Twenty-four Hours of Sleep, Sedentary Behavior, and Physical Activity with Nine Wearable Devices

MARY E. ROSENBERGER1, MATTHEW P. BUMAN2, WILLIAM L. HASKELL3, MICHAEL V. MCCONNELL4, and LAURA L. CARSTENSEN1

1Stanford Center on Longevity and Psychology Department, Stanford University, Stanford, CA; 2School of Nutrition and Health Promotion, Arizona State University, Phoenix, AZ; 3Stanford Prevention Research Center, Stanford University, Stanford, CA; and 4Cardiovascular Medicine, Stanford School of Medicine, Stanford University, Stanford, CA

ABSTRACT
ROSENBERGER, M. E., M. P. BUMAN, W. L. HASKELL, M. V. MCCONNELL, and L. L. CARSTENSEN. Twenty-four Hours of Sleep, Sedentary Behavior, and Physical Activity with Nine Wearable Devices. Med. Sci. Sports Exerc., Vol. 48, No. 3, pp. 457–465, 2016. Getting enough sleep, exercising, and limiting sedentary activities can greatly contribute to disease prevention and overall health and longevity. Measuring the full 24-h activity cycle—sleep, sedentary behavior (SED), light-intensity physical activity (LPA), and moderate-to-vigorous physical activity (MVPA)—may now be feasible using small wearable devices. Purpose: This study compared nine devices for accuracy in a 24-h activity measurement. Methods: Adults (n = 40, 47% male) wore nine devices for 24 h: ActiGraph GT3X+, activPAL, Fitbit One, GENEactiv, Jawbone Up, LUMOback, Nike Fuelband, Omron pedometer, and Z-Machine. Comparisons (with standards) were made for total sleep time (Z-machine), time spent in SED (activPAL), LPA (GT3X+), MVPA (GT3X+), and steps (Omron). Analysis included mean absolute percent error, equivalence testing, and Bland–Altman plots. Results: Error rates ranged from 8.1% to 16.9% for sleep, 9.5% to 65.8% for SED, 19.7% to 28.0% for LPA, 51.8% to 92% for MVPA, and 14.1% to 29.9% for steps. Equivalence testing indicated that only two comparisons were significantly equivalent to standards: the LUMOback for SED and the GT3X+ for sleep. Bland–Altman plots indicated GT3X+ had the closest measurement for sleep, LUMOback for SED, GENEactiv for MVPA, Fitbit for SED, and GT3X+ for steps. Conclusions: Currently, no device accurately captures activity data across the entire 24-h day, but the future of activity measurement should aim for accurate 24-h measurement as a goal. Researchers should continue to select measurement devices on the basis of their primary outcomes of interest. Key Words: ACTIGRAPH, GENEACTIV, ACTIVPAL, ACCELEROMETERS, ACTIVITY MONITORS, FITBIT

S

ubstantial evidence has led to recommendations for adequate exercise, healthy sleep habits, and limited sedentary behavior for increased longevity, improved health, and disease prevention (7,14,21). Health research has focused intensely on these different daily activities, but for researchers, clinicians, and consumers to understand better these activity–health relations, it is important to study the complete 24-h activity cycle. Combined measurement of sleep, sedentary behavior, and physical activity may be an important step in guiding activity recommendations throughout a 24-h cycle. Current activity and sleep guidelines are limited to 30 min d−1 of exercise and 7–8 h of sleep, leaving about 16 h of unaccounted time with a nonquantified recommendation to avoid too much sitting.

The components of the 24-h model, organized into domains of activity intensity, are sleep, sedentary behaviors (SED), light-intensity physical activity (LPA), and moderate-to-vigorous physical activities (MVPA or “exercise”). For all nonsleep activities, SED is defined as sitting or lying with energy expenditure less than 1.5 METs (32), LPA would include activities with energy expenditure between 1.5 and 3 METs (1), and MVPA includes moderate activity (3–6 METs) and vigorous activity (any activity greater than 6 METs) (1). A 24-h model of activity was previously difficult to measure, and incorporating the model into medical research was limited because of the error associated with the measurement. First, sleep, SED and physical activity are traditionally studied in separate laboratories. Second, measurement technology had both limited memory and short battery life. Lastly, there has been a lack of analytical methods to consider time spent in different activity levels and the relative relations to health outcomes.

Sleep recommendations—to sleep for 7–8 h per night—are based on observations that shorter or longer sleep duration is associated with risk factors for a range of diseases.
SED recommendations are sparse (31), but objective monitoring of SED has revealed relations to several health outcomes (21), and several general recommendations have been published (13,39). Exercise is also related to multiple health outcomes (14), and this has led to public health recommendation of 150 min wk\(^{-1}\) of MVPA to contribute substantially to longevity and disease prevention (14). Increased LPA is associated with improved energy expenditure (26,27) and physical health and well-being measures in older adults (5). Decreased LPA contributes to several health risks including elevated plasma glucose (15) and higher blood pressure and lower HDL cholesterol (8) but not mortality rates (24). There are no recommendations for how much of the day should be spent in LPA compared with SED.

Importantly, the relations among these activity domains are not well understood. For example, physical activity can be used as a treatment for poor sleep (6), but research has not addressed the need for more sleep (or sedentary time) as recovery after several days of extended vigorous-intensity exercise. The relation among activity domains is also probably not stagnant, but changes across the life span, during specific physiological or disease states (i.e., pregnancy, diabetes), and with heavier physical training loads. Accurate and reliable measurement of the 24-h cycle could answer many of these specific research questions that cannot be addressed with current measurement methods.

The collection of objective measures of sleep, SED, LPA, and MVPA has traditionally been costly, difficult, or nonexistent. Technological advances now make these measurements possible using small wearable devices. There has been a proliferation of wearable devices for the various components of daily activity, but there has been minimal research into how these devices compare with one another and how valid and reliable they are compared with common research measurement methods. Figure 1 provides a representation of a 24-h cycle of activity with current recommendations and a rough estimate of the proportion of SED to LPA. The 24-h model is used in this study as a framework to help evaluate what these devices measure in each of the four activity domains. It should be noted that most devices are unable to produce this 24-h chart with their current reported data. The purpose of this study was to compare the output from commercially available wearable devices using current standards for objective measurement of sleep, SED, LPA, and MVPA in the field. The ultimate goal of this research was to determine the best ways to measure the full 24 h of activity behavior to guide future clinical studies and recommendations.

**METHODS**

**Participants.** Participants were recruited from the Stanford University community and surrounding areas through word of mouth with an effort to include equal numbers of men and women over a wide age range. Before participation, all participants signed a written informed consent approved by the Stanford University institutional review board. Participants (n = 40, 21 women) came to the laboratory for instructions, initialization, and device fitting, then wore the devices for 24 consecutive hours during normal activities, and returned to the laboratory on the second day to return the devices. The mean age of the participants was 36 yr (the range was 21–76 yr).

**Standards for free-living activity measurements.** Measuring activity domains over the 24-h day cannot be limited to specific activities that can be measured in a laboratory but is dependent on measuring free-living activities. The standards selected for comparisons in this study were not laboratory-based gold-standard devices but the closest standard that could be conveniently worn during a complete 24-h cycle in a free-living environment. The Z-machine measures brain activity with an electroencephalogram in a portable monitor and is thus a more comparable measurement with polysomnography, the laboratory-based measure of brain activity, than actigraphy, or an accelerometer on the wrist (20). For SED, posture measurement is a key component of the definition, which involves sitting or lying while awake with an energy expenditure of less than 1.5 METs, so the actiPAL monitor was the standard for this domain (23). The ActiGraph GT3X+ is a frequently used device for LPA and MVPA measurement (40). The Omron pedometer was selected as the standard because it has been validated as an accurate measure of steps (17) and is independent of our other standard devices. For example, the GT3X+ is a standard for
other domains in this comparison but it is not regularly used as the step counter in epidemiological studies.

**Measurements.** In addition to the above devices used as standards, the following wearable devices were studied: the Fitbit One, Jawbone Up, Nike Fuelband, GENEactiv, and LUMOback. Table 1 shows a listing and description of the nine devices worn in this study. Devices were selected to represent both research devices and commercially available devices that were in widespread use at the outset of the study and measured at least one domain of the 24-h cycle with some specificity.

At the beginning of the study, participants came to the laboratory where height, weight, age, and gender were collected and recorded. Software described in Table 1 was used to submit participant-specific information to each device for initialization. Participants also received both written and oral instructions of when to put on the devices and how to wear them. The LUMOback also required initial calibration, where the participant walks and then sits in a slouching position while following directions on the mobile device. This was performed in the laboratory using an iPhone 4S, connected to the LUMOback via Bluetooth, and the participant was guided directly by the application on the phone. After initialization, a study kit was prepared for the participant. It included all nine of the devices plus both a hip and wrist strap for the GT3X+; one clip and one strap for the Fitbit; alcohol wipes, extra electrodes, electrode cables, and the user manual (supplied by General Sleep Corporation) for the Z-Machine; a clip and a leash for the Omron; and several stickies for the activPAL.

Participants were asked to wear all nine devices for a day consisting of one full day of activity and one full night of sleep. Devices were worn from approximately the time a participant woke up until the participant woke up the next morning. If the participant did not wake up at the same time on the two consecutive days, more or less than 24 h are recorded. A daily log was used to record when the participant woke up, what time the devices were put on, if they were taken off for bathing or water activities, when the participant got into bed for the purpose of sleeping, and when the participant woke up and took off the devices. A verbal follow-up was also conducted when the participants returned the devices to confirm whether times were accurately recorded.

During daily and nightly wear, device feedback was not provided to the participant except in cases where the data were presented on the device itself. Omron has a steps display, the Fuelband displays steps, and Nike Fuel, and the Fitbit displays steps, floors climbed, calories burned, and activity level. All other devices did not provide feedback to the user. No interventions were introduced such as step goals, vibrations to interrupt SED, or other guidelines for the participant.

Device data were downloaded after the participant returned the study kit. Participants could view their data after the conclusion of their participation if they were willing to stay through data download. No written reports were provided to the participant. Data were downloaded either to the computer (Fitbit, GT3X+, Fuelband, and activPAL) or through the phone application (LUMOback and Jawbone) for devices that lack desktop software. In addition, a separate research portal, provided by the company, was used to download data from the LUMOback to obtain 5-min epoch summaries, which are not provided by the consumer phone application.

**Sleep.** Devices compared with the Z-machine for measuring sleep duration included the Fitbit, Jawbone, GENEactiv, and GT3X+. All of these were worn for the entire 24-h period with the exception of the Z-Machine (only during sleep periods). The Z-Machine uses three electrodes on the head/neck. Calibration of the Z-machine included inputs of height, weight, and age through a computer connected to the device. Once initialized, the user could apply the electrodes, check electrode connection, and start sleep measurement independently.

All other sleep measurement devices were worn on the wrist and rely on an accelerometer-based measurement algorithm to estimate total sleep time. Commercial devices have proprietary algorithms for sleep, so total sleep time was recorded directly from the summary. LUMOback and activPAL do not have specific sleep measurement because sedentary time and sleep are recorded on the basis of posture; therefore, these devices were not analyzed for total sleep time measurement. The Fitbit was moved from the trunk to the wrist and placed in a sweatband-style sleeve for sleep measurement. A button on the device was also pressed and held, putting the device into sleep mode, when the user got into bed for the purpose of sleeping. Similar buttons were used on the Jawbone and the GENEactiv to start sleep measurement. The GT3X+ was also moved from the waistband to the wrist in a specially designed sweat-style band with a pocket designed to hold the device. The GT3X+ does

---

**TABLE 1. A list of devices included in this study by company, versions used, and location worn.**

<table>
<thead>
<tr>
<th>Company</th>
<th>Device/Version</th>
<th>Company Location</th>
<th>Location Worn</th>
<th>Software</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Sleep</td>
<td>Z-Machine</td>
<td>Euclid, OH</td>
<td>Head electrodes</td>
<td>Z-machine Data Viewer</td>
</tr>
<tr>
<td>PAL Technologies Ltd</td>
<td>activPAL v2</td>
<td>Pensacola, FL</td>
<td>Right thigh</td>
<td>ActivPAL version 7.1.18</td>
</tr>
<tr>
<td>Actigraph LLC</td>
<td>GT3X+</td>
<td></td>
<td>Day: Right hip</td>
<td>Actilife 6</td>
</tr>
<tr>
<td>Omron Healthcare, Inc.</td>
<td>HJ-112 Pocket Pedometer</td>
<td>Lakeforest, IL</td>
<td>Right hip</td>
<td>On-screen summary</td>
</tr>
<tr>
<td>Fitbit</td>
<td>One</td>
<td>San Francisco, CA</td>
<td>Day: right hip or pocket</td>
<td>Desktop sync, online feedback, and iPhone app</td>
</tr>
<tr>
<td>Activinsights, Ltd.</td>
<td>GENEactiv Original</td>
<td>Kimbolton, Cambs, United Kingdom</td>
<td>Right wrist</td>
<td>GENExentic PCSoftware, version 2.2</td>
</tr>
<tr>
<td>Jawbone</td>
<td>Jawbone UP</td>
<td>San Francisco, CA</td>
<td>Right wrist</td>
<td>iPhone App</td>
</tr>
<tr>
<td>Lumo Bodytech, Inc.</td>
<td>LUMOback</td>
<td>Palo Alto, CA</td>
<td>Lower back</td>
<td>iPhone App</td>
</tr>
<tr>
<td>Nike, Inc.</td>
<td>Fuelband</td>
<td>Beaverton, OR</td>
<td>Right wrist</td>
<td>Desktop sync, online feedback, and iPhone App</td>
</tr>
</tbody>
</table>

---

**REFERENCES**


---

**Copyright © 2016 by the American College of Sports Medicine. Unauthorized reproduction of this article is prohibited.**
not “log” sleep with a button push; sleep time started when the participant started logging sleep on the Z-machine and stopped when the electrodes stopped recording. If the Z-machine malfunctioned because of user error, the sleep log as recorded by the participant was used to determine start and stop times of sleep.

Sleep can be measured using a variety of variables, but this comparison was limited to total sleep time because this is the variable universally measured by sleep devices and has also been shown to have a relation to health outcomes (7). A sleep-specific algorithm, specifically, the Sadeh sleep algorithm (36), was used to analyze data for the research devices (GT3X+ and GENeActiv). The commercial device summaries (Fitbit, Jawbone) were downloaded using the device-associated software. The raw data extracted from the activPAL on the thigh cannot be analyzed with the same Sadeh algorithm because it was developed for actigraphy on the wrist and the activPAL is worn on the thigh.

**Sedentary behavior.** Devices compared with the activPAL for measuring SED duration included the GT3X+, GENeActiv, LUMOback, and Fitbit. Total minutes spent in SED were found using the GT3X+ with a cut point of 150 counts per minute (23); the GENeActiv was worn on the right wrist with a cut point <217 gravity-subtracted minutes (g*min) (10); the LUMOback, with time spent in a sitting or lying posture; and the Fitbit, with sedentary time defined on the dashboard (this feature was included in the original reporting but was removed when “tiles” were added to the dashboard).

The activPAL was used as the standard and adheres to the definition of SED, which includes sitting or lying. Devices that are accelerometer-based (GT3X+, GENeActiv, and Fitbit) will be measuring a lack of motion, not posture. Early sedentary research relied on motion measurement, yet a posture-based definition has evolved. This comparison will provide insight into the differences between posture and motion-based sedentary measurement. Therefore, they could not be included in comparisons of time spent in LPA.

**LPA.** Devices compared with the GT3X+ for measuring LPA duration included the Fitbit and GENeActiv. A GT3X+ cut point of >150 and <1580 counts per minute was used as the standard (11), and was compared to a GENeActiv cut point of 217–644 g*min (10) and time spent in light activity from the Fitbit. None of the other devices measured LPA, nor could it be derived from time spent in other behaviors.

**MVPA.** Devices compared with the GT3X+ for measuring MVPA duration included the Jawbone, Fitbit, GENeActiv, and Fuelband. A GT3X+ cut point of ≥1580 counts per minute was used as the standard (11). The comparisons include active minutes from the Fuelband, active time from Jawbone, moderate plus vigorous minutes from the Fitbit, and a cut point of >644 from the GENeActiv (10). Other devices were not included in this comparison because they did not measure time spent in MVPA.

**Steps.** Devices compared with the Omron for measuring steps included the Jawbone, Fitbit, Fuelband, GT3X+, LUMOback, and activPAL. All devices reported total steps per day.

**Statistical analysis.** Table 2 summarizes the measurements provided by each device, which variables were used in this analysis, and what device was used as a criterion measure for each activity domain. Standard sample calculations were conducted to set goals for subject recruitment, and alpha was set at 0.05, with the confidence interval set to 95%. Separate sample calculations were conducted for each domain. Statistical analyses were performed to determine statistically significant differences and agreement among devices. Mean absolute percent errors (MAPE) are reported to establish differences between the devices and the “field-based” measurements and determine accuracy. In addition, equivalence testing is reported to establish similarities between the devices and measurement standards. Bland–Altman plots were used to test biases between the standards and the other measurement devices. These measurements of differences, similarities, and biases are similar to a recent study comparing devices with laboratory-based measurement of energy expenditure (25).

## Results

**Sleep duration.** Figure 2 illustrates the mean error analysis for the devices measuring sleep, ranging from 8.1% to 9.9% for the Fitbit and Jawbone, respectively, and from 8.1% to 11.1% for the Fuelband and GENeActiv, respectively. The differences between the devices are significant in terms of the percentage of error compared to the standard. The Fitbit has the lowest error rate, while the GENeActiv has the highest, indicating that the Fitbit is the most accurate device for measuring sleep duration.

### Table 2. Device variables reported from the nine devices included in the study.

<table>
<thead>
<tr>
<th>Device</th>
<th>Sleep</th>
<th>Sedentary</th>
<th>Light</th>
<th>Moderate/Vigorous</th>
<th>Steps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z Machine (Sleep)</td>
<td>Sleep/wake time*</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>activPAL</td>
<td>Included as Sedentary</td>
<td>Total sleep time</td>
<td>Sedentary, &lt;150</td>
<td>Standing time</td>
<td>N/A</td>
</tr>
<tr>
<td>Actigraph GT3X+</td>
<td>N/A</td>
<td>Sedentary time</td>
<td>Cut points, 150–1579*</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Omron</td>
<td>N/A</td>
<td>Sedentary time</td>
<td>Light time</td>
<td>Stepping time</td>
<td>N/A</td>
</tr>
<tr>
<td>Fitbit One</td>
<td>N/A</td>
<td>Sedentary time</td>
<td>Cut points, 150–1579*</td>
<td>Stepping time</td>
<td>N/A</td>
</tr>
<tr>
<td>GENeActiv</td>
<td>N/A</td>
<td>Sedentary time</td>
<td>Light time</td>
<td>Moderate steps/time</td>
<td>N/A</td>
</tr>
<tr>
<td>Jawbone Up</td>
<td>Total sleep time, % of goal, light sleep, latency, deep sleep, awake time, number of awakenings</td>
<td>Sedentary cut point, &gt;217</td>
<td>N/A</td>
<td>Step counts, 644+</td>
<td>N/A</td>
</tr>
<tr>
<td>LUMOback</td>
<td>Sleep, % of goal, light sleep, latency, deep sleep, awake time, number of awakenings</td>
<td>Longest idle</td>
<td>Cut points, 217–643</td>
<td>Cut points, 644+</td>
<td>N/A</td>
</tr>
<tr>
<td>Fuelband</td>
<td>Total sleep time, right, left, side and back</td>
<td>Sitting time, slouching time, straight time, standing time (Fuel by hour)</td>
<td>Walk</td>
<td>Walk</td>
<td>Steps</td>
</tr>
</tbody>
</table>

*Variables used as a criterion measure.

Variables used in analysis are set in bold. The units for GT3X+ cut points are counts per minute and for GENeActiv are g*min.

---

**Copyright © 2016 by the American College of Sports Medicine. Unauthorized reproduction of this article is prohibited.**
for GT3X+ to 16.9% for GENEactiv. Equivalence analysis, Figure 3, indicates the GT3X+ was equivalent to the Z-machine for sleep measurement, but the other devices showed significant differences. Bland–Altman plots had mean differences in measured sleep duration ranging from 4 min for GT3X+ to 36 min for Fitbit and GENEactiv. Summary data are provided in Table 3, and the original plots are contained in Supplemental Digital Content 1 (see Document, Supplemental Digital Content 1, Bland–Altman plots, including regression lines and average differences between the standard and the comparison device, http://links.lww.com/MSS/A581). The GT3X+ also had the lowest SD on Bland–Altman analysis.

**Sedentary behavior.** Figure 2 illustrates the mean error for SED (i.e., sitting time), which ranged from 9.5% for LUMOback to 65% for GENEactiv. Equivalence testing (Fig. 3) highlighted that LUMOback accurately measured SED. All other devices produced significantly different estimates. Bland–Altman plots had mean differences ranging from 18 min for LUMOback to 162 min for GENEactiv (Table 3, and Supplemental Digital Content 1, Bland–Altman plots, including regression lines and average differences between the standard and the comparison device, http://links.lww.com/MSS/A581), with LUMOback also having the smallest SD. Because these numbers highlight a difference between posture-based measurement and motion-based measurement, results not reported here show that if the GT3X+ was used as the standard, the GENEactiv would have significantly underreported SED but the Fitbit produced sedentary measurements equivalent to that of the GT3X+.

**LPA.** For LPA, MAPE from the GENEactiv was 20% and was 28% from Fitbit, as shown in Figure 2. Figure 3 illuminates significant differences in minutes of LPA from

![Figure 2](image-url) **FIGURE 2—MAPE for the various devices and five activity domains.**

![Figure 3](image-url) **FIGURE 3—Equivalence testing for all of the devices in all domains.** Shaded areas are equivalence zones (±10% of the mean), and error bars indicate the 90% confidence interval for the mean measurement. *Equivalent measures.
TABLE 3. Bland–Altman plot summaries for all of the domains and all of the devices.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Device</th>
<th>Mean</th>
<th>SD</th>
<th>Slope</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sleep (min)</td>
<td>UP</td>
<td>32</td>
<td>37.7</td>
<td>0.02</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>One</td>
<td>36</td>
<td>37.8</td>
<td>-0.09</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td>GT3X+</td>
<td>4</td>
<td>35.8</td>
<td>-0.14</td>
<td>0.05</td>
</tr>
<tr>
<td>Sedentary behavior (min)</td>
<td>GENEactiv</td>
<td>-36</td>
<td>51.8</td>
<td>-0.17</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>LUMOback</td>
<td>18</td>
<td>52.1</td>
<td>-0.16</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td>GENEactiv</td>
<td>-162</td>
<td>110.0</td>
<td>-0.30</td>
<td>0.028</td>
</tr>
<tr>
<td></td>
<td>One</td>
<td>34</td>
<td>81.2</td>
<td>-0.34</td>
<td>0.0006</td>
</tr>
<tr>
<td></td>
<td>GT3X+</td>
<td>48</td>
<td>100.2</td>
<td>-0.47</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>LPA (min)</td>
<td>GENEactiv</td>
<td>43</td>
<td>91.8</td>
<td>-0.40</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>One</td>
<td>-64</td>
<td>47.7</td>
<td>-0.18</td>
<td>0.63</td>
</tr>
<tr>
<td>MVPA (min)</td>
<td>One</td>
<td>76</td>
<td>39.2</td>
<td>-0.05</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td>UP</td>
<td>48</td>
<td>33.7</td>
<td>-0.06</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>GENEactiv</td>
<td>170</td>
<td>89.3</td>
<td>-0.07</td>
<td>0.86</td>
</tr>
<tr>
<td>Steps</td>
<td>Fuelband</td>
<td>598</td>
<td>134.2</td>
<td>-0.63</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>LUMOback</td>
<td>1281</td>
<td>1692</td>
<td>0.02</td>
<td>0.733</td>
</tr>
<tr>
<td></td>
<td>GT3X+</td>
<td>679</td>
<td>1267</td>
<td>-0.07</td>
<td>0.160</td>
</tr>
<tr>
<td></td>
<td>activPAL</td>
<td>2258</td>
<td>1452</td>
<td>-0.07</td>
<td>0.178</td>
</tr>
</tbody>
</table>

The actual plots are presented in Supplemental Digital Content 1.

both GENEactiv and Fitbit. Lastly, the Bland–Altman summary in Table 3 gives an overestimation in LPA of 43 min for GENEactiv and underestimation of 64 min for Fitbit, with Fitbit having the smaller SD. The plots are contained in the Supplemental Digital Content 1 (see Document, Supplemental Digital Content 1, Bland–Altman plots, including regression lines and average differences between the standard and the comparison device, http://links.lww.com/MSS/A581).

**MVPA.** For MVPA, MAPE is illustrated in Figure 2 as ranging from 52% for Jawbone to 92% for Fuelband. All measurements were significantly different from the standard measure of MVPA. Mean differences from the monitors as determined by the Bland–Altman plots ranged from 48 min for Jawbone to 598 min for Fuelband, with the Jawbone also having the lowest SD.

**Steps.** Error rates for steps (as total steps per day) ranged from 14% for GT3X+ to 29% for Fuelband (Fig. 2). All devices were significantly different from the standard for measuring steps (Fig. 3) and total step differences as large as 2500 steps. Bland–Altman plots had the smallest mean difference for GT3X+ at 698 steps, with the largest difference for activPAL at 2258 steps (Table 3) and the lowest SD for the GT3X+.

**DISCUSSION**

Objective measurement of sleep, SED, and physical activity is an important component of both research and feedback from consumer wearables. All of the activity domains are related to disease outcomes. This study suggests that measurement of these domains is highly varied among wearable devices when tested outside the laboratory. Although this may sound discouraging, the ability to measure very specific behaviors has greatly increased with the introduction of a large number of wearable devices. For sleep, this study shows that many of the devices can measure total sleep time with the predictable error that comes from comparing actigraphy to polysomnography. For SED, this study highlights the differences between posture measurement (LUMOback being similar to activPAL) and an accelerometer measurement indicating a lack of motion (GT3X+, Jawbone, Fitbit, and GENEactiv). For LPA and MVPA, this study also suggests that there are major differences between the devices and that these devices may be using different measures of the behavior of interest. For example, LPA is usually defined as 1.5–3.0 METs, but not all devices may be trying to identify that intensity as LPA. For steps, many of the devices were different from the standard but gave results similar to each other, implying some predictable agreement among devices.

Currently, 24-h activity measurement is only possible with research devices, such as the GT3X+. None of the commercial devices provide all the measures of the 24-h model. Tapping into richer data from application programming interfaces from commercial devices may allow complete 24-h measurement, but it may be significantly different from previous measurement standards. For this reason, choosing a device specific to the primary outcome measure of interest will be of utmost importance. Calibration and evaluation of devices will be an ongoing research area because of the rapid changes in wearable technology. Evaluating devices for their ability to determine time spent at different intensities is highly relevant to optimal health, yet many devices are not created specifically with this focus in mind. This study highlights a lack of standards among commercial devices for important health-related objective activity measurement. The following discussion will highlight areas of interest in each activity domain and propose recommendations for manufacturers and device calibration experts.

**Sleep.** Actigraphy has previously been used to measure sleep/wake patterns with some reliability (37). In addition, a single-channel electrode is an accurate method for sleep/wake detection relative to full polysomnography (20), and this was the method used with the Z-machine. The portable electrode method of the Z-machine produced a similar difference in total sleep time as the scoring of polysomnography (20,30), and further exploration of the Z-machine may lead to better portable electroencephalogram sleep measurement in the field. Although there are published algorithms for sleep scoring (36), none of the consumer accelerometer-based devices publish their algorithms for measuring sleep, creating an issue with comparisons of the devices. Previously, the Fitbit was found to overestimate total sleep time and lacked sleep/wake specificity similar to how other accelerometer-based devices compared with polysomnography (30). These results were replicated in this study, and in general, the sleep devices overestimated total sleep time. Because this study highlights some agreement between the sleep/wake measurement of consumer devices and research devices, the use of these devices in research should be explored further. Algorithm development work is currently ongoing in this regard for the activPAL.

Sleep measurement from consumer devices covers aspects of sleep that were not examined in this study. Total sleep time was evaluated because stages of sleep, sleep efficiency, and measurement of circadian rhythms are not recommended using actigraphy on the wrist (37). For
example, the Jawbone Up has several sleep variables (light vs deep sleep) that contradict the recommendation for measurement with wrist actigraphy from sleep experts (37). Other variables that could be explored in future research include sleep latency, number of awakenings, time spent in different stages of sleep, and sleep efficiency. The evaluation of all sleep variables from these devices is dependent on either polysomnography in the laboratory or creation of a portable standard measure. In addition, the sleep/wake measurement should be evaluated with different devices in broader populations.

Sedentary behavior. Sedentary behavior measurement is complicated by varying definitions used to describe the lack of activity. Current definitions rely on a combination of posture (i.e., sitting) (32,38), low levels of energy expenditure (32,33), or specific activities (such as TV viewing, but not including sleep) (33). A promising outcome of this article is the addition of LUMOback as an accurate measure of daily posture. Many health outcome studies that highlight the importance of limiting SED found associations without the postural measurement defined in this article (3,15,16,29), creating a debate on which measurement (postural or lack of motion) is important for health (32). Unfortunately, postural measurement devices are not necessarily the best devices for other components of the full 24-h activity cycle, because they lack specificity in measurement of activity intensity. The design and goal of a study will determine whether a postural device should be used (e.g., sedentary interventions to reduce sitting) or whether 24-h measurement should be prioritized (e.g., controlling for sedentary behavior in physical activity studies).

LPA. Relatively little is known about LPA because of the difficulty in obtaining accurate objective measurement (also true for assessing LPA by questionnaire) (4). In the past, LPA has been measured using a 7-d recall and subtracting sleep, sedentary time, and MVPA from 24 h as opposed to having a direct estimate of LPA (4). Measuring LPA in the 24-h cycle can be done with any device that can separate SED and MVPA from LPA, but because there is no device that accurately captures LPA, a recommendation cannot be made on the basis of the results presented here. An important part of creating an accurate 24-h measurement device will be the improved measurement of LPA during daily activities. Activity measurements for 24 h could lead to a recommendation of how much time should be spent in LPA (which is also a major displacement of sedentary time) on a daily basis to optimize disease prevention.

MVPA. A surprising result of this study is that MVPA was not accurately measured by several devices. Given the small percentage of time spent in MVPA in many populations, even modest measurement error is clinically significant in a 24-h period. One reason for the discrepancy in measurement could simply be the definition of MVPA. Many commercial device companies do not provide a definition of what they are measuring, so although the official definition of moderate activity includes any activity ≥3 METs and <6 METs (1), there is no confirmation that this is what the devices are attempting to measure. For example, the Jawbone UP defines their activity measurement only as “time spent moving” (19). In this study, MVPA had 51%–91% error, most likely because the devices were measuring different activities from the official definition. One recommendation of standardizing activity measurement would be to adhere to commonly used definitions of intensity. Alternatively, the calibration of the ActiGraph on the hip was one of the earliest calibration studies (11) and is still used as the standard in epidemiological research (2,28). Research shows the relation between these standards and health outcomes (16,24,28), making this an appropriate standard to use while calibrating devices.

The results of this study also call into question the ability of field methods to accurately measure MVPA. In recent evaluations of these devices for predicting energy expenditure, Jawbone and Fitbit were more accurate than the GT3X+ (25). The GT3X+ provides a measure of MVPA different from the measures of MVPA provided by the other devices, but it is not necessarily more accurate at measuring activities with energy expenditure above 3 METs. A recent study concluded that the cut point analysis of GT3X+ data underestimates the time spent in MVPA compared with other methods (22). Cut point analysis is also not universally applicable and has known limitations; for example, cut points for younger adults are not the same as those specifically created for older adults (28). This limitation is specific to the algorithm used, not to the device overall. In this case, a useful follow-up will be to see whether other device measures of MVPA have the same relation to health outcomes as cut points on the GT3X+. Luckily, large databases of activity measurement are being created by the users of these devices. Defining the optimal amount of MVPA on the basis of objective measurement may have to become device specific, or, at the very least, current methods in physical activity epidemiology should consider additional standardization.

Steps. In this study, none of the wearables measured steps in the same way as the Omron, but a recent article found that the Fitbit One might be the most accurate device for measuring steps compared with researcher step counts (9). Many of the devices are dependent on the “bout” or number of steps you take in order for the device to count those steps and the “time out” or the time between steps that will reset the “bout” (17). Given those two variables, a recommendation should be developed as to what type of walking is most beneficial for health. For example, researchers may determine whether a one-step bout requirement has a different relation to health outcomes at 10,000 steps a day compared with an algorithm that requires a four-step bout requirement.

RECOMMENDATIONS AND CONCLUSIONS

Research has identified areas of our daily activity cycle that relate to health in many ways. Sleep research has focused on finding a healthy amount of sleep to prevent
disease and optimize performance in our daily activities. Sedentary behavior research has cautioned about the detrimental health outcomes and metabolic disturbances that come from inactivity. LPA research has focused on the added benefit of burning extra calories through more movement in a 24-h day. MVPa research, based primarily on survey data, has a very specific relation to health in a dose–response manner, with most benefits coming from getting 30 min or more of moderate-intensity physical activity in a day. At present, the most common activity intervention is to increase daily exercise, but for those who are sleeping less than 6 h a night, increasing exercise may prove to be less important than increasing sleep to over 7 h a night.

Given what we know about activity and the link to better health, these domains should be measured objectively, with accuracy, and in ways that can be compared with guidelines defined by the biomedical community. We should strive to make these activity definitions and measures match as closely as possible for both feedback to the user and for researchers to gain a better understanding of the rich data sets being generated by a barrage of new wearable users. The importance of 24-h measurement in medical research, as well as for consumer application, raises a number of areas that should be considered for future device research. The explosion of new wearables, along with the addition of new devices, software upgrades, and other changes, demands continuous updating of device evaluations. The expanding measurement capabilities of devices, with multiple physiological and contextual measures, will continue to expand how research can be conducted. HR, for example is a common theme among the upcoming Apple Watch, Jawbone 3, Basis Peak, Microsoft Band, Fitbit Surge, and a number of other “smart watch” devices. The addition of HR to the motion data presents a new avenue for defining sleep, SED, and all levels of physical activity. Not only is there research to expand the analytical techniques that are used to combine information when examining relations among activities and health outcomes. Multiple data inputs from various devices can be quite complicated, and the field lacks consensus about how to combine devices for an optimal daily activity cycle focused on promoting health while preventing negative health outcomes. An optimal activity cycle will be exceptionally important for quantifying activity as well as in designing and evaluating interventions to promote health.

Financial support for this project was provided by Grant R37-AG008816 from the National Institute on Aging to Laura L. Carstensen. Dr. Rosenberger was a postdoctoral fellow supported by the same grant. Stanford Cardiovascular Medicine has received in-kind mobile health research support from Apple, Inc. Contributions to data collection were made by Brent LaStofka. The authors have no other potential conflicts of interest to disclose. The results of the present study do not constitute endorsement by the American College of Sports Medicine.

REFERENCES


