

Relationship Between Physical Activity Intensity and Physiological Stress Indicators in Wearable Sensor Data: A Correlational and Regression Analysis

Shourya Vanaja Aravind
shourya.va@gmail.com

ABSTRACT

This study aimed to investigate the ability of the intensity of physical activity measured with an accelerometer to predict three physiological signals, which are typically considered stress indicators in wrist-worn wearables. The 100 participant-session recordings in a public wearable-sensor dataset of an Empatica E4 (Empatica, Milan, Italy) were screened for artifacts (e.g., implausible physiological values, sensor faults, duplicated data), and means were computed per session for 96 participant sessions that passed quality control for the following signals: accelerometer, heart rate, electrodermal activity (EDA), and skin temperature. Three hypotheses related to the correlation of accelerometer magnitude with HR, EDA, and TEMP were tested using Pearson correlation and simple linear regression. Three hypotheses were tested regarding the correlation of the three measures: HR, EDA, and TEMP with magnitude on the accelerometer using Pearson correlation and simple linear regression. No significant correlation was found between activity intensity and any of the three physiological variables ($r = -.005$ to $-.057$, all $ps \geq .58$), nor were any of the three regression models significant, accounting for less than 1% of the variance for all of the outcomes. There was a moderate correlation between heart rate and EDA ($r = .424$, $p < .001$), which are both linked to the sympathetic branch. These results do not suggest that physical activity does not affect HR, EDA, or TEMP, but rather that the session-level aggregation approach is limited by condition heterogeneity and/or statistical power. It is likely that the signal collected from a wearable sensor is a combination of physical activity and psychological stress, and that these two components can be separated using dynamic, time-resolved, and condition-specific analyses.

Keywords: wearable sensors, physical activity, electrodermal activity, heart rate, stress detection, accelerometer

INTRODUCTION

Physiological sensors that can be worn and are capable of continuously monitoring psychological stress have become important tools in the field (Can et al., 2019). With heart rate (HR), electrodermal activity

(EDA), skin temperature (TEMP), and tri-axial acceleration, researchers were able to combine physiological signals with machine-learning techniques for stress detection (Schmidt et al., 2018; Gedam & Paul, 2021). Even with these advances, a great challenge still exists because many of the physiological changes that occur in the body during psychological stress are also seen during physical activity, as both involve activation of the sympathetic nervous system (Anwer et al., 2020). The identification of the stress-related responses from the exercise-induced physiological changes is thus a very important challenge for wearable-based stress monitoring.

One of the most commonly used measures of physiological arousal is heart rate. It rises with acute psychosocial stress, as in the Trier Social Stress Test (Kirschbaum et al., 1993), and during physical activity as cardiac output is increased to meet the metabolic demands. Additionally, HRV is related to ANS activity, but is also heavily influenced by movement, and should be interpreted with physical activity data (Shaffer & Ginsberg, 2017). Electrodermal activity is an indicator of a peripheral origin of physiology which is driven by the sweat gland activity under the direct control of the sympathetic nervous system (SNS); it is regarded as one of the most direct measures of physiology in the peripheral nervous system (Critchley, 2002; Dawson et al., 2017). Although EDA is very widely used in affective-computing research (Posada-Quintero et al., 2016), it is still vulnerable to artefacts due to movement and temperature (Taylor et al., 2015), which makes the removal of artefacts crucial before stress classification. Skin temperature responds slowly to changes in HR and EDA and is indicative of thermoregulatory processes. It usually rises during prolonged exercise in response to heat dissipation, but can fall transiently during acute psychological stress, due to peripheral vasoconstriction (Anwer et al., 2020; Blokker et al., 2022).

Wearable-based stress detection systems could falsely detect exercise as stress due to the effects of both physical activity and psychological arousal on HR, EDA, and TEMP. To solve this problem, some previous studies have used accelerometer data to detect and eliminate artifacts due to movements before stress classification (Can et al., 2019; Taylor et al., 2015). However, there are still ongoing free-living validation studies that show that physical activity is one of the primary sources of classification error (Abd Al-Aalim et al., 2024). The majority of the previous studies have instead focused on movement as a confounder and not explored the impact of the intensity of physical activity on the physiology of stress.

This is addressed in the current study by analysing a public dataset of wearable physiological data recorded with the Empatica E4 during a set of structured stress induction and exercise sessions (Hongn et al., 2025). After a known artifact removal and quality control pipeline was applied to the raw sensor signals, 96 quality-controlled participant sessions (of 100 originally recorded) were analyzed to determine any relationship between intensity of physical activity (measured with an accelerometer) and three parameters of physiological stress (HR, EDA, and TEMP) measured at the session level. Three hypotheses were proposed based on previous literature: hypothesis (H1), that intensity of physical activity would be positively correlated with HR due to the cardiovascular requirements of physical activity; hypothesis (H2), that physical activity intensity would be more strongly correlated with HR than with EDA, reflecting the greater specificity of EDA with sympathetic arousal; and hypothesis (H3), that acceleration derived from physical activity would be more strongly correlated with HR than with skin

temperature because cardiovascular responses occur more rapidly than thermoregulatory responses. The study aims to quantify these relationships, and to better understand the degree to which physical activity is associated with common physiological measures used as markers of psychological stress.

LITERATURE REVIEW

The importance of wearable technologies has grown in the field of continuous psychological stress monitoring as they enable the measurement of physiological responses in real-world settings, not just in the laboratory. The Empatica E4 can collect several physiological signals from the autonomic nervous system, such as heart rate (HR), electrodermal activity (EDA), skin temperature (TEMP), and physical activity using triaxial accelerometers, offering researchers multiple measures of autonomic nervous system functioning (Can et al., 2019). These signals can be combined and incorporated into machine-learning models that have improved as they become more accurate in detecting stress (Gedam & Paul, 2021). A significant problem, however, is the fact that many physiological responses triggered by psychological stress are also triggered by exercise-induced physiological arousal, making it hard to differentiate emotional stress from exercise-induced physiological arousal. Under the influence of psychological stress, the sympathetic part of the autonomic nervous system is activated, resulting in physiological changes like increased heart rate, elevated activity of sweat glands, and changes in peripheral blood flow (McEwen, 2007). One of the most common measures of stress is heart rate, which elevates rapidly when the emotional challenge is present and cardiac output increases to meet the physiological needs (Shaffer & Ginsberg, 2017). Likewise, ED is a sensitive measure of emotional arousal (Critchley, 2002; Dawson et al., 2017), as it is a measure of changes in activity of the sweat glands, which are regulated only by activation of the sympathetic nervous system. Another physiological parameter that has been frequently included in wearable stress monitoring systems is skin temperature. Peripheral vasoconstriction and thermoregulatory response may cause temperature variation during stress periods, which can be measured and detected (Anwer et al., 2020). These physiological variables are commonly employed as indicators of stress, but they can also result from other factors unrelated to psychological stress. Exercise elicits many of the same physiological responses as are seen with stress. Exercise increases heart rate (HR), sympathetic nervous system activity, sweating, and skin temperature (ST) via metabolic heat production and thermoregulatory mechanisms (Anwer et al., 2020). If movement is not taken into account, a physical activity period might be confused with psychological stress by an algorithm used by the wearer. To overcome this drawback, researchers are now starting to integrate the accelerometer information into stress-detection algorithms to detect changes in physiology derived from movement (Can et al., 2019). Additionally, wrist movements create motion artifacts in EDA signals, which is particularly problematic for wrist-worn sensors, so there is a need for specific preprocessing methods to reduce noise in the signal (Taylor et al., 2015). With recent improvements in wearable stress detection, multi-modal has been highlighted as a way of combining physiological and movement data to enhance classification accuracy. Available datasets like WESAD (Schmidt et al., 2018) and the Wearable Device Dataset from Induced Stress and Structured Exercise Sessions (Hongn et al., 2025) have allowed researchers to assess stress-detection algorithms in different stress-induction and exercise scenarios.

However, many of the available studies are based on maximizing machine learning accuracy, rather than directly studying the physical activity intensity-individual physiological stress indicator relationship. Therefore, the evidence base of whether physical activity as measured by an accelerometer is a good independent predictor of heart rate, electrodermal activity, or skin temperature is relatively sparse when the physiological data are aggregated over sessions of activity.

In the present investigation, this gap is addressed by examining the correlation between the intensity of physical activity and three physiological stress indicators using Pearson correlation and simple linear regression analyses. It is crucial to learn whether such variation is explained by movement alone, or if there is significant variation in these physiological signals that reflects a psychological stress response, to enhance the validity of the wearable stress-monitoring systems and to differentiate between a true psychological stress response and a movement-related physiological response.

METHODOLOGY

Data Source

The data used were from the publicly available dataset on PhysioNet (Goldberger et al., 2000; Hongn et al., 2025): Wearable Device Data from Induced Stress and Structured Exercise Sessions. Physiological data from healthy controls were recorded with the Empatica E4 wristband during (a) a structured acute stress induction protocol (ASI), involving a series of mathematical tasks followed by emotional tasks, and rest intervals, and (b) a 2-way stationary aerobic and anaerobic exercise protocol. The device was recording the tri-axial acceleration (ACC), blood-volume-pulse-derived heart rate (HR), electrodermal activity (EDA), and peripheral skin temperature (TEMP) during each recording session. This resource was used as the basis for 100 participant-session recordings; after a systematic preprocessing and quality-control pipeline (described in Procedure and Data Preparation) removed sessions with corrupted, invalid, or missing sensor data, 96 sessions were retained, each of which was condensed into a mean score for each of the four channels of interest.

Participants

The session contributions were made by healthy adult volunteers who participated in the original data collection protocol, which involved a stress induction protocol followed by structured aerobic and anaerobic exercise sessions (Hongn et al., 2025). As this is a secondary analysis of anonymized, archived data that is publicly accessible, no further recruitment of participants, consent, or screening of demographic characteristics took place beyond that reported by the original data providers.

Measures

- ***Accelerometer Magnitude***

The predictor variable was the mean value of the tri-axial accelerometer signal (ACC channel) as a measure of physical activity intensity for each session. The validity of this type of triaxial acceleration data has been generally demonstrated to be valid when compared to criterion measures of physical activity and energy expenditure in healthy adults, with accuracy depending on the type and intensity of activity (Wu et al., 2023).

- ***Heart Rate***

The session mean HR was calculated as the first outcome variable as derived from the blood-volume-pulse signal captured from the HR sensor of the Empatica E4. Empatica E4-derived heart rate has proven to be well correlated with electrocardiography in resting or in low-movement conditions (Schuurmans et al., 2020).

- ***Electrodermal Activity***

EDA is a continuously recorded measure of skin conductance that reflects changes mediated by the sympathetic nervous system (Critchley, 2002). Session means EDA was used as the second outcome variable.

- ***Skin Temperature***

Peripheral skin temperature was measured continually from the sensor attached to the wrist in degrees Celsius, and the third outcome variable was session mean TEMP.

Procedure and Data Preparation

The four raw signal channels (tri-axial accelerometer, HR, EDA, TEMP) underwent screening and cleaning before aggregation, and the processing was documented and implemented in Python (pandas, NumPy, and SciPy). The raw accelerometer counts (units per axis: 1/64 g) were converted to g, and the per-sample vector magnitude was calculated using the Euclidean Norm Minus One ($ENMO = \max[0, \sqrt{(x^2 + y^2 + z^2) - 1g}]$), which is a standard measure derived from accelerometer data that captures acceleration due to movement but excludes the static gravitational force component (Wu et al., 2023). Both channels were de-spiked with a five-sample centered median filter (≈ 156 ms when sampled at 32 Hz by the accelerometer; ≈ 1.25 s when sampled at 4 Hz by EDA), as recommended for wrist-worn EDA data (Taylor et al., 2015), and then averaged to reduce brief, high-amplitude artifacts from movement. Out-of-range samples were also removed from the session mean: those outside 30-220 beats per minute for HR and those outside 0.01-100 μS for EDA (the physiologically plausible range for the skin conductance; Boucsein, 2012), and TEMP values outside 25-40°C were marked as sensor artifacts. Then, a quality-control rule was applied at the session level - a session was accepted only if 50% of the samples survived the cleaning in each of the four channels; otherwise, it was completely rejected and not added to

the sample; a session was only included if it had passed this threshold in all four channels. This pipeline was also utilized to address the Data Quality issues that were documented for this dataset. For participant S02, the mean was calculated excluding the period where the duplication occurred because all four channels were truncated at the point where the duplication started ($\approx 1,548$ s into the session); thus, only the non-duplicated portion of the session was used to calculate the mean. The protection dock was not removed from the wristband of the participant f07, so this session was dropped entirely, and the PPG channel that was used to read the HR and the TEMP channel used for the temperature reading were not valid. Three sessions (AEROBIC-condition for S07 and both ANAEROBIC-condition sessions for both S02 and S05) also had skin-temperature readings that were physiologically impossible (above 200°C), likely due to a detached or malfunctioning sensor, and were excluded on that basis. 96 of the 100 participant-session recordings were used in the analysis after having been preprocessed and screened for artifacts in all four channels; the remaining channel values were averaged into a single session mean for each of the four variables (accelerometer magnitude, HR, EDA, and TEMP). Three main reasons were used for the primary analysis, which was conducted as session-level analyses instead of time-resolved or lagged analyses. The exercise and stress protocols used in the source dataset vary significantly in duration, structure, and event timing between participants and conditions, and the accompanying tags in the dataset do not provide continuous fine-grained labels for any activity, but instead only mark major protocol transitions, meaning that a single-level summary provides the most directly comparable unit of analysis across the entire sample. Second, the level of aggregation for the session matches the level at which physiological stress indices are typically reported and interpreted in applied wearable-sensor research, which allows the present results to be most directly compared with this literature. Third, aggregating to the session level gave a conservative starting point for determining whether there is any correlation between physical activity intensity and stress markers at a physiological level, as suggested by the hypotheses, and then went on to perform more computationally-intensive time-resolved modeling. This has a significant constraint, however: Session-level analyses reported here do not test for moment-to-moment coupling between movement and physiology, but rather determine whether a stable, session-wide relationship exists.

Figure 1: *Analytical Workflow schematic, from Raw WSD Data Acquisition, to Preprocessing, QC, Aggregation and Statistical Analysis*



Data Analysis

The aggregated data over the session level was used for all analyses. The first step was to calculate Pearson product-moment correlation coefficients between each of the four variables to describe the linear relationship between physical activity intensity and the three physiological outcomes as well as between the three physiological outcomes. Second, three separate simple linear regression models were estimated, with respect to HR, EDA, and TEMP, and with the same predictor (accelerometer magnitude). The standardized regression coefficient (β), coefficient of determination (R^2), and the significance test (p) were analyzed for each model. All inferential tests used an alpha level of .05. The analyses were based on standard correlation and ordinary least-squares regression methods and did not involve the use of multivariate or multilevel models, as the data were session-aggregated and cross-sectional in design. Unstandardized regression slopes and Pearson correlation coefficients were computed with a 95% confidence interval using the Fisher z-transformation and ordinary-least-squares formulas, respectively. The residual diagnostics of each of the regression models were checked by means of the Breusch–Pagan

test for heteroscedasticity and the Jarque–Bera test for normality of the residuals. Spearman rank-order correlations and a sensitivity analysis (excluding this session) were performed to assess if the main Pearson results were sensitive to this high-leverage session or to non-linearity. Lastly, the dataset also included psychological stress sessions and structured aerobic and anaerobic exercise sessions, therefore, separate descriptive statistics for each of the different types of stress and exercise sessions and correlations between the accelerometer and the outcome measures were calculated, and one-way analyses of variance (ANOVAs) were used to determine if a difference existed between mean HR, EDA, TEMP, and accelerometer magnitude for each of the different types of stress and exercise sessions.

RESULTS

Preliminary Analysis

All 100 participant-session recordings were collected under the experimental conditions and were retained after preprocessing and quality control (above) for further analysis (96 sessions, complete and artifact-free on all four channels). The mean was computed for each session of the data for mean accelerometer magnitude, HR, EDA, and TEMP, where the predictor of the intensity of the physical activity was the mean accelerometer magnitude, and the outcome variables were mean HR, mean EDA, and mean TEMP. Data collected were screened for completeness, then summed at the session level and subsequently analysed by Pearson correlation to determine the relationship between physical activity intensity and physiological measures, and three separate simple linear regression models were used to determine if physical activity intensity could predict HR, EDA, and TEMP.

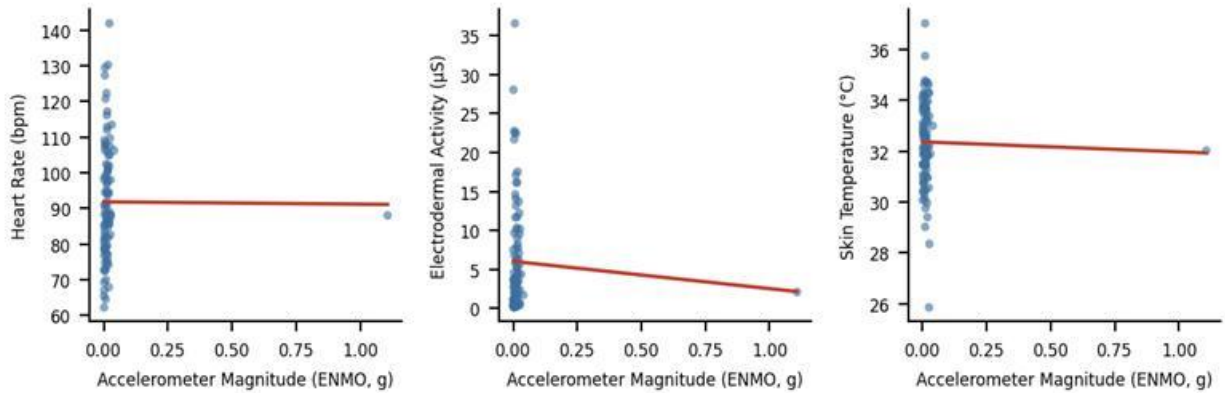
Correlation Analysis

The Pearson's correlation coefficient was calculated to examine the relationships between physical activity intensity and the physiological indicators (Table 1). There was no meaningful association between intensity of the physical activity and heart rate (HR; $r = -.005$, all $ps \geq .58$), electrodermal activity (EDA; $r = -.057$), or skin temperature (TEMP; $r = -.027$) based on the magnitude of the accelerometer. There was a moderate positive correlation between HR and EDA ($r = .424$, $p < .001$); there was a weak positive correlation between EDA and skin temperature ($r = .256$, $p = .012$); and HR was not significantly correlated with skin temperature ($r = .143$, $p = .166$). All three correlations between accelerometer and the other measures were compatible with a modestly moderate true correlation, both positive and negative, given the sample size used here and 95% confidence intervals (Fisher z) were provided (ACC–HR: 95% CI $[-.205, .196]$; ACC–EDA: 95% CI $[-.254, .145]$; ACC–TEMP: 95% CI $[-.226, .175]$). The HR–EDA correlation was estimated more precisely and stayed positive for its entire range, 95% CI $[.244, .576]$. Overall, there were modest associations between the physiological measures, and no meaningful association between physical activity intensity and the physiological measures at the session level.

Table 1: Pearson Correlation Matrix

Variable	1	2	3	4
1. Accelerometer Magnitude	—			
2. Heart Rate	−.005	—		
3. Electrodermal Activity	−.057	.424	—	
4. Skin Temperature	−.027	.143	.256	—

Figure 2: Scatterplots of Accelerometer Magnitude Predicting Heart Rate, Electrodermal Activity, and Skin Temperature



***Note.** The acceleration in each session is plotted against a physiological outcome ($N = 96$) along with the regression line fit with the ordinary least-squares method. Data are the session-level means after cleaning and artifact screening of the original sensor recordings from PhysioNet (Hongn et al., 2025). One anaerobic-condition session (participant f09) has a much larger magnitude than the rest of the sample and is an outlier in each panel; therefore, it was not excluded based on any documented or quality control exclusion criterion. Overall, non-significant, small associations between the level of physical activity and each physiological measure are shown visually in the near-horizontal regression lines in all three panels.*

Robustness Checks: Rank-Based and Outlier-Sensitivity Analyses

To assess if the absence of significant accelerometer results was due to true lack of association or a methodological limitation of the main analysis, two robustness checks were performed: one with the top 90% of the sample and another by excluding the single extreme session from the sample (the participant with the greatest magnitude of accelerometer data for the condition was an outlier in the original data set; visible in Figure 2). Spearman rank order correlations were calculated for every physiological outcome

with the magnitude of the acceleration, which are less sensitive to extreme values and do not require linearity. The other correlation analyses between the accelerometer magnitude and HR, EDA, and skin temperature were statistically significant and positive for the Pearson correlations ($p = .002$, $.029$, and $-.028$, respectively) and non-significant for the Spearman correlations ($p = .152$, $.360$, and $.788$, respectively). Second, the Pearson correlation and regression were re-calculated without the one high-leverage session. This correlation was then made significant and positive when the session was removed, $r = .275$, $p = .007$, whereas the correlation with EDA ($r = .009$, $p = .928$) and TEMP ($r = -.080$, $p = .440$) remained non-significant. These robustness checks demonstrate that the lack of a monotonic relationship between heart rate and physical activity intensity in the primary analysis was partially due to the disproportionate influence of one extreme session and not due to the absence of any monotonic relationship, and that physical activity intensity may have a positive relationship with heart rate after this analytical limitation is corrected. The correlations with EDA and TEMP, by contrast, were not strong and not significant in either of the robustness checks, suggesting more consistent evidence that the magnitude of the accelerometer measures specifically is not a strong linear or rank-based predictor of the two outcomes. The analyses presented in Tables 1 and 2 are presented with the full quality-controlled sample ($N = 96$) for consistency and transparency; these are presented as supplementary robustness checks.

Regression Analysis

Three simple linear regression analyses were performed to test the significance of the relationship between the intensity of the physical activity and heart rate, electrodermal activity, and skin temperature (Table 2 and Figure 2).

- ***Heart Rate***

A first regression model was used to determine if the magnitude of the accelerometer predicted the average heart rate. The model was not statistically significant as it accounted for almost none of the variance in heart rate ($R^2 = .00002$). No significant relationship was found between heart rate and the magnitude of the accelerometer, $\beta = -.005$, $p = .965$.

- ***Electrodermal Activity***

The second regression model examined the possibility that the magnitude of the accelerometer would predict electrodermal activity. The model accounted for 0.32% of the variance in EDA, $R^2 = .00323$. The magnitude of the accelerometers was not significantly related to electrodermal activity, $\beta = -.057$, $p = .583$.

- ***Skin Temperature***

The third regression model looked at the relationship between the intensity of the accelerometer's readings and skin temperature. Only 0.07% of the variance in skin temperature was explained by the model, $R^2 = .00071$. There was no significant relationship between the magnitude of the accelerometer and skin temperature, $\beta = -.027$, $p = .797$.

Regression Diagnostics

95% confidence intervals were calculated for each unstandardized slope and residual diagnostics were checked for each model to determine whether the null regression results were due to a truly flat relationship or a violation of the ordinary-least-squares assumptions. These 95% CIs for the slope of the relationship between accelerometer magnitude and each of HR, EDA, and TEMP included zero and were wide with respect to the point estimate, reflecting the high level of uncertainty around the current estimates with the small size of the present sample. Heteroscedasticity was tested using Breusch–Pagan tests which showed no evidence of unequal residual variance in any of the three models (all p s > .49), suggesting that this is not the reason for the null findings. The Jarque–Bera tests, however, showed that the residuals in all three models (all p s < .05) were not normally distributed, most strongly for the EDA model (skewness = 1.90) in keeping with the skewed distribution of electrodermal activity data. This departure from normality does not affect the regression coefficients themselves, but it may indicate that exact p values should be taken with a pinch of salt and that rank-based or transformed-variable analyses (which are reported in the Robustness Checks section) are useful supplements to the linear model reported here.

Table 2: Summary of Regression Analyses

Outcome Variable	β	R^2	p	Interpretation
Heart Rate	-.005	.00002	.965	Not significant
Electrodermal Activity	-.057	.00323	.583	Not significant
Skin Temperature	-.027	.00071	.797	Not significant

Hypothesis Testing

Hypothesis 1 proposed that there would be a positive correlation between intensity of physical activity and heart rate. This hypothesis was not confirmed; the higher the acceleration, the higher the HR, no significant correlation was found ($r = -.005$); if anything, a negative correlation was found, and regression analysis did not show that the intensity of the physical activity was significantly correlated with the HR, $p = .965$. This association was found to be statistically significant and positive after excluding a single high-leverage session (see Robustness checks) ($r = .275$, $p = .007$) and with a rank-based Spearman correlation ($\rho = .314$, $p = .002$), which casts doubt on the validity of the present null result as definitive evidence against H1.

Hypothesis 2 stated that the physical activity intensity would show a lower correlation with the electrodermal activity than with the heart rate. This was partially confirmed because the correlation between the magnitude of the accelerometer and the electrodermal activity was more negative than the correlation between the magnitude of the accelerometer and the heart rate ($r = -.057$, $r = -.005$, respectively), both very low and not statistically significant, and not easily comparable.

Hypothesis 3 was that the intensity of the physical activity would better predict heart rate than skin temperature. This hypothesis was not supported because the regression analyses did not find that the magnitude of the accelerometer was a significant predictor of HR or skin temperature, and both models explained very little variance.

Summary of Findings

Overall, the results showed that there was not a significant relationship between the physical activity intensity (accelerometer magnitude) and average measures of HR, EDA, and skin temperature at the session level. On the other hand, moderate positive associations were found between cardiovascular and electrodermal responses, indicating some synchrony between these two systems. Thus, the findings suggest that session-average physical activity metrics from the accelerometer might not be sufficient to account for the variability in physiological measures of stress and that a significant lack of predictive relationships may be due to the loss of temporal information that occurs with data aggregation.

Condition-Specific Analysis

Due to the nature of the data set (both psychological stress-induction sessions and structured aerobic and anaerobic exercise sessions), condition-specific patterns were explored as an additional analysis. Mean heart rate differed substantially by condition, one-way ANOVA $F(2, 93) = 30.50, p < .001$, and was highest during AEROBIC sessions ($M = 101.12, SD = 16.25$), followed by ANAEROBIC sessions ($M = 97.96, SD = 11.71$), and lowest during STRESS sessions ($M = 78.96, SD = 9.46$). Mean EDA also differed significantly by condition, $F(2, 93) = 3.91, p = .024$, again highest during AEROBIC sessions ($M = 8.28, SD = 8.43$) and lowest during STRESS sessions ($M = 3.69, SD = 6.07$). Mean skin temperature did not differ significantly by condition, $F(2, 93) = 1.25, p = .292$, nor did mean accelerometer magnitude, $F(2, 93) = 1.51, p = .226$. When the accelerometer–outcome correlations were recomputed separately within each condition, none reached statistical significance (STRESS, $n = 36$: ACC–HR $r = .103, p = .549$, ACC–EDA $r = -.071, p = .683$, ACC–TEMP $r = .242, p = .155$; AEROBIC, $n = 30$: ACC–HR $r = -.012, p = .948$, ACC–EDA $r = -.224, p = .234$, ACC–TEMP $r = -.258, p = .169$; ANAEROBIC, $n = 30$: ACC–HR $r = -.166, p = .382$, ACC–EDA $r = -.152, p = .422$, ACC–TEMP $r = -.011, p = .952$), although the smaller per-condition sample sizes substantially reduced statistical power relative to the combined analysis. These findings together suggest that, while there was no linear relationship between the magnitude of the accelerometer signal and levels of physiological arousal in each session, there were significant differences in the absolute levels of physiological arousal for each session, especially when considering HR and EDA.

DISCUSSION

In this study, whether the physical activity intensity, which was measured by an accelerometer at the session level, was correlated with three physiological parameters most frequently employed in wearable stress research: heart rate (HR), electrodermal activity (EDA), and skin temperature (TEMP) was examined. Contrary to the hypotheses, there were no significant relationships between physical activity

intensity and any of the measures, and physical activity intensity did not significantly predict any of the physiological outcomes. Based on these results, it may be inferred that, at the session level, there is a low correlation between these stress-related variables and the intensity of the physical activity. This pattern of null results calls for a careful differentiation between the absence of the relationship and the restriction in the present type of analysis. The latter seems not impossible because of several considerations. First, as reported above, the robustness checks showed that the correlation between the level of the accelerometer and the level of heart rate only became significant after excluding one of the most strongly leveraged sessions and when using a rank-based correlation (Spearman) rather than a Pearson correlation, suggesting that the linear estimate that was obtained using the full sample may have underestimated an underlying relationship. Second, because a minimum of $N = 96$ (and as low as 30 per condition) was needed for the power, only small-to-moderate effects could be detected; as indicated above with the wide confidence intervals surrounding each estimate from the accelerometers, even a true correlation as high as $r \approx .2$ may not be discerned as different from zero, depending on the sample size. Thirdly, combining each session into an average mean is a conservative test in that the greatest amount of within-session variation in movement and physiology is lost and therefore is most likely to have a detectable signal. The present results should therefore be viewed as an indication that the session-level magnitudes of accelerometers did not predict HR, EDA, and TEMP in this sample and with this analytical approach, and not as an indication that there was no effect of physical activity on those physiological signals.

The non-association of physical activity with HR is especially noteworthy as increases in HR are a well-documented physiologic response to exercise. One possible reason is that session-level averages are used, concealing the time series relationship between the short-term bursts of activity and the short-term cardiac reaction. The variability that is essential for detecting changes associated with activity in physiological signals may therefore be lost if averaged over the whole session of the trial. This interpretation aligns with past wearable-sensing studies which have advocated for more fine-grained analyses of movement over shorter time periods, instead of using session-level summary metrics (Can et al., 2019). Beyond the averaged HR, exercise also has some effects on other aspects of stress physiology, such as cortisol response and autonomic regulation, which cannot be explained by HR alone (Hill et al., 2008; Caplin et al., 2021).

The second hypothesis, that physical activity would have a weaker relationship with EDA than with HR, was only partially supported as there was essentially no relationship between physical activity and EDA. EDA is mostly associated with sympathetic nervous system activity and is generally seen as a sensitive response to psychological arousal (Critchley, 2002; Posada-Quintero et al., 2016). There are some concerns that physical motion may cause artifacts in EDA recordings, but previous research has identified that these artifacts are typically short-lived and can be filtered with specialized software (Taylor et al., 2015). This means that, on the one hand, true physiological responses and artifacts due to the movement were dampened by averaging over sessions.

Physical activity was also not a significant predictor of skin temperature. TEMP is less reactive than HR or EDA since it is a measure of thermoregulatory processes, not of autonomic responses. The results from previous studies have indicated that there was an increase in skin temperature during continuous exercise

(Anwer et al., 2020), but the effects observed might have been obscured by averaging over a session with varying levels of activity. There was no significant correlation of physical activity with the physiological variables, but HR and EDA had some moderate positive correlation. This discovery aligns with psychophysiological studies which indicate that although both are associated with different physiology, they represent common sympathetic nervous system activation (Critchley, 2002). This observed correlation implies that the physiological measures kept a coherent autonomic signal despite this lack of an explanation by session-level physical activity intensity.

The results may be relevant to wearables that detect stress. Physical activity is a very common confounding factor when considering HR, EDA, and temperature signals (Can et al., 2019; Gedam & Paul, 2021). The current findings indicate that this confounding effect is likely to be most apparent at the sub-session level, but not necessarily at the session level. Thus, session averages only would provide an underestimation of the effects of physical activity and would mask meaningful physiological responses. Interpretation of the results of session-level analysis with wearable devices like E4 should therefore be done with caution, especially when studying the association between physical activity and physiological responses to stress. There are a few caveats. First, the averages of the sessions were used, so they could not examine the temporal coupling between movement and physiological responses. Second, the dataset that included stress-induction and exercise sessions did not separate the effects of stress and exercise sessions from each other, possibly causing relationships between stress and exercise sessions to be masked. Third, a systematic pipeline for artifact removal and quality control was applied to all channels (including de-spiking, detection of physiologically implausible values, and a minimum number of valid data for each channel), but there were three instances of undocumented artifactual issues with sensors; these efforts may have prevented some of the null results, but they could not resolve them entirely. Fourth, the pattern of near-zero, non-significant associations in the cleaned and artifact-screened data was nearly the same as in the uncleaned session averages, indicating that these null findings are not due to uncorrected measurement noise but are instead the result of the loss of information that occurs with aggregation at the session level. Finally, quality control resulted in 96 sessions being retained from 100 sessions in the original study to limit generalizability. Other restrictions also should be recognized. The present study was exclusively secondary, meaning that the original recruitment, instrumentation, and protocol design were out of the control of the analysis. It was unable to gather further physiological, contextual, or self-reported information which might help to account for the variation in the movement–physiology relationship in individuals. Other demographic data, including fitness level, medication use, psychiatric history, and more, were not available and would have allowed for additional moderators to be investigated. Because the overall sample size ($N = 96$) was small and the per-condition sample sizes ($n = 30\text{--}36$) were even smaller, statistical power was limited to detect small effects, and more complex multilevel and time-series modeling was not feasible. Additionally, wrist-worn photoplethysmographic and electrodermal sensors are also known to be less accurate than research-grade or chest-worn sensors, especially when moving, which may have added measurement error beyond the scope of the current pipeline used to remove artifacts. Last, all data were obtained in structured supervised laboratory settings; results might not apply to the natural environment in which consumer wearables are used. This is confirmed by the condition-specific analysis above (see Condition-Specific Analysis).

Aerobic and anaerobic exercise yielded significantly higher average HR and EDA than the stress induction protocol, indicating that the dataset consists of two physiologically different experimental states at the same level of a single session. Having combined these, as in the primary correlation and regression models, the relationship between movement and physiology by condition may obscure the relationships of physiological arousal and movement within each particular condition; for example, a positive relationship between movement and physiology in the aerobic situation may be lost when compared to the other conditions because of high cognitive and emotional demands without an increase in movement. The present condition-specific correlations were not significant; however, the sample size for each condition ($n = 30-36$) was too low to yield sufficient power for these comparisons. Experimental condition should thus not be considered as a nuisance variable but rather as a potential moderator of the relationship between movement and the physiologic signals used to detect stress, and repeated-measures or multilevel time-series models that estimate the relationship between movement and physiologic signal separately in each condition and each participant over time would be more appropriate than session-level cross-sectional correlation to identify condition-specific coupling between movement and physiologic signal.

Future studies should utilize time-resolved or repeated measures designs to investigate HR, EDA, and TEMP in the immediate aftermath of increased levels of physical activity. The introduction of an experimental condition (e.g., rest, stress, aerobic exercise, anaerobic exercise) as a factor could better elucidate the time at which physical activity is a significant factor in wearable physiological stress indicators. These would offer a more realistic assessment of the dynamics of the movement – physiological stress relationship.

Implications for Wearable Stress-Detection Systems

The results have several implications for designing consumer- and research-grade wearable stress-detection algorithms, but should not be confused with validation or evaluation of any such system. But far from proving that movement is unimportant in measuring stress, the results point to several preprocessing and interpretation problems that algorithm developers will need to solve: summaries of the accelerometer at the level of a session or otherwise heavily aggregated may be an inadequate foundation for detecting movement-related from stress-related physiological changes, while the robust analyses and condition-specific associations reported here indicate that finer-grained, condition-aware or rank-based models may be more likely to recover associations that simple linear, fully aggregated models do not. This suggests the development of multimodal algorithm designs that can utilize the data from the accelerometer as a parallel time-resolved input to the model that are incorporated with HR, EDA and TEMP, because these signals are not ultimately affected by movement and can therefore be used to differentiate physiological responses to psychological stress from responses to physical activity once the movement context is available in parallel and at the same temporal resolution as the physiological measures. Adding individual-level calibration, condition or activity-state labels, or temporal features (e.g., recent activity history, rate of change in HR or EDA) could further improve the ability of such systems to distinguish psychological stress from exercise-induced arousal. To sum up, the present study is not research to suggest the acceptance or rejection of any particular stress-detection algorithm; rather, it is

research that can help to guide the preprocessing and interpretive assumptions that are built into the wearable stress-detection research.

CONCLUSION

The goal of this study was to assess the validity of the prediction of three physiological measures that are frequently interpreted as stress measures in wearable sensor studies: heart rate, electrodermal activity, and skin temperature, using physical activity intensity aggregated at the session level. There were small correlations between physical activity intensity and all three outcomes that were not statistically significant, and moderate correlations between heart rate and electrodermal activity but not with any of the three outcomes across 96 quality-controlled participant sessions (from 100 originally recorded) in a wearable dataset that included structured stress-induction and exercise protocols. The findings indicate that these measures of movement intensity are not necessarily a reliable indicator of the physiological changes that are generally believed to underlie stress, as the aggregation of sessions removes temporal information, which may mask a relationship between movement and physiology that occurs on a much finer timescale. In future, physical activity should be modeled directly with HR, EDA, and TEMP, in order to improve the capacity of wearable monitoring systems to separate out true psychological responses from physiological responses to physical activity, and to support the evidence for applied uses of consumer and research-grade wearables in stress-monitoring.

REFERENCES

- Abd Al-Alim, M., Mubarak, R., Salem, N. M., & Sadek, I. (2024). A machine-learning approach for stress detection using wearable sensors in free-living environments. *Computers in Biology and Medicine*, 179, Article 108918. <https://doi.org/10.1016/j.compbimed.2024.108918>
- Anwer, S., Li, H., Antwi-Afari, M. F., Umer, W., & Wong, A. Y. L. (2020). Cardiorespiratory and thermoregulatory parameters are good surrogates for measuring physical fatigue during a simulated construction task. *International Journal of Environmental Research and Public Health*, 17(15), Article 5418. <https://doi.org/10.3390/ijerph17155418>
- Blokker, T., Bucher, E., Steiner, T., & Wehrlin, J. P. (2022). Effect of cold ambient temperature on heat flux, skin temperature, and thermal sensation at different body parts in elite biathletes. *Frontiers in Sports and Active Living*, 4, Article 966203. <https://doi.org/10.3389/fspor.2022.966203>
- Boucsein, W. (2012). *Electrodermal activity* (2nd ed.). Springer.
- Can, Y. S., Chalabianloo, N., Ekiz, D., & Ersoy, C. (2019). Continuous stress detection using wearable sensors in real life: Algorithmic programming contest case study. *Sensors*, 19(8), Article 1849. <https://doi.org/10.3390/s19081849>
- Caplin, A., Chen, F. S., Beauchamp, M. R., & Puterman, E. (2021). The effects of exercise intensity on the cortisol response to a subsequent acute psychosocial stressor. *Psychoneuroendocrinology*, 131, Article 105336. <https://doi.org/10.1016/j.psyneuen.2021.105336>

- Chiang, J.-K., Lin, Y.-C., Hung, T.-Y., Kao, H.-H., & Kao, Y.-H. (2024). The impact on autonomic nervous system activity during and following exercise in adults: A meta-regression study and trial sequential analysis. *Medicina*, 60(8), Article 1223. <https://doi.org/10.3390/medicina60081223>
- Childs, E., & de Wit, H. (2014). Regular exercise is associated with emotional resilience to acute stress in healthy adults. *Frontiers in Physiology*, 5, Article 161. <https://doi.org/10.3389/fphys.2014.00161>
- Critchley, H. D. (2002). Electrodermal responses: What happens in the brain. *The Neuroscientist*, 8(2), 132–142. <https://doi.org/10.1177/107385840200800209>
- Dawson, M. E., Schell, A. M., & Filion, D. L. (2017). The electrodermal system. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of psychophysiology* (4th ed., pp. 217–243). Cambridge University Press.
- Gedam, S., & Paul, S. (2021). A review on mental stress detection using wearable sensors and machine learning techniques. *IEEE Access*, 9, 84045–84066. <https://doi.org/10.1109/ACCESS.2021.3085502>
- Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23), e215–e220. <https://doi.org/10.1161/01.CIR.101.23.e215>
- Hill, E. E., Zack, E., Battaglini, C., Viru, M., Viru, A., & Hackney, A. C. (2008). Exercise and circulating cortisol levels: The intensity threshold effect. *Journal of Endocrinological Investigation*, 31(7), 587–591. <https://doi.org/10.1007/BF03345606>
- Hongn, A., Bosch, F., Prado, L., & Bonomini, P. (2025). *Wearable device dataset from induced stress and structured exercise sessions* (Version 1.0.1) [Data set]. PhysioNet. <https://doi.org/10.13026/he0v-tf17>
- Kirschbaum, C., Pirke, K.-M., & Hellhammer, D. H. (1993). The ‘Trier Social Stress Test’—A tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology*, 28(1–2), 76–81. <https://doi.org/10.1159/000119004>
- Nelson, B. W., & Allen, N. B. (2019). Accuracy of consumer wearable heart rate measurement during an ecologically valid 24-hour period: Intraindividual validation study. *JMIR mHealth and uHealth*, 7(3), Article e10828. <https://doi.org/10.2196/10828>
- Posada-Quintero, H. F., Florian, J. P., Orjuela-Cañón, A. D., Aljama-Corrales, T., Charleston-Villalobos, S., & Chon, K. H. (2016). Power spectral density analysis of electrodermal activity for sympathetic function assessment. *Annals of Biomedical Engineering*, 44(10), 3124–3135. <https://doi.org/10.1007/s10439-016-1606-6>
- Schmidt, P., Reiss, A., Duerichen, R., Marberger, C., & Van Laerhoven, K. (2018). Introducing WESAD, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction* (pp. 400–408). Association for Computing Machinery. <https://doi.org/10.1145/3242969.3242985>
- Schuurmans, A. A. T., de Looft, P., Nijhof, K. S., Rosada, C., Scholte, R. H. J., Popma, A., & Otten, R. (2020). Validity of the Empatica E4 wristband to measure heart rate variability (HRV) parameters: A comparison to electrocardiography (ECG). *Journal of Medical Systems*, 44, Article 190. <https://doi.org/10.1007/s10916-020-01648-w>

- Shaffer, F., & Ginsberg, J. P. (2017). An overview of heart rate variability metrics and norms. *Frontiers in Public Health*, 5, Article 258. <https://doi.org/10.3389/fpubh.2017.00258>
- Taylor, S., Jaques, N., Chen, W., Fedor, S., Sano, A., & Picard, R. (2015). Automatic identification of artifacts in electrodermal activity data. In *Proceedings of the 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 1934–1937). IEEE.
- Wu, W.-J., Yu, H.-B., Tai, W.-H., Zhang, R., & Hao, W.-Y. (2023). Validity of Actigraph for measuring energy expenditure in healthy adults: A systematic review and meta-analysis. *Sensors*, 23(20), Article 8545. <https://doi.org/10.3390/s23208545>