Accuracy of a smartphone application in estimating sleep in children

Pious Patel[1] · Ji Young Kim[2] · Lee J. Brooks[1,2]

Abstract

Purpose Chronic sleep problems can lead to difficulties for both the individual and society at large, making it important to effectively measure sleep. This study assessed the accuracy of an iPhone application (app) that could potentially be used as a simple, inexpensive means to measure sleep over an extended period of time in the home.

Methods Twenty-five subjects from the ages of 2–14 who were undergoing overnight polysomnography (PSG) were recruited. The phone was placed on the mattress, near their pillow, and recorded data simultaneously with the PSG. The data were then downloaded and certain parameters were compared between the app and PSG, including total sleep time, sleep latency, and time spent in various defined "stages."

Results Although there seemed to be a visual relationship between the graphs generated by the app and PSG, this was not confirmed on numerical analysis. There was no correlation between total sleep time or sleep latency between the app and PSG. Sleep latency from the PSG and latency to "deep sleep" from the app had a significant relationship ($p = 0.03$). No combination of PSG sleep stages corresponded with app "stages" in a meaningful way.

Conclusions The Sleep Cycle App may have value in increasing the user's awareness of sleep issues, but it is not yet accurate enough to be used as a clinical tool.

Work performed at Children's Hospital of Philadelphia.

✉ Lee J. Brooks
  brooksl@email.chop.edu

1 University of Pennsylvania, Philadelphia, PA 19104, USA

2 Division of Pediatric Pulmonary and Sleep Medicine, Children's Hospital of Philadelphia, 3401 Civic Center Blvd, Philadelphia, PA 19104, USA

Introduction

Sleep deprivation is one of the most common problems in children. Approximately 69 % of American adolescents report insufficient sleep [1]. Lack of sleep can cause daytime drowsiness, impair memory and attention, and influence other cognitive and emotional functions [2], which can result in poor academic performance [3]. In addition, sleep deprivation has been linked to other physiological problems, including obesity and diabetes, in children [4, 5]. Therefore, it is important to be able to measure sleep.

Studies in the sleep laboratory are very accurate but expensive, and only measure one night's sleep in an artificial environment. Sleep diaries are inexpensive, but depend entirely on an accurate report by the patient. Some actigraphy devices can accurately measure sleep in the home over a prolonged (1–2 weeks) period [6], but they are expensive and specialized equipment is needed to download and interpret the data. This creates a great need for a simple, inexpensive means to measure the quantity and quality of sleep in the home.

The use of smartphones in medicine has become increasingly common, as almost two-thirds of Americans own them [7], creating a vastly accessible platform through which patients can be reached. Applications (apps) have been created to keep track of Alzheimer's patients at risk of wandering away [8], to monitor blood pressure [9], to track diet and fitness, and to monitor patients' progress during rehabilitation [10]. Apps for measuring sleep have also been developed. Many of these require an external device connected to the smartphone, including a Bluetooth-linked external single channel ECG [11], a snoring detector [12], an application that combines snoring

data with sleep apnea education [13], several wristwatch devices [14], and a single channel EEG [15].

Integrating medical devices with smartphones is a promising idea, but most of the aforementioned apps still require patients to buy a separate medical device. Therefore, a smartphone application that accurately measures sleep without the aid of external devices would be useful. Apps like this have been developed and tested against actigraphy [16, 17], but only recently have they begun to be compared with results from polysomnography (PSG). One study found one of these apps to correlate poorly with PSG when it comes to basic sleep parameters and sleep staging [18]. Another app was found to be a weak differentiator of wake and sleep [19].

Maciek Drejak Labs (now Northcube AB) developed a smartphone app for the iPhone (Sleep Cycle App, (SCA)) that purports to assess sleep for people of all ages non-invasively by using the accelerometer of the phone to detect motion. Using proprietary software, it claims to be able to provide a report on the timing of different "stages" of sleep (classified into "wake", "sleep", and "deep sleep") by sensing the motion of the bed. A new feature has recently been added that claims to make use of the phone's microphone to estimate sleep. Our study focused solely on the motion-detecting function since this would compare more directly with medical-grade actigraphy, as well as consumer-oriented wrist devices. We assessed the accuracy of this app by comparing it with sleep staging determined simultaneously in children who are undergoing overnight PSG for clinical reasons.

Methods

Patients were recruited from the sleep laboratory at the Children's Hospital of Philadelphia. All had been referred for a clinical suspicion of sleep disorders, usually obstructive sleep apnea (OSA). This study was approved by the Institutional Review Board and informed consent was obtained from all individual participants included in the study. Anthropometric data, including height, weight, sex, and age were obtained in the sleep lab at the time of recruitment. At the start of each subject's sleep study, an iPhone 4S running SCA was placed on the bed at a consistent location near the subject's head, as recommended on the developer's website. The same device was used throughout the study.

The PSGs were scored by an experienced sleep technologist using standard pediatric scoring criteria and reviewed by a Board-certified sleep physician. The following parameters were recorded (Rembrandt, Embla, Broomfield, CO): electroencephalogram (EEG) (C3/A2, C4/A1, F3/A2, F4/A1, O1/A2, O2/A1), electrooculogram (EOG) (left and right), submental electromyogram (EMG), tibial EMG, modified lead 2 electrocardiogram, chest and abdominal wall motion by respiratory inductance plethysmography (SensorMedics, Yorba Linda, CA), airflow by nasal pressure (Pro-Tech Services, Inc, Mukilteo, WA) and three-pronged oronasal thermistor (Pro-Tech Services, Inc, Mukilteo, WA); end-tidal PCO_2 (E_TCO_2) by capnography (Novametrix 7000; Novametrix, Wallingford, CT), arterial oxyhemoglobin saturation (SpO_2; Novametrix 7000 or Masimo, Irvine, CA), oximeter pulse waveform, and digital video. Studies were scored utilizing the American Academy of Sleep Medicine standards [20, 21]. Sleep latency was calculated as the number of minutes from "lights out" to sleep onset. The apnea/hypopnea index (AHI) was calculated as the number of apneas + hypopneas per hour of sleep. The arousal index (AI) was calculated as the number of arousals + awakenings per hour of sleep.

The total time in each of the app's sleep state classifications, the latency to "sleep," and the latency to "deep sleep" were acquired using the graphs generated by the SCA. The lengths of the graph segments were converted into minutes by comparing them to the length of the entire graph, which represented a known total number of minutes. The investigators reviewing these data and performing these calculations had no prior knowledge of the PSG results.

The calculated data from the SCA were compared with analogous parameters from the PSG data. The SCA's latency to "sleep" and latency to "deep sleep" were compared with the PSG's sleep latency, and the SCA's total sleep time (TST) was compared with the PSG's TST. Bivariate linear regression and Bland-Altman plots were used for these comparisons. The concordance correlation coefficient for each PSG/SCA pair and its associated 95 % confidence interval were computed to assess the agreement between PSG and SCA. In addition, paired t test and Wilcoxon signed-rank test were performed to evaluate the difference in means and medians between PSG and SCA.

We hypothesized that certain stages of sleep might contribute more to the app designation than others. For example, we hypothesized that "deep sleep" might represent REM and/or N3 sleep. To assess the relationship between app staging and the PSG, stepwise regression was used with app "stage" as the dependent variable (i.e., minutes spent "awake," "sleep," and "deep sleep"), and PSG stages as independent variables (i.e. minutes spent in "Wake," "N1," "N2," N3," "REM," and all combinations of summing them together (e.g., "Wake + N1," "Wake + N1 + N2," etc.)).

Results

Twenty-five patients participated in the study, ranging from 2 to 14 years of age. Twenty-two had been referred for suspected OSA, and three were participating as volunteers for other studies. Subject demographics are displayed in Table 1, PSG parameters are shown in Table 2, and SCA parameters are shown in Table 3. Eight subjects had

Table 1 Anthropometric data of subjects

N	25 subjects
Age (years)	8.0+/−3.6
Height (cm)	130.6+/−18.5
Weight (kg)	37.8+/−25.2
Percent Male	56
BMI	20.4+/−7.3
BMI Z-Score	0.85+/−1.2

Statistics presented as mean +/− standard deviation

Table 3 SCA results

Latency to "sleep" (min.)	27.4+/−19.8
Latency to "deep sleep" (min.)	45.8+/−24.7
Total Sleep Time (min.)	406.2+/−100.7
% of sleep time in "awake"	19.8+/−14.2
"sleep"	53.6+/−13.7
"deep sleep"	26.6+/−10.3

Statistics presented as mean +/− standard deviation

AHI ≤ 1.5, nine had between 1.5 and 5, one had between 5 and 10, three had between 10 and 20, and four had AHI ≥20.

Most graphs from the app visually resembled the hypnograms from the PSG (Fig. 1), but some did not (Fig. 2). There was no significant relationship between the app and PSG measurements of TST (Fig. 3), sleep latency (Fig. 4), or sleep efficiency. The differences were random and without a systematic bias, as determined by Bland-Altman plots.

There was a significant relationship between sleep latency on the PSG and the latency to "deep sleep" on the app (Fig. 5; r-squared = 0.33, $p = 0.03$). There remained no relationship between TST and total "deep sleep" time on the app. There were no effects of AHI, weight, age, or AI on app accuracy.

When comparing sleep stages in PSG versus the app, there was a borderline ($p = 0.056$) relationship between "deep sleep" and minutes in "N1 + N3." No other matches between PSG sleep stages and app "stages" were found.

Discussion

Although the visual images suggest that this simple, consumer-oriented app may yield a gross approximation of the amount and quality of sleep, our hard numerical data do not support its use in a quantitative fashion. Only latency to "deep sleep" on the app showed a significant relationship with any PSG parameters (sleep latency).

Table 2 General sleep statistics from PSG

Sleep latency (min.)	35.9+/−22.3
Total sleep time (min.)	426.2+/−69.1
Sleep efficiency (%)	83.0+/−9.9 %
% of sleep time in N1	3.6+/−2.5
N2	42.9+/−8.0
N3	32.2+/−8.7
REM	21.35+/−5.9
Arousal Index	11.3+/−5.9
Apnea-Hypopnea Index	10.3+/−17.6

Statistics presented as mean +/− standard deviation

We considered other factors that might affect the accuracy of the app. As this was a highly selected population being investigated for OSA, we considered whether abnormal sleep architecture in the subjects could have affected its accuracy. For example, patients with OSA might have more movement/arousals than a healthy individual, which could cause deviation from the model used to develop the algorithm. Weight was considered, as heavier subjects would move the mattress with more intensity and movements might be more easily detected by the app. AI was also considered, as more arousals throughout the night without consequent movement could create discrepancies between the PSG and app wakefulness readings. However, none of these variables affected the discrepancy between app and PSG readings for the parameters analyzed.

There are other possible explanations for the lack of accuracy. The SCA measures movement. It is possible that the bed dampens movement enough to impair the accuracy of the app (we can only speculate on the effects of a waterbed!). The sensitivity of the accelerometer, at least, does not seem to be the problem, as we found the phone could detect even the most miniscule motions from the opposite edge of a normal mattress. The website for the Sleep Cycle App claims that it works on all types of beds, including water beds and memory foam mattresses. It even claims that it works when two people share a bed. We used only firm, hospital-style mattresses, and there was no co-sleeping so we were unable to evaluate this claim. It is possible that a wrist device might yield better results without the potentially dampening effects of the bed.

Even though there was a significant relationship between the app's latency to "deep sleep" and the PSG's sleep latency, the correlation between the two parameters was low. Because they did not have a 1:1 relationship, sleep latency cannot be predicted directly by the app's latency to "deep sleep," but rather as a fraction of this variable. This finding could be useful for developers who are building sleep measuring algorithms and suggests that better, perhaps less stringent criteria for "sleep" might improve the app's accuracy in measuring sleep latency.

Although there was little numerical agreement between the PSG and the app, the strong relationship between "deep sleep" and stages N1 + N3 is encouraging, since N3 sleep usually has
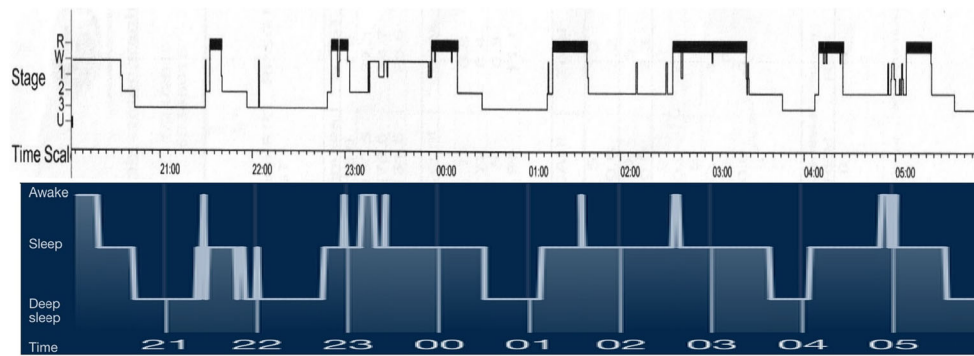
Fig. 1 Visual comparison between PSG hypnogram and SCA graph: example of good visual concordance. Progression of sleep stages throughout the night is shown, where top graph is PSG, with *y*-axis sleep stages of REM (R), wake (W), N1 (1), N2 (2), and N3 (3), and bottom graph is SCA, with *y*-axis sleep states of "awake," "sleep," and "deep sleep"

less movement than other stages, suggesting that with further refinement of the algorithm, and/or the accelerometer sensitivity, a closer relationship between the app and PSG may emerge.

This study has some limitations. Firstly, the sample size was only 25 subjects. Perhaps a larger sample might validate a relationship between "deep sleep" and N3. Our subjects were all children, varying in weight, height, and age. However, there was no effect of age or weight on accuracy of the app. Finally, the subjects recruited for this study were already scheduled for an overnight sleep study, meaning some of them were suspected for OSA or some other sleep disorder. Subjects with complicated conditions such as genetic or craniofacial disorders were excluded, and neither AHI nor AI was found to contribute to the variability of the app.

The app was unable to provide actual minutes spent in each reported sleep state. These data were gleaned by physically measuring the lengths of the graphs in an analog fashion to determine number of minutes. This may introduce an element of error caused by the extra step of human measurement.

Since the PSG assesses sleep stage in 30 s epochs, brief state changes and arousals were recorded, while the app graphs had

more smoothing (Fig. 1). A noticeable consequence of this was in the app's measurement of time spent awake. There were many moments where in the PSG, a subject had multiple arousals but spent most of the actual time asleep. The app registered these entire periods of time as "wake." Some REM periods were also registered as "wake" by the app because of arousals before and after the stage. This observation follows a similar trend among sleep measuring devices [22], including actigraph devices [23], of having low specificity in detecting wake from sleep.

Our use of the iPhone 4S as our study instrument was not ideal, as new generations of Apple and Android phones have since emerged. These new phones, which have more sensitive accelerometers and greater hardware capabilities, could perhaps allow for more accurate results from sleep measurement apps. We used the technology available at the time we started the study; Apple will always be able to make phones faster than the IRB can approve studies, so this will remain a problem.

Finally, it is important to emphasize that these results are only directly applicable to the Sleep Cycle App, as algorithms and design vary among the different sleep-measuring apps
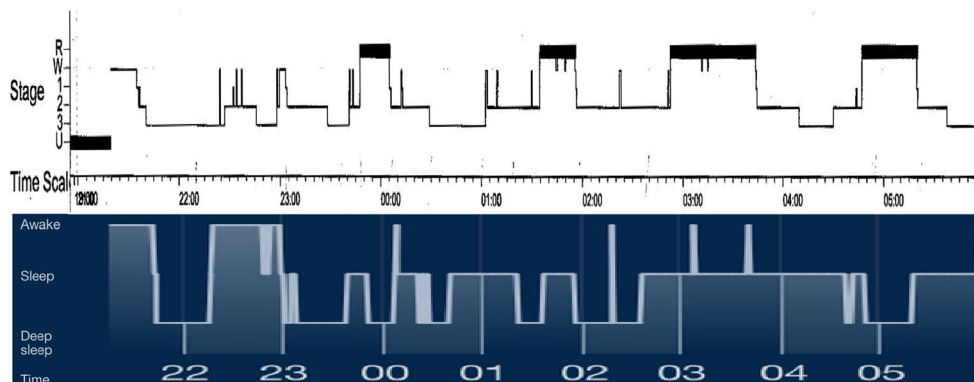

Fig. 2 Visual comparison between PSG hypnogram and SCA graph: example of poor visual concordance. Progression of sleep stages throughout the night is shown, where top graph is PSG, with *y*-axis sleep stages of REM (R), wake (W), N1 (1), N2 (2), and N3 (3), and bottom graph is SCA, with *y*-axis sleep states of "awake," "sleep," and "deep sleep"
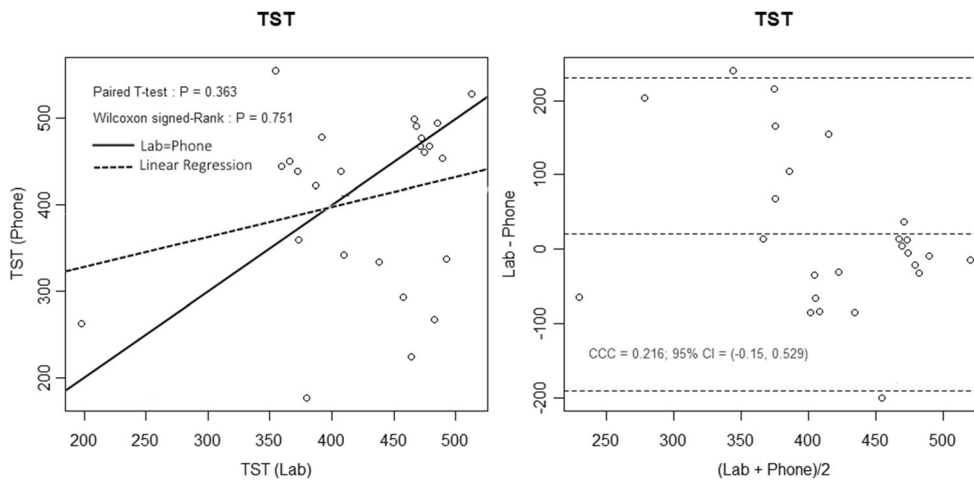
Fig. 3 Comparison of PSG (Lab) and SCA (Phone) measurement of total sleep time (TST). Each point represents a subject. **a** Linear regression is *dotted line*, line of identity (if SCA were to perfectly match PSG) is *solid line*. **b** Bland-Altman plot, concordance correlation coefficient (CCC), and 95 % confidence interval (95 % CI) values given

available on the market. However, this app was the most widely purchased at the start of the study.

This study is not meant to promote or defame a particular device. The use of these consumer sleep-measuring devices is a reality in the patient base that we are trying to investigate. In clinical settings, patients increasingly show up with data from these devices [24]. Our data suggest that they should not be used as diagnostic tools. Medical devices using similar principles [6] have been well studied and validated. Until consumer-oriented devices are shown to be accurate, physicians might unfairly dismiss their graphs or give them more legitimacy than deserved. Likewise, consumers need to know if these apps are usable for self-diagnosis, or if they should be treated as nothing more than entertainment.

Further research into the field of smartphone medical tools may include potential integration with other sleep applications, such as snoring detectors or utilization of other phone hardware such as the camera. Finding innovative ways to use them to measure sleep parameters and then integrating them into full sleep architecture could be promising, especially as smartphones become increasingly advanced. In addition, integration with smartwatches could potentially offer a more accurate measure of movement, as it would not be dependent on the movement of the bed.

There is evidence that the poor performance of consumer-oriented sleep measuring devices is mainly due to imperfect algorithms. While some actigraph devices estimate sleep with some degree of validity, their low specificity is a point of concern, particularly in pediatric samples [25]. Even though other consumer devices work similarly by measuring motion, they do not measure sleep with the same accuracy as medical-grade devices [22]. Even actigraphy tends to be more accurate
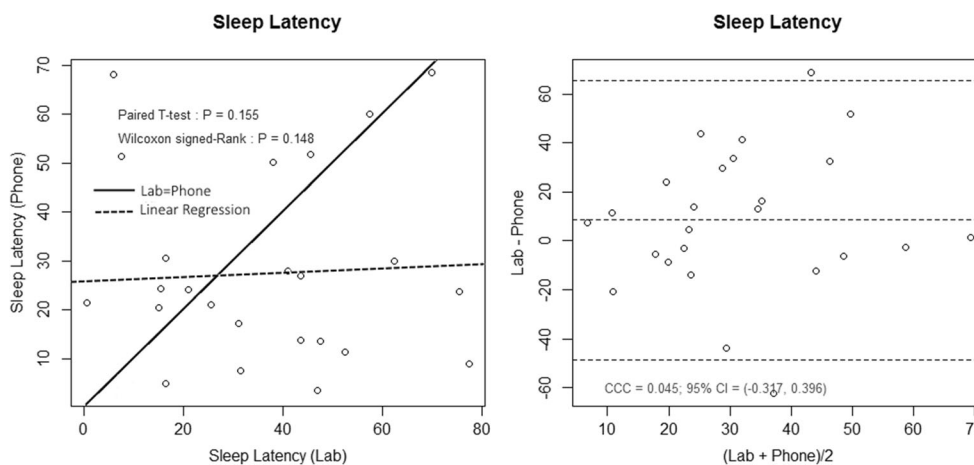

Fig. 4 Comparison of PSG (Lab) and SCA (Phone) measurement of sleep latency. **a** Linear regression is *dotted line*, line of identity is *solid line*. **b** Bland-Altman plot, concordance correlation coefficient (CCC) and 95 % confidence interval (95 % CI) values given

Fig. 5 Comparison of PSG (Lab) sleep latency and SCA (Phone) latency to "deep sleep." **a** Linear regression is *dotted line*, line of identity is *solid line*. **b** Bland-Altman plot, concordance correlation coefficient (CCC), and 95 % confidence interval (95 % CI) values given
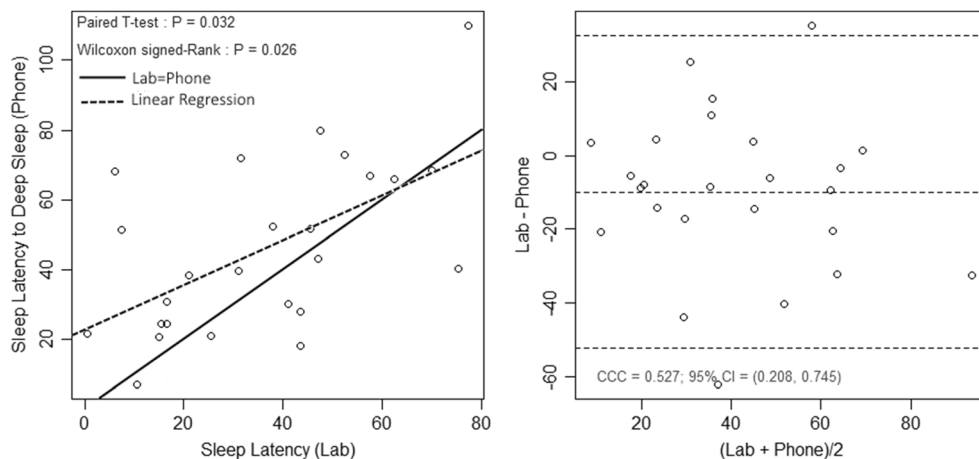
with human—rather than computerized—scoring [26]. Its algorithms seem to be improving over the long term, which suggests an optimistic potential for consumer sleep-measuring devices. In fact, one fitness tracker device was shown to have good concordance with PSG in adolescents [27].

Finally, even if the app may not be accurate, it may at least be useful in educating users about sleep. In today's technology-driven society, a flashy app might just be able to make the patient population more scientifically literate. As people get excited about measuring their own sleep, they will learn enough to have more meaningful conversations with physicians, increasing the quality of their healthcare. If the app can be improved, it has the potential to encourage users to fix poor sleep habits on their own, reducing unnecessary trips to the clinic. It could also be used to improve outcomes from behaviorally based interventions, which are already highly effective [28], or as a critical detection tool for early sleep problems such as behavioral insomnia of childhood [29], as it could allow physicians to monitor their patients' long-term sleep patterns more reliably than with sleep diaries or self-reports. From keeping track of bedtimes to checking for abnormal awakenings or sleep quality, the app could provide the physician with better data while also eliminating a chore for the patient.

``We conclude that the Sleep Cycle App is not currently useful as a tool for sleep estimation. The difficulty is likely with the computation of the raw data picked up by the app. Some of these devices are marketed as "alarm clocks" that claim to awaken the user during a "light stage" of sleep. This study calls into question the ability of these "alarm clocks" to act as advertised. Our hope is that the results of this study can give developers and scientists new information to refine their models and that someday a cheap, accurate measurement for sleep is publicly available.

References

1. McKnight-Eily LR, Eaton DK, Lowry R, Croft JB, Presley-Cantrell L, Perry GS (2011) Relationships between hours of sleep and health-risk behaviors in US adolescent students. Prev Med 53(4): 271–273

2. Vriend J, Davidson F, Rusak B, Corkum P (2015) Emotional and cognitive impact of sleep restriction in children. Sleep Med Clin 10(2):107–115

3. Curcio G, Ferrara M, De Gennaro L (2006) Sleep loss, learning capacity and academic performance. Sleep Med Rev 10(5):323–337

4. Owens J, Au R, Carskadon M et al (2014) Insufficient sleep in adolescents and young adults: an update on causes and consequences. Pediatrics 134(3):e921–e932

5. Koren D, Katz LEL, Brar PC, Gallagher PR, Berkowitz RI, Brooks LJ (2011) Sleep architecture and glucose and insulin homeostasis in obese adolescents. Diabetes Care 34:1–6

6. Marino M, Li Y, Rueschman MN et al (2013) Measuring sleep: accuracy, sensitivity, and specificity of wrist actigraphy compared to polysomnography. Sleep 36(11):1747–1755

7. Smith A. US smartphone use in 2015. (2015) Pew Res Center; 18-29

8. Sposaro F, Danielson J, Tyson G (2010) iWander: an android application for dementia patients. EMBC, Annual International Conference of the IEEE 3875-3878

9. Lin YH, Lin ZM, Hsu YT, Ku HY. (2011) A mobile intelligent blood pressure monitor based on the android smartphone. InSecurity-Enriched Urban Computing and Smart Grid 178-87

10. Ozdalga E, Ozdalga A, Ahuja N (2012) The smartphone in medicine: a review of current and potential use among physicians and students. J Med Internet Res 14(5):e128

11. Bsoul M, Minn H, Tamil L (2011) Apnea MedAssist: real-time sleep apnea monitor using single-lead ECG. Inf Technol Biomed, IEEE Trans on 15(3):416–427

12. Shin H, Cho J (2014) Unconstrained snoring detection using a smartphone during ordinary sleep. Biomed Eng Online 13(1):116

13. Mosa AS, Yoo I, Sheets L (2012) A systematic review of healthcare applications for smartphones. BMC Med Inform Decis Mak 12(1):1

14. Kelly JM, Strecker RE, Bianchi MT (2012) Recent developments in home sleep-monitoring devices. ISRN Neurol 2012:768794

15. Bianchi MT (2015) Consumer sleep apps: when it comes to the big picture, it's all about the frame. J Clin Sleep Med 11(7):695–696

16. Chen Z, Lin M, Chen F et al. (2013) Unobtrusive sleep monitoring using smartphones. Pervasive Computing Technologies for Healthcare, 7th International Conference 145-52

17. Natale V, Drejak M, Erbacci A, Tonetti L, Fabbri M, Martoni M (2012) Monitoring sleep with a smartphone accelerometer. Sleep Biol Rhythms 10:287–292

18. Bhat S, Ferraris A, Gupta D et al (2015) Is there a clinical role for smartphone sleep apps? Comparison of sleep cycle detection by a smartphone application to polysomnography. J Clin Sleep Med 11(7):709–715

19. Toon E, Davey MJ, Hollis SL, Nixon GM, Horne RS, Biggs SN (2015) Comparison of commercial wrist-based and smartphone accelerometers, actigraphy, and PSG in a clinical cohort of children and adolescents. J Clin Sleep Med 12(3):343–350

20. Iber C, Ancoli-Israel S, Chesson AL, Quan SF. (2007) The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications. Am Acad Sleep Med

21. Berry RB, Budhiraja R, Gottlieb DJ et al (2012) 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 8(5):597–619

22. Meltzer LJ, Hiruma LS, Avis K, Montgomery-Downs H, Valentin J (2014) Comparison of a Commercial Accelerometer with Polysomnography and Actigraphy in Children and Adolescents. Sleep 38(8):1323–1330

23. Meltzer LJ, Walsh CM, Traylor J, Westin AM (2012) Direct comparison of two new actigraphs and polysomnography in children and adolescents. Sleep 35(1):159–166

24. de Zambotti M, Godino JG, Baker FC, Cheung J, Patrick K, Colrain IM (2016) The boom in wearable technology: cause for alarm or just what is needed to better understand sleep? Sleep 39(9):1761–1762

25. Meltzer LJ, Montgomery-Downs HE, Insana SP, Walsh CM (2012) Use of actigraphy for assessment in pediatric sleep research. Sleep Med Rev 16(5):463–475

26. Boyne K, Sherry DD, Gallagher PR, Olsen M, Brooks LJ (2013) Accuracy of computer algorithms and the human eye in scoring actigraphy. Sleep Breath 17(1):411–417

27. de Zambotti M, Baker FC, Colrain IM (2014) Validation of sleep-tracking technology compared with polysomnography in adolescents. Sleep 38(9):1461–1468

28. Mindell JA, Kuhn B, Lewin DS, Meltzer LJ, Sadeh A (2006) Behavioral treatment of bedtime problems and night wakings in infants and young children. Sleep 29(10):1263–1276

29. Meltzer LJ (2010) Clinical management of behavioral insomnia of childhood: treatment of bedtime problems and night wakings in young children. Behav Sleep Med 8(3):172–189