Published in final edited form as:

Chronobiol Int. 2015 August ; 32(7): 1024–1028. doi:10.3109/07420528.2015.1054395.

Evaluation of a Consumer Fitness-Tracking Device to Assess Sleep in Adults:

Evaluation of Wearable Technology to Assess Sleep

Massimiliano de Zambotti, PhD[1], **Stephanie Claudatos, BS**[1], **Sarah Inkelis, BS**[1], **Ian M. Colrain, PhD**[1,2], and **Fiona C. Baker, PhD**[1,3,*]

[1]Center for Health Sciences, SRI International, Menlo Park, CA, USA

[2]Melbourne School of Psychological Sciences, University of Melbourne, VIC, Australia

[3]Brain Function Research Group, University of the Witwatersrand, Johannesburg, South Africa

Abstract

Wearable fitness-tracker devices are becoming increasingly available. We evaluated the agreement between Jawbone UP and polysomnography (PSG) in assessing sleep in a sample of twenty-eight midlife women. As shown previously for standard actigraphy, Jawbone UP had high sensitivity in detecting sleep (0.97) and low specificity in detecting wake (0.37). However, it showed good overall agreement with PSG with a maximum of two women falling outside Bland-Altman plot agreement limits. Jawbone UP overestimated PSG total sleep time (26.6±35.3min) and sleep onset latency (5.2±9.6min), and underestimated wake after sleep onset (31.2±32.3min) (p's<0.05), with greater discrepancies on nights with more disrupted sleep. The low-cost and wide-availability of these fitness-tracker devices may make them an attractive alternative to standard actigraphy in monitoring daily sleep-wake rhythms over several days.

Keywords

Motion; wristbands; actigraphy; sleep; activity trackers

Introduction

Smart “wearables” are integrated into our lifestyle, with real-time data linking our physiologic and electronic worlds by providing on-line and off-line feedback of our behavior. In particular, the use of fitness-tracking devices recently exploded and it has become normal to track one's daily activity (Lowe & Olaighin, 2014). Several companies quickly understood the importance of measuring sleep as an integrated component of wellness, and sleep patterns can now be easily and inexpensively tracked for extended periods.

[*]Corresponding author (FB): SRI International, 333 Ravenswood Avenue, Menlo Park, CA-94025; Tel.+1(650)859-3062; Fax: +1(650)859-2743; fiona.baker@sri.com.

Authors declare no conflicts of interest. This study was performed at the SRI International (Menlo Park, CA, USA) and was supported by National Institutes of Health, Grant HL103688 to FCB.

Fitness trackers mainly use the “actigraphic method”, in which an accelerometer detects motion, from which sleep patterns are inferred. Actigraphy has been established as a reliable alternative to the “gold standard” polysomnographic (PSG) method of assessing sleep, being comparatively inexpensive, nonintrusive, less time consuming for individuals and evaluators, and easily accessible. Standard actigraphs generally show high sensitivity (accuracy in detecting sleep) and low specificity (accuracy in detecting wake) with different levels of agreement with PSG mainly depending on the scoring algorithm and population observed (Ancoli-Israel et al., 2003; de Souza et al., 2003; Sadeh, 2011; Van de Water, Holmes, & Hurley, 2011).

Recently, studies have begun to investigate the validity of fitness-tracking consumer devices to assess sleep (de Zambotti, Baker, & Colrain, in press; Montgomery-Downs, Insana, & Bond, 2012). Fitbit® Flex™ and standard actigraphy both overestimated sleep efficiency (SE) and total sleep time (TST), showing high sensitivity for sleep and low specificity for wake, compared with PSG in 24 healthy adults (Montgomery-Downs et al., 2012).

We recently tested the validity of another commercialized wristband (Jawbone UP) against standard PSG in a group of 65 adolescents (de Zambotti et al., in press). Our results showed overall good agreement between methods, with Jawbone UP overestimating PSG TST by, on average, 10.0 min and underestimating wake after sleep onset (WASO) by, on average, 10.6 min.

We aimed here to assess the validity of Jawbone UP compared with PSG in assessing sleep in a sample of adult women on one night. We also investigated wake-dependent differences in agreement between methods in a subgroup of participants who had two overnight recordings, one with a greater amount of PSG-WASO than the other.

Materials and Methods

The study was approved by the Institutional Review Board at SRI International and participants gave written informed consent. Twenty-eight midlife women (age 50.1±3.9y, Body Mass Index 24.6±3.6kg.m^{-2}) participated. Women were screened as described in Sassoon et al. (2014). Based on clinical interview, twelve women met DSM-IV criteria for insomnia disorder. Based on clinical PSG assessment, two participants had periodic leg movement index (PLMI) >10, and two participants had PLMI >10 and apnea-hypopnea index >5.

All participants spent at least one night in the sleep lab at SRI International. Eighteen women had a second overnight recording. PSG and Jawbone UP data were simultaneously collected. PSG lights-out and lights-on times were self-selected by participants and Jawbone UP bands were synchronized accordingly.

Standard PSG was performed using Compumedics amplifiers and Profusion 3 software (Compumedics, Abbotsford, Victoria, Australia) and sleep stages (Wake, N1, N2, N3, and rapid-eye-movement (REM) sleep) were scored in 30-s epochs accordingly to American Academy of Sleep Medicine (AASM) rules (Iber, Ancoli-Israel, Chesson, & and Quan SF for the American Academy of Sleep Medicine, 2007). Time in bed (time from lights-out to lights-on, TIB, min), TST (min), sleep onset latency (SOL, time from lights-out to the first epoch of any sleep stage, min), and WASO (min) were calculated.

Jawbone® UP™ is a novel fitness-tracking wearable and easy-to-use device, suitable to wear 24h/day. Raw data are not readily accessible and information on the algorithm is not publicly available. In order to perform the epoch-by-epoch comparison between Jawbone UP and PSG data, we manually derived min-by-min wake and sleep periods from the Jawbone UP graphs through the Jawbone UP mobile app. PSG data were re-coded to match the Jawbone UP time resolution ("wake" was assigned if one or both of the PSG 30-s epochs were scored as wake; "sleep" was assigned when both 30-s epochs were scored as N1, N2, N3 or REM) (Sadeh, Sharkey, & Carskadon, 1994). Next, sensitivity (proportion of PSG sleep epochs scored as "sleep" by Jawbone UP) and specificity (proportion of PSG wake epochs scored as "wake" by Jawbone UP) were calculated (Ancoli-Israel, Cole et al., 2003).

To compare the overall sleep outcomes, we used standard PSG-equivalent outcomes provided by the Jawbone® UP™'s mobile App: "*In bed for*" (min; the equivalent of PSG-TIB), "*You slept*" (min; the equivalent of PSG-TST), "*Fell asleep*" (min; the equivalent of PSG-SOL). We also derived WASO (calculated as "*Awake for*" minus "*Fell asleep*", min; the equivalent of PSG-WASO).

Differences between Jawbone UP and PSG sleep outcomes were compared with Wilcoxon signed-rank tests. Agreement between methods for each variable was estimated using Bland and Altman plots of the difference between PSG and Jawbone measures against the mean of the two measures for each participant (1986). Mean (or Bias) and SD of the differences between Jawbone UP and PSG outcomes, lower and upper agreement limits (mean difference ±1.96SD) and 95%CI for mean differences and agreement limits are provided. Positive values of the mean difference between Jawbone UP and PSG sleep indicate that Jawbone UP underestimates PSG while negative values indicate that Jawbone UP overestimates PSG sleep. Mann–Whitney U tests were used to compare average discrepancies (PSG minus Jawbone UP) in the estimation of TST, WASO and SOL between women with and without an insomnia diagnosis. The Kolmogorov-Smirnov tests confirmed the normality of the distribution of the PSG-Jawbone UP differential values.

We used the subgroup of 18 women with two recordings to determine wake-dependent discrepancy between methods. Mean differences between PSG and Jawbone UP variables were compared between the night with higher PSG-WASO and the night with lower PSG-WASO for each woman using Wilcoxon signed-rank tests. Results are reported as mean ±SD. $P<0.05$ was considered significant.

Results

For the overnight recordings analyzed, women spent 442.4±47.4 min in bed. Jawbone UP compared with PSG showed higher values for TST (UP: 393.2±59.7 min; PSG: 366.6±61.2 min, $z=3.64$, $p<0.001$) and SOL (UP: 14.3±10.1 min; PSG: 9.1±6.9 min, $z=2.78$, $p=0.005$) and lower values for WASO (UP: 35.6±35.9 min; PSG: 66.8±38.8 min, $z=3.97$, $p<0.001$).

Epoch-by-epoch analysis indicated that Jawbone UP had high sensitivity (0.96) and low specificity (0.37).

As shown in Figure 1, Bland-Altman plots highlight that the PSG-Jawbone UP differences were more dispersed in participants with more disrupted sleep (lower TST and higher SOL and WASO). Jawbone UP overestimated PSG TST by 26.6±35.3 min (low, -95.9 min and high, 42.6 min agreement limits), overestimated PSG SOL by 5.2±9.6 min (low, -24.1 min and high, 13.7 min agreement limits), and underestimated PSG WASO by 31.2±32.3 min (low, -32.2 min and high, 94.5 min agreement limits). Analysis of all measures shows a maximum of two participants falling outside Bland-Altman plot agreement limits.

Discrepancies between Jawbone UP and PSG measures of TST, SOL, and WASO were similar in women with and without an insomnia diagnosis (all p's >0.05). Epoch-by-epoch analysis results were also similar in women with (sensitivity=0.96, specificity=0.41) and without insomnia (sensitivity=0.97, specificity=0.33).

In the sample of 18 women with two recordings, there was a greater discrepancy between Jawbone UP and PSG on the night when women had a higher amount of PSG-WASO (60.4±19.5min) compared to the night with a lower amount of PSG-WASO (31.6±18.1 min) for TST (higher PSG-WASO night: -33.5±17.9 min vs lower PSG-WASO night: -16.8±22.2 min, z=2.12, p=0.034) and WASO (higher PSG-WASO night: 35.4±16.6 min vs lower PSG-WASO night :15.4±20.4 min, z=2.59, p=0.010). There was no significant discrepancy between Jawbone UP and PSG in SOL measures between the higher PSG-WASO and lower PSG-WASO nights (-2.7±7.7 min vs 0.9±14.1 min, z=0.71).

Discussion

Bland-Altman plots showed that Jawbone UP had good agreement with PSG in the overall estimation of sleep in this female adult population with a maximum of two participants falling outside Bland-Altman plot agreement limits. With its poor ability to detect wake (low specificity), Jawbone UP showed less accuracy in the estimation of WASO, particularly on nights of more disrupted PSG sleep. Similarly to standard actigraphy (Ancoli-Israel et al., 2003; de Souza et al., 2003; Sadeh, 2011; Van de Water et al., 2011), Jawbone UP overestimated TST and SOL, while it underestimated WASO. These results are in agreement with our previous study assessing the validity of Jawbone UP compared with PSG in adolescents (de Zambotti et al., in press) and with the study of Montgomery-Downs et al. (2012) investigating the validity of another commercial device (Fitbit) in 24 adults.

The epoch-by-epoch comparison between Jawbone UP and PSG data confirms that similarly to standard actigraphy (Ancoli-Israel et al., 2003; de Souza et al., 2003; Marino et al., 2013; Sadeh, 2011; Van de Water et al., 2011) and to Fitbit® Flex™ (Montgomery-Downs et al., 2012), Jawbone UP had high sensitivity and low specificity, and therefore may be less accurate in evaluating sleep quality in people with fragmented sleep. Interestingly, in our study, two participants had PLMI >10, and two participants had PLMI >10 and apnea-hypopnea index >5 but none of them showed extreme values in the overall distribution of

Jawbone UP-PSG discrepancies. The use of Jawbone UP in populations with highly fragmented sleep (e.g. obstructive sleep apnea) needs to be further investigated.

Our sample included women with insomnia. We did not find differences in the accuracy of Jawbone UP in assessing PSG sleep between women with and without an insomnia diagnosis, similarly to previous research with standard actigraphy, which concluded that insomnia did not modify the association between actigraphy and PSG WASO (Marino et al., 2013). The accuracy of actigraphic-based sleep systems highly depends on the individuals' sleep architecture (e.g. the amount of wakefulness) (Paquet, Kawinska, & Carrier, 2007). Indeed, in our study, independently of the diagnosis of insomnia, the dispersion of the PSG-Jawbone UP differences was wider for women with low TST and high SOL and WASO. In addition, in a sub-group of women having two overnight recordings, there were greater discrepancies between Jawbone UP and PSG in the assessment of WASO and TST on the night with higher PSG-WASO compared to the night with lower PSG-WASO. The overall greater discrepancy between Jawbone UP and PSG in the present data from midlife women relative to that previously reported in adolescents (de Zambotti et al., in press) is probably related to the overall age-related difference in sleep quality between the two groups.

Several features of Jawbone UP, limit its utility in clinical and sleep research: 1) claims of being able to record sleep parameters (e.g. "*light sleep*", "*deep sleep*") other than acceptable standard motion-dependent sleep indices and the introduction of commercially attractive but scientifically poorly defined sleep outcomes (e.g. "*sleep goal*") create skepticism among sleep specialists; 2) raw data are not readily accessible, limiting the ability to easily evaluate sensitivity and specificity on an epoch by epoch basis, a gold standard analysis to evaluate the reliability of actigraphic methods against PSG; 3) algorithms are proprietary.

Despite these limitations, our data suggest that Jawbone UP provides acceptable levels of agreement with PSG measures, when the overall night is considered, and thus may be a feasible alternative for ecologically monitoring sleep-wake rhythms over several days, for example to track changes in sleep timing in large samples of adolescents or shift-workers. Recently, *a priori* defined clinically satisfactory ranges (e.g. TST ≤30 min) for the mean differences of the discrepancies between methods have been proposed (Meltzer, Walsh, Traylor, & Westin, 2012). Even if our data comply with these criteria, we suggest that the determination of the clinical usefulness of these actigraphic-based devices should rely more on the dispersion of the discrepancies (using for example the Bland-Altman lower and upper agreement limits as the reference measures) rather than on the average of the discrepancies. Given our laboratory findings of a wake-dependent change in the Jawbone UP-PSG discrepancy, the reliability of Jawbone UP should be tested in the home environment over several days when sleep is more likely to be variable and thus the Jawbone UP-PSG discrepancies could be exacerbated. Further validation of clinical utility will require study of larger groups of clinical populations.

Acknowledgments

We thank Justin Greco, David Sugarbaker, David Dresser, and Lena Kardos.

References

Ancoli-Israel S, Cole R, Alessi C, Chambers M, Moorcroft W, Pollak C. The role of actigraphy in the study of sleep and circadian rhythms. Sleep. 2003; 26(3):342–392. [PubMed: 12749557]

Bland J, Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986; 1(8476):307–310. [PubMed: 2868172]

de Souza L, Benedito-Silva A, Pires M, Poyares D, Tufik S, Calil H. Further validation of actigraphy for sleep studies. Sleep. 2003; 26(1):81–85. [PubMed: 12627737]

de Zambotti M, Baker F, Colrain I. Validation of sleep-tracking technology compared with polysomnography in adolescents. Sleep. (in press).

Iber, C.; Ancoli-Israel, S.; Chesson, A. Quan SF for the American Academy of Sleep Medicine. The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications. 1st. Westchester, Illinois: American Academy of Sleep Medicine; 2007. American Academy of Sleep Medicine

Lowe S, Olaighin G. Monitoring human health behaviour in one's living environment: A technological review. Med Eng Phys. 2014; 36(2):147–168. [PubMed: 24388101]

Marino M, Li Y, Rueschman MN, Winkelman J, Ellenbogen J, Solet J, … Buxton OM. Measuring sleep: accuracy, sensitivity, and specificity of wrist actigraphy compared to polysomnography. Sleep. 2013; 36(11):1747–1755. [PubMed: 24179309]

Meltzer L, Walsh C, Traylor J, Westin A. Direct comparison of two new actigraphs and polysomnography in children and adolescents. Sleep. 2012; 35(1):159–166. [PubMed: 22215930]

Montgomery-Downs H, Insana S, Bond J. Movement toward a novel activity monitoring device. Sleep Breath. 2012; 16(3):913–917. [PubMed: 21971963]

Paquet J, Kawinska A, Carrier J. Wake detection capacity of actigraphy during sleep. Sleep. 2007; 30(10):1362–1369. [PubMed: 17969470]

Sadeh A. The role and validity of actigraphy in sleep medicine: an update. Sleep Med Rev. 2011; 15(4):259–267. [PubMed: 21237680]

Sadeh A, Sharkey K, Carskadon M. Activity-based sleep-wake identification: an empirical test of methodological issues. Sleep. 1994; 17(3):201–207. [PubMed: 7939118]

Sassoon S, de Zambotti M, Colrain I, Baker F. Association between personality traits and DSM-IV diagnosis of insomnia in peri-and postmenopausal women. Menopause. 2014; 21(6):602–611. [PubMed: 24448105]

Van de Water A, Holmes A, Hurley D. Objective measurements of sleep for non-laboratory settings as alternatives to polysomnography-a systematic review. J Sleep Res. 2011; 20(1 Pt 2):183–200. [PubMed: 20374444]

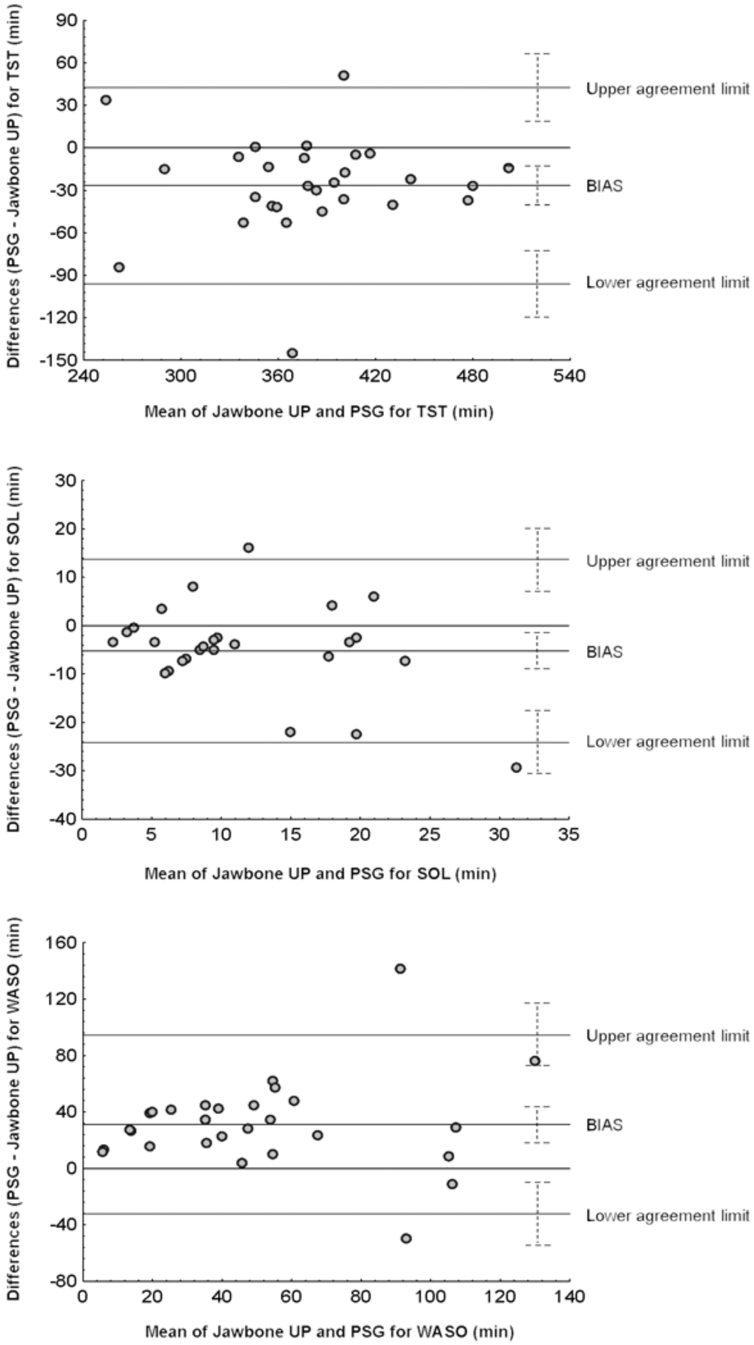

Figure 1.

Agreement (Bland-Altman plots) between Jawbone UP and polysomnography (PSG) for total sleep time (TST), sleep onset time (SOL) and wake after sleep onset (WASO) for 28 women who had a PSG recording. In the sample of 18 women with two recordings, only the first PSG night was used. Average, mean differences (or bias) between Jawbone UP and PSG outcomes, lower and upper agreement limits (mean difference ±1.96SD) and 95%CI for mean differences and agreement limits (dotted line) are displayed.