Wearable Estimation of Heart Rate Recovery to Physical Activity During Daily Life in Patients with Recurrent Major Depressive Disorder

Alberto Barquero[1], Esther García[2,3], Spyridon Kontaxis[1]. Sara Siddi[4], Josep Maria Haro[4] Nicholas Cummins[5], Srinivasan Vairavan[6], Matthew Hotopf[5,7], Femke Lamers[8,9], Brenda Penninx[8,9], Richard Dobson[5], Vaibhav Narayan[6], Raquel Bailón[1,2], Pablo Armañac-Julián[1,2], the RADAR-CNS consortium

[1] BSICoS Group, I3A, IIS Aragón, Universidad de Zaragoza, Spain
[2] Center for Biomedical Research Network – Bioengineering, (CIBER-BBN), Spain
[3] Microelectronics and Electronic Systems, Autonomous University of Barcelona, Spain
[4] Parc Sanitari Sant Joan de Déu, Sant Joan de Déu Foundation, CIBERSAM, University of Barcelona, Spain
[5] King's College London, Institute of Psychiatry, Psychology and Neuroscience, London, UK
[6] Research and Development in Information Technology, Janssen Research & Development, LLC, Titusville, NJ, USA
[7] South London and Maudsley NHS Foundation Trust, UK
[8] Department of Psychiatry, Amsterdam UMC, Vrije Universiteit, the Netherlands
[9] Amsterdam Public Health Research Institute, the Netherlands

Abstract

This study investigates wearable estimation of heart rate recovery (HRR) to free-living physical activity in 529 patients with recurrent Major Depressive Disorder (MDD) using data from wrist-worn wearable devices (Fitbit) over a 2-year period. Depression is associated with autonomic nervous system dysregulation and increased cardiovascular risk. Our hypothesis was that HRR would be lower in patients with more severe depressive symptoms, and that surrogate physiological markers derived from wearables could complement clinical evaluations. Heart rate (HR) and step count data were continuously collected from the wearables. Depression severity was assessed biweekly using the Patient Health Questionnaire-8 (PHQ-8). Periods of physical activity were automatically detected from step count data using predefined criteria. To analyze HRR, we applied bivariate phase rectified signal averaging (BPRSA), estimating parameters characterizing HR response to physical exertion for each patient and PHQ-8 score. Univariate analyses did not show statistically significant differences in HRR across depression severity levels consistently, and a multivariate TabPFN model was able to classify patients with and without depressive symptoms with 53.82% accuracy (AUC = 0.5876). Our results suggest that the relationship between HRR to free-living physical activity and depression is not straightforward.

1. Introduction

Major Depressive Disorder (MDD) is a serious condition that affects 6% of the global adult population, associated with consequences such as disability, decreased quality of life, premature mortality, and suicide, with a higher prevalence in women than in men [1]. Its etiology is multifactorial and presents with complex symptomatology, such as mood and sleep disturbances, many of them related to the autonomic nervous system (ANS). In fact, depression has been linked to alterations in the ANS [2], which might be also related to the higher cardiovascular risk observed in MDD patients [3]. Wearable devices which allow the recording of physiological parameters, such as heart rate (HR) and physical activity , can provide objective measures related to these symptoms and can help identify early changes in health status often missed in sporadic clinical evaluations [4]. In fact, previous studies have shown that resting mean HR at night was higher in patients with more severe depression (also linked to sleep disturbances) while resting mean HR during the day was lower [5, 6].

This study combines HR and physical activity to analyze wearable estimation of heart rate recovery to free-living physical activity (HRR) in MDD patients during daily life. Our hypothesis is that HRR after exercise will be slower in more severe MDD patients, maybe reflecting cardiovascular impairment, and that a surrogate of this measurement can be obtained analyzing HRR to automatically detected

ISSN: 2325-887X DOI: 10.22489/CinC.2025.267

activity periods during daily life. We will investigate i these surrogate HRR are different according to the severit of depression and could be used to complement classica clinical evaluations.

2. Materials and Methods

2.1. Database

A subset of 529 MDD patients from the RADAR-CN project [7] was analyzed. These patients utilized Fitbi wristbands for continuous daily-life data collection of HI and step count over a period of 2 years. Number of steps are provided every minute while estimates of HR are provided with uneven sampling (one estimate every 5-15 s in favourable scenarios). Depression severity was assessed every 2 weeks with the Patient Health Questionnarie 8 items (PHQ-8), [8], delivered through an app installed in patients' smartphones.

Table 1. Characteristics of the Analyzed Population. Continuous variables are presented as median [interquartile range, IQR].

Characteristic	Value
Gender (N, %Female)	564 (70.95%)
Age	55 [26.5]
Baseline PHQ-8	10 [9]
PHQ-8 per patient	18 [26]

2.2. Average HR Response to Exercise

In order to identify periods of similar activity during daily life, this study employs the unintentional 6-minute walk test (6MWT) methodology, proposed by Sokas et al. [9]. Using 6-minute sliding windows with a 1-minute overlap (see Figure 1), activity onset and offset points are determined based on some thresholds imposed on cadence.

The analysis defines three cadence levels for detecting continuous activity windows (at least 6 minutes): (1) >0 steps/min (continuous movement); (2) $\geq$ 60 steps/min (standardized moderate activity) [9]; and (3) a personalized threshold based on each patient's median step count.

In order to characterize the HRR for each patient and PHQ-8 score, the bivariate phase-rectified signal averaging (BPRSA) method is used[10]. First, both step count and HR series are resampled with a sampling period of 5 s. Then, taking the offset of each detected activity window as the reference point, the HR series from 1 minute before to 3 minutes after the reference point were averaged across all activity windows detected in the interval of two weeks prior to the PHQ-8 completion for each patient.

Then, the following variables are defined to characterize HRR: maximum HR in the average curve (HR_{max}), minimum HR in the average curve (HR_{min}), cardiac capacity (C) calculated as HR_{max} minus HR_{min}, and the time to reach 95% of HR_{min} in the average curve (T_{95}), HR at each minute after the offset, and capacity at each minute (e.g., HRR_1 calculated as HR_{max}-HR_1 (see Figure 2). Other variables have also been investigated, such as the curve slope.

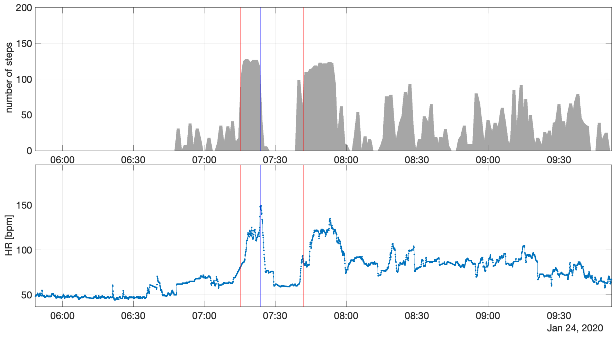

Figure 1. Onset and offset defined on the step series with custom threshold, and carried over to the synchronized HR series. Onsets are shown in red, and offsets in blue.

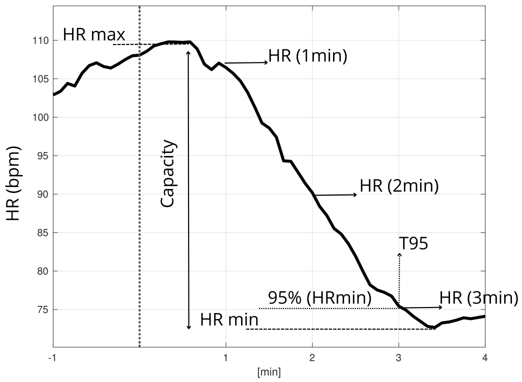

Figure 2. Variables characterizing BPRSA curve

2.3. Statistical Analysis and Classification Method

To evaluate significant differences in each variable across depression levels, the Mann-Whitney U test was employed. To leverage the multiple PHQ-8 measurements available for each patient and enhance result robustness, a Bootstrap technique was applied. In each replicate, one measurement per patient was randomly selected. The 5% and 95% confidence intervals for p-value were provided based on 5000 replicates. Spearman correlation analysis, suitable for non-linear associations, explored relationships between variables and PHQ-8 levels using Bootstrap resampling for improved precision.

To further investigate the link between physiological variables and depression status, a multivariate classification model using the efficient TabPFN model was developed. Trained on extracted physiological biomarkers, it classified patients into no significant depression (PHQ-8 $<$ 10) referred to as level 0, and depression (PHQ-8 $\geq$ 10) referred to as level 1. Model generalization was evaluated via a train-test split (80% train 20% test), ensuring patient data weren't mixed between the training and test sets. For robust performance estimation, mean accuracy, AUC, average precision, and the confusion matrix, are reported with k-fold cross-validation (k=5).

3. Results and Discussion

Statistical analysis was performed on a total of 5936 level 0 (PHQ-8 $<$ 10) cases (PHQ-8 score with corresponding HRR parameters) and 5529 level 1 (PHQ-8 $\geq$ 10) cases. In both groups HRR biomarkers varied according to the defined physical activity threshold. For instance, the capacity and T95 values consistently increased with higher activity threshold, supporting higher increases in HR and longer recovery periods with more intense activities.

Variable	**Level**	**No restriction**	**>60 steps**	**Custom**
Capacity	0	1.5 (1, 2.5)	3.5 (1.3, 6.6)	4.5(2, 7.5)
(bpm)	1	1 (0.9, 2.5)	3 (1.3, 6.4)	4.5 (2.5, 7.6)
HR_{max}	0	79 (69, 88)	83 (64, 94)	83 (68, 93)
(bpm)	1	81 (70, 90)	83 (68, 95)	84 (70, 94)
HR_{min}	0	77 (67, 85)	77 (57, 86)	76 (61, 86)
(bpm)	1	79 (67, 87)	77 (64, 87)	77 (63, 87)
Curve slope	0	-3 (-4.7, -1.9)	-7.4 (-13.8, -2.7)	-9.7 (-16.7, -5)
(beats/min)	1	-3.1 (-4.9, -1.9)	-7.1 (-13.9, -2)	-9.6 (-17.2, -2.6)
T_{95}	0	0.05 (-0.25, 0.20)	1.5 (0.75, 2.08)	1.83(1, 2.58)
(min)	1	-0.25 (-0.75, 0.17)	1.42 (0.67, 2)	1.75 (0.92, 2.42)

Table 2. Distribution of the proposed variables studied to characterize cardiac recovery. Data are presented as median (25th, 75th percentiles) for the depression levels according to PHQ-8 (Levels 0, 1), separated by the three thresholds used to define physical activity for the 6MWT.

Variable	**No restiction**	**>60**	**Custom**
Capacity	(0.112, 0.959)	(**0.009**, 0.952)	(**0.017**, 0.968)
$\mathrm{HR}_{\mathrm{max}}$	(0.148, 0.985)	(0.070, 0.982)	(**0.043**, 0.973)
$\mathrm{HR}_{\mathrm{min}}$	(0.153, 0.988)	(0.126, 0.982)	(0.115, 0.985)
Slope	(0.053,0.980)	(0.067,0.976)	(0.097,0.981)
T95	(0.087, 0.979)	(**0.025**, 0.969)	(0.085, 0.978)

Table 3. 5th and 95th percentiles of the p-values for the Mann-Whitney U test, comparing each of the variables according to the two groups of depression levels.

The conditions of over 60 steps per minute and the custom percentile, used to define activity in the 6MWT, tended to yield lower p-values when discriminating between level 0 and 1 compared to the no-restriction condition. This suggests that a minimal level of physical activity might be necessary to induce a heart rate response in the studied population. However, these differences were not statistically significant across all bootstrap iterations (see Table 3). .

Furthermore, the variability observed in the bootstrap-derived p-values underscores the complex and potentially fluctuating relationship between the HRR variables and depression scores, possibly stemming from individual physiological response variations and shifts in emotional states during the study. In line with the findings in Tables 2 and 3, correlation analyses between depression levels and individual HRR variables also failed to reveal consistent significant associations. Therefore, these univariate findings warrant cautious interpretation, and further investigation with larger datasets may uncover more consistent patterns.

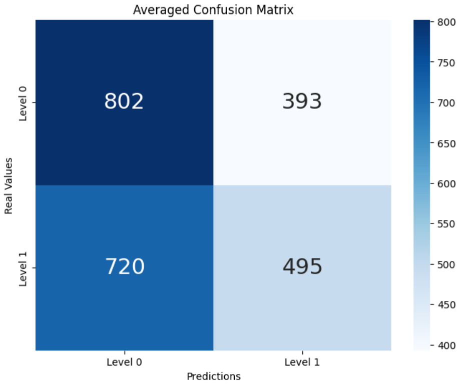

Figure 3. Confusion matrix illustrating the number of true negatives (TN), false positives (FP), false negatives (FN), and true positives (TP) for the depression classification model on the test set. The positive class (depression, level 1), and the negative class (no significant depression, level 0)

A TabPFN classifier was developed to explore the combined discriminative power of the derived HRR parameters. The model was trained using a 5-fold cross-validation approach, achieving an average accuracy of 72.87% and an average AUC of 0.8118. The averaged confusion matrix across the 5 folds further illustrates its performance. It correctly identified 717 out of 855 cases with no significant depression, but incorrectly classified 300 out of 759 individuals with the condition as not depressed. This high number of false negatives directly contributed to a limited average F1-score of 0.6769 and highlights a key challenge in detecting the positive class. The model's ability to identify cases without the condition was strong, with an average specificity of 83.89%.

Despite these robust cross-validation results, the model showed a substantial drop in performance when evaluated on an independent test set. With an accuracy of 53.82% and an AUC of 0.5874, the model's ability to generalize to completely new data was limited. This discrepancy is largely driven by a high number of false negatives, as demonstrated by the confusion matrix which shows that

720 out of 1,215 patients with the condition were incorrectly classified as not depressed. This suggests that the dataset's inherent variability is a key limitation, rather than an issue with the model's design itself.

While the classification model's results showed a limited ability to distinguish between individuals with and without depression based on heart rate recovery, it's important to consider the growing body of evidence linking depression to increased cardiovascular risk[3]. From this perspective, the physiological parameters analyzed could serve a dual purpose: not only as potential predictors of depression but also as possible early indicators or monitoring tools for cardiovascular risk in this vulnerable population. This dual justification validates the importance of the research and the need for further studies.

The primary limitation of this study is the absence of a control group without depression, which hinders direct comparisons of the results. Additionally, the PHQ-8, being a subjective test based on patient self-assessment, may not capture the full spectrum of depressive symptoms, potentially complicating its correlation with objective physiological data. Another limitation is that patients on the borderline of the classification threshold (around PHQ-8 = 10) may introduce noise, as their inclusion in either group is based on a single cutoff point. Furthermore, the data were not stratified by age or gender, both of which are known to influence individuals' physiological responses [11]. These limitations call for a cautious interpretation of the findings. For future research, it would be valuable to conduct a longitudinal analysis and explore the possibility of using personalized models for each patient, although these approaches were not feasible with the current dataset. The rigorous analysis itself revealed a key limitation: while the cross-validated model demonstrated good generalization performance, the significant drop in accuracy observed in a single-split test highlights the dataset's inherent variability, which can affect the model's stability when presented with a single, unique test set.

4. Conclusion

In a sample of 529 MDD patients, wearable estimates of heart rate recovery to physical activity automatically detected during daily life did not show significant differences across depression severity levels when examined separately. However, a multivariate machine learning approach achieved moderate accuracy to classify patients with and without MDD symptoms. This suggests that a combination of wearable-derived biomarkers may complement traditional assessments, offering a more comprehensive understanding of the disease. Future research should explore links to cardiovascular risk and use clustering to identify subgroups.

Acknowledgments

This work was partially funded by Spanish national and regional grants (TED2021-131106B-I00, PID2021-126734OBC21, BSICoS T39-23R) and the European RADAR-CNS project (grant No. 115902). Computations were performed by ICTS NANBIOSIS (HPC of the University of Zaragoza).

References

[1] Bromet E, et al. Cross-national epidemiology of dsm-iv major depressive episode. BMC Medicine 2011;9.

[2] Schiweck C, et al. Heart rate and high frequency heart rate variability during stress as biomarker for clinical depression. a systematic review. Psychological Medicine 2019; 49(2).

[3] O'Connor CM, et al. Depression and ischemic heart disease. American heart journal 2000;140(4).

[4] Charlton PH, et al. Wearable photoplethysmography for cardiovascular monitoring. Proceedings of the IEEE 2022; 110(3):355–381.

[5] Siddi S, et al. The usability of daytime and night-time heart rate dynamics as digital biomarkers of depression severity. Psychol Medicine 2023;53(8).

[6] Condominas E, et al. Exploring the dynamic relationships between nocturnal heart rate, sleep disruptions, anxiety levels, and depression severity over time in recurrent major depressive disorder. Journal of Affective Disorders 2025;.

[7] Matcham F, et al. Remote assessment of disease and relapse in major depressive disorder (radar-mdd): recruitment, retention, and data availability in a longitudinal remote measurement study. BMC Psychiatry 2022;22(1).

[8] Kroenke K, et al. The phq-8 as a measure of current depression in the general population. Journal of Affective Disorders 2009;114(1-3):163–173.

[9] Sokas D, et al. Detection of walk tests in free-living activities using a wrist-worn device. Frontiers in Physiology 2021;12:706545.

[10] Schumann AY, et al. Bivariate phase-rectified signal averaging. Physica A Statistical Mechanics and its Applications 2008;387(21).

[11] Seedat S, et al. Cross-national associations between gender and mental disorders in the world health organization world mental health surveys. Archives of General Psychiatry 2009;66(7):785–795.

Address for correspondence:

Alberto Barquero Ruiz. I3A, Universidad de Zaragoza, C. de Mariano Esquillor Gómez, s/n, 50018 Zaragoza.
761145@unizar.es