Published in final edited form as:
J Affect Disord. 2017 August 01; 217: 299–305. doi:10.1016/j.jad.2017.04.030.

Utility of the Fitbit Flex to Evaluate Sleep in Major Depressive Disorder: A comparison against polysomnography and wrist-worn actigraphy

Jesse D. Cook, **Michael L. Prairie**, and **David T. Plante**
Department of Psychiatry, School of Medicine and Public Health, University of Wisconsin-Madison, Madison, Wisconsin

Abstract

Background—Sleep disturbance is a common and important component of affective illness. Fitness activity trackers are emerging as alternative means to estimate sleep in psychiatric patients; however, their ability to quantify sleep in mood disorders has not been empirically evaluated. Thus, this study sought to evaluate the utility of the Fitbit Flex (FBF) to estimate sleep in patients with major depressive disorder (MDD) relative to gold standard polysomnography (PSG) and a widely-used actigraph (Actiwatch-2; AW-2).

Methods—Twenty-one patients with unipolar MDD wore the FBF and AW-2 during in-laboratory PSG. Bland-Altman analysis compared sleep variables among devices. Epoch-by-epoch analysis further evaluated sensitivity, specificity, and accuracy for the FBF and AW-2 relative to PSG.

Results—The FBF demonstrated significant limitations in quantifying sleep and wake, relative to PSG. In the normal setting, the FBF significantly overestimated sleep time and efficiency, and displayed poor ability to correctly identify wake epochs (i.e. low specificity). In the sensitive setting, the FBF significantly underestimated sleep time and efficiency relative to PSG. Performance characteristics of the FBF were more similar to the AW-2 in the normal compared to sensitive setting.

Limitations—Participants were young to middle aged and predominantly female, which may limit generalizability of findings. Study design also precluded ability to assess longitudinal performance of FBF.

Conclusions—The FBF is not an adequate substitute for PSG when quantifying sleep in MDD, and the settings of the device sizably impact its performance relative to PSG and other standard actigraphs. The limitations and capabilities of the FBF should be carefully considered prior to clinical and research implementation.

[*] Corresponding author. Jesse D. Cook, B.S., Wisconsin Psychiatric Institute and Clinics, 6001 Research Park Blvd., Madison, WI 53719, USA. Tel.: (608) 262 0130. jdcook4@wisc.edu.

Keywords

Fitbit; Depression; Sleep; Actigraphy; Polysomnography

1. Introduction

Sleep disturbance is very common in patients with Major Depressive Disorder (MDD). It has been estimated that up to 90% of individuals with MDD experience reduced sleep quality during a depressive episode (Tsuno et al., 2005). Depression can be accompanied by a diverse range of sleep disturbances including insomnia (difficulty falling asleep, maintaining sleep, or waking up too early) and/or hypersomnolence (excessive daytime sleepiness and/or sleep duration (Soehner et al., 2014). Objective changes in sleep continuity and duration in MDD, as measured by polysomnography, are robust physiological indicators of sleep disturbance in the disorder (Benca et al., 1992; Steiger and Kimura, 2010; Pillai et al., 2011; Plante et al., 2017). Sleep disturbance is also associated with treatment resistance, symptomatic relapse, suicidality, and impaired daytime function, underscoring its importance in the course of affective illness (Baglioni et al., 2011; McCall et al., 2010; Nadorff et al., 2013; Perlis et al., 1997; Riemann and Voderholzer, 2003; Szklo-Coxe et al., 2010). Thus, the ability to quantify sleep duration and continuity in patients with MDD is of potentially high value in the assessment and treatment of patients with mood disorders.

Polysomnography (PSG) is considered the gold standard for objective sleep measurement, however its widespread applicability is limited by its time-intensiveness, high cost, and intrusiveness (Meltzer et al., 2015; Montgomery-Downs et al., 2012). Furthermore, PSG is typically unable to provide information on longitudinal sleep-wake patterns over a multiple night assessment period (Meltzer et al., 2015). The use of validated actigraphs that utilize wrist-worn accelerometry to quantify movement as a surrogate measure for sleep and wake can circumvent some of these shortcomings of PSG due to their relatively low-cost, nonintrusiveness, and ambulatory capabilities (de Souza et al., 2003; Montgomery-Downs et al., 2012). Although actigraphy has been validated in its ability to identify sleep/wake times and patterns in adult populations (Morgenthaler et al., 2007) actigraphic devices may tend to overestimate total sleep time (de Souza et al., 2003; Montgomery-Downs et al., 2012). This deficiency stems largely from an inability to correctly identify wake epochs (Marino et al., 2013) as these devices rely upon an accelerometer (movement detector) as the sole measurement for sleep/wake designation, which inherently leaves them vulnerable to classifying periods of inactivity as sleep regardless of vigilance state. In addition, most validated actigraphs used in clinical and research settings generally require patients to return the device periodically (typically 2–4 weeks) in order for data to be retrieved for evaluation, making their use over more extended periods cumbersome.

The rise of commercially-available fitness activity trackers has provided another low-cost, field-based, and user-friendly alternative that may prove useful in evaluating sleep for both clinical and research purposes (de Zambotti et al., 2015; Meltzer et al., 2015; Montgomery-Downs et al., 2012). These mass-marketed devices are gaining a broader acceptance in both general and patient populations, and practitioners have begun to integrate their use in the assessment and treatment of affective disorders, despite limited research evaluating their use in patients with psychiatric disorders (Vahia and Sewell, 2016). Besides their low cost, these devices typically leverage direct-to-consumer cloud-based platforms and/or mobile technologies to allow for continuous data collection and retrieval over time. Considering the widespread availability of these devices and their potential impact on the management of psychiatric illness, comparison of their performance in estimating sleep against gold standard PSG and other validated actigraphs is a vital area of inquiry.

In a previous investigation conducted in healthy adults, a fitness activity tracker (the inaugural version of the Fitbit) overestimated total sleep time and congruently demonstrated an inability to correctly identify wake epochs when compared against a commonly used brand of actigraphy (Actiwatch 64) and PSG (Montgomery-Downs et al., 2012). Although not many validation studies have been performed on fitness activity trackers in the domain of sleep, this demonstration of an overestimation for total sleep time and inability to accurately identify wake epochs has been corroborated by studies on women with insomnia (de Zambotti et al., 2015) and adolescents referred for clinical PSG (Meltzer et al., 2015). Contrary to the results of these investigations, one study in healthy, young adults demonstrated comparable results in estimation of total sleep time for multiple fitness activity trackers, relative to PSG (Mantua et al., 2016). However, because epoch-by-epoch comparisons were not performed, the full performance characteristics of the fitness trackers utilized in this study, relative to PSG, could not be determined. The inconsistent results of existing validation studies - particularly in regards to the estimation of total sleep time and identification of wake periods - suggests a need to further investigate the true capabilities of these devices, with an emphasis on elucidating their utility within specific disorders. To our knowledge, no prior research has evaluated the validity of a commercially-available fitness activity tracker in persons with affective illness.

Thus, to further extend this line of inquiry into patients with affective illness, the primary aim of this investigation was to examine the utility of a commercially-available fitness activity tracking device, the Fitbit Flex (FBF, Fitbit Inc.; San Francisco, CA), against both PSG and validated actigraphy, the Actiwatch 2 (AW-2; Phillips Respironics), in a well-characterized cohort of adult patients with MDD.

2. Methods

2.1. Participants, Inclusion/Exclusion Criteria, and Study Design

A convenience sample of twenty-one, right-handed unmedicated patients with unipolar Major Depressive Disorder (MDD) was recruited as part of a larger study investigating electroencephalographic biomarkers of sleep disturbance in neuropsychiatric disorders. After an initial phone screening, participants completed an in-person medical, sleep, and psychiatric evaluation that included the Structured Clinical Interview for DSM-IV (SCID) (First et al., 2002), semi-structured sleep disorders evaluation, and physical exam, performed by a physician board certified in psychiatry and sleep medicine (DTP). Exclusion criteria included the following: smoking of greater than 15 cigarettes per day; >3 caffeinated beverages per day; significant sleep, neurologic, or medical disorder; history of significant head trauma or loss of consciousness > 30 minutes; and imminent risk of self-harm or suicide. Women who were pregnant, breastfeeding, <6 months post-partum, or planning to become pregnant during the study were also excluded. Participants were also excluded if they met DSM-IV criteria for alcohol or substance abuse/dependence within the preceding 6 months. Additionally, if patients met criteria for other Axis I psychiatric disorders, MDD had to be considered the primary disorder for study inclusion. Participants completed additional self-report instruments including the Beck Depression Inventory (BDI-II) (Beck et al., 1996), Pittsburgh Sleep Quality Index (PSQI) (Buysse et al., 1989), and Insomnia Severity Index (ISI) (Bastien et al., 2001). Eligible participants were then scheduled for an in-laboratory PSG at least one week but no more than one month after their in-person screening visit. All participants provided informed consent and were instructed to maintain their usual sleep-wake schedules for the duration of their time in the study. This study was approved by the Institutional Review Board of the University of Wisconsin-Madison.

2.2. In Laboratory Overnight Visit Procedures

Participants arrived at approximately 18:00 on the night of their PSG for set-up, at which point, a wrist-worn Actiwatch 2 (AW-2) and Fitbit Flex (FBF) were both placed adjacently on the participant's non-dominant (left) wrist. Polysomnographic data were collected using an integrated recording system that utilized a 256-channel EEG net (Electrical Geodesics, Eugene, OR) along with other standard recording sensors including electrooculogram (EOG), sub-mental electromyogram (EMG), electrocardiogram (ECG, bilateral tibial EMG, respiratory inductance plethysmography, pulse oximetry, and a positition sensor (Alice® Sleepware; Phillips Respironics, Murrysville, PA). A registered sleep technologist, blind to the FBF and AW-2 staging output, staged all sleep recordings using 30-second epochs according to standard criteria based on 6 EEG channels at approximate 10–20 locations (F3, F4, C3, C4, O1, and O2) referenced to the mastoids, electrooculogram, and sub-mental electromyogram according to American Academy of Sleep Medicine criteria (Berry et al., 2014). Bedtimes were tailored to each participant's habitual sleep pattern, with lights-off (participant actively trying to fall asleep) occurring between approximately 22:00 and 23:00. Participants were allowed to sleep *adlibitum*, remaining undisturbed throughout the night and not awoken at a prescribed time the following morning. Lights on was determined based on the participant's stated desire to terminate the nocturnal sleep period upon awakening. Polysomnography and accelerometer data were collected within a local network of computers time synchronized to an external atomic clock through frequent restart.

2.3. Data Analysis

PSG was considered the gold standard measure of sleep duration and continuity. PSG lights-off and lights-on times were used as the start and end points for the AW-2 and FBF rest periods to maintain consistency (Meltzer et al., 2015). The following sleep variables were calculated for PSG, FBF, and AW-2: total sleep time (TST; total duration of sleep during period of time in bed), sleep onset latency (SOL; time from lights-off to the first epoch of sleep), wake after sleep onset (WASO; total duration of wake time after sleep onset), and sleep efficiency (SE; equal to TST divided by total time in bed). AW-2 data were analyzed utilizing the medium threshold (value = 40) with five minute immobility time for sleep onset/offset since this setting has been shown to produce the most accurate output, relative to PSG (Chae et al., 2009). FBF data were analyzed using both the normal and sensitive settings, since prior studies in pediatric sleep apnea have suggested the significant effects of sensitivity settings for this device (Meltzer et al., 2015). The FBF was left in default (Normal) settings during the sleep recording; however, outputs for both Normal and Sensitive settings were available during off-line analyses for evaluation since the setting at the time of recording does not limit the off-line sensitivity outputs available. The FBF sleep variable data was extracted from the Fitbit web interface after the device synchronized with the interface through the Bluetooth capabilities of a USB-connected dongle. After device synchronization, the start/end points of sleep periods were manually adjusted to correspond to PSG lights-off and -on, and outputs from both Normal (FBF-N) and Sensitive (FBF-S) algorithms were obtained.

Bland-Altman analysis (Altman and Bland, 1983) was utilized to calculate the mean difference between devices for each comparison of interest (AW-2 vs. PSG; FBF-N vs. PSG; FBF-N vs. AW-2; FBF-S vs. PSG; and FBF-S vs. AW-2).

Further analyses explored the overall congruency between individually staged epochs among the devices. The PSG and AW-2 data were collected and staged in 30-second epochs. FBF epochs were extracted for both normal and sensitive settings using Fitabase (Small Steps Labs LLC, San Diego, CA). In order to compare the 60-second FBF epochs with PSG and AW-2, each FBF epoch was split into two equivalent 30-second epochs to correspond with the PSG and AW-2 30-second epoch values (Montgomery-Downs et al., 2012). Sensitivity (ability to correctly detect PSG-scored sleep epochs), specificity (ability to correctly detect PSG-scored wake epochs), and accuracy (ability to correctly detect PSG-scored sleep and wake epochs) were calculated for the AW-2, FBF-N, and FBF-S (Meltzer et al., 2015; Montgomery-Downs et al., 2012).

Epoch-by-epoch comparisons were conducted utilizing MATLAB (Mathworks; Natick, MA) with all other statistical analyses performed using JMP Pro 11 (SAS; Cary, NC). Alpha equaled 0.05 for statistical significance for all comparisons. Results are presented as ± standard deviation unless otherwise noted.

3. Results

3.1. Participants

The 21 participants consisted of 17 women and 4 men (mean age = 26.5 ± 4.6), who had mild to moderate MDD (mean BDI-II score = 22.9 ± 6.8). Participants also demonstrated moderate sleep disturbance as evidenced by their PSQI (mean score = 8.4 ± 2.5) and ISI (mean score = 14.3 ± 5.6) scores. Overall results including sleep variables quantified from each measure, mean differences resulting from Bland-Altman analyses, as well as sensitivity, specificity, and accuracy relative to PSG are summarized in Table 1.

3.2. AW-2 versus PSG

When the AW-2 was compared to gold standard PSG, AW-2 significantly overestimated TST (mean difference of 40.6 min, p=0.0004) and SE (mean difference of 7.0%, p=0.0003), while significantly underestimating SOL (mean difference of −13.5 min, p=0.012) and

WASO (mean difference of −27.1 min, p=0.005). Corresponding Bland-Altman plots are presented in Figure 1.

When compared epoch-by-epoch against PSG, the AW-2 displayed relatively good sensitivity (0.97 ± 0.02) and accuracy (0.87 ± 0.06), with poor specificity (0.31 ± 0.15).

3.3. FBF-N versus PSG and AW-2

When the FBF-N was compared to PSG, like the AW-2, FBF-N significantly overestimated TST (mean difference of 46.0 min, p<0.0001) and SE (mean difference of 8.1%, p<0.0001), while significantly underestimating WASO (mean difference of −44.0 min, p<.0001). However, SOL assessed by FBF-N and PSG were quite similar (mean difference of −2.0 min, p =0.72) (Figure 2A). When compared epoch-by-epoch against PSG, again like the AW-2, the FBF-N showed a high sensitivity (0.98 ± 0.02) and accuracy (0.88 ± 0.05), with low specificity (0.35 ± 0.13).

Direct comparison of the FBF-N to AW-2 demonstrated significantly higher estimates of SE (mean difference of 1.1%, p=0.042) and SOL (mean difference of 11.5 min, p=0.0003) for the FBF-N, as well as significantly lower estimates of WASO (mean difference of −16.9 min, p<0.0001) (Figure 3A). FBF-N and AW-2 had comparable estimates of TST (mean difference of 5.4 min, p=0.08) (Figure 3A).

3.4. FBF-S versus PSG and AW-2

When the FBF-S was compared to PSG, findings were quite different from those derived using the normal (non-sensitive) mode for the device. Relative to PSG, FBF-S significantly underestimated TST (mean difference of −86.3 min, p<0.0001) and SE (mean difference of −16.0%, p<0.0001), while significantly overestimating SOL (mean difference of 11.5 min, p=0.012) and WASO (mean difference of 74.8 min, p<0.0001) (Figure 2B). When compared epoch-by-epoch against PSG, the FBF-S displayed a modest sensitivity (0.78 ± 0.09), specificity (0.80 ± 0.17), and accuracy (0.78 ± 0.08).

Similarly, when the FBF-S was compared to the AW-2, FBF-S had significantly lower estimates of TST (mean difference of −126.8 min, p<.0001) and SE (mean difference of −22.9%, p<.0001), with significantly higher estimates of SOL (mean difference of 24.9 min, p=0.0006) and WASO (mean difference of 101.9 min, p<0.0001) (Figure 3B).

4. Discussion

To our knowledge, this is the first investigation to examine the utility of the Fitbit Flex as a sleep measurement device in adult patients with major depressive disorder, and as such may significantly impact both clinical care and research in affective illness. As it becomes more commonplace in clinical practice for patients to report data obtained from commercially available fitness activity trackers (Meltzer et al., 2015; Vahia and Sewell, 2016), elucidating the capabilities and shortcomings of these devices is increasingly important in the delivery of care. Furthermore, from a research perspective, clarifying the potential for these devices to accurately estimate sleep is a vital step in interpreting longitudinal, field-based assessments of sleep-wake patterns in large cohorts of patients with psychiatric disorders that may employ these devices. The results of our investigation suggest that the FBF has some utility in quantifying sleep and wake in patients with MDD; however, there are several limitations of the device, particularly when compared to gold standard polysomnography.

The primary results of our study indicate that the FBF, under the normal (non-sensitive) settings, when compared to PSG, significantly overestimates sleep duration and efficiency, while underestimating wake after sleep onset. Furthermore, the FBF-N displays an inability to accurately identify wake epochs relative to PSG staging (specificity = 0.35). Interestingly, the FBF, under normal setting, performed comparably to a validated actigraph, the AW-2, in assessing total sleep time. Both devices significantly overestimated total sleep time when compared against PSG, but did not significantly differ when evaluated against one another, with a difference of only approximately 5 minutes between devices. Similarities in performance between the FBF-N and AW-2 were further substantiated by epoch-by-epoch comparisons relative to PSG, where the FBF-N and AW-2 displayed similar sensitivity, specificity, and accuracy.

It is noteworthy that the specificity of the FBF-N and AW-2 observed in our investigation were relatively similar (0.35 and 0.31, respectively). These results differ from prior studies that utilized the inaugural version of the Fitbit and Actiwatch-64 and demonstrated poorer specificities for Fitbit compared to standard actigraphy (0.198 and 0.389, respectively) (Montgomery-Downs et al., 2012). Since other investigations have demonstrated similarly poor specificity for standard actigraphs relative to PSG, frequently below 0.5 (Blood et al., 1997; Sivertsen et al., 2006; de Souza et al., 2003; Paquet et al., 2007), our findings suggest the specificity of the FBF (used under the normal setting) may be improved relative to older models of the device, potentially due to changes in the device algorithm or the hardware itself. However, the interpretation and generalizability of these findings must be considered in the context of the patient population studied, as well as the use of *ad libitum* overnight recordings, which could theoretically increase the potential for misidentification of wake-epochs due to longer evaluation periods with increased wake after sleep onset. Regardless, these results call into question the necessity of using more costly standard actigraphs over this commercially-available fitness activity tracker for field-based estimates of total sleep time in patients with affective disorders.

Despite these similarities between FBF-N and AW-2, there were important differences in their abilities to assess sleep onset latency and wake after sleep onset. The FBF-N more closely approximated sleep onset latency in comparison to PSG, but the AW-2 significantly overestimated this variable. Estimates of wake after sleep onset for both devices were significantly lower than PSG, but AW-2 estimates were closer to PSG than FBF-N. These inter-device differences are noteworthy as they may impact device selection for clinical assessment or research depending on the primary sleep variable of interest. For instance, if a patient’s main sleep complaint is difficulty falling asleep, then the FBF-N would theoretically be more optimal than the AW-2 given its stronger performance in assessing sleep onset latency. Ultimately, our data suggest neither device can replace PSG as a measure of nocturnal sleep; however, both devices make similar estimates of sleep duration to each other, with more variable estimates of sleep onset and continuity.

Notably, relative performance similarities between FBF and AW-2 only apply to the FBF in the normal (non-sensitive setting). FBF-S had markedly different performance relative to PSG, with large underestimates of sleep duration and efficiency, as well as over estimates of SOL and WASO. Our results are similar to a recent investigation in children with sleep disordered breathing that demonstrated the sensitive setting significantly underestimated sleep duration and efficiency in comparison to PSG (Meltzer et al., 2015). Thus, based on our results, the sensitive setting of the FBF should not be utilized either in the clinical treatment of MDD or research studies in these patients, as estimates of sleep are not accurate, particularly relative to the normal setting.

There are limitations of our study that merit discussion. Participants were predominantly young to middle-aged, and thus our results may not extend to other age groups with MDD, such as pediatric or geriatric populations. Second, our findings may not generalize equally to both sexes, as participants in this study were predominantly female. Moreover, findings cannot be directly extended to other mood and/or sleep disorders as our study specifically examined outpatients with unipolar MDD that was generally mild to moderate in severity. Also, due to logistical and security issues, collection and analysis of analysis of accelerometer-based data was performed on a computer that was not identical to that used to record polysomnography, which despite our best efforts to synchronize the devices within the local network, may have introduced error into epoch by epoch comparisons. Finally, our results cannot be extended to other fitness trackers, or more current generations of the same model (i.e. Fitbit Flex 2), as these devices may have different performance characteristics.

The rapid evolution of fitness activity trackers presents a major complication for the use of these devices in clinical and scientific contexts, and creates a necessity for frequent reevaluation of the validity of the most current generations of these devices, as underlying software and algorithmic improvements may yield more accurate estimations of sleep and wake. A major hurdle towards accomplishing this goal will be obtaining sufficient funding that is free from commercial influence to conduct research that informs personal consumers, clinical practitioners, and scientific researchers about the capabilities of these devices. In addition, the validity of a given device to measure sleep and wake remains continuously in question if the algorithm used to define vigilance states is proprietary and can be altered without notice to consumers. Thus, before these devices are adopted for scientific research on a broad scale, these issues will need to be addressed to ensure adequate scientific rigor of results obtained, which may include open access to user data and algorithmic transparency.

Similar to prior validation studies comparing PSG to wrist-worn accelerometry to measure sleep, only a single night of PSG was utilized in analysis. This research design, while highly pragmatic and cost-effective, leaves in question the true capabilities of the FBF as a long-term, longitudinal sleep measurement device. Previous research studies utilizing fitness activity trackers as a longitudinal sleep assessment tool have encountered data acquisition loss stemming mostly from user implementation error (Baroni et al., 2015), which may be an unavoidable limitation of these devices. Since commercial fitness trackers may be particularly useful for field-based, epidemiological research studies, further research is warranted to determine the reliability of these devices assessed over multiple nights, and develop implementation methods that might improve reliability.

In summary, this study demonstrates that the Fitbit Flex cannot fully serve as a proxy for gold standard polysomnography in the quantification of sleep and wake in major depressive disorder. However, this commercially-available fitness tracker does demonstrate similar performance characteristics to a standard actigraph, particularly in the estimation of total sleep duration, when used in the normal (non-sensitive) mode. Clinicians and researchers should consider the capabilities, limitations, and settings of these devices when interpreting data and/or treating patients, as it will likely impact treatment decisions and interpretation of empiric results. To provide a more substantial evidence base for the application of fitness trackers in clinical and research settings will require ongoing investigation to evaluate the utility of an ever-expanding number of these devices across the spectrum of psychiatric illness.

Acknowledgments

Role of Funding Source: This work was supported by grants from the National Institute of Mental Health (K23MH099234), the Brain and Behavior Research Foundation, and the American Sleep Medicine Foundation to Dr. Plante. The sources of funding for this investigation had no further role in the study design, data collection, analysis and interpretation of the data, and the decision to submit the paper for publication. The study authors have no relationship with Fitbit, and Fitbit did not supply any funding, supplies, or guidance towards this investigation, or have any bearing on the decision to submit this manuscript for publication.

The authors thank Sydney Notermann, Lara Rotar, and Sam Boroumand for their assistance with data collection. We additionally thank the research participants for their time and effort, as well as the clinical and laboratory staff at Wisconsin Sleep Center for their assistance.

References

Altman DG, Bland JM. Measurement in Medicine: the Analysis of Method Comparison Studies. The Statistician. 1983; 32:307–317.

Baglioni C, Battagliese G, Feige B, Spiegelhalder K, Nissen C, Voderholzer U, Lombardo C, Riemann D. Insomnia as a predictor of depression: a meta-analytic evaluation of longitudinal epidemiological studies. J Affect Disord. 2011; 135:10–19. [PubMed: 21300408]

Baroni A, Bruzzese JM, Di Bartolo CA, Shatkin JP. Fitbit Flex: an unreliable device for longitudinal sleep measures in a non-clinical population. Sleep Breath. 2016; 20:853–854. [PubMed: 26449552]

Bastien CH, Vallières A, Morin CM. Validation of the Insomnia Severity Index as an outcome measure for insomnia research. Sleep Med. 2001; 2:297–307. [PubMed: 11438246]

Beck, AT., Steer, RA., Brown, GK. Manual for the Beck Depression Inventory-II. 2. San Antonio, TX: 1996.

Benca RM, Obermeyer WH, Thisted RA, Gillin JC. Sleep and psychiatric disorders. A meta-analysis. Arch Gen Psychiatry. 1992; 49:651–668. discussion 669–670. [PubMed: 1386215]

Berry, RB., Brooks, R., Gamaldo, CE., Harding, SM., Lloyd, RM., Marcus, CL., Vaughn, BV. The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology, and Technical Specifications, Version 2.1. Darien, Illinois: 2014.

Blood ML, Sack RL, Percy DC, Pen JC. A comparison of sleep detection by wrist actigraphy, behavioral response, and polysomnography. Sleep. 1997; 20:388–395. [PubMed: 9302721]

Buysse DJ, Reynolds CF, Monk TH, Berman SR, Kupfer DJ. The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. Psychiatry Res. 1989; 28:193–213. [PubMed: 2748771]

Chae KY, Kripke DF, Poceta JS, Shadan F, Jamil SM, Cronin JW, Kline LE. Evaluation of immobility time for sleep latency in actigraphy. Sleep Med. 2009; 10:621–625. [PubMed: 19103508]

de Souza L, Benedito-Silva AA, Pires ML, Poyares D, Tufik S, Calil HM. Further validation of actigraphy for sleep studies. Sleep. 2003; 26:81–85. [PubMed: 12627737] de Zambotti M, Claudatos S, Inkelis S, Colrain IM, Baker FC. Evaluation of a consumer fitness-tracking device to assess sleep in adults. Chronobiol Int. 2015; 32:1024–1028. [PubMed: 26158542]

Krouwer JS. Why Bland-Altman plots should use X, not (Y+X)/2 when X is a reference method. Stat Med. 2008; 27:778–780. [PubMed: 17907247]

First, M., Spitzer, R., Gibbon, M., Williams, J. Structured clinical interview for DSM-IV-TR axis I disorders, research version, patient edition. Biometrics Research, New York State Psychiatric Institute; New York: 2002.

Mantua J, Gravel N, Spencer RM. Reliability of Sleep Measures from Four Personal Health Monitoring Devices Compared to Research-Based Actigraphy and Polysomnography. Sensors (Basel). 2016:16.

Marino M, Li Y, Rueschman MN, Winkelman JW, Ellenbogen JM, Solet JM, Dulin H, Berkman LF, Buxton OM. Measuring sleep: accuracy, sensitivity, and specificity of wrist actigraphy compared to polysomnography. Sleep. 2013; 36:1747–1755. [PubMed: 24179309]

McCall WV, Blocker JN, D'Agostino R, Kimball J, Boggs N, Lasater B, Rosenquist PB. Insomnia severity is an indicator of suicidal ideation during a depression clinical trial. Sleep Med. 2010; 11:822–827. [PubMed: 20478741]

Meltzer LJ, Hiruma LS, Avis K, Montgomery-Downs H, Valentin J. Comparison of a Commercial Accelerometer with Polysomnography and Actigraphy in Children and Adolescents. Sleep. 2015; 38:1323–1330. [PubMed: 26118555]

Montgomery-Downs HE, Insana SP, Bond JA. Movement toward a novel activity monitoring device. Sleep Breath. 2012; 16:913–917. [PubMed: 21971963]

Morgenthaler T, Alessi C, Friedman L, Owens J, Kapur V, Boehlecke B, Brown T, Chesson A, Coleman J, Lee-Chiong T, Pancer J, Swick TJ. Committee So.P Medicine A.A.o.S. Practice parameters for the use of actigraphy in the assessment of sleep and sleep disorders: an update for 2007. Sleep. 2007; 30:519–529. [PubMed: 17520797]

Nadorff MR, Nazem S, Fiske A. Insomnia symptoms, nightmares, and suicide risk: duration of sleep disturbance matters. Suicide Life Threat Behav. 2013; 43:139–149. [PubMed: 23278677]

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

Perlis ML, Giles DE, Buysse DJ, Tu X, Kupfer DJ. Self-reported sleep disturbance as a prodromal symptom in recurrent depression. J Affect Disord. 1997; 42:209–212. [PubMed: 9105962]

Pillai V, Kalmbach DA, Ciesla JA. A meta-analysis of electroencephalographic sleep in depression: evidence for genetic biomarkers. Biol Psychiatry. 2011; 70:912–919. [PubMed: 21937023]

Plante DT, Cook JD, Goldstein MR. Objective Measures of Sleep Duration and Continuity in Major Depressive Disorder with Comorbid Hypersomnolence: A Primary Investigation with Contiguous Systematic Review and Meta-Analysis. J Sleep Res. 2017 In Press.

Riemann D, Voderholzer U. Primary insomnia: a risk factor to develop depression? J Affect Disord. 2003; 76:255–259. [PubMed: 12943956]

Sivertsen B, Omvik S, Havik OE, Pallesen S, Bjorvatn B, Nielsen GH, Straume S, Nordhus IH. A comparison of actigraphy and polysomnography in older adults treated for chronic primary insomnia. Sleep. 2006; 29:1353–1358. [PubMed: 17068990]

Soehner AM, Kaplan KA, Harvey AG. Prevalence and clinical correlates of co-occurring insomnia and hypersomnia symptoms in depression. J Affect Disord. 2014; 167:93–97. [PubMed: 24953480]

Steiger A, Kimura M. Wake and sleep EEG provide biomarkers in depression. J Psychiatr Res. 2010; 44:242–252. [PubMed: 19762038]

Szklo-Coxe M, Young T, Peppard PE, Finn LA, Benca RM. Prospective associations of insomnia markers and symptoms with depression. Am J Epidemiol. 2010; 171:709–720. [PubMed: 20167581]

Tsuno N, Besset A, Ritchie K. Sleep and depression. J Clin Psychiatry. 2005; 66:1254–1269. [PubMed: 16259539]

Vahia IV, Sewell DD. Late-Life Depression: A Role for Accelerometer Technology in Diagnosis and Management. Am J Psychiatry. 2016; 173:763–768. [PubMed: 27477136]

Highlights

- Activity monitors may prove useful as a sleep measurement tool in mood disorders
- The Fitbit Flex estimates sleep duration comparably to actigraphy in depression
- Fitbit Flex settings dramatically alter estimates of sleep continuity and duration

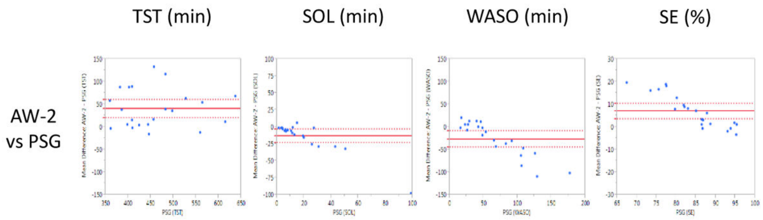

Figure 1.

Bland-Altman Plots presenting the mean difference values of the Actiwatch-2 (AW-2) and Polysomnography (PSG) on the Y-Axis against the PSG values on the X-Axis (Krouwer, 2008) across Total Sleep Time (TST; minutes), Sleep Onset Latency (SOL; minutes), Wake After Sleep Onset (WASO; minutes), and Sleep Efficiency (SE; percent). Horizontal, solid red line denotes the average mean difference with dotted lines representing 95% confidence interval.

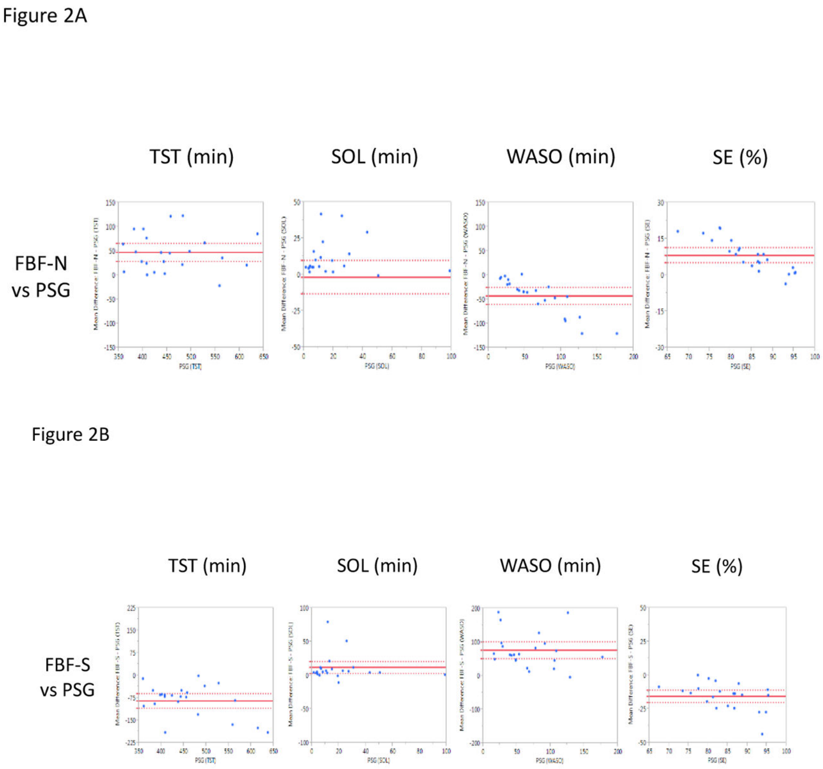

Figure 2.

Bland-Altman Plots presenting the mean difference values of the Fitbit Flex (FBF) and Polysomnography (PSG) on the Y-Axis against the PSG values on the X-Axis (Krouwer, 2008) across Total Sleep Time (TST; minutes), Sleep Onset Latency (SOL; minutes), Wake After Sleep Onset (WASO; minutes), and Sleep Efficiency (SE; percent). Horizontal, solid red line denotes the average mean difference with dotted lines representing 95% confidence interval. Figure A presents the normal setting (FBF-N) vs. PSG. Figure B presents the sensitive setting (FBF-S) vs. PSG.

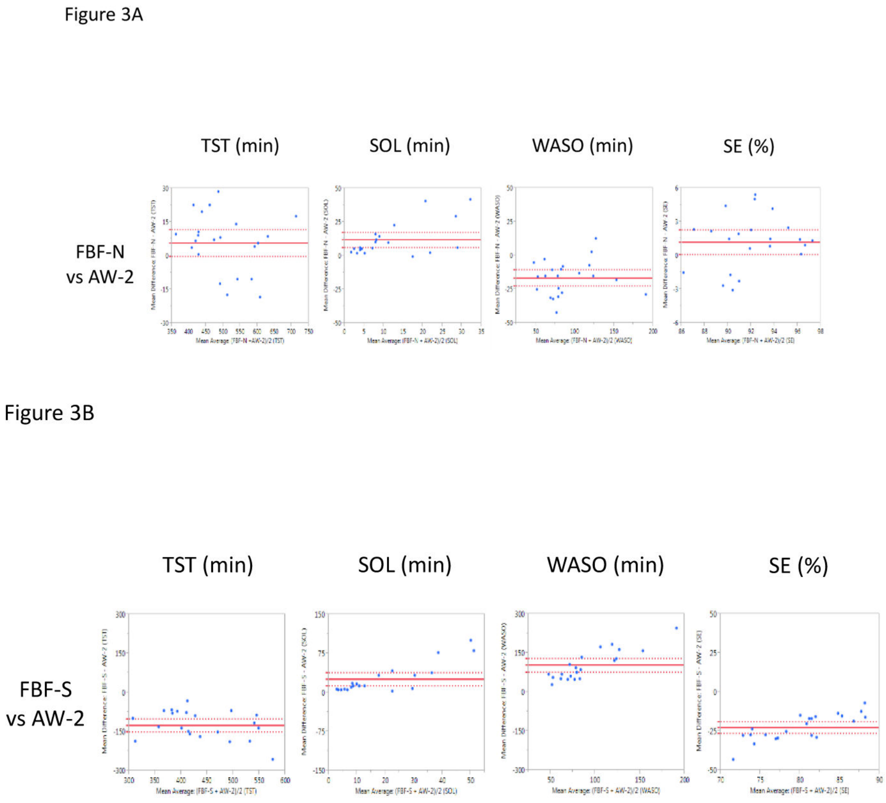

Figure 3.

Bland-Altman Plots presenting the mean difference values of the Fitbit Flex (FBF) and the Actiwatch-2 (AW-2) on the Y-Axis against the mean average values of the Fitbit Flex (FBF) and the Actiwatch-2 (AW-2) on the X-Axis (Krouwer, 2008) across Total Sleep Time (TST; minutes), Sleep Onset Latency (SOL; minutes), Wake After Sleep Onset (WASO; minutes), and Sleep Efficiency (SE; percent). Horizontal, solid red line denotes the average mean difference with dotted lines representing 95% confidence interval. Figure A presents the normal netting (FBF-N) vs. PSG. Figure B presents the sensitive setting (FBF-S) vs. PSG.

Table 1

Sleep Variables, Mean Differences, and Epoch-by-Epoch Comparisons

		Value (SD)		
Measure	TST (min)	SE (%)	SOL (min)	WASO (min)
PSG	457.8 (80.8)	84.4 (7.7)	19.2 (22.7)	68.3 (44.2)
AW-2	498.3 (20.3)	91.4 (3.1)	5.8 (7.7)	41.2 (17.0)
FBF-N	503.7 (89.6)	92.5 (3.5)	17.2 (14.2)	24.3 (19.1)
FBF-S	371.5 (72.0)	68.4 (8.8)	30.7 (28.6)	143.1 (62.6)
		Bland-Altman Mean Difference (p-value)		
Comparison	TST (min)	SE (%)	SOL (min)	WASO (min)
AW-2 v. PSG	40.6 (.0004)	7.0 (.0003)	−13.5 (.012)	−27.1 (.005)
FBF-N v. PSG	46.0 (<.0001)	8.1 (<.0001)	−2.0 (.72)	−44.0 (<.0001)
FBF-N v. AW-2	5.4 (.08)	1.1 (.042)	11.5 (.0003)	−16.9 (<.0001)
FBF-S v. PSG	−86.3 (<.0001)	−16.0 (<.0001)	11.5 (.012)	74.8 (<.0001)
FBF-S v. AW-2	−126.8 (<.0001)	−22.9 (<.0001)	24.9 (.0006)	101.9 (<.0001)
		Sensitivity, Specificity, and Accuracy (SD)		
Comparison	Sensitivity	Specificity		Accuracy
AW-2 v. PSG	0.97 (0.02)	0.31 (0.15)		0.87 (0.06)
FBF-N v. PSG	0.98 (0.02)	0.35 (0.13)		0.88 (0.05)
FBF-S v. PSG	0.78 (0.09)	0.80 (0.17)		0.78 (0.08)

Sleep variables of interest derived from Polysomnography (PSG), Actiwatch-2 (AW-2), Fitbit Flex Normal Setting (FBF-N), and Fitbit Flex Sensitive Setting (FBF-S). TST = Total Sleep Time (minutes); SE = Sleep Efficiency (percentage); SOL = Sleep Onset Latency (minutes); WASO = Wake After Sleep Onset (minutes). Values reported as mean ± standard deviation. Mean difference between comparisons derived from Bland-Altman analysis with corresponding p-value reported for each variable of interest. Sensitivity (ability to detect PSG-scored sleep epochs), Specificity (ability to detect PSG-scored wake epochs), and Accuracy (ability to detect PSG-scored sleep and wake epochs) mean values with standard deviation are reported for each device.