On the Reliability of Consumer Devices for the Assessment of Sleep

Manuel Schabus, Mohamed Ameen

I. Introduction

We spend around one third of our day sleeping, yet we still know little about the nature, the core function or most importantly the necessity of sleep for our well-being and life expectancy. This lack of knowledge has motivated scientists and researchers from various backgrounds to investigate and better understand the biological foundations of sleep. As a consequence, the past years have added significantly to the basic knowledge about sleep. Recent research has provided us with unprecedented knowledge about the structure of sleep, its influence on cognitive and motor abilities as well as ways to directly manipulate sleep experimentally.

What is still lacking however is the reliable measurement of sleep outside the laboratory, that is, ambulatory sleep assessment directly in the habitual sleeping environment; i.e. the participants home. Thus, in the past few years we have witnessed a surge on commercial sleep aids, sleep devices or mobile phone applications that aim to assess and ultimately improve sleep. In order that such devices or applications can provide credible and reliable information about the quality and structure of sleep, they must be tested in a robust scientific way and against the "gold-standard". That is, their performance must be described and assessed in terms of their agreement with the current gold-standard which is polysomnography (PSG) (i.e., EEG, EOG and EMG data). Only this can ensure the meaningfulness and validity of the new methods advertised by the industry.

Unfortunately, this vital step is skipped by literally all the consumer devices available on the market.

For this reason, we decided to start with the assessment of some of the common consumer devices available on the market which are claiming to monitor sleep and provide reliable information about sleep quality of the end-user night-by-night.

II. Methods

We recorded polysomnography data (PSG) from 18 healthy participants (13 Females, mean age: 29 ±13) using a 256-electrode GSN HydroCel Geodesic Sensor Net (Electrical 478 Geodesics Inc., Eugene, Oregon, USA) and a Net Amps 400 amplifier. Simultaneous electrocardiography, electromyography and electrooculography were recorded using bipolar electrodes. Sleep staging was performed using the sleep classification system developed by the SIESTA group (Somnolyzer 24×7; The SIESTA Group Schlafanalyse GmbH., Vienna, Austria) following the standard criteria described by the American Association for Sleep Medicine (AASM). The SIESTA output was considered our gold standard and compared to the consumer devices/application. Specifically we assessed sleep data from two devices and one application against the gold-standard: (1) a commercial activity tracker: the Mifit band v2 (Xiaomi, BJ, ROC); (2) a scientific actigraph (Actiwatch): Motionwatch 8 (CamNTech, CB, UK), and (3) a readily used mobile phone application for sleep assessment: Sleep Cycle v3.0.1.2511-release (Northcube, GOT, SE). For the later we used the option to estimate sleep via sounds not movement.

Four main sleep parameters were evaluated: (I) sleep onset latency (SOL), (II) sleep efficiency (SE), (III) total sleep time (TST), and (IV) wake after sleep onset (WASO). SOL, by definition, measures the difference between the time when the participant went to bed and the time when the participant actually fell asleep. SE was calculated as Sleep Efficiency = Total Sleep Time / Total Time in Bed. Measurements from all the devices were synchronized to the start of the PSG recording and correlations were computed non-parametrically using spearman correlations and means and standard deviations are reported. For Mifit measurements: we calculated SOL, SE, TST, while we report WASO values provided by the device, for Sleep Cycle measurements: we calculated SOL, SE, TST and WASO and for the actiwatch we report SOL, SE, TST and WASO values provided by the device. Values provided by the devices are reported in red while calculated values are reported in blue.

III. Results

A. Sleep Onset Latency (SOL)

No significant correlation was detected between SOL estimates from our gold standard sleep scoring and any of the three devices (*Mifit: r=0.24, p=0.42* (n=13); *Sleep Cycle: r= -0.41, p=0.23* (n=10); *Actiwatch: r= -0.09*, *p=0.77* (n=12). In addition mean values vary widely (M_{siesta}*: 30 ± 20.49,* M_{mifit}*: 15.53 ± 36.88,* $M_{Sleep\ cycle}$*: 36.4 ±21.64;* M_{Acti}*: 19.33 ± 21.95*).

B. Sleep Efficiency (SE)

A positive correlation was found between SE values derived from the Mifit band and that of our PSG gold standard (*r=0.57, p=0.04).* No correlation was revealed with the Actiwatch or the Sleep Cycle application (*r= 0.34, p=0.28 and r= -0.20, p=0.58, respectively*). In addition mean values vary again widely (M_{siesta}*: 82.40 ± 13.82,* M_{mifit}*: 97.97±3.74,* $M_{Sleep\ cycle}$*: 92.56±4.41;* M_{Acti}*: 88.13 ±5.56).*

Manuel Schabus is with the University of Salzburg, Department of Psychology, Laboratory for Sleep, Cognition and Consciousness, and with the Centre for Cognitive Neuroscience Salzburg. (email: manuel.schabus@sbg.ac.at)

Mohamed Ameen is with the University of Salzburg, Department of Psychology, Laboratory for Sleep, Cognition and Consciousness.

C. Total Sleep Time (TST)

No correlation was found between the total sleep time measured by the three devices/applications and our gold standard (*all p's >0.1*). In addition the mean values for estimated TST time again vary immensely (M_{siesta}: *378.5±95.2,* M_{mifit}*: 463.8 ± 56.7,* $M_{Sleep\ cycle}$*: 461.9 ±53.4;* M_{Acti}*: 417.20 ± 35.80).*

D. Wake After Sleep Onset (WASO)

Significant positive correlations between PSG estimates of WASO time in minutes and that of the Mifit band (*r=0.57, p=0.03*) as well as that of the actiwatch (*r=0.75, p=0.004*) were revealed. For the Sleep Cycle we only had 4 data sets available. WASO time estimates vary again widely between the evaluated devices and the PSG gold-standard (M_{siesta}*: 44.44 ± 38.66,* M_{mifit}*: 11.23 ± 23.60;* $M_{Sleep\ cycle}$*: 26.25 ± 7.50,* M_{Acti}*: 34.17 ± 20.70).*

IV. Discussion

Our preliminary results suggest that current consumer devices for sleep-monitoring are only able to deliver sleep data in a very inaccurate and general manner. Correlations between our PSG gold-standard derived data and these consumer devices were only found for sleep efficiency (MiFit) and wake after sleep onset time (MiFit and scientific actigraphy). However, also there the absolute values provided by consumer devices are far from the true gold-standard values. For example, for sleep efficiency (SE) these values vary between 82% (in PSG), to 93% (Sleep Cycle App) or even 98% in the MiFit device. The large amount of disagreement is crucial as a SE below 85% would be indicative of a real, clinically significant sleep problem, whereas a SE of 95% and above would on the other hand indicate sleep of very good quality. The low variance which is evident between nights for the consumer devices indicates that these devices and apps cannot to date capture the natural variation between individual nights which is present in real-world data.

For wake after sleep onset (WASO) we likewise found a wide range of sleep estimates with an average of 44 min WASO in our gold-standard (PSG) to 26 min in the Sleep Cycle App or even only an average of 11 min WASO in the MiFit device. This finding is however not surprising as devices mainly relying on activity measures are known to capture especially (transient) wake states poorly. In principal the MiFit device would in addition come with a photoplethysmogram (PPG) sensor and would have heart rate data available for its estimations, however, this does not seem to benefit the overall WASO estimation of this device to date. The important sleep measures "total sleep time" (TST) and "sleep onset latency" (SOL) of the consumer devices were especially poor in our analyses and revealed no relations to the PSG gold-standard; derived mean values for TST indicate that TST is strongly overestimated in the tested consumer devices and apps to date.

In our view the next necessary step to take would be to improve the accuracy of these devices and apps. Key measures of sleep (such as reported here) need to be compared to the PSG gold-standard, and ideally in big sleep laboratory studies with healthy young, as well as older poor sleeper controls in order to verify an acceptable accuracy also of consumer devices on the market.

Ultimately these devices should also allow reliable sleep scoring minute by minute sleep. To date the sleep staging output of these devices (although not statistically evaluated yet) seems rather arbitrary with highly implausible percentages of wake, light and deep sleep as well as a circadian variation of these states over the night which are literally impossible in human sleepers.

What these devices probably will not be able to capture in the near future is fine-grained sleep scoring and scoring accuracy as known from the PSG standard. For example, the differentiation in wake, transitional non-Rem sleep (N1), light non-Rem sleep (N2), deep non-Rem sleep (N3) and REM ("dreaming") sleep – although being standard in any sleep laboratory or polysomnography – is not yet implemented in any of the tested consumer devices. Likely this will also not be possible without a full polysomnography (PSG), that is without brain (electroencephalography), muscle (electromyogram) as well as eye (electrooculogram) activity from the sleeper.

Analysis-wise pending is the evaluation of the specificity and sensitivity of the sleep staging provided by these consumer which needs to be derived from epoch-wise agreement with the PSG gold-standard. That is, the agreement of the sleep staging information provided by the Mifit device and Sleep Cycle App with our PSG staging in 30sec or 1min epochs over the whole night. Scientifically it is still under investigation which degree of accuracy can actually maximally be derived if such consumer devices rely on movements and/or cardiac activity (Mifit), or simply sounds (or mattress vibrations) at bedside (Sleep Cycle).

In summary, we believe that current "sleep monitoring" consumer devices on the market must undergo a more robust validation process before being made available and distributed in the general public. This is especially noteworthy as there have been first reports in the literature that inaccurate feedback of such consumer devices can worry subjects and may even lead to compromised well-being of the user.