IOP Publishing Physiol. Meas. 39 (2018) 105004 (23pp) https://doi.org/10.1088/1361-6579/aae021

Physiological Measurement

IPEM Institute of Physics and Engineering in Medicine

RECEIVED
9 May 2018

REVISED
31 August 2018

ACCEPTED FOR PUBLICATION
10 September 2018

PUBLISHED
12 October 2018

PAPER

An open source benchmarked toolbox for cardiovascular waveform and interval analysis

Adriana N Vest[1,2], Giulia Da Poian[1], Qiao Li[1], Chengyu Liu[1], Shamim Nemati[1], Amit J Shah[2,3] and Gari D Clifford[1,4]

[1] Department of Biomedical Informatics, Emory University School of Medicine, Woodruff Memorial Research Bldg, Suite 4100, 101 Woodruff Circle, Atlanta, GA 30322, United States of America
[2] Department of Epidemiology, Rollins School of Public Health at Emory University, 1518 Clifton Road NE, Room 3053, Atlanta, GA 30322, United States of America
[3] Department of Medicine, Division of Cardiology, Emory University School of Medicine, 101 Woodruff Circle Suite, 319 Woodruff Memorial Research Building, Atlanta, GA 30322, United States of America
[4] Department of Biomedical Engineering, Georgia Institute of Technology, 1760 Haygood Drive, HSRB Suite W200, Atlanta, GA 30322, United States of America

E-mail: gari@physionet.org

Keywords: heart rate variability, toolbox validation, peak detection, physiological signal processing

Abstract

Objective: This work aims to validate a set of data processing methods for variability metrics, which hold promise as potential indicators for autonomic function, prediction of adverse cardiovascular outcomes, psychophysiological status, and general wellness. Although the investigation of heart rate variability (HRV) has been prevalent for several decades, the methods used for preprocessing, windowing, and choosing appropriate parameters lacks consensus among academic and clinical investigators. Moreover, many of the important steps are omitted from publications, preventing reproducibility. *Approach*: To address this, we have compiled a comprehensive and open-source modular toolbox for calculating HRV metrics and other related variability indices, on both raw cardiovascular time series and *RR* intervals. The software, known as the PhysioNet Cardiovascular Signal Toolbox, is implemented in the MATLAB programming language, with standard (open) input and output formats, and requires no external libraries. The functioning of our software is compared with other widely used and referenced HRV toolboxes to identify important differences. *Main results*: Our findings demonstrate how modest differences in the approach to HRV analysis can lead to divergent results, a factor that might have contributed to the lack of repeatability of studies and clinical applicability of HRV metrics. *Significance*: Existing HRV toolboxes do not include standardized preprocessing, signal quality indices (for noisy segment removal), and abnormal rhythm detection and are therefore likely to lead to significant errors in the presence of moderate to high noise or arrhythmias. We therefore describe the inclusion of validated tools to address these issues. We also make recommendations for default values and testing/reporting.

1. Introduction

Interest in heart rate variability (HRV) and signal processing of cardiovascular dynamics has seen a recent resurgence due to the increased availability of devices and wearables that record physiological signals. It has been widely reported that metrics which quantify cardiovascular dynamics can be used to estimate basal states and detect changes in the autonomic nervous system (Task Force of the European Society of Cardiology the North American Society of Pacing Electrophysiology 1996, Clifford 2002, Pan *et al* 2016) and consequently they hold promise as tools that can aid in disease tracking, wellness promotion, and risk stratification. The non-invasive nature of HRV measurement makes it particularly attractive as a long-term health tracking tool, or as a component of a more comprehensive health monitoring framework.

Despite its popularity in research and relatively long history, there is still much disagreement and ambiguity surrounding the methods employed by researchers to estimate HRV. In particular, peak (event) detection techniques vary (or detectors performances are not detailed), filter choices are variable and ad hoc, noise removal is generally undocumented, and detection of non-sinus beats is often poorly described. These issues limit meaningful comparisons between studies and scientific repeatability, especially when in-house, custom, non-public software are used. Unfortunately, few HRV software programs are rigorously designed and tested with methods that are clear and open to inspection. Additionally, of the open-source HRV programs available, many are poorly documented, no longer supported by their original authors, or have broken dependencies that require extensive troubleshooting. Regardless, no existing HRV software toolbox, to our knowledge, provides a comprehensive suite of validated tools encompassing both raw data and derived beat-to-beat intervals. More specifically, such software should undergo a validation process in which the output is rigorously compared with expected values based on a standardized input. Traditionally, the expected values for validation are either from a model or a gold standard algorithm. Since there is no gold standard in the HRV toolbox landscape and few models of HRV exist, a chicken-and-egg situation persists.

To address the issues of validation, standardization, and repeatability, we have developed an open-source cardiovascular signal and HRV analysis toolbox (The PhysioNet Cardiovascular Signal Toolbox) and have validated its component tools in two ways. First, for a subset of algorithms, where a model could be considered acceptable, artificial data has been used. Where no realistic and validated model is available, the toolbox is compared to other existing toolboxes to identify consistencies. (The assumption here is that the toolboxes, published by reputable authors, should converge to similar estimates.) The toolbox described here has been designed to accept a wide range of cardiovascular signals and analyze those signals with a variety of classic and modern signal processing methods. The toolbox was written in the MATLAB programming language and does not have any dependencies on external software or libraries.

Preprocessing and data cleaning is an important aspect of signal processing that often is overlooked or poorly documented in HRV-related publications. The PhysioNet Cardiovascular Signal Toolbox presented here employs several methods to prepare data for HRV estimation, including assessing signal quality and detecting arrhythmias, erroneous data, and noise. These segments of data, which must be excluded from HRV analysis, can then be systematically removed based on threshold settings selected by the user or recommended in previously validated studies.

The goal of this work was to advance the standardization, reproducibility, and clinical applicability of HRV and cardiovascular variability research. This publication has outlined the current HRV analysis tool landscape alongside our new suite of open-source tools contained within the PhysioNet Cardiovascular Signal Toolbox. We have presented the considerations necessary to invoke the use of these tools in a repeatable and standardized manner. The consequences of divergent approaches to HRV analysis are presented in a series of studies that systematically vary methodology and input data. Finally, a standard model by which HRV analysis packages may be judged in the future are presented along with a discussion of the recommendations by which HRV analysis should be conducted by researchers and clinicians alike.

2. HRV tool landscape

Publicly available tools for HRV analysis are scattered throughout the internet and have varying levels of sophistication. Here, we have reviewed a subset of the most popular toolboxes available and the HRV metrics that they generate. Perhaps the most used and trusted HRV toolbox is the PhysioNet HRV Toolkit, written by Mietus and Moody, available from PhysioNet.org (Mietus and Goldberger 2014). This toolbox is an open-source package that is written in C and performs time domain and spectral HRV statistics. It has the unique feature of compatibility with PhysioNet's Waveform Database (WFDB) Software Library (also written in C). This allows the user to leverage PhysioNet's many QRS detectors, data libraries, and processing and evaluation tools. However, installation is nontrivial and the default preprocessing and other variables associated with it are not well documented. Nevertheless, it is considered the standard in the field. The proprietary Kubios HRV software (Tarvainen *et al* 2014) is another frequently used and cited HRV analysis tool. At the time of this publication, Kubios is available in both a no-cost 'Standard' version and a licensed 'Premium' version available for $329 per seat license. Both versions of the Kubios software offer an extensive user interface and the ability to process *RR* intervals. As with the PhysioNet HRV Toolkit, the Premium version can also process ECG waveform data and perform a Lomb-Scargle periodogram (Lomb 1976, Press *et al* 1992), both of which are essential functions, as we explain in this article. Running the Kubios HRV software is strictly through a proprietary user interface which is susceptible to human 'tweaking' of data and tools, and does not support batch processing of input data. This can make analysis time consuming for moderate sized datasets and unfeasible for large datasets and it introduces additional opportunities for human error. Two less commonly referenced MATLAB-based toolboxes available are Kaplan's HRV toolbox (Kaplan and Staffin 1998) and Vollmer's HRV toolbox (Vollmer 2018). Both these

toolboxes are open-source and were written for MATLAB. Additionally, Vollmer's HRV toolbox employs a user interface, but does not require it.

All of the aforementioned HRV toolboxes, including the PhysioNet Cardiovascular Signal Toolbox described in this publication, compute classic HRV metrics including the mean of normal-to-normal (*NN*) intervals, the standard deviation of *NN* intervals (SDNN), the square root of the mean squared differences of successive *NN* intervals (RMSSD), the proportion of interval differences of successive *NN* intervals greater than 50 ms (pNN50), or more generally pNNx (where *x* is a variable between five and 100 ms), the total power of the power spectral density across various frequency bands, and the ratio of low frequency to high frequency power. Additional HRV metrics are available in the various toolboxes per table 1. See Clifford *et al* (2006) for a detailed description of these statistics.

It is worth noting that the PhysioNet Cardiovascular Signal Toolbox includes both the main established HRV metrics (e.g. LF/HF-ratio, pNNx, RMSSD, etc) and more recent HRV metrics which have been shown to be highly promising (e.g. multiscale entropy and phase rectified signal averaging) (Norris *et al* 2008a, 2008b, Lin *et al* 2014, Kantelhardt *et al* 2007, Campana *et al* 2010, Kisohara *et al* 2013, Lobmaier *et al* 2015). The toolbox also includes versions of existing metrics which have been shown to be more computationally efficient than available versions (e.g. entropy metrics), an important feature when using an interpreted language, and more accurate (e.g. the version of detrended fluctuation analysis provided). We have chosen to omit the less well-founded/more ad hoc statistics, such as the TINN (which is just a poor estimate of a moment of a distribution) in favor of more acceptable and more descriptive statistics (such as the first four moments of a distribution).

3. PhysioNet cardiovascular signal toolbox design

3.1. Overview

The PhysioNet Cardiovascular Signal Toolbox developed by the authors utilizes a standardized approach to preprocess data and compute HRV metrics that provides a unique and comprehensive approach:

1. An initialization file (*InitializeHRVparams.m*) sets up global variables that deal with thresholds, window settings, noise limits, and spectral analysis limits (listed in appendix C). The default parameters are as used in this article. However, we strongly recommend the user consult an expert to identify a reasonable choice of parameters for their population. For example, children or smaller animals will require significantly different thresholds on almost all parameters.
2. Data identification and formatting is then the next step. The toolbox does not assume any format of data, except that the *RR* interval data are two equal length vectors (time and *RR* interval in units of seconds). Additionally, the 'raw' ECG, blood pressure waveform and photoplethysmographic/pulsatile data should be in the standard physical units (mV, mmHg or normalized units respectively). Note that we have included native support for loading WFDB-compliant annotation files (often denoted by an '*atr*' file extension on PhysioNet). However, we have deliberately dissociated the toolbox from any library dependencies outside of the required MATLAB toolboxes (listed in appendix A).
3. If raw waveforms are to be analyzed, the QRS complex or pulsatile beat onsets must be detected first using one of the in-built beat detectors (*jqrs.m*, *wabp.m*, *qppg*). We do supply other ECG beat detectors such as *sqrs* and *wqrs* for benchmarking and signal quality analysis, but we do not suggest using the results derived from their use unless the input data is perfectly clean.
4. Subsequently, the signal quality of the raw waveform data (either windowed or beat-by-beat) must be evaluated. A signal quality index (SQI) is calculated on a rolling window (Default is 10 s, with one s increment, *HRVparam.sqi.windowlength* = 10 and *HRVparams.sqi.increment* = 1) for the duration of the ECG waveform using *bsqi.m*, or on a beat by beat basis for blood pressure and pulsatile data using *jsqi.m* and *PPG_SQI_buf.m* respectively. HRV analysis should not be performed on noisy data (data that drops below some predefined threshold–*HRVparam*.sqi.LowQualityThreshold = 0.9 by default) that leads to false positive beat detections.
5. If desired, ventricular fibrillation/ventricular tachycardia (VF/VT) can be detected on the waveform based on the method discussed in section 3.2.3.
6. The time series is next converted to *RR* intervals by taking the consecutive differences of the beat locations in contiguous data (where segments have not been removed). If the user desires to use *RR* interval data instead of the raw waveforms, the *RR* interval time series can be loaded into the HRV Toolbox directly, although signal quality and VF detection cannot then be performed.
7. Before calculating HRV statistics, arrhythmic and noisy periods of data must be removed. Once the time series is in interval form, atrial fibrillation classification and ectopy (premature ventricular contraction (PVC)) can be performed on the *RR* interval time series. Any data that is deemed undesirable for HRV

Table 1. Summary of functionalities of various HRV toolboxes. See section 3 for definition of HRV metrics.

Software origin → ↓Functionality	PhysioNet HRV toolkit (v 10.5.24) (last update 4 August 2009)	PhysioNet Cardiovascular Signal Toolbox (v 1.0)	Kubios (v 3.0.2) (premium)	Kaplan (last update 3 February 1998)	Vollmer (v 0.98 last update 4 October 2017)
Data formats accepted	Intervals or waveforms	Intervals or waveform	Intervals or waveform	Intervals	Intervals or waveform
Dependencies	WFDB Libs (C Version)	None	None	None	WFDB Libs (MATLAB Version)
Waveforms analyzed	ECG, ABP	ECG, ABP, PPG	ECG	None	ECG
Can operate independent from GUI	Yes	Yes	No	Yes	Yes
Open-source	Yes	Yes	No	Yes	Yes
Preprocessing	Intervals that vary by more than 20% of the average interval measured on 20 beats before and after the current interval are removed	Intervals that vary more than 20% from the median interval measured over five intervals before and after the current interval and nonphysiological intervals are removed (default $RR > 2$ s or $RR < 0.375$ s)	Proprietary and unknown	Statistical outlier removal and spline interpolation	Filter function available but not integrated in HRV metric calculations
Simulator	None	*rrgen.c*	None	*makerr.m*	None
HRV metrics	Mean NN, SDNN, pNN50, pNNxx, RMSSD, ULF, VLF, LF, HF, LF/HF, Total Power, SDANN, SDNNI, MSE, DFA	Mean NN, SDNN, pNN50, pNNxx, RMSSD, skewness, variance, ULF, VLF, LF, HF, LF/HF, total power, SDANN, SDNNI, SD1, SD2, SD2/SD1, DFA, ApEn, SampEn, MSE, PRSA (AC and DC), HRT	Mean NN, SDNN, pNN50, pNNxx, RMSSD, VLF, LF, HF, LF/HF, total power, SDANN, SDNNI, SD1, SD2, SD2/SD1, DFA, ApEn, SampEn, MSE, triangular index, TINN, peak frequency, ECG derived respiration, recurrence plot analysis	Mean NN, SDNN, pNN50, pNNxx, RMSSD, VLF, LF, HF, LF/HF, total power, SDANN, SD1, SD2, SD2/SD1, DFA, ApEn	Mean HR, SDNN, pNN50, pNNxx, RMSSD, VLF, LF, HF, LF/HF, total power, SD1, SD2,SD2/SD1, DFA, ApEn, triangular index, TINN, StDev of successive differences (SDSD), correlation dimension (CD), euclidean distance based on relative *RR* intervals

analysis (arrhythmia, low SQI, ectopy, artefact, noise) is excluded from analysis and HRV metrics are calculated on the remaining data (*NN* intervals).

A high-quality analysis of HRV starts with a thoughtful selection of data and input parameters. The length of the data source, the appropriateness of the method and extent of preprocessing, and the metrics to be generated all must be considered before, during, and after analysis. Poor choice of analysis parameters can result in the generation of erroneous results that are representative of noise instead of physiology. The following sections address the most common considerations of any HRV analysis. For a more detailed overview of the signal processing issues related to HRV, we refer the reader to Clifford *et al* (2006).

A set of demonstration files (listed in appendix D) are made available to the user for testing the toolbox and verifying the correct 'installation' of required MATLAB packages.

3.2. Waveform preprocessing routines

3.2.1. Peak detection

The toolbox can accept electrocardiogram (ECG), blood pressure (ABP), and photoplethysmogram (PPG) data and has validated beat detectors for each of these signals. The available beat detectors for ECG include MATLAB versions of the PhysioNet tools *sqrs.c* (Engelse and Zeelenberg 1979, Moody 2015b), *wqrs.c* (Zong *et al* 2003), and *jqrs* (Behar *et al* 2014, Johnson *et al* 2014). The performance of these peak detectors has been shown to be comparable to previously published detectors *wqrs.c* (Zong *et al* 2003), *sqrs.c* (Moody 2015a), and *gqrs.c* (Moody 2015a), available with the WFDB software package. (The data from the performance comparison is included in appendix B for convenience (Vest *et al* 2017).) Interested readers can learn more about how each detector functions from their respective citations.

The MATLAB version of *wabp.c*, *wabp.m*, is used for pulse detection on ABP waveforms (Sun 2006). This program detects the onset of each beat in the ABP signal using the slope sum function which amplifies the rising edge of the waveform. The same algorithm was also adapted and optimized to be used on PPG waveforms, establishing *qppg.m* as the toolbox's PPG peak detector.

3.2.2. SQI

To determine if the data is of high enough quality to analyze, a quantitative and objective signal quality measurement should be employed. The toolbox uses *bsqi* (Li *et al* 2008) for ECG, *jSQI* (Sun *et al* 2005, Sun 2006, Johnson *et al* 2015) for ABP, and *PPG_SQI_buf.m* (Li and Clifford 2012) for PPG. Published by Li *et al* (2008), *bsqi* provides the percentage of beats that match when detected by multiple annotation generators with highly differing noise responses. The SQI is typically given as a percentage or normalized value, and a threshold below which data is removed should be chosen (or rather optimized) and reported. *jSQI* measures the quality of the ABP waveform on a beat by beat basis, returning a binary signal quality assessment based on a set of measured features on the ABP pulse, including onset time and pressure values. *psqi* also measures quality of the PPG waveform on a beat by beat basis based on beat template correlation. After determining the fit of the current beat to the template, the beat is assigned an assessment of excellent ('E'), acceptable ('A'), or unacceptable ('Q').

3.2.3. VF/VT classification

Ventricular tachycardia/fibrillation detection is performed using a state-of-the-art method published by Li *et al* (2014), *VF_Classification.m*. In the published method, a support vector machine (SVM) model was trained on three annotated public domain ECG databases (the American Heart Association Database, the Creighton University Ventricular Tachyarrhythmia Database, and the MIT-BIH Malignant Ventricular Arrhythmia Database) and 14 different VF features. After training, the model was optimized for use of only two features on 5 s windows.

3.2.4. PVC classification

Premature ventricular contraction (PVC or VPC) detection is essential to HRV analysis, although PVC detection is not provided in any of the current open source toolboxes. In our toolbox we provide a new software package for this which is based on the application of a convolutional neural network (CNN) to the wavelet transform (WT) of the raw ECG (Li *et al* 2018). The WT is used to map short segments of a single channel (1D) ECG waveform into a 2D time-frequency 'image'. The images are then passed into the CNN to optimize convolutional filters to improve classification. Using ten-fold cross validation, an overall F1 score of 84.94% and an accuracy of 97.96% was achieved on the MIT BIH Arrhythmia Database. The American Heart Association ECG Database (AHA 2018) was then used as an out-of-sample validation database. Without retraining, the PVC detector achieved an F1 score of 84.94% and an accuracy of 97.33% on this second database. We note that the identification of ectopic beats (as opposed to noise identification or other abnormal beats) is needed for not only for abnormal *RR* interval removal but for the evaluation of heart rate turbulence, for which it is important not to confuse noise with ectopy. Once an ectopic beat is identified, the researcher has the option to insert a 'phantom' beat or remove the *RR* intervals corresponding to the ectopic beat (both the preceding and following *RR* interval).

3.2.5. AF classification

Atrial fibrillation (AF) is detected on the *RR* interval time series using the method published by Oster *et al* (Oster and Clifford 2015). The method uses a support vector machine (SVM) trained on features from the *RR* interval time series which reflect the unpredictability of the heartbeat. The classifier has been shown to produce an AUC of 96.76% on windows containing 60 beats, 95.27% on windows containing 30 beats, and 92.72% on windows containing 12 beats (Liu *et al* 2018, Li *et al* 2016). We recommend 30 s windows with a 10 s overlap to minimize the amount of data removed, and a bias of the data away from high variability.

3.3. *RR* interval preprocessing routines

3.3.1. Non-sinus beat identification and removal/replacement

Additional preprocessing steps are taken to address noise and artefact that occur at a scale smaller than the signal quality index window or in data that has already been translated into *RR* intervals. Since HRV metrics are meant to measure the activity of the sinoatrial node, all intervals associated with non-sinus beats must be removed. Outside of beat classification in the ECG, a notoriously difficult issue which is highly error prone or impossible in non-ECG or noisy ambulatory conditions, non-sinus beats can be identified with reasonable certainty using statistics of the *RR* interval time series itself.

In the absence of waveform data, we may identify non-sinus *RR* intervals as those that occur prematurely or late. The most common method to identify such intervals (and the method employed in this work) involves measuring changes in the current *RR* interval from the previous *RR* interval or an average of the last *N* intervals

and excluding intervals that change by more than a certain percentage. In this toolbox (and the work presented here) we chose N to be a default value of five (five beats before and five beats after the current interval) and the standard threshold of 20%. We note however, that a threshold of 15% balances the need to remove aberrant data with the desire to keep sinus beats and has shown to exclude at least 80% of ectopic beats and 93% of the noise-induced (extra beat) detections at the expense of 2% sinus beats in the MIT NSR database (Clifford 2002, Clifford *et al* 2002). If the non-sinus beats are infrequent, the PhysioNet Cardiovascular Signal Toolbox has the ability to perform interpolation to add a beat where a sinus beat would have been expected to occur. The term 'interpolation' is usually referred to the process by which the unevenly sampled *RR* interval data is resampled to an evenly sampled time series, usually prior to the use of the FFT. In this article, we follow Clifford *et al* (Clifford 2002, Clifford *et al* 2006) and use resampling to refer to the conversion to an evenly sampled time series (see section 3.3.3).

Additional checks and corrections include flagging and removing non-physiologic data (*RR* intervals above 2 s or below 0.375 s, outside of physiologically possible range) and data that is labeled as non-normal per a supplied annotation file (if applicable).

3.3.2. Manual correction

The PhysioNet Cardiovascular Signal Toolbox does not enable manual correction of annotations or R peak locations. Although automated peak detectors do not always accurately classify the location of QRS complexes, manual correction of the location is a subjective procedure at best and inter-reader variability is a well-documented phenomenon that contributes to the inability to reproduce results amongst studies. Statistics on inter-reader variability have been measured to be greater than 20% (Sparrow *et al* 1988, Pinedo *et al* 2010, Zhu *et al* 2014). We explicitly advise against 'expert' or 'hand' modification of data, since it invalidates scientific repeatability of the research.

The question of whether erroneous detections cause significant changes in specific HRV estimates has been addressed previously (Clifford and Tarassenko 2005), but whether this affects a final downstream classifier is another issue. The only real way to know is to stress test the classifier or predictor under varying levels of noise. The toolbox provides robust and repeatable methods for dealing with noise, providing users with a level of trust in the output. Automatic methods for dealing with erroneous detections and identification of unreliable segments of data are incorporated in the pre-processing tools and signal quality index stage of our toolbox (see section 3.2.2).

3.3.3. Resampling

Resampling the *RR* interval time series involves interpolating through the signal (such as by linear or cubic spline interpolation) and resampling at regular intervals specified by the resampling frequency. Most of the papers in the field of HRV report on the use of resampling rates between 1 Hz and 10 Hz (Malik and Camm 1995, Hilton *et al* 1998, Task Force of the European Society of Cardiology the North American Society of Pacing Electrophysiology 1996). Since the human heart rate can sometimes exceed 3 Hz (180 bpm), then a sample rate of at least 6 Hz may be required to satisfy the Nyquist criterion. However, if one knows that the *RR* tachogram is unlikely to exceed 120 bpm then a resampling rate of 4 Hz is sufficient. Resampling introduces an implicit assumption about the form of the underlying variation in the *RR* tachogram; for example, cubic spline techniques assume that the variation between beats can be modelled accurately by a cubic polynomial.

3.3.4. Thresholding on data loss

A threshold can be applied for how much data can be thrown out before a segment is rendered unusable, but this of course depends on the analysis being performed. Mølgaard *et al* (Mølgaard 1991) demonstrate how certain time series metrics (such as RMSSD) are extremely sensitive to missed beats especially in patients with reduced HRV and therefore it is extremely important to consider whether the data in such cases should be used at all. There is much variation in how researchers address the issue of removed beats or missing data (due to noise, missed detections, etc). The calculation of time domain metrics may withstand large losses of data, but the results will vary based on the length of the segment analyzed.

3.4. Frequency domain analysis

3.4.1. Power spectral density estimation

For frequency domain calculations, the power spectral density (PSD) of the *NN* interval time series can be generated using several methods, including: the Lomb periodogram, the Welch PSD estimate, the Burg PSD estimate, and the discrete fast Fourier transform (FFT). FFT- or wavelet-based PSD estimates require resampling to an evenly sampled time series, and cubic spline interpolation is often preferred to linear interpolation because the latter increases LF power (due to flattening) and HF power (due to sharp edges at each beat). Resampling functionality is provided to users in the toolbox. Error in the PSD estimate and frequency domain metrics grows linearly with the amount of data removed. Previous studies have shown that losses of data up to 20%

will not significantly alter results generated with the Lomb periodogram, as long as the data are not missing in concentrated clusters (Clifford 2002). Moreover, Clifford and Tarassenko (Clifford and Tarassenko 2005) showed that although phantom beat insertion does provide marginal improvements for FFT-based metrics, using more appropriate techniques that can handle unevenly sampled time series (such as the Lomb periodogram (Lomb 1976, Scargle 1982, Press *et al* 1992)) are far superior. We therefore do not recommend the use of interpolation, phantom beat insertion, or techniques that require evenly sampled time series such as the FFT and wavelet analysis. Thus, we use the Lomb periodogram as the default method for frequency analysis.After the PSD is calculated, various frequency domain HRV metrics are calculated. The sum of power in the various frequency bands is calculated as is the total power in the spectrum. These spectral metrics can be normalized to the variance of the *NN* interval time series, or to another measure. As stated above, many researchers normalize the sum of the power spectral density plot to variance because of the mathematical equivalency of the two. The choice of normalization is up to the user, but explicitly specified in the set-up of the analysis. All PSD estimates calculated by the HRV Toolbox described here can accept frequency bin size specification, which improves control over the reproducibility of the resulting analysis.

We note that some researchers work in the 'beatquency' domain in order to avoid resampling issues. However, missing data due to poor QRS detection or data excision due to noise disrupts this sequence and leads to false peaks in the spectra. Additionally, the axes are then a function of the data itself and causality/stability of the metric becomes an issue. We note that it is unclear whether several ventricular beats could be replaced by estimates of sinus beats without causing significant issues, but in reality, the baroreflex response due to ectopy (which is exploited by heart rate turbulence measures) creates a nonstationarity in the time series. Therefore, any analysis using methods that assume stationarity should be truncated at such a point and restarted after the discontinuity.

In summary, if the incidence of artifact is high within a given segment then it is preferable to eliminate the segments from the analysis. If the incidence of artifact is low, removal of the artefact without replacement is recommended (Clifford and Tarassenko 2005). The exact regions of data removed and percentage of removed or missing data should be reported.

3.4.2. *Frequency bands for spectral content estimation*

The frequency bands of interest for analyzing HRV are generally defined as

ULF—Ultra Low Frequency:	$0.0001\ \text{Hz} \leqslant \text{ULF} < 0.003\ \text{Hz}$
VLF—Very Low Frequency:	$0.003\ \text{Hz} \leqslant \text{VLF} < 0.04\ \text{Hz}$
LF—Low Frequency:	$0.04\ \text{Hz} \leqslant \text{LF} < 0.15\ \text{Hz}$
HF—High Frequency:	$0.15\ \text{Hz} \leqslant \text{HF} < 0.4\ \text{Hz}$

The frequency bands are thought to capture different physiological mechanisms, but the bands can be redefined and do not perfectly map to a particular physiological process (Cerutti *et al* 1995). The bands can also shift lower in the case of a very fit clinical study population with lower baseline heart rates, or higher in the case of a pediatric or adolescent clinical study population with higher baseline heartrates. It is generally accepted in the clinical community that the HF band is mostly a measure of the parasympathetic activity (Cerutti *et al* 1995) with some sympathetic activity, while the LF band contains mostly sympathetic activation (Eckberg 1997). Researchers may want to measure the power in the HF and LF frequency bands as a measure of sympathovagal balance. The LF/HF ratio is used often and simplifies the units of the measurement (i.e. it is unitless). However, we note that this ratio can change depending on whether the power is estimated in the logarithmic domain or not. The PhysioNet Cardiovascular Signal Toolbox defaults to normal domain and not logarithmic domain.

3.4.3. *Normalization method*

Common normalization factors used for HRV metrics include the length of the data segment analyzed and the variance of the *NN* interval data. Variance is mathematically equal to total power of the *NN* interval time series, so many researchers normalize the total power by dividing by variance. No matter the normalization method, it is important that the chosen method is reported because it can contribute to inter-study differences.

3.5. Length of data

The user needs to decide if a long-term (i.e. ~24 h or longer) or short-term (i.e. ~5 min) recording is desired. (This can be defined by modifying the *HRVparams.windowlength* and *HRVparams.increment* parameters in the initialization file.) However, certain considerations and limits should be kept in mind. The choice depends on the research being performed and the availability and quality of data. Long-term recordings capture circadian rhythm variations that have been valued for diagnostic value (Task Force of the European Society of Cardiology the North American Society of Pacing Electrophysiology 1996) and short term metrics have been shown to be capable of assessing neurological activity (Malik and Camm 1995, Task Force of the European Society of

Cardiology the North American Society of Pacing Electrophysiology 1996). Confounders for long-term HRV metrics can include temperature (Malik and Camm 1995), quality of sleep (Cooper *et al* 2000), and large gaps in data (Clifford 2002). Moreover, short-term HRV can be influenced by changes in mental, emotional, or physical state (Bernardi *et al* 2000). Both long- and short-term recordings can suffer when data quality is low and only a fraction of the recording is useable, but to different extents. Care should be taken to control for these confounders when possible, and to assess their influence on the results when not.

The length of data analyzed has implications on the appropriateness of the HRV metrics being employed. In order to choose the best window size for the given analysis, the researcher must balance the requirement of stationarity (if required) versus the time required to resolve the information present. For most time domain HRV statistics, previous researchers have recommended long-term recordings. Haaksma *et al*'s (1998) study led to recommendations of 20 h of data be collected to estimate time domain variables or for total power (calculated between 0.0001 Hz and 0.4 Hz) calculations (Haaksma *et al* 1998). The Task Force on standards in HRV (Task Force of the European Society of Cardiology the North American Society of Pacing Electrophysiology 1996) recommends applying frequency domain methods to recordings at least 10 times the inverse of the lower frequency bound of the investigated component, but no longer. This is to ensure stability of the signal. During a short-term period, the data can be considered to be stationary or quasi-stationary and is therefore amenable to estimation of the power spectral density (PSD). However, it is unlikely that the *RR* interval time series remains stationary for more than a few minutes, and this makes the above recommendation rather impractical.

As an example, if the research is to determine if the *RR* interval time series contains a 0.01 Hz oscillation, at least 100 s of data (the length of one period of a 0.01 Hz oscillating signal) is necessary, although in practice 300 s or more are needed. The European and North American Task Force on standards in HRV (Task Force of the European Society of Cardiology the North American Society of Pacing Electrophysiology 1996) suggested that the shortest time period over which HRV metrics should be assessed is 5 min. This results in a limitation of the lowest frequency that can be resolved being $1/300 \approx 0.003$ Hz (just above the lower limit of the VLF region). In practice the limit is higher since noise affects the estimation. A 5 min segment can therefore only be used to evaluate higher frequency bands, i.e. LF and HF. The upper frequency limit of the highest band for HRV analysis is generally quoted as being 0.4 Hz (Malik and Camm 1995), but in reality, frequencies can be estimated (only) up to the reciprocal of twice the shortest *RR* interval. In general, we quote the average Nyquist frequency as $f_N = \frac{1}{2\Delta t_{av}} = \frac{N}{2T}$ where Δt_{av} is the mean *RR* interval, T is the length of the window in seconds and N the number of *RR* intervals in the window. Thus, a 5 min window ($T = 300$) leads to the constraint of $N/2\ T \geqslant 0.4$ Hz on the number of points and hence to a lower limit on N of 240 beats (an average lower heart rate limit of 48 bpm if all beats in a 5 min segment are used) (Clifford 2002, Clifford *et al* 2006).

Finally, it should be noted that metrics should only be compared between subjects when the data lengths are the same (Clifford 2002) and they cover the same period of the circadian cycle (Clifford and Tarassenko 2004, Clifford *et al* 2006). The latter is particularly important, because diurnal or momentary changes in activity, both psychophysical (e.g. after lunch, exercise or a stressful event like driving) and consciousness-related (such as sleep) can be one of the most dominant factors confounding any HRV comparison.

3.6. Long range scaling metrics: DFA and MSE

3.6.1. Detrended fluctuation analysis

Detrended fluctuation analysis (DFA) is included as a part of this toolbox as a method for quantifying long-term self-similarities in *RR* interval time series (Peng *et al* 1995). Such self-similarity can be described as a $1/f^{\beta}$ scaling in the log–log power-frequency spectrum, where the β is the slope of this spectrum. An alternative method used to compute the fractal scaling exponent, $\alpha = (\beta + 1)/2$, is by using the DFA, which is briefly summarized in the following paragraph. For a detailed description see Peng *et al* (1995).

Given a time series $x(n)$, the first step of DFA consists of integrating the original time series in order to obtain a self-similar process $y(k)$, $y(k) = \sum_{i=1}^{k} (x(i) - \bar{x})$, where $\bar{x}$ is the mean of x. The next step consists of dividing the integrated time series into boxes of equal length m and for each box performing a least squares line fit to the data. The time series is then detrended by subtracting the local trend $y_n(k)$ in each box. At this point, for a given box size m, the characteristic size of the fluctuation $F(m)$ for this integrated and detrended time series is calculated by

$$F(m) = \sqrt{\frac{1}{N}\sum_{k=1}^{N}[y(k) - y_m(k)]^2}.$$

The procedure is repeated over different time scales (box sizes) to provide a relationship between $F(m)$ and the box size m.

The code for DFA included in the PhysioNet Cardiovascular Signal Toolbox (i.e. *dfaScalingExponent.m*), provided by McSharry (McSharry and Malamud 2005), has been integrated into the toolbox with no significant

modification. New features introduced in this version include an option for the user to change the minimum and maximum box sizes and a *midBoxSize* parameter for the optional computation of scaling exponents α_1 and α_2. (Default parameters in the code mirror *dfa.c* and are set to: *minBoxSize* = 4; *maxBoxSize* = $L/4$, where L is the length of the input series; and *midBoxSize* = 16.) (Moody 2015a) The scaling exponent, α_1 reflects power related to short-term fluctuations (LF and HF) and α_2 reflects power related to long-term fluctuations (VLF and ULF) (Willson *et al* 2002).

3.6.2. Multiscale entropy

Multiscale entropy (MSE) analysis was first introduced by Costa *et al* (2002, 2005) as a method for analyzing the dynamic complexity of a system by quantifying its entropy over a range of temporal scales. Traditional methods use entropy-based algorithms to quantify the degree of regularity of a time series. However, there is no straightforward correspondence between regularity and complexity. MSE relies on sample entropy (SampEn) (Richman and Moorman 2000), which quantifies the likelihood that two sequences similar for m points remain similar at the next point (i.e. match within a tolerance of r), not taking into account self-matches. This metric is included in the PhysioNet WFDB libraries and therefore is provided in our toolbox.

MSE can be summarized as a two-step procedure. The first step consists of generating a coarse-grained time series by averaging the data points of the original time series $x(n)$ within non-overlapping windows of increasing length, τ. For scale one, the coarse-grained time series $y(1)$ corresponds to the original signal. The length of the coarse-grained time series is M/τ, where M is the length of $x(n)$. The second step consists of computing the sample entropy on each coarse-grained time series.

All the parameters used for MSE analysis can be changed in the *InitializeHRVparametrs.m* file (Default settings include the following: *RadiusOfSimilarity* = 0.15 (r), *patternLength* = 2 (m), *maxCoarseGrainings* = 20 ($\max \tau$)).

Two implementations of the SampEn algorithms are provided, a normal speed and a fast speed. The fast speed version is an implementation of the traditional SampEn (*FastSampEn.m*) which provides equivalent results. Currently the program switches automatically to *FastSampEn.m* when the size of the time series is less than 34 000 points. This default was chosen based on the memory required for MATLAB R2017a running on an Intel Core i7 processor equipped with 16 GB memory to execute the function. The user can modify this parameter in the function *ComputeMultiscaleEntropy.m*.

3.7. Phase-rectified signal averaging

Phase-rectified signal averaging (PRSA) is a method used for identifying short-term quasi-periodicities that are normally masked by non-stationarities and provide information on the deceleration (DC) and acceleration (AC) capacity of the heart (Bauer *et al* 2006). The code made available in the PhysioNet Cardiovascular Signal Toolbox implements the simplest version of the PRSA algorithm, where the anchor points correspond to increases in the signal (or decreases): $x_i > x_{i-1} (x_i < x_{i-1})$. In order to avoid anchor points at the positions of artifacts, a threshold parameter ensures that increases or decreases larger than such a threshold are discarded (Default = *HRVparams.prsa.thresh_per* = 20%; as suggested in Campana *et al* (2010)). The length (L) of the PRSA signal before and after the anchor points can be changed in the initialization file and should exceed the period of the slowest oscillation that is of interest (Default = *HRVparams.prsa.win_length* = 30). Wavelet analysis using Haar mother wavelet function is employed to derive the AC or DC from the central part of the PRSA signal (with scale parameter s defined by *HRVparams.prsa.scale* = 2 by default):

$$AC\,(DC) = \sum_{i=1}^{s} \frac{prsa\,(L+i)}{2s} - \sum_{i=1}^{s} \frac{prsa\,(L-i)}{2s}.$$

For a more detailed description of the algorithm we refer the reader to Bauer *et al* (2006).

3.8. Heart rate turbulence

Heart rate turbulence (HRT) is a method used to analyze the fluctuations in sinus-rhythm cycle length after PVCs (Schmidt *et al* 1999, Bauer *et al* 2008). Two parameters are used to characterize the response of sinus rhythm to a PVC: the turbulence onset (TO) and turbulence slope (TS). TO is used as a measure of the initial acceleration after the PVC, and it is derived by comparing the relative changes of *NN* intervals immediately after and before a PVC:

$$TO = 100 * \frac{(NN_{+2} + NN_{+1}) - (NN_{-1} + NN_{-2})}{(NN_{-1} + NN_{-2})},$$

where *NN*+ is the *i*th sinus rhythm after the compensatory pause of the PVC and *RR*-indicates the coupling interval of the PVC. The TO value is first computed for each single PVC (figure 1) and subsequently averaged to

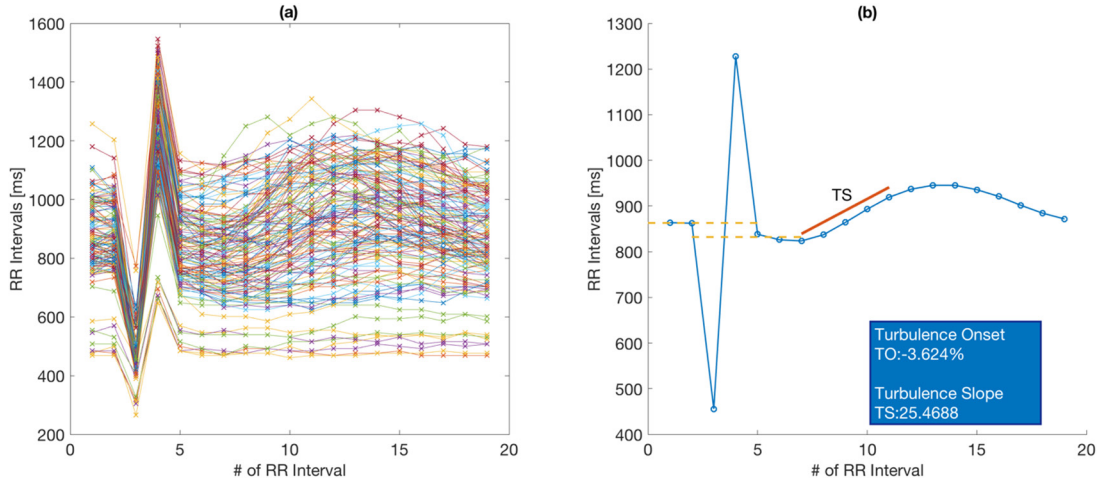

Figure 1. Visualization of HRT analysis using the PhysioNet Cardiovascular Signal Toolbox. Left (a) figure shows an example of 145 *RR* interval sequences (of 20 beats) taken from one patient, all aligned at the PVC (at *RR* interval three); the TO is computed for each single PVC and subsequently averaged to obtain the value characterizing the subject. Right (b) figure shows the average tachogram used to compute the TS. TS is the maximum of regression slopes (solid red line) computed for five consecutive *NN* intervals starting at *RR* interval seven (3rd *NN* interval after the compensatory pause). In the example, the regression lines for *RR* intervals seven through eleven corresponding to maximum slope is shown (red line).

obtain the value characterizing the patient (Bauer and Schmidt 2003). TO can be also calculated on the averaged tachogram which leads to very similar values (Bauer *et al* 2008).

The second measure, the TS, quantifies the deceleration rate after a PVC. TS is the maximum positive slope of a regression line assessed over any sequence of five subsequent sinus rhythm *NN* intervals within the first 20 sinus rhythm *NN* intervals after a PVC (Bauer *et al* 2008). In figure 1, the red line corresponds to the line of regression fit to the five consecutive *NN* intervals that result in the largest gradient.

TO values below zero and TS values above 2.5 are considered normal, and abnormal otherwise (i.e. a healthy response to PVCs is a strong sinus acceleration followed by a rapid deceleration (Clifford *et al* 2006). Because the HRT pattern might be masked by heart rate variability (HRV) of other origins, the TS is computed on the PVC tachogram, obtained by aligning and averaging the *R–R* interval sequences surrounding isolated PVCs, for a sufficient number of PVCs (i.e. >5) (Bauer *et al* 2008). Despite this accumulation of data around numerous PVCs, performing HRT analysis on very short ECG recordings may not lead to meaningful results (Berkowitsch *et al* 2004). It is important to ensure that the sinus rhythm preceding and following a PVC is free of arrhythmia, artifacts, and false beat classification due to artifact. Thus, a set of exclusion criteria was implemented according to Clifford *et al* (2006):

- Remove all *RR* intervals <300 ms or >2000 ms.
- Remove all RR_n where $|RR_{n-1} - RR_n| > 200$ ms.
- Remove all *RR* intervals that change by more than >20% with respect to the mean of the five previous sinus intervals (the reference interval) (Alternative: *RR* intervals that change by more than >20% with respect to the previous one).
- Only use PVCs with a minimum prematurity of 20%.
- Exclude extrasystolic pauses greater than 20% longer than the normal interval.

The function *HRT_Analysis.m* computes the TO and TS value given a time series of *RR* intervals and related labels (annotations) following the PhysioNet standard[6], the number of *NN* intervals to consider before the PVC (*BeatsBefore*), and after the PVC plus a compensatory pause (*BeatsAfter*). The function also returns the number and position of the PVCs used for the analysis, the average tachogram, and the graphical representations of the HRT analysis shown in figure 1.When computing the average tachogram or the mean TO, the user should aim to include a minimum of 15–20 tachograms containing a single PVC.

4. Methods

In order to elucidate the consequences of divergent approaches to HRV analysis, a series of benchmarking studies were performed that systematically vary methodology and input data. The studies were conducted on sample data using the PhysioNet Cardiovascular Signal Toolbox and the four other HRV toolboxes described in table 1.

[6] www.physionet.org/physiobank/annotations.shtml

Table 2. Differences between HRV analysis methods in the five HRV toolboxes benchmarked. Default options were selected (L = data length; t_{max} = last time index in timeseries; t_{min} = first time index in data series; Fs = sampling frequency; PSD = power spectral density estimate; FFT = fast Fourier transform, *ofac* = desired oversampling factor in computing the Lomb periodogram, *hifac* = multiplier of the average Nyquist frequency that defines the sequence of frequencies in computing the Lomb periodogram).

	PhysioNet HRV Toolkit	PhysioNet Cardiovascular Signal Toolbox	Kubios	Kaplan	Vollmer
QRS detection	*gqrs*, *wqrs*, *sqrs*	*jqrs*, *wqrs*, *sqrs*	Unknown QRS detector	No QRS detection	(Requires WFDB)
Noise and artifact identification method	Identify successive intervals whose difference exceeds threshold (20% of the value of adjacent 20 intervals on either side); identify non-physiologic intervals ($RR < 0.4$ s) ($RR > 2$ s)	Identify successive intervals whose difference exceed a threshold (20%); Identify PVCs, AF, VF; Identify Non-physiologic Intervals ($RR < 0.375$ s) ($RR > 2$ s)	Identify successive intervals whose difference exceed a threshold (20%)	'Glitches' identified using AR model	None
Artifact correction method	Remove non-physiologic *RR* intervals and intervals that exceed a threshold	Remove *RR* intervals that exceed threshold (default = 20%), PVCs, suspected AF/VF/VT, non-physiologic beats, and segments with SQI lower than 0.9 (when applicable) no interpolation	Interpolate through *RR* intervals that exceed a threshold	Spline Interpolation through data labeled 'glitches'	None
Frequency vector (Hz)	$df : df : df \times 2 \times L$ $df = \frac{1}{4(t_{max}-t_{min})}$	$\frac{1}{1024} : \frac{1}{1024} : \mathbf{0.5}$	$\frac{1}{300} : \frac{1}{300} : \mathbf{0.5}$	$\frac{Fs}{L}(\mathbf{0:1:}\frac{L}{2})$	$\frac{Fs}{2}(\mathbf{0:1:}\frac{NFFT}{2}+\mathbf{1})$ $NFFT = 2\text{^nextpow2}(L)$
Frequency Transformation	**PSD** Lomb periodogram	**PSD** Lomb periodogram	**FFT** Welch periodogram	**FFT**	**FFT**
Power Calculation	Squares PSD and sums bins in band	Squares PSD and sums bins in band	Sums bins in band	Squares then doubles FFT and sums bins in band	Doubles FFT and sums bins in band
Normalization	$\sqrt{\frac{PSD}{nout}}$ $nout = 0.5 * ofac * hifac * L$ $ofac = 4$, $hifac = 2$	$\sqrt{\frac{PSD}{nout}}$ $nout = 0.5 * ofac * hifac * L$ $ofac = 4$, $hifac = 2$	Normalizes to the total power	Normalizes to square of length of data segment analyzed	Normalizes to length of data segment analyzed

These toolboxes were chosen for their popularity, open-source availability, regard amongst experts in the field, or a combination of these factors. The studies in the benchmarking analysis, their purpose, and their sub-studies are described here.

4.1. Study A: comparison to a known standard LF/HF ratio

The aim of Study A was to compare the results generated by each toolbox on one HRV metric, the LF/HF ratio, using a known standard value. The LF/HF ratio is sensitive to small differences between populations. A time series with a known LF/HF ratio was generated using an *RR* interval generator detailed in Clifford (2002), hereafter called LFHFGEN. The default options for each toolbox were used to simulate the results achieved by a typical user of the HRV toolboxes, employing the software 'off the shelf'. (Studies have shown up to 95% of software users will not alter the default settings (Spool 2011).)

The LF/HF ratio generated from the various toolboxes were compared by calculating the normalized root mean square error (NRMSE) using the method of *mxm.c*, a WFDB routine that calculates the root mean squared error and normalizes it per the equation

$$\text{NRMSE} = {}^{100*}\sqrt{\frac{\sum_{i=1}^{n}(X_T - X_S)^2}{n}} \Big/ \frac{1}{n}\sum_{i=1}^{n} X_S$$

where n equals the number of windows considered, X_T is the metric generated by the test toolbox on the ith window, and X_S is standard compared against. The NRMSE value is reported back as a percentage. Default parameters and settings for each toolbox (per table 2) were used unless otherwise specified in the Methods.

One hundred synthetic 300 s *RR* interval time series were created with randomly assigned LF/HF ratios between 0.5 and 10 using the *RR* interval generator *LFHFGEN*. This generator produces an *RR* time series evenly sampled at 7 Hz composed of two sine waves at specific LF and HF frequencies (here we use the defaults of 0.095 Hz

and 0.275 Hz respectively). The frequencies are then slowly shifted to smear out the LF and HF frequency bands to generate a specific known LF/HF ratio. Finally, the time series was unevenly sampled in a realistic manner by searching for (and keeping only) each consecutive *RR* interval that is at least as large as the time from the previously selected *RR* interval. The time series were then analyzed with the various toolboxes according to table 2 to estimate the LF/HF ratio and the NRMSE was calculated using a standard that is found before frequency shifting or down-sampling. Two standards were used, one generated using the FFT and one generated using the Lomb periodogram. The PhysioNet toolboxes were compared to the Lomb standard and the Kubios, Kaplan, and Vollmer toolboxes were compared to the FFT standard. This provides a conservative outcome of the result.

4.2. Study B: the significance of collective processing differences

To expand the comparative analysis performed in Study A, the aim of Study B was to compare the results of the toolboxes on a wider selection of commonly assessed HRV metrics on both synthetic data and real patient data. The metrics generated include mean *NN* interval, PNN50, RMSSD, SDNN, HF, LF, LF/HF ratio, and total power. The default options for each toolbox were used to simulate the results achieved by a typical user of the HRV toolboxes. In addition to the default parameters, the artifact correction option (default: off) was also enabled on the Kubios toolbox analysis in order to determine the effects on HRV metrics. Each subsequent trial performs an evaluation on data with increasing amounts of noise. Trial 1 compares the HRV metric results from an analysis of synthetic *RR* interval data. Trial 2 compares the HRV metric results from an analysis of patient data from the MIT Normal Sinus Rhythm (NSR) database (Goldberger *et al* 2000). Trial 3 compares the HRV metric results from an analysis of patient waveform data from the MIT BIH Arrhythmia database (Moody and Mark 2001). The standard in all three trials was taken to be the PhysioNet HRV Toolkit, the most well published and validated of the available toolboxes.

4.2.1. Trial 1: synthetic RR interval data analysis

One hundred segments of synthetic *RR* interval data were generated using *RRGEN* (a method developed by McSharry *et al* (2002, 2003)) with the probability of ectopy set to 0.03% (Pe = 0.0003) and the probability of noise set to 0.48% (Pn = 0.0048). The segments were analyzed in full and were 600 s long. No segments were excluded from the analysis.

4.2.2. Trial 2: MIT NSR database RR interval data analysis

All 18 *RR* interval records from the MIT NSR database were segmented into 5 min windows with 4 min of overlap between windows, resulting in 23 103 windows. Non-normal annotations were removed.

Windows with possible AF (according to our detector described in section 3.2.5) or with greater than 15% of the data missing were not analyzed, reducing the dataset to 22 230 segments. An additional 182 segments, containing mostly noise and artifact, were eliminated by the PhysioNet HRV Toolkit as un-analyzable.

To determine the cause of diverging results from the toolboxes, a step by step comparison was performed using the PhysioNet HRV Toolkit and the PhysioNet Cardiovascular Signal Toolbox. The MIT NSR database was analyzed and normalized RMS error was calculated after each step of the analysis for each HRV metric. In the interest of using cleaner data to determine the cause of processing differences, windows with greater than 5% of the data missing were not analyzed. The windows were minimally preprocessed with the PhysioNet HRV Toolkit and the data was then fed into both the PhysioNet HRV Toolkit and the PhysioNet Cardiovascular Signal Toolbox.

The first comparison (Comparison 1) involved only varying the toolbox for calculating HRV statistics. This involved keeping the preprocessing steps and definition of the frequency bins constant. The frequency bins were assigned by the PhysioNet HRV Toolkit. The mean was removed before calculating spectral metrics. Mean *NN* interval, PNN50, RMSSD, SDNN, HF, LF, LF/HF ratio, and total power were calculated on each window over the entirety of the 24 h recording for each patient ($n = 18$). The spectral metrics were calculated using the Lomb-Scargle periodogram and normalized per the method in the C implementation of the function in Numerical Recipes in C (Press *et al* 1992).

The second comparison (Comparison 2) involved varying the toolbox for calculating HRV statistics and frequency bin assignment. The third comparison (Comparison 3) involved varying the toolbox for calculating HRV statistics, frequency bin assignment, and preprocessing algorithm.

4.2.3. Trial 3: waveforms of the MIT BIH arrhythmia database

All 48 records from the MIT BIH Arrhythmia Database (Goldberger *et al* 2000) were processed using the waveform analysis methods in the respective toolboxes which possess this functionality (namely the PhysioNet HRV Toolkit, Kubios, PhysioNet Cardiovascular Signal Toolbox, and Vollmer). Each approximately 30 min record was broken up into five minute segments with four minutes of overlap between them and then HRV metrics were

estimated on each segment. Segments from all 48 records were compiled and NRMSE was computed on the compiled segments.

4.3. Study C: long range scaling metrics—DFA

The aim of this study was to compare the results from the PhysioNet HRV Toolkit's detrended fluctuation analysis (DFA) algorithm, *dfa.c*, to the remainder of the toolboxes described in table 1. One hundred segments of synthetic *RR* interval data were generated using *RRGEN* (McSharry *et al* 2002, 2003) with the probability of ectopy and noise set to 0% (Pe = 0, Pn = 0). The segments were 24 h long and were analyzed in their entirety. The same dataset was analyzed for Studies D and E. Default options were used for all the toolboxes. The effect of different ranges of box size m used for the computation of the scaling exponents α_1 and α_2 was assessed as well as the proprietary detrending option implemented by the Kubios toolbox.

4.4. Study D: long range scaling metrics—MSE

The goal of this study was to compare the results of MSE analysis of the PhysioNet Cardiovascular Signal Toolbox and Kubios to the results generated with the PhysioNet HRV Toolkit. The Kaplan and Vollmer toolboxes do not provide MSE estimates, so those toolboxes were not analyzed in this study. The effects of detrending options on the entropy values calculated with Kubios were also compared (Kubios MSE calculations use the default of detrending).

The MSE implementation from the PhysioNet HRV Toolkit is preset to use a default pattern length of $m = 2$ and a similarity criterion of $r = 0.15$, the same defaults as in the PhysioNet Cardiovascular Signal Toolbox. The maximum number of coarse-grained time series is defined by the parameter τ_{max}, which by default is set to be equal to 20. The scaling exponents of synthetic *RR* interval data were also estimated.

At each scale, the relative error, defined as $\epsilon = \frac{|X_T - X_S|}{|X_S|}$, where X_T is the metric generated by the test toolbox and X_S is standard compared against, was computed.

4.5. Study E: PRSA

The aim of this study was to show equivalency between the PRSA algorithm from the PhysioNet Cardiovascular Signal Toolbox and code from the original authors of PRSA (Bauer *et al* 2006). PRSA is not available in any other toolbox, so no other comparisons are made. One hundred synthetic signals, previously used for Studies C and D, were used for the comparison of the code included in the PhysioNet Cardiovascular Signal Toolbox to the code provided by Bauer *et al* (2006) in order to ensure that the code between the two were consistent and free from implementation errors.

4.6. Study F: HRT

The aim of this study was to show equivalency between the HRT algorithm from the PhysioNet Cardiovascular Signal Toolbox (*HRT_Analysis.m)* and the HRT code provided by Raphael Schneider (Bauer *et al* 2008). No other toolboxes perform this analysis, so no other comparisons were performed. The comparison with Schneider's code was performed on data from the MIT NSR database (Goldberger *et al* 2000). Since the two code bases under evaluation use different preprocessing methods, both preprocessing methods were used in different tests: (i) removal of *RR* intervals that change by more than >20% with respect to the mean of the five last sinus intervals and, (ii) removal of *RR* intervals that change by more than >20% with respect to the previous interval.

5. Results

5.1. Study A

The PhysioNet Cardiovascular Signal Toolbox and Kaplan toolboxes achieved negligible error in the LF/HF ratio, with errors between 3.5% and 5.7% (table 3). (The rationale to indicate these are negligible here is that the LF–HF ratio changes by approximately 20%–100% during different activities or between different medical conditions (Bernardi *et al* 2000, Otzenberger *et al* 1998).) Although the Kaplan toolbox exhibited the lowest average error in the LF/HF ratio estimate on this dataset compared to the known LF/HF ratio, it not possible to say that it has definitively performed in a superior manner to the Lomb periodogram for two reasons. First, the difference was only around 2%, which is trivial in terms of the LF/HF ratio. Second, the simulated data was a synthetically generated combination of a sine waves whereas real HRV data is much more complex with stochastic noise and nonstationarities that may have been exacerbated by the resampling procedures required by the FFT.

Kubios's default calculation using FFT results in a 33.6% error. When the option is engaged to use the Lomb periodogram method the error drops to 6.1%. Vollmer's toolbox has the highest error at 58.2%. We note that these errors may be consistent offsets, which, although they prevent comparison between studies, can still provide valid comparisons within studies. Nevertheless, we strongly suggest using a toolbox with settings that pro-

Table 3. Study A. The normalized RMS error generated among different toolboxes on LF/HF ratio when compared to a known (artificial) standard.

	PhysioNet HRV toolkit (%)	PhysioNet Cardiovascular Signal Toolbox (%)	Kubios FFT method (%)	Kubios Lomb method (%)	Kaplan (%)	Vollmer (%)
LF/HF	25.0	5.7	33.6	6.1	**3.5**	58.2

Table 4. Study B—Trial 1. The normalized RMS error (or discrepancy) generated on various HRV metrics compared to the metric calculated by the PhysioNet HRV Toolbox on synthetic data.

Metric	PhysioNet Cardiovascular Signal Toolbox (%)	Kubios: no artifact correction FFT (%)	Kubios: artifact correction FFT (%)	Kubios: no artifact correction Lomb (%)	Kubios: artifact correction Lomb (%)	Kaplan (%)	Vollmer (%)
Mean NN	0.4	0.4	0.4	0.4	0.4	0.4	0.4
pNN50	**4.2**	4.9	4.6	4.9	4.6	**4.2**	**4.2**
RMSSD	2.0	1.1	**1.0**	1.1	**1.0**	1.9	1.9
SDNN	9.4	34.5	34.5	34.5	34.5	**8.3**	9.3
VLF	48.7	94.0	94.0	87.5	87.5	**26.4**	4.3×10^5
LF	**28.5**	36.1	36.1	51.4	51.2	39.4	1.1×10^6
HF	70.8	38.0	38.0	**34.1**	34.3	45.6	1.6×10^6
TTLPWR	49.3	65.5	65.5	59.2	59.2	**11.4**	6.0×10^5
LF/HF	137.1	102.9	102.9	114.4	114.4	139.7	**35.5**

vides an error below 5% or 10%, since this may still allow the user ability to distinguish between mental and physical activities. Note that from here on in this article, all comparisons will be made with the PhysioNet HRV toolkit (written in C). This is not because this is necessarily correct, but because it is the most well-known open source HRV toolbox, and one to which we would like to closely map in order to allow the interchange of C and MATLAB functions when computational efficiency is important.

5.2. Study B

5.2.1. Trial 1: synthetic data

The entire dataset of synthetic data was analyzed with no records eliminated. The calculated error between the toolboxes when compared to the results from the PhysioNet Cardiovascular Signal Toolbox are shown in table 4. Note that since the data are synthetic with no artifact, the artifact correction in the Kubios software leads to a negligible difference to the results calculated with the same software and no artifact correction.

5.2.2. Trial 2: patient data

Of the 23 103 segments created from the database, 22 994 had annotations marked 'N' (normal). A total of 2835 segments were not analyzed because AF was detected (2366 segments) or too little data was present in the segment (more than 5% of the window was missing or noisy).

The calculated error between the toolboxes when compared to the results from the PhysioNet HRV Toolkit are shown in table 5. The PhysioNet Cardiovascular Signal Toolbox operates most closely to the PhysioNet HRV Toolkit, as is seen by its low NRMSE values.

Although table 5 shows large differences exist for all toolboxes, the PhysioNet Cardiovascular Signal Toolbox provided the closest correspondence to the PhysioNet HRV Toolbox. To determine the origin of the differences, the PhysioNet Cardiovascular Signal Toolbox and PhysioNet HRV Toolkit were compared side by side on the MIT NSR database. In Comparison A, the PhysioNet Cardiovascular Signal Toolbox generated results which were within 3.4% NRMSE of the PhysioNet HRV Toolbox (table 6) on all metrics tested. The metrics with the highest error were PNN50 and RMSSD. The minor differences in these metrics can be largely attributed to the fact that the PhysioNet HRV Toolbox removed additional data points on the edge of the windows compared to the method by the PhysioNet Cardiovascular Signal Toolbox. To a lesser extent, the remainder of the error is due to round off of constants that can be performed differently in MATLAB and in C (integers can be defined differently). None of these errors are clinically significant compared to any studies that have leveraged HRV metrics, and therefore we consider the toolboxes equivalent in this benchmark test.

Frequency binning (Comparison B) added significant error to the calculation of spectral metrics. The LF/HF ratio was least impacted by this effect, but the error still increased on this metric to almost 2%. Once the preprocessing was varied (Comparison C), the errors continued to climb.

Table 5. Study B—Trial 2. The normalized RMS error generated among different toolboxes on standard HRV metrics when compared to the values of the same metrics calculated by the PhysioNet HRV Toolkit on expert beat-labelled *RR* interval data taken from the MIT NSR database.

Metric	PhysioNet Cardiovascular Signal Toolbox (%)	Kubios: no artifact correction FFT (%)	Kubios: artifact correction FFT (%)	Kubios: no artifact correction Lomb (%)	Kubios: with artifact correction Lomb (%)	Kaplan (%)	Vollmer (%)
Mean NN	**1.5**	4.2	3.7	4.2	3.7	3.9	4.2
pNN50	**17.1**	55.1	38.7	55.1	38.7	44.3	54.4
RMSSD	**31.7**	165.7	113.6	165.7	113.6	128.2	171.6
SDNN	**18.3**	67.1	58.4	67.1	58.4	52.3	71.1
VLF	**67.3**	158.6	159.5	880.9	157.0	146.9	2.5×10^5
LF	**90.2**	298.2	184.0	802.2	200.4	184.8	6.6×10^5
HF	**163.6**	1.9×10^3	1.1×10^3	961.7	555.9	785.4	1.4×10^6
TTLPWR	**71.0**	325.0	217.9	711.5	155.7	186.8	4.6×10^5
LF/HF	**49.2**	72.3	67.5	72.8	50.6	50.3	102.8

Table 6. Study B—Trial 2. The calculated differences between the PhysioNet HRV Toolkit and the PhysioNet Cardiovascular Signal Toolbox as determined by the NRMSE method. Comparison A used identical settings for both toolboxes. Comparison B introduced the variability due to the different frequency binning methods between the two toolboxes. Comparison C introduced the variability due to preprocessing differences between the two toolboxes. N/A indicates the fact that trial B affected only spectral metrics.

Comparison → HRV metric ↓	*A* (%)	*B*	*C* (%)
Mean NN	0.0	N/A	0.6
pNN50	3.4	N/A	11.8
RMSSD	2.6	N/A	8.3
SDNN	0.0	N/A	10.0
VLF	0.0	8.6%	41.0
LF	0.0	3.8%	27.4
HF	0.0	4.0%	32.4
LF/HF ratio	0.0	1.8%	42.4
TTLPWR	0.0	4.8%	24.0

Table 7. Study B—Trial 3. The NRMSE difference generated among different toolboxes on standard HRV metrics when compared to the values of the same metrics calculated by the PhysioNet HRV Toolkit.

Metric	PhysioNet Cardiovascular Signal Toolbox (%)	Kubios: no artifact correction FFT (%)	Kubios: with artifact correction FFT (%)	Kubios: no artifact correction Lomb (%)	Kubios: with artifact correction Lomb (%)	Vollmer (%)
Mean NN	**2.1**	8.8	8.5	8.8	8.5	11.7
pNN50	**36.4**	91.0	76.5	91.0	76.5	86.3
RMSSD	**86.6**	976.5	189.2	976.5	189.2	299.3
SDNN	**74.1**	1.3×10^3	127.6	1.3×10^3	127.6	166.0
VLF	**243.6**	8.0×10^4	507.5	5.8×10^4	401.7	1.0×10^5
LF	**603.3**	4.2×10^5	1.6×10^3	2.3×10^5	467.3	3.1×10^5
HF	918.7	1.4×10^4	1.1×10^3	3.9×10^5	**601.1**	5.8×10^5
TTLPWR	380.9	1.1×10^5	572.9	1.5×10^5	**352.1**	2.1×10^5
LF/HF	793.1	824.3	793.7	792.4	**791.4**	797.1

5.2.3. Trial 3: waveform data

The discrepancy between the toolboxes being tested when compared to the results generated by the PhysioNet HRV Toolkit were calculated and are shown in table 7. Windows that did not meet minimal requirements for the PhysioNet Cardiovascular Signal Toolbox were not analyzed, resulting in the loss of 92 out of 1248 windows. Those minimal requirements include greater than 90% SQI and less than 15% of data lost to cleaning. Only the Kubios software with artifact correction and Lomb frequency domain metrics compared with the PhysioNet Cardiovascular Signal Toolbox in terms of mapping to the existing PhysioNet HRV Toolbox.

Table 8. Study C. The NRMSE generated among different toolboxes on DFA scaling coefficients α_1 and α_2 compared to the values calculated by the PhysioNet HRV Toolkit.

	PhysioNet Cardiovascular Signal Toolbox (%)	Kubios (default settings) (%)	Kubios (no detrending) (%)	Kaplan (%)	Vollmer (%)
α_1[a]	1.5	5.4	0.6	0.7	**0.1**
α_2[b]	**3.1**	34.6	16.1	18.6	17.4

[a] Short-term scaling coefficient: all toolboxes $4 \leqslant n \leqslant 16$.

[b] Long-term scaling coefficient: PhysioNet HRV Toolkit and PhysioNet Cardiovascular Signal Toolbox: $16 \leqslant n \leqslant N/4$; Other toolboxes: $16 \leqslant n \leqslant 64$.

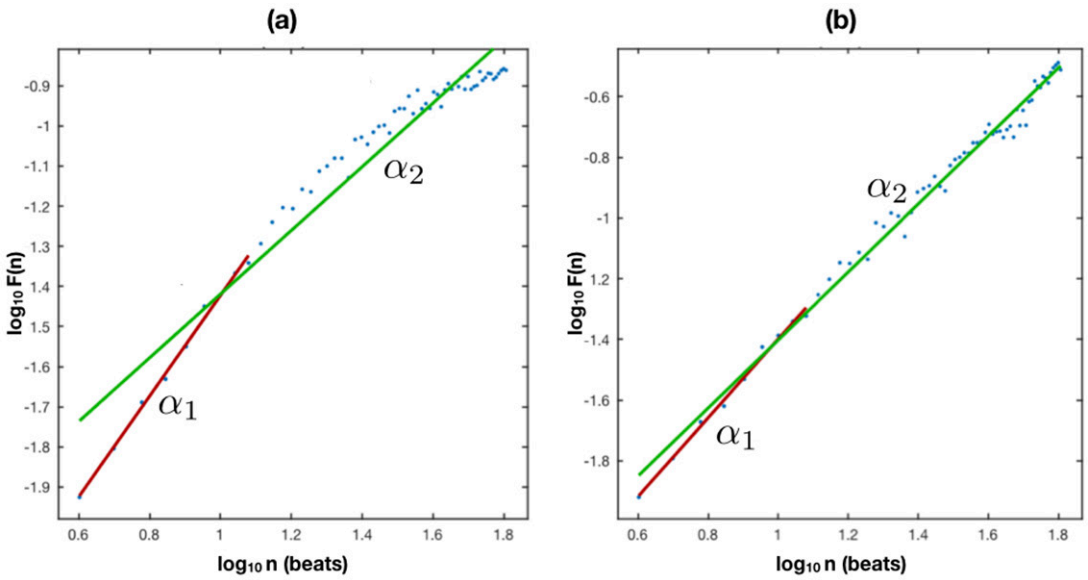

Figure 2. Study C—results from DFA generated using Kubios with (a) detrending and smoothness priors ($\alpha_1 = 1.26, \alpha_2 = 0.79$) and (b) without detrending ($\alpha_1 = 1.29, \alpha_2 = 1.12$).

5.3. Study C: long range scaling metrics—DFA

The differences between the toolboxes when compared to the results from the PhysioNet HRV Toolkit are calculated in table 8. Note that the large difference for the coefficient α_2 found for the Kubios software could be a consequence of the default detrending option using the method called smoothness priors, which basically corresponds to a time-varying high pass filter with $f_c = 0.035$ Hz using default parameters. Figure 2 highlights the effect of the detrending option on the estimation of α_2.

5.4. Study D: long range scaling metrics—MSE

Figure 3 shows results for MSE computed on 24 h synthetic *NN* tachograms, which reports the relative error ε for each MSE scale calculated with the PhysioNet Cardiovascular Signal Toolbox and Kubios in comparison to the MSE scale calculated by the PhysioNet HRV Toolkit. The error was shown to be lower than 0.0004 at all scales for the PhysioNet Cardiovascular Signal Toolbox whereas the Kubios MSE implementation, with and without detrending, shows significantly higher error.

5.5. Study E: PRSA

The results of this study show that the code implemented in the PhysioNet Cardiovascular Signal Processing Toolbox provide the same results as the one provided by Bauer *et al* (2006). Both DC and AC measures on 100 synthetic signals generated using *RRGEN* achieve an average NRMSE of 0%.

5.6. Study F: HRT

Comparison of HRT algorithms on the MIT NSR database for the PhysioNet Cardiovascular Signal Toolbox using the default filtering option against the code provided by Raphael Schneider (Bauer *et al* 2006) resulted in a NRMSE value of 9.4% for the TO and 8.5% for the TS. Using the second filtering option (removal of *RR* intervals that change by more than >20% with respect to the previous), as implemented in the original code provided by Raphael Schneider, resulted in an NRMSE value of 6.5% for the TO and of 1.0% for the TS.

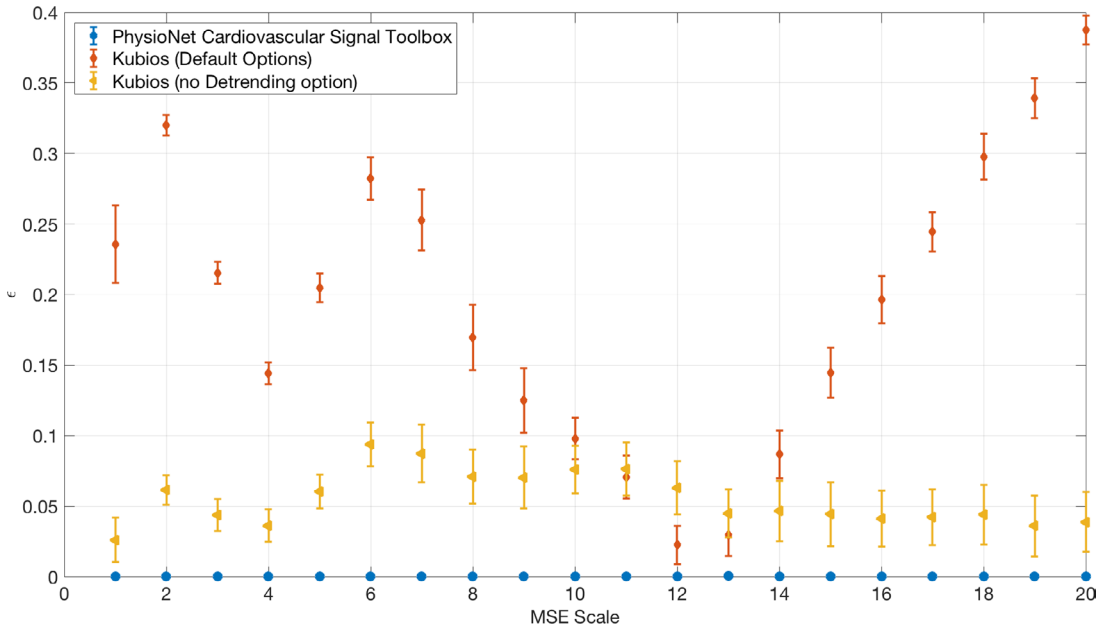

Figure 3. Study D—plot of the mean relative error ε at different scales of multiscale entropy using the PhysioNet Cardiovascular Signal Toolbox and Kubios with respect to results from the PhysioNet HRV Toolkit for 100 synthetic *NN* interval signals generated using *RRGEN* (Pn = 0, Pe = 0). Error bars show standard deviation from the mean.

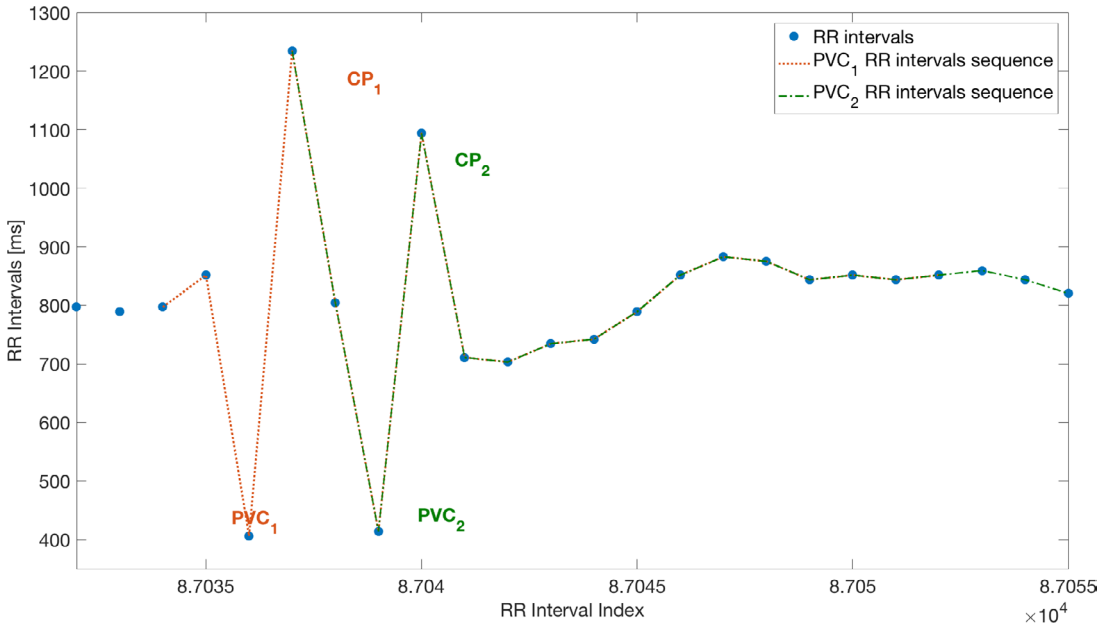

Figure 4. Study F—example of two consecutive PVCs in signal nsr010. Tachogram related to the second PVC contains the compensatory pause CP of the preceding PVC and thus is excluded by the HRT algorithm implemented in the PhysioNet Cardiovascular Signal Toolbox.

We investigated the reason of the larger error for the TO value using the second filtering setting. On the analyzed dataset, for some recordings, a larger number of *NN* intervals before and after PVCs have been 'filtered' by the PhysioNet Cardiovascular Signal Toolbox than the Schneider code. When two or more PVCs are separated by only a small amount of time the rejection is performed differently. The PhysioNet Cardiovascular Signal Toolbox excludes PVCs for which one of the two *RR*-intervals before the current PVC is a compensatory pause of the previous PVC, while in Schneider's implementation those *RR* intervals before or after the PVCs are considered valid.

An example is reported in figure 4, where the intervals related to the second PVC exhibit a compensatory pause (CP) from the preceding PVC. Since a tachogram is considered valid for HRT analysis if it has sinus rhythm interval preceding and following a PVC, both sequences of *RR* intervals are excluded by our implementation. TO was computed using the two *NN* intervals preceding the PVC and two *NN* intervals following the CP, thus including a CP in the computation of the TO might lead to different results.

6. Discussion and recommendations

The benchmarking results detailed in this work demonstrated that significant errors result from seemingly small and inconsequential choices in analysis methods. Moreover, the earlier in the process pipeline that the choices begin to differ, the larger the overall effects. The differences in analysis methods, parameter choices, and data preprocessing have yielded a field of HRV results that are impossible to compare between patient populations and research groups, and perhaps even within research groups. The results have shown that it is imperative that future studies adhere to a consistent method of reporting how an analysis has been performed, particularly in terms of the many parameter settings possible. (Although Malik *et al* (Task Force of the European Society of Cardiology the North American Society of Pacing Electrophysiology 1996, Clifford 2002, Pan *et al* 2016) attempted to encourage this practice, their prescription of what to report was too vague, and did not detail any requirements on preprocessing, apart from some basics of interval rejection.)

We note that the analysis in this article has some limitations. First, although we have quantified the differences in each HRV metric, and in some cases, these were enormous, this does not necessarily mean that the use of one toolbox over another, in classification tasks for example, would result in a significant difference in the algorithm's accuracy (or other pertinent metric). This is particularly true for a multivariate or nonlinear classifier. Conversely, a small difference may, for some tasks, result in large classification differences. What we can say, however, is that one should not compare results between articles that have used different toolboxes, let alone home-grown or unspecified/closed software.

When considering the use of HRV analysis in research, it is important that researchers carefully consider the data to be analyzed and the assumptions of the analysis. An essential part to that consideration is identifying the methods and settings used for the analysis and providing this listing in the subsequent publication along with the data. The PhysioNet Cardiovascular Signal Toolbox initialization file can be used as a template when publishing this information. Researchers should compare subjects with similar length recordings to minimize the effect of metrics sensitive to temporal recording length (such as scaling metrics). Moreover, longer recordings can lead to larger averaging, or the capture of behaviors at different points in the circadian or daily rhythm. Subjects should also be exposed to similar psychosocial scenarios, where stress, environment, and mental state can be carefully controlled variables. Sleep is a good normalization approach, as shown in Clifford and Tarassenko (2004).

How a preprocessing algorithm addresses noise, ectopy, or artifact can have either a subtle or a significant effect on the results of analysis and depends to a large extent on how reliable or corrupt the data is to begin with. When a comparison was made between data pre-processed with the PhysioNet Cardiovascular Signal Toolbox and the PhysioNet HRV Toolkit, two toolboxes with markedly similar approaches to HRV analysis, the differences observed ranged from 0.6% on the Mean *NN* interval to over 40% on LF/HF ratio (table 6, Comparison C). When investigating the cause of error in the non-spectral metrics (RMSSD, pNN50, and SDNN), it was observed that a single window with just one or two removed non-physiologic data points can dramatically affect the estimated value, particularly the NRMSE. More markedly, table 7 shows that even simple time domain statistics can differ by significant amounts when different QRS detectors or different abnormal interval filters are employed.

The normalization of the PSD estimation is seldom reported, and the method employed can have a very large influence on spectral results, especially when they are not reported as ratios. It is usually very difficult to retrospectively determine how an author has normalized data if only a select handful of parameters are reported. The effect of differing frequency bins on the results of spectral analysis can also be a significant source of error between two different methods analyzing the same data. When the PhysioNet HRV Toolkit and PhysioNet Cardiovascular Signal Toolbox were allowed to define the frequency bins separately, the RMS error on LF/HF ratio, a metric that is buffered from error because of the nature of ratios, was over 2% (table 6, Comparison B). The error for the identical power calculations with slightly different frequency bands was nearly 4% at best and 8% at worst. Especially at VLF, where the binning may leave these bands with only one to five bins, changes in can lead to significant differences in the outcome.

When using frequency domain analysis, the Lomb periodogram has demonstrated to be the superior choice for *RR* interval data (Clifford and Tarassenko 2005). Therefore, it should be standard practice to present results using the Lomb periodogram when referencing a spectral metric. However, it is important to note that the *RR* interval time series is not a stationary time series and therefore, sliding a window across data and using a technique that assumes stationarity is somewhat flawed. Although there has been much attention paid to time-frequency tools over the last two decades, little work has been done on unevenly sampled data and so we do not currently include such tools in this toolbox (since the effect of resampling on such tools has not been rigorously tested). Instead we recommend segmenting data into stationary blocks.

We recommend that the PhysioNet Cardiovascular Signal Toolbox be used to perform HRV analysis because of the following advantages.

1. A close correspondence to the C code of PhysioNet's HRV Toolkit. This allows the user to swap between code bases for embedded applications or fast execution on servers for a subset of the algorithms.
2. Parameters chosen are not arbitrary and have been justified in this publication.
3. Parameters have been refactored into one initialization file so the user can change this as suited and report the changes efficiently.
4. Extensive benchmarked waveform analysis tools are included.
5. It is the only software suite that includes signal quality and arrhythmia analysis tools to help remove noisy and non-sinus periods of data.

Comparison to standard models and other available software demonstrate that the PhysioNet Cardiovascular Signal Toolbox can even be used itself as a benchmarking system for other HRV studies, FDA filings, and industrial applications (due to the BSD licensing).

Other toolboxes that are characterized in this paper may produce adequate analysis. In particular we found that, with certain potential clinically significant differences in long range metrics, Kubios software was similar to our toolbox and the PhysioNet C toolbox and is sufficient for clinicians to use if they are willing to hand operate the software on a per-file basis (since no scripting facility is available in Kubios at this time) and as long as the default parameters are not selected. However, due to the dangers of hand-processing data, we could only really recommend Kubios if a batch scripting version were made available. We recommend that when using other toolboxes, users report the differences between their code and other HRV tools to avoid erroneous conclusions when comparing with the literature. Also, by comparing to our toolbox, it will persuade others to report the many thresholds that are swept under the carpet in other studies, but which have such an enormous effect on the output. We also note that none of the toolboxes presented are as comprehensive as the PhysioNet Cardiovascular Signal Toolbox.

Finally, the full potential of HRV analysis, the subject of so many studies over the last 40 years or more, will not be realized without further contributions to the open source tools. We encourage benchmarked contributions to our toolbox, which is freely available from PhysioNet and Github (Vest *et al* 2018).

7. Conclusions

This article presents evidence in support of standardizing HRV analysis methods and demonstrates how the PhysioNet Cardiovascular Signal Toolbox makes advances towards such standardization. Using in-house code that has not been thoroughly benchmarked and failing to report all parameter settings will continue to hold the field back. We caution against the use of default parameters, particularly when dealing with raw ECG or other pulsatile data. We recommend that researchers use our MATLAB toolbox except where fast implementation is needed, and then to use the PhysioNet C implementation where code is available. Rigorously applying the standards described in this article and working with common, benchmarked code such as that provided with this publication, will improve the science of HRV analysis and, we hope, should provide a significant boost to its clinical utility.

Acknowledgments

The authors wish to acknowledge the National Institutes of Health (Grant # NIH K23 HL127251, R01 HL136205) the National Science Foundation Award 1636933, the Fogarty International Center and the Eunice Kennedy Shriver National Institute of Child Health and Human Development, grant number 1R21HD084114-01, the Rett Syndrome Research Trust, and the One Mind Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, the National Institutes of Health, the Rett Syndrome Research Trust or the One Mind Foundation. The authors also wish to thank Minxuan Huang for contributing to data analysis, and George Moody and Joe Mietus for creating and posting the PhysioNet HRV toolkit that serves as the baseline comparison point for this article. GC wishes to extend a big thanks to Danny Kaplan for suggesting this toolbox in Oxford during his sabbatical at the Maths Institute around 1999, to Patrick McSharry for introducing us, and for code, collaborations, friendship and inspiration, and to Lionel Tarassenko, Roger Mark and Ary Golberger for patience, patronage and mentorship.

Appendix A. Software requirements to Use the PhysioNet Cardiovascular Signal Toolbox

The current version (1.0) of the HRV toolbox was tested with the following MATLAB configuration: MATLAB (v 9.3), Signal Processing Toolbox (v 7.3), Neural Network Toolbox (v 7.0), and Statistics and Machine Learning Toolbox (v 11.0). The Toolbox has been tested using Windows, OSX, and Unix systems.

Table A1. Performance of peak detectors when tested on the MIT BIH Arrhythmia Database (taken from Vest *et al* (2017)).

Peak detector	Recommended application	F1	St dev
wqrs.c	Low-noise scenarios or as a comparator to detect noise	99.00	1.89
wqrs.m	Low-noise scenarios or as a comparator to detect noise	99.04	1.84
sqrs.c	Low-noise scenarios or as a comparator to detect noise	98.19	4.22
sqrs.m	Low-noise scenarios or as a comparator to detect noise	96.33	6.38
jqrs.m	Long-term moderate to high noise recordings, such as in ICU Holter or exercise.	93.02	12.27
gqrs.c	Moderate-noise ICU or Holter recordings	95.72	14.84

Table A2. Default parameters in the PhysioNet Cardiovascualr Signal Toolbox. au indicates arbitrary units.

Parameter	Value	Unit	Description
data_confidence_level	1	au	NOT YET IN USE
Windowlength	300	s	HRV statistics analysis window length
Increment	60	s	HRV statistics sliding window increment
Numsegs	5	au	Number of segments to collect with lowest HR
RejectionThreshold	0.2	au	Amount of data that can be rejected before a window is considered too low quality for analysis. 0.2 = 20%
MissingDataThreshold	0.15	au	Maximum percentage of data allowable to be missing from a window. 0.15 = 15%
sqi.LowQualityThreshold	0.9	au	Threshold for which SQI represents good data
sqi.windowlength	10	s	SQI analysis window length
sqi.increment	1	s	SQI sliding window increment
sqi.TimeThreshold	0.1	s	Maximum absolute difference in annotation times that is permitted for matching annotations.
sqi.margin	2	s	Margin time not include in comparison
preprocessg.aplimit	2	s	Maximum believable gap in *RR* intervals
preprocess.per_limit	0.2	au	Percent limit of change from one interval to the next. 0.2 = 20%
preprocess.forward_gap	3	s	Maximum tolerable gap at beginning of timeseries in seconds
preprocess.method_outliers	'rem'	—	Method of dealing with outliers
preprocess.lowerphysiolim	0.375	s	Lower physiological limit, minimum *RR* interval
preprocess.upperphysiolim	2	s	Upper physiological limit, maximum *RR* interval
preprocess.method_unphysio	'rem'		Method of dealing with unphysiologically low beats. 'rem' = removal
Preprocess.threshold1	0.9	au	Threshold for which SQI represents good data
preprocess.minlength	30	s	The minimum length of a good data segment in seconds
af.windowlength	30	s	AFib analysis window length, set to include ~30 beats in each window
af.increment	30	s	AFib sliding window increment
timedomain.alpha	50	ms	Alpha value for PNN analysis method
timedomain.win_tol	0.15	au	Maximum percentage of data allowable to be missing from a window. 0.15 = 15%
prsa.thresh_per	20	%	Percent difference that one beat can differ from the next in the PRSA code
prsa.win_length	30	s	The length of the PRSA signal before and after the anchor points
pPrsa.scale	2	au	Scale parameter for wavelet analysis (to compute AC and DC)
ulf	0–0.0033	Hz	ULF band, requires window > 300 s
vlf	0.0033–0.04	Hz	VLF band, requires at least 300 s window
lf	0.04–0.15	Hz	LF band, requires at least 25 s window
hf	0.15–0.4	Hz	HF band, requires at least 7 s window
freq.zero_mean	1	—	Option for subtracting the mean from the input data
freq.method	'lomb'	—	Frequency estimation method, Options: 'lomb', 'burg', 'fft', 'welch'
freq.normalize_lomb	0	—	When selected, adds a normalization step to frequency domain analysis

(*Continued*)

Table A2. (*Continued*)

Parameter	Value	Unit	Description
freq.burg_poles	15	au	Number of coefficients for spectral estimation using the Burg method (not recommended)
freq.resampling_freq	7	Hz	Resampling frequency for 'welch', 'fft', or 'burg'
freq.resample_interp_method	'cub'	—	Resampling interpolation method for 'welch', 'fft', or 'burg'
freq.resampled_burg_poles	100	au	Number of poles for burg method
sd,segmentlength	300	s	Windows length for SDANN and SDNNI analysis
PeakDetect,REF_PERIOD	0.25	s	Assumed refractory period after a natural sinus beat
PeakDetect.THRES	0.6	au	Energy threshold of the peak detector
PeakDetect.fid_vec	[]	—	If some subsegments should not be used for finding the optimal threshold of the P&T then input the indices of the corresponding points here
PeakDetect.SIGN_FORCE	[]	—	Force sign of peaks (positive value/negative value). Particularly useful in a window by window detection with uncertain peak polarity. Could be used to build an Fetal ECG template.
PeakDetect.ecgType	'MECG'	—	Use QRS detector for Adult ECG analysis
PeakDetect.windows	15	s	Size of the window onto which to perform QRS detection
MSE.windowlength	[]	s	Window size in seconds. Default [] performs MSE on the entire signal
MSE.increment	[]	s	MSE window increment. Default [] performs MSE on the entire signal
MSE.RadiusOfSimilarity	0.15	au	Radius of similarity (% of std)
MSE.patternLength	2	au	Pattern length for SampEn computation
MSE.maxCoarseGrainings	20	au	Maximum number of coarse-grainings
Entropy.RadiusOfSimilarity	0.15	au	Radius of similarity (% of standard deviation)
Entropy.patternLength	2	au	Pattern length for SampEn computation
DFA.windowlength	[]	s	Windows size for DFA analysis Default [] performs DFA on entire signal
DFA.increment	[]	s	Sliding window increment for DFA analysis Default [] uses no sliding window
DFA.minBoxSize	4	au	Smallest box width for DFA analysis
DFA.maxBoxSize	[]	au	Largest box width for DFA analysis Default [] uses the signal length/4
DFA.midBoxSize	16	au	Medium time scale box width for DFA analysis
HRT.BeatsBefore	2	au	Number of beats before PVC
HRT.BeatsAfter	16	au	Number of beats after PVC and CP
HRT.windowlength	24	h	Window size for HRT analysis. Default 24 h
HRT.increment	24	h	Sliding window increment or HRT analysis Default 24 h
HRT.filterMethod	'mean5before'	—	HRT analysis filtering option

Appendix B. QRS detection benchmark testing for PhysioNet Cardiovascular Signal Toolbox and PhysioNet HRV Toolkit

Appendix table A1 provides results detailed in Vest *et al* (2017) for a comparison of the standard QRS detectors available in the PhysioNet Cardiovascular Signal Toolbox and PhysioNet HRV Toolkit when tested on the MIT BIH Arrhythmia Database. Note that the database on which they are tested is largely free from noise and artifact. The F1 scores therefore reflect how well they perform in ideal circumstances. When noise is present, only *jqrs* and *gqrs* are able to maintain accuracy.

Appendix C. Default parameters in the PhysioNet Cardiovascular Signal Toolbox

Appendix table A2 provides the default parameters utilized in the PhysioNet Cardiocascular Signal Toolbox. Note that parameters related to file extension, demo visualization, and saving options are not reported. Only analysis related parameters are summarized below.

Appendix D. Demonstration code available in the PhysioNet Cardiovascular Signal Toolbox

D.1. Atrial fibrillation detection demo: *DemoRawDataAF.m*

This demonstration analyzes a segment of raw (or filtered) ECG signal with known atrial fibrillation to show the operation of the AF detection algorithm and its use in removing segments of arrhythmia during HRV analysis.

D.2. Annotated data demo: *DemoAnnotatedData.m*

This demonstration uses the PhysioNet Cardiovascular Signal Toolbox on *RR* intervals with annotations. After pre-processing the *RR* intervals—taking into account the beat annotations—and removal of windows containing AF, the HRV analysis is performed on the clean *NN* (normal-to-normal) time series and the resulting output is saved in a.csv file.

D.3. ECG, ABP, and PPG data demo: DemoRawDataICU.m

This demonstration analyzes a segment of data collected in the intensive care unit (ICU) which contains ECG, ABP, and PPG signals. This demo will perform HRV analysis on the raw ECG signals as well as detection of fiducial points of PPG and ABP signals. It will also display the pulse transit time (PPT) graph (Blood Pressure versus PTT).

D.4. *RRGEN* data demo: *DemoStandardizedData.m*

This function demonstrates the function of the synthetic *RR* interval generator *RRGEN* and the calculation of HRV metrics.

ORCID iDs

Adriana N Vest https://orcid.org/0000-0001-7889-1956
Giulia Da Poian https://orcid.org/0000-0002-8960-1077

References

American Heart Association 2018 ECG Database www.ecri.org/components/Pages/AHA_ECG_USB.aspx (Accessed: 2018)
Bauer A and Schmidt G 2003 Heart rate turbulence *J. Electrocardiol.* **36** 89–93
Bauer A, Kantelhardt J W, Bunde A, Barthel P, Schneider R, Malik M and Schmidt G 2006 Phase-rectified signal averaging detects quasi-periodicities in non-stationary data *Physica* A **364** 423–34
Bauer A, Malik M, Schmidt G, Barthel P, Bonnemeier H, Cygankiewicz I, Guzik P, Lombardi F, Müller A and Oto A 2008 Heart rate turbulence: standards of measurement, physiological interpretation, and clinical use: International Society for Holter and Noninvasive Electrophysiology Consensus *J. Am. Coll. Cardiol.* **52** 1353–65
Behar J, Johnson A, Clifford G D and Oster J 2014 A comparison of single channel fetal ECG extraction methods *Ann. Biomed. Eng.* **42** 1340–53
Berkowitsch A, Zareba W, Neumann T, Erdogan A, Nitt S M, Moss A J and Pitschner H F 2004 Risk stratification using heart rate turbulence and ventricular arrhythmia in MADIT II: usefulness and limitations of a 10 min holter recording *Ann. Noninvasive Electrocardiol.* **9** 270–9
Bernardi L, Wdowczyk-Szulc J, Valenti C, Castoldi S, Passino C, Spadacini G and Sleight P 2000 Effects of controlled breathing, mental activity and mental stress with or without verbalization on heart rate variability *J. Am. Coll. Cardiol.* **35** 1462–9
Campana L M, Owens R L, Clifford G D, Pittman S D and Malhotra A 2010 Phase-rectified signal averaging as a sensitive index of autonomic changes with aging *J. Appl. Physiol.* **108** 1668–73
Cerutti S, Bianchi A M and Mainardi L T 1995 Spectral Analysis of Heart Rate Variability Signal *Heart Rate Variability* ed M Malik and A J Camm (Armonk, NY: Futura) pp 63–74
Clifford G D 2002 Signal processing methods for heart rate variability *PhD Thesis* University of Oxford
Clifford G D and Tarassenko L 2004 Segmenting cardiac-related data using sleep stages increases separation between normal subjects and apnoeic patients *Physiol. Meas.* **25** 27–35
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
Clifford G D, Azuaje F and McSharry P 2006 *Advanced Methods And Tools for ECG Data Analysis* (Norwood, MA: Artech House, Inc.)
Clifford G D, McSharry P E and Tarassenko L 2002 Characterizing artefact in the normal human 24 h *RR* time series to aid identification and artificial replication of circadian variations in human beat to beat heart rate using a simple threshold *Comput. Cardiol.* **29** 129–32
Cooper A B, Thornley K S, Young G B, Slutsky A S, Stewart T E and Hanly P J 2000 Sleep in critically ill patients requiring mechanical ventilation *Chest* **117** 809–18
Costa M, Goldberger A L and Peng C-K 2002 Multiscale entropy analysis of complex physiologic time series *Phys. Rev. Lett.* **89** 068102
Costa M, Goldberger A L and Peng C-K 2005 Multiscale entropy analysis of biological signals *Phys. Rev.* E **71** 021906
Eckberg D L 1997 Sympathovagal balance *Circulation* **96** 3224–32
Engelse W A H and Zeelenberg C 1979 A single scan algorithm for QRS-detection and feature extraction *Comput. Cardiol.* **6** 37–42
Goldberger A, Amaral L, Glass L, Hausdorff J, Ivanov P, Mark R, Mietus J, Moody G, Peng C-K and Stanley H 2000 PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals *Circulation* **101** e215–20
Haaksma J, Dijk W, Brouwer J, Van den Berg M, Dassent W, Mulder B and Crijns H 1998 The influence of recording length on time and frequency domain analysis of heart rate variability *Comput. Cardiol.* **25** 377–80
Hilton M F, Bates R A, Godfrey K R and Cayton R M 1998 A new application for heart rate variability: diagnosing the sleep apnoea syndrome *Comput. Cardiol.* **25** 1–4
Johnson A E, Behar J, Andreotti F, Clifford G D and Oster J 2014 *R*-peak estimation using multimodal lead switching *Comput. Cardiol.* **41** 281–4
Johnson A E, Behar J, Andreotti F, Clifford G D and Oster J 2015 Multimodal heart beat detection using signal quality indices *Physiol. Meas.* **36** 1665

Kantelhardt J W, Bauer A, Schumann A Y, Barthel P, Schneider R, Malik M and Schmidt G 2007 Phase-rectified signal averaging for the detection of quasi-periodicities and the prediction of cardiovascular risk *Chaos* **17** 015112
Kaplan D and Staffin P 1998 *Software for Heart Rate Variability* (www.macalester.edu/~kaplan/hrv/doc/index.html) (Accessed: 1 September 2016)
Kisohara M, Stein P K, Yoshida Y, Suzuki M, Iizuka N, Carney R M, Watkins L L, Freedland K E, Blumenthal J A and Hayano J 2013 Multi-scale heart rate dynamics detected by phase-rectified signal averaging predicts mortality after acute myocardial infarction *Europace* **15** 437–43
Li Q and Clifford G D 2012 Signal quality and data fusion for false alarm reduction in the intensive care unit *J. Electrocardiol.* **45** 596–603
Li Q, Liu C, Li Q, Shashikumar S P, Nemati S, Shen Z and Clifford G D 2018 Ventricular beat detection using a wavelet transform and a convolutional neural network *Physiol. Meas.* submitted
Li Q, Liu C, Oster J and Clifford G D 2016 *Signal Processing and Feature Selection Preprocessing for Classification in Noisy Healthcare Data Machine Learning for Healthcare Technologies* ed D A Clifton (Hertfordshire, England: IET) pp 33–58
Li Q, Mark R G and Clifford G D 2008 Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a Kalman filter *Physiol. Meas.* **29** 15–32
Li Q, Rajagopalan C and Clifford G D 2014 Ventricular fibrillation and tachycardia classification using a machine learning approach *IEEE Trans. Biomed. Eng.* **61** 1607–13
Lin Y-H *et al* 2014 Multi-scale symbolic entropy analysis provides prognostic prediction in patients receiving extracorporeal life support *Crit. Care* **18** 548
Liu C Y, Oster J, Reinertsen E, Li Q, Zhao L N, Nemati S and Clifford G D 2018 A comparison of entropy approaches for AF discrimination *Physiol. Meas.* **39** 074002
Lobmaier S M, Ortiz J U, Schmidt G and Schneider K 2015 Phase-rectified signal averaging method to predict intermediate perinatal outcomes in infants with very preterm fetal growth restriction–a secondary analysis of TRUFFLE-trial *Z. Geburtshilfe Neonatologie* **219** FV01_1
Lomb N R 1976 Least-squares frequency analysis of unequally spaced data *Astrophys. Space Sci.* **39** 447–62
Malik M and Camm A J 1995 *Heart Rate Variability* (Armonk, NY: Futura Pub. Co. Inc.)
McSharry P and Malamud B 2005 Quantifying self-similarity in cardiac inter-beat interval time series *Comput. Cardiol.* **32** 459–62
McSharry P E, Clifford G D, Tarassenko L and Smith L A 2002 Method for generating an artificial *RR* tachogram of a typical healthy human over 24 h *Comput. Cardiol.* **29** 225–8
McSharry P E, Clifford G D, Tarassenko L and Smith L A 2003 A dynamical model for generating synthetic electrocardiogram signals *IEEE Trans. Biomed. Eng.* **50** 289–94
Mietus J E and Goldberger A L 2014 Heart Rate Variability Analysis with the HRV Toolkit—PhysioNet https://physionet.org/tutorials/hrv-toolkit/ (Accessed: August 2016)
Mølgaard H 1991 Evaluation of the Reynolds Pathfinder II system for 24 h heart rate variability analysis *Eur. Heart J.* **12** 1153–62
Moody G and Mark R 2001 The impact of the MIT-BIH Arrhythmia Database *IEEE Eng. Med. Biol.* **20** 45–50
Moody G B 2015a WFDB Applications Guide www.physionet.org/physiotools/wag/ (Accessed: August 2016)
Moody G B 2015b WFDB Programmer's Guide https://physionet.org/physiotools/wpg/ (Accessed: March 2017)
Norris P R, Stein P K and Morris J A 2008 b Reduced heart rate multiscale entropy predicts death in critical illness: A study of physiologic complexity in 285 trauma patients *J. Crit. Care* **23** 399–405
Norris P, Anderson S, Jenkins J, Williams A and Morris J 2008a Heart rate multiscale entropy at three hours predicts hospital mortality in 3154 trauma patients *Shock* **30** 17–22
Oster J and Clifford G D 2015 Impact of the presence of noise on *RR* interval-based atrial fibrillation detection *J. Electrocardiol.* **48** 947–51
Otzenberger H, Gronfier C, Simon C, Charloux A, Ehrhart J, Piquard F and Brandenberger G 1998 Dynamic heart rate variability: a tool for exploring sympathovagal balance continuously during sleep in men *Am. J. Physiol.* **275** H946–50
Pan Q, Zhou G and Wang R 2016 Do the deceleration/acceleration capacities of heart rate reflect cardiac sympathetic or vagal activity? A model study *Med. Biol. Eng. Comput.* **54** 1921–33
Peng C K, Havlin S, Stanley H E and Goldberger A L 1995 Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series *Chaos* **5** 82–7
Pinedo M, Villacorta E, Tapia C, Arnold R, López J, Revilla A, Gómez I, Fulquet E and San Román J A 2010 Inter-and intra-observer variability in the echocardiographic evaluation of right ventricular function *Rev. Esp. Cardiol.* **63** 802–9
Press W H, Teukolsky S A, Vetterlin W T and Flannery B P 1992 *Ch 13: Fourier and Spectral Applications Numerical Recipes in C: The Art of Scientific Computing* 2nd edn (Cambridge: Cambridge University Press) pp 537–608
Richman J S and Moorman J R 2000 Physiological time-series analysis using approximate entropy and sample entropy *Am. J. Physiol. Heart Circ. Physiol.* **278** H2039–49
Scargle J D 1982 Studies in astronomical time series analysis. II-Statistical aspects of spectral analysis of unevenly spaced data *Astrophys. J.* **263** 835–53
Schmidt G, Malik M, Barthel P, Schneider R, Ulm K, Rolnitzky L, Camm A J, Bigger J T Jr and Schömig A 1999 Heart-rate turbulence after ventricular premature beats as a predictor of mortality after acute myocardial infarction *Lancet* **353** 1390–6
Sparrow J M, Ayliffe W, Bron A J, Brown N P and Hill A R 1988 Inter-observer and intra-observer variability of the Oxford clinical cataract classification and grading system *Int. Ophthalmol.* **11** 151–7
Spool J 2011 *Do Users Change Their Settings? User Interface Engineering (blog)* (https://archive.li/X9hdp) (Accessed: 20 June 2018)
Sun J X 2006 Cardiac output estimation using arterial blood pressure waveforms *PhD Thesis* Massachusetts Institute of Technology
Sun J, Reisner A, Saeed M and Mark R 2005 Estimating cardiac output from arterial blood pressure waveforms: a critical evaluation using the MIMIC II database *Comput. Cardiol.* **32** 295–8
Tarvainen M P, Niskanen J-P, Lipponen J A, Ranta-Aho P O and Karjalainen P A 2014 Kubios HRV–heart rate variability analysis software *Comput. Methods Programs Biomed.* **113** 210–20
Task Force of the European Society of Cardiology the North American Society of Pacing Electrophysiology 1996 Heart rate variability: standards of measurement, physiological interpretation, and clinical use *Circulation* **93** 1043–65
Vest A N, Li Q, Liu C, Nemati S, Shah A and Clifford G D 2017 Benchmarking heart rate variability toolboxes *J. Electrocardiol.* **50** 744–7
Vest A N, Poian G D, Li Q, Liu C, Nemati S, Shah A and Clifford G D 2018 *PhysioNet Cardiovascular Signal Toolbox* (https://doi.org/10.5281/zenodo.1243111)
Vollmer M 2018 HRVTool https://marcusvollmer.github.io/HRV/ (Accessed: 30 March 2017)
Willson K, Francis D P, Wensel R, Coats A J and Parker K H 2002 Relationship between detrended fluctuation analysis and spectral analysis of heart-rate variability *Physiol. Meas.* **23** 385
Zhu T, Johnson A E, Behar J and Clifford G D 2014 Crowd-sourced annotation of ECG signals using contextual information *Ann. Biomed. Eng.* **42** 871–84
Zong W, Moody G and Jiang D 2003 A robust open-source algorithm to detect onset and duration of QRS complexes *Comput. Cardiol.* **30** 737–40