



# BMJ Open Prospective cohort study to evaluate the accuracy of sleep measurement by consumer-grade smart devices compared with polysomnography in a sleep disorders population

Claire M Ellender <sup>1,2</sup>, Syeda Farah Zahir,<sup>3</sup> Hailey Meaklim,<sup>4</sup> Rosemarie Joyce,<sup>4</sup> David Cunningham <sup>4</sup>, John Swieca <sup>4</sup>

**To cite:** Ellender CM, Zahir SF, Meaklim H, *et al*. Prospective cohort study to evaluate the accuracy of sleep measurement by consumer-grade smart devices compared with polysomnography in a sleep disorders population. *BMJ Open* 2021;**11**:e044015. doi:10.1136/bmjopen-2020-044015

► Prepublication history for this paper is available online. To view these files, please visit the journal online (<http://dx.doi.org/10.1136/bmjopen-2020-044015>).

Received 26 August 2020  
Accepted 15 October 2021



© Author(s) (or their employer(s)) 2021. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

For numbered affiliations see end of article.

## Correspondence to

Dr Claire M Ellender;  
[claire.ellender@health.qld.gov.au](mailto:claire.ellender@health.qld.gov.au)

## ABSTRACT

**Objectives** Consumer-grade smart devices are now commonly used by the public to measure waking activity and sleep. However, the ability of these devices to accurately measure sleep in clinical populations warrants more examination. The aim of the present study was to assess the accuracy of three consumer-grade sleep monitors compared with gold standard polysomnography (PSG).

**Design** A prospective cohort study was performed.

**Setting** Adults undergoing PSG for investigation of a suspected sleep disorder.

**Participants** 54 sleep-clinic patients were assessed using three consumer-grade sleep monitors (Jawbone UP3, ResMed S+ and Beddit) in addition to PSG.

**Outcomes** Jawbone UP3, ResMed S+ and Beddit were compared with gold standard in-laboratory PSG on four major sleep parameters—total sleep time (TST), sleep onset latency (SOL), wake after sleep onset (WASO) and sleep efficiency (SE).

**Results** The accelerometer Jawbone UP3 was found to overestimate TST by 28 min (limits of agreement, LOA=−100.23 to 157.37), with reasonable agreement compared with gold standard for TST, WASO and SE. The doppler radar ResMed S+ device underestimated TST by 34 min (LOA=−257.06 to 188.34) and had poor absolute agreement compared with PSG for TST, SOL and SE. The mattress device, Beddit underestimated TST by 53 min (LOA=−238.79 to 132) on average and poor reliability compared with PSG for all measures except TST. High device synchronisation failure occurred, with 20% of recordings incomplete due to Bluetooth drop out and recording loss.

**Conclusion** Poor to moderate agreement was found between PSG and each of the tested devices, however, Jawbone UP3 had relatively better absolute agreement than other devices in sleep measurements compared with PSG. Consumer grade devices assessed do not have strong enough agreement with gold standard measurement to replace clinical evaluation and PSG sleep testing. The models tested here have been superseded and newer models may have increase accuracy and thus potentially powerful patient engagement tools for long-term sleep measurement.

## Strengths and limitations of this study

- Consumer grade devices were compared with gold standard in clinic patients.
- More than one device was included for comparison.
- This study includes measure of sleep parameters that clinicians frequently need to review in daily practice, such as total sleep time and sleep efficiency.
- High device failure was found in this study, confirming that consumer grade devices cannot be used to replace high fidelity diagnostic measurement.
- This sample had patients with sleep apnoea, insomnia or hypersomnia as their final sleep diagnosis.

## BACKGROUND

Poor sleep quality and duration has been shown to be an independent risk to overall mortality and for many chronic diseases.<sup>1</sup> The gold standard test for the measurement of sleep and diagnosis of sleep disorders is attended polysomnography (PSG). However, this is an involved and costly test, that requires complex equipment, dedicated space, trained staff and does not lend itself well to multi-night monitoring.

Sales of consumer sleep monitors and wearable consumer-grade smart devices have dramatically increased in recent years, with 33 million units estimated to have been sold in the USA in 2015<sup>2</sup> and the estimated value of the wearable industry in the USA expected to grow to US\$8.5 billion in 2020.<sup>3 4</sup> Consumer-grade devices fall into three major categories (i) wrist based devices (eg, Jawbone, FitBit); (ii) Bedside devices (eg, ResMed S+, Touch-Free Life Care) and (iii) Mattress-based devices (eg, Beddit, EarlySense Mattress, Emfit Bed Sensor). Each of the categories of devices use unique proprietary algorithms

for inferring wake/sleep, body position and measures of sleep quality.

The Jawbone UP (the precursor to the UP3 used in this study) has been compared with PSG in adolescents and concluded to have good agreements for total sleep time (TST), sleep efficiency (SE) and wake after sleep onset (WASO), however, the tendency to underestimate TST and SE increased with age.<sup>5</sup> In a study of adult women, the FitBitChargeHR overestimated TST by 27 min, and was found to have significantly different SOL and WASO compared with PSG.<sup>5</sup> Similarly in adolescents the Jawbone UP tended to overestimate TST and SOL, while underestimating WASO. The researchers also found greater discrepancies in nights when participants had more disrupted sleep (ie, lower TST and greater SOL and WASO).<sup>5</sup> In patients with suspected central disorders of hypersomnolence, the Jawbone UP3 was found to significantly overestimate TST by an average of 39.6 min compared with PSG and was not able to discriminate stages of sleep adequately.<sup>6</sup> Interestingly, the Jawbone UP3 performed similarly to actigraphy in this study. Another clinical study found that the FitBit Flex overestimated TST more in a group of insomnia patients compared with good sleepers (32.9 min vs 6.5 min).<sup>7</sup> Taken together, these two studies suggest that consumer-grade sleep devices are less accurate at measuring TST in a clinical sleep disorder population, than they are for good sleepers.

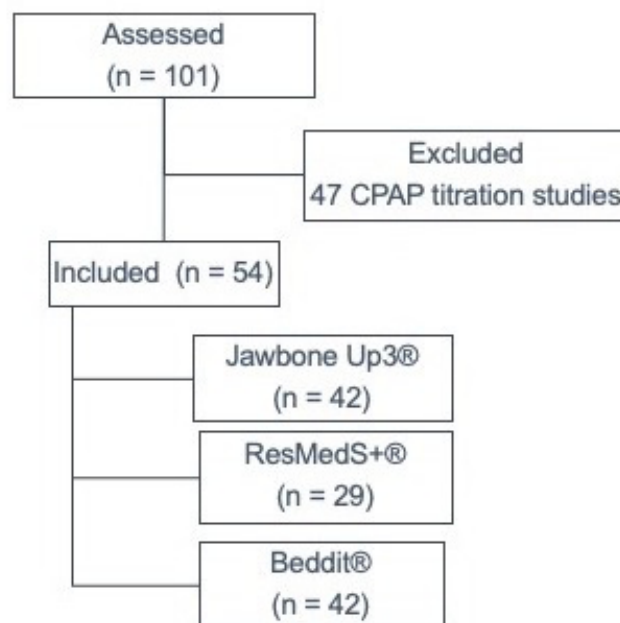
The Beddit mattresses based device has been found in 10 health controls to have poor agreement with TST (overestimated by 43.5 min), WASO and SE.<sup>8</sup> SOL was the only measure to have agreement, but had a wide variance.<sup>8</sup> The sensor technology used in the ResMed S+ device has been shown to have moderate accuracy in measuring TST and SE in healthy volunteers compared with PSG and high specificity.<sup>9,10</sup> Furthermore its utility in measuring sleep disordered breathing has been investigated and found to have reasonable accuracy in detecting moderate obstructive sleep apnoea, with a sensitivity of 89% and specificity 92%.<sup>11</sup>

Patients are increasingly attending sleep clinics with downloads from consumer-grade devices for discussion with primary care physicians and sleep specialists. These commonly encountered situations in the sleep clinic raise the questions: how reliable are consumer-grade devices, and which type of technology is most comparable to gold standard? This study aims to answer these questions with an in-laboratory comparison of PSG with the three consumer devices—Jaw Bone UP3, Beddit and ResMed S+ in a sleep clinic population. It was hypothesised that these devices would have similar accuracy in detecting TST, SOL, WASO and SE.

## METHODS

### Study population

Fifty-four adult patients were consecutively recruited through a private sleep disorders centre in Melbourne, Australia from June 2015 to February 2016. Inclusion



**Figure 1** Consolidated Standards of Reporting Trials statement of included participants. CPAP, continuous positive airway pressure.

criteria were age >18 years and any patient who required overnight PSG as standard investigation following sleep physician review to either confirm or exclude sleep disordered breathing. All patients attending the laboratory for a polysomnogram were screened for inclusion. Exclusion criteria were age <18 years, positive airway pressure titration study, pregnancy and cognitive impairment. **Figure 1** demonstrates the Consolidated Standards of Reporting Trials statement.

### Procedure

All assessments took place at an attended sleep laboratory in Melbourne, Australia. Sleep laboratory staff were trained to set up the three devices in addition to regular overnight PSG monitoring; lights out time was noted for synchronisation across all devices. The primary outcome measure was TST and secondary outcomes were sleep onset latency (SOL, min), SE (%) as TST/(TST+total wake time) and WASO (min). Other measures from the consumer grade devices such as time spent in light, deep or rapid eye movement sleep was not compared in this analysis.

### Polysomnographic recording

PSG was measured using a standard six-channel electroencephalography, submental electromyography and electrooculography, ECG, airflow (thermistor and nasal cannula), respiratory effort, oximetry, snoring (dB sound metre), body position, pulse rate, leg electromyography and digital video, recorded according to American Academy of Sleep Medicine standards.<sup>12</sup> The following standard sleep parameters were recorded via PSG: TST, SOL (min), total wake time (TWT, min), SE (%) as TST/

(TST +TWT) and WASO (min). Participants were classified as having obstructive sleep apnoea if the apnoea hypopnoea index was >5 events/hour. A single registered polysomnographic technologist scoring the PSG was blinded to the download of consumer grade devices and raw data were scored using Compumedics amplifiers and Profusion software V.3 (Compumedics, Abbotsford, Victoria, Australia).

### JawBone UP3

Participants were fitted with the JawBone Up3 on the participant's non-dominant wrist with the Jawbone UP3 shortly before lights out time. Data were collected via a dedicated iPod Touch, synced to the Jawbone app V.4.0.0.<sup>13</sup> This consumer-grade actigraphy device has a three-axis accelerometer and heart rate monitor, which together measure TST, SOL, WASO and SE which were exported by a technician the following morning after the PSG was complete.

### ResMed S+

The ResMed S+ is a non-contact radio-frequency sensor that continuously measures the biomotion due to breathing and body-movement in bed. The sensor operates in a license-free band at 5.8 GHz, emits an average power less than 1 mV and is capable of sensing movement and breathing over a distance ranging from 0.3 to 1.5 m. The device was positioned by the bedside and synced shortly before lights out time to a dedicated iPod with the ResMed S+ app V.1.2.1.<sup>14</sup> Measurements from the ResMed S+ were TST, SOL, WASO, SE which were exported by a technician the following morning after the PSG was complete.

### Beddit

The primary sensor in the Beddit is a piezoelectric 70 cm band that was attached to the mattress prior to patients getting into bed. The device detects micro-movements of the chest wall from heartbeats and respiration and uses ballistocardiography to infer sleep stage and time. Ballistocardiography is a non-invasive measurement of cardiac output and respiration by converting mechanical motion (eg, movement generated by a heartbeat) to a digital signal. Measurements from the Beddit were taken each night using the device synced to a dedicated iPod running the Beddit app V.1.<sup>15</sup> Output from the app included TST, SOL, WASO, SE and HR which were exported by a technician the following morning after the PSG was complete.

### Statistical analyses

Each of the three non-invasive devices was compared with PSG as the gold standard on an intention to treat basis. The primary and secondary outcomes were compared on total measurements over the night, not epoch-by-epoch method. Summary statistics of the study population are presented. For all normally distributed continuous variables mean and SD, whereas for non-normally distributed variables median and IQR were presented. Normality was assessed using the Shapiro-Wilk test. Frequencies

and proportions are presented for categorical variables. Extent of agreement and reliability between gold standard and each of the selected test devices, was assessed using intraclass correlation coefficients (ICCs) with two-way random-effects model. Agreement was considered moderate, good and excellent if the ICC values were between 0.5 and 0.75, 0.75 and 0.9 and >0.9, respectively.<sup>16</sup>

Additionally, Bland-Altman plots<sup>17</sup> were used to visualise the agreement between gold standard PSG and each of the selected devices. The average of two measurements was plotted on x-axis and difference between the two along y-axis. The mean of the differences provided an estimate of average bias between the methods. The upper and lower limits of agreement (LOA) were calculated which correspond to the mean difference (gold standard–selected method)±2 SD. LOA estimated the interval that a given proportion of differences between the measurements is likely to lie within and will be used to determine if the methods can be used interchangeably. Cohen's d is reported for the magnitude of the effect size. In case of non-normally distributed data, effect size 'r' was calculated by dividing Z statistic by the square root of the sample size (N). Interpretation of r is 0.10 to <0.3 (small effect), 0.30 to <0.5 (moderate effect) and ≥0.5 (large effect).<sup>18</sup> Data were analysed using R (V.4.0.4) (<https://www.r-project.org/>) (R Core Team, 2017).

### Patient and public involvement

Patients at our sleep disorders centre sparked the interest to assess the accuracy of consumer-grade sleep monitors. Our clinicians were often asked about the accuracy of home sleep monitors. To answer this question our team invited the patients to be involved in evaluating three commonly available consumer-grade smart devices. Participants were not paid for their involvement but did provide written consent. The findings of this research suggest that consumer-grade sleep monitors can give insights into trends in sleep but are not accurate enough to replace laboratory measurement.

### RESULTS

Fifty-four adult patients (57% females) with a mean age of 48.09 (±SD 18.05) years participated in this study. **Table 1** presents demographics of study population. The final sleep diagnosis found was obstructive sleep apnoea in 33 (61%), insomnia 9 (17%) and central hypersomnolence disorder in 12 (22%) participants. The mean PSG detected TST was 371 min (SD ±69), SOL of 16 min (SD ±15), WASO 63 min (SD ±56) and SE of 82% (SD ±13%). The absolute values of the measurements for each device are summarised in **table 2**. The results of the Bland-Altman analyses and intraclass correlation are summarised in **table 3** and displayed in **figures 2–4**.

### JawBone UP3

On average JawBone UP3 overestimated TST by 28.57 min (LOA=−100.23 to 157.37). By inspecting the



**Table 1** Patient demographics

Variable	Results (n=54)
Age in years, mean (SD)	48.09 ( $\pm$ SD 18.05)
Gender	31 (57%) women 23 (43%) men
BMI kg/m <sup>2</sup> , median (IQR)	27 (24–31)
PSG AHI events/hour, median (IQR)	9 (3–18.75)
Indication for PSG	
Rule in suspected OSA	32 (60%)
Rule out OSA	22 (40%)
Final clinical diagnosis	
OSA syndrome	33 (61%)
Insomnia	9 (17%)
Hypersomnia	12 (22%)

AHI, apnoea hypopnoea index; BMI, body mass index; OSA, obstructive sleep apnoea; PSG, polysomnogram.

Bland-Altman plots (shown in [figure 2A](#)), the cluster of points surrounded the mean tightly between 300 and 400 min and there was greater variability with TST below 300 min and above 400 min. The magnitude of effect size was small ( $d=0.44$ ). A moderate degree of reliability for recording TST was found between PSG and Jawbone UP3 with an ICC of 0.6 (95% CI 0.34 to 0.77;  $p<0.001$ ).

Bland-Altman plot ([figure 2B](#)) suggests that the mean difference in SOL between two methods was very small and on average Jawbone UP3 measured SOL 0.14 min (LOA= $-39.95$  to  $40.23$ ) more than the gold standard. The cluster of points surrounded the mean tightly on the left, with greater variability for values over 20 min. The magnitude of difference was small ( $r=0.13$ ). The reliability between the two methods was between poor to moderate (ICC=0.29; 95% CI  $-0.04$  to  $0.57$ ;  $p=0.04$ ).

Jawbone UP3 overestimated WASO only slightly, 1.7 min (LOA= $-102.32$  to  $105.71$ ,  $d=0.03$ ) compared with PSG. Greater variability was seen for measurements over 50 min (as shown in [figure 2C](#)), indicating better estimation of WASO by Jawbone UP3 at lower values. The agreement

**Table 2** Mean sleep duration

Variable	Device			
	PSG	Jawbone UP3 (N=42)	ResMed S+ (N=29)	Beddit (N=42)
TST (min SD $\pm$ )	371 $\pm$ 69	397 $\pm$ 83	345.8 $\pm$ 120	321 $\pm$ 107
SOL (min)	16 $\pm$ 15	18 $\pm$ 16	50 $\pm$ 44	60 $\pm$ 57
WASO (min)	63 $\pm$ 56	65 $\pm$ 55	80 $\pm$ 72	–
SE (%)	82.4 $\pm$ 13	82.9 $\pm$ 11	68.8 $\pm$ 21	81 $\pm$ 17

PSG, polysomnography; SE, sleep efficiency; SOL, sleep onset latency; TST, total sleep time; WASO, wake after sleep onset.

between Jawbone UP3 and PSG for WASO was poor to moderate (ICC=0.55; 95% CI 0.29 to 0.73;  $p<0.001$ ).

The mean difference in SE between two methods indicated that on an average Jawbone UP3 measures SE 0.51% (LOA:  $-18.96$  to  $19.99$ ) less than the gold standard. This bias seems to be due to measurements less than 85%, with better estimation of SE by Jawbone UP3 at higher SE, as seen in [figure 2D](#). The magnitude of difference was small ( $d=0.05$ ). The ICC for agreement between Jawbone UP3 and PSG regarding SE was 0.66 (95% CI 0.41 to 0.81;  $p<0.001$ ) indicating poor to good reliability between the two measures based on 95% CI.

### ResMed S+

As shown in [figure 3A](#), on average ResMed S+ underestimated TST by 34 min (95% CI  $-257$  min to  $188$  min). The mean difference between ResMed S+ measured and PSG measured TST was offset (lying below) zero, suggesting a bias. The points remained in the same general pattern for all x-axis values, except for few outliers at lower mean values. The magnitude of difference was moderate ( $r=0.4$ ). ICC of 0.36 (95% CI 0.02 to 0.63;  $p=0.02$ ) indicating poor to moderate reliability.

Conversely, ResMed S+ overestimated SOL by 35.6 min (LOA= $-57.68$  to  $-128.89$ ) and effect size was large ( $r=0.8$ ). Cluster of points go from below the mean at short SOL, to above the mean with increasing SOL, showing proportional error, suggesting overestimation of SOL by ResMed S+ at increasing SOL duration, as shown in [figure 3B](#). A poor agreement for SOL was seen between the two methods (ICC= $-0.01$ ; 95% CI  $-0.21$  to  $0.26$ ;  $p=0.51$ ).

Similarly, ResMed S+ recorded WASO 27 min more than PSG (LOA= $-73.53$  to  $127.91$ ) and a large effect was found ( $r=0.52$ ). Visual inspection of Bland-Altman plot ([figure 3C](#)) suggested that ResMed S+ increasingly overestimating WASO with increasing time. Reliability between methods was between poor to excellent (ICC=0.61; 95% CI 0.28 to 0.8,  $p<0.01$ ).

Visual inspection of the Bland-Altman plot [figure 3D](#) suggests that on average ResMed S+ underestimated SE by 16% (LOA= $-54.06$  to  $22.31$ ). The effect size was large ( $r=0.8$ ) and an ICC value of 0.28 (95% CI  $-0.06$  to  $0.58$ ;  $p=0.06$ ) was found. Moreover, the mean difference was not constant, with greater variability at lower values (particularly below 80%), showing proportional bias.

### Beddit

The Beddit and PSG had the least agreement for all outcomes except TST compared with other devices. TST was underestimated by 53 min (LOA= $-238.79$  to  $132$ ). As demonstrated in [figure 4A](#), the cluster of points shifted from below mean to above mean with increasing TST, showing a proportional error depending on the duration of sleep. The magnitude of difference was large ( $r=0.55$ ) and reliability poor to moderate (ICC=0.40; 95% CI 0.09 to 0.63;  $p=0.01$ ).

SOL was overestimated by 45 min (LOA= $-74.09$  to  $163.33$ ) by the Beddit compared with PSG. The points

**Table 3** Comparison of the outcomes between polysomnography (gold standard) and each of the selected methods

	TST (min)	SOL (min)	WASO (min)	Percentage
<b>Jawbone vs PSG Bland-Altman analysis</b>				
N	42	36	41	35
Bias	28.57	0.14	1.70	-0.51
LOA	-100.23 to 157.37	-39.95 to 40.23	-102.32 to 105.71	-19.99 to 18.96
<i>Cohen's d or r (magnitude)</i>	0.44 (small)	0.13* (small)	0.03 (small)	0.05 (small)
ICC	0.6 (95% CI 0.34 to 0.77; p<0.001)	0.29 (95% CI -0.04 to 0.57; p=0.04)	0.55 (95% CI 0.29 to 0.73; p<0.001).	0.65 (95% CI 0.41 to 0.81; p<0.001)
<b>ResMed S+ vs PSG Bland-Altman analysis</b>				
N	29	29	29	29
Bias	-34.36	35.60	27.19	-15.88
LOA	-257.06 to 188.34	-57.68 to -128.89	-73.53 to 127.91	-54.06 to 22.31
<i>Cohen's d or r (magnitude)</i>	*0.41 (moderate)	*0.81 (large)	*0.52 (large)	*0.8 (large)
ICC	0.36 (95% CI 0.02 to 0.63; p=0.02)	-0.01 (95% CI -0.21 to 0.26; p=0.51)	0.61 (95% CI 0.28 to 0.8; p<0.01)	0.06 (95% CI -0.06 to 0.58; p=0.06)
<b>Beddit vs PSG Bland-Altman analysis</b>				
N	42	42	NA	44
Bias	-53.39	44.62	NA	-1.35
LOA	-238.79 to 132	-74.09 to 163.33	NA	-38.81 to 36.11
<i>Cohen's d or r (magnitude)</i>	*0.55 (large)	*0.78 (large)	NA	*0.31 (small)
ICC	0.40 (95% CI 0.09 to 0.63; p=0.01)	0.004 (95% CI -0.173 to 0.22; p=0.48)	NA	0.26; 95% CI -0.04 to 0.51; p=0.06

\*Effect size=r.

Bias, the mean differences between test device minus PSG; LOA, limits of agreement (MD±2 SD); n, count of pairwise complete cases in groups; PSG, polysomnography; SE, sleep efficiency; SOL, sleep onset latency; TST, total sleep time; WASO, wake after sleep onset.

were tightly clustering above the mean, and go from above, to below the mean, from left to right (figure 4B), showing error proportional to the duration of SOL. The effect size was large (r=0.78) and reliability poor (ICC=0.004; 95% CI -0.173 to 0.22; p=0.48).

Beddit slightly underestimated SE by 1.35% (LOA=-38.81 to 36.11). As shown in figure 4C, variability of points was constant around the mean at values below 80%. This suggests that at higher values, Beddit estimated SE more closely to the PSG gold standard. The effect size was small (r=0.13) and poor agreement (ICC 0.26; 95% CI -0.04 to 0.51; p=0.06).

### Consumer-grade recording failure

Consumer-grade devices were set-up by Sleep Scientist staff each night at the time of the standard PSG set-up. Despite this, device or recording failure resulting in inability to record sufficient data, on the single night of recording, in the consumer-grade devices was common. Failure to synchronise with the dedicated Bluetooth device was the most common reason for device failure. The ResMed S+ failed to synchronise the most, with 25/54 nights (46%) resulting in recording failure. The Jawbone and Beddit had similar rates of synchronisation failure (12/54, 22%), however, not usually in the same room or on the same

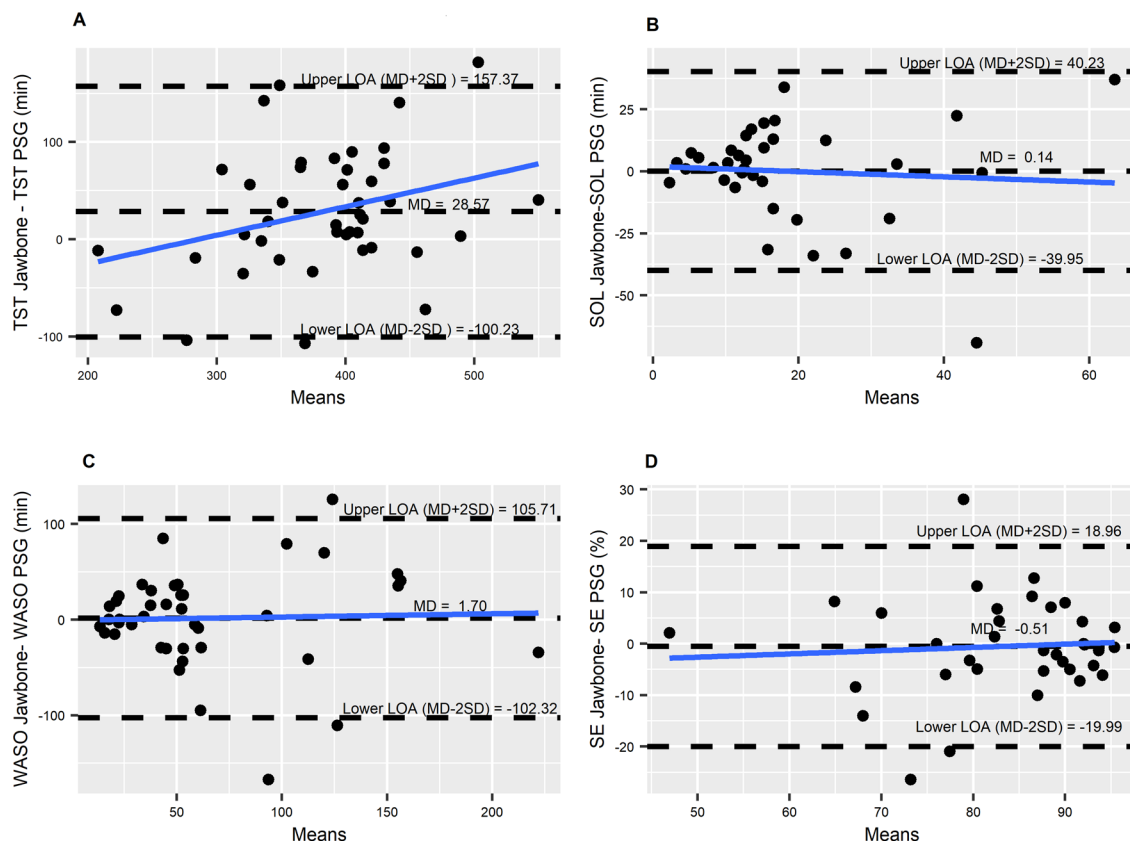
patient. Comparisons were made on an intention to treat analysis, even where large differences in TST were seen.

### DISCUSSION

The agreement of these three consumer-grade smart devices have simultaneously been compared with gold standard attended PSG in an adult sleep clinic cohort. For each of the devices, there were components of sleep measurement with poor to moderate agreement with the gold standard. This study found the primary outcome measure of TST was overestimated by, Jawbone UP3 whereas both ResMed S+ and Beddit underestimated it. The Jawbone UP3 also overestimated SOL and WASO, however, the magnitude of difference was very small. Generally Jawbone UP3 had better agreement across all outcomes, however for SE agreement was better between ResMed S+ and PSG. The Beddit had the least agreement with PSG, all components having poor agreement when compared with gold standard PSG.

Wearable devices, particularly wrist-worn accelerometers have now been widely compared with PSG. Similar to the results of this study, the accelerometers have been shown to overestimate TST by around 20–30 min,

## Bland-Altman plots for various outcomes measured by Jawbone and PSG



**Figure 2** Bland-Altman plot of the four outcomes (TST, SOL, WASO and SE) recorded by the Jawbone UP3 and PSG. The middle line represents the mean difference, and the upper and lower dotted line represents the upper and lower limits of agreement (mean difference $\pm$ 2 SD). The blue line is the line of best fit quantifying the difference between gold standard and new devices. (A) TST; (B) SOL; (C) WASO and (D) SE. MD, mean difference (or bias, in this panel a positive value indicates overestimation); LOA, lower limits of agreement; PSG, polysomnography; TST, total sleep time; SOL, sleep onset latency; WASO, wake after sleep onset; SE, sleep efficiency.

particularly in sleep disordered populations compared with healthy controls.<sup>5 7 19</sup> Previous investigations into consumer grade accelerometers in clinical populations found TST overestimated by 32.9 min<sup>7</sup> in a population of 33 insomnia patients and 39 min in 43 hyper-somnolence patients.<sup>6</sup> In our study, SOL had a large CI, with bias found with measurements over 15 min, consistent with findings of a recent systematic review and meta-analysis.<sup>20</sup>

The Beddit device and mattress devices in general are one of the least studied consumer grade devices. Tuominen *et al*<sup>8</sup> found in 10 healthy controls the Beddit overestimated TST by 43 min, whereas our data suggest a significant underestimation (PSG TST 371 min vs Beddit TST 321 min) with a larger sample size (n=42). Tuominen *et al*<sup>8</sup> were also able to access WASO data, which was not available with the model of Beddit tested in this study and found to underestimate WASO by 32 min. Non-wearable devices have a potential growing market as non-intrusive home monitors of sleep, as they can be applied in a ‘set and forget’ method. Thus, further refinement and evaluation of bed-based devices would be desirable.

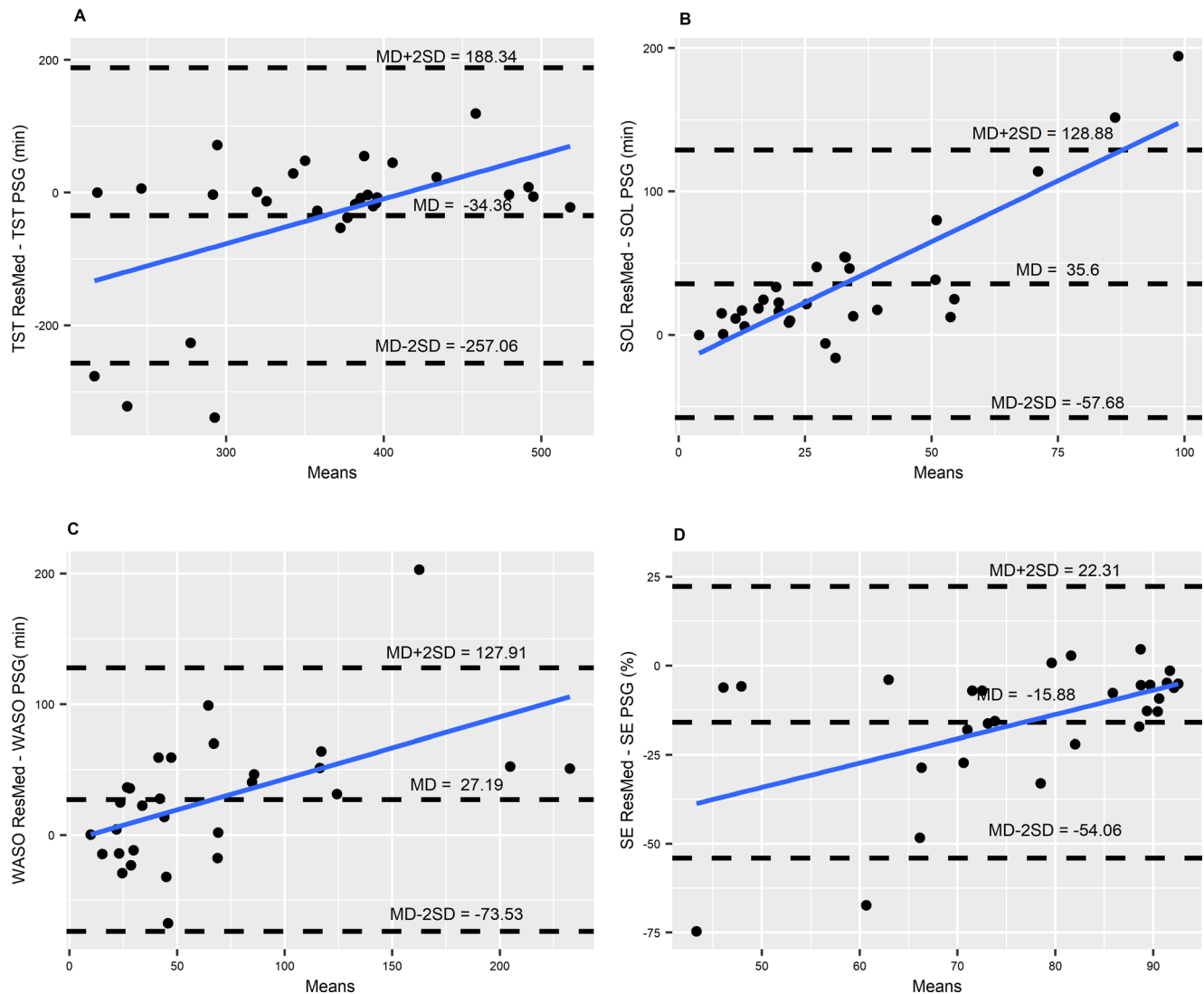
Chinoy *et al*<sup>10</sup> recently compared PSG to ResMed S+ and to SleepScore Max with a population of 19 young ‘healthy

normal’ individuals. The ResMed S+ was found to have underestimated TST by only 0.3 min (95% CI -70.7 to 70.2) and the SleepScore Max overestimate TST by 7.5 min (95% CI -60.7 to 75.7). A likely explanation for the difference these findings and the present study is the difference in population—‘healthy normal’ participants versus sleep clinic population. There is growing literature that consumer grade devices have lower accuracy in clinical population compared with control populations.<sup>21</sup> Notably, Chinoy *et al*<sup>10</sup> found 2/19 nights (10.5%) using the ResMed S+ were impacted by device synchronisation issues, requiring device re-synchronisation.

The high device synchronisation failure rate also observed in our study is concerning, despite the set-up being performed by sleep laboratory scientific staff. There is no way to calibrate these consumer-grade devices over time and it is difficult to monitor device connectivity to the Bluetooth device until the next morning. The high failure rate further confirms the role of these consumer devices is not to replace that of a diagnostic sleep study.

The main strength of this study was the sample size and that it was conducted in a clinical adult sleep population with a range of suspected sleep disorders. This makes

## Bland-Altman plots for various outcomes measured by ResMed S+ and PSG



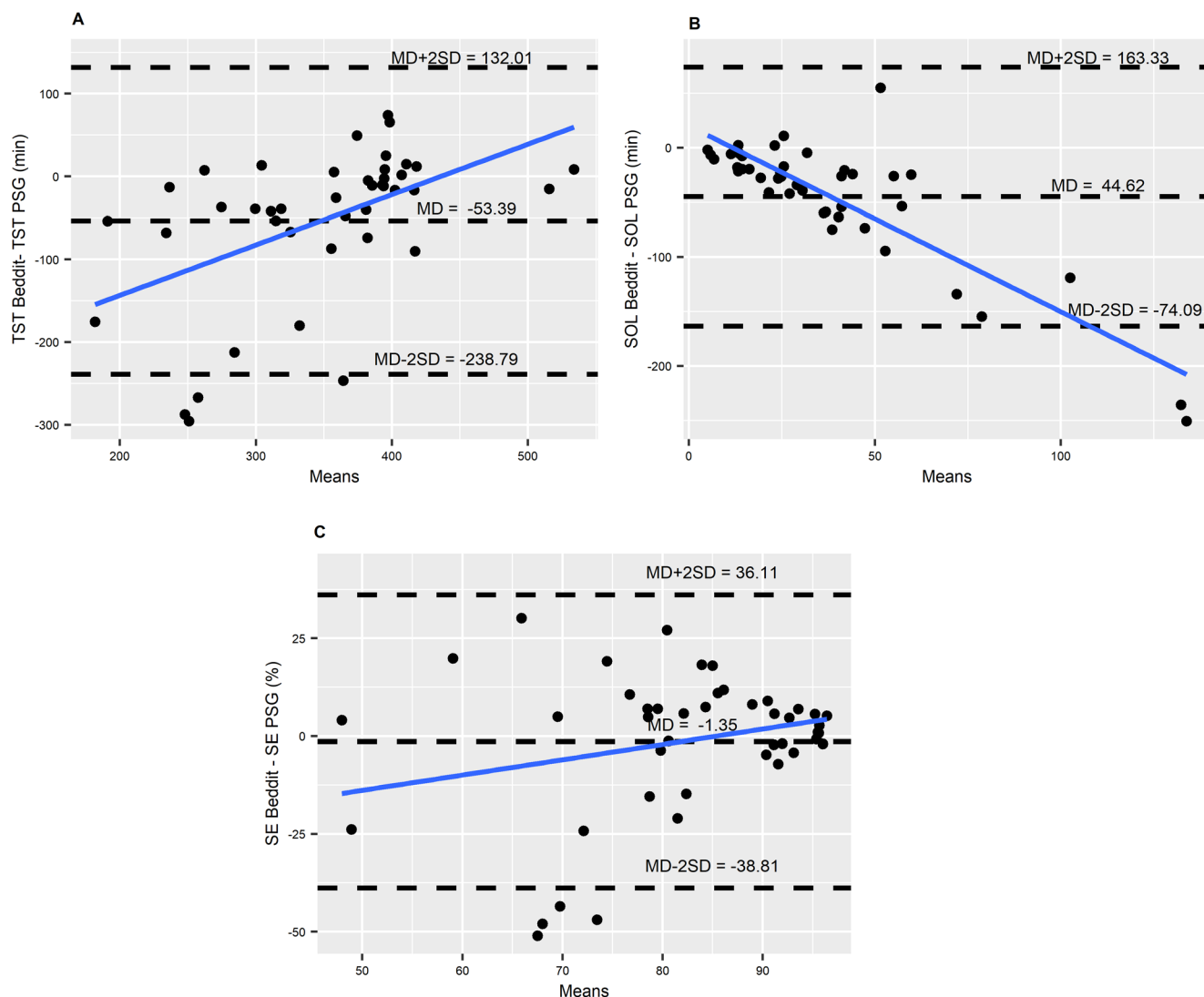
**Figure 3** Bland-Altman plot of the four outcomes (TST, SOL, WASO and SE) recorded by the ResMed S+ and PSG. The middle line represents the mean difference, and the upper and lower dotted line represents the upper and lower limits of agreement (mean difference  $\pm$  2 SD). The blue line is the line of best fit quantifying the difference between gold standard and new devices. (A) TST; (B) SOL; (C) WASO and (D) sleep efficiency. MD, mean difference (or bias, in this panel a positive value indicates overestimation); LOA, lower limits of agreement; PSG, polysomnography; TST, total sleep time; SOL, sleep onset latency; WASO, wake after sleep onset; SE, sleep efficiency.

the findings more translatable to clinicians managing patients with sleep disorders. Further, assessing a number of different devices is a novel approach. The weaknesses of the study include a high device recording failure rate, predominantly with Bluetooth synchronisation failure. Epoch-by-epoch analysis was not performed. Further, sales of devices tested in this study have since been discontinued. Beddit was acquired by Apple Inc in May 2017 and relaunched an updated device, the Beddit 3.5 which has reportedly improved integration with mobile phone health kits.<sup>22</sup> The ResMed S+ was discontinued and subsequently a similar device was launched in 2017 as Sleep-Score labs, which is similarly Apple iOS and Android integrated.<sup>23</sup> Jawbone however has gone into liquidation

with no subsequent models leading on from the UP3 device.<sup>24</sup>

This study indicates that the wrist worn Jawbone UP3 had the best agreement in measuring sleep compared with gold standard and can provide useful information about commonly measured parameters of sleep quality. For Sleep Medicine Clinicians, the translation of these findings, is that when our patients present with longitudinal measurements of sleep from their consumer grade devices, we can be reassured that wrist worn devices have reasonable accuracy and can be harnessed as an engagement tool for behavioural sleep interventions. This is consistent message with the American Academy of Sleep Medicine's position statement about the use of

## Bland-Altman plots for various outcomes measured by Beddit and PSG



**Figure 4** Bland-Altman plot of three outcomes (TST, SOL and SE) recorded by the Beddit and PSG. The middle line represents the mean difference, and the upper and lower dotted line represents the upper and lower limits of agreement (mean difference  $\pm 2$  SD). The blue line is the line of best fit quantifying the difference between gold standard and new devices. (A) TST; (B) SOL; (C) SE. MD, mean difference (or bias, in this panel a positive value indicates overestimation); LOA, lower limits of agreement; PSG, polysomnography; TST, total sleep time; SOL, sleep onset latency; WASO, wake after sleep onset; SE, sleep efficiency.

consumer-grade sleep devices stating that these devices cannot be used for clinical diagnosis, however they allow for meaningful discussions with patients about sleep and encourage active participation in sleep-related healthcare.<sup>25</sup>

## CONCLUSION

Given the large body of literature linking sleep quality to mortality and many chronic diseases, patient-collected longitudinal sleep data provides a powerful insight into a patient's overall health. This study adds to the data of consumer grade wearable sleep monitors, showing they can provide some reliable information compared with gold standard PSG, however do not replace clinical evaluation

and gold-standard PSG sleep testing. In reviewing sleep data collected by patients with consumer-grade devices, clinicians are encouraging measurement and quantification of sleep, which in turn will likely emphasise the importance of quality sleep in maintaining good health.

## Author affiliations

<sup>1</sup>Department of Respiratory and Sleep Medicine, Princess Alexandra Hospital, Woolloongabba, Queensland, Australia

<sup>2</sup>Faculty of Medicine, The University of Queensland, Saint Lucia, Queensland, Australia

<sup>3</sup>QCIF Facility for Advanced Bioinformatics, The University of Queensland, Saint Lucia, Queensland, Australia

<sup>4</sup>Melbourne Sleep Disorders Centre, East Melbourne, Victoria, Australia

**Twitter** Hailey Meaklim @SleepPsych\_Aus



**Acknowledgements** Sleep laboratory staff at St Vincent's Private Hospital, East Melbourne for their set up efforts. Telstra Corporation Ltd (Australia) for the provisions of the Jawbone UP3, ResMed (San Diego) for the ResMed S+ and Beddit Ltd (Finland) for the supply of the test devices used. The authors acknowledge the statistical support received through the Metro South Health Biostatistics Service.

**Contributors** CME was involved in the protocol preparation, participant consent, data collection, analysis, manuscript preparation and is the manuscript guarantor. SFZ was involved in the data curation and analysis and manuscript preparation. HM was involved in data analysis and manuscript preparation. RJ was involved with participant consent, data collection and manuscript preparation. DC and JS were involved in protocol preparation, data analysis and manuscript preparation.

**Funding** The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

**Competing interests** The Telstra Corporation Ltd (Australia) provided the Jawbone UP3 test devices used in the study, ResMed (San Diego) provided the ResMed S+ and Beddit Ltd (Finland) provided the Beddit device.

**Patient and public involvement** Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

**Patient consent for publication** Consent obtained directly from patient(s)

**Ethics approval** The study was approved by the Human Research and Ethics Committee of St Vincent's Hospital, Melbourne (LRR141/15).

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data availability statement** Data are available upon reasonable request. The dataset will be available upon emailed request to the corresponding author.

**Open access** This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

#### ORCID iDs

Claire M Ellender <http://orcid.org/0000-0002-1727-576X>

David Cunnington <http://orcid.org/0000-0002-8403-0420>

John Swieca <http://orcid.org/0000-0001-8281-4048>

#### REFERENCES

- Cai H, Shu X-O, Xiang Y-B, *et al*. Sleep duration and mortality: a prospective study of 113 138 middle-aged and elderly Chinese men and women. *Sleep* 2015;38:529–36.
- Davona T. The Wearables report: growth trends, consumer attitudes and why smart watches will dominate, business insider, February 12, 2015. *Business Insider Australia* 2015 <http://www.businessinsider.com.au/the-wearable-computing-market-report-bii-2015-7>
- de Zambotti M, Baker FC, Willoughby AR, *et al*. Measures of sleep and cardiac functioning during sleep using a multi-sensory commercially-available wristband in adolescents. *Physiol Behav* 2016;158:143–9.
- Intelligence C. US enterprise Wearables market: 5-year forecast, 2014–2019, 2015. Available: <http://www.marketwired.com/press-release/compass-intelligence-forecasts-wearables-enterprise-grow-exponentially-us-device-revenue-2032309.htm>
- de Zambotti M, Claudatos S, Inkelis S, *et al*. Evaluation of a consumer fitness-tracking device to assess sleep in adults. *Chronobiol Int* 2015;32:1024–8.
- Cook JD, Prairie ML, Plante DT. 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. *J Clin Sleep Med* 2018;14:841–8.
- Kang S-G, Kang JM, Ko K-P, *et al*. Validity of a commercial wearable sleep tracker in adult insomnia disorder patients and good sleepers. *J Psychosom Res* 2017;97:38–44.
- Tuominen J, Peltola K, Saaresranta T, *et al*. Sleep parameter assessment accuracy of a consumer home sleep monitoring Ballistocardiograph Beddit sleep Tracker: a validation study. *J Clin Sleep Med* 2019;15:483–7.
- De Chazal P, Fox N, O'Hare E, *et al*. Sleep/wake measurement using a non-contact biometric sensor. *J Sleep Res* 2011;20:356–66.
- Chinoy ED, Cuellar JA, Huwa KE, *et al*. Performance of seven consumer sleep-tracking devices compared with polysomnography. *Sleep* 2021;44. doi:10.1093/sleep/zsaa291. [Epub ahead of print: 14 05 2021].
- Zaffaroni A, de Chazal P, Heneghan C. SleepMinder: an innovative contact-free device for the estimation of the apnoea-hypopnoea index. *Conference proceedings : Annual International Conference of the IEEE Engineering in Medicine and Biology Society IEEE Engineering in Medicine and Biology Society Annual Conference 2009*, 2009:7091–4.
- Berry RB, Budhiraja R, Gottlieb DJ, *et al*. Rules for scoring respiratory events in sleep: update of the 2007 AASM manual for the scoring of sleep and associated events. deliberations of the sleep apnea definitions Task force of the American Academy of sleep medicine. *J Clin Sleep Med* 2012;8:597–619.
- AliphCom dba Jawbone. Jawbone UP3 San Francisco, 2016. Available: <https://jawbone.com/support/articles/000001027/download-the-app> [Accessed 19 Apr 2016].
- ResMed. San Diego, 2016. Available: <https://itunes.apple.com/us/app/s+-by-resmed/id883611019?mt=8> [Accessed 19 Apr 2016].
- Beddit Ltd. Beddit sleep Tracker Helsinki, Finland, 2016. Available: <http://support.beddit.com/hc/en-us/articles/201422237-Downloading-the-Beddit-app> [Accessed 19 Apr 2016].
- Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med* 2016;15:155–63.
- Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;1:307–10.
- Tomczak M, Tomczak E. The need to report effect size estimates revisited. An overview of some recommended measures of effect size. *Trends in Sport Sciences* 2014;21:19–25.
- de Zambotti M, Baker FC, Colrain IM. Validation of Sleep-Tracking technology compared with polysomnography in adolescents. *Sleep* 2015;38:1461–8.
- Scott H, Lack L, Lovato N. A systematic review of the accuracy of sleep wearable devices for estimating sleep onset. *Sleep Med Rev* 2020;49:101227.
- Kahawage P, Jumabhoy R, Hamill K, *et al*. Validity, potential clinical utility, and comparison of consumer and research-grade activity trackers in insomnia disorder I: In-lab validation against polysomnography. *J Sleep Res* 2020;29:e12931.
- Lee D. Apple releases new Beddit sleep tracker, 2018. Available: <https://www.theverge.com/2018/12/7/18131220/apple-beddit-3-5-sleep-monitor> [Accessed 22 Jul 2021].
- Dignan L. SleepScore max review 2017, 2021. Available: <https://www.zdnet.com/article/sleepscore-max-review-sleep-improvement-system-with-big-data-backing/> [Accessed 22 Jul 2021].
- Smith C. Rise and fall of the Jawbone UP24: the tracker that changed wearable Tech 2019, 2021. Available: <https://www.wearable.com/fitness-trackers/remembering-the-jawbone-up24-7320> [Accessed 22 Jul 2021].
- Khosla S, Deak MC, Gault D, *et al*. Consumer sleep technology: an American Academy of sleep medicine position statement. *J Clin Sleep Med* 2018;14:877–80.