



RESEARCH ARTICLE

Metrology of two wearable sleep trackers against polysomnography in patients with sleep complaints

Justine Fria ^{1,2,3}  | Jimmy Mullaert ^{4,5} | Laure Abensur Vuillaume ⁶ |
 Mathieu Grajosz ^{1,7} | Ruben Wanono ¹ | Hélène Benzaquen ¹ | Fedja Kerzabi ¹ |
 Pierre Alexis Geoffroy ^{2,3} | Boris Matrot ² | Théo Trioux ⁴ | Thomas Penzel ⁸  |
 Marie-Pia d'Ortho ^{1,2,7}

¹Explorations Fonctionnelles et Centre du Sommeil- Département de Physiologie Clinique, APHP, Hôpital Bichat, Paris, France

²Université de Paris, NeuroDiderot, Inserm U1141, Paris, France

³Département de psychiatrie et d'addictologie, GHU Paris Nord, DMU Neurosciences, APHP, Hôpital Bichat Claude Bernard, Paris, France

⁴AP-HP, Hôpital Bichat, DEBRC, Paris, France

⁵Université de Paris, IAME, INSERM, Paris, France

⁶Emergency Department CHR Metz-Thionville, Metz, France

⁷Digital Medical Hub SAS, Assistance Publique Hôpitaux de Paris AP-HP, Hotel Dieu, Place du Parvis Notre Dame, Paris, France

⁸Interdisciplinary Sleep Medicine Center, Charité Universitätsmedizin Berlin, Berlin, Germany

Correspondence

Justine Fria, Physiologie-Explorations fonctionnelles, Hôpital Bichat-Claude Bernard, 46 rue Henri Huchard, 75018 Paris, France.
 Email: justine.fria@aphp.fr

Funding information

MSD Avenir Foundation

Summary

Sleep trackers are used widely by patients with sleep complaints, however their metrological validation is often poor and relies on healthy subjects. We assessed the metrological validity of two commercially available sleep trackers (Withings Activité/Fitbit Alta HR) through a prospective observational monocentric study, in adult patients referred for polysomnography (PSG). We compared the total sleep time (TST), REM time, REM latency, nonREM1 + 2 time, nonREM3 time, and wake after sleep onset (WASO). We report absolute and relative errors, Bland-Altman representations, and a contingency table of times spent in sleep stages with respect to PSG. Sixty-five patients were included (final sample size 58 for Withings and 52 for Fitbit). Both devices gave a relatively accurate sleep start time with a median absolute error of 5 (IQR -43; 27) min for Withings and -2.0 (-12.5; 4.2) min for Fitbit but both overestimated TST. Withings tended to underestimate WASO with a median absolute error of -25.0 (-61.5; -8.5) min, while Fitbit tended to overestimate it (median absolute error 10 (-18; 43) min. Withings underestimated light sleep and overestimated deep sleep, while Fitbit overestimated light and REM sleep and underestimated deep sleep. The overall kappas for concordance of each epoch between PSG and devices were low: 0.12 (95%CI 0.117-0.121) for Withings and VPSG indications 0.07 (95%CI 0.067-0.071) for Fitbit, as well as kappas for each VPSG indication 0.07 (95%CI 0.067-0.071). Thus, commercially available sleep trackers are not reliable for sleep architecture in patients with sleep complaints/pathologies and should not replace actigraphy and/or PSG.

KEYWORDS

hypersomnia, insomnia, polysomnography, sleep apnea, sleep trackers

Justine Fria and Jimmy Mullaert both contributed equally.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2024 The Authors. *Journal of Sleep Research* published by John Wiley & Sons Ltd on behalf of European Sleep Research Society.

1 | INTRODUCTION

Sleep accounts for approximately one third of a person's lifetime. Personal activity and sleep trackers have become increasingly popular among both healthy individuals and those with sleep disorders, providing a means to monitor sleep quality and quantity, but also circadian rhythms (Smith et al., 2018). Poor sleep quality and quantity, and their variability, are common complaints in patients suffering from chronic physical and mental conditions (Geoffroy et al., 2014; Smith et al., 2018). While sleep trackers could be of help to diagnose and follow-up these complaints in real-life settings and over time, most are not currently classified as medical devices and therefore not reimbursed to patients. Though promising in healthy individuals and selected populations (Moreno-Pino et al., 2019), the lack of metrological validation for personal connected devices limits their use in medical settings, preventing healthcare professionals from confidently recommending them to patients and analysing patients' data (de Zambotti et al., 2019; Frija-Masson et al., 2021; Moreno-Pino et al., 2019). They are also used widely in research studies to assess sleep in a non-invasive and continuous way. Data regarding their validity in selected populations is scarce, however (Sargent et al., 2018; Stucky et al., 2021). Most studies using PSG as a comparator report the global concordance for each sleep stage, but do not focus on sleep architecture (Kim et al., 2022; Moreno-Pino et al., 2019; Stucky et al., 2021). This is of importance, since several sleep diseases are characterised by a poor sleep architecture (insomnia, periodic limb movements, sleep apnea...). Some studies do not use PSG as a gold standard to assess the performance of the device (Xie et al., 2018). Finally, many studies have focussed either on healthy subjects or on only one disease (Kahawage et al., 2020; Kang et al., 2017; Moreno-Pino et al., 2019; Sargent et al., 2018; Stucky et al., 2021). This limits their use in the clinical setting, since their performance is likely to be different in unhealthy subjects, and could be different according to the disease.

Therefore, our aim was to evaluate the metrological validity of two widely used sleep trackers among patients seeking care for a sleep complaint at a sleep centre. The Fitbit Alta HR and Withings Activité were chosen, since at the time of the study, they accounted for an important number of sold wearable activity trackers in France, but their metrological validation was not found in the literature at the time of the study.

2 | METHODS

2.1 | Study design

We conducted a transversal study at a tertiary hospital sleep centre (Centre du sommeil, hôpital Bichat Claude Bernard, Paris, France). Consecutive adult patients referred for videopolysomnography (VPSG) were asked to participate. Patients were not included if they had known cardiac rhythm disorder, pacemaker, diabetes mellitus, or did not consent to participate. Two devices were tested during the

same night (Fitbit Alta HR and Withings Activité). A total of 65 patients were planned for each device, according to previous studies (Frija-Masson et al., 2021).

2.2 | Devices

Fitbit Alta HR and Withings Activité were chosen since they account for an important market in France and are supposed to focus on activity and sleep rather than sport. Both use an accelerometer and ballistocardiography to determine heart rate and activity, and then to deduce wake or sleep stage. Fitbit gives the results as an hypnogram, as well as the total sleep time, and the time spent in wake, light sleep, deep sleep, and REM sleep. Withings gives the result as a simplified hypnogram stating wake, light sleep and deep sleep, and a "sleep score" comprising the amount of deep sleep and the total sleep time. Both devices need a smartphone- or tablet-based application (Withings Health Mate and Fitbit) to extract data every day. Data are sent on a cloud and processed by the dedicated algorithms. No information is available on how these algorithms work.

These devices were bought on a research grant, Fitbit and Withings had no part in the study.

2.3 | Data acquisition

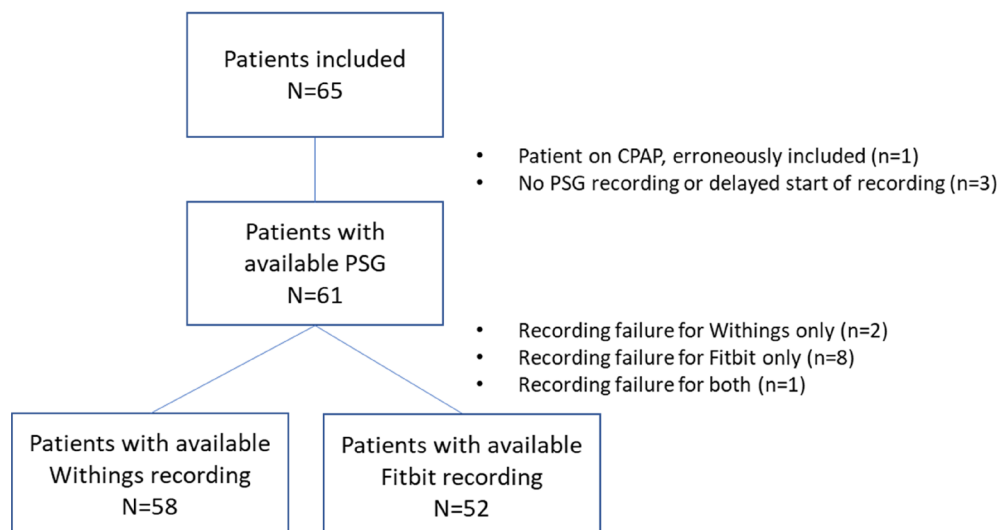
After obtaining informed consent, patients were instructed to wear both devices (Fitbit Alta HR and Withings Activité) on their non-dominant hand before and during full videopolysomnography (VPSG), so that both recordings could be compared. Patients were asked to take a walk (at least 15 minutes) shortly after the devices were installed to ensure proper detection of movement by the devices.

Full VPSG (Alice 6, Philips Respironics, or Morpheus, Micromed) was performed and scored following American Association of Sleep Medicine (AASM) guidelines by trained technicians and doctors (Berry et al., 2012). This scoring was considered as the gold standard.

2.4 | Data treatment

After manual scoring, raw VPSG data on Alice 6 were extracted (edf format); due to technical limitations of the software, raw data from Morpheus could not be extracted. Hypnograms from the connected devices were extracted with a screenshot from the manufacturer applications (Fitbit and Withings Health Mate, available on Google play), and then processed with image tools on R software to derive the sequence of sleep stages during the recording together with the transition times.

The following durations were compared between the gold standard and each device: total sleep time, REM time, REM latency, non-REM 1 + 2 time, non-REM 3 time and wake time during sleep. Of note, REM was not categorised by the Withings device and thus could not be compared with the gold standard. We also reported, separately for each device, a contingency table of times spent in sleep stages

FIGURE 1 Study flow chart.

with respect to the gold standard. This allowed us to estimate the agreement between VPSG and both devices on a finer scale, compared with aggregated measures.

2.5 | Statistics

Patient characteristics were described as median and interquartile range for quantitative variables and percentages for categorical variables. Absolute and relative errors for each outcome were reported with median and interquartile range. Bland–Altman representations were used to show systematic bias or trend in measurement error. Univariate linear models were used to estimate and test the association between measurement error and possible associated variables. For this analysis, we reported the slope estimation with its 95% confidence interval and the *p* value corresponding to the Wald statistic. Contingency tables containing the time in minutes spent in each sleep stage according to VPSG as the gold standard and each device were reported. A kappa coefficient with its 95% confidence interval is reported to assess agreement of sleep stages categorisation between both devices and the VPSG. We also reported kappa coefficients stratified by VPSG clinical indication: SAS, hypersomnia or insomnia. Missing data were not replaced. The significance threshold was 0.05. All analyses were performed using R software version 4.1.2 (R Foundation for Statistical Computing).

2.6 | Ethical aspects

This study is part of the Evaluation of the Metrological Reliability of Connected Objects in the Measurement of Medical Physiological Parameters (EvalExplo) study [NCT03803098]. Ethics approval was obtained from Comité de Protection des Personnes Sud Est VI (approval number AU 1443), and written non-opposition was obtained according to the Jardé decree in France.

3 | RESULTS

3.1 | Patient characteristics

In total 65 patients (men 56%, median age 54 years [40–62]) were included, with 61 VPSG available. After accounting for missing data (such as the device not retrieving data or technical failures during the VPSG), the final sample size for analysis was 58 for Withings and 52 for Fitbit (Figure 1). The median [Q1, Q3] body mass index was 26 [24–30] kg/m². 11% of patients were under cardiotropic medication. The Epworth sleepiness scale was available in 42 patients; the median was 10 [5–12], with 17 patients (40%) having a pathological score > 11. The indication for PSG was suspected sleep apnea syndrome (SAS) in 33 (54%) patients, hypersomnia in 13 (26%), insomnia in 13 (26%) and other in 2 (3%). The total sleep time was 385 [344–429] min, the apnea–hypopnea index (AHI) was 13 [7–25] events/hour.

3.2 | Accuracy of the devices for aggregated measures

Absolute and relative errors for aggregated measures and both devices are presented in Table 1. Both devices gave a relatively accurate sleep start time with a median absolute error of 5 (–43; 27) min for Withings and –2.0 (–12.5; 4.2) min for Fitbit. The absolute error for the sleep end time was comparable for both devices. However, both devices overestimated the total sleep time. Withings tended to underestimate wake after sleep onset (WASO) with an absolute error of –25.0 (–61.5; –8.5) min, while Fitbit tended to overestimate it (absolute error 10 (–18; 43) min). Withings underestimated light sleep and overestimated deep sleep, while Fitbit overestimated light and REM sleep and underestimated deep sleep. Fitbit was the only device to detect REM sleep, but tended to overestimate it (absolute error 12 (–12; 26) min).

TABLE 1 Absolute and relative errors for VPSG parameters.

Withings (N = 58)					Fitbit (N = 52)			
	PSG (min)	Device (min)	Absolute error (min)	Relative error (%)	PSG (min)	Device (min)	Absolute error (min)	Relative error (%)
Sleep start time			5 (−43; 27)	Not meaningful			−2.0 (−12.5; 4.2)	Not meaningful
Sleep end time			6 (4; 25)				6 (0; 56)	
Total sleep time	384 (344; 428)	422 (360; 483)	28 (−14; 63)	7.4 (−4.0; 16.6)	390 (345; 440)	414 (356; 472)	10 (−18; 43)	3.1 (−4.1; 11.4)
Wake after sleep onset	40 (23; 80)	15 (4; 31)	−25.0 (−61.5; −8.5)	−69 (−91; −33)	44 (24; 81)	58 (46; 75)	16 (−12; 32)	56 (−14; 114)
Non REM 1 + 2 time	215 (170; 254)	179 (146; 235)	−26 (−86; 26)	−14 (−38; 12)	227 (174; 268)	262 (235; 328)	52 (15; 90)	19.5 (5.9; 43.7)
Non REM 3 time	91 (58; 121)	220 (165; 276)	116 (65; 198)	139 (66; 300)	92 (57; 123)	56 (39; 76)	−29 (−69; −2)	−45.0 (−63.1; −2.6)
REM time	Not available for this device				72 (56; 93)	92 (58; 108)	12 (−12; 26)	18 (−14; 46)
REM latency					99 (73; 154)	116 (88; 181)	5.5 (−15.1; 58.1)	5.2 (−13.2; 71.0)

Note: Variables are described as median (Q1–Q3). Of note, the median absolute error cannot be obtained by taking the difference of medians from the PSG recording and the device.

Bland–Altman graphs are presented in Figure 2, which demonstrates that there is no linear trend for total sleep time for either device. However, significant linear trends for WASO and non-REM sleep for both devices, as well as for REM sleep on Fitbit, lead to high errors that are not clinically acceptable.

3.3 | Concordance with VPSG scoring

We extracted the raw data for PSG on Alice 6, and compared the epoch-to-epoch classification from PSG and each device. A total of 24 patients (169 hours of simultaneous recording of VPSG and device) were available for Withings and 21 patients (136 hours of simultaneous recording of VPSG and device) for Fitbit. This difference is explained by failed recordings, that were more frequent for Fitbit. Table 2 reports the time spent in each sleep stage according to the Withings device, for different stages of the VPSG: wake, non-REM sleep 1–2 and 3, and REM-sleep. REM sleep stages were included even for Withings, to assess how Withings would categorise them. The overall kappa was very low: 0.12 (95%CI 0.117–0.121), as well as kappas for each VPSG indication (kappa = 0.10 95%CI 0.097–0.102, kappa = 0.12 95%CI 0.116–0.123 and kappa = 0.23 95%CI 0.222–0.232 for SAS, insomnia and hypersomnia, respectively). Kappa values near 0 mean that the agreement between VPSG and categorisation of the device is mostly explained by chance. The same results for the Fitbit device are reported, including REM sleep. The overall kappa was 0.07 (95%CI 0.067–0.071) and kappas stratified by indication were 0.16 (95%CI 0.158–0.163), 0.00 (95%CI 0.00–0.00) and −0.15 (95%CI −0.154 to −0.147) for SAS, insomnia, and hypersomnia, respectively. This discrepancy is illustrated in Figure 3, a sample depicting a representative example of a hypnogram given by PSG and by each device for the same patient.

4 | DISCUSSION

This study prospectively evaluated the reliability of two different sleep trackers against polysomnography. Our findings demonstrate that these devices are not reliable in clinical practice for measuring global sleep parameters (such as total sleep time, WASO, NonREM3, and REM amount), nor for accurately assessing sleep architecture when comparing individual sleep stages. To our knowledge, this is the first study to focus on sleep architecture and to assess sleep stages individually in patients with a sleep complaint, regardless of the suspected underlying disease.

4.1 | Global performances of wearable devices

Data comparing wearable devices and VPSG are conflicting. Kim et al. found a significant correlation with VPSG for deep sleep and WASO only (Kim et al., 2022). Some studies found good correlations with actigraphy with selected devices (Henriksen et al., 2022; Kanady et al., 2020; Kang et al., 2017). Chinoy et al. assessed seven different devices and found significant differences in reliability, underlining the need for an individual validation of each tracker (Chinoy et al., 2021). This is problematic since there are multiple brands, each having several devices, and thus each patient should be counselled individually about their choice.

Xie et al. found an acceptable sleep duration estimation by different commercial devices, but the total sleep time comparator was self-reporting of the bed-time and arousal hours by the patients (Xie et al., 2018). Chen et al. found a good correlation between PSG and a forehead wearable device for sleep stages, but this type of device is not comparable to smart watches because it uses EEG (single channel) for scoring (Chen et al., 2023).

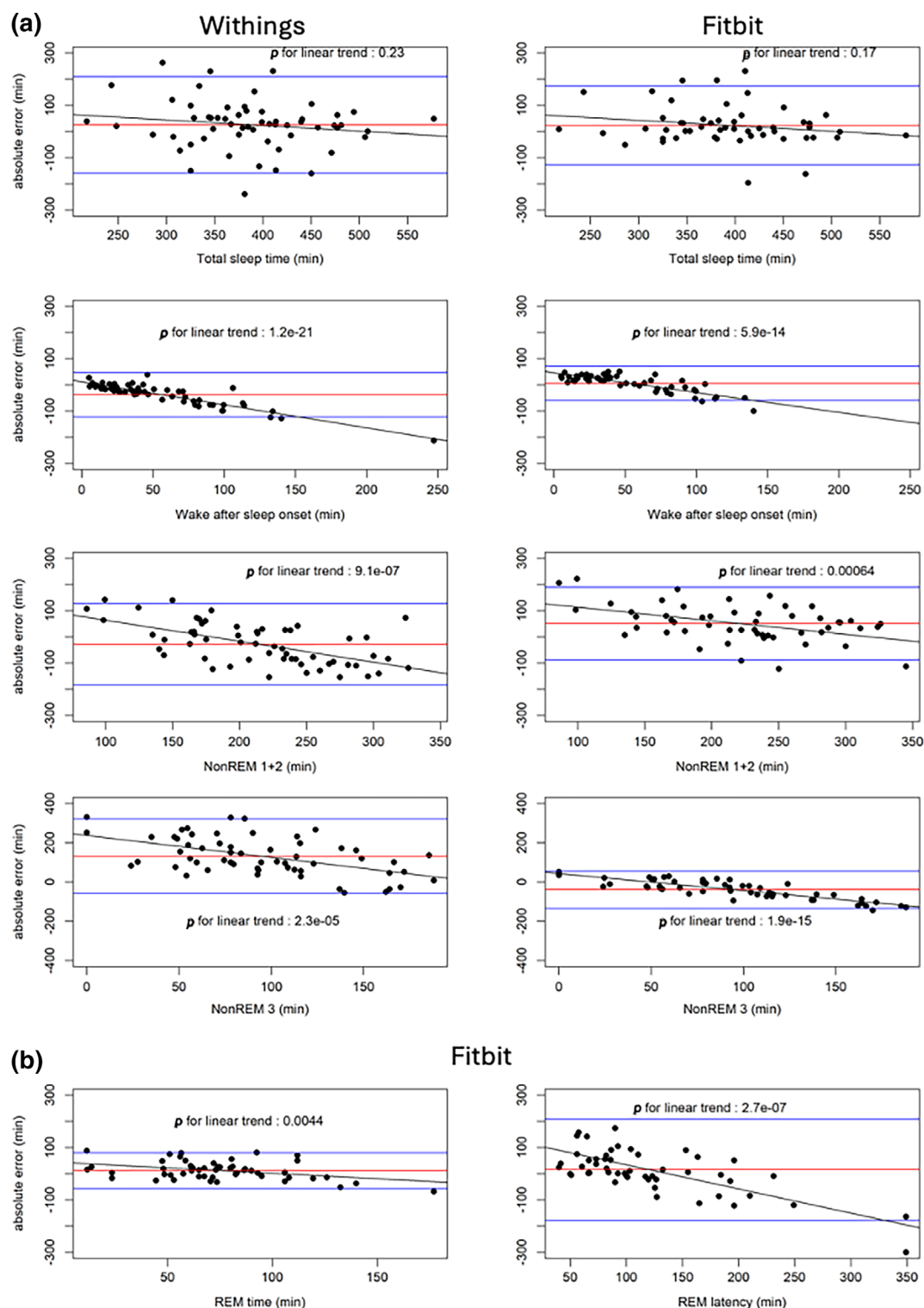


FIGURE 2 (a) Bland-Altman plots of both devices for Total sleep time (TST), wake after sleep onset (WASO), nonREM1 + 2, nonREM3. (b) Bland-Altman plots of Fitbit Alta HR for REM sleep time and latency. The red line indicates mean error, blue lines indicate 2.5 and 97.5 percentiles of the error distribution, and the black line represents a linear fit. For (a), left panel is for Withings Activite and right panel for Fitbit.

We did not focus on children since GDPR in Europe prevents the collection of health data in minors using non-medical devices, but similar results were found in other studies. Pesonen et al. found an underestimation of sleep in healthy children and adults by the

Polar Electro Oy compared with PSG (Pesonen & Kuula, 2018). Hakim et al. did not find a good correlation between Fitbit Charge and PSG in children with sleep disordered breathing (Hakim et al., 2022).

TABLE 2 Time spent in each sleep stage according to the device with respect to the gold standard categorisation ($N = 24$ patients, 169 h of simultaneous recording of VPSG and device for Withings and $N = 21$ patients, 136 h of simultaneous recording of VPSG and device for Fitbit).

PSG categorisation	Total cumulative recording time* (h)	Withings device categorisation (% relative to total recording time)			Total cumulative recording time* (h)	Fitbit device categorisation (% relative to total recording time)			
		Wake	Light	Deep		Wake	REM	Light	Deep
Wake	32	7.8	50.7	41.5	21	23.4	14.9	56.2	5.5
REM	28	1.1	43.4	55.6	25	10.7	20.0	53.5	15.8
NonREM1	19	2.9	50.6	46.6	16	11.4	19.2	61.1	8.3
NonREM2	59	2.7	46.8	50.6	51	7.7	21.3	58.1	12.9
NonREM3	31	3.0	25.0	71.9	24	8.8	17.9	52.7	20.6

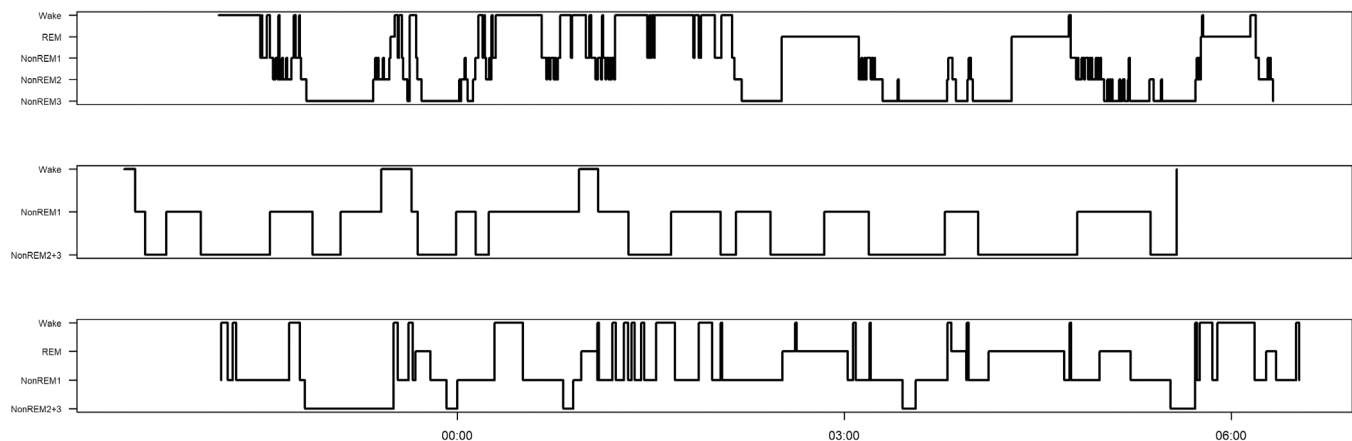


FIGURE 3 Comparison of the hypnograms given by VPSG (top), Withings Activité (middle) and Fitbit Alta HR (bottom) for one patient.

Sleep trackers are used widely in research studies to assess sleep in a non-invasive and continuous way. Data regarding their validity in selected populations is scarce, however. Sargent et al. found that Fitbit HR Charge did not detect well daytime naps in young athletes (Sargent et al., 2018). In shift workers, Stucky et al. found a good correlation between Fitbit Charge 2 and PSG for TST and heart rate, allowing a reasonable estimate of sleep (Stucky et al., 2021). But the authors underline the lack of concordance for sleep stages, limiting the use of such a device in a research study focussing on sleep in shift workers.

4.2 | Usability of wearable devices in insomnia

Wearable devices rely on accelerometer and ballistocardiography to assess sleep stages and wake detection. However, in a study by Kholghi, ballistocardiography was not accurate to detect sleep-wake transitions (Kholghi et al., 2022). Thus, in subjects complaining of poor sleep quality and insomnia, these devices may not accurately report the perturbations in hypnograms. Indeed, Kang et al. focussed on insomnia and compared the performances of Fitbit Flex to VPSG in insomnia patients and good sleepers (Kang et al., 2017). They found a

good agreement only for TST in insomnia, with an overestimation of TST by 32.9 min insomnia patients (versus 6.5 min in good sleepers) and a poor specificity to stage epochs correctly. The same results were found by Kahawage et al. in insomnia with Fitbit Alta HR (Kahawage et al., 2020). This poor epoch-by-epoch correlation is in accordance with our results and should discourage the use of sleep trackers for subjects with insomnia, although those subjects are the most likely to buy such a device.

4.3 | Usability of wearable devices in sleep apnea syndrome

We did not find a good correlation for sleep stages in patients with sleep apnea syndrome, neither with Fitbit nor with Withings. As discussed earlier, sleep-wake transitions are poorly recognised by devices; also, microarousals that are typical of sleep apnea syndrome are manually scored on VPSG but it is likely that their presence on sleep epochs limits the ability of the device to correctly assess the sleep stage. This is in accordance with the results of Moreno-Pino et al., who found a pretty good sensibility but a poor specificity for sleep stages in patients with OSA, including patients under CPAP

treatment, for two Fitbit devices (Moreno-Pino et al., 2019). The authors, however, did not assess sleep stages individually as we did in the present study. On the same level, Gruwez et al. did not find a good correlation for sleep stages and arousals in patients with OSA for Withings Pulse O2 (Gruwez et al., 2019). Again, no epoch-by-epoch comparison was available.

4.4 | Usability of wearable devices in central hypersomnolence

To our knowledge, only two studies focussed specifically on patients with suspected central hypersomnolence. Cook et al. evaluated the Fitbit Alta HR in 49 patients and show that the device was unable to detect any nocturnal SOREMPs, which are critical for diagnosis, and overestimated total sleep time, sleep efficiency, and deep sleep (Cook et al., 2019). The same authors found that the Jawbone 3 also had poor performances on night sleep, and also failed to detect SOREMPs during day multiple sleep latency tests (MLST), which are mandatory for diagnosing narcolepsy (Cook et al., 2018). These data are in accordance with our findings, although we did not focus on hypersomnolence, and does not allow for usage of these devices in patients complaining of hypersomnolence.

4.5 | Strengths and limits of the study

The strengths of our study are the systematic comparison epoch by epoch, the use of polysomnography as a gold standard, the use of hypnograms which allow for an easy comparison between devices and VPSG, and the inclusion of patients regardless of the suspected diagnosis and dominant sleep complaint. The limits are the one-night recording, the selected population of patients coming for in-hospital polysomnography, and the study of only two devices. Indeed, in the studies by Van Someren et al. and by Ravindran et al., single-night actigraphy had poor performances in sleep analysis, and acceptable reliability was reached after 7 days of recording (Ravindran et al., 2023a; Van Someren, 2007). In another study by Ravindran et al., the total sleep time was overestimated by the contactless sleep tracker and by actigraphy without a sleep diary, compared with actigraphy with a sleep diary (Ravindran et al., 2023b). These results emphasise the fact that actigraphy alone should not be used without clinical records, and that single-night actigraphy is not reliable for sleep analysis. Nonetheless, smart watches use ballistography as well as actigraphy and one could hope that their accuracy would be better than actigraphy alone. Unfortunately, it is not possible in the research setting to perform 7 nights in a row of both PSG and smart watch, but devices such as the Somno Art that give an accurate sleep analysis and are not invasive could be used as a control for several days at home in future studies.

The use of such devices during in-hospital PSG does not seem a limit in the evaluation of their reliability, since in-hospital and in-home PSG have similar performances.

4.6 | Future developments

The overestimation of TST by wearable devices is demonstrated by several studies using different devices (Kanady et al., 2020; Kang et al., 2017; Kholghi et al., 2022) and does not allow the use of such devices for at-home follow-up (“electronic sleep diary”). Still, in a recent meta-analysis by Haghayegh, focussing on Fitbit devices, the authors state that recent trackers, using heart rate variability and body positions and staging sleep, perform better than older non-sleep-staging devices (Haghayegh et al., 2019). The technological developments expected in such wearable devices will probably lead to better performances, and some firms such as Withings have started to focus on medical-grade device sleep trackers, for which performance is likely to be better and guaranteed by clinical studies. Indeed, the use of sleep trackers from a commercial point of view will allow firms to collect a lot of personal data, despite poor reliability, and some of those devices are clearly intended to be sold as well-being objects for profit only; on the contrary, doctors should focus on choosing and advising medical-grade devices that are accurate in the clinical setting. This is of great importance since these devices are more and more often worn by patients complaining of their sleep, but also since they are starting to be used widely in clinical studies to assess sleep in a non-invasive and continuous manner, not only in healthy subjects but also patients with chronic diseases (Mendelsohn et al., 2019; Niotis et al., 2021; Robbins et al., 2020). Finally, future developments in sleep tracking technology may focus on the use of individual sets of biomarkers, tailored to each patient's unique sleep characteristics (Ricka et al., 2023). This personalised approach could improve the accuracy of sleep screening and follow-up, leading to better clinical outcomes for patients.

In the future, good clinical validation studies using various devices, meta-analysis but also a reflection on sleep stages scoring by devices (Menghini et al., 2021; Nguyen et al., 2021; Penzel et al., 2021) will be necessary to ensure the construction of proper guidelines (Pires et al., 2023).

5 | CONCLUSION

Although sleep trackers have a wide public usage, our study shows that they are not clinically reliable for total sleep time and sleep architecture in patients with a sleep complaint and cannot replace PSG. Future studies should focus on evaluating specific sleep disorders and exploring alternative devices to identify reliable sleep trackers that can be recommended for clinical use.

AUTHOR CONTRIBUTIONS

Justine Frija: Conceptualization; investigation; writing – original draft; methodology; formal analysis; data curation; validation. **Jimmy Mullaert:** Conceptualization; writing – original draft; software; formal analysis. **Laure Abensur Vuillaume:** Writing – review and editing. **Mathieu Grajoszex:** Conceptualization; investigation; project administration; writing – review and editing. **Ruben Wanono:**

Conceptualization; investigation; writing – review and editing; formal analysis. **Hélène Benzaquen**: Investigation; writing – review and editing. **Fedja Kerzabi**: Investigation; data curation; project administration. **Pierre Alexis Geoffroy**: Writing – review and editing; supervision. **Boris Matrot**: Writing – review and editing; software. **Théo Trioux**: Formal analysis; data curation. **Thomas Penzel**: Writing – review and editing; supervision. **Marie-Pia d'Ortho**: Funding acquisition; writing – review and editing; conceptualization; investigation; supervision; validation.

ACKNOWLEDGEMENTS

The authors are grateful to the patients who agreed to participate. The authors thank the sleep technicians (Rémi Cellot, Béatrice Guy, Marie-Cécile Flottes, Carine Ecourtemer, Axelle Lefebvre-Roque, Rosine Zana, Claire Petrovic).

FUNDING INFORMATION

The Evalexpo study was funded through a donation from MSD Avenir Foundation to the Assistance Publique Hôpitaux de Paris Foundation.

CONFLICT OF INTEREST STATEMENT

None of the authors has a conflict of interest to declare.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

PATIENT CONSENT

Written non-opposition was obtained according to the Jardé decree in France.

PERMISSION TO REPRODUCE MATERIAL FROM OTHER SOURCES

Not applicable.

ORCID

Justine Fria  <https://orcid.org/0000-0001-5575-3913>

Thomas Penzel  <https://orcid.org/0000-0002-4304-0112>

REFERENCES

- Berry, R. B., Brooks, R., Gamaldo, C., Harding, S., Marcus, C., & Vaughn, B. (2012). *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications* (Vol. 176). American Academy of Sleep Medicine.
- Chen, X., Jin, X., Zhang, J., Ho, K. W., Wei, Y., & Cheng, H. (2023). Validation of a wearable forehead sleep recorder against polysomnography in sleep staging and desaturation events in a clinical sample. *Journal of Clinical Sleep Medicine*, 19(4), 711–718. <https://doi.org/10.5664/jcsm.10416>
- 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. P. A., & Markwald, R. R. (2021). Performance of seven consumer sleep-tracking devices compared with polysomnography. *Sleep*, 44(5) zsa291, 1–16. <https://doi.org/10.1093/sleep/zsaa291>
- Cook, J. D., Prairie, M. L., & Plante, D. T. (2018). Ability of the multisensory jawbone UP3 to quantify and classify sleep in patients with suspected central disorders of Hypersomnolence: A comparison against polysomnography and Actigraphy. *Journal of Clinical Sleep Medicine*, 14(5), 841c848. <https://doi.org/10.5664/jcsm.7120>
- Cook, J. D., Eftekar, S. C., Dallmann, E., Sippy, M., & Plante, D. T. (2019). Ability of the Fitbit Alta HR to quantify and classify sleep in patients with suspected central disorders of hypersomnolence: A comparison against polysomnography. *Journal of Sleep Research*, 28(4), e12789. <https://doi.org/10.1111/jsr.12789>
- de Zambotti, M., Cellini, N., Goldstone, A., Colrain, I. M., & Baker, F. C. (2019). Wearable sleep Technology in Clinical and Research Settings. *Medicine and Science in Sports and Exercise*, 51(7), 1538–1557. <https://doi.org/10.1249/MSS.0000000000001947>
- Frija-Masson, J., Mullaert, J., Vidal-Petiot, E., Pons-Kerjean, N., Flamant, M., & d'Ortho, M. P. (2021). Accuracy of smart scales on weight and body composition: Observational study. *JMIR mHealth and uHealth*, 9(4), e22487. <https://doi.org/10.2196/22487>
- Geoffroy, P. A., Boudebesse, C., Bellivier, F., Lajnef, M., Henry, C., Leboyer, M., Scott, J., & Etain, B. (2014). Sleep in remitted bipolar disorder: A naturalistic case-control study using actigraphy. *Journal of Affective Disorders*, 158, 1–7. <https://doi.org/10.1016/j.jad.2014.01.012>
- Gruwez, A., Bruyneel, A. V., & Bruyneel, M. (2019). The validity of two commercially-available sleep trackers and actigraphy for assessment of sleep parameters in obstructive sleep apnea patients. *PLoS One*, 14(1), e0210569. <https://doi.org/10.1371/journal.pone.0210569>
- Haghayegh, S., Khoshnevis, S., Smolensky, M. H., Diller, K. R., & Castriotta, R. J. (2019). Accuracy of wristband Fitbit models in assessing sleep: Systematic review and meta-analysis. *Journal of Medical Internet Research*, 21(11), e16273. <https://doi.org/10.2196/16273>
- Hakim, M., Miller, R., Hakim, M., Tumin, D., Tobias, J. D., Jatana, K. R., & Raman, V. T. (2022). Comparison of the Fitbit® charge and polysomnography for measuring sleep quality in children with sleep disordered breathing. *Minerva Pediatrics (Torino)*, 74(3), 259–263. <https://doi.org/10.23736/S2724-5276.18.05333-1>
- Henriksen, A., Svartdal, F., Grimsgaard, S., Hartvigsen, G., & Hopstock, L. A. (2022). Polar vantage and Oura physical activity and sleep trackers: Validation and comparison study. *JMIR Formative Research*, 6(5), e27248. <https://doi.org/10.2196/27248>
- Kahawage, P., Jumabhoy, R., Hamill, K., de Zambotti, M., & Drummond, S. P. A. (2020). Validity, potential clinical utility, and comparison of consumer and research-grade activity trackers in insomnia disorder I: In-lab validation against polysomnography. *Journal of Sleep Research*, 29(1), e12931. <https://doi.org/10.1111/jsr.12931>
- Kanady, J. C., Ruoff, L., Straus, L. D., Varbel, J., Metzler, T., Richards, A., Inslicht, S. S., O'Donovan, A., Hlavín, J., & Neylan, T. C. (2020). Validation of sleep measurement in a multisensor consumer grade wearable device in healthy young adults. *Journal of Clinical Sleep Medicine*, 16(6), 917–924. <https://doi.org/10.5664/jcsm.8362>
- Kang, S. G., Kang, J. M., Ko, K. P., Park, S. C., Mariani, S., & Weng, J. (2017). Validity of a commercial wearable sleep tracker in adult insomnia disorder patients and good sleepers. *Journal of Psychosomatic Research*, 97, 38–44. <https://doi.org/10.1016/j.jpsychores.2017.03.009>
- Kholghi, M., Szollosi, I., Hollamby, M., Bradford, D., & Zhang, Q. (2022). A validation study of a ballistocardiograph sleep tracker against polysomnography. *Journal of Clinical Sleep Medicine*, 18(4), 1203–1210. <https://doi.org/10.5664/jcsm.9754>
- Kim, K., Park, D. Y., Song, Y. J., Han, S., & Kim, H. J. (2022). Consumer-grade sleep trackers are still not up to par compared to polysomnography. *Sleep & Breathing*, 26(4), 1573–1582. <https://doi.org/10.1007/s11325-021-02493-y>
- Mendelsohn, D., Despot, I., Gooderham, P. A., Singhal, A., Redekop, G. J., & Toyota, B. D. (2019). Impact of work hours and sleep

- on well-being and burnout for physicians-in-training: The resident activity tracker evaluation study. *Medical Education*, 53(3), 306–315. <https://doi.org/10.1111/medu.13757>
- Menghini, L., Cellini, N., Goldstone, A., Baker, F. C., & de Zambotti, M. (2021). A standardized framework for testing the performance of sleep-tracking technology: Step-by-step guidelines and open-source code. *Sleep*, 44(2) zsa170, 1–12. <https://doi.org/10.1093/sleep/zsa170>
- Moreno-Pino, F., Porras-Segovia, A., López-Esteban, P., Artés, A., & Baca-García, E. (2019). Validation of Fitbit charge 2 and Fitbit Alta HR against polysomnography for assessing sleep in adults with obstructive sleep apnea. *Journal of Clinical Sleep Medicine*, 15(11), 1645–1653. <https://doi.org/10.5664/jcsm.8032>
- Nguyen, Q. N. T., Le, T., Huynh, Q. B. T., Setty, A., Vo, T. V., & Le, T. Q. (2021). Validation framework for sleep stage scoring in wearable sleep trackers and monitors with polysomnography ground truth. *Clocks Sleep*, 3(2), 274–288. <https://doi.org/10.3390/clockssleep3020017>
- Niotis, K., Saif, N., Simonetto, M., Wu, X., Yan, P., Lakis, J. P., Ariza, I. E., Buckholz, A. P., Sharma, N., Fink, M. E., & Isaacson, R. S. (2021). Feasibility of a wearable biosensor device to characterize exercise and sleep in neurology residents. *Expert Review of Medical Devices*, 18(11), 1123–1131. <https://doi.org/10.1080/17434440.2021.1990038>
- Penzel, T., Dietz-Terjung, S., Woehle, H., & Schöbel, C. (2021). New paths in respiratory sleep medicine: Consumer devices, e-health, and digital health measurements. *Sleep Medicine Clinics*, 16(4), 619–634. <https://doi.org/10.1016/j.jsmc.2021.08.006>
- Pesonen, A. K., & Kuula, L. (2018). The validity of a new consumer-targeted wrist device in sleep measurement: An overnight comparison against polysomnography in children and adolescents. *Journal of Clinical Sleep Medicine*, 14(4), 585–591. <https://doi.org/10.5664/jcsm.7050>
- Pires, G. N., Arnardóttir, E. S., Islind, A. S., Leppänen, T., & McNicholas, W. T. (2023). Consumer sleep technology for the screening of obstructive sleep apnea and snoring: Current status and a protocol for a systematic review and meta-analysis of diagnostic test accuracy. *Journal of Sleep Research*, 32(4), e13819. <https://doi.org/10.1111/jsr.13819>
- Ravindran, K. K. G., Della Monica, C., Atzori, G., Lambert, D., Hassanin, H., Revell, V., & Dijk, D.-J. (2023a). Three contactless sleep technologies compared with Actigraphy and polysomnography in a heterogeneous Group of Older men and Women in a model of mild sleep disturbance: Sleep laboratory study. *JMIR mHealth and uHealth*, 11, e46338. <https://doi.org/10.2196/46338>
- Ravindran, K. K. G., Della Monica, C., Atzori, G., Lambert, D., Hassanin, H., Revell, V., & Dijk, D.-J. (2023b). Contactless and longitudinal monitoring of nocturnal sleep and daytime naps in older men and women: A digital health technology evaluation study. *Sleep*, 46(10) zsad194, 1–18. <https://doi.org/10.1093/sleep/zsad194>
- Ricka, N., Pellegrin, G., Fompeyrine, D. A., Lahutte, B., & Geoffroy, P. A. (2023). Predictive biosignature of major depressive disorder derived from physiological measurements of outpatients using machine learning. *Scientific Reports*, 13(1), 6332. <https://doi.org/10.1038/s41598-023-33359-w>
- Robbins, R., Affouf, M., Seixas, A., Beaugris, L., Avirappattu, G., & Jean-Louis, G. (2020). Four-year trends in sleep duration and quality: A longitudinal study using data from a commercially available sleep tracker. *Journal of Medical Internet Research*, 22(2), e14735. <https://doi.org/10.2196/14735>
- Sargent, C., Lastella, M., Romy, G., Versey, N., Miller, D. J., & Roach, G. D. (2018). How well does a commercially available wearable device measure sleep in young athletes? *Chronobiology International*, 35(6), 754–758. <https://doi.org/10.1080/07420528.2018.1466800>
- Smith, M. T., McCrae, C. S., Cheung, J., Martin, J. L., Harrod, C. G., Heald, J. L., & Carden, K. A. (2018). Use of Actigraphy for the evaluation of sleep disorders and circadian rhythm sleep-wake disorders: An American Academy of sleep medicine systematic review, meta-analysis, and GRADE assessment. *Journal of Clinical Sleep Medicine*, 14(7), 1209–1230. <https://doi.org/10.5664/jcsm.7228>
- Stucky, B., Clark, I., Azza, Y., Karlen, W., Achermann, P., Kleim, B., & Landolt, H. P. (2021). Validation of Fitbit charge 2 sleep and heart rate estimates against polysomnographic measures in shift workers: Naturalistic study. *Journal of Medical Internet Research*, 23(10), e26476. <https://doi.org/10.2196/26476>
- Van Someren, E. J. W. (2007). Improving actigraphic sleep estimates in insomnia and dementia: How many nights? *Journal of Sleep Research*, 16(3), 269–275. <https://doi.org/10.1111/j.1365-2869.2007.00592.x>
- Xie, J., Wen, D., Liang, L., Jia, Y., Gao, L., & Lei, J. (2018). Evaluating the validity of current mainstream wearable devices in fitness tracking under various physical activities: Comparative study. *JMIR mHealth and uHealth*, 6(4), e94. <https://doi.org/10.2196/mhealth.9754>

How to cite this article: Frija, J., Mullaert, J., Abensur
Vuillaume, L., Grajoszex, M., Wanono, R., Benzaquen, H.,
Kerzabi, F., Geoffroy, P. A., Matrot, B., Trioux, T., Penzel, T., &
d'Ortho, M.-P. (2025). Metrology of two wearable sleep
trackers against polysomnography in patients with sleep
complaints. *Journal of Sleep Research*, 34(2), e14235. <https://doi.org/10.1111/jsr.14235>