IOP Publishing | Institute of Physics and Engineering in Medicine Physiological Measurement

Physiol. Meas. **38** (2017) 1061–1076 https://doi.org/10.1088/1361-6579/aa6e9c

Heart rate-based window segmentation improves accuracy of classifying posttraumatic stress disorder using heart rate variability measures

Erik Reinertsen[1], Shamim Nemati[2], Adriana N Vest[2,3], Viola Vaccarino[3,4], Rachel Lampert[5], Amit J Shah[3,4] and Gari D Clifford[1,2]

[1] Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA, United States of America
[2] Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, GA, United States of America
[3] Department of Epidemiology, Rollins School of Public Health at Emory University, Atlanta, GA, United States of America
[4] Division of Cardiology, Department of Medicine, Emory University School of Medicine, Atlanta, GA, United States of America
[5] Division of Cardiology, Department of Medicine, Yale University School of Medicine, New Haven, CT, United States of America

E-mail: er@gatech.edu

Received 13 December 2016, revised 3 April 2017
Accepted for publication 21 April 2017
Published 9 May 2017

Abstract

Objective. Heart rate variability (HRV) characterizes changes in autonomic nervous system function and varies with posttraumatic stress disorder (PTSD). In this study we developed a classifier based on heart rate (HR) and HRV measures, and improved classifier performance using a novel HR-based window segmentation.

Approach. Single-channel ECG data were collected from 23 subjects with current PTSD, and 25 control subjects with no history of PTSD over 24 h. RR intervals were derived from these data, cleaned, and used to calculate HR and HRV metrics. These metrics were used as features in a logistic regression classifier. Performance was assessed via repeated random sub-sampling validation. To reduce noise and activity-related effects, we calculated features from five non-overlapping ten-minute quiescent segments of RR intervals defined by lowest HR, as well as random ten-minute segments as a control.

Main Results. Using a combination of the four most predictive features derived from quiescent segments we achieved a median area under the

receiver operating curve (AUC) of 0.86 on out-of-sample test set data. This was significantly higher than the AUC using 24 h of data (0.72) or random segments (0.67).

Significance. These results demonstrate our segmentation approach improves the classification of PTSD from HR and HRV measures, and suggest the potential for tracking PTSD illness severity via objective physiological monitoring. Future studies should prospectively evaluate if classifier output changes significantly with worsening or effective treatment of PTSD.

Keywords: post-traumatic stress disorder, heart rate variability, machine learning, segmentation, electrocardiogram

Supplementary material for this article is available online

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

1. Introduction

Posttraumatic stress disorder (PTSD) can develop after exposure to traumatic events such as violence, natural disasters, or combat. Symptoms include nightmares of the trauma, hypervigilance, difficulty sleeping, poor concentration, and avoidance of places, activities, or persons that remind the affected individual of the causal incident (Yehuda *et al* 2015). PTSD has a lifetime prevalence of about 8% in the US general population (Resnick *et al* 1993, Yehuda 2002). The prevalence of PTSD is higher in developing or war-afflicted countries, in which people are exposed to more severe and/or more numerous traumas (Karam *et al* 2014). Lifetime prevalence is thus especially high in veterans, ranging from 6–30% (Dohrenwend *et al* 2006, Kok *et al* 2012, Sundin *et al* 2014, Marmar *et al* 2015).

Patients with PTSD have significantly different measures of heart rate variability (HRV) compared to healthy controls (Minassian *et al* 2014, Liddell *et al* 2016). HRV—changes in beat-to-beat heart rate—can be used to assess changes in the autonomic nervous system (Clifford 2002, Pan *et al* 2016). Recently, twins with PTSD were reported to have 49% lower low frequency (LF) HRV compared to their brothers without PTSD (Shah *et al* 2013). When attempting to identify differences in autonomic function as measured by HRV, it is important to control for other factors such as stress, affect, physical activity, and cardiovascular or neurological disease other than PTSD.

Evaluating HRV during sleep can account for confounding from stress, affect, and physical activity (Germain *et al* 2005). Furthermore, some reports show HRV reductions due to PTSD are greatest during the night (Woodward *et al* 2009, Kobayashi *et al* 2014), suggesting that analyzing data only during nocturnal sleep could improve classifier performance. However, HRV metrics vary by sleep stage due to changes in vagal and sympathetic activity during REM, light and deep sleep (Vanoli *et al* 1995, Viola *et al* 2002). Segmentation by sleep stage may thus improve the signal to noise ratio. For example, in earlier work using this novel methodology, we showed that comparing HRV metrics in REM sleep, and separately in deep sleep, better separated sleep apneic patients from healthy controls (Clifford 2002, Clifford and Tarassenko 2004). This approach may also apply to other illnesses associated with changes in HRV measures, such as PTSD. However, accurately measuring sleep status or estimating sleep stage from other data such as HR is difficult.

PTSD has been classified using self-reported data and demographics (Kessler *et al* 2014, Karstoft *et al* 2015, Galatzer-Levy *et al* 2014). However, a multivariate classifier separating

PTSD patients and controls using HRV measures or other objective physiological data has not yet been developed. Additionally, the utility of thresholding on individual HRV measures to identify PTSD has yet to be evaluated.

Here we propose a novel method of controlling for activity by only evaluating quiescent segments of RR intervals, with quiescence determined by lowest median HR for each subject. This segmentation approach may reduce random error from mental and physical activity, highlight involvement of the autonomic nervous system, and approximate restfulness in the absence of validated sleep stage data.

The objectives of this work were to: (1) calculate features from HR and HRV measures indicative of PTSD in male veterans using 24 h Holter ECG recordings, (2) use these features to train a multivariate classifier whose output—a probability of membership in either the PTSD or control group—could potentially be used as a proxy for illness severity in a patient already diagnosed with PTSD, and (3) improve classifier performance using a novel segmentation method on RR intervals to reduce noise and potential confounders.

2. Methods

All data processing, feature extraction, and classifier training was performed using Matlab R2016b (Mathworks, Natick, MA).

2.1. Subject enrollment

ECG recordings were obtained from 24 male subjects with current PTSD (symptoms within the past 30 days) and 26 healthy control subjects in a dataset derived from the Emory Twins Studies first reported by Shah *et al* (2013). This smaller cohort was selected to balance classes, i.e. a similar number of subjects with PTSD as controls. Participants were subjects with clinical diagnoses of PTSD, and healthy control subjects examined at the same time at the Emory University General Clinical Research Center. Individuals lacking sufficient ECG data were excluded (see exclusion criteria in later section). All participants wore an ambulatory ECG (Holter) monitor (GE Marquette SEER digital system; GE Medical Systems, Waukesha, WI) for 24 h. Participants had matched recording times and schedules. Activity was restricted to non-strenuous walking around the university campus and medical center, and participants were told to refrain from smoking or drinking alcohol or coffee. This study was approved by the Emory Institutional Review Board (81004), and all subjects signed an informed consent.

2.1.1. Data recording. The ECG signal was sampled at 125 Hz. Data were downloaded to a local HIPAA-compliant data repository using a MARS SEER Light digital recorder. QRS complexes were detected and annotated in the ECG automatically using the GE MARS software. RR intervals were calculated from the time difference between adjacent annotated beats.

2.1.2. Data pre-processing and exclusion criteria. RR intervals obtained later than 24 h after the start of recording were discarded. Ectopic beats and artifacts were removed via established methods (Malik 1996); non-physiological RR intervals with values >1.5 s or <0.33 s were discarded, and RR intervals 20% shorter or longer than the previous RR interval, or 20% shorter or longer than the overall mean RR interval were discarded. Gaps in the time series were interpolated via linear spline. RR intervals were re-sampled at 3.413 Hz (1024 samples per five minute segment) to create a uniformly spaced time series for spectral HRV measures. One subject with PTSD and one subject without PTSD had fewer than 22 h of ECG

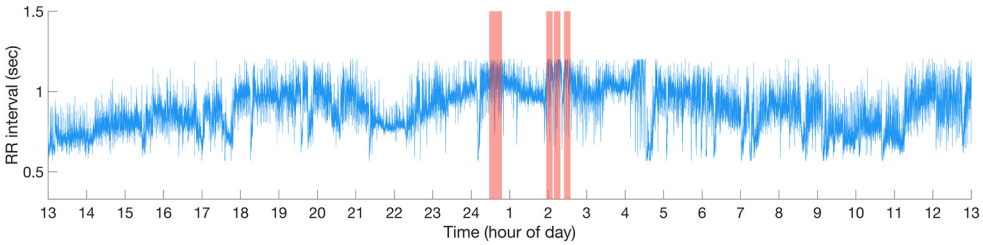

Figure 1. Representative time series of RR interval data from a single subject. Shaded red areas are ten-minute quiescent segments. Horizontal axis is time of day in hours; 13 corresponds to 1 PM, 1 corresponds to 1 AM, etc. ECG recording started at the origin of the *x*-axis (approximately 1 PM).

recordings; both were excluded from further analysis. Cleaned RR intervals were obtained from 23 subjects with PTSD and 25 control subjects (48 total). To demonstrate the utility of data pre-processing, we also used uncleaned RR intervals as a comparison.

2.1.3. Identification of quiescent segments. To reduce confounding effects of mental and physical activity, five non-overlapping ten-minute periods with the lowest median HR—hereafter referred to as 'quiescent segments'—were identified from cleaned RR data for each subject. Figure 1 illustrates a representative 24 h RR tachogram from a study subject, with quiescent segments indicated by shaded regions. Healthy humans cycle through each of the five defined sleep stages with a period of approximately 100 min, and each sleep stage lasts up to 20 min; this informed our selection of segment length (Clifford and Tarassenko 2004). For each subject, each feature was calculated for each of five quiescent segments, resulting in $5 \times m$ total features per subject. For each feature, the median feature value from the five segments was calculated, resulting in m features per subject to be used for training a logistic regression model. Feature extraction was also performed on ten-minute segments chosen at random, excluding quiescent segments of lowest HR, to serve as a control and to investigate if segment length was a confounder.

2.2. Feature extraction and heart rate variability measures

Cleaned RR intervals from either (a) all 24 h of ECG recordings, (b) random control segments, or (c) quiescent segments were used to calculate features. These features included the median quiescent window time converted to radians, basic RR interval statistics (mean, median, mode, standard deviation (σ_{rr}), interquartile range ($\mathrm{IQR}_{\mathrm{rr}}$), skewness, and kurtosis), AC, DC, power spectral measures (VLF, LF, HF, total power), and other measures of the distribution of RR intervals (NNN, MNN, PNN, PNN50, RMSSD, and SDNN) (Malik 1996).

2.2.1. Power spectral measures of HRV. HRV power spectral measures were computed from cleaned RR interval time series with a fast Fourier transform (FFT) and a Parzen window, following our previous methodology (Shah *et al* 2013). The FFT and spectra were corrected for window attenuation and boxcar sampling. The power spectrum was integrated over four discrete frequency bands: ultra-low frequency (ULF) $<$0.0033 Hz; very low frequency (VLF) 0.0033–0.04 Hz; low frequency (LF) 0.04–0.15 Hz; and high frequency (HF) 0.15–0.40 Hz (Bigger *et al* 1992). These frequency bands measure the renin-angiotensin, sympathetic, and parasympathetic cardiovascular control systems (Akselrod *et al* 1981). Total power, incorporating the full spectrum from 0–0.40 Hz was also estimated.

2.2.2. Phase-rectified signal averaging. Phase-rectified signal averaging (PRSA) was performed on cleaned RR intervals to quantify acceleration and deceleration capacity of HR. This method can be used to detect quasi-periodic oscillations and to separate processes occurring during increasing and decreasing parts of the signal (Bauer *et al* 2006a). Furthermore, PRSA is robust to noise and non-stationarity. Heartbeat interval shortenings are used as anchors for acceleration-related PRSA signals, whereas heartbeat interval lengthenings are used as anchors for deceleration-related PRSA signals. Sampling frequency was set to 512 Hz, and the window length around anchors was set to 30 elements.

2.3. Assessment of PTSD

The structured clinical interview for psychiatry disorders was administered to classify subjects into two classes: (1) current PTSD with symptoms within the past 30 days, or (2) no history of PTSD (control subjects).

2.4. Feature selection and classification

All twenty features, as well one feature at a time, were used to train a logistic regression. The output of this binary classifier was the probability of membership in the PTSD class. L1L2 (elastic net) regularization was performed to reduce coefficient values for collinear or non-predictive features and create a sparser and more generalizable model. Unconstrained differentiable multivariate optimization was performed using `minFunc`. Specifically, maximum likelihood estimation was performed via Quasi-Newton limited-memory Broyden–Fletcher–Goldfarb–Shanno updating (Bishop 1995). Distributions of features from PTSD and control subjects were visualized and compared via two-sided Kolmogorov–Smirnov tests. Additionally, given the relatively low number of features, a grid search was performed to assess combinations of features.

To assess classifier performance on out-of-sample data, we performed bagging with replacement, an ensemble method to reduce variance and avoid overfitting (Breiman 1996, Arlot and Celisse 2010). Data were randomly split into training and test data at a 70:30 ratio, with the class balance in training and test sets maintained to reflect the class balance in the entire data set. By random sampling with replacement, some data may be used more than once between models, or not be used at all. Features in the training set were transformed to have Gaussian distributions using either the identity, square root and logarithmic transformations. The transformation which provided the lowest k-statistic using the Lilliefors test was used on both training and test sets. Data were then z-scored to by subtracting the training mean and dividing by the training standard deviation on both the training and test sets. A grid search was performed to select the value of λ ranging from 0.001–5.0 that maximized the test set AUC within the model. The classifier thus learned solely from training data, and was evaluated solely on test data. Sampling, feature transformation, learning, and classifier evaluation was repeated nine more times for a total of ten models. AUC, accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were calculated for training and test sets within each model.

3. Results

3.1. Temporal distribution of quiescent segments

The temporal distribution of quiescent segments does not differ by PTSD status ($P = 0.23$ via two-sided Kolmogorov–Smirnov test; figure 2). Box plots are not associated with the y-axis; $+$ indicates the mean, the middle line indicates the median, the box denotes the interquartile

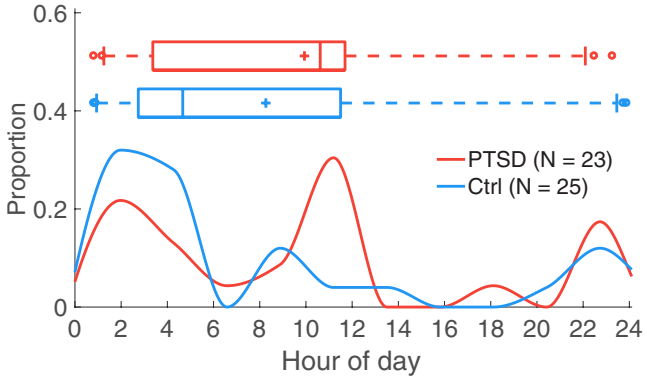

Figure 2. Temporal distribution of quiescent segments does not differ by PTSD status ($P = 0.23$). The x-axis denotes hour of the day (i.e. hours past midnight), ranging from 0 to 24; 12 corresponds to noon. Red indicates quiescent segments from subjects with PTSD (23 subjects); blue indicates quiescent segments for healthy controls (25 subjects).

range (IQR) flanked by the 25th and 75th percentiles, the vertical lines outside of the box indicate the 9th and 91th percentiles, and circles indicate outliers.

3.2. Classifier trained on all features

All twenty features were used to train an L1L2 regularized logistic regression. Classifier performance was assessed for three different segmentation approaches, and using either uncleaned or cleaned RR intervals. Using quiescent segments of cleaned RR intervals results in greater predictivity compared to other segmentation approaches, with a training AUC of 0.87 and a test AUC of 0.70 (table 1).

3.3. Classifier trained on individual features and combinations of features

To improve classifier performance, individual features and combinations of features were used to train a regularized logistic regression. Testing many combinations of features is computationally inefficient, but was feasible here given the small number of features and fast speed of training a logistic regression model. Classifier performance was assessed for three different segmentation approaches, using uncleaned or cleaned RR intervals. A classifier trained on the most predictive combination of four features derived from quiescent segments of RR intervals achieves greater predictivity (training AUC = 0.85, test AUC = 0.84) compared to when using features derived from random segments or 24 h of RR intervals (table 2).

The most predictive combination of four features derived from 24 h of RR intervals is σ_{rr}, IQR_{rr}, LF power, and SDNN (table 3). The most predictive combination of four features derived from quiescent segments of RR intervals were AC, DC, LF power, and SDNN (table 4). The β coefficients of these most predictive models are shown in table 5, and other classifier performance metrics are shown in table 6.

3.4. Distributions of predictive features

Distributions of predictive features were visualized (figures 3–8). Box plots are not associated with the y-axis; + indicates the mean, the middle line indicates the median, the box denotes

Table 1. AUCs of L1L2 regularized logistic regression models using all HR and HRV features extracted from RR intervals. Values shown are medians and IQR bounds in brackets.

	Train AUC		Test AUC	
	No RR cleaning	RR cleaning	No RR cleaning	RR cleaning
24 h	0.77 [0.75 0.82]	0.75 [0.70 0.78]	0.54 [0.46 0.64]	0.58 [0.46 0.64]
Random segments	0.76 [0.73 0.80]	0.78 [0.77 0.80]	0.50 [0.45 0.57]	0.56 [0.50 0.71]
Quiescent segments	0.89 [0.87 0.91]	0.87 [0.83 0.89]	0.73 [0.70 0.80]	0.75 [0.71 0.82]

Table 2. AUCs of L1L2 regularized logistic regression models using the top four features extracted from RR intervals. Values shown are medians across sub-samples and IQR bounds in brackets.

	Train AUC		Test AUC	
	No RR cleaning	RR cleaning	No RR cleaning	RR cleaning
24 h	0.74 [0.73 0.78]	0.73 [0.69 0.74]	0.66 [0.45 0.66]	0.67 [0.62 0.71]
Random segments	0.70 [0.66 0.76]	0.76 [0.72 0.77]	0.61 [0.50 0.64]	0.72 [0.62 0.77]
Quiescent segments	0.85 [0.84 0.88]	0.85 [0.83 0.88]	0.81 [0.70 0.84]	0.86 [0.75 0.88]

Table 3. Features extracted from 24 h of of RR intervals, shown as medians and IQR bounds in brackets. CTRL refers to the control group. Test AUC reports performance of univariate classifier trained solely on one feature.

	PTSD status		
Feature	PTSD	CTRL	Test AUC
AC (s)	−8.28 [−1.27e1 −6.31]	−1.04e1 [−1.33e1 −8.18]	0.54 [0.52 0.68]
DC (s)	8.19 [6.55 1.23e1]	1.05e1 [8.89 1.38e1]	0.58 [0.54 0.73]
LF power (s^2)[a,b]	3.51e2 [1.37e2 4.91e2]	5.86e2 [3.76e2 8.76e2]	0.71 [0.64 0.80]
σ_{rr} (s)[b]	1.15e−1 [9.15e−2 1.34e−1]	1.29e−1 [1.14e−1 1.51e−1]	0.65 [0.59 0.73]
IQR_{rr} (s)[b]	1.76e−1 [1.26e−1 2.11e−1]	2.08e−1 [1.52e−1 2.34e−1]	0.63 [0.59 0.67]
SDNN (s)[b]	3.89e1 [2.97e1 5.42e1]	5.07e1 [4.09e1 6.32e1]	0.61 [0.55 0.75]

[a] $P < 0.05$ comparing feature values from PTSD versus control subjects via two-sided Kolmogorov–Smirnov test.
[b] Feature among combination that maximizes training set AUC.

the IQR flanked by the 25th and 75th percentiles, the vertical lines outside of the box indicate the 9th and 91th percentiles, and circles indicate outliers.

Segmentation improves separability of some features as determined by two-sided Kolmogorov–Smirnov tests. AC does not significantly differ by PTSD status when evaluating 24 h of data ($P = 0.24$), but is significantly higher in subjects with PTSD versus controls when analyzing quiescent segments ($P = 0.04$; figure 3). Similarly, DC does not significantly differ by PTSD status when evaluating 24 h of data ($P = 0.13$), but is significantly lower in subjects with PTSD versus controls when analyzing quiescent segments ($P = 0.01$; figure 4). LF power is lower in PTSD for both 24 h data ($P = 0.01$) and quiescent segments of data ($P = 0.01$; figure 5). σ_{rr} does not differ by PTSD status for 24h of data ($P = 0.25$) but is higher in control subjects versus subjects with PTSD ($P < 0.05$; figure 6). Similarly, IQR_{rr} does not differ by PTSD status for 24h of data ($P = 0.47$) but is higher in control subjects versus subjects with PTSD ($P < 0.05$; figure 7). SDNN does not differ by PTSD status for 24 h of data ($P = 0.06$), but is significantly lower in PTSD when analyzing quiescent segments ($P = 0.04$; figure 8).

Table 4. Features extracted from quiescent segments of of RR intervals, shown as medians and IQR bounds in brackets. CTRL refers to the control group. Test AUC reports performance of univariate classifier trained solely on one feature.

	PTSD status		
Feature	PTSD	CTRL	Test AUC
AC (s)[a,b]	−9.62 [−1.26e1 −6.22]	−1.28e1 [−1.91e1 −9.72]	0.77 [0.73 0.82]
DC (s)[a,b]	9.43 [6.64 1.22e1]	1.40e1 [1.11e1 2.06e1]	0.82 [0.73 0.84]
LF power (s^2)[a,b]	3.31e2 [1.52e2 5.78e2]	8.71e2 [4.44e2 1.47e3]	0.81 [0.75 0.88]
σ_{rr} (s)[a]	4.14e−2 [3.44e−2 5.34e−2]	7.12e−2 [4.9e−2 8.06e−2]	0.82 [0.73 0.84]
IQR_{rr} (s)[a]	5.40e−2 [3.55e−2 5.60e−2]	7.20e−2 [5.50e−2 9.38e−2]	0.78 [0.71 0.81]
SDNN (s)[a,b]	4.68e−1 [3.16e1 5.97e1]	6.47e1 [4.32e1 7.70e1]	0.75 [0.57 0.86]

[a] $P < 0.05$ comparing feature values from PTSD versus control subjects via two-sided Kolmogorov–Smirnov test.
[b] Feature among combination that maximizes training set AUC.

Table 5. β coefficients of L1L2 regularized logistic regression models trained on four most predictive features from either 24 h or quiescent segments of RR intervals. Values shown are medians across sub-samples and IQR bounds in brackets.

	24 h		Quiescent segments	
	Feature	Coefficient value	Feature	Coefficient
β_1	Intercept	0.06 [0.05 0.06]	Intercept	0.08 [0.08 0.11]
β_2	σ_{rr}	0.46 [0.35 0.51]	AC	1.12 [1.03 1.60]
β_3	IQR_{rr}	0.29 [0.22 0.54]	DC	0.80 [0.61 1.06]
β_4	LF power	0.00 [-0.03 0.07]	LF power	0.32 [0.00 0.67]
β_5	SDNN	-0.04 [-0.31 -0.00]	SDNN	0.30 [0.01 0.39]

Table 6. Classifier performance on test set data using most predictive logistic regression models trained on features extracted from RR intervals after using three different segmentation approaches. Values shown are medians across sub-samples and IQR bounds in brackets. PPV is positive predictive value and NPV is negative predictive value.

	Segmentation approach		
Metric	24 h	Random segments	Quiescent segments
AUC	0.67 [0.62 0.71]	0.70 [0.62 0.79]	0.86 [0.75 0.88]
Accuracy	0.73 [0.67 0.73]	0.73 [0.67 0.80]	0.80 [0.73 0.80]
Sensitivity	0.57 [0.43 0.71]	0.43 [0.43 0.57]	0.71 [0.57 1.00]
Specificity	0.94 [0.75 1.00]	1.00 [0.88 1.00]	0.94 [0.88 1.00]
PPV	0.92 [0.71 1.00]	1.00 [0.78 1.00]	0.94 [0.83 1.00]
NPV	0.69 [0.67 0.75]	0.67 [0.64 0.73]	0.79 [0.73 0.88]

4. Discussion

In this study on 23 subjects with current PTSD and 25 controls, HR and HRV features were calculated and used to train an L1L2 regularized logistic regression to classify PTSD status. A classifier trained on a combination of the four most predictive features—LF power, σ_{rr}, IQR_{rr}, and SDNN for 24 h of RR intervals, and AC, DC, LF power, and SDNN for quiescent segments—achieved out-of-sample test AUCs of 0.67 using 24 h of RR interval data, 0.72 using random segments, and 0.86 using quiescent segments.

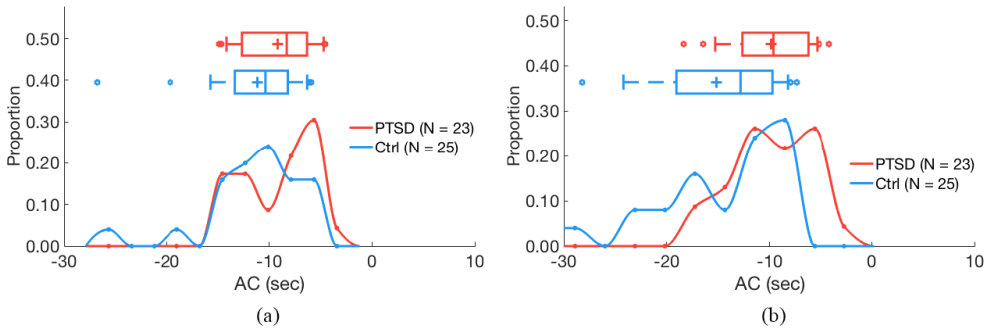

Figure 3. Acceleration capacity (AC) does not differ by PTSD status for 24 h of RR intervals (a; $P = 0.18$) but is higher in subjects with PTSD for quiescent segments (b; $P < 0.05$). (a) 24 hours of data. (b) Quiescent segments.

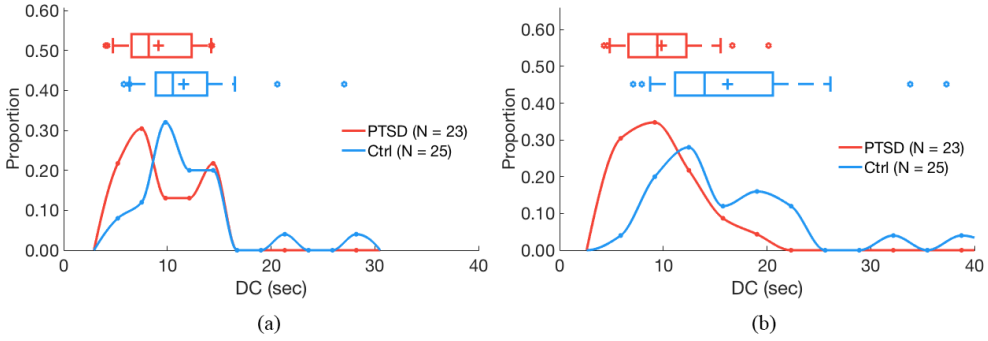

Figure 4. Deceleration capacity (DC) does not differ by PTSD status for 24 h of RR intervals (a; $P = 0.09$) but is lower in subjects with PTSD for quiescent segments (b; $P < 0.05$). (a) 24 hours of data. (b) Quiescent segments.

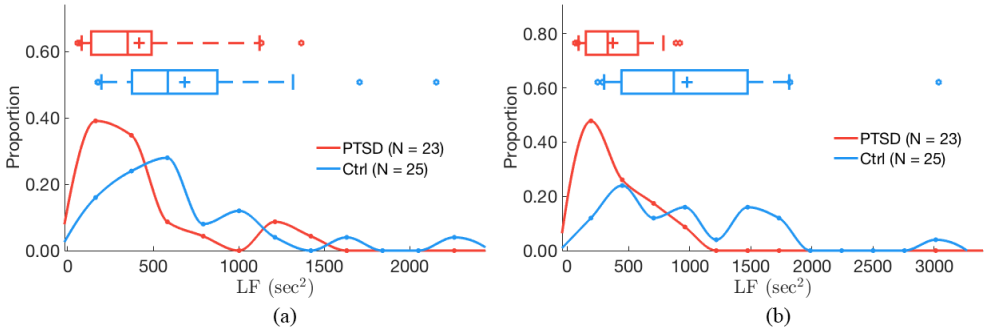

Figure 5. Low frequency (LF) power differs by PTSD status for both 24 h of RR intervals (a; $P < 0.05$) and quiescent segments (b; $P < 0.05$). (a) 24 hours of data. (b) Quiescent segments.

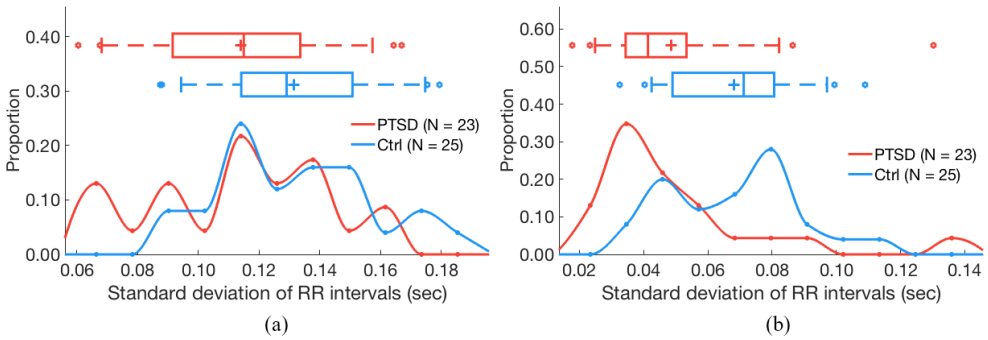

Figure 6. σ_{rr} (standard deviation of RR intervals) does not differ by PTSD status for 24 h of RR intervals (a; $P = 0.25$) but but is higher in control subjects for quiescent segments (b; $P < 0.05$). (a) 24 hours of data. (b) Quiescent segments.

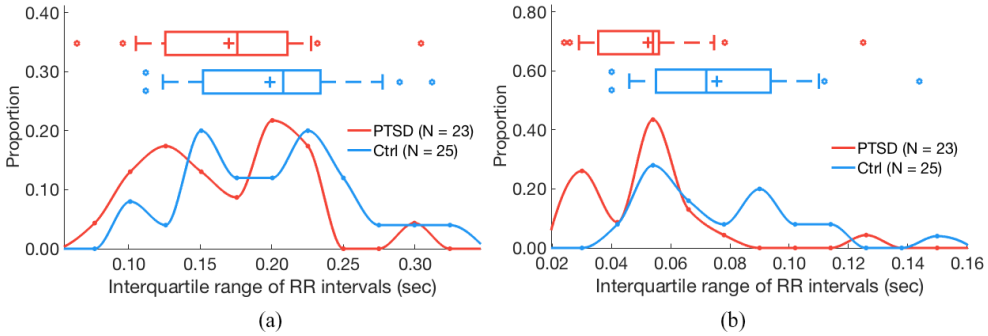

Figure 7. IQR_{rr} (interquartile range of RR intervals) does not differ by PTSD status for 24 h of RR intervals (a; $P = 0.47$) but is higher in control subjects for quiescent segments (b; $P < 0.05$). (a) 24 hours of data. (b) Quiescent segments.

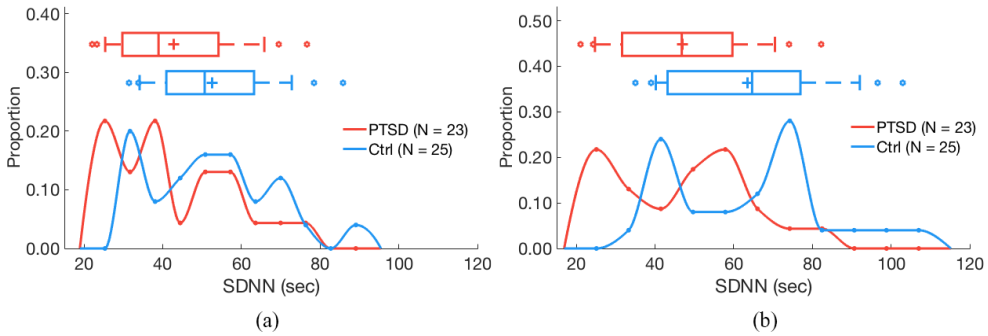

Figure 8. Standard deviation of normal-to-normal RR intervals (SDNN) does not differ by PTSD status for 24 h of RR intervals (a; $P = 0.06$) but is higher in control subjects for quiescent segments (b; $P < 0.05$). (a) 24 hours of data. (b) Quiescent segments.

Sleep disordered breathing and sleep disruption are both associated with PTSD, so proxies of sleep are expected to differ by PTSD status (Germain 2013, Yesavage *et al* 2014). However, the time of median quiescent segments did not significantly differ with PTSD status ($P = 0.23$; figure 2), indicating these factors were not significant in this cohort. Most quiescent segments

occurred from midnight to early morning in control subjects. A larger portion of segments were distributed closer to noon in subjects with PTSD. Periods of low HR—a measure of restfulness, not sleep stage—can occur at any time and may reflect differences in sleep patterns, differences in activity, or both. Quiescent segments may contain less noise and movement artifact, as well as reflect lower levels of mental and physical activity, and thus improve the performance of a classifier trained on features from those segments.

Next, we used all HR and HRV measures as features for a logistic regression classifier. L1L2 regularization was performed to reduce coefficient values associated with collinear or redundant features, and to isolate predictive features. A classifier trained on all 20 features from 24 h of RR intervals achieved a low test AUC of 0.58 (table 1). Using features extracted from quiescent segments improved the test AUC to 0.75, whereas the use of randomly selected control segments resulted in a low test AUC of 0.56. Compared to these low test AUCs, training AUCs were 0.75, 0.78, and 0.87 for 24 h, random segments, and quiescent segments of RR intervals respectively. These results show a model using all features over-fits training data and would not generalize to out-of-sample data despite regularization. Classifier performance was similar when using uncleaned RR interval data.

Regularization attempts to reduce co-linearity by effectively placing a prior on model coefficients, forcing sparsity with small weights. However, the posterior—formed by updating the prior with evidence—determines the final form of a model. Thus, with small data sets, even regularized models trained with many features may not work well compared to the use of a hard prior via manual feature selection. Therefore, we tested individual features and combinations of features to train lower-dimensional models.

Given $m = 20$ total features and a subset of $k = 1, 2, \ldots, m$ features, the number of possible combinations (i.e. possible arrangements of k features) is the binomial coefficient $\binom{m}{k}$. To ensure feasible computation time and a parsimonious and interpretable model, we limited the maximum number of features used in a combination to four, i.e. $k = 1, 2, \ldots, 4$. Furthermore, using more than four features led to the selection of colinear features and overfitting on the training data (results not shown).

Values of some individually predictive features, and test set AUC and accuracy for classifiers trained these features, are shown in table 3. We compared distributions of some features via a two-sided Kolmogorov–Smirnov test, but selected the most predictive comnbination features on the basis of maximizing training AUC. For classification, features should be chosen on the basis of predictability rather than significance, because significance alone does not guarantee predictability (Lo *et al* 2015). For 24 h of RR intervals, LF power significantly differed by PTSD status and was one of the four most predictive combination of features (table 3). The other most predictive features were σ_{rr}, $\mathrm{IQR}_{\mathrm{rr}}$, and SDNN, but these did not significantly differ by PTSD status. For quiescent segments, the median value of the four most predictive combination of features—AC, DC, LF power, and SDNN—significantly differed by PTSD status (table 4).

AC did not differ by PTSD status for 24 h of RR intervals, but was higher in subjects with PTSD for quiescent segments (figure 3). Similarly, DC did not differ by PTSD status for 24 h of RR intervals, but was lower in subjects with PTSD for quiescent segments (figure 4). AC may reflect physiologic performance when parasympathetic withdrawal occurs, whereas DC measures general parasympathetic augmentation (Bauer *et al* 2006b, Pan *et al* 2016). Although some literature suggests that AC also measures sympathetic activation, this is unlikely because sympathetic modulations occur at 0.1 Hz, which may be four times faster than the modulation frequency of AC, depending on the underlying heart rate (Julien 2006).

LF power differed by PTSD status for both 24 h and quiescent segments of RR intervals (figure 5). Differences in these measures by PTSD status may be exacerbated in quiescent

segments. In PTSD, vagal augmentation is expected during slow wave sleep, which may be altered by increased insomnia or sleep-disordered breathing. Other physiologic pathways may also be affected during abnormal sleep episodes; low LF may reflect baroreceptor insensitivity (Khoury *et al* 2012). These findings underscore physiologic changes that occur with PTSD.

When shifting from 24 h to quiescent segments, σ_{rr} and IQR_{rr} became less predictive, whereas AC and DC became more predictive. In quiescent segments, σ_{rr} in controls was greater than σ_{rr} in subjects with PTSD (figure 6). σ_{rr}, IQR_{rr}, and SDNN measure variability of RR intervals, and were all significantly lower in quiescent segments from subjects with PTSD. This finding is consistent with previous reports of lower variability of HR being associated with PTSD (Tan *et al* 2009, 2011). Additionally, the lack of significance or predictivity of these features (aside from SDNN, which was a predictive feature) from 24 h of RR intervals is unsurprising because quiescent segments were selected on the basis of low resting HR values, which excludes periods with higher variability. Concerning AC and DC, quiescent segments approximate restfulness rather than sleep state, but may also correspond to slow-wave sleep, during which vagal activity may be augmented and the predictivity of PRSA measures increased.

We calculated β coefficients of L1L2 regularized logistic regressions trained on four most predictive combination of features from 24 h or quiescent segments of RR intervals (table 5). Although LF power and SDNN were among the most predictive features when using either 24 h or quiescent segments of RR intervals, the β coefficients of these features significantly differed depending on the segmentation approach. For example, the median β coefficient for LF power computed from 24 h of RR intervals was close to zero, but for quiescent segments, the median β coefficient was 0.32. This difference suggests interactions between features that reflect the complexity of the underlying physiology, and/or a dependence on time scale.

Although our objective was to accurately classify PTSD status—particularly on out-of-training-sample data—rather than quantify individual features, we also estimated β coefficients using a logistic regression with no regularization, and all data at once rather than bagging with sub-sampling. When using 24 h of RR intervals, the β coefficients of the most predictive combination of four features were 0.08 (σ_{rr}), 0.10 (IQR_{rr}), -0.36 (LF), and -0.01 (SDNN) (table S1) (stacks.iop.org/PM/38/1061/mmedia). When using quiescent segments of RR intervals, the β coefficients of the most predictive combination of four features were: 0.89 (AC), -1.15 (DC), -0.44 (LF), and 0.15 (SDNN) (table S2). *P*-values for all β coefficient were >0.05, likely due to collinearity. We note aggregating several nearly significant predictors into one overall model can still result in high discrimination performance. The training AUC using the most predictive combination of features from 24 h of RR intervals was 0.71, and the training AUC using quiescent segments was 0.87. Using all 20 features from quiescent segments resulted in an training AUC of 0.94. Assessments were performed on the same data used to train the models, resulting in overfitting. The performance of these models would not generalize to out-of-sample data.

A regularized classifier trained on the most predictive combination of four features from (a) quiescent segments outperformed classifiers using (b) all 24 h of RR intervals, or (c) on random control segments, with test AUCs of (a) 0.86, (b) 0.67, and (c) 0.70 respectively (table 6). Using quiescent segments instead of 24 h of RR intervals improved every performance metric except specificity, which did not change. Using quiescent segments instead of random segments improved every performance metric except specificity and PPV, which decreased. This suggests classifier performance depends on the information within segments rather than the quantity of data.

We also compared the distribution of probabilistic classifier output using a Wilcoxon signed rank test to account for the paired nature of these data, and found a statistically significant difference between $P_{\text{estimated}}(\text{PTSD}|\text{features from subjects with PTSD})$ and $P_{\text{estimated}}(\text{PTSD}|\text{features from control subjects})$ ($P < 0.001$). This suggests the classifier accurately discriminated PTSD status.

Here the AUC can be interpreted as the ability of a model to classify PTSD status using disease-associated physiological changes. Although learning was done with data from healthy controls, this approach would be suited for monitoring patients with established PTSD. It would not be a screening test for the general population. Future studies could assess how treatments affect physiology, and classify or even predict post-intervention recovery.

We note several limitations of our study. First, our cohort consisted only of 23 subjects with PTSD and 25 controls. This small sample size may not have been adequately powered to detect smaller effect sizes. Our study design would be more elegant with discordant pairs only; however, this would eliminate ten unpaired twins and could reduce statistical power. To evaluate this we compared classifier performance using all subjects ($N = 48$) versus using only paired twins ($N = 38$) (table S3). We found no statistical differences using a two-sided Wilcoxon rank-sum test between all subjects and only paired twins cohorts in training or test AUCs for any segmentation approach. This may be due to two competing effects. Reducing sample size could diminish the ability of the classifier to learn predictive features, and decrease out-of-sample test set performance by learning features not representative of the population distribution. Furthermore, HRV may be about 50% heritable (Su *et al* 2010). If HRV and PTSD share an underlying genetic and physiological cause, and our approach evaluates features related to this mechanism, adding paired twins could confound the study, enrich both positive and negative classes with similar physiology-based features, and reduce classifier performance. However, focusing on twins could reduce the random error caused by differences in cardiovascular or autonomic physiology between subjects. Our results suggest the inclusion of non-twins does not reduce the impact of our findings, since we aimed to developed a system for monitoring physiology of subjects with PTSD rather than for screening a correlated population.

A second limitation of our work was only recording 24 h of ECG data per subject. Our approach could potentially enable home-based continuous physiologic monitoring of the efficacy of a PTSD intervention. However, doing so would require longer monitoring than 24 h and additional validation studies. Additionally, longitudinal monitoring could necessitate a specific, rather than sensitive assay, to prevent alarm fatigue driven by false positives. We emphasize the importance of prospective studies with larger sample sizes and a testable intervention in order to determine clinical utility.

A third limitation is our lack of locomotor activity data, which if present may have enhanced the accuracy of our classifier. Previously we have shown the addition of locomotor activity to HRV metrics improves accuracy of classification of schizophrenia (Osipov *et al* 2015). This could also be the case for PTSD; locomotor activity may improve signal quality assessment or directly indicate disturbed sleep, sedentary behavior, or avoidance of traumatic stimuli.

A fourth limitation is model output being probability of a PTSD diagnosis, which is a coarse proxy for illness severity. Our method would estimate a low probability of illness for a subject who is diagnosed with PTSD yet has atypically low levels of ANS dysfunction. Other aspects of PTSD symptomatology described in the DSM-V—such as negative alterations in mood or problems concentrating—have yet to be evaluated in the context of HRV measures. Estimating particular manifestations of PTSD severity may be more clinically useful than estimating PTSD status. However, doing so would require larger studies with multimodal data including high-resolution ECG recordings, locomotor activity, and clinical questionnaires.

Despite several limitations, this approach of classifying mental illness from physiological data has applications beyond PTSD. Changes in ANS function and psychological stress occur in other psychiatric illnesses such as bipolar disorder and depression, and are detectable using noninvasive physiological sensors (Burns *et al* 2011, Sano and Picard 2013, Tsanas *et al* 2016, Palmius *et al* 2016). Previously we used HRV measures and locomotor activity to accurately separate subjects with schizophrenia from healthy controls (Osipov *et al* 2015). Our novel approach of extracting features from quiescent segments of RR intervals could also be applied to locomotor activity, which correlates with illness status and HR. Techniques that improve the signal-to-noise ratio and enable fusing of complementary data sources could aid the classification of other mental illnesses. Other possible applications of this approach are to monitor adherence to medication, or to assess efficacy of an intervention. Interpreting model output as illness severity rather than a probability of class membership could alert a caregiver of deterioration or a sustained problem in a patient.

The utility of computational approaches to interpret multiple statistical and dynamic features of physiological signals has become increasingly apparent in all fields of biomedicine. Complex, information-rich settings such as critical care or sleep medicine are especially fertile sources of data with which to build tools and address clinical questions (Monasterio *et al* 2012, Behar *et al* 2013).

5. Conclusion

We classified PTSD in 48 male veterans using L1L2 regularized logistic regression trained on HR and HRV features. Classifiers trained on the most predictive four features from 24 h or random ten-minute control segments of RR intervals achieved test AUCs of 0.67 and 0.70, respectively. We improved test AUC to 0.86 by segmenting RR intervals into quiescent ten-minute segments to filter out activity- or noise-related effects. To our knowledge this is the first report of classification of PTSD status using non-invasive physiological features. This approach may provide a long-term ambulatory index of PTSD severity, have applications in the study and management of other mental illnesses, and be useful for other clinical disciplines where cardiovascular disease and stress are significant factors.

Acknowledgments

This study was funded in part by NIH grants UL1TR000454, K24HL077506, R01HL68630, R01AG026255, R01HL125246, R01HL109413, P01HL101398, and K23HL127251; AHA grants 0245115N and 15SDG25310017; and Emory University GCRC MO1-RR00039 and KL2TR000455. A.N.V., A.S, and G.D.C also thank Medibio Limited for support.

References

Akselrod S *et al* 1981 Power spectrum analysis of heart rate fluctuation: a quantitative probe of beat-to-beat cardiovascular control *Science* **213** 220–22

Arlot S and Celisse A 2010 A survey of cross-validation procedures for model selection *Stat. Surv.* **4** 40–79

Bauer A *et al* 2006a Deceleration capacity of heart rate as a predictor of mortality after myocardial infarction: cohort study *Lancet* **367** 1674–81

Bauer A *et al* 2006b Phase-rectified signal averaging detects quasi-periodicities in non-stationary data *Phys. A: Stat. Mech. Appl.* **364** 423–34

Behar J *et al* 2013 A review of current sleep screening applications for smartphones *Physiol. Meas.* **34** R29–46
Bigger J T *et al* 1992 Frequency domain measures of heart period variability and mortality after myocardial infarction *Circulation* **85** 164–71
Bishop C M 1995 *Neural Networks for Pattern Recognition* (Oxford: Oxford University Press)
Breiman L 1996 Bagging predictors *Mach. Learn.* **24** 123–40
Burns M N *et al* 2011 Harnessing context sensing to develop a mobile intervention for depression *J. Med. Internet Res.* **13** e55
Clifford G D 2002 Signal processing methods for heart rate variability *PhD Thesis* University of Oxford
Clifford G D and Tarassenko L 2004 Segmenting cardiac-related data using sleep stages increases separation between normal subjects and apnoeic patients *Physiol. Meas.* **25** N27–35
Dohrenwend B P *et al* 2006 The psychological risks of Vietnam for U.S. veterans: a revisit with new data and methods *Science* **313** 979–82
Galatzer-Levy I R *et al* 2014 Quantitative forecasting of PTSD from early trauma responses: a machine learning application *J. Psychiatric Res.* **59** 68–76
Germain A 2013 Sleep disturbances as the Hallmark of PTSD: where are we now? *Am. J. Psychiatry* **170** 372–82
Germain A *et al* 2005 A brief sleep scale for posttraumatic stress disorder: pittsburgh sleep quality index addendum for PTSD *J. Anxiety Disorders* **19** 233–44
Julien C 2006 The enigma of Mayer waves: facts and models *Cardiovascular Res.* **70** 12–21
Karam E G *et al* 2014 Cumulative traumas and risk thresholds: 12-month Ptsd in the world mental health (WMH) surveys *Depression Anxiety* **31** 130–42
Karstoft K I *et al* 2015 Bridging a translational gap: using machine learning to improve the prediction of PTSD *BMC Psychiatry* **15** 30
Kessler R C *et al* 2014 How well can post-traumatic stress disorder be predicted from pre-trauma risk factors? an exploratory study in the WHO world mental health surveys *World Psychiatry* **13** 265–74
Khoury N M *et al* 2012 The renin-angiotensin pathway in posttraumatic stress disorder: angiotensin-converting enzyme inhibitors and angiotensin receptor blockers are associated with fewer traumatic stress symptoms *J. Clin. Psychiatry* **73** 849–55
Kobayashi I, Lavela J and Mellman T A 2014 Nocturnal autonomic balance and sleep in PTSD and resilience *J. Traumatic Stress* **27** 712–6
Kok B C *et al* 2012 Posttraumatic stress disorder associated with combat service in Iraq or Afghanistan: reconciling prevalence differences between studies *J. Nervous Mental Dis.* **200** 444–50
Liddell B J *et al* 2016 Heart rate variability and the relationship between trauma exposure age, and psychopathology in a post-conflict setting *BMC Psychiatry* **16** 133
Lo A *et al* 2015 Why significant variables aren't automatically good predictors *Proc. Natl Acad. Sci.* **112** 13892–7
Malik M 1996 Heart rate variability: standards of measurement, physiological interpretation, and clinical use *Eur. Heart J.* **17** 354–81
Marmar C R *et al* 2015 Course of posttraumatic stress disorder 40 years after the vietnam war: findings from the National Vietnam Veterans Longitudinal Study *JAMA Psychiatry* **10016** 875–81
Minassian A *et al* 2014 Heart rate variability characteristics in a large group of active-duty Marines and relationship to posttraumatic stress *Psychosomatic Med.* **76** 292–301
Monasterio V, Burgess F and Clifford G D 2012 Robust classification of neonatal apno ea-related desaturations *Physiol. Meas.* **33** 1503–16
Osipov M *et al* 2015 Objective identification and analysis of physiological and behavioral signs of schizophrenia *J. Mental Health* **24** 276–82
Palmius N *et al* 2016 Detecting bipolar depression from geographic location data *IEEE Trans. Biomed. Eng.* https://doi.org/10.1109/TBME.2016.2611862
Pan Q *et al* 2016 Do the deceleration/acceleration capacities of heart rate reflect cardiac sympathetic or vagal activity? a model study *Med. Biol. Eng. Comput.* **54** 1–13
Resnick H S *et al* 1993 Prevalence of civilian trauma and posttraumatic stress disorder in a representative national sample of women *J. Consulting Clin. Psychol.* **61** 984–91
Sano A and Picard R W 2013 Stress recognition using wearable sensors and mobile phones *Humaine Association Conf. on Affective Computing and Intelligent Interaction Stress* https://doi.org/10.1109/ACII.2013.117

Shah A J *et al* 2013 Posttraumatic stress disorder and impaired autonomic modulation in male twins *Biol. Psychiatry* **73** 1103–10
Su S *et al* 2010 Common genes contribute to depressive symptoms and heart rate variability: the twins heart study *Twin Res. Hum. Genet.* **13** 1–9
Sundin J *et al* 2014 Mental health outcomes in US and UK military personnel returning from Iraq *Br. J. Psychiatry* **204** 200–7
Tan G *et al* 2011 Heart rate variability (HRV) and posttraumatic stress disorder (PTSD): a pilot study *Appl. Psychophysiol. Biofeedback* **36** 27–35
Tan G *et al* 2009 Associations among pain, PTSD, mTBI, and heart rate variability in veterans of operation enduring and Iraqi freedom: a pilot study *Pain Med.* **10** 1237–45
Tsanas A *et al* 2016 Daily longitudinal self-monitoring of mood variability in bipolar disorder and borderline personality disorder *J. Affective Disorders* **205** 225–33
Vanoli E *et al* 1995 Heart rate variability during specific sleep stages *Circulation* **91** 1918–22
Viola A U *et al* 2002 Sleep processes exert a predominant influence on the 24 h profile of heart rate variability *J. Biol. Rhythms* **17** 539–47
Woodward S H *et al* 2009 Autonomic activation during sleep in posttraumatic stress disorder and panic: a mattress actigraphic study *Biol. Psychiatry* **66** 41–6
Yehuda R 2002 Post-traumatic stress disorder *New Engl. J. Med.* **346** 108–14
Yehuda R *et al* 2015 Post-traumatic stress disorder *Nat. Rev. Dis. Primers* **1** 15057
Yesavage J A *et al* 2014 Longitudinal assessment of sleep disordered breathing in Vietnam veterans with post-traumatic stress disorder *Nat. Sci. Sleep* **6** 123–7