# Module 2. Evaluating the performance of technology

## Using sleep technology as an example

### Algorithms

Actigraphy data is generally scored and recorded in one-minute epochs, providing an activity value for each minute. The first published actograms summarised data by the hour using the activity of two psychiatric inpatients – one with anorexia nervosa and the other with psychotic depression. The researchers were able to quickly discern patterns of sleep and wake, and, while accuracy was around 80% of that measured via PSG, it provided a good starting point for wearable actigraphic devices used to measure periods of sleep.

The first scoring technique reported the use of raw movement counts from the EEG and actigraphy (Kupfer *et al.*, 1972). Scorers using the MediLog system used "the best guess which could be derived" (Kripke *et al.*, 1978, p. 85). By 1980, scoring was considered to be the result of the best guess using two scorers to minimise error (Mullaney *et al.*, 1980).

The need for developing a scoring algorithm that could operationalise and automate this process became evident. While actigraphy could not replace polysomnography, it could provide a cheaper, easier, lighter alternative for capturing objective sleep data.

An effective algorithm must:

- incorporate data from several epochs before and after the one being assessed to determine probability of sleep versus wake.
- Account for time of day to strengthen sleep versus wake predictability.
- Discern between sleep and quiet wakefulness in the absence of movement.
- Account for time the device is off the wrist, such as during bathing or exercise.
- Put brief changes in activity levels within a greater context.
- Account for limitations of the device itself.

The first algorithms only considered broad levels of activity, with high activity presumably occurring during wake and low activity occurring during sleep. But they also began to explore moment-to-moment changes in activity. Periods of movement interspersed with a few minutes of stillness were considered likely to occur during wake. Therefore, all stillness is not likely to be indicative of sleep. Effective algorithms use context. Whether a person was moving during the previous one to two minutes is more important for predicting the likelihood of sleep or wake than movement 10 minutes ago. Time of day

context is also important. Absence of movement during the daytime when a person is normally awake is unlikely to be a sleep episode.

The first actigraphy scoring was published in 1972 and referred to by the researchers as 'telemetry' (Kupfer *et al.*, 1972). Eight patients had 24-hour motor activity recorded and compared to EEG. While there were some discrepancies (as seen in table 1), the data were well correlated.

The word 'actigraphy' first appeared in the sleep literature in 1978 (Kripke *et al.*, 1978), though it had been used in the United States military prior to that. The researchers used one-minute epochs to estimate time in bed, sleep duration, and WASO which demonstrated good agreement with PSG (table 1). So, at this time, you will start seeing the metrics of how does one define agreement between devices. We are going to get to this. But, at first, all they had essentially were correlation coefficients. Eventually, that was deemed too loose, not accurate enough, and so that is where we have other metrics we use today.

**Table 1. MediLog vs. PSG**

| Comparison of EEG and actigraphic scoring[a] | | |
|---|---|---|
| | EEG scoring (min) | |
| | Sleep | Wake |
| Actigraphic scoring (min) | | |
| Sleep | 378 | 24 |
| Wake | 8 | 172 |
| [a]Average for 102 recordings. | | |

Source: Mullaney *et al.*, 1980, p. 88.

In 1980, a large-scale study compared 39 patients to 63 health control subjects, which happened to be college students and hospital staff (Mullaney *et al.*, 1980). The MediLog device recorded one-minute epochs and was compared to PSG also scored in one-minute epochs (table 1).

The discrepancies occurred where the device recorded sleep, while the EEG detected wake and, in general, in an issue of misclassification that persists today, wearable devices underscore wake.

The data displayed in table 2 from this same study appeared to capture many of the most challenging obstacles that emerged over the next several decades in the field. The first of these was that the wearables were more effective at capturing sleep in non-patients. This is because patient sleep tends to be more variable and less predictable in patients.

Therefore, it is harder to score because there tends to be more noise. Unexpected scenarios can occur in a patient setting.

## Table 2. Actigraphy vs EEG

| Correlation between actigraph and EEG scoring | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Group variable** | **TSP** | | **TST** | | **WASO** | | **MSA** | | **Agreement (%)** | |
| **All subjects (n = 102)** | 0.90[a] | | 0.89[a] | | 0.70[a] | | 0.25[c] | | 94.5 | |
| **Patients (n = 39)** | 0.82[a] | | 0.81[a] | | | | | | | |
| **Nonpatients (n = 63)** | 0.97[a] | (p < 0.0001) | 0.95[a] | (p < 0.001) | 0.56[a] | (p < 0.001) | 0.09[e] | (p < 0.05) | 91.6 | (p < 0.001) |
| | | | | | 0.87[a] | | 0.46[a] | | 96.3 | |
| **Age** | | | | | | | | | | |
| ≥50 (n = 17) | 0.52[a] | | 0.39[e] | | 0.82[a] | | 0.01[e] | | 88.5 | |
| <50 (n = 85) | 0.95[a] | (p < 0.0001) | 0.95[a] | (p < 0.0001) | 0.65[a] | (ns) | 0.32[b] | (ns) | 95.7 | (p < 0.001) |
| **EEG TSP** | | | | | | | | | | |
| ≥390 (n = 66) | 0.85[a] | | 0.83[a] | | 0.72[a] | | 0.19[e] | | 94.7 | |
| <390 (n = 36) | 0.74[a] | (ns) | 0.78[a] | (ns) | 0.66[a] | (ns) | 0.52[a] | (ns) | 94.2 | (ns) |
| **Sleep log TST** | | | | | | | | | | |
| ≥390 (n = 53) | 0.96[a] | | 0.94[a] | | 0.72[a] | | 0.26[d] | | 96.3 | |
| <390 (n = 49) | 0.79[a] | (p < 0.0001) | 0.69[a] | (p < 0.0001) | 0.70[a] | (ns) | 0.23[e] | (ns) | 92.5 | (p < 0.001) |
| **Recording quality** | | | | | | | | | | |
| Low (n = 20) | 0.78[a] | | 0.79[a] | | 0.74[a] | | -0.07[e] | | 92.8 | |
| High (n = 24) | 0.97[a] | (p < 0.001) | 0.97[a] | (p < 0.001) | 0.75[a] | (ns) | 0.59[c] | (p < 0.05) | 95.9 | (ns) |
| **Interpretability** | | | | | | | | | | |
| Low (n = 27) | 0.77[a] | | 0.83[a] | | 0.62[a] | | -0.16[e] | | 90.4 | |
| High (n = 24) | 0.97[a] | (p < 0.001) | 0.97[a] | (p < 0.01) | 0.75[a] | (ns) | 0.59[c] | (p < 0.01) | 95.9 | (p < 0.005) |
| [a] $p < 0.0001$ (one-tailed), [b] $p < 0.001$, [c] $p < 0.01$, [d] $p < 0.05$, [e] not significant. | | | | | | | | | | |
| Abbreviations: TSP, total sleep period; TST, total sleep time; WASO, wake after sleep onset; and MSA, number of midsleep awakenings. | | | | | | | | | | |

Source: Mullaney et al., 1980, p. 87.

Secondly, the devices were better able to measure movement during sleep in younger people, therefore giving rise to an age bias.

Thirdly, there was a bias with total sleep duration. Those with shorter sleep have fewer total epochs. Fewer epochs result in greater weight on any one epoch. This means that scoring one epoch incorrectly in a short sleeper can skew the results to a greater degree than a longer sleeper. Therefore, those who got more than six and a half hours sleep (390 minutes) had greater accuracy.

Lastly, recording quality and interpretability pose problems to this day. Difficult records are more likely to result in disagreement between scorers. So too is a low-quality EEG recording.

After it was found that actigraphic recordings agreed fairly consistently with PSG on a minute-by-minute basis, researchers began questioning whether scoring could it be automated. They did this by entering the following factors into a prediction equation:

1. Total activity in current epoch

2. The most active sample of the minute
3. The sum of the two most active samples more than 30 seconds apart
4. The sum of the eight most active samples
5. Weights for total activity in the preceding four epochs and subsequent two epochs to account for context

After first experimenting with this paradigm, they showed that (1), (3), and (4) could be eliminated. Therefore, the most active sample of the minute and the total activity in the four previous, and the two subsequent epochs, accounted for the highest accuracy in predicting the sleep-wake behaviour of any one epoch. Kripke and colleagues (1978) used 20 records, developed an algorithm based on 17 of them, and tested it on three individuals.

Webster's Algorithm, with 'I' being the epoch of interest, with four epochs preceding it and two following it.

**Webster's algorithm**

$$D = 0.025 \times [0.15T_{(i-4)} + 0.15T_{(i-3)} + 0.15T_{(i-2)} + 0.08T_{(i-1)} + 0.21T_{(i)} + 0.12T_{(i+1)} + 0.13T_{(i+2)}]$$

This resulted in agreement of 94.46% in the development group and 96.02% in the three people that comprised the validation sample. It over-scored sleep more than it over-scored wake, as occurred in prior studies. This is because sleep onset occurs after movement has ceased. This is where using movement alone has its limitations. However, the algorithm offered better prediction of sleep versus wake by incorporating the activity from preceding epochs into the model. A lack of movement preceded by lack of movement was more likely to be a recording of sleep, for example.

Some follow-up rules were created to account for the problem of over-reporting sleep:

1. if you have 4-9 minutes of wake, recode the first 1 minute of sleep to wake.
2. If you have 10-14 minutes of wake, recode the first 3 minutes of sleep to wake.
3. If you have 15+ minutes of wake, recode the first 4 minutes of sleep to wake.
4. If you have 10-19 minutes of wake surrounding a sleep period, any sleep period of 6 minutes or less should be recoded as wake.
5. If you have 20 minutes of wake surrounding a sleep period, any sleep period of 10 minutes or less should be recoded as wake.

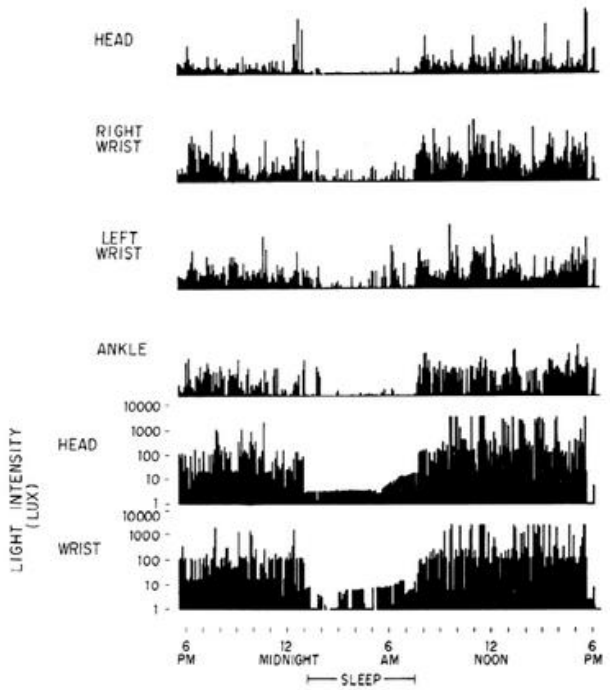
Webster's algorithm was then slightly modified for use in a portable recording device. The VitaLog was still wrist-based, but the external recorder was worn on a belt. This was a different transducer, with different electronics and a different amplifier. Despite the upgraded technology, Webster's algorithm still resulted in 93% accuracy with PSG.

The next iteration of this work came in 1983 when VitaLog recorders were used on the forehead, non-dominant wrist and ankle (Okudaira *et al.*, 1983). The researchers also

recorded ambient light levels through a separate wearable light transducer, which served as a circadian marker of movement. Many consumer wearables do not record light; however, research devices used in a laboratory setting still record this important parameter.

As seen in figure 1, the most effective location for wearable light-recording devices was the ankle, with locations such as the wrist less likely to be covered by bedding during the night. While the head is even more likely to exposed to light at night, it does not tend to move as much as the limbs, so the results tend to be more reliable than devices worn on the wrist.

**Figure 1. Locations of actigraphs**




Source: Okudaira *et al.*, 1983.

The next advancement came in 1989 with the Sadeh algorithm, when paediatric sleep specialist Avi Sadeh investigated the clinical potential of actigraphy (Sadeh *et al.*, 1991). This algorithm was developed using a similar process and found good minute by minute agreement.

- Controls (age 20-76; n = 13): 90.2%
- Sleep apnoea (age 20-76; n = 24): 85.7%
- Insomnia (age 20-76; n = 16): 78.2%
- Children (age 3-13; n = 13): 89.9%

This diverse sample was in less agreement with PSG, but was more flexible for patients with sleep disorders and is still the gold standard among paediatric populations today.

The Sadeh algorithm fared slightly better for insomnia than Webster's, but, statistically, the two were not significantly different from each other or hand scoring. Hand scoring differed from PSG by about 50 minutes, with 25% of records more than an hour discrepant from PSG. The Webster algorithm differed from PSG by approximately 63 minutes, with discrepancies of more than an hour 45% of the time. The Sadeh algorithm also differed from PSG by about an hour, with a third of samples discrepant by more than 60 minutes.

Actigraphic devices were next compared to PSG for their ability to detect sleep versus wake among patients with both primary and secondary insomnia. Primary, or psychophysiological, insomnia occurs when insomnia symptoms arise from no known existing cause. Secondary insomnia, on the other hand, stems from a primary medical illness, mental disorder, or other sleep disorder (Morgenthaler *et al.*, 2006).

While the average error, when compared to PSG at detecting sleep versus wake, was 49 minutes, the actigraphy data was found to be better aligned with the sleep diary estimates of sleep duration than PSG was, which may be more appropriate for capturing insomnia. There are many reasons why PSG can be problematic for diagnosing insomnia. Primarily, patients often wakefulness while a PSG reports sleep. In insomnia diagnoses, sleep diaries tend to be the primary measure.

In comparison with PSG, a sleep diary typically reports many fewer awakenings per night. This is because although micro-awakenings often occur throughout the night; most individuals do not remember these. A typical sleep diary will record one to three awakenings, while a PSG might report 10 or 20, or even more. Actigraphy tends to underreport nighttime awakenings, and thus, in cases of insomnia, actigraphy may be more reliable than PSG as well as avoiding the common issue of an aversion to attempting to sleep inside a laboratory environment.

The next development came when the Webster algorithm was refined in 1992 (Cole *et al.*, 1992). The Cole-Kripke algorithm remains the standard actigraphy algorithm to date. The study included 41 subjects, 15 of which were controls, alongside several psychiatric, sleep apnoea, pain, and insomnia patients in an attempt to diversify the sample as Sadeh had done.

This was the first validation study performed in a commercially available device (Motionlogger), newer version of which are available for purchase today. The original device sampled 30 times per second using two-second counts for movement. It evaluated mean activity and maximum 2, 6, 10, 20, and 30 second periods within each epoch. They found that the maximum 10-second period was the one that was most predictive.

Cole-Kripke Algorithm, with 'I' being the epoch of interest, uses the four epochs preceding the epoch in question, and two that follow it, to score sleep versus wake. If $D \geq 1$, the epoch is scored as wake.

**Cole-Kripke algorithm**

$$D = 0.00001 \times [404T_{(i-4)} + 598T_{(i-3)} + 326T_{(i-2)} + 441T_{(i-1)} + 1408T_{(i)} + 508T_{(i+1)} + 350T_{(i+2)}]$$

It is important to mention that this algorithm was developed using a movement transducer, and despite the leap to actigraphy in modern devices, this algorithm is still applied, using the epoch of interest along with the four epochs preceding it and the two epochs that followed to predict sleep versus wake.

This algorithm resulted in agreement of 87.05% in the development group and 87.91% in the validation sample, increasing to 87.93% and 88.25% when the Webster follow-up rules were applied. The whole night sleep parameters can be seen in table 2.

**Table 3. Cole-Kripke**

| | Training sample | | | Validation sample | | |
|---|---|---|---|---|---|---|
| | PSG[b] | Actigraph[b] | *r*[c] | PSG | Actigraph | *r* |
| Minutes scored | 427.0 ± 74.7 | 427.0 ± 74.7 | | 445.5 ± 43.1 | 445.5 ± 43.1 | |
| Total sleep time (minutes)[d] | 308.0 ± 93.5 | 329.5 ± 95.1* | 0.91*** | 344.5 ± 52.5 | 363.9 ± 54.4* | 0.77*** |
| Percent sleep (%)[d,e] | 71.4 ± 20.3 | 76.7 ± 20.1* | 0.89*** | 77.9 ± 13.0 | 82.4 ± 13.5* | 0.82*** |
| Sleep efficiency (%)[d,e] | 80.4 ± 18.7 | 84.4 ± 19.4 | 0.85** | 85.0 ± 11.6 | 88.6 ± 11.4 | 0.71** |
| Sleep latency (minutes)[f] | 75.6 ± 89.7 | 70.9 ± 102.2 | 0.94*** | 59.2 ± 46.1 | 50.1 ± 50.7 | 0.90*** |
| Wake time after sleep onset (minutes)[f] | 51.3 ± 45.6 | 40.2 ± 32.7 | 0.49* | 49.9 ± 37.2 | 36.6 ± 30.6* | 0.63** |

Source: Cole *et al.*, 1992, p. 465.

[a] Scoring was performed using the algorithm for a maximum 10-second overlapping epoch per minute. Rescoring was based on the Webster follow-up rules.

[b] Means ± SD. Asterisks indicate significant differences between PSG and actigraph by paired *t* test.

[c] Pearson correlation between PSG and actigraph scores. Asterisks indicate significant correlations.

[d] Includes all minutes scored as sleep, including those which occurred before 20-minute criterion for sleep onset was met.

[e] Percent sleep is based on the entire record. Sleep efficiency is based on time in bed.

[f] Both PSG and actigraph sleep onset defined as beginning of the first interval containing 20 minutes scored as sleep, with no more than one minute of wakefulness intervening. Latencies computed from the start of simultaneous actigraph/PSG recording, regardless of whether the subject was attempting to sleep. * $p < 0.05$, ** $p < 0.002$, *** $p < 0.0001$

Different transducers work in different ways, but older algorithms have been applied to newer data with surprising success. When the Webster algorithm, calculated using the original movement transducer, was applied to the Cole-Kripke Motionlogger data, there were agreements that were only marginally shifted from 86.44% in the training sample and 87.54% in the validation sample to 87.24% and 87.73%, respectively.

The first actigraphy review paper, published in 1995, summarised the studies done to date and consolidated the rates of agreement with PSG (Sadeh *et al.*, 1995).

**Table 4. Summary of actigraphy rates of agreement until 1995**

| Study[a] | Sample[b] | Sample size | Age | S/W[c] | SEF[c] | DUR[c] | Comments[d] |
|---|---|---|---|---|---|---|---|
| Kripke (3) | N | 5 | NA | NA | 0.98 | 0.95 | |
| Mullaney (4) | N | 53 | 18–66 yr | 96.3 | 0.81 | 0.97 | HS TST |
| | P | 32 | 18–66 yr | 91.6 | 0.95 | 0.82 | HS TST |
| Webster (5) | N | 14 | College | 93.9 | NA | NA | AS IDT |
| | N + P | 14 | College | 93.4 | NA | NA | AS IDT |
| Sadeh (6) | N | 13 | 20–76 yr | 90.2 | 0.91 | NA | AS |
| | SAS | 25 | 20–76 yr | 85.7 | 0.63 | NA | AS |
| | INS | 16 | 20–76 yr | 78.2 | 0.79 | NA | AS |
| | P/C | 13 | 3–13 yr | 89.9 | 0.81 | NA | AS |
| Sadeh (7) | N + P/C | 11 | 12–48 mo | 85.3 | NA | NA | AS |
| Hauri (8) | INS | 36 | 24–69 yr | 82.1 | NA | NA | AS |
| Cole (11) | N + P | 51 | NA | 88.0 | 0.82 | 0.90 | AS SLT |
| Sadeh (12) | N | 36 | 10–25 yr | 91.2 | NA | NA | AS |
| Sadeh (13) | N | 41 | Newborn–12 mo | 95.3 | NA | 0.95 | AS |

Source: Sadeh *et al.*, 2005, p. 290.

In 2001, Jean-Louis and colleagues developed the Jean-Louis algorithm using the Actillume device, which was the first to include accelerometery. While the weights given to the epochs surrounding the one in question were slightly different to the algorithms that had some previously, this was the first designed for an accelerometer rather than a movement transducer, and had an agreement of 87.1% with PSG, increasing to 88.3% after applying the Webster follow-up rules.

### Jean Louis algorithm

$$D = 0.13 \times [0.010T_{(i-4)} + 0.015T_{(i-3)} + 0.028T_{(i-2)} + 0.31T_{(i-1)} + 0.085T_{(i)} + 0.015T_{(i+1)} + 0.010T_{(i+2)}]$$

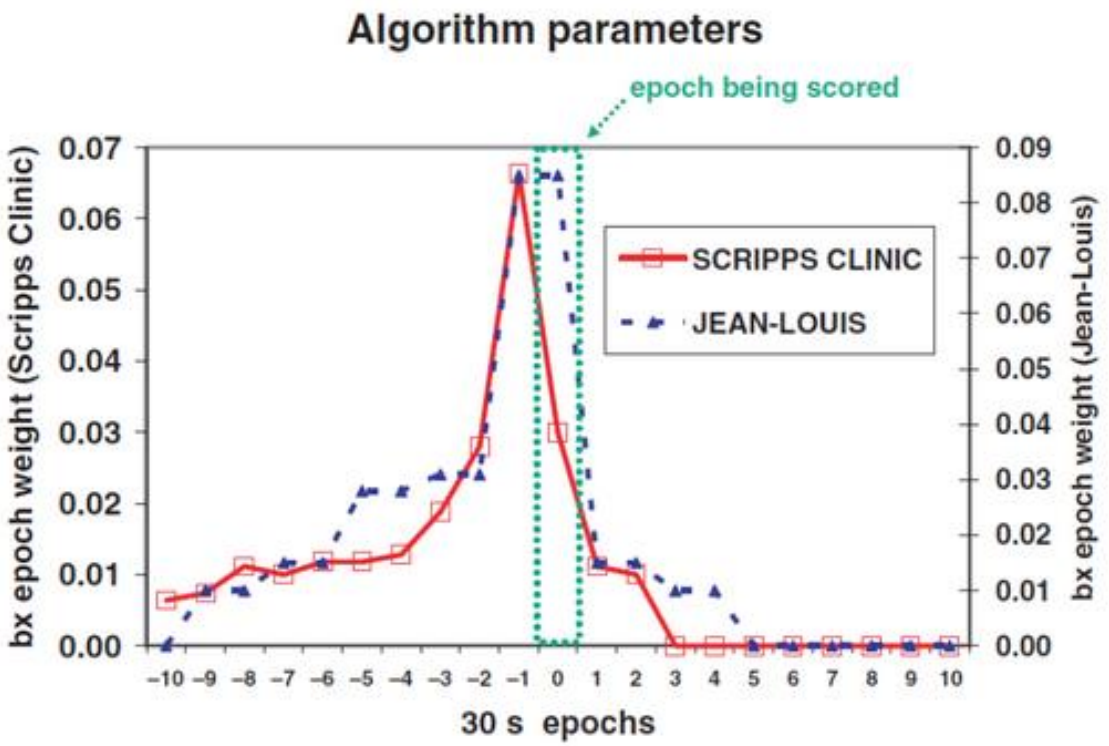
Many other devices appeared on the market in the 2000s, with relatively similar rates of agreement, including the Actiwatch, Sleepwatch, Actical, Actiwatch-L, and Spectrum.

By 2010, the Actiwatch-L was one of the most widely-used devices, but the algorithm was not made public. Kripke and colleagues advanced previous methods in a sample of 49 individuals and 30 second epochs to align with PSG (Kripke *et al.*, 2010). The new Scripps algorithm incorporated 10 preceding epochs and two subsequent epochs, resulting in 87% agreement with PSG. Agreement with PSG for the Actiwatch Spectrum was slightly lower, at 85%, the reason for which remains unclear.

**Scripps algorithm**

$$D = 0.30 \times [0.0064T_{(i-10)} + 0.0074T_{(i-9)} + 0.0112T_{(i-8)} + 0.0112T_{(i-7)} + 0.0118T_{(i-6)} + 0.0118T_{(i-5)} + 0.0128T_{(i-4)} + 0.0188T_{(i-3)} + 0.0280T_{(i-2)} + 0.0664T_{(i-1)} + 0.0300T_{(i)} + 0.0112T_{(i+1)} + 0.1000T_{(i+2)}]$$

**Figure 2. Scripps vs Jean-Louis algorithms**




Source: Kripke *et al.*, 2010.

Compared to the Jean-Louis algorithm, the Scripps algorithm was weighted slightly more towards the previous epoch than the current one, but the two are similarly aligned, both

appearing to plateau between 85-90% as a result of inherent performance limitations (Kripke *et al.*, 2010).

### Performance

Actigraphy performance is evaluated through three vital parameters: sensitivity, specificity, and accuracy.

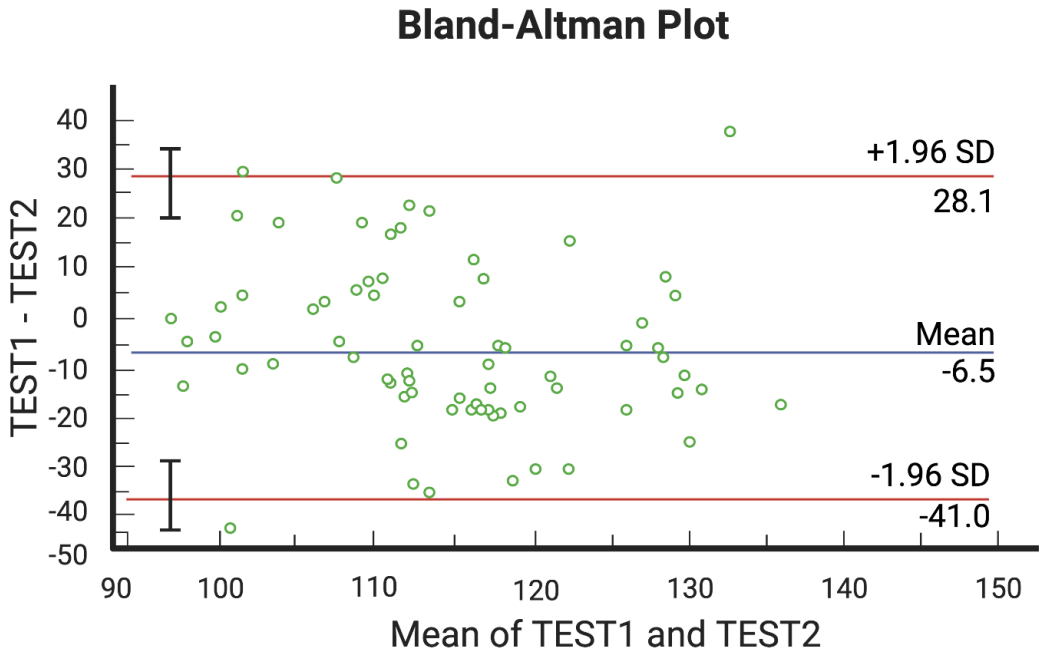
- Sensitivity – the percentage of 'true' sleep epochs scored as 'sleep'
- Specificity – the percentage of 'true' wake epochs scored as 'wake'
- Accuracy – the percentage of epochs where the device score agrees with the PSG recording

Sensitivity is usually high, if most of the recording episode consists of sleep. Specificity is often the weakest metric, reflecting the inability to detect wake epochs that take place amidst a period of sleep. This has emerged as the true test of how effective a device is. While accuracy is a critical metric, agreement may be biased in certain situations, such as insomnia patients, and this must be taken into consideration when selecting the best device for an individual.

Kappa scores are also typically reported in studies assessing performance of wearable devices, whether it be sleep, physical activity, or performance wearable technology. This represents the agreement between two 'raters' or 'scorers' of the data used for two different methodologies, i.e. PSG and a wristwatch. Scores range from 0 (chance) to 1 (perfect agreement) using the following equation:

$K$ = (Observed – expected) / (1 – expected)

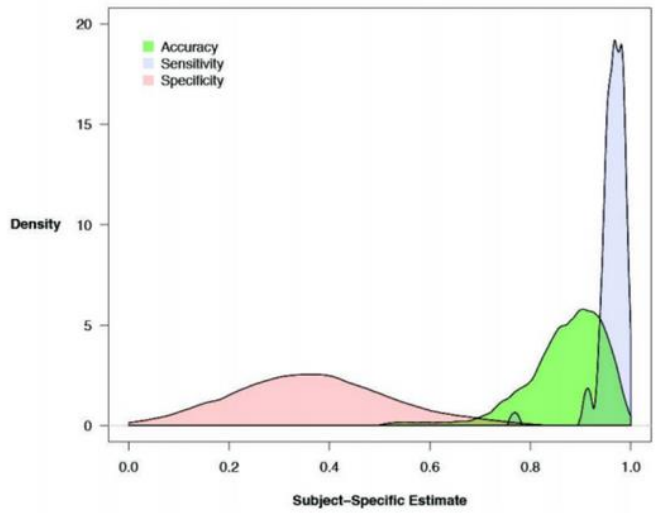
**Figure 3. Example Bland-Altman plot**




Source: [Untitled image about example Bland-Altman plot], n. d.

Bland Altman plots are also frequently used to compare methodologies. The X axis displays the mean and the Y axis displays the difference between the two methods. In the example in figure 3, the mean difference between the scores of two different methods is -6.5 units. However, a negative slope appears in the scatterplot data. This means that, as the score increases, the agreement between the two methods decreases.

**Figure 4. Performance of AW-64 and Actiwatch Spectrum**




Source: Marino *et al.*, 2013, https://bit.ly/3D1rP4L.

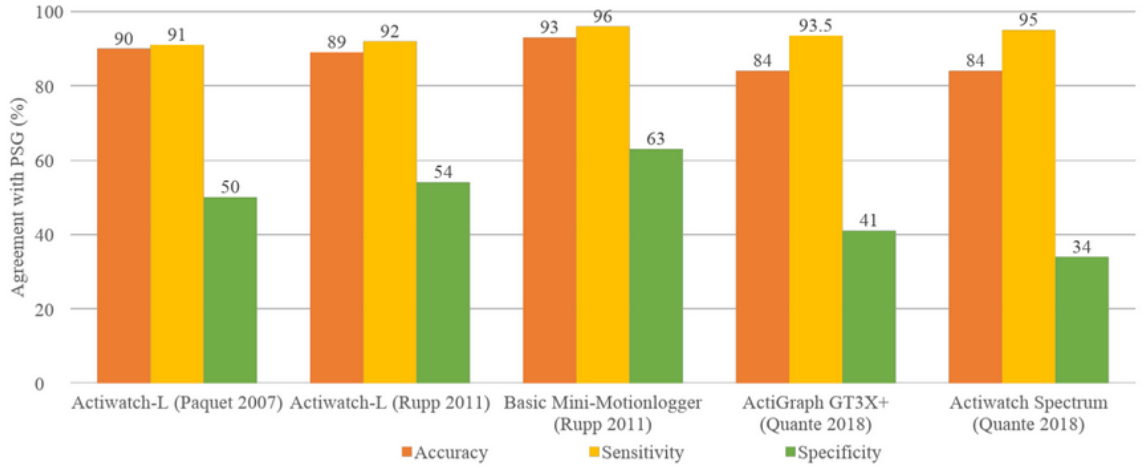
The AW-64 (Minimitter, Inc, Bend, OR) and Actiwatch Spectrum (Philips/Respironics, Murrysville, PA) were validated in 2013 in a sample of 90 individuals (Marino *et al.*, 2013). Sensitivity and accuracy were high, but specificity was comparably low (figure 4).

This was the first study to address the importance of temporal synchronisation between actigraphy and PSG during recordings. The researchers compared specific epochs throughout the sleep period between the actigraphy and PSG recordings, aligning the two for the highest proportion of matched epochs. The study found no significant differences in results stratified by gender, time of day, or the presence of insomnia; but there was a bias in WASO scoring.

A night of sleep with proportionally more awakenings increases the disagreement between actigraphy and PSG. This is because actigraphy tends to miss wake amidst periods of sleep due to the general lack of movement. In this, the Bland Altman plot shows that, as WASO increases, this increases the risk of the actigraphy devices misreporting nocturnal wakefulness as sleep. The PSG continues to capture the wake correctly, but the actigraphy overscores sleep. This resulted in the lower specificity score.

The accuracy, sensitivity, and specificity scores for a range of devices that are shown in figure 5.

**Figure 5. Performance of multiple devices**




Source: Lujan *et al.*, 2021

The GENEActiv recorder (ActivInsights Ltd., Kimbolton, UK) was compared to the Actiwatch, in a 2013 study that included 15 participants (te Lindert and Van Someren, 2013). The researchers applied the Actiwatch algorithm to the scores generated by the GENEActiv device and found stronger agreement between two GENEActiv devices, or one

Geneactiv device and an Actiwatch than between two Actiwatches (table 3). The algorithm used was as follows:

$$A_0 = [0.04E_{-(8\text{-}5)} + 0.20E_{-(4.1)} + 0.40E_{(0)} + 0.20E_{+(1\text{-}4)} + 0.04E_{+(5.8)}]$$

Where $A_0$ is the total rescored activity for the 15-second epoch of interest, $E_0$ is the activity scored in the epoch, and $E_n$ is the activity scored in the two minutes before and after the scored epoch. If $A > T$, the epoch is scored as wake. If $A \leq T$, the epoch is scored as sleep.

**Table 5. GENEActiv vs Actiwatch**

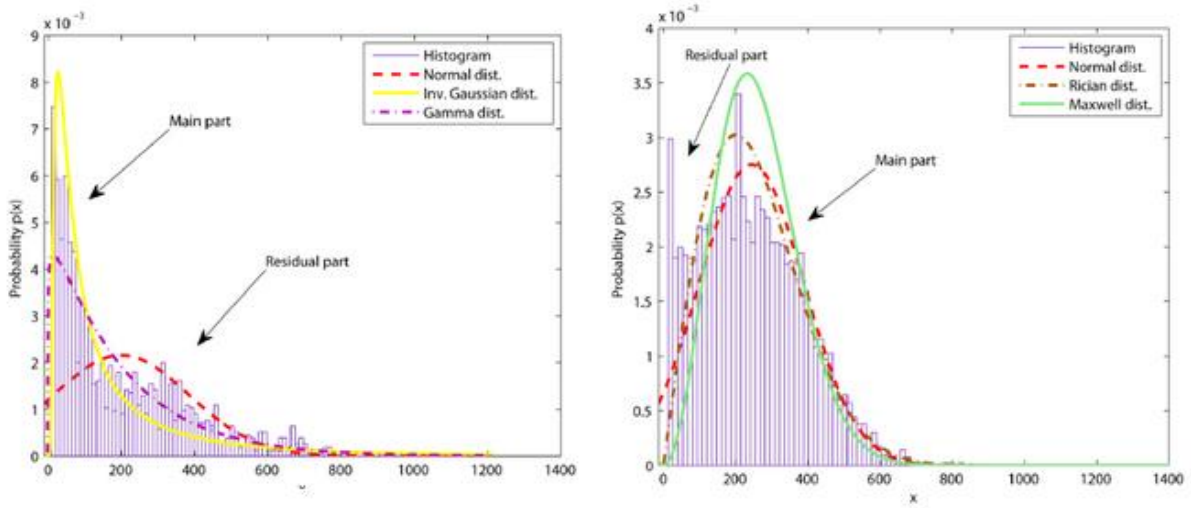
| | MEMS scores sleep | MEMS scores wake |
|---|---|---|
| **Actiwatch scores sleep** | 91.6 ± 3.3% | 1.0 ± 0.4% |
| **Actiwatch scores wake** | 1.0 ± 0.6% | 0.4 ± 2.9% |

Source: own source based on Te Lindert and Van Someren, 2013.

Many of the current devices on the market use custom, proprietary algorithms that were not disclosed to third parties, such as researchers using the devices. This makes them challenging to rigorously evaluate. Another issue is when these algorithms are updated part-way through ongoing research studies. While these algorithms tend to be based on the Cole-Kripke or other previously published algorithms, relatively minor changes can result in significant differences in sleep scores. This not only makes it difficult to collect data from participants and athletes, if some of their data have been scored differently, but it also raises the question about whether new algorithms nullify previous validation studies.

One way of avoiding some of these issues is by understanding mathematically how actigraphic data distribute themselves in order to better predict sleep versus wake within a given window. Looking at circadian behaviours can help to predict quiet wakefulness during the day and night (figure 6; Adamec *et al.*, 2010).

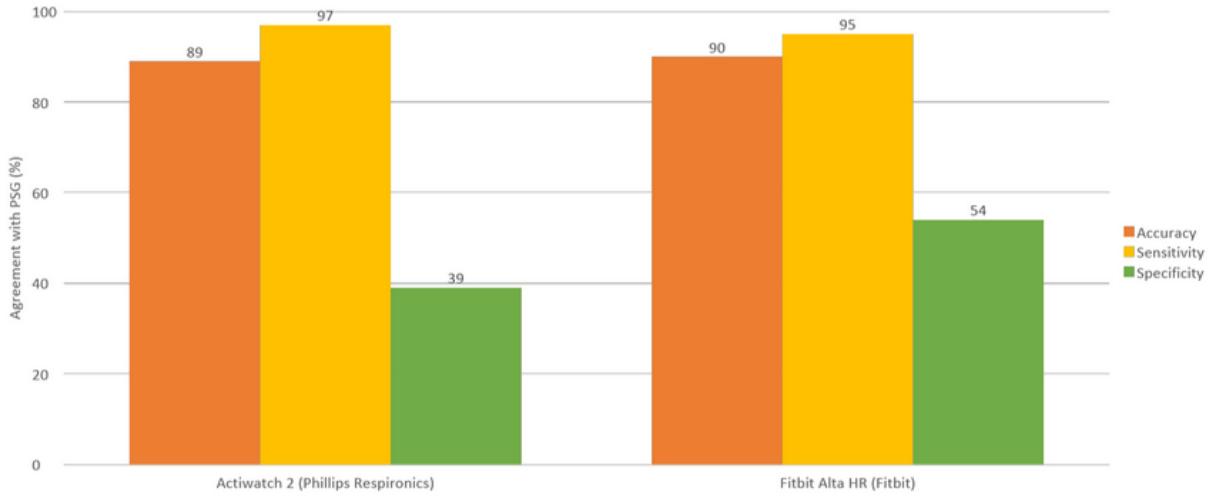
**Figure 6. Sleep versus wake**




Source: Adamec *et al.*, 2010, p. 1.

Adding heart rate as an extra measure beyond movement alone can significantly improve specificity. In figure 7, the Fitbit Alta HR (Fitbit) is 15% better at detecting wake than the Actiwatch 2 (Philips/Respironics).

**Figure 7. Fitbit Alta HR vs Actiwatch 2**

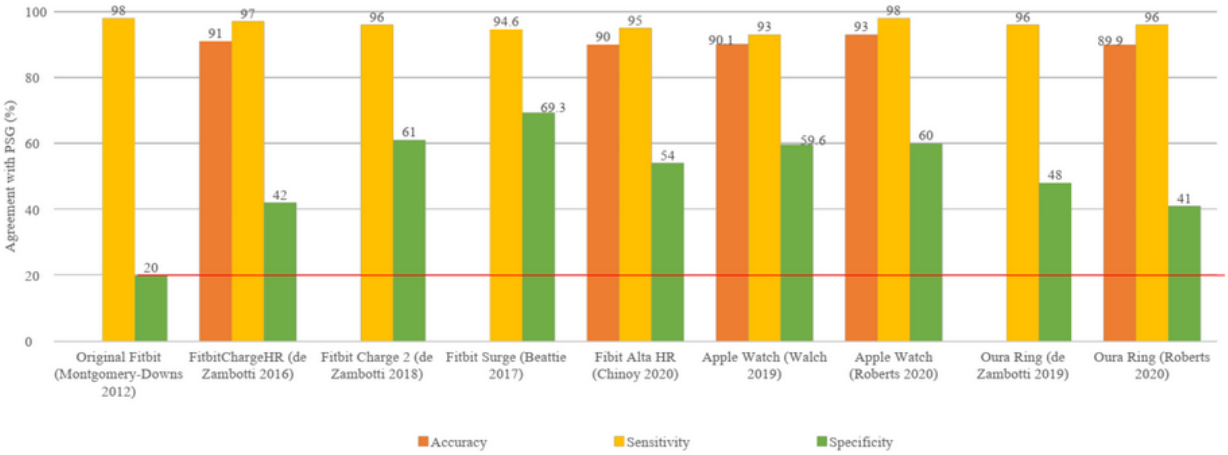



Source: Lujan *et al.*, 2021.

A comparison of today's next-generation devices compared to the original Fitbit (which did not record heart rate) can be seen in figure 8.

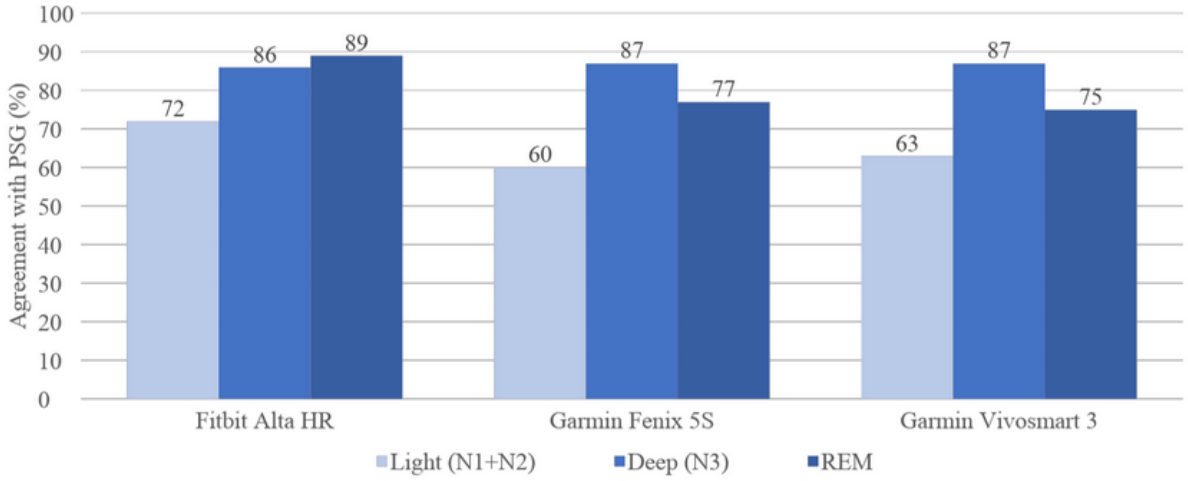
When it comes to detecting sleep stages, Fitbit validation data from 2017 showed a general prediction, but with a high margin of error (Beattie *et al.*, 2017). A more recent comparison of three popular devices can be seen in figure 9. While there is some variability, there remains an accuracy cap around 90%.

**Figure 8. Comparison of multiple devices**




Source: Lujan *et al.*, 2021.

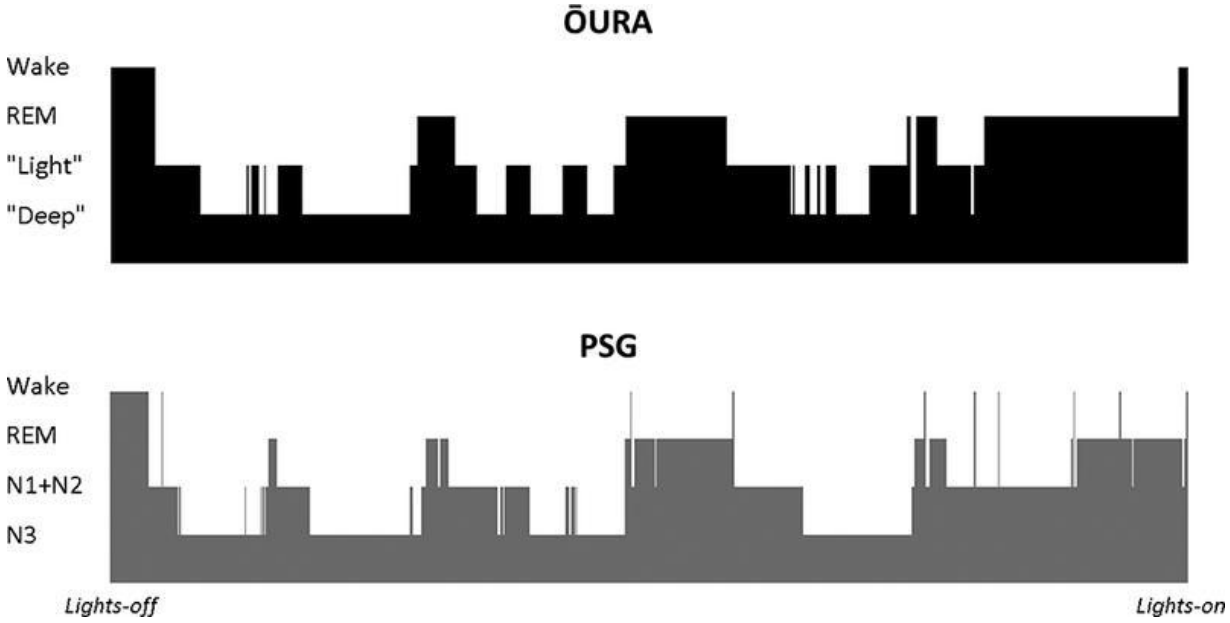
**Figure 9. Comparison of Fitbit and Garmin devices**




Source: Lujan *et al.*, 2021.

Figure 10 and table 6 show the performance of the Oura ring at detecting sleep versus wake, along with sleep architecture, compared with PSG data. The Oura data were calculated based on the first version of the Oura algorithm in a study that enrolled 41 adolescents and young adults, with sensitivity, specificity, and accuracy scores that were comparable to many wrist worn devices on the market.

**Figure 10. Oura validation data**




Source: De Zambotti *et al.*, 2019.

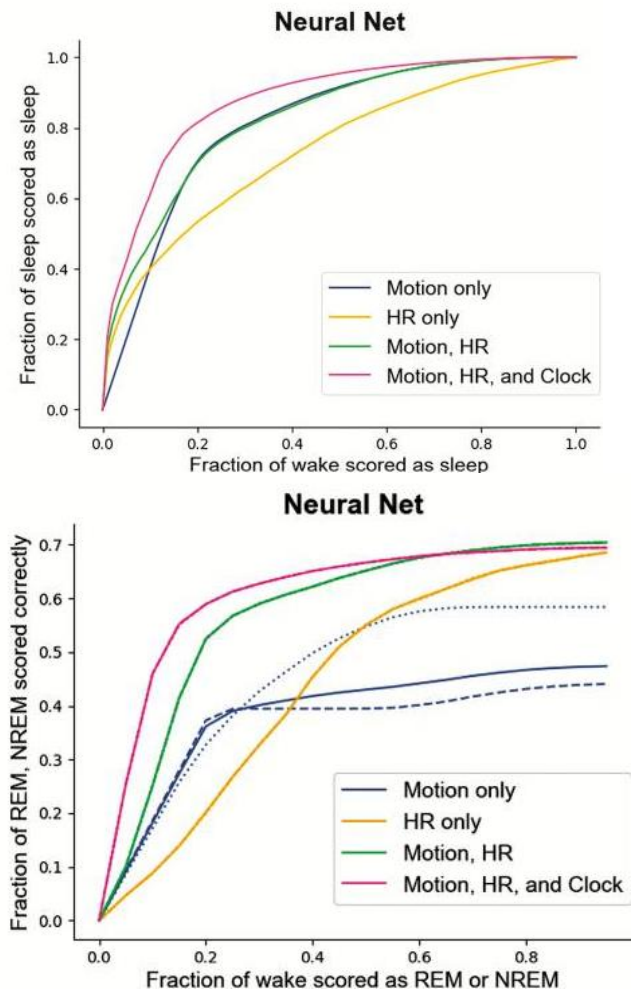
**Table 6. Oura validation data**

| | Mean ± SD | ± 95% CI of the bias |
|---|---|---|
| **Sensitivity** | 95.5 ± 4.5 | 94.1-96.9 |
| **Specificity** | 48.1 ± 19.1 | 42.0-54.1 |
| **Agreement for N1+N2 ("light sleep")** | 64.6 ± 13.9 | 60.3-69.0 |
| **Agreement for N3 ("deep sleep")** | 50.9 ± 24.5 | 43.2-58.6 |
| **Agreement for REM sleep** | 61.4 ± 22.8 | 54.2-68.6 |

Source: De Zambotti *et al.*, 2019.

Data comparing the Apple Watch (Apple Inc.) to PSG are shown in figure 11. This study enrolled 39 healthy participants using an algorithm designed by the researchers for the purpose of the study (Walch *et al.*, 2019). At the time of writing, there remains no native app for this device, so the researchers harnessed the raw data output. Using heart rate data and the clock as a proxy for circadian rhythm, the results demonstrated a good estimation of sleep versus wake, and moderate estimation of REM versus non-REM sleep.

**Figure 11. Apple Watch compared to PSG**




Source: Walch *et al.*, 2019, https://bit.ly/3SaFJpn.

Incorporating the clock time into the algorithm allowed the researchers to weight the likelihood of movement as representing wake (or stillness representing sleep) differently depending on time of day and at different stages of sleep. During stage 3/N3/slow-wave sleep, which typically occurs intermittently during the first half of the night, the body is most relaxed and movement is less likely to occur. Heart rate is also the slowest during this stage of sleep. During REM sleep, which generally occurs during the second half of the night, the muscles are essentially paralysed, but physiologic data shows a heart rate more similar to that which occurs during wake. Combining these multisensory measures can approximate sleep stages with some degree of accuracy, particularly when clock time is incorporated into the scoring algorithm.

The Whoop was compared to PSG in 2021 and it performed reasonably well in scoring sleep versus wake (90% sensitivity, 60% specificity, 86% accuracy.) The sleep staging data was also scored with between 60-66% accuracy (table 7).

**Table 7. Whoop validation**

| Four-Stage Error Matrix for WHOOP-AUTO and PSG | | | | | |
|---|---|---|---|---|---|
| | | WHOOP-AUTO | | | |
| | Stage | Wake | Light sleep | SWS | REM |
| PSG | Wake | 60% | 26% | 1% | 12% |
| | Light sleep | 14% | 61% | 10% | 15% |
| | SWS | 6% | 28% | 64% | 2% |
| | REM | 6% | 27% | 1% | 66% |
| **Notes: This matrix presents the percentage of each sleep stage that WHOOP-AUTO has correctly or incorrectly classified compared to PSG. Shaded cells indicate correctly classified sleep. SWS; slow wave sleep, REM; rapid eye movement sleep. | | | | | |

Source: [Untitled image about whoop validation], n. d.

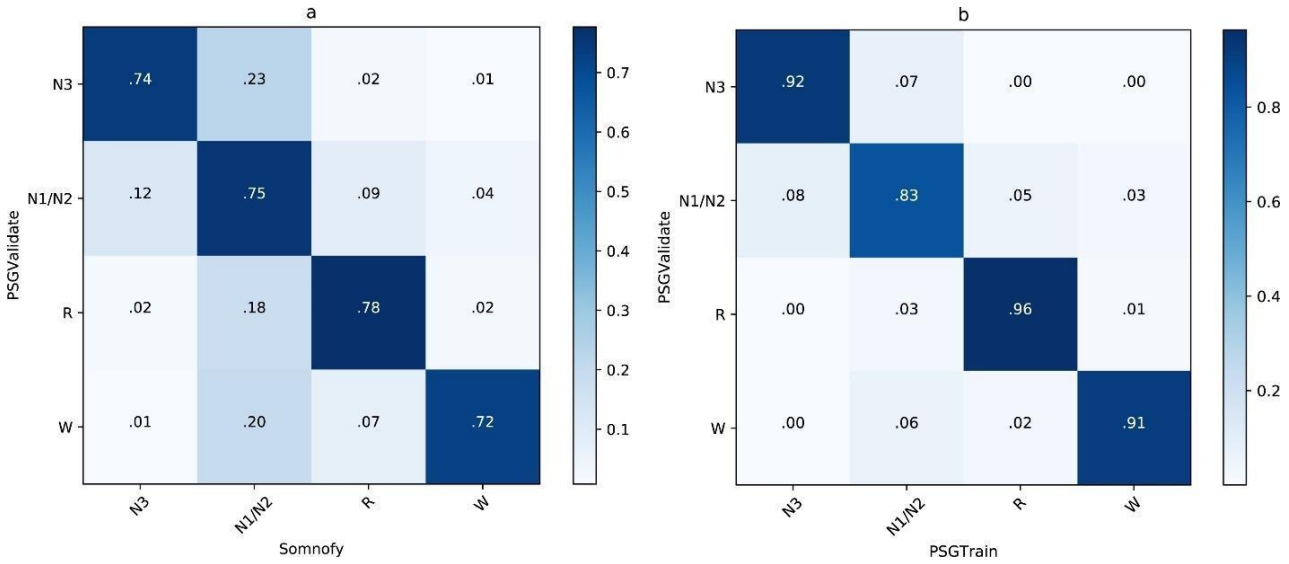
### Nearables

While sleep tracking typically happens via wearable devices, others assess sleep versus wake based on devices that are kept near the sleeper ('nearables'). These include sensors in, on, or near the bed.

Radio frequency, such as the S+/SleepScore (ResMed) was shown to be able to detect respiration with moderate accuracy from the bedside, with potential use for sleep apnoea diagnoses. A 2017 study assessing its ability to score sleep versus wake demonstrated 93-94% sensitivity, 70-73% specificity, and 61-62% sleep staging accuracy.

Somnofy (VitalThings) uses radar technology to monitor sleep from the bedside. Results from a 2020 performance study can be seen in table 8 and figure 12 (Toften *et al.*, 2020).

**Figure 12. Somnofy performance**




Source: Toften *et al.*, 2020, p. 59.

**Table 8. Somnofy performance**

| | PSG | Somnofy |
|---|---|---|
| | Mean (SD) | Mean (SD) |
| **Cohen's kappa** | 0.82 (0.10) | 0.63 (0.10) |
| **Accuracy** | 0.88 (0.06) | 0.76 (0.07) |
| **Sensitivity** | 0.99 (0.02) | 0.97 (0.03) |
| **Specificity** | 0.85 (0.11) | 0.72 (0.19) |

Source: Toften *et al.*, 2020, p. 59.

Under-bed sensors are another way of non-invasively tracking sleep and are more discrete than technologies used on the nightstand. These devices have their limitations; however, in that they do not capture out-of-bed activity. This may be useful, if nocturnal sleep is the only duration of interest. However, naps taken in places other than the bed, for example, would be missed. The EarlySense Live (EarlySense) performance compared to PSG can be seen in table 9 (Tal *et al.*, 2017). Sleep stages are approximated using pressure changes and heartbeat detection.

**Table 9. EarlySense performance**

| | | Reference Values (full PSG) | | | |
|---|---|---|---|---|---|
| **A** | **Piezoelectric contact-free system** | **Awake** | **REM** | **LS** | **SWS** |
| | **Awake** | 9,482 (80.4%) | 588 (5.4%) | 3,710 (9.7%) | 114 (1.1%) |
| | **REM** | 569 (4.8%) | 5,844 (53.7%) | 4,224 (11.1%) | 308 (3.0%) |
| | **LS** | 1,477 (12.5%) | 4,190 (38.5%) | 24,771 (64.9%) | 4,014 (39.7%) |
| | **SWS** | 265 (2.2%) | 254 (2.3%) | 5,471 (14.3%) | 5,684 (56.2%) |

| | | PSG Reference | |
|---|---|---|---|
| **B** | **Piezoelectric contact-free system** | **Wake** | **Asleep** |
| | **Wake** | 9,482 (80.4%) | 4,412 (7.5%) |
| | **Asleep** | 2,311 (19.6%) | 54,760 (92.5%) |

Contingency tables comparing values obtained with the piezoelectric contact-free system to those acquired with full polysomnography for all subjects in all three setups (85 nights).

Source: Tal *et al.*, 2017, p. 519.

Beddit is another sensor that is placed beneath bedding on top of the mattress. It records heart rate, snoring, breathing, and pressure and is able to approximate multiple sleep parameters. Data comparing the device to PSG can be seen in tables 10 and 11 (Tuominen *et al.*, 2019).

**Table 10. Beddit performance**

| | PSG | | | BST | | | |
|---|---|---|---|---|---|---|---|
| | n | Mean (SD) | Range | n | Mean (SD) | Range | *P* |
| **TST (mins)** | 19 | 412.8 (58.1) | 333.0-524.4 | 19 | 456.3 (52.0) | 363.0-531.6 | <0.001 |
| **SOL (mins)** | 17 | 34.1 (18.0) | 5.5-70.0 | 17 | 30.9 (20.4) | 8.0-86.0 | NS |
| **WASO (mins)** | 18 | 46.1 (33.0) | 13.0-121.5 | 18 | 13.5 (17.3) | 0.0-56.0 | <0.001 |
| **SE (%)** | 18 | 83.6 (7.9) | 68.0-91.8 | 18 | 90.6 (6.7) | 78.2-98.2 | <0.001 |

Source: Tuominen *et al.*, 2019, p. 485.

Abbreviations: TST, total sleep time; SOL, sleep onset latency; WASO, wake after sleep onset; SE, sleep efficiency.

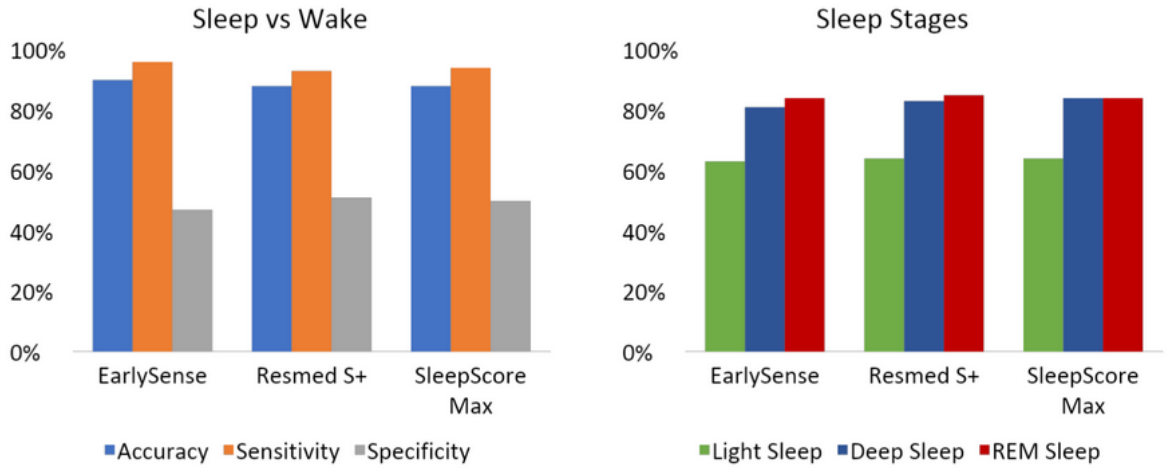
**Table 11. Beddit performance**

| | | PSG | | | |
|---|---|---|---|---|---|
| | | Wake (%) | N1 (%) | N2 (%) | N3 (%) |
| **BST** | Wake (%) | 42.1 | 9.9 | 40.2 | 7.8 |
| | N1 = Light (%) | 31.4 | 9.6 | 45.3 | 13.7 |
| | N2 = Light (%) | 15.4 | 8.7 | 49.6 | 28.3 |
| | N3 = Deep (%) | 5.8 | 4.8 | 51.9 | 37.5 |

Source: Tuominen *et al.*, 2019, p. 485.

A 2021 study compared the performance of the EarlySense Live, Resmed S+ and SleepScore in terms of scoring sleep versus wake and approximating sleep stages when compared to PSG (Chinoy *et al.*, 2021). As seen in figure 13, the data are relatively similar to many of the popular wearable devices currently on the market.

**Figure 13. EarlySense, Resmed S+ and Sleepscore performance**




Source: [Untitled image about EarlySense, Resmed S+ and Sleepscore performance], n. d.

Portable EEG headbands are a lesser-discussed category of wearables that approximate sleep by recording brain activity in addition to accelerometery. The Sleep Profiler (Advanced Brain Monitoring) records EEG, EOG, and EMG, along with head position and movement. A 2016 validation study by Finan and colleagues (2016) showed a 92% agreement with wake, 89% agreement with N1 sleep, 78% agreement with N2 sleep, 88% agreement with N3 sleep, and 85% agreement with REM. Sleep staging data tend to be markedly better in these devices as compared to those worn on the wrists or fingers.

The Dreem headband (Dreem) device is another that has gained popularity in recent years. Arnal and colleagues (2020) evaluated the performance of the Dreem headband compared to PSG The performance was evaluated, as compared to PSG, by Arnal and colleagues in 2020. The agreement scores can be seen in figure 14.

Another measure being incorporated into wearable devices is oxygen saturation. Intermittent decreases below the typical baseline of somewhere between 95-98% can indicate sleep apnoea (Ryan *et al.*, 1989).

**Figure 14. Confusian matrix of Dreem headband vs PSG**


Source: Arnal *et al.*, 2020.

Skin conductance and temperature can also be used to better approximate sleep. Skin conductance decreases during sleep, with the lowest point typically occurring during SWS. Core body temperature follows a similar pattern (Barrett *et al.*, 1993). When combined with movement and heart rate data, these parameters can provide greater insight into sleep physiology.

Other wearables are marketed as being able to assess inflammatory markers and hormonal changes throughout the night. Interleukin-6 (CRP), for example, is a cytokine that exhibits rhythmic behaviour and is partially sleep-dependent. C-reactive protein (CRP) is another sleep dependent molecule that is elevated in sleep deprivation (Haack *et al.*, 2007).

Cortisol is a hormone that is intrinsically tied to innate circadian rhythms and exhibits a peak during morning wakefulness and its lowest point during the deepest stages of sleep. Devices that can incorporate biological parameters such as this into their sleep scoring may be more accurate. These will inevitably be less discrete, however, which may be less preferred for certain studies or consumer preferences.

### Standards for validation

A 2020 article for manufacturers of new sleep technologies was created by sleep researchers as a 'roadmap' of standards that includes information about various metrics

that must be considered when assessing the performance of a new device (Depner *et al.*, 2020). It discusses the various proxies for measuring sleep, relevant circadian and sleep metrics, and the correct process for assessing new devices compared to PSG. It highlights the importance of having access to epoch-by-epoch data in order to perform rigorous analyses, rather than only having summary statistics.

Other important considerations include the algorithm that was used to predict sleep versus wake or sleep staging data. As of yet, there is no consensus on what is considered acceptable agreement with PSG. Because most devices struggle to detect quiet wakefulness and thus specificity is often low, devices are often comparable to one another, but may not be robustly comparable to PSG. As long as researchers, clinicians, and consumers are well-informed about device limitations and put recordings in a greater context, they may serve a valuable role, but often do not capture data with the level of accuracy that marketing claims often make.

Other parameters of performance include the device failure rate, because some may malfunction more often than others. It is also important to highlight who was tested using the device in a given study. Diversity in age, weight, skin tone, and other parameters may all affect outcomes. Also, was the study performed on devices worn in a laboratory, or at home? Studies performed in a laboratory can be more carefully controlled, but results from home are more generalisable.

Some studies have participants wear multiple devices at the same time. This may be fine for devices targeted to different areas of the body (such as one headband, one ring, and one wrist worn device) but will be less effective for multiple rings or multiple wrist worn devices. Ultimately, the results from devices worn in the usual area will be preferential.

Historically, the word 'validation' has been used in performance studies assessing wearable sleep technology against PSG. A 2021 commentary by Cathy A. Goldstein and Christopher Depner, two leading sleep researchers, suggested the rephrasing of this to "performance in context." The very specific criteria often employed in these studies may not allow the results to be generalisable to the population at large. Rather than using the word 'validation', which suggests a single event is required to deem wearable devices as acceptable for widespread use, 'performance evaluation' continues to investigate efficacy under various conditions and populations. However, a more rapid publication rate is needed to help close the gap between the speed of innovation and the slow pace of the peer review process for scientific articles.

Another limitation is the current resolution by which algorithms score sleep versus wake. Classic algorithms generally perform as low-pass filters, with values above or below a specific cut-off point classified as either sleep or wake. By aggregating epochs in a finite impulsive response (FIR) filter, this may improve the capacity of these devices to accurately score sleep-wake behaviour (Bieganski *et al.*, 2021).

# References


**Adamec, O., Domingues, A., Paiva, T., Sanches, J. M**. (2010). Statistical characterisation of actigraphy data during sleep and wakefulness states. *IEEE Engineering in Medicine and Biology Society*. 2342–2345.

**Arnal, P., Thorey, V., Debellemaniere, E., Ballard, M., Hernandez, A., Guillot, A., & Jourde, H., Harris, M., Guillard, M., Beers, P., Chennaoui, M., & Sauvet, F**. (2020). The Dreem Headband compared to Polysomnography for EEG Signal Acquisition and Sleep Staging. *Sleep*. http://dx.doi.org/10.1093/sleep/zsaa097.

**Barrett, J., Lack, L., & Morris, M**. (1993). The Sleep-Evoked Decrease of Body Temperature. *Sleep*, 16(2), 93-9. http://dx.doi.org/10.1093/sleep/16.2.93.

**Beattie, Z., Pantelopoulos, A., Ghoreyshi, A., & Oyang, Y**. (2017). Estimation of sleep stages using cardiac and accelerometer data from a wrist-worn device. *Sleep*. http://dx.doi.org/10.1093/sleepj/zsx050.067.

**Biegański, P., Stróż, A., Dovgialo, M., Duszyk-Bogorodzka, A., & Durka, P**. (2021). On the Unification of Common Actigraphic Data Scoring Algorithms. *Sensors (Basel, Switzerland)*, 21(18), 6313. https://doi.org/10.3390/s21186313.

**Chinoy, E. D., Cuellar, J. A., Huwa, K. E., Jameson, J. T., Watson, C. H., Bessman, S. C., Hirsch, D. A., Cooper, A. D., Drummond, S., & Markwald, R. R.** (2021). Performance of seven consumer sleep-tracking devices compared with polysomnography. *Sleep*, 44(5), zsaa291. https://doi.org/10.1093/sleep/zsaa291.

**Cole, R. J., Kripke, D. F., Gruen, W., & Mullaney, D. J.** (1992). Technical Note Automatic Sleep/Wake Identification From Wrist Activity. *Sleep,* 15(5), 461-9. http://dx.doi.org/10.1093/sleep/15.5.461.

**Depner, C., Cogswell, D., Bisesi, P., Markwald, R., Cruickshank-Quinn, C., Quinn, K., Melanson, E., Reisdorph, N., & Jr, K.** (2020). Developing preliminary blood metabolomics-based biomarkers of insufficient sleep in humans. *Sleep*, 43. http://dx.doi.org/10.1093/sleep/zsz321.

**Finan, P. H., Quartana, P. J., Remeniuk, B., Garland, E. L., Rhudy, J. L., Hand, M., Irwin, M. R., & Smith, M. T.** (2016). Partial sleep deprivation attenuates the positive affective system: Effects across multiple measurement modalities. *Sleep*, 40. https://doi.org/10.1093%2Fsleep%2Fzsw017.

**Haack, M., Sanchez, E., & Mullington, J. M.** (2007). Elevated inflammatory markers in response to prolonged sleep restriction are associated with increased pain experience in healthy volunteers. *Sleep,* 30, 1145.

**Kripke, D. F., Mullaney, D. J., Messin, S., & Wyborney, V. G.** (1978). Wrist actigraphic measures of sleep and rhythms. *Electroencephalography and Clinical Neurophysiology*, 44(5), 674-6. https://doi.org/10.1016/0013-4694(78)90133-5.

**Kripke, D. F., Hahn, E. K., Prorock Grizas, A., & Wadiak, K. H.** (2010). Wrist actigraphic scoring for sleep laboratory patients: Algorithm development. *Journal of Sleep Research*, 19(4), 612-9. http://dx.doi.org/10.1111/j.1365-2869.2010.00835.x.

**Kupfer, D. J., Himmelhoch, J. M., Swartzburg, M., Anderson, C., Byck, R., Detre, T. P.** (1972). Hypersomnia in manic-depressive disease: A preliminary report. *Diseases of the Nervous System*, 33(11), 720–724.

**Leproult, R., Copinschi, G., Buxton, O., & Van Cauter, E**. (1997). Sleep loss results in an elevation of cortisol levels the next evening. *Sleep*, 20(10), 865–870.

**Marino, M., Li, Y., Rueschman, M. N., Winkelman, J. W., Ellenbogen, J. M., Solet, J. M., Dulin, H., Berkman, L. F., & Buxton, O. M**. (2013). Measuring Sleep: Accuracy, Sensitivity, and Specificity of Wrist Actigraphy Compared to Polysomnography. *Sleep*, 36(11), 1747–1755. https://doi.org/10.5665%2Fsleep.3142.

**Morgenthaler, T., Kramer, M., Alessi, C., Friedman, L., Boehlecke, B., Brown, T. Coleman, J., Kapur, V., Lee Chiong, T., Owens, J., Pancer, J., & Swick, T.** (2006). Practice Parameters for the Psychological and Behavioural Treatment of Insomnia: An Update. An American Academy of Sleep Medicine Report. *Sleep*, 29, 1415–1419.

**Mullaney, D. J., Kripke, D. F., & Messin, S**. (1980). Wrist-Actigraphic Estimation of Sleep Time. *Sleep*, 3(1), 83-92.

**Okudaira, N., Fukuda, H., Nishihara, K., Ohtani, K., Endo, S., & Torii, S**. (1983). Sleep apnoea and nocturnal myoclonus in elderly persons in Vilcabomba, Ecuador. *Journal of Gerontology*, (38), 436-438.

**Ryan, T., Mlynczac, S., Ericson, T., Paul Man, S. F., & Godfrey, C. W**. (1989). Oxygen Consumption During Sleep: Influence of Sleep Stage and Time of Night. *Sleep*, 12(3), 201-210.

**Sadeh, A., Lavie, P., Scher, A., Tirosh, E., & Epstein, R**. (1991). Actigraphic home-monitoring of sleep-disturbed and control infants and young children: A new method for paediatric assessment of sleepwake patterns. *Pediatrics*, 87,494-499.

**Sadeh, A., Hauri, P. J., Kripke, D. F., Lavie, P**. (2005). The role of actigraphy in the evaluation of sleep disorders. *Sleep*, 18, 288–302.

**Tal, A., Shinar, Z., Shaki, D., Codish, S., & Goldbart, A**. (2017). Validation of Contact-Free Sleep Monitoring Device with Comparison to Polysomnography. *Journal of Clinical Sleep Medicine*, 13(3), 517-522. https://doi.org/10.5664/jcsm.6514.

**Te Lindert, B. H. W., & Van Someren, E. J. W**. (2013). Sleep estimates using microelectromechanical systems (MEMS). *Sleep*, 36(5), 781-9.

**Toften, S., Pallesen, S., Hrozanova, M., Moen, F., Grønli, J**. (2020). Validation of sleep stage classification using non-contact radar technology and machine learning (Somnofy®). *Sleep Medicine*, 75, 54-61.

**Tuominen, J., Peltola, K., Saaresranta, T. & Valli, K**. (2019). Sleep Parameter Assessment Accuracy of a Consumer Home Sleep Monitoring Ballistocardiograph Beddit Sleep Tracker: A Validation Study. *Journal of Clinical Sleep Medicine*, 15, 483-487. http://dx.doi.org/10.5664/jcsm.7682.

**Walch, O., Huang, Y., Forger, D., & Goldstein, C**. (2019). Sleep stage prediction with raw acceleration and photoplethysmography heart rate data derived from a consumer wearable device. *Sleep*. https://doi.org/10.1093%2Fsleep%2Fzsz180.