Validation of the Samsung Smartwatch for Sleep–Wake Determination and Sleep Stage Estimation

Dongyeop Kim1, Eun Yeon Joo2, Su Jung Choi3

1Department of Neurology, Seoul Hospital, Ewha Womans University College of Medicine, Seoul, 2Department of Neurology, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, 3Graduate School of Clinical Nursing Science, Sungkyunkwan University, Seoul, Korea

Received March 1, 2023
Revised March 28, 2023
Accepted April 21, 2023

Address for correspondence
Su Jung Choi, RN, PhD
Graduate School of Clinical Nursing Science, Sungkyunkwan University, 115 Irwon-ro, Gangnam-gu, Seoul 06355, Korea
Tel: +82-2-2148-9927
Fax: +82-2-2148-9949
E-mail: sujungchoi@skku.edu

Objectives: Galaxy Watch 3 (GW3) is a commercially available smartwatch equipped with a sleep-tracking function capable of collecting longitudinal sleep data in a real-world environment. We aimed to investigate the validity of GW3 for estimating sleep stages compared with reference data from polysomnography (PSG).

Methods: Thirty-two healthy adults (mean age 37.8, male 87.5%) were recruited to wear a GW3 concurrently with in-laboratory overnight PSG recording. Sleep parameters, including total sleep time (TST) and the duration of each sleep stage (light, deep, and rapid eye movement [REM] sleep), were calculated for both GW3 and PSG. Sleep parameters were compared using intraclass correlation coefficients (ICCs) and Bland–Altman plots. The epoch-by-epoch classification performance was evaluated to determine the sensitivity, specificity, accuracy, kappa values, and confusion matrices.

Results: Bland–Altman plots showed moderate agreement between GW3 and PSG for TST (ICC=0.640), light sleep (ICC=0.518), and deep sleep (ICC=0.639), whereas REM sleep duration was not reliably estimated using the GW3. The GW3 overestimated TST by a mean of 9.5 min. The sensitivity of epoch-by-epoch sleep detection was 0.954; however, the specificity was 0.524. The sensitivity of each sleep stage estimation was 0.695 for light sleep, 0.612 for deep sleep, and 0.598 for REM sleep. The overall accuracy of GW3 in distinguishing the four-stage sleep epochs was 0.651.

Conclusions: GW3 demonstrated high performance in sleep detection but moderate performance in wake determination and sleep stage estimation compared with PSG results, which were comparable to previously reported results for other consumer wearable devices.

Keywords: Wearable electronic devices; Actigraphy; Polysomnography; Sleep stages; Validation.

INTRODUCTION

Activity and sleep trackers have become popular among the general population. Over the last two decades, actigraphy has become a major assessment tool in sleep research and sleep medicine.1 Actigraphy is used to estimate sleep parameters over multiple nights in a home sleep environment, rather than measuring sleep overnight in a laboratory setting. However, conventional actigraphy does not provide information about sleep to the user while wearing the measurement device and is not updated in real time. Recently, various consumer-grade wearable devices operating with smartphone-specific applications are available.2 These applications focus on activity and sleep and enable individuals to monitor their overall health.3 They can also be used to improve patient empowerment when treating sleep disorders.

Most people are unable to objectively assess sleep quality when they want to know whether they are sleeping well. Although total sleep time (TST), sleep latency, number of awakenings, and sleep efficiency are recognized as objective indicators representing sleep quality, sleep architecture such as the percentage of sleep stages is also evaluated as one of the important factors of sleep quality.4 Polysomnography (PSG) is the standard method for measuring sleep stages; however, it has limitations in terms of accessibility, inconvenience, cost, and first-night effects. Wearable devices have the potential to compensate for the weaknesses of PSG, and various wearable smartwatches equipped with advanced sleep stage estimation algorithms are being released on the market.
Previous studies have reported a correlation between consumer wearable devices and PSG. For example, for Fitbit wearable devices, which have been widely investigated, de Zambotti et al.\(^7\) reported a sensitivity of 0.96 and specificity of 0.61 for sleep detection with Fitbit Charge 2\(^7\), correctly identifying 81% of light sleep, 49% of deep sleep, and 74% of rapid eye movement (REM) sleep, respectively. Additionally, studies evaluating and comparing the sleep stage classification functions of several consumer wearable devices have been published.\(^8\) The Galaxy Watch 3 (GW3) is a commercially available wrist-type wearable device equipped with an accelerometer and photoplethysmography sensor that can measure heart rate. Until now, sleep stage estimation performance using a Galaxy Watch has not been reported. In this study, we aimed to investigate the validity of GW3 in evaluating sleep stages and measuring sleep duration parameters compared with PSG results as the ground truth in healthy adults.

**METHODS**

**Participants**

Thirty-two healthy adults aged between 20 and 60 without complaints of sleep disturbance were prospectively recruited between September and December 2020 through advertisements at the Samsung Medical Center in Seoul, Korea. Participants with the following conditions were excluded: cardiovascular diseases (including myocardial infarction, congestive heart failure, and arrhythmia), neurological diseases (including stroke, epilepsy, and neurodegenerative disease), known sleep disorders (including insomnia, obstructive sleep apnea [OSA], narcolepsy, and restless legs syndrome), shift workers (at risk of sleep disorders), limited mobility due to muscular problems, and psychiatric diseases or treatment with psychotropic drugs.

**Study procedure**

On the day of the visit to the laboratory, questionnaires were issued to the participants to evaluate subjective sleep-related problems using the Epworth Sleepiness Scale (ESS),\(^7\) Insomnia Severity Index (ISI),\(^7\) and Pittsburgh Sleep Quality Index (PSQI).\(^9\) Participants with an ESS score of >10 were considered to have clinically significant daytime sleepiness. A total ISI score of 15 indicated significant insomnia, and a PSQI score of >5 as poor sleep quality. All participants wore a GW3 concurrently with PSG recording devices. For data analysis, sleep variables of PSG were measured beginning at the time set by the technicians for “lights on.” The participants were allowed to wake up at any time, and the end of monitoring was set by technicians for “lights on.” The study was approved by the Institutional Review Board of the Samsung Medical Center (SMC 2020-08-004) and conducted according to the Declaration of Helsinki.

**PSG and sleep parameters**

The PSG studies were performed with standard electrodes and sensors using an Embla N7000 (Medcare Flaga, Reykjavik, Iceland) by trained technicians. Electroencephalography electrodes were applied at C3-A2, C4-A1, F3-A2, F4-A1, O1-A2, and O2-A1, and four electrooculography electrodes were applied at both lateral sides, the superior and inferior of one eye, to record horizontal and vertical eye movements. Electromyography and electrocardiography sensors were used. Two plethysmography belts were used to monitor the thoracic and abdominal movements. Nasal and oral airflow were measured using a nasal pressure transducer and thermistor. Oxygen saturation was measured using a pulse oximeter attached to the index finger. Synchronized video monitoring was performed to monitor abnormal sleep breathing and movements. The wake and sleep stages for each 30-s epoch were assessed according to the rules of the American Academy of Sleep Medicine: N1, N2, and N3 non-REM sleep and REM sleep.\(^10\)

**Tested wearable device**

GW3 (Samsung Electronics, Suwon, Republic of Korea) is a commercially available wrist-type wearable device equipped with motion and heart rate sensors. The motion sensor measures the acceleration of the motion or vibration transmitted to the device. Heart rate data are obtained using photoplethysmography, which measures changes in blood flow in the microvascular bed of the skin. GW3 classified wake and sleep stages based on accelerometer data, and sleep stages were determined using an algorithm derived from a combination of plethysmography signals, heart rate variability, and accelerometer data. The sleep classifications were categorized as awake, light sleep (equivalent to N1 and N2 sleep from PSG), deep sleep (equivalent to N3 sleep from PSG), and REM sleep.

**Performance evaluation of GW3**

Sleep stages were scored every 30 s based on the American Academy of Sleep Medicine guidelines\(^10\) using both PSG and GW3. First, sleep parameters per subject were calculated using both PSG and GW3: TST, and the duration of each sleep stage was classified as light, deep, and REM sleep. Each sleep parameter was analyzed for agreement. Second, the estimated epoch from GW3 was compared with that from the PSG epoch by epoch after lights off across all epochs and per subject. The classification performance of sleep stages using GW3 was evaluated.

**Statistical analyses**

Baseline clinical characteristics of the participants were sum-

---

\(^6\) Kim D et al.

\(^7\) Zambotti et al.

\(^8\) The Galaxy Watch 3 (GW3) is a commercially available wrist-type wearable device equipped with an accelerometer and photoplethysmography sensor that can measure heart rate.

\(^9\) The Galaxy Watch has not been reported. In this study, we aimed to investigate the validity of GW3 in evaluating sleep stages and measuring sleep duration parameters compared with PSG results as the ground truth in healthy adults.

---

https://www.e-jsm.org 29
marized using descriptive statistics. The agreement between the two sleep assessment methods was examined using intra-class correlation coefficients (ICCs) and 95% confidence intervals (CIs), which were computed using a two-way analysis of variance. ICCs values less than 0.5, between 0.5 and 0.75, between 0.75 and 0.9, and greater than 0.90 indicate poor, moderate, good, and excellent reliability, respectively. Bland–Altman plots were created to provide a graphic representation of the observed differences between the paired measurements. The classification performance of the sleep stages was evaluated epoch-by-epoch using the following indicators: sensitivity, specificity, accuracy, kappa value, positive predictive value, and negative predictive value. The confusion matrices of the two- and four-sleep stage classifications were presented. All statistical analyses were conducted using the Statistical Package for Social Science (SPSS) for Windows (version 22.0, IBM Corp., Armonk, NY, USA), and statistical significance was defined as \( p < 0.05 \).

**RESULTS**

**Characteristics of participants**

The demographic and polysomnographic parameters of the participants are summarized in Table 1. The mean age was 37.8±8.1 years and 87.5% of participants were males. Although the participants did not complain about their sleep, sleep-related problems were identified based on the questionnaire and PSG results. Ten (31.3%) were poor sleepers and seven (21.9%) had excessive daytime sleepiness. Based on PSG results, 16 participants were diagnosed with OSA: mild OSA (n=6), moderate OSA (n=7), and severe OSA (n=3).

**Agreement of sleep parameters between GW3 and PSG**

Moderate agreement between PSG and GW3 was found for TST (ICC=0.640, 95% CI 0.263–0.824), light sleep (ICC=0.518, 95% CI 0.013–0.765), deep sleep (ICC=0.639, 95% CI 0.260–0.824), and REM sleep (ICC=0.153, 95% CI -0.736–0.586) (Table 2). Regarding the TST, the PSG results and GW3 estimates differed only by a mean of 9.5 min. GW3 overestimated deep sleep by a mean of 11.6 min, REM sleep by a mean of 26.3 min, and underestimated light sleep by a mean of 28.6 min more than PSG. Among the sleep stages, a proportional bias was observed, showing the tendency for GW3 to overestimate REM sleep duration when the duration increased.

**Classification performance of sleep stages with epoch-by-epoch analysis**

The results of the confusion matrices were displayed as

| Table 1. Baseline characteristics of the participants (n=32) |
|-------------|-------------|-------------|-------------|-------------|-------------|
| Variables   | Values      |
| Age (yr)    | 37.8±8.1 (27–53) |
| Sex, male   | 28 (87.5) |
| Body mass index (kg/m²) | 25.7±4.3 (19.0–35.8) |
| PSQI        | 4.4±1.5 (2–7) |
| Poor sleeper, PSQI>5 | 10 (31.3) |
| ISI         | 5.8±3.9 (1–12) |
| ESS         | 7.5±3.9 (1–18) |
| Excessive daytime sleepiness, ESS >10 | 7 (21.9) |

<table>
<thead>
<tr>
<th>Polysomnographic parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>TST (min)</td>
</tr>
<tr>
<td>Sleep latency (min)</td>
</tr>
<tr>
<td>Sleep efficiency (%)</td>
</tr>
<tr>
<td>Wakefulness after sleep onset (%)</td>
</tr>
<tr>
<td>NREM1/TST (%)</td>
</tr>
<tr>
<td>NREM2/TST (%)</td>
</tr>
<tr>
<td>NREM3/TST (%)</td>
</tr>
<tr>
<td>REM/TST (%)</td>
</tr>
<tr>
<td>Arousal index (/h)</td>
</tr>
<tr>
<td>Oxygen desaturation index (/h)</td>
</tr>
<tr>
<td>Apnea-hypopnea index (/h)</td>
</tr>
</tbody>
</table>

Data are shown as mean±standard deviation (range) or n (%) unless otherwise indicated. PSQI, Pittsburg Sleep Quality Index; ISI, Insomnia Severity Index; ESS, Epworth Sleepiness Scale; TST, total sleep time; NREM, non-rapid eye movement sleep; REM, rapid eye movement sleep.

| Table 2. Agreement of sleep parameters between Galaxy Watch 3 and polysomnography using Bland-Altman plot results and intra-class correlation coefficient |
|-------------|-------------|-------------|-------------|-------------|-------------|
| Variables   | PSG         | GW3         | Bias        | 95% LoA     | ICC         |
| Total sleep time (min) | 370.0±47.1 | 379.5±42.8 | 9.5±46.3    | -81.2–100.3 | 0.640*      |
| Light sleep (min)      | 261.0±30.9  | 232.4±42.4 | -28.6±42.4  | -111.6–54.4 | 0.518*      |
| Deep sleep (min)       | 28.0±24.3   | 39.7±18.2  | 11.6±22.1   | -31.7–55.0  | 0.639*      |
| REM sleep (min)        | 80.8±26.7   | 107.1±43.6 | 26.3±49.0   | -69.7–122.2 | 0.153       |

Data are shown as mean±standard deviation unless otherwise indicated. Bias was calculated as the time estimated by GW3 minus the time measured by PSG. *\( p < 0.05 \). PSG, polysomnography; GW3, Galaxy Watch 3; LoA, limits of agreement; ICC, intraclass correlation coefficient; REM, rapid eye movement sleep.
Figure 1. Bland-Altman plots for evaluating the agreement between Galaxy Watch 3 (Samsung Electronics, Suwon, Republic of Korea) and polysomnography during total sleep time (A), light sleep (B), deep sleep (C), and rapid eye movement sleep (REM) sleep (D). The mean bias was depicted as a solid line and upper and lower limits of agreement (±1.96 standard deviations) as dotted lines. GW3, Galaxy Watch 3; PSG, polysomnography.

Figure 2. Confusion matrices showing the two-stage (A) and four-stage (B) sleep classifying ability of Galaxy Watch 3 (Samsung Electronics, Suwon, Republic of Korea) across all epochs. The intensity of the color reflects the accuracy of sleep stage estimation. REM, rapid eye movement sleep.

The overall accuracy of GW3 in distinguishing four stages of sleep was 0.651.

The epoch-by-epoch analyses for each subject are summarized in Table 3. For detailed sleep stage estimation, light sleep had the highest performance, with a mean sensitivity of 0.7, followed by deep sleep, REM sleep, and wake with mean sensitivities of 0.63, 0.58, and 0.51, respectively. Among sleep stages, the highest mean specificity of 0.93 was found at determining deep sleep, followed by REM and light sleep with mean specificities of 0.86 and 0.66, respectively.

https://www.e-ism.org 31
In this study, we evaluated the ability of GW3 to estimate sleep parameters and score sleep stages compared with the gold-standard PSG. GW3 exhibited moderate reliability in measuring sleep-related parameters, such as TST and duration of light and deep sleep, but did not reliably estimate REM sleep duration. When validating the ability to classify sleep stages using epoch-by-epoch analysis, GW3 had a high sensitivity of 0.954 but a low specificity of 0.524 for sleep detection. Among the sleep stages, it had the best performance for estimating light sleep, with a sensitivity of 0.695, followed by deep and REM sleep with sensitivities of 0.612 and 0.598, respectively. We attempted to recruit participants without any subjective sleep complaints and clinical sleep disorders. However, some participants were found to be poor sleepers and had significant daytime sleepiness based on the PSQI and ESS scores. OSA was confirmed with PSG in 16 participants. In addition, ten had N3 sleep for less than 1% of their TST, and the average arousal index was 19.5/h.

Making a direct comparison between GW3 and other sleep-tracking wearable devices is inadequate because these studies were not conducted using the same participants in the same experimental environment. Nevertheless, when compared with the sensitivity and specificity values reported in previous studies, GW3 can be considered comparable to other devices for assessing sleep. Previous studies have reported high sensitivity and low specificity for sleep detection, overestimating TST and sleep efficiency but underestimating wakefulness after sleep onset (WASO).

In a report comparing actigraphy (Philips Respironics Actiwatch 2) and four consumer wearable devices (Fatigue Science Readiband, Fitbit Alta HR, Garmin Fenix 5S, and Garmin Vivosmart 3) with PSG results, an epoch-by-epoch analysis revealed that the sensitivity for sleep detection was 0.94–0.99, and the specificity was measured as low as 0.18–0.54. For the devices providing information on further stages of sleep for light, deep, and REM sleep, the sensitivity was confirmed to be 0.68–0.76, 0.53–0.56, and 0.50–0.69, respectively. Consumer wearable devices performed better than actigraphy devices in estimating sleep parameters and stages. In addition, it has been reported that the estimation performance tends to be poor for cases with a shorter TST, lower sleep efficiency, longer sleep latency, and longer WASO, which implies a longer wake time. In another study comparing three consumer devices (Mi Band 2, Gearfit 2, and Fitbit Alta HR), the ICCs for the agreement of the TST were 0.20–0.297. Among them, the Fitbit Alta HR performed best for detailed sleep duration estimation, with ICCs of −0.19, 0.301, and 0.323 for light, deep, and REM sleep, respectively.

Because EEG is a critical tool for determining sleep stages under the current scoring system, it is inevitable that the sleep stage estimation function of wrist-worn devices is imperfect. Overestimation of sleep is a major limitation observed when attempting to assess sleep stages using wearable devices, and quiet wakefulness is a major contributor to this, as many previous studies have indicated. As the sleep–wake determination of GW3 is based on accelerometric data, GW3 is prone to errors in determining sleep when the user is lying still without moving. Recently, autonomic feature data, including heart rate variability, have been used to improve the ability to classify waking and sleeping states.

Sleep-tracking wearable devices can play an important role in future sleep research and sleep medicine. Conventional actigraph devices, such as the Actiwatch with movement-based sleep detection algorithms, have been commonly used in scientific research; however, sleep detection algorithms have not improved for years. Moreover, various wearable sensing modalities have been used for sleep staging, and they have exhibited progressively improved performance by updating the artificial intelligence-based algorithms for sleep detection in the recently developed consumer wearable devices. It has been previously reported that new consumer wearable devices perform as well as or better than conventional actigraphy devices. Consumer wearable devices cannot be used for diagnostic purposes because of a lack of U.S. Food and Drug Administration (FDA) approval and rigorous validation data; however, they are promising for enhancing patient–clinician interactions. Smartwatches are perceived as more convenient to use because of their user-friendly interface and ability to provide feedback to users through applications. They may also be helpful in identifying night-to-night and intraindividual variability in sleep metrics. If these devices are validated, real-world sleep
data for the daily environment can be obtained longitudinally in a large population.

This study had some limitations. First, several important sleep parameters, such as sleep latency or WASO, were not collected from GW3. Because GW3 has a low performance in distinguishing wake from sleep, we predicted that GW3 will be prone to errors in estimating wake-time-related parameters. Second, this study included participants with various clinical characteristics, from healthy subjects to patients with OSA, and we did not identify the clinical factors that affected the estimation performance of GW3. For example, episodes of sleep apnea may be accompanied by cyclic variation in heart rate; therefore, the estimation of sleep stages based on heart rate may be affected by sleep-disordered breathing.20,21 In addition, frequent sleep stage shifts owing to respiratory events may increase the probability that GW3 incorrectly estimates sleep stages. Many participants who had difficulty maintaining sleep and had poor sleep quality were enrolled in this study, and these factors may have contributed to the erroneous estimation of GW3. A previous validation study of a commercially available wearable device conducted in patients with OSA confirmed poor validity and limited performance in predicting sleep parameters.32 Third, because this study was conducted at a single clinic, different performance results may have been obtained for the same participants depending on the laboratory. This is owing to the potential inter-rater variability in PSG scoring, which can result in variations in the scoring of sleep stages between different raters. Moreover, this study did not directly compare the performance of GW3 with that of other wearable devices. Therefore, based solely on the results of this study, it is not possible to determine which device has a superior performance.

In conclusion, compared with the wearable devices used in previous studies, GW3 has comparable potential for predicting sleep parameters and classifying sleep stages. Because the classification performance of a device can vary depending on different settings and populations, it is necessary to identify the factors that may affect the performance under various sleep conditions. Although the accuracy of sleep stage estimation algorithms compared with standard PSG in consumer wearable devices still requires improvement, and the devices are not licensed for clinical diagnosis, the devices may play an important role in the future of sleep research because of their convenience, ability to assess consecutive nights, and potential for use in big data analysis.

Conflicts of Interest

Eun Yeon Joo, a contributing editor of the Journal of Sleep Medicine, was not involved in the editorial evaluation or decision to publish this article. All remaining authors have declared no conflicts of interest.

ORCID iDs

Dongyeop Kim https://orcid.org/0000-0001-6045-0693
Eun Yeon Joo https://orcid.org/0000-0003-1233-959X
Su Jung Choi https://orcid.org/0000-0003-2171-7441

Author Contributions


Funding Statement

This study was supported by a Samsung Medical Center Grant (OTC 1190671) and the smartwatch used in the study were provided by Samsung Electronics. The funders had no role in the data analysis, or the decision to publish.

Acknowledgments

The GW3 sleep data are not publicly available. We thank Jeong Yup Han from Samsung Electronics for helping with data collection.

REFERENCES

13. Walch O, Huang V, Forger D, Goldstein C. Sleep stage prediction with raw acceleration and photoplethysmography heart rate data derived

https://www.e-jsm.org