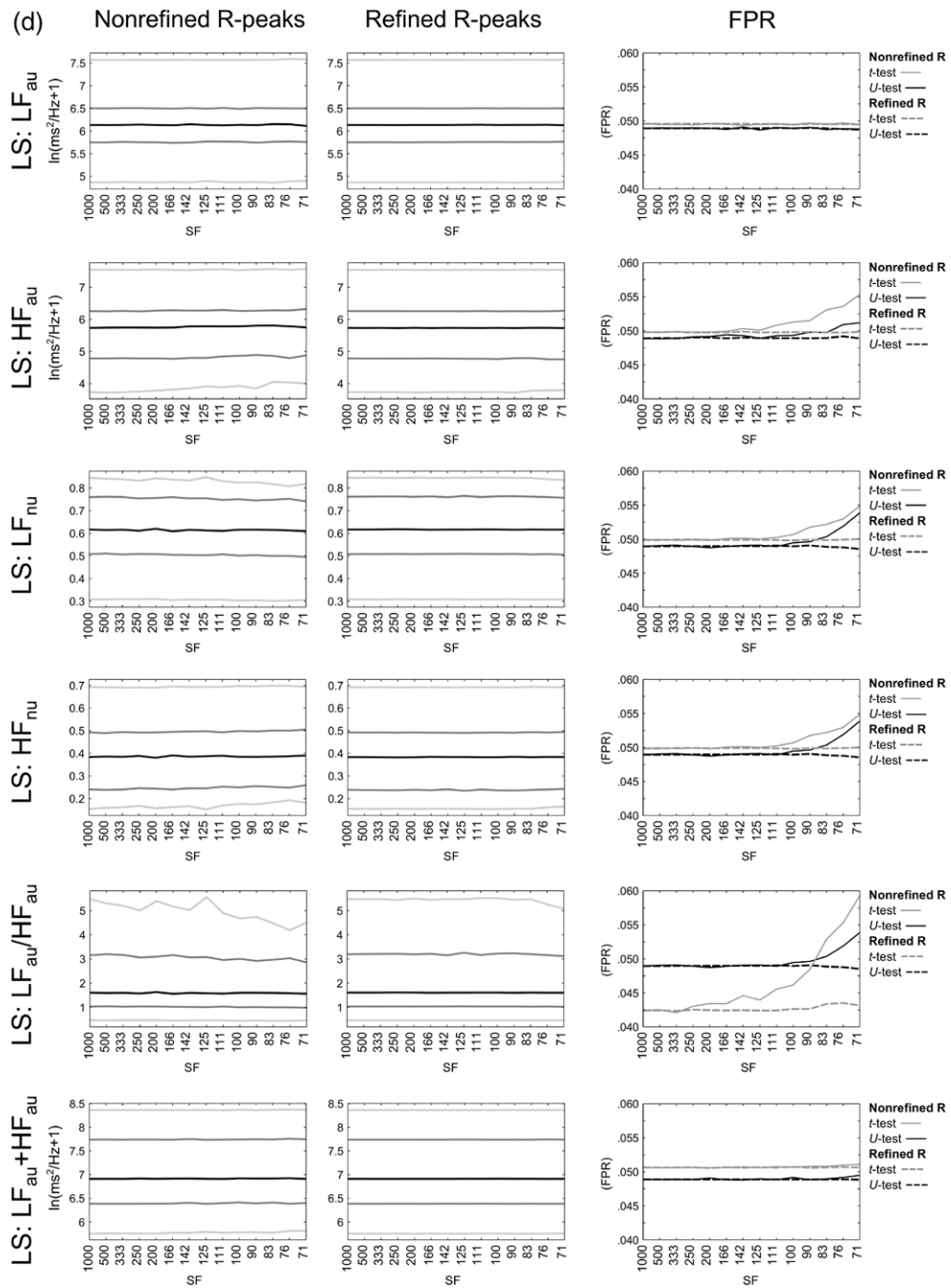


Figure 5. (Continued)

By contrast, when R-peak interpolation *was* applied prior to HRV measure calculation, the percentile-based summary statistics are far more stable across the SF<sub>EX</sub> values examined: inflation/deflation is nearly absent in the time-domain measures (SDSD, RMSSD, and SD1)

and frequency-domain measures ( $HF_{au}$ ,  $LF_{nu}$ , and  $LF_{au}/HF_{au}$ ) that all showed clear inflation/deflation in the absence of R-peak interpolation.

Although these patterns and similarities are interesting, the results of the Monte Carlo-based FPR analysis give a better intuition about the experimental consequences of comparing samples at different SFs. We turn to those results next.

### 3.2. General comments about FPRs

Two general comments about FPRs may be offered before discussing individual HRV measures. First, in most cases,  $U$ -test FPRs were consistently lower than  $t$ -test FPRs. This was not unexpected, due to the noncontinuous relationship between rank-based statistics and their 'exact'  $p$ -values (see Bergmann *et al* 2000). For example, when  $n = 20$ , a  $U$ -test is significant if  $U \leq 127$ , and yields an expected  $\alpha = 0.0491$  rather than  $\alpha = 0.05$ . The *actual* difference between the  $t$ - and  $U$ -test FPRs tends to exceed the expected difference of  $(0.05 - .0491) = 0.0009$ , particularly when R-peak refinement was used. This suggests that use of R-peak refinement in conjunction with a  $U$ -test conveys *true* protection against false positive two-sample test results for many HRV measures.

Second, FPRs showed some variation around the expected  $\alpha = 0.05$ , even when both sets of HRV values were calculated at  $SF_{Or}$ . As with any Monte Carlo-based operation, such variation is inevitable, as is illustrated by the simulation detailed in section 3.2 of the supplementary material ([stacks.iop.org/PM/36/091827/mmedia](http://stacks.iop.org/PM/36/091827/mmedia)). Thus, for sake of simplicity, the amount of FPR inflation at a given  $SF_{Ex}$  will be described relative to the FPR level at  $SF_{Or}$ . For example, 'inflation of +0.001 at 200Hz' would indicate that a particular FPR that was higher by 0.001 at 200Hz relative to that same FPR at 1000Hz (e.g. 0.052 versus 0.051).

### 3.3. SF, R-peak interpolation, and FPRs

Figure 5 reveals clear differences for nonrefined HRV measures and refined HRV measures. The benefits of R-peak interpolation on the stability of HRV values, noted above, extended to the pattern of observed FPRs. Specifically, all HRV measures showed inflation of less than +0.001 across all  $SF_{Ex}$  values.

When R-peak interpolation was not utilized, however, FPRs showed varying patterns and rates of inflation. Some measures performed much like their refined counterparts, with inflation of less than +.001 (AVNN, SD2, all three  $HF_{au}$ , and all three  $LF_{au} + HF_{au}$ ) or just over +.001 (SDNN) across all  $SF_{Ex}$ . The remaining time-domain measures (SDSD, RMSSD, and SD1) showed FPR inflation of +.005 at 125 Hz and +.01 (or more) by 100 Hz. The remaining frequency-domain measures (all three  $HF_{au}$ ,  $LF_{nu}$ , and  $HF_{nu}$ , and  $LF_{au}/HF_{au}$ ) all showed rising FPR inflation below 100 Hz, reaching +.005 (or more) by 71 Hz. (The substantially lower FPRs for  $LF_{au}/HF_{au}$   $t$ -test results are likely a consequence of the positively skewed distribution of values, which are not typically log-transformed prior to calculation. The nonparametric  $U$ -test yielded a more typical pattern of FPRs.)

Thus, to summarize, negligible FPR inflation (i.e. less than 0.005 higher than the expected 0.05 level) was maintained under the following conditions:

1. When R-peak refinement *was* utilized: down to 71 Hz for all time-domain and frequency measures.
2. When R-peak refinement was *not* utilized, (a) down to 100 Hz for frequency-domain measures inflation; (b) down to 125 Hz for SDSD, RMSSD, and SD1; and (c) down to 71 Hz for AVNN, SDNN, and SD2.

The implications of these findings will be discussed next.

## 4. Discussion

The present paper aimed to take a fresh look at the relationship between sampling frequency (SF), R-peak interpolation, and common time- and frequency-domain measures of HRV. Prior investigations into this topic (see table 1) typically used a *within-subjects* approach to quantify the SF-induced changes in HRV values. In our view, the results of such an analysis would offer insights into a rather limited experimental scenario: changing the sampling rate over the course of a longitudinal trial. By contrast, a *between-subjects* analysis—specifically, testing whether HRV values from two different sets of subjects at two different SFs are consistently different—offers insights that would be highly relevant for the growing number of meta-analytic studies of HRV, or for studies wishing to compare ambulatory HRV with normative HRV values collected at a potentially different SF.

Two specific hypotheses were evaluated. First, that repeated two-sample comparisons of HRV values at 1000 Hz in one healthy sample and increasingly lower SFs in an independent healthy sample would lead to progressively higher false positive rates (FPRs). Second, that the use of cubic spline interpolation to ‘refine’ the R-wave fiducial point would reduce the severity of FPR inflation. Previous in-depth discussions of FPR in the context of psychophysiology have been largely restricted to techniques designed to maintain a valid familywise error rate when performing multiple comparisons within the context of repeated-measures analysis of variance (e.g. Jennings 1987, Vasey and Thayer 1987). To our knowledge, the present study marks the first use of Monte Carlo methods to derive the observed FPR of a particular statistical test within the domain of psychophysiology.

The present set of results adds some concrete insight into the behavior of HRV measures when comparing values measured at different SFs. The language used in two classic reference papers on HRV methodology (Task Force 1996a, pp 1047–1048; Berntson *et al* 1997, p 630) was rather cautious with respect to SF recommendations: 128 Hz ‘may be useable’ and 250 Hz ‘may be adequate’ in some cases, but 500 Hz ‘or perhaps higher’ would be ‘optimal and generally applicable’. Also, that R-peak interpolation of ECGs at lower SFs ‘may’ result in HRV measures that ‘behave satisfactorily’. Our own results can be stated more concretely: without R-peak interpolation, all examined HRV measures showed negligible inflation of FPRs down to 125 Hz; with R-peak interpolation, inflation was negligible down to 71.43 Hz (i.e. decimating a 1000 Hz signal by 14).

That FPRs remain statistically valid even when comparing HRV values at 1000 Hz and 125 Hz (and, by inference, at SF values in between them) is particularly notable, as the bulk of prior published work on HRV uses SFs within this range (see table 2), including contemporary studies at 1000 Hz and numerous legacy studies utilizing a Holter monitor at 128 Hz. It would also imply that values recorded in the laboratory at a high SF in one sample could be compared with values recorded ‘in the field’ at a battery-saving 128 Hz in another sample from the same population without systematically inflating the FPR above the target  $\alpha = 0.05$ —under the assumption that any differences in sample demographics, signal recording, and data analysis would not *themselves* induce a confound.

### 4.1. Caveats

Several design aspects of the present study stand in marked departure from previous investigations. Although we believe that these changes result in cleaner experimental methods and inferentially stronger results, they should (at a minimum) be noted as caveats.

A first caveat pertains to the chosen dataset of healthy subjects. The ECG records selected from the PTBDB database were high quality and relatively free of recording artifacts, were digitized at a high original sampling frequency ( $SF_{Or} = 1000\text{Hz}$ ), and sampled a wide age

range (17 to 81). They were also more numerous (47 unique subjects) than nearly all previous forays into this topic (see table 1). Nevertheless, certain properties of this dataset (e.g. the wide age range of subjects) may translate into undetected experimental confounds. We also note that we opted not to use simulated ECGs (e.g. as are available using the compelling and comprehensive ECGSYN model proposed by McSharry and colleagues (McSharry *et al* 2003, McSharry and Clifford 2006). Although doing so would have solved the ‘low- $N$ ’ problem, it would have introduced its own caveats. ECGSYN, for example, generates ECGs by first simulating an RR interval series with precise spectral characteristics (which the user is required to specify) and *then* generating the associated ECG using a series of time-varying differential equations which capture different waveform components. We chose the PTBDB dataset so as to make our approach as data-driven as possible, mirroring an actual experiment. However, the relatively short duration of ECG records (2 min) precluded studying several widely used HRV measures (see section 2.5.3). Were such a database (e.g. 20 min ECGs at 1000 Hz) made available in the future, our approach could be used to explore how SF influences geometric and nonlinear measures of HRV.

A second caveat pertains to the chosen methods of R-peak detection (Pan and Tompkins 1985) and R-peak refinement (cubic spline interpolation; e.g. Daskalov and Christov 1997; see figure 2). Numerous methods for both R-peak detection (e.g. for a review, see Kohler *et al* 2002) and R-peak refinement (see table 1) are available, and may yield improvements in R-peak detection power or R-peak refinement accuracy over the two ‘classic’ operations used here. A systematic comparison of how different R-peak detection and R-peak refinement options affect FPRs is beyond the scope of one paper. Thus, our choice was driven by a desire to make the present methodology (and thus its findings) aligned with the broadest possible set of experimental studies dating from the 1990s to the present.

A third caveat pertains to the exclusion of ECG records which yielded spurious R-peaks, missing R-peaks, or other potential outliers (as identified using the algorithm by Berntson *et al* 1990) across the entire set of  $SF_{Ex}$  values examined (i.e. relative to detected R-peaks in the ECGs at  $SF_{Or}$ ). This decision was made to ensure that the observed FPR patterns were driven by the target experimental manipulations of ECG downsampling and R-peak refinement, and not confounded with an *additional* source of error (i.e. possible outliers). Importantly, this choice may mean that our results reflect a ‘best case scenario’: FPRs would almost certainly be higher, for example, if outliers due to missing R-peaks were present at low SFs. Furthermore, although we found no R-peak outliers in the *current* sample of 47 ECGs at 71.43 Hz, Task Force (1996a) and Berntson *et al* (1997) note that SFs below 100 Hz have the *potential* to lead to missed or partial R-spikes.

A fourth caveat pertains to the reliance on FPR *itself* as a means to assess the ‘significance’ of SF-induced changes in HRV values. Although we have argued in section 1.3 that an FPR enables more practically significant insights than does a traditional within-subjects approach quantifying SF-induced error in HRV values, FPR is but one side of the coin—or, perhaps, one side of the dice—as is discussed in the next section.

#### 4.2. False positives and true positives

The FPR of a binary classification test is only one of several statistics that can be calculated from a  $2 \times 2$  (see figure 3). FPR is the probability that two samples from the *same* population are incorrectly labeled as statistically different. Another statistic is the true positive rate (TPR): the probability that two samples from *different* populations are correctly labeled as statistically different. Put another way, TPR is the probability of obtaining a statistically significant difference when a real difference is present. In the statistics literature, TPR is more



often referred to as *statistical power* (e.g. Cohen 1992); in the clinical medicine literature, TPR is more often referred to as *diagnostic sensitivity*.

Evaluating the role of SF on diagnostic sensitivity is a substantially more challenging task than the one addressed in the current paper, as the *specific* disease or condition being examined introduces a new set of challenges. It may be, for example, that measure  $m_1$  has greater sensitivity than measure  $m_2$  when comparing populations  $p_1$  and  $p_2$ , whereas measure  $m_2$  has greater sensitivity than measure  $m_1$  when comparing populations  $p_1$  and  $p_3$ —even before differences in SF enter the picture. Evaluating the effect of SF on sensitivity becomes a complex task when considering a large set of HRV measures  $\{m_1, m_2, \dots\}$  and populations  $\{p_1, p_2, \dots\}$ . Such work must be left to future studies.

#### 4.3. Conclusion

A careful look at the how sampling frequency (SF) and R-peak interpolation influence HRV was presented. Building upon repeated tests of *statistical* significance (i.e. a Monte Carlo analysis of false positive rates in two-sample tests), findings of *practical* significance are offered regarding the statistical validity of comparing samples of subject data collected at different SFs. Given the recent increasing attention to the importance of false positive results across numerous scientific disciplines and methodologies (e.g. Ioannidis 2005), we hope that the present results will provide researchers with useful insights regarding FPR inflation for one widely explored physiological construct: heart rate variability.

#### Author notes

Kind thanks to Tobias Kaufmann for his Matlab implementation of the Berntson *et al* (1990) outlier detection algorithm, to Graham Percival for discussions regarding frequency-domain analytics, and to Boyd Anderson for discussions regarding inferences about false positive rate analysis. This research was supported by a grant from the Singapore Ministry of Education (grant R-252-000-516-112) to Y W.

#### References

- Abboud S and Barnea O 1995 Errors due to sampling frequency of the electrocardiogram in spectral analysis of heart rate signals with low variability *Comput. Cardiol.* 461–3 ([http://ieeexplore.ieee.org/xpls/abs\\_all.jsp?arnumber=482685](http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=482685))
- Berger J O 2003 Could Fisher, Jeffreys and Neyman have agreed on testing? *Stat. Sci.* **18** 1–32
- Bergmann R, Ludbrook J and Spooren W P J M 2000 Different outcomes of the Wilcoxon–Mann–Whitney test from different statistics packages *Am. Stat.* **54** 72–7
- Berntson G G *et al* 1997 Heart rate variability: origins, methods, and interpretive caveats *Psychophysiol.* **34** 623–48
- Berntson G G, Quigley K S, Jang J F and Boysen S T 1990 An approach to artifact identification: application to heart period data *Psychophysiol.* **27** 586–98
- Bhatia V, Rarick K R and Stauss H M 2010 Effect of the data sampling rate on accuracy of indices for heart rate and blood pressure variability and baroreflex function in resting rats and mice *Physiol. Meas.* **31** 1185
- Bianchi A M, Mainardi L, Petrucci E and Signorini M G 1993 Time-variant power spectrum analysis for the detection of transient episodes in HRV signal *IEEE Trans. Biomed. Eng.* **40** 136–44
- Blair R C and Higgins J J 1980 The Power of t and wilcoxon statistics a comparison *Eval. Rev.* **4** 645–56
- Boulos M N, Wheeler S, Tavares C, Jones R *et al* 2011 How smartphones are changing the face of mobile and participatory healthcare: an overview, with exmple from eCAALYX *Biomed. Eng. Online* **10** 24

- Bousseljot R, Kreiseler D and Schnabel A 1995 Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das internet *Biomed. Tech. Eng.* **40** 317–8
- Bragge T, Tarvainen M P, Ranta-aho P O and Karjalainen P A 2005 High-resolution QRS fiducial point corrections in sparsely sampled ECG recordings *Physiol. Meas.* **26** 743
- Bravi A et al 2011 Review and classification of variability analysis techniques with clinical applications *Biomed. Eng. Online* **10** 1–27
- Brennan M, Palaniswami M and Kamen P 2001 Do existing measures of Poincare plot geometry reflect nonlinear features of heart rate variability? *IEEE Trans. Biomed. Eng.* **48** 1342–7
- Castiglioni P, Piccini L and Di Rienzo M 2003 Interpolation technique for extracting features from ECG signals sampled at low sampling rates *Comput. Cardiol.* 481–4 ([http://ieeexplore.ieee.org/xpls/abs\\_all.jsp?arnumber=1291197](http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1291197))
- Chalmers J A, Quintana D S, Abbott M J-A and Kemp A H 2014 Anxiety disorders are associated with reduced heart rate variability: a meta-analysis *Affect Disord. Psychosom. Res.* **5** 80
- Chellakumar P J, Brumfield A, Kunderu K and Schopper A W 2005 Heart rate variability: comparison among devices with different temporal resolutions *Physiol. Meas.* **26** 979
- Clifford G D 2003 rpeakdetect.m ([www.mit.edu/~gari/CODE/ECGtools/ecgBag/rpeakdetect.m](http://www.mit.edu/~gari/CODE/ECGtools/ecgBag/rpeakdetect.m))
- Clifford G D and Tarassenko L 2005 Quantifying errors in spectral estimates of HRV due to beat replacement and resampling *IEEE Trans. Biomed. Eng.* **52** 630–8
- Cohen J 1992 A power primer *Psychol. Bull.* **112** 155–9
- Conover W J and Iman R L 1981 Rank transformations as a bridge between parametric and nonparametric statistics *Am. Stat.* **35** 124–9
- Cumming G 2012 *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis* (New York: Routledge)
- Daskalov I and Christov I 1997 Improvement of resolution in measurement of electrocardiogram RR intervals by interpolation *Med. Eng. Phys.* **19** 375–9
- Dekker J M, Crow R S, Folsom A R, Hannan P J, Liao D, Swenne C A and Schouten E G 2000 Low heart rate variability in a 2 min rhythm strip predicts risk of coronary heart disease and mortality from several causes: the ARIC Study *Circulation* **102** 1239–44
- Dieter W R, Datta S and Kai W K 2005 Power reduction by varying sampling rate *Proc. of the 2005 Int. Symp. on Low power Electronics and Design (ACM)* pp 227–32
- Dinh H A N, Kumar D K, Pah N D and Burton P 2001 Wavelets for QRS detection *Proc. of the IEEE 23rd Ann. Int. Conf. of the Engineering in Medicine and Biology Society* pp 1883–7 ([http://ieeexplore.ieee.org/xpls/abs\\_all.jsp?arnumber=1020593](http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1020593))
- Ellis R J, Sollers J J III, Edelstein E A and Thayer J F 2008 Data transforms for spectral analyses of heart rate variability *Biomed. Sci. Instrum.* **44** 392–7 (PMID: 19141947)
- Fay M P and Proschan M A 2010 Wilcoxon–Mann–Whitney or *t*-test? On assumptions for hypothesis tests and multiple interpretations of decision rules *Stat. Surv.* **4** 1–39
- Free C, Phillips G, Watson L, Galli L, Felix L, Edwards P, Patel V and Haines A 2013 The effectiveness of mobile-health technologies to improve health care service delivery processes: a systematic review and meta-analysis *PLoS Med.* **10** e1001363
- Friesen G M, Jannett T C, Jadallah M A, Yates S L, Quint S R and Nagle H T 1990 A comparison of the noise sensitivity of nine QRS detection algorithms *Biomed. Eng. IEEE Trans.* **37** 85–98
- García-González M A, Fernández-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
- García-González M A, Fernández-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 *Biomed. Eng. IEEE Trans.* **56** 345–51
- García-González M A, Ramos-Castro J and Fernández-Chimeno M 2011 The effect of electrocardiographic lead choice on RR time series *Proc. of the 33rd Annual Int. Conf. IEEE Engineering in Medical and Biology Society* pp 1933–6
- Goldberger A, Amaral L, Glass L, Hausdorff J, Ivanov P, Mark R, Mietus J, Moody G, Peng C and Stanley H 2000 PhysioBank, PhysioToolkit, and PhysioNet—components of a new research resource for complex physiologic signals *Circulation* **101** E215–20
- Hall M, Vasko R, Buysse D, Ombao H, Chen Q, Cashmere J D, Kupfer D and Thayer J F 2004 Acute stress affects heart rate variability during sleep *Psychosom. Med.* **66** 56–62
- Hart A 2001 Mann–Whitney test is not just a test of medians: differences in spread can be important *BMJ* **323** 391–3



- Hejfel L and Roth E 2004 What is the adequate sampling interval of the ECG signal for heart rate variability analysis in the time domain? *Physiol. Meas.* **25** 1405–11
- Hogue C W, Domitrovich P P, Stein P K, Despotis G D, Re L, Schuessler R B, Kleiger R E and Rottman J N 1998 RR interval dynamics before atrial fibrillation in patients after coronary artery bypass graft surgery *Circulation* **98** 429–34
- Huikuri H V, Mäkikallio T H and Perkiömäki J 2003 Measurement of heart rate variability by methods based on nonlinear dynamics *J. Electrocardiol.* **36** 95–9
- Ioannidis J P 2005 Why most published research findings are false *PLoS Med.* **2** e124
- Jennings J R 1987 Editorial policy on analyses of variance with repeated measures *Psychophysiol.* **24** 474–5
- Kaufmann T, Sütterlin S, Schulz S M and Vögele C 2011 ARTiiFACT: a tool for heart rate artifact processing and heart rate variability analysis *Behav. Res. Methods* **43** 1161–70
- Kay M 2011 mHealth: New horizons for health through mobile technologies *World Health Organ.* ([http://www.who.int/entity/ehealth/mhealth\\_summit.pdf](http://www.who.int/entity/ehealth/mhealth_summit.pdf))
- Kemp A H, Quintana D S, Gray M A, Felmingham K L, Brown K and Gatt J M 2010 Impact of depression and antidepressant treatment on heart rate variability: a review and meta-analysis *Biol. Psychiatry* **67** 1067–74
- Kim K K, Kim J S, Lim Y G and Park K S 2009 The effect of missing RR-interval data on heart rate variability analysis in the frequency domain *Physiol. Meas.* **30** 1039–50
- Kim K K, Lim Y G, Kim J S and Park K S 2007 Effect of missing RR-interval data on heart rate variability analysis in the time domain *Physiol. Meas.* **28** 1485–94
- Kirk R E 1996 Practical significance: a concept whose time has come *Educ. Psychol. Meas.* **56** 746–59
- Kleiger R E, Stein P K and Bigger J T 2005 Heart rate variability: measurement and clinical utility *Ann. Noninvasive Electrocardiol.* **10** 88–101
- Kohler B-U, Hennig C and Orglmeister R 2002 The principles of software QRS detection *IEEE Eng. Med. Biol. Mag.* **21** 42–57
- Lehmann E L 1993 The Fisher, Neyman–Pearson theories of testing hypotheses: one theory or two? *J. Am. Stat. Assoc.* **88** 1242–9
- Lotufo P A, Valiengo L, Benseñor I M and Brunoni A R 2012 A systematic review and meta-analysis of heart rate variability in epilepsy and antiepileptic drugs *Epilepsia* **53** 272–82
- Maser R, Mitchell B, Vinik A and Freeman R 2003 The association between cardiovascular autonomic neuropathy and mortality in individuals with diabetes—A meta-analysis *Diabetes Care* **26** 1895–901
- McSharry P E and Clifford G D 2006 Models for ECG and RR interval processes *Advanced Methods and Tools for ECG Data Analysis* ed G D Clifford *et al* (Boston: Artech House) pp 101–33
- McSharry P E, Clifford G D, Tarassenko L and Smith L A 2003 A dynamical model for generating synthetic electrocardiogram signals *Biomed. Eng. IEEE Trans. On* **50** 289–94
- 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
- Mietus J E, Peng C K, Henry I, Goldsmith R L and Goldberger A L 2002 The pNNx files: re-examining a widely used heart rate variability measure *Heart* **88** 378–80
- Nunan D, Sandercock G R H and Brodie D A 2010 A quantitative systematic review of normal values for short-term heart rate variability in healthy adults *Pacing Clin. Electrophysiol. PACE* **33** 1407–17
- Pan J and Tompkins W J 1985 A real-time QRS detection algorithm *IEEE Trans. Biomed. Eng.* **32** 230–6
- Patel S, Park H, Bonato P, Chan L and Rodgers M 2012 A review of wearable sensors and systems with application in rehabilitation *J. Neuroeng. Rehabil.* **9** 21
- Pinna G D, Maestri R, Cesare A D and Colombo R 1994 The accuracy of power-spectrum analysis of heart-rate variability from annotated RR lists generated by Holter systems *Physiol. Meas.* **15** 163–79
- Rajendra Acharya U, Paul Joseph K, Kannathal N, Lim C M and Suri J S 2006 Heart rate variability: a review *Med. Biol. Eng. Comput.* **44** 1031–51
- Ramshur J 2014 HRVAS (<http://sourceforge.net/projects/hrvas/>)
- Riniolo T and Porges S W 1997 Inferential and descriptive influences on measures of respiratory sinus arrhythmia: sampling rate, R-wave trigger accuracy, and variance estimates *Psychophysiol.* **34** 613–21
- Riniolo T C and Porges S W 2000 Evaluating group distributional characteristics: why psychophysiologicalists should be interested in qualitative departures from the normal distribution *Psychophysiol.* **37** 21–8
- Ruxton G D 2006 The unequal variance *t*-test is an underused alternative to Student's *t*-test and the Mann–Whitney *U* test *Behav. Ecol.* **17** 688–90

- Sandercock G R H, Bromley P D and Brodie D A 2005 The reliability of short-term measurements of heart rate variability *Int. J. Cardiol.* **103** 238–47
- Schroeder E B, Whitsel E A, Evans G W, Prineas R J, Chambless L E and Heiss G 2004 Repeatability of heart rate variability measures *J. Electrocardiol.* **37** 163–72
- Serlin R C 2000 Testing for robustness in Monte Carlo studies *Psychol. Methods* **5** 230
- Singh J P, Larson M G, Tsuji H, Evans J C, O'Donnell C J and Levy D 1998 Reduced heart rate variability and new-onset hypertension insights into pathogenesis of hypertension: the Framingham heart study *Hypertension* **32** 293–7
- Smith A-L, Owen H and Reynolds K J 2013a Heart rate variability indices for very short-term (30 beat) analysis. Part 1: survey and toolbox *J. Clin. Monit. Comput.* **27** 569–76
- Smith A-L, Owen H and Reynolds K J 2013b Heart rate variability indices for very short-term (30 beat) analysis. Part 2: validation *J. Clin. Monit. Comput.* **27** 577–85
- Tak L M, Riese H, de Bock G H, Manoharan A, Kok I C and Rosmalen J G 2009 As good as it gets? A meta-analysis and systematic review of methodological quality of heart rate variability studies in functional somatic disorders *Biol. Psychol.* **82** 101–10
- Tarkoma S, Siekkinen M, Lagerspetz E and Xiao Y 2014 *Smartphone Energy Consumption: Modeling and Optimization* (Cambridge: Cambridge University Press) ISBN 978-1-107-04233-9
- Tarvainen M P and Niskanen J-P 2012 Kubios HRV version 2.1 user's guide ([http://kubios.uef.fi/media/Kubios\\_HRV\\_2.1\\_Users\\_Guide.pdf](http://kubios.uef.fi/media/Kubios_HRV_2.1_Users_Guide.pdf))
- Tarvainen M P, Niskanen J-P, Lipponen J A, Ranta-Aho P O and Karjalainen P A 2009 Kubios HRV: a software for advanced heart rate variability analysis *Proc. 4th European Conf. of the Int. Federation for Medical and Biological Engineering* pp 1022–5
- Tarvainen M P, Ranta-aho P O and Karjalainen P A 2002 An advanced detrending method with application to HRV analysis *IEEE Trans. Biomed. Eng.* **49** 172–5
- Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology 1996a Heart rate variability: standards of measurement, physiological interpretation, and clinical use *Circulation* **93** 1043–65
- Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology 1996b Heart rate variability: standards of measurement, physiological interpretation, and clinical use *Eur. Heart J.* **17** 354–81 (<http://eurheartj.oxfordjournals.org/content/17/3/354.long>)
- Thayer J F and Fischer J E 2009 Heart rate variability, overnight urinary norepinephrine and C-reactive protein: evidence for the cholinergic anti-inflammatory pathway in healthy human adults *J. Intern. Med.* **265** 439–47
- Thayer J F, Ahs F, Fredrikson M, Sollers J J III and Wager T D 2012 A meta-analysis of heart rate variability and neuroimaging studies: implications for heart rate variability as a marker of stress and health *Neurosci. Biobehav. Rev.* **36** 747–56
- Vanderlei L C M, Silva R A, Pastre C M, Azevedo F M and Godoy M F 2008 Comparison of the Polar S810i monitor and the ECG for the analysis of heart rate variability in the time and frequency domains *Braz. J. Med. Biol. Res.* **41** 854–9
- Vasey M W and Thayer J F 1987 The continuing problem of false positives in repeated measures ANOVA in psychophysiology: a multivariate solution *Psychophysiol.* **24** 479–86
- Vikman S, Mäkikallio T H, Yli-Mäyry S, Pikkujämsä S, Koivisto A-M, Reinikainen P, Airaksinen K J and Huikuri H V 1999 Altered complexity and correlation properties of RR interval dynamics before the spontaneous onset of paroxysmal atrial fibrillation *Circulation* **100** 2079–84
- Voss A, Schulz S, Schroeder R, Baumert M and Caminal P 2009 Methods derived from nonlinear dynamics for analysing heart rate variability *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.* **367** 277–96
- Ward S, Shouldice R B, Flanagan M and Heneghan C 2004 Electrocardiogram sampling frequency errors in PR interval spectral analysis *Proc. IEEE Postgraduate Conf. on Biomedical Engineering Medical Physics* (<http://ee.ucd.ie/~mark/papers/PGBIOMED.pdf>)
- Wilkinson L 1999 Statistical methods in psychology journals: Guidelines and explanations *Am. Psychol.* **54** 594
- Ziemssen T, Gasch J and Ruediger H 2008 Influence of ECG sampling frequency on spectral analysis of rr intervals and baroreflex sensitivity using the EUROBAVAR data set *J. Clin. Monit. Comput.* **22** 159–68