Proceedings of the
13th International Conference on Applied Informatics
Eger, Hungary, February 19–21, 2026
DOI: 10.17048/icai.2026.48
pp. 48–54

# Accuracy of Wearable Heart Rate Monitors under Static and High-Intensity Activity Conditions

**János Dávid Balogh, Attila Adamkó**

Faculty of Informatics, University of Debrecen
davidkm2@mailbox.unideb.hu
adamko.attila@inf.unideb.hu

**Abstract.** The widespread adoption of wearable devices has created new opportunities for continuous monitoring of physical activity, cardiovascular health, and sleep patterns. However, it remains an open question to what extent the physiological and kinematic data provided by commercial smartwatches can be considered accurate and clinically interpretable under different conditions. The aim of this study is to compare heart rate (HR), heart rate variability (HRV), step count, and sleep parameters measured by four commercial smartwatches against clinical-grade reference devices under both static/resting and dynamic activity conditions.

In a controlled study involving 30 participants, we evaluated four popular device models: Apple Watch Series 8, Garmin Forerunner 265, Samsung Galaxy Watch 6, and Xiaomi Smart Band 8. For heart rate in a static environment, the devices showed high correlation with the reference ($r > 0.92$) and low Mean Absolute Error (MAE $< 3$ bpm). During high-intensity physical activity, the HR error rate increased significantly across all models (MAE $= 8$–$16$ bpm). Step count accuracy was generally high during steady-state walking (error $< 3\%$), but variability increased during running. Regarding sleep parameters, the devices estimated Total Sleep Time (TST) with moderate accuracy, but sleep stage determination showed poor reliability compared to clinical polygraphy.

The results suggest that smartwatches are highly suitable for detecting long-term health trends and tracking general physical activity. However, their use for critical clinical decision-support requires careful interpretation, particularly regarding dynamic heart rate tracking and sleep architecture analysis. The study highlights the need for standardized multi-parameter validation protocols.

*Keywords:* Wearable devices, Heart rate monitoring, Photoplethysmography (PPG), Heart rate variability (HRV), Step count, Sleep tracking, Measurement accuracy, Motion artifacts, Clinical validation

# 1. Introduction

Wearable devices equipped with photoplethysmography (PPG) sensors and triaxial accelerometers have revolutionized the continuous monitoring of human physiology and daily activity. These consumer devices are increasingly utilized as preliminary screening tools for detecting cardiovascular anomalies, tracking physical exertion, and monitoring sleep quality [5, 13].

Despite their popularity, the clinical reliability of wrist-worn sensors remains controversial. Previous literature highlights that while PPG technology performs admirably under static or resting conditions, its accuracy degrades significantly during physical exertion due to motion artifacts, changes in skin contact pressure, and variations in peripheral blood perfusion [3, 6, 8]. Similarly, while accelerometer-based step counting is well-established, its accuracy can fluctuate depending on walking speed and arm kinematics. Furthermore, sleep tracking relies on a combination of actigraphy and HR/HRV algorithms, which often struggle to accurately differentiate between specific sleep stages when compared to clinical references [7].

While numerous studies have validated specific devices in isolated scenarios, there is a lack of comprehensive, multi-parameter evaluations focusing on HR, steps, and sleep across multiple devices in a unified methodological framework [12]. Therefore, the primary objective of this empirical study is to present a structured validation framework that quantitatively evaluates the reliability of four highly popular commercial smartwatches under strictly controlled conditions against standard clinical reference instruments.

# 2. Methods

## 2.1. Participants

A total of 30 healthy adult volunteers (15 male, 15 female; mean age $26 \pm 4$ years, mean BMI $23.5 \pm 2.1$) participated in the study. All participants provided informed consent prior to the measurements. The experimental protocol was approved by the institutional ethics committee.

## 2.2. Devices and Reference Standards

To ensure a comprehensive analysis, four commercially available smartwatches from leading manufacturers were evaluated simultaneously: Apple Watch Series 8, Garmin Forerunner 265, Samsung Galaxy Watch 6, and Xiaomi Smart Band 8.

To validate the smartwatch measurements, we utilized the following standard clinical reference devices:

- **Heart Rate & HRV:** A clinically validated ambulatory Holter ECG monitor, widely used in hospital settings for reliable beat-to-beat cardiovascular monitoring [11].
- **Step Count:** A clinically validated pedometer, supplemented by manual video-based step tallying to ensure ground-truth accuracy during the treadmill protocol.
- **Sleep Parameters:** A clinically validated polygraphy device (cardiorespiratory monitor), commonly used in clinical settings for sleep assessment, recording airflow, respiratory effort, and continuous oxygen saturation.

### 2.3. Experimental Protocol

The data collection was divided into three distinct phases:

1. **Static/Resting HR and HRV Phase:** Participants remained seated in a resting state for 15 minutes to establish a baseline with low temporal dynamics.
2. **Dynamic Activity and Step Count Phase:** Participants performed a treadmill protocol consisting of a 10-minute walk (5 km/h) followed by a 10-minute high-intensity run (10–12 km/h). This phase evaluated both the step-counting algorithms and the robustness of the HR sensors against motion artifacts.
3. **Sleep Monitoring Phase:** A subset of participants ($n = 15$) wore the smartwatches and the portable polygraphy system overnight in a controlled environment for a minimum of 7 hours.

### 2.4. Statistical Analysis

Data were synchronized using timestamps. Performance was evaluated using Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE), and Pearson's correlation coefficient ($r$). Statistical significance between measurement errors under different conditions was determined using paired t-tests, with $p < 0.05$ considered significant. Bland–Altman analysis was applied to assess the limits of agreement (LoA) [4].

## 3. Results

### 3.1. Heart Rate and HRV Accuracy

The evaluation yielded a clear dichotomy in HR measurement performance based on the physical activity level. Table 1 summarizes the quantitative metrics for each device compared to the ECG reference.

**Table 1.** Device-level HR Accuracy: Static vs. Dynamic Conditions.

| Device | Static MAE | Static $r$ | Dynamic MAE | Dynamic $r$ |
|---|---|---|---|---|
| Apple Watch Series 8 | 1.5 bpm | 0.98 | 8.5 bpm | 0.85 |
| Garmin Forerunner 265 | 1.8 bpm | 0.96 | 9.2 bpm | 0.82 |
| Samsung Galaxy Watch 6 | 2.2 bpm | 0.95 | 14.5 bpm | 0.72 |
| Xiaomi Smart Band 8 | 2.5 bpm | 0.92 | 16.1 bpm | 0.68 |
| **Average** | **2.0 bpm** | **0.95** | **12.1 bpm** | **0.77** |

In the static environment, all smartwatches demonstrated a very high correlation with the reference and an exceptionally low MAE. Conversely, high-intensity running resulted in a statistically significant increase in measurement error across all models ($p < 0.001$). The Apple Watch and Garmin managed motion artifacts slightly better than the Samsung and Xiaomi devices, but the degradation in accuracy was universal. Regarding HRV, dynamic activity profoundly distorted the pulse-to-pulse intervals (PPIs), making HRV data derived during intense exercise statistically unreliable compared to resting HRV [2, 10].

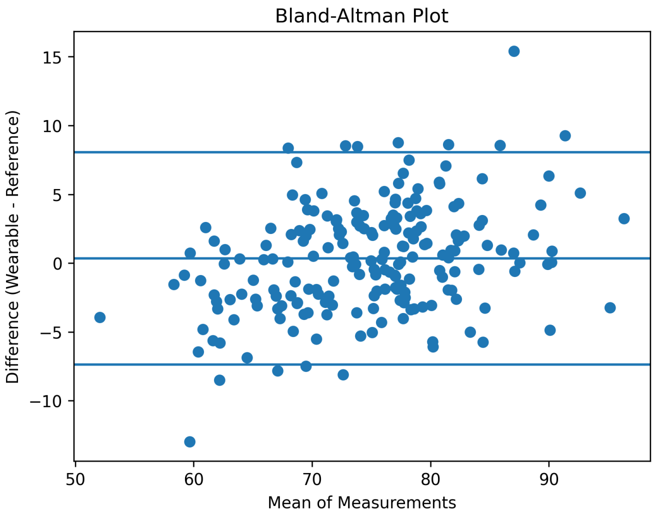


**Figure 1.** Bland–Altman plot for HR measurements. Wider limits of agreement during dynamic activity (dashed lines) illustrate the increased error margin and vulnerability to motion artifacts.

## 3.2. Step Count Accuracy

Step counting proved to be the most reliable metric across the evaluated devices. Table 2 shows the Mean Absolute Percentage Error (MAPE) against the clinically validated pedometer and video reference.

**Table 2.** Step Count Accuracy (MAPE %).

| Device | Walking (5 km/h) | Running (12 km/h) |
|---|---|---|
| Apple Watch Series 8 | 1.2% | 2.5% |
| Garmin Forerunner 265 | 1.5% | 2.8% |
| Samsung Galaxy Watch 6 | 2.1% | 4.2% |
| Xiaomi Smart Band 8 | 2.5% | 4.8% |

During steady-state walking, the error was negligible (MAPE $< 3\%$ for all devices). During running, the error increased slightly due to changes in arm swing kinematics, but overall step detection remained highly robust and suitable for daily physical activity tracking.

### 3.3. Sleep Parameter Accuracy

Sleep tracking accuracy varied significantly depending on the metric evaluated. Table 3 compares the wearable estimates of Total Sleep Time (TST) and Sleep Efficiency (SE) against the clinical polygraphy data [7].

**Table 3.** Sleep Tracking Accuracy vs. Clinical Polygraphy.

| Device | TST Error (Bias ± SD) | SE Error (Bias ± SD) |
|---|---|---|
| Apple Watch Series 8 | +15 ± 20 min | +2.5 ± 4.0% |
| Garmin Forerunner 265 | +12 ± 18 min | +1.8 ± 3.5% |
| Samsung Galaxy Watch 6 | +22 ± 25 min | +3.2 ± 5.1% |
| Xiaomi Smart Band 8 | +30 ± 28 min | +4.5 ± 6.2% |

All devices systematically overestimated Total Sleep Time by 12 to 30 minutes, primarily due to misclassifying periods of quiet wakefulness as light sleep. Furthermore, while basic sleep duration was moderately accurate, sleep stage determination showed poor agreement with the clinical reference, with Cohen's kappa values ranging from 0.35 to 0.48 across the devices.

## 4. Discussion

The results of this study confirm that commercial smartwatches offer a mixed profile of reliability depending on the measured parameter and the physical state of the user. Step counting remains highly accurate regardless of intensity, validating these devices as excellent tools for general activity promotion.

For cardiovascular metrics, our findings align with previous literature [6, 9]: in static environments, PPG sensors closely match clinical ECG references, supporting their use for resting HR and baseline HRV monitoring. However, the MAE increase to 8–16 bpm during high-intensity running indicates that these devices should not be solely relied upon for real-time clinical decision-support or precise sports-health diagnostics during exertion.

Sleep tracking results present a similar duality. While smartwatches provide a reasonable estimation of overall sleep duration (TST), their reliance on wrist actigraphy and basic HR patterns limits their ability to accurately map complex sleep architecture [7]. Therefore, clinical sleep diagnostics still firmly require clinical-grade polygraphy or polysomnography.

Future improvements in the wearable industry must target robust motion-compensation algorithms for PPG and multi-sensor fusion for sleep staging. Furthermore, the validation of physiological data streams should be integrated into modular system architectures that can continuously assess data quality and flag anomalies caused by sensor noise before the data is fed into predictive health informatics models [1].

# 5. Conclusion

Wearable devices demonstrate high accuracy in tracking step counts and resting heart rates, making them valuable tools for general wellness tracking and long-term physiological trend detection.

However, during high-intensity physical activity, HR measurement error increases significantly due to motion artifacts. Similarly, while overnight sleep duration is estimated with acceptable accuracy, advanced sleep staging remains unreliable compared to clinical standards.

Future research should focus on:

- Developing advanced adaptive filtering to reduce PPG noise during vigorous activity.
- Enhancing sleep algorithms by fusing continuous HRV, temperature, and actigraphy data.
- Standardizing ambulatory validation protocols across diverse populations and physical activities.
- Enhancing algorithmic transparency, allowing independent scientific verification of proprietary signal processing methodologies.

In conclusion, while commercial smartwatches are excellent for non-critical health monitoring, integrating their dynamic HR and sleep measurements into high-stakes clinical workflows necessitates a cautious approach, robust validation, and awareness of their technological limitations.

# References


[1] J. D. Balogh, A. Adamkó: *eHealth and Smart Solutions Framework for Health Monitoring in the Course of the Pandemic*, Annales Mathematicae et Informaticae (2023), doi: 10.33039/ami.2023.08.013.

[2] C. Bellenger, D. Miller, S. Halson, G. Roach, C. Sargent: *Wrist-Based Photoplethysmography Assessment of Heart Rate and Heart Rate Variability*, Sensors 21.10 (2021), p. 3571, doi: 10.3390/s21103571.

[3] B. Bent, B. Goldstein, W. Kibbe, J. Dunn: *Investigating Sources of Inaccuracy in Wearable Optical Heart Rate Sensors*, npj Digital Medicine 3 (2020), p. 18, doi: 10.1038/s41746-020-0226-6.

[4] J. M. Bland, D. G. Altman: *Statistical methods for assessing agreement between two methods of clinical measurement*, The Lancet 327.8476 (1986), pp. 307–310, doi: 10.1016/S0140-6736(86)90837-8.

[5] M. Etiwy, Z. Akras, S. Gillinov, D. Phelan, R. Wang, G. Blackburn, P. Houghtaling, H. Javadikasgari, M. Desai: *Erratum to accuracy of wearable heart rate monitors in cardiac rehabilitation*, Cardiovascular Diagnosis and Therapy 10.3 (2020), pp. 644–645, doi: 10.21037/cdt.2019.10.05.

[6] S. Gillinov, M. Etiwy, et al.: *Variable accuracy of wearable heart rate monitors during aerobic exercise*, Medicine & Science in Sports & Exercise 49.8 (2017), pp. 1697–1703, doi: 10.1249/MSS.0000000000001284.

[7] S. Haghayegh, S. Khoshnevis, M. H. Smolensky, K. R. Diller: *Accuracy of PurePulse photoplethysmography technology of Fitbit Charge 2 for assessment of heart rate during sleep*, Chronobiology International 36.7 (2019), pp. 927–932, doi: 10.1080/07420528.2019.1596947.

[8] J. Parak, I. Korhonen: *Evaluation of wearable consumer heart rate monitors based on photoplethysmography*, in: 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, Chicago, IL, USA, 2014, pp. 3670–3673, doi: 10.1109/EMBC.2014.6944419.

[9] R. Quinn, N. Leader, G. Lebovic, C.-M. Chow, P. Dorian: *Accuracy of Wearable Heart Rate Monitors During Exercise in Sinus Rhythm and Atrial Fibrillation*, Journal of the American College of Cardiology 83.12 (2024), pp. 1177–1179, doi: 10.1016/j.jacc.2024.01.024.

[10] F. Shaffer, J. P. Ginsberg: *An Overview of Heart Rate Variability Metrics and Norms*, Frontiers in Public Health 5 (2017), p. 258, doi: 10.3389/fpubh.2017.00258.

[11] R. Z. Ur Rehman, M. Chatterjee, N. V. Manyakov, et al.: *Assessment of Physiological Signals from Photoplethysmography Sensors Compared to an Electrocardiogram Sensor: A Validation Study in Daily Life*, Sensors 24.21 (2024), p. 6826, doi: 10.3390/s24216826.

[12] C. N. Van Oost, F. Masci, A. Malisse, et al.: *Accuracy of Heart Rate Measurement Under Transient States: A Validation Study of Wearables for Real-Life Monitoring*, Sensors 25.20 (2025), p. 6319, doi: 10.3390/s25206319.

[13] M. P. Wallen, S. R. Gomersall, S. E. Keating, U. Wisløff, J. S. Coombes: *Accuracy of heart rate watches: Implications for weight management*, PLoS ONE 11.5 (2016), e0154420, doi: 10.1371/journal.pone.0154420.