Title: An Open Source Benchmarked Toolbox for Cardiovascular Waveform and Interval Analysis

Authors: Adriana N. Vest[1,2], Giulia Da Poian[1], Qiao Li[1], Chengyu Liu[1], Shamim Nemati[1], Amit Shah[2,4], Gari D. Clifford[1,5]

Contact: gari@physionet.org

[1] Department of Biomedical Informatics, Emory University School of Medicine
Woodruff Memorial Research Bldg, Suite 4100, 101 Woodruff Circle, Atlanta, GA, 30322

[2] Department of Epidemiology, Rollins School of Public Health at Emory University
1518 Clifton Road NE, Room 3053, Atlanta, GA,30322

[3] Department of Electrical and Computer Engineering, Georgia Institute of Technology
266 Ferst Drive, Atlanta, GA, 30332

[4] Division of Cardiology, Department of Medicine, Emory University School of Medicine
101 Woodruff Circle Suite, 319 Woodruff Memorial Research Building, Atlanta, GA 30322

[5] Department of Biomedical Engineering, Georgia Institute of Technology
1760 Haygood Drive, HSRB Suite W200, Atlanta, GA, 30322

Corresponding Author Gari D. Clifford, gari@physionet.org

Phone: 251-459-9342, fax: 509-695-2437

An Open Source Benchmarked Toolbox for Cardiovascular Waveform and Interval Analysis

Adriana N. Vest[1,2], Qiao Li[1], Chengyu Liu[1], Shamim Nemati[1], Giulia Da Poian[1], Amit Shah[2,4], Gari D. Clifford[1,5]
[1] Department of Biomedical Informatics, Emory University School of Medicine
Woodruff Memorial Research Bldg, Suite 4100, 101 Woodruff Circle, Atlanta, GA, 30322
[2] Department of Epidemiology, Rollins School of Public Health at Emory University
1518 Clifton Road NE, Room 3053, Atlanta, GA,30322
[3] Department of Electrical and Computer Engineering, Georgia Institute of Technology
266 Ferst Drive, Atlanta, GA, 30332
[4] Division of Cardiology, Department of Medicine, Emory University School of Medicine
101 Woodruff Circle Suite, 319 Woodruff Memorial Research Building, Atlanta, GA 30322
[5] Department of Biomedical Engineering, Georgia Institute of Technology & Emory University,
1760 Haygood Drive, HSRB Suite W200, Atlanta, GA, 30322

E-mail: gari@physionet.org

Abstract. We have compiled a comprehensive and open-source modular toolbox for calculating heart rate variability (HRV) metrics and other related variability indices, implemented in Matlab with evidence-based algorithms and output formats. Variability metrics hold promise as potential indicators for autonomic function, prediction of adverse cardiovascular outcomes, psychophysiological status, and general wellness. Although the investigation of HRV has been prevalent for several decades, the methods used for preprocessing, windowing, and choosing appropriate parameters lacks consensus among academic and clinical investigators. The functioning of our software, the PhysioNet Cardiovascular Signal Toolbox (this work), is compared with other widely used and referenced HRV toolboxes. Our findings demonstrate how differences in the methodology of HRV analysis can lead divergent results, a factor that might have contributed to the lack of repeatability of studies and clinical applicability of HRV metrics. Existing HRV toolboxes do not include standardized preprocessing, signal quality indices 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 for performing preprocessing, signal quality, and arrhythmia detection. We also make recommendations for default values and reporting.

Key Terms. Heart rate variability, toolbox validation, peak detection, physiological signal processing

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 (Malik, 1996; Clifford, 2002; Pan *et al.*, 2016) and consequently 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 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 by which researchers apply HRV signal processing. This issue limits meaningful comparisons between studies and scientific repeatability, especially when in-house, custom, non-public software are used. Unfortunately, few HRV programs are rigorously designed and tested with methods that are clear and open access. 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. More specifically, such software should undergo a validation process in which the output is rigorously compared with expected values based on a standardized input; furthermore, it should be tested against other HRV tools for consistency.

To address the issues of validation, standardization, and repeatability, we have developed a validated open-source cardiovascular signal and HRV analysis toolbox. The software suite 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 includes many features not offered in other programs, including peak and pulse detection, signal quality analysis, rhythm detection, beat classification, standard HRV statistics, and more recent HRV metrics, such as phase rectified signal averaging (PRSA). The toolbox is written in the Matlab programming language and does not have any dependencies on external software or libraries. (A list of minimal default Matlab toolboxes are provided in Appendix A - the toolbox was designed to use the minimal number of dependencies and the most basic operators to future-proof the code base as much as possible.) The toolbox can process raw waveform data (such as electrocardiograms) as well as derived RR-interval data. Although the toolbox was not designed to deal with file formats (to avoid specific dependencies), it does natively support MAT, CSV, or WFDB-compatible annotation formats without relying on PhysioNet's WFDB libraries (or other libraries). If users wish to export results from the HRV Toolbox, a function is included that allows for standard WFDB compatible output annotation files or vanilla CSV output files.

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 described 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. In particular, our toolbox contains one initialization (or header) file which lists all the options available, with typical default settings. In this way, a user may easily identify which settings need to be given considerable thought (all the ones listed) and provide this listing in a publication. This file fully defines all the parameter options to ensure a fully repeatable experiment.

The goal of this work is to advance the standardization of HRV and cardiovascular variability research and clinical applicability. This publication outlines the current HRV analysis tool landscape alongside our new suite of open-source tools contained within the PhysioNet Cardiovascular Signal Toolbox. We present 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, we present a standard model by which HRV analysis packages may be judged in the future and 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 review a subset of the most popular toolboxes available and the HRV metrics that they generate. Perhaps the most used and trusted HRV toolbox is that written by Mietus and Moody, available from PhysioNet.org (Mietus and Goldberger, 2014). The PhysioNet HRV Toolkit is an open-source package that is written in C and performs time domain and spectral HRV statistics. This toolbox has the unique feature of compatibility with the 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. Never-the-less, 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 does not support batching input data, and can therefore be time consuming for moderate sized datasets and unfeasible for large datasets. Moreover, the fact that users can set the many preprocessing parameters by hand, means the results may be unrepeatable (since humans are prone to errors when clicking buttons in a repeated manner, and the documentation of the exact parameters selected may not be fully recorded). Two less commonly referenced Matlab-based toolboxes available are Kaplan's HRV toolbox (Kaplan and Staffin, Updated Feb 3 1998) and Vollmer's HRV toolbox (Vollmer). 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 RR intervals, the standard deviation of normal-to-normal (NN) RR 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 the pNNx (where x is a variable between 5 and 100ms) 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.* (Clifford *et al.*, 2006) for a detailed description of these statistics.

Table 1. Summary of Functionality of Various HRV Toolboxes. See section 3 for definition of HRV metrics

Software Origin → **↓Functionality**	**PhysioNet HRV Toolkit** (v 10.5.24) (Last Update Aug 4 2009)	**PhysioNet Cardiovascular Signal Toolbox** (v 1.0)	**Kubios** (v 3.0.2) (Premium)	**Kaplan** (Last Update Feb 3, 1998)	**Vollmer** (v 0.98)
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 are greater than 20% different than the average interval measured on 20 beats before and after beat in question are removed	Intervals that vary greater than 20% from the preceding interval are removed	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

It is worth noting that the PhysioNet Cardiovascular Signal Toolbox includes more recent HRV statistics which have been shown to be highly predictive (MSE, PRSA), and versions of existing metrics which have been shown to be more computationally efficient (sampEn, ApEn), and important feature for Matlab versions, and more accurate (DFA). We have also ignored the less well-founded/more ad-hoc statistics, such as the TINN (which is just a poor estimate of a distribution) in favor of more acceptable statistics (such as the first four moments).

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.

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 consults and 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 a 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 (usually denoted by an '*atr*' file extension on PhysioNet, but not always). 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 your 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 1 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. Noisy data (below some defined threshold – 0.9 or 0.7 for example) must be removed from the analysis (*HRVparam*.sqi.LowQualityThreshold = 0.9 by default).
5. Before calculating HRV statistics, arrhythmic periods of data must be removed.
6. If desired ventricular fibrillation/ventricular tachycardia (VF/VT) can be detected on the waveform based on the method discussed in section III.B.3. 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.

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

For frequency domain calculations, the power spectral density (PSD) of the RR interval time series can be generated using several methods. Those methods include: the Lomb Periodogram, the Welch PSD estimate, the Burg PSD estimate, and the discrete or fast Fourier transform. An option to resample the RR interval time series is provided to users since the methods other than the Lomb Periodogram assume that the time series is uniformly sampled. All PSD estimates calculated by the HRV Toolbox described here can accept frequency bin delineation, which improves control over the reproducibility of the resulting 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 RR 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.

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. (Clifford *et al.*, 2006).

A set of demo files (listed in Appendix D) are made available to the user for testing the toolbox and verify 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* (Moody, 2015a; Zong *et al.*), 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* (Moody, 2015a), sqrs.c(Moody, 2015a), and *gqrs.c* (Moody, 2015a), available from 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, but an overview of the approximate trigger locations on the ECG is shown in Figure 1.

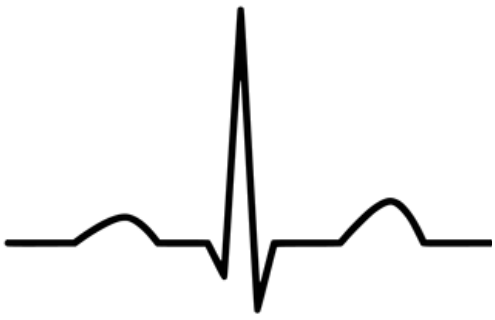

Figure 1. Typical trigger points of each QRS detector.

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. (Li *et al.*, 2008), *bsqi* provides the percentage of beats that match when detected by multiple annotation generators with highly differing noise responses. The signal quality index (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. (Li *et al.*, 2014), *VF_Classification*. 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 fourteen different VF features. After training, the model was optimized for use of only two features on 5 second windows, thus the classification algorithm is rapid and provides real time operation of VF detection.

3.2.4. PVC Classification

Premature ventricular contraction (PVC) 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.*, In Submission). The WT is used to map short segments of a single channel (1-D) ECG waveform into a 2-D 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, Accessed 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) as described below.
(It should also be noted that for metrics that are sensitive to missing data (such as those involving Fourier analysis and resampling), it is important to remove both associated RR intervals and insert a phantom beat at the point where the RR interval would have been 'expected' to be under sinus rhythm. For intervals associated with noise, the interval can simply be removed and the adjacent intervals recalculated accurately.)

3.2.5. RR Interval Preprocessing Routines

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.

3.2.6. 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.*, In Submission; 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.

3.2.7. Non-sinus beat identification and removal/replacement

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 work we chose N to be 1 and a threshold of 20 %. 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 normal sinus rhythm 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 III.C.4).

Additional checks and corrections include flagging and removing non-physiologic data (RR intervals above 2 seconds or below .375 seconds, outside of physiologically possible range) and data that is labeled as non-normal per a supplied annotation file (if applicable).

3.2.8. 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 interreader variability have been measured to be greater than 20% (Zhu *et al.*, 2014; Sparrow *et al.*, 1988; Pinedo *et al.*, 2010). The impact of this variability has not been measured for HRV, and is not within the scope of this paper, since we are specifically attempting to remove such subjective vagueness from HRV analysis through the publication of this toolbox. We explicitly advise against 'expert' or 'hand' modification of data, since it destroys scientific repeatability of the research.

3.2.9. Resampling

Re-sampling the RR interval time series involves interpolating through the signal (such as by linear or cubic spline interpolation) and re-sampling at regular intervals specified by the resampling frequency. Most of the papers in the field of HRV report on the use of re-sampling rates between 1 Hz and 10 Hz (Malik and Camm, 1995; Hilton *et al.*, 1998; Malik, 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 re-sampling rate of 4 Hz is sufficient.

Re-sampling 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.

3.2.10. 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.

Error in the PSD estimate and frequency domain metrics grows linearly with the amount of data removed when interpolation is used prior to taking an 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). Linear interpolation is more susceptible to generating erroneous results with small amounts of artefactual data. On the other hand, cubic spline interpolation, while creating a more smoothly resampled time series, can become unstable at moderate levels of missing data and lead to unconstrained oscillations between nodes in the spline, which artificially elevate the HF power (Clifford, 2002; McSharry *et al.*, 2003). FFT-based time domain techniques are therefore highly susceptible to noise, ectopy and missing data, as shown in Clifford et al. (Clifford, 2002; McSharry *et al.*, 2003). Moreover, Clifford and Tarassenko (Clifford and Tarassenko, 2004) 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(Scargle, 1982; Lomb, 1976; Press *et al.*, 1992)) are far superior. Previous studies have shown 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). 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. 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 is it 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.

3.3. Parameter Selection

3.3.1. Length of Data

The user needs to decide if a long term (~24 hours or longer) or short term (~5 minutes) recording is desired. (This can be done 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 (Malik, 1996) and short term metrics have been shown to be capable of assessing neurological activity (Malik, 1996; Malik and Camm, 1995). Confounders for long term metrics HRV can include temperature (Malik and Camm, 1995), quality of sleep (Cooper et al., 2000), and large gaps in data and short term HRV can by influenced by changes in mental, emotional, or physical state. 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.

3.3.2. Window Size Depends on the HRV Metric Being Calculated

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's 1998 study led to recommendations of 20 hours 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 (Malik, 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 (Malik, 1996) suggested that the shortest time period over which HRV metrics should be assessed is 5 minutes. This results in a limitation of the lowest frequency that can be resolved being $\frac{1}{300} \approx 0.003$ Hz (just above the lower limit of the VLF region). In practice the limit is higher since noise affects the estimation. A 5-minute 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-minute window (T = 300) leads to the constraint of $\frac{N}{2T} \geq 0.4$ 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-minute 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.

3.3.3. Frequency Bands for Spectral Content Estimation

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

ULF – Ultra Low Frequency: 0.0001 Hz $\leq$ ULF $<$ 0.003 Hz

VLF – Very Low Frequency: 0.003 Hz ≤ VLF < 0.04 Hz

LF – Low Frequency: 0.04 Hz ≤ LF < 0.15 Hz

HF – High Frequency: 0.15 Hz ≤ HF < 0.4 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 youth 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) while the LF band contains 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.

3.3.4. Normalization Method

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

3.4. Long Range Scaling Metrics: DFA and MSE

3.4.1. 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-intervals 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, α=(β+1)/2, is by using the DFA, which is briefly summarized in the following paragraph. (For a detailed description see Peng et al. (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:

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

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 minimal 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 α1

and α2. (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)

3.4.2. Multiscale Entropy

Multiscale entropy (MSE) analysis was first introduced by Costa et al. (Costa *et al.*, 2005, 2002) 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 N/τ, where N 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 τ))

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*).

3.5. Phase-Rectified Signal Averaging

Phase-rectified signal averaging (PRSA) is a recently introduced method 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. (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.

3.6. Heart Rate Turbulence

Heart Rate Turbulence (HRT) is a method used to analyze the fluctuations in sinus-rhythm cycle length after ventricular premature complexes (VPCs or PVCs)(Schmidt et al., 1999; Bauer et al., 2008). Two parameters are used to characterize the response of sinus rhythm to a VPC: the Turbulence Onset (TO) and Turbulence Slope (TS). TO is used as a measure of the initial acceleration after the VPC, and it is derived by comparing the relative changes of RR intervals immediately after and before a VPC:

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

where RR+ is the i-th sinus rhythm after the compensatory pause of the VPC and RR- indicates the coupling interval of the VPC. The TO value is first computed for each single VPC and subsequently averaged to 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 Turbulence Slope (TS), quantifies the deceleration rate after a VPC. TS is the maximum positive slope of a regression line assessed over any sequence of 5 subsequent sinus rhythm RR intervals within the first 20 sinus rhythm RR intervals after a VPC (Bauer et al., 2008).

TO values below 0 and TS values above 2.5 are considered normal, and abnormal otherwise (i.e., a healthy response to VPCs 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 VPC tachogram, obtained by aligning and averaging the R-R interval sequences surrounding isolated VPCs, for a sufficient number of VPCs (i.e., >5) (Bauer et al., 2008). Despite this accumulation of data around numerous VPCs, 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 VPC 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 VPCs 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[1], the number of NN intervals to consider before the VPC (*BeatsBefore*), and after the PVC plus a compensatory pause (*BeatsAfter*). The function also returns the number and position of the VPCs used for the analysis, the average tachogram, and the graphical representations of the HRT analysis shown in Figure 2. When computing the average tachogram or the mean TO, the user should aim for include a minimum of 15 to 20 VPC tachograms.

[1] https://www.physionet.org/physiobank/annotations.shtml

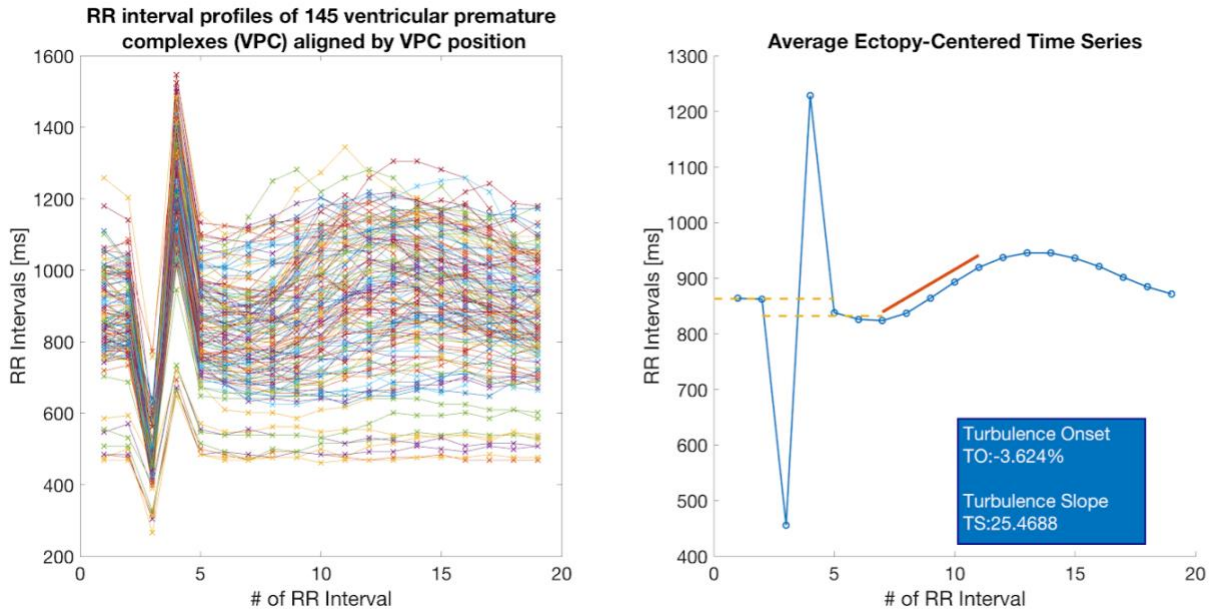

Figure 2. Visualization of HRT analysis using the PhysioNet Cardiovascular Signal Toolbox.
Left figure shows an example of NN (normal to normal) interval sequences in one patient, all aligned at the VPC; the Turbulence Onset is computed for each single VPC and subsequently averaged to obtain the value characterizing the subject . Right figure shows the average tachogram used to compute the Turbulence Slope (TS). TS is the maximum of regression slopes computed for 5 consecutive NN sequences. In the example, the regression lines for beats 7~11 corresponding to maximum slope is shown (red line).

4. Methods

A series of benchmarking studies were conducted on sample data using the PhysioNet Cardiovascular Signal Toolbox and the four other HRV toolboxes described in Table 1. 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

Study A compares the results generated by each toolbox on one HRV metric, the LF/HF ratio, using a known standard value. The standard LF/HF ratio is generated using an RR interval generator detailed in Clifford (Clifford, 2002), hereafter called LFHFGEN. The default options for each toolbox are used to simulate the results achieved by a typical user of the HRV toolboxes.

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 PhysioNet routine that calculates the root mean squared error and normalizes it per the equation

$$NRMSE = 100 * \sqrt{\frac{\sum_{i=1}^{n}(X_{Test} - X_{Standard})^2}{n}} \Big/ \frac{1}{n}\sum_{i=1}^{n} X_{Standard}$$

where n equals the number of windows considered, X_{Test} is the metric generated by the test toolbox on the i^{th} window, and $X_{Standard}$ is standard compared against. The NRMSE value is reported back as a percentage. Default parameters and settings for each toolbox (per Table 3) were used unless otherwise specified in the Methods.

100 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 .095 Hz and .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 is 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 **3** to estimate the LF/HF ratio and the normalized RMSE was calculated using a standard that is found before frequency shifting or downsampling.

4.2. Study B: The significance of collective processing differences

Study B expands the comparative analysis performed in Study A to include a wider selection of commonly assessed HRV metrics and their performance 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 are 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 is 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

100 segments of synthetic RR interval data were generated using *RRGEN* (a method developed by McSharry et al. (McSharry *et al.*, 2002; McSharry *et al.*, 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 Normal Sinus Rhythm Database RR Interval Data Analysis

All 18 RR interval records from the MIT NSR database were segmented into 5 minute windows with 4 minutes 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 3.2.6) 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 Normal Sinus Rhythm 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-hour 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 and third comparisons added the effects of assigning frequency bins using the PhysioNet Cardiovascular Signal Toolbox method (Comparison 2) and preprocessing using the PhysioNet Cardiovascular Signal Toolbox method (Comparison 3) respectively. An overview of the differences between the comparisons can be seen in Table 2.

Table 2. A summary of the stepwise comparison performed between the PhysioNet HRV Toolkit and the PhysioNet Cardiovascular Signal Toolbox.

	Comparison 1	**Comparison 2**	**Comparison 3**
Preprocessed with:	PhysioNet HRV Toolkit	PhysioNet HRV Toolkit	PhysioNet HRV Toolkit & PhysioNet Cardiovascular Signal Toolbox
Frequency Bins Assigned in:	PhysioNet HRV Toolkit	PhysioNet HRV Toolkit & PhysioNet Cardiovascular Signal Toolbox	PhysioNet HRV Toolkit & PhysioNet Cardiovascular Signal Toolbox
HRV calculated with:	PhysioNet HRV Toolkit & PhysioNet Cardiovascular Signal Toolbox	PhysioNet HRV Toolkit & PhysioNet Cardiovascular Signal Toolbox	PhysioNet HRV Toolkit & PhysioNet Cardiovascular Signal Toolbox

4.2.3. Trial 3: Waveforms of the MIT Arrhythmia Database

All 48 records from the MIT Arrhythmia (BIH) Database were processed using the waveform analysis methods in the toolboxes which have this functionality (PhysioNet HRV Toolkit, Kubios, PhysioNet Cardiovascular Signal Toolbox, and Vollmer). Each roughly 30 min record was broken up into 5 min segments with 4 min of overlap between them and then each segment was analyzed for HRV metrics. Segments from all 48 records were compiled and NRMSE was computed on the compiled segments.

Table 3. Differences between HRV analysis methods in the 5 HRV toolboxes benchmarked. Default options were selected.

	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	Identify successive intervals whose difference exceed	Identify successive intervals whose difference	'Glitches' identified using AR model	None

	the value of adjacent 20 intervals on either side); Identify Non-physiologic Intervals (RR < 0.4 s) (RR > 2 s)	a threshold (20%); Identify PVCs, AF, VF; Identify Non-physiologic Intervals (RR < 0.375 s) (RR > 2 s)	exceed a threshold (20%)		
Artifact Correction Method	Remove Non-physiologic RR intervals and intervals that exceed 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 threshold	Spline Interpolation through data labeled "glitches"	None
Frequency Vector (Hz)	$df : df : df \times 2 \times L$ $df = \frac{1}{4 \times (t_{max} - t_{min})}$ $L = data\ length$	$\mathbf{\frac{1}{1024} : \frac{1}{1024} : 0.5}$	$\mathbf{\frac{1}{300} : \frac{1}{300} : 0.5}$	$\mathbf{\frac{Fs}{L}(0:1:\frac{L}{2})}$	$\mathbf{\frac{Fs}{2}(0:1:\frac{NFFT}{2}+1)}$ NFFT = 2^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	(Verify)	Squares then Doubles FFT and Sums Bin in Band	Doubles FFT and Sums Bins in Band
Normalization	$\sqrt{\frac{PSD}{nout}}$ nout = 0.5 * ofac * hifac * n ofac = 4, hifac = 2	$\sqrt{\frac{PSD}{nout}}$ nout = 0.5 * ofac * hifac * n ofac = 4, hifac = 2	(Verify)	Normalizes to square of length of data segment analyzed	Normalizes to length of data segment analyzed

4.3. Study C: Validation of long range scaling metrics - DFA

The DFA implementation from both the PhysioNet HRV Toolkit and the PhysioNet Cardiovascular Signal Toolbox were used to estimate the scaling exponents of time series with normal distributions according to the experiments reported in (McSharry and Malamud, 2005). Synthetic RR interval data with given scaling coefficients (β) were generated (for β = -2: 0.25 :2, time series length 4096, for each β 20 time series where used).

The standard output for the PhysioNet HRV Toolkit's implementation, *dfa.c*, contains two columns of numbers, which are the base 10 logarithms of *n* and *F(n)*. The program does not compute the scaling exponent. To obtain such a scaling exponent the output has been fitted to a line using custom code written for the PhysioNet Cardiovascular Signal Toolbox. The slope of the line relating log $F(n)$ to log n determines the scaling exponent (self-similarity parameter), α ($4 \leq n \leq N/4$, where N is the signal length in samples).

100 segments of synthetic RR interval data were generated using *RRGEN* (McSharry *et al.*, 2002; McSharry *et al.*, 2003) with the probability of ectopy and noise set to 0 % (Pe = 0, Pn=0). The segments were analyzed in full and were 24 hours long. The same dataset was analyzed for Studies D and E. Default options were used for all the toolboxes.

4.4. Study D: Validation of long range scaling metrics - MSE

The MSE values were computed and compared for 100 stochastic and 100 deterministic signals using the PhysioNet Cardiovascular Signal Toolbox and the PhysioNet HRV Toolkit. The MSE implementation from the PhysioNet HRV Toolkit Toolkit uses a default pattern length $m = 2$ and a similarity criterion $r = 0.15$, the same parameters that are set as default in the PhysioNet Cardiovascular Signal Toolbox. The maximum number of coarse-grained time series is defined by the parameter *maxTau*, which by default is set to be equal to 20. The scaling exponents of synthetic RR interval data were also estimated. A total of one hundred 24-hour synthetic RR tachograms were generated using *rrgen.m* (Pn = 0, Pe=0) and used for the validation of PhysioNet Cardiovascular Signal Toolbox with respect to PhysioNet HRV Toolkit and to Kubios with and without the detrending preprocessing option. (Kubios MSE calculations default to detrending.)

4.5. Study E: Validation of PRSA

The PRSA algorithm from the PhysioNet Cardiovascular Signal Toolbox was evaluated against code provided by the original authors of PRSA (Bauer *et al.*, 2006). One hundred synthetic, 24 hour RR interval time series were generated using *rrgen.m* for Pn = 0, Pe=0 were used for the validation of the code included in the PhysioNet Cardiovascular Signal Toolbox with respect to the one provided by Schmidt et al. (Schmidt *et al.*, 1999).

4.6. Study F: Validation of HRT

The HRT algorithm from the PhysioNet Cardiovascular Signal Toolbox (*HRT_Analysis.m)* was evaluated against HRT code provided by Raphael Schneider (Bauer *et al.*, 2008). The comparison was done on data from the Normal Sinus Rhythm RR Interval Database (Goldberger *et al.*, 2000). Because of differences between the two implementations with preprocessing, both methods of preprocessing were tested (removal of RR intervals that change by more than > 20% with respect to the mean of the five last sinus intervals and removal of RR intervals that change by more than > 20% with respect to the previous one).

5. Results

5.1. Study A

The PhysioNet Cardiovascular Signal Toolbox and Kaplan toolboxes achieve negligible error in the LF/HF ratio, with errors between 3.52 % and 4.98 %. (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).) Kubios's default calculation using FFT achieves 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 % (see Table 4). We note that these errors may be consistent offsets, which, although prevent comparison between studies, can still provide valid comparisons within studies. Never-the-less, we strongly suggest using a toolbox with settings that provides 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.

Table 4. 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 %

5.2. Study B

5.2.1. Trial 1: Synthetic Data

The entire dataset 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 5. 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.

Table 5. The normalized RMS error 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 With Artifact Correction Lomb**	**Kaplan**	**Vollmer**
Mean RR	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 %

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 2,835 segments were not analyzed because AF was detected (2,366 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 6. The PhysioNet Cardiovascular Signal Toolbox operates most closely to the PhysioNet HRV Toolkit, as is seen by its low NRMSE values.

Table 6. 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 Normal Sinus Rhythm 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 RR	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 %

Although large difference exists 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% normalized RMSE of the PhysioNet HRV Toolbox (Table 7) 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 7. The calculated differences between the PhysioNet HRV Toolkit and the PhysioNet Cardiovascular Signal Toolbox as determined by the NRMSE method. Comparison A uses identical settings for both toolboxes. Comparison B introduces the variability due to the different frequency binning methods between the two toolboxes. Comparison C introduces the variability due to preprocessing differences between the two toolboxes. N/A indicates not applicable (because the trial affected only spectral metrics).

Comparison → HRV Metric ↓	A	B	C
Mean NN interval	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 %

5.2.3. Trial 3: Waveform Data

The calculated error between the toolboxes when compared to the results generated by the PhysioNet HRV Toolkit are shown in Table 8. 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 compared with the PhysioNet Cardiovascular Signal Toolbox in terms of mapping to the existing PhysioNet HRV Toolbox.

Table 8. The normalized RMSE 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 RR	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 %
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 %

5.3. Study C: Validation of long range scaling metrics - DFA

Results for the estimation of the scaling exponents of time series with normal distribution and known scaling coefficients are reported in Figure 3.

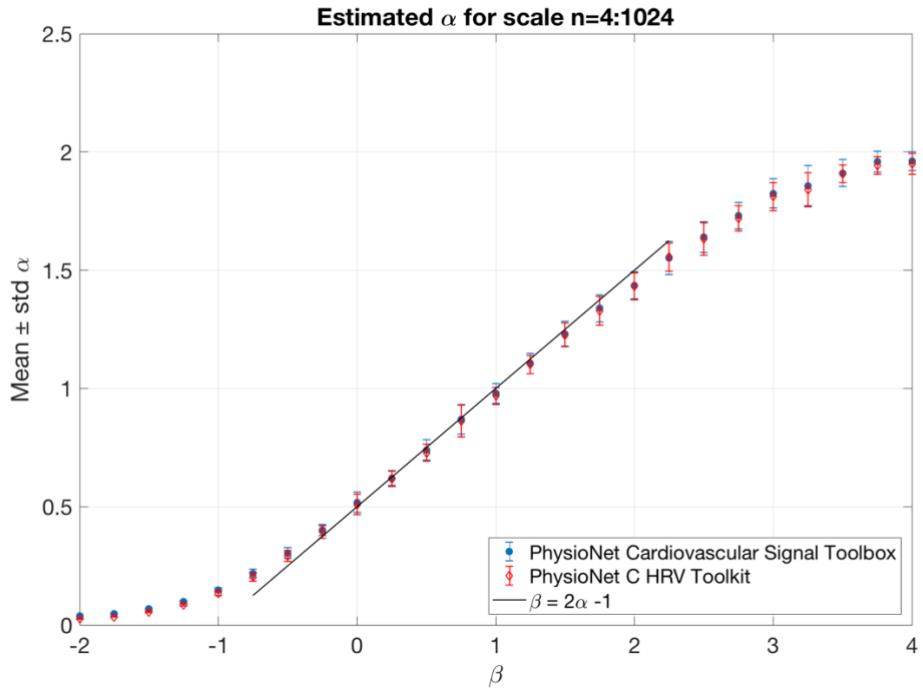

Figure 3. Mean and standard deviation (over 20 time series for each β) of estimated scaling exponents α, calculated using the PhysioNet HRV Toolkit (red diamonds) and the PhysioNet Cardiovascular Signal Toolbox (blue dots), compared to the theoretical exponent β. For both the methods the estimated exponent is linked to the theoretical exponent by β = 2α - 1 for -0.75 <= β <= 2.25. It should be noted that the characteristic RR-interval series is usually in the region between β = 0 and β =2 (McSharry and Malamud, 2005).

The calculated difference between the toolboxes when compared to the results from the PhysioNet HRV Toolkit are shown in Table **9**.

Table 9. The NRMSE generated among different toolboxes on DFA scaling coefficients α1 and α2 compared to the values calculated by the PhysioNet HRV Toolkit.

	PhysioNet Cardiovascular Signal Toolbox	**Kubios (default settings)**	**Kubios (no detrending)**	**Kaplan**
α1*	1.5 %	5.4 %	0.6 %	0.7 %
α2**	3.1 %	34.6%	16.1%	18.6 %
* PhysioNet HRV Toolkit: 4-16; PhysioNet Cardiovascular Signal Toolbox: 4-16; Kubios: 4-16; Kaplan: 4-16 ** PhysioNet HRV Toolkit: 16-N/4; PhysioNet Cardiovascular Signal Toolbox: 16-N/4; Kubios: 16-64; Kaplan: 16-64				

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 fc = 0.035 Hz using default parameters.

Figure **4** highlights the effect of the detrending option on the estimation of α_2.

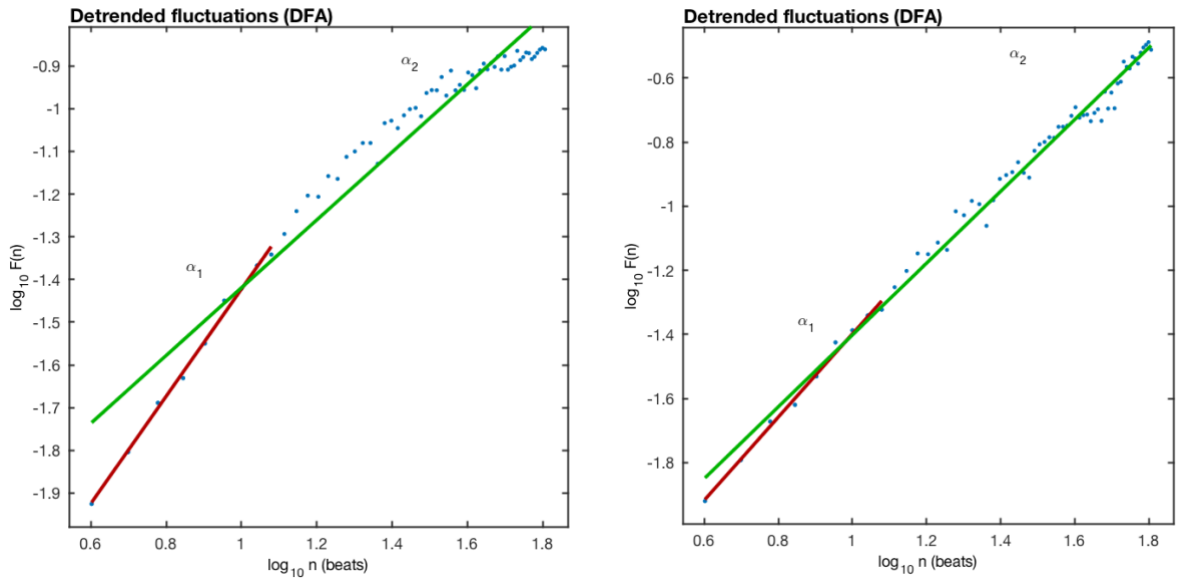

Figure 4. Results from DFA generated using Kubios with detrending and smoothness priors (left; α1 = 1.26, α2 = 0.79) and without detrending (right; α1 = 1.29, α2 = 1.12).

5.4. Study D: Validation of long range scaling metrics - MSE

The comparison of Multiscale Entropy values using the PhysioNet Cardiovascular Signal Toolbox and the PhysioNet HRV Toolkit for 100 stochastic and 100 deterministic signals shows very close correspondence and variance at all scales (Figure 5).

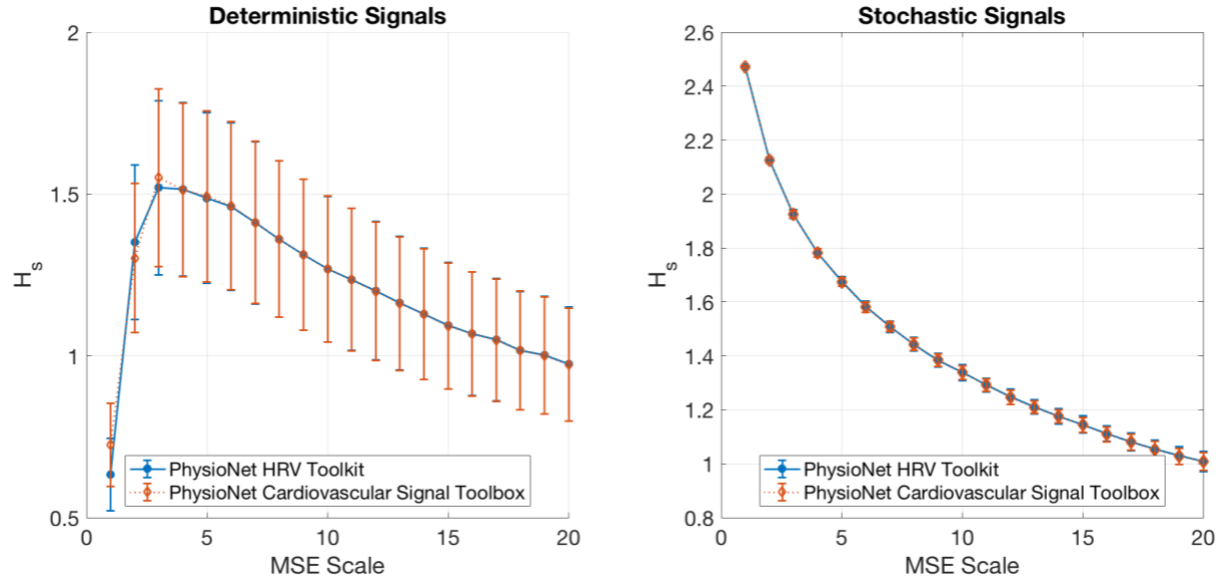

Figure 5. Multiscale Entropy values using PhysioNet Cardiovascular Signal Toolbox and PhysioNet HRV Toolkit for 100 stochastic and 100 deterministic signals.

Results for MSE computed on 24-hour synthetic RR tachograms are shown in Figure 6, which reports the NRMSE for each MSE scale calculated with the PhysioNet Cardiovascular Signal Toolbox in comparison to the MSE scale calculated by the PhysioNet HRV Toolkit. The error was shown to be lower than 0.05% at all scales for the PhysioNet Cardiovascular Signal Toolbox. The Kubios MSE implementation, with and without detrending, shows significantly higher error.

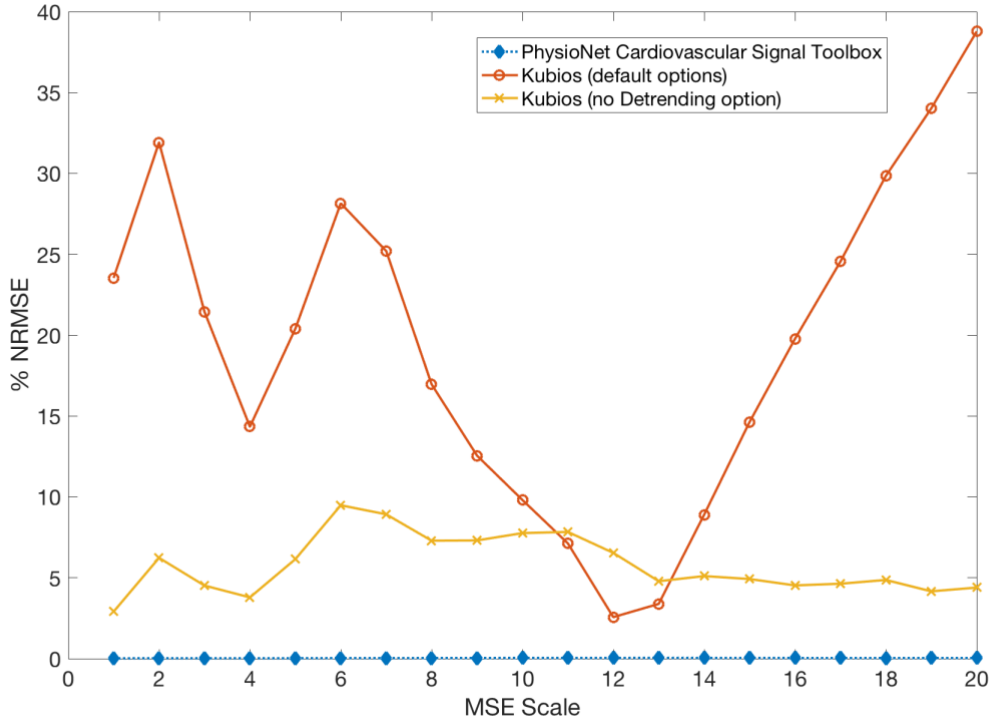

Figure 6. Plot of the NRMSE at different scales of Multiscale Entropy using the PhysioNet Cardiovascular Signal Toolbox, the PhysioNet HRV Toolkit, and Kubios for 100 synthetic RR interval signals generated using *RRGEN* (Pn = 0, Pe=0).

5.5. Study E: Validation of PRSA

Summary results for AC and DC on synthetic RR interval time series are reported in Table 10.

Table 10. The normalized RMS difference generated among different toolboxes on PRSA coefficients DC and AC compared to the values calculated by the PRSA implementation provided by *Bauer et al. (Bauer et al., 2006)*.

	PhysioNet Cardiovascular Signal Toolbox	**Bauer *et al.***	**NRMSD**
DC (ms)	5.9 ± 0.1	5.9 ± 0.1	0.0%
AC (ms)	-6.0 ± 0.1	-6.0 ± 0.1	0.0%

5.6. Study F: Validation of HRT metrics

Comparison of HRT algorithms on the Normal Sinus Rhythm RR Interval 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.

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 VPC tachograms have been 'filtered' by the PhysioNet Cardiovascular Signal Toolbox than the Schneider code. When two or more VPCs are separated by only a small amount of time the rejection is performed differently. The PhysioNet Cardiovascular Signal Toolbox excludes VPCs for which one of the two RR-intervals before the current VPC is a compensatory pause of the previous VPC, while in Schneider's implementation those VPCs tachograms are considered valid.

An example is reported in Figure 7 where the tachogram related to the second VPC contains the compensatory pause (CP) of the preceding VPC. Since a tachogram is considered valid if it has sinus rhythm preceding and following a VPC, both tachograms are excluded by our implementation. TO is computed using the two RR intervals preceding the VPC and following the CP, thus including a CP in the computation of the TO might lead to different results.

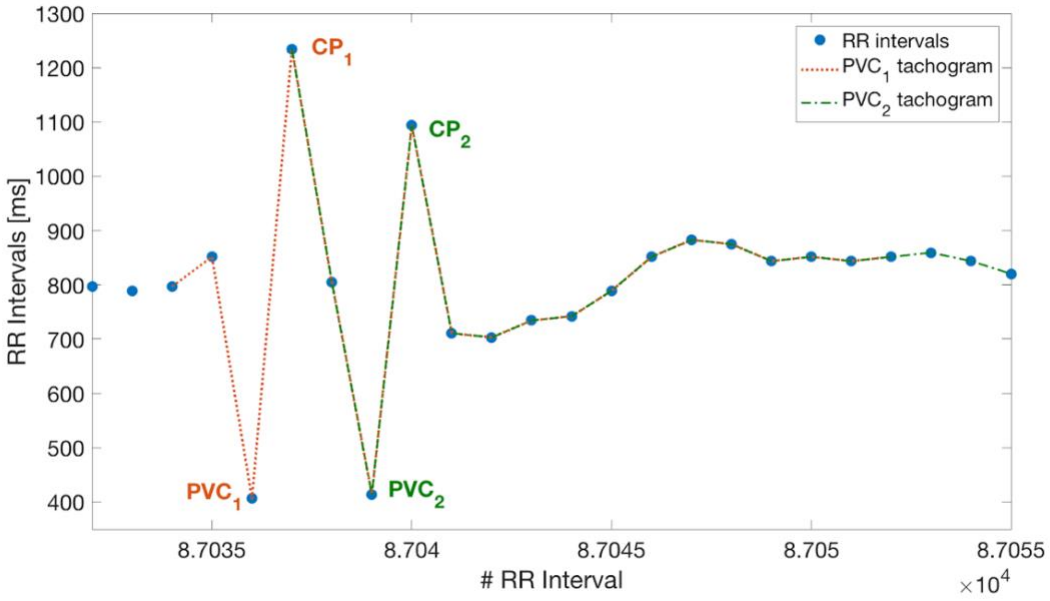

Figure 7. Example of two consecutive VPCs in signal nsr10. Tachogram related to the second VPC contains the compensatory pause CP of the preceding VPC and thus is excluded by the HRT algorithm implemented in the PhysioNet Cardiovascular Signal Toolbox.

6. Discussion

The benchmarking results demonstrate 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 show that it is imperative that future studies adhere to a consistent method of reporting upon how an analysis has been set up. The following discussion will review the effects in greater detail along with some warnings and recommendations from the authors.

6.1. Effects of Preprocessing: Beat Deletion or Insertion

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 7, Comparison C). The calculations of RMSSD and PNN50 are particularly sensitive to noisy RR intervals. When investigating the cause of this error, it was observed that a single window with just one or two removed non-physiologic data points can dramatically affect the estimated value of the RMSE. More markedly, Table 8 shows that even simple time domain statistics can differ by significant amounts when different QRS detectors or abnormal interval filters are employed.

6.2. Effects of Frequency Bin Choice on Spectral Analysis

The effect of differing frequency bins on the results of spectral analysis can 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 2, C Generated Frequency Bins vs Matlab Generated Frequency Bins). The error for the identical power calculations with slightly different frequency bands was nearly 4% at best and 8% at worst. Especially at the Ultra Low (ULF) and Very Low Frequencies (VLF), where the binning may leave these bands with only 1 to 5 bins, changes in those bins can lead to significant differences in the outcome.

6.3. Effects of Normalization of Spectral Metrics

Normalization of the Power Spectral Density estimation is a seldom reported parameter that 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. Notably, in comparing two different (but frequently employed) normalization methods on a common PSD estimation, it was found that LF/HF ratios, a theoretically robust metric to normalization, differed substantially between the two methods.

6.4. Recommendations

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. (A subsequent article in preparation will address the issue of just how much two recordings can differ in length before the metrics become incomparable.) 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).

When using frequency domain analysis, the Lomb Periodogram has been 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 assume 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.

7. Conclusions

This article presents evidence in support of standardizing HRV analysis methods and demonstrates how the PhysioNet Cardiovascular Signal Toolbox achieves such a standardization. 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). 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 should provide a significant boost to its clinical utility. Using in-house code that has not been thoroughly benchmarked and failing to report all parameter settings will continue to hold the field back. 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). We caution against the use of default parameters, particularly when dealing with raw ECG. 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. We note that none of the toolboxes presented are as comprehensive the toolbox described here, and we encourage benchmarked contributions to our software, which is freely available from PhysioNet[2] and Github (Vest *et al.*, 2018).

Acknowledgments

The authors wish to acknowledge the National Institutes of Health (Grant # NIH K23 HL127251) the National Science Foundation Award 1636933, the National Institutes of Health, 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.

References

American Heart Association ECG Database Accessed 2018 (https://www.ecri.org/components/Pages/AHA_ECG_USB.aspx)

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: Statistical Mechanics and its Applications* **364** 423-434

[2] https://physionet.org/physiotools/physionet-cardiovascular-signal-toolbox/

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 *Journal of the American College of Cardiology* **52** 1353-1365

Bauer A and Schmidt G 2003 Heart rate turbulence *Journal of Electrocardiology* **36** 89-93

Behar J, Johnson A, Clifford G D and Oster J 2014 A comparison of single channel fetal ECG extraction methods *Annals of Biomedical Engineering* **42** 1340-1353

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-Minute Holter Recording *Annals of Noninvasive Electrocardiology* **9** 270-279

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 *Journal of the American College of Cardiology* **35** 1462-1469

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 *Journal of Applied Physiology* **108** 1668-1673

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 Pub. Co. Inc.) pp 63-74

Clifford G 2002 Signal processing methods for heart rate variability *Thesis* University of Oxford

Clifford G, McSharry P and Tarassenko L 2002 Characterizing artefact in the normal human 24-hour RR time series to aid identification and artificial replication of circadian variations in human beat to beat heart rate using a simple threshold *Computers in Cardiology, 2002* (Memphis, TN)

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 and Tarassenko L 2004 Segmenting cardiac-related data using sleep stages increases separation between normal subjects and apnoeic patients *Physiological Measurement* **25** 27-35

Clifford G D and Tarassenko L 2005 Quantifying errors in spectral estimates of HRV due to beat replacement and resampling *IEEE Transactions on Biomedical Engineering* **52** 630-638

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** 809-818

Costa M, Goldberger A L and Peng C-K 2002 Multiscale entropy analysis of complex physiologic time series *Physical Review Letters* **89** 068102

Costa M, Goldberger A L and Peng C-K 2005 Multiscale entropy analysis of biological signals *Physical Review E* **71** 021906

Eckberg D L 1997 Sympathovagal Balance *Circulation* **96** 3224-3232

Engelse W A H and Zeelenberg C 1979 A single scan algorithm for QRS-detection and feature extraction *Computers in Cardiology 6:37-42* **6**

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-e220

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 *Computers in Cardiology, 1998* (Cleveland, OH)

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 *IEEE Computers in Cardiology, 1998* (Cleveland, OH)

Johnson A E, Behar J, Andreotti F, Clifford G D and Oster J 2014 R-peak estimation using multimodal lead switching *Computing in Cardiology, 2014* (Cambridge, MA)

Johnson A E, Behar J, Andreotti F, Clifford G D and Oster J 2015 Multimodal heart beat detection using signal quality indices *Physiological Measurement* **36** 1665

Kaplan D and Staffin P Updated Feb 3 1998 Software for Heart Rate Variability Accessed Sept 1 2016 (https://www.macalester.edu/~kaplan/hrv/doc/index.html)

Li Q and Clifford G D 2012 Signal quality and data fusion for false alarm reduction in the intensive care unit *Journal of Electrocardiology* **45** 596-603

Li Q, Liu C, Li Q, Shashikumar S, Nemati S, Shen Z and Clifford G In Submission Ventricular Beat Detection Using a Wavelet Transform and a Convolutional Neural Network *Physiological Measurement*

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 *Physiological measurement* **29** 15-32

Li Q, Rajagopalan C and Clifford G D 2014 Ventricular fibrillation and tachycardia classification using a machine learning approach *IEEE Transactions on Biomedical Engineering* **61** 1607-1613

Liu C Y, Oster J, Li Q, Zhao L N, Nemati S and Clifford G D In Submission Comparison of Measures of Entropy for Atrial Fibrillation Detection

Lomb N R 1976 Least-squares frequency analysis of unequally spaced data *Astrophysical and Space Science* **39** 447-462

Malik M 1996 Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. Heart rate variability. Standards of measurement, physiological interpretation, and clinical use *Circulation* **93** 1043-1065

Malik M and Camm A J 1995 *Heart Rate Variability* (Armonk, N.Y.: Futura Pub. Co. Inc.)

McSharry P and Malamud B 2005 Quantifying self-similarity in cardiac inter-beat interval time series *Computers in Cardiology 2005* **32** 459-462

McSharry P E, Clifford G, Tarassenko L and Smith L A 2002 Method for generating an artificial RR tachogram of a typical healthy human over 24-hours *Computers in Cardiology* **29** 225-228

McSharry P E, Clifford G D, Tarassenko L and Smith L A 2003 A Dynamical Model for Generating Synthetic Electrocardiogram Signals *IEEE Transactions on Biomedical Engineering* **50** 289-294

Mietus J E and Goldberger A L 2014 Heart Rate Variability Analysis with the HRV Toolkit - PhysioNet Accessed August 2016 (https://physionet.org/tutorials/hrv-toolkit/)

Mølgaard H 1991 Evaluation of the Reynolds Pathfinder II system for 24 h heart rate variability analysis *European heart journal* **12** 1153-1162

Moody G and Mark R 2001 The impact of the MIT-BIH Arrhythmia Database *IEEE Engineering in Medicine and Biology* **20** 45-50

Moody G B 2015a WFDB Applications Guide Accessed: August 2016 (https://www.physionet.org/physiotools/wag/)

Moody G B 2015b WFDB Programmer's Guide Accessed: March 2017 (https://physionet.org/physiotools/wpg/)

Oster J and Clifford G D 2015 Impact of the presence of noise on RR interval-based atrial fibrillation detection *Journal of Electrocardiology* **48** 947-951

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 *American Journal of Physiology* **275** 946H-950

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 *Medical and Biological Engineering and Computing* **54** 1921-1933

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: An Interdisciplinary Journal of Nonlinear Science* **5** 82-87

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 *Revista Española de Cardiología (English Edition)* **63** 802-809

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 Ed*, (Cambridge, MA: Cambridge University Press) pp 537-608

Richman J S and Moorman J R 2000 Physiological time-series analysis using approximate entropy and sample entropy *American Journal of Physiology-Heart and Circulatory Physiology* **278** H2039-H2049

Scargle J D 1982 Studies in astronomical time series analysis. II-Statistical aspects of spectral analysis of unevenly spaced data *The Astrophysical Journal* **263** 835-853

Schmidt G, Malik M, Barthel P, Schneider R, Ulm K, Rolnitzky L, Camm A J, Bigger Jr J T and Schömig A 1999 Heart-rate turbulence after ventricular premature beats as a predictor of mortality after acute myocardial infarction *The Lancet* **353** 1390-1396

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 *International Ophthalmology* **11** 151-157

Sun J, Reisner A, Saeed M and Mark R 2005 Estimating cardiac output from arterial blood pressurewaveforms: a critical evaluation using the MIMIC II database *Computers in Cardiology* **32** 295-298

Sun J X 2006 Cardiac output estimation using arterial blood pressure waveforms *Thesis* Massachusetts Institute of Technology

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 *Computer Methods and Programs in Biomedicine* **113** 210-220

Vest A N, Li Q, Liu C, Nemati S, Shah A and Clifford G D 2017 Benchmarking Heart Rate Variability Toolboxes *Journal of Electrocardiology* **50** 744-747

Vest A N, Li Q, Liu C, Nemati S, Shah A and Clifford G D 2018 PhysioNet Cardiovascular Signal Toolbox Accessed: May 2018 (https://github.com/cliffordlab/PhysioNet-Cardiovascular-Signal-Toolbox) DOI 10.5281/zenodo.1243111

Vollmer M HRVTool Accessed March 30 2017 (https://marcusvollmer.github.io/HRV/)

Zhu T, Johnson A E, Behar J and Clifford G D 2014 Crowd-sourced annotation of ECG signals using contextual information *Annals of Biomedical Engineering* **42** 871-884

Zong W, Moody G and Jiang D A robust open-source algorithm to detect onset and duration of QRS complexes *Computers in Cardiology 2003* **30** 737-740

Appendices

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.

B. QRS Detection Benchmark Testing for PhysioNet Cardiovascular Signal Toolbox and PhysioNet HRV Toolkit

Appendix Table A1 provides results detailed in Vest et al. (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 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

C. Default parameters in the PhysioNet Cardiovascular 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 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	au	Lower physiological limit, minimum RR interval
preprocess.upperphysiolim	2	au	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	m	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
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 VPC
HRT.BeatsAfter	16	au	Number of beats after VPC 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

D. Demonstration Code Available in the PhysioNet Cardiovascular Signal Toolbox

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.
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-norma) time series and the resulting output is saved in a .csv file.
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 vs PTT).
4. RRGEN Data Demo: *DemoStandardizedData.m*

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