Performance Evaluation of the Circadia Contactless Breathing Monitor and Sleep Analysis Algorithm for Sleep Stage Classification

Timo Lauteslager, PhD[1], Stylianos Kampakis, PhD[1], Adrian J. Williams, FRCP FAASM[1,2], Michal Maslik[1], and Fares Siddiqui[1]

Abstract— **Although polysomnography (PSG) remains the gold standard for studying sleep in the lab, the development of wearable and 'nearable' non-EEG based sleep monitors has the potential to make long-term sleep monitoring in a home environment possible. However, validation of these novel technologies against PSG is required. The current study aims to evaluate the sleep staging performance of the radar-based Circadia Contactless Breathing Monitor (model C100) and proprietary Sleep Analysis Algorithm, both in a home and sleep lab environment, on cohorts of healthy sleepers. The C100 device was initially used to record 17 nights of sleep data from 9 participants alongside PSG, with a subsequent 24 nights of PSG data for validation purposes. Respiration and body movement features were extracted from sensor data, and a machine learning algorithm was developed to perform sleep stage prediction. The algorithm was trained using PSG data obtained in the initial dataset (n=17), and validated using leave-one-subject-out cross-validation. An epoch-by-epoch recall (true positive rate) of 75.0 %, 59.9 %, 74.8 % and 57.1 %, was found for 'Deep', 'Light', 'REM' and 'Wake' respectively. Highly similar results were obtained in the independent validation dataset (n=24), indicating robustness of results and generalizability of the sleep staging model, at least in the healthy population. The device was found to outperform both a consumer and medical grade wrist-worn monitoring device (Fitbit Alta HR and Philips Respironics Actiwatch) on sleep metric estimation accuracy. These results indicate that the developed non-contact monitor forms a viable alternative to existing clinically used wrist-worn methods, and that longitudinal monitoring of sleep stages in a home environment becomes feasible.**

I. Introduction

Sleep tracking, both through wrist-worn actigraphy and polysomnography (PSG), is extensively used in sleep medicine to diagnose sleep disorders and evaluate treatment outcomes. Actigraphy data are used for clinical evaluation of insomnia, circadian rhythm disorders and to exclude insufficient sleep as a possible cause for complaints [1]. PSG data are used to diagnose or exclude sleep-disordered breathing and periodic limb movement disorder. In addition, the portion of the night spent in different sleep stages and their timing are generally accepted to be of clinical significance: Adequate non-rapid eye movement (NREM) sleep is considered essential for general restoration, to avoid hypertension, diabetes, weight gain, and cognitive problems. The rapid-eye-movement (REM) sleep stage is important for memory, and its timing is a marker for depression, circadian phase, and neurological disorders such as narcolepsy [2].

Although PSG remains the gold standard for sleep assessment in a lab setting, PSG studies are inconvenient and uncomfortable to the patient. In addition, the high cost is prohibitive for long-term data collection. Various non-EEG consumer sleep tracking devices exist, which use biosignals such as respiration, heart rate and levels of movement to distinguish between discrete sleep stages. Wearable devices, including clinical grade actigraphy, struggle to accurately determine time-in-bed and sleep efficiency, which are both of clinical importance for patients suffering from insomnia [3]. As opposed to wearables, non-contact devices (such as [4]–[7]) are typically able to measure bed occupancy, and offer superior patient comfort and compliance.

The Circadia Contactless Breathing Monitor (model C100) is a novel 'nearable' device that uses pulsed ultra-wideband radar for contactless monitoring of respiration and body motion. The Circadia Sleep Analysis Algorithm predicts bed occupancy, sleep stages, and derives standardized sleep metrics. The device offers the advantage of range gating through a pulsed radar architecture, with sufficient down-range resolution to distinguish between the user and a potential bed partner. Unlike mattress-based sensors, a radar sensor is able to measure overall bedroom presence, and radar-based fall detection has been shown to be possible [8].

The purpose of the current study is to evaluate the sleep staging performance of the novel C100 device and algorithms against the gold standard PSG. Epoch-by-epoch sleep staging performance will be assessed, as well as prediction accuracy of derived sleep metrics. In addition, sleep metric prediction accuracy will be compared to accuracy of two commonly used wrist-worn devices, through a direct comparison.

II. Methods

A. Initial data collection

The C100 device was used to record 17 nights of sleep data from nine healthy participants (three females; mean age 25.3, SD ±1.73). Alongside the device, PSG data were recorded (Somté PSG, Compumedics). To allow for direct comparison, participants were asked to wear both a consumer grade and a clinical grade tracking device: The Fitbit Alta HR and the Philips Respironics Actiwatch Spectrum Plus. Recordings were performed at the participant's home, participants were free to choose their bed and wake time. Preparation of

This work was funded by Circadia Technologies Ltd.

[1]Timo Lauteslager, Stylianos Kampakis, Adrian J. Williams, Michal Maslik and Fares Siddiqui are with Circadia Technologies Ltd, 6 Delmey Close, London, CR0 5QD, United Kingdom. Email: {timo, stylianos, adrian, michal, fares}@circadia.health

[2]Adrian J. Williams is Professor of Sleep Medicine, King's College London, United Kingdom, and with the Sleep Disorders Centre, Guy's and St Thomas' NHS Foundation Trust, London, United Kingdom. Email: adrian.williams6@nhs.net

participants, as well as PSG data scoring was performed by a RPSGT-certified sleep technician, according to standards set forth by the American Academy of Sleep Medicine (AASM): Each 30-second epoch was scored as either 'Wake', 'N1', 'N2', 'N3', or 'REM'. For the purpose of this study, stages N1 and N2 were grouped together into 'Light sleep'. N3 was relabelled as 'Deep sleep'. The initial data collection was performed on friends and family of the investigators, and no IRB approval was obtained. The data collection protocol was designed and executed in accordance with good clinical practice guidelines, all participants provided written informed consent, and all devices used were CE compliant.

B. Data processing and analysis

Respiration and body movement features were obtained through the C100 device using embedded digital signal processing algorithms, and processed by the proprietary Sleep Analysis Algorithm. An ensemble of machine learning algorithms was trained on PSG data, and evaluated using leave-one-subject-out cross-validation. One of four sleep stages (Wake, REM, Light, Deep) was predicted for each 30 s epoch of sensor data. An epoch-by-epoch comparison of predicted sleep stages and PSG data was made. Precision and recall were used to evaluate model performance. To evaluate the performance of sleep metric estimation (such as 'total sleep time', 'sleep efficiency' and 'sleep latency', all derived according to AASM standards), mean absolute percentage error (MAPE) was used. Prediction accuracy was defined as 100-MAPE %. Sleep metrics were also obtained from two wrist-worn devices. Philips Actiwatch data were analyzed in Actiware software, using the 'medium' wake threshold and default settings. PSG recording onset and offset were used to define an analysis window for the Actiwatch data. Fitbit sleep metrics were exported from the user dashboard, after the sleep window was manually updated to correspond to the PSG data collection window. Sleep metric 'sleep onset latency' is not provided by Fitbit, but was manually obtained from hypnogram data.

C. Independent validation

Despite the use of cross-validation for model training, a risk of overfitting exists when using machine learning techniques for sleep stage prediction. To truly validate performance, the system was independently tested by a sleep and memory research group based at University of Fribourg, Switzerland. Sleep data were collected from 24 participants (healthy sleepers, 18 females, mean age 23.1, SD ±3.42) in a sleep lab using PSG (Brain Products, Germany), and the C100 device. The study was reviewed and approved by the institutional ethics committee, and written informed consent was obtained from all participants. PSG data were scored according to AASM standards. Obtained datasets were only used for performance evaluation and were not used for model training. The validation dataset was thus completely independent from the training dataset. Sleep staging performance was assessed through epoch-by-epoch comparison with PSG data.

III. Results

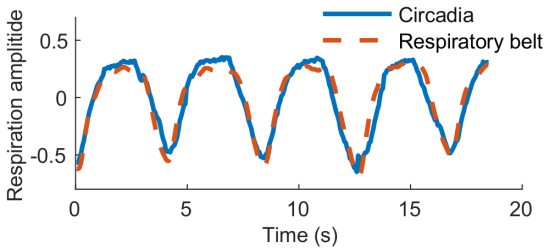

Fig. 1. A comparison of respiratory chest excursion measured using the C100 device (blue line) and a PSG respiratory belt (red dashed line).

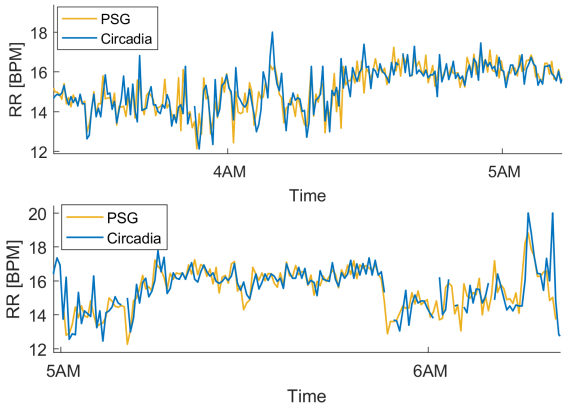

Fig. 2. Respiratory rate from PSG data (yellow) and as predicted by the C100 device, for two representative datasets. Overall accuracy was 93.2 % (SD ±1.82 pp).

Non-contact measurement of respiratory rate obtained using the C100 device was found to be 93.2 % (SD ±1.82 pp) accurate in comparison with PSG respiratory belt data (100 - MAPE). An example respiration trace (normalized amplitude) is shown in Fig. 1, demonstrating the device ability to accurately track chest excursion. A comparison of respiratory rate between the C100 device and PSG data is shown in Fig. 2, for two representative datasets. Sleep stage classification performance of the initial dataset (17 nights of data, 9 participants) was evaluated against PSG data using cross-validation. An epoch-by-epoch recall (true positive rate) of 75.0 %, 59.9 %, 74.8 % and 57.1 %, was found for 'Deep', 'Light', 'REM' and 'Wake' respectively (confusion matrix in Fig. 3. Predicted hypnograms for three representative datasets are given in Fig. 4). Precision was 70.3 %, 67.0 %, 69.2 % and 51.5 %, for 'Deep', 'Light', 'REM' and 'Wake' respectively. Overall accuracy (recall for each sleep stage weighted by the number of occurrences) was 66.7 %.

Prediction accuracy of sleep metrics by the C100 device was compared to accuracy of two wrist-worn devices: The Fitbit Alta HR and Philips Respironics Actiwatch Spectrum Plus. Data were obtained in a direct comparison, as part of the initial 17-nights trial. Not all metrics are provided by both wrist-worn devices, available results are given in Table I.

From the independent validation dataset, an epoch-by-epoch recall (true positive rate) of 70.7 %, 52.5 %, 83.0 % and

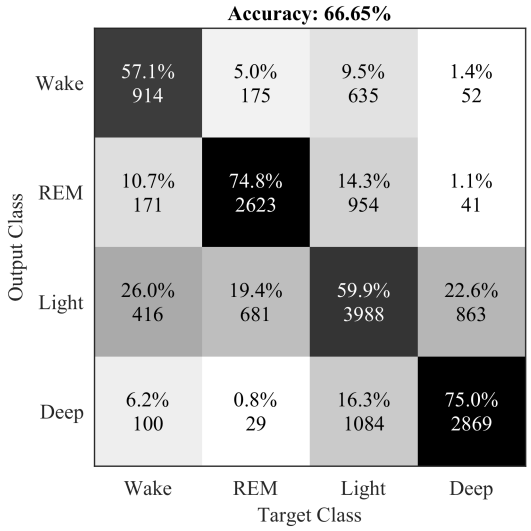

Fig. 3. Confusion matrix summarizing the epoch-by-epoch sleep stage classification performance of the C100 device and sleep analysis algorithm. Results were obtained using a leave-one-subject-out cross-validation approach, from 17 nights of data (initial dataset). Rows represent instances in the C100 (Output) class, whereas columns represent instances in the PSG (Target) class. Each cell shows the percentage of instances as a fraction of total number in the target class. Recall is shown in the matrix diagonal.

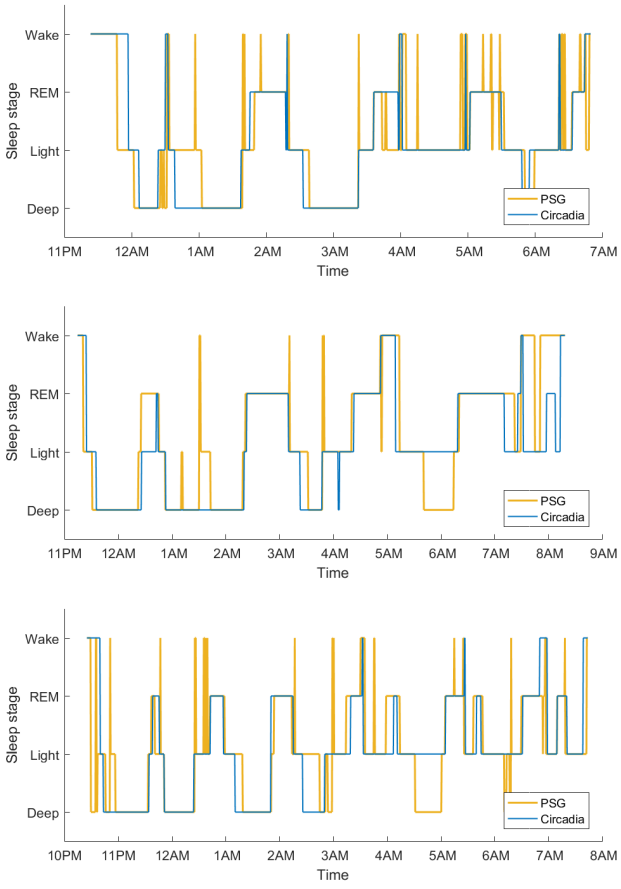

Fig. 4. Hypnograms from PSG data (yellow) and as predicted by the C100 device, for three representative datasets. Hypnograms from the contactless monitor predict almost all major sleep events throughout the night.

TABLE I

DIRECT COMPARISON OF SLEEP METRIC PREDICTION ACCURACY, WITH PSG AS GROUND TRUTH (N=17).

Sleep metric	Accuracy		
	Circadia C100	Fitbit Alta HR	Philips Actiwatch
Total Sleep Time	**95.5 %**	93.8 %	94.7 %
Sleep Efficiency	**94.6 %**	95.4 %	94.8 %
Sleep Onset Latency	**42.7 %**	-12.7 %	37.0 %
Wake After Sleep Onset	**60.0 %**	-	52.2 %
Percentage REM Sleep	**71.7 %**	70.1 %	-
Percentage Deep Sleep	**76.3 %**	61.6 %	-
REM Latency	**76.6 %**	-	-

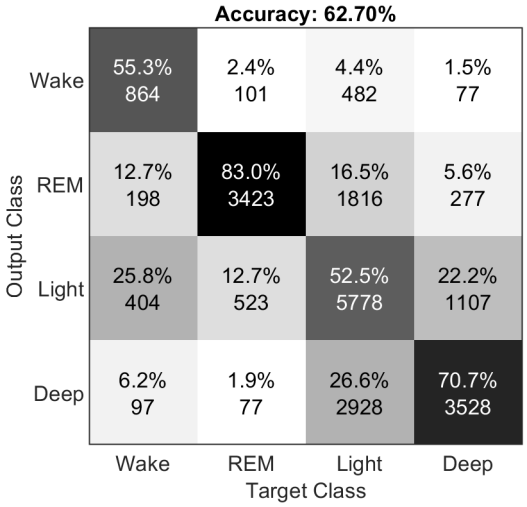

Fig. 5. Confusion matrix summarizing the epoch-by-epoch sleep stage classification performance from the independent validation cohort (n=24). Training and testing data were fully independent. Rows represent instances in the C100 (Output) class, whereas columns represent instances in the PSG (Target) class.

55.3 %, was found for 'Deep', 'Light', 'REM' and 'Wake' respectively (confusion matrix in Fig. 5). Precision was 53.2 %, 74.0 %, 59.9 % and 56.7 %, for 'Deep', 'Light', 'REM' and 'Wake' respectively. Overall accuracy was 62.7 %.

IV. DISCUSSION

A novel, non-contact, radar-based continuous monitor was used in combination with the proprietary Sleep Analysis Algorithm, to predict sleep stages in two cohorts of healthy sleepers (n=17 and n=24 nights of data, home and sleep lab environment, respectively). Sleep staging performance and sleep metric prediction accuracy was assessed using PSG data.

Non-contact respiration monitoring using the C100 device was found to be highly accurate and robust. Inaccuracies in respiratory rate monitoring occurred during participant motion, both on the novel contactless monitor and on the PSG device.

Sleep stage prediction accuracy was found to be high, and PSG accuracy was being approached: Due to the subjective nature of scoring PSG data (inter-rater-agreement on average is 82 % [9]), a sleep staging accuracy of 100 % is

unattainable. An average sleep staging accuracy of 66.7 % suggests that the developed non-contact monitor forms a reasonably accurate alternative to clinically used sleep monitoring devices, with the additional advantages of low cost and long-term monitoring in a home environment. Sleep stage classification performance results obtained using cross-validation on the initial dataset (n=17) were highly similar to results obtained from the independent validation dataset (n=24). These results provide a high level of confidence in the generalizability of the sleep staging model, at least within the healthy population. In addition, results indicate adequate performance of the device, irrespective of environment (sleep lab versus home). Recall from the independent validation data did show increased sensitivity to Deep and REM sleep, indicating that the model could benefit from retraining on a larger dataset. Sleep staging performance of the C100 device was compared against published performance data of the following non-wearable consumer sleep tracking devices: SleepScore Labs (formerly sold as ResMed S+), Beddit, and Early Sense. The Oura Ring was included as an example of a wearable device. A direct comparison is complicated by the fact that experimental conditions may differ, and not all manufacturers publish device performance data. Performance data for the various devices were obtained from online sources. varying from manufacturer white papers, to independent device validation studies. An attempt was made to include the most recent and objective performance assessments. Epoch-by-epoch sleep staging true positive rate (identical to recall) was compared across devices, for sleep stages 'Wake', 'REM', 'Light', and 'Deep'. Results are given in Table II. Overall accuracy of the C100 device and Sleep Analysis Algorithm was found to be higher than any of the devices included in this comparison.

In a direct comparison against two commercially available wrist-worn monitoring devices, the C100 device was found to outperform both on sleep metric prediction accuracy. Estimation of Sleep Onset Latency remains a challenge for any device due to the gradual transition of wake to sleep and the level of ambiguity in determining exact sleep onset. Relative error was large also due to the low sleep onset latency values, leading to a Fitbit prediction accuracy of -12.7 % (112.7 % MAPE). Compared to the medical grade Philips Respironics actigraphy device, the C100 device was found to be more accurate in distinguishing sleep from wake, and provided additional sleep staging metrics.

Both the initial data collection and independent validation have been performed on cohorts of young and healthy sleepers. Additional validation in patient populations diagnosed with sleeping disorders, as well as further model improvements, are currently ongoing.

TABLE II

COMPARISON OF SLEEP STAGING RECALL ACROSS VARIOUS NON-CONTACT CONSUMER SLEEP TRACKERS.

Device	Sleep staging recall (True positive rate)			
	Wake	Rem	Light	Deep
Circadia C100	**57.1 %**	**75.8 %**	**59.9 %**	**75.0 %**
SleepScore Sonar [4]	66 %	59 %	58 %	60 %
SleepScore RF [4]	56 %	65 %	66 %	55 %
ResMed S+ [5]	43.8 %	21.5 %	71.9 % (NREM)	
ResMed S+ [10]	73.1 %	61.6 %	65.1 %	52.2 %
Early Sense [6]	80.4 %	53.7 %	64.9 %	56.3 %
Beddit* [7]	42.1 %	55.6 %	37.4 %	-
Oura Ring [11]	48.1 %	61.4 %	64.6 %	50.9 %

*Reported Beddit performance metrics are precision, not recall. For reference: C100 precision for Wake, REM, Light and Deep sleep was 51.5 %, 69.2 %, 67.0 %, and 70.3 %, respectively.

V. CONCLUSION

In a direct comparison against gold standard PSG, it was found that sleep stage classification and sleep metric estimation using the C100 device and Sleep Analysis Algorithm is highly accurate. High sleep staging accuracy was being replicated in an independent validation study, indicating robustness of results and generalizability of the sleep staging model. C100 sleep staging performance was compared to published performance data of various consumer sleep tracking devices, and was found to be high in comparison across sleep stages. In addition, respiration monitoring was found to be highly accurate. Results suggest that the developed non-contact monitor forms a viable alternative to existing clinically used wrist-worn methods, and that longitudinal monitoring of respiration and sleep stages in a home environment is feasible.

ACKNOWLEDGEMENTS

The authors would like to thank Jonas Beck, Prof. Björn Rasch, Aditi Satija, Adam Birdseye, Sara Stevens, and Sakina Dastagir for their help with data collection. The authors would also like to thank Dr. Guy Leschziner and Dr. Saad Marfani for manuscript review.

REFERENCES

[1] M. T. Smith *et al.*, "Use of actigraphy for the evaluation of sleep disorders and circadian rhythm sleep-wake disorders...." *J. Clin. Sleep Med.*, vol. 14, no. 07, pp. 1231–1237, 2018.

[2] M. Hirshkowitz, "Polysomnography and beyond," in *Principles and Practice of Sleep Medicine*. Elsevier, 2017, pp. 1564–1566.

[3] K. R. Evenson *et al.*, "Systematic review of the validity and reliability of consumer-wearable activity trackers," *Int. J. Behav. Nutr. Phys. Act.*, vol. 12, no. 1, p. 159, 2015.

[4] S. Labs, "Evaluation of the sleepscore labs sonar and radiofrequency sleep sensor technologies by resmed," 2018, www.sleepscore.com/media/white-papers/, viewed March 2019.

[5] K.-Y. Chung *et al.*, "Noncontact sleep study by multi-modal sensor fusion," *Sensors*, vol. 17, no. 7, p. 1685, 2017.

[6] A. Tal *et al.*, "Validation of contact-free sleep monitoring device with comparison to polysomnography," *J. Clin. Sleep Med.*, vol. 13, no. 03, pp. 517–522, 2017.

[7] J. Tuominen *et al.*, "Sleep parameter assessment accuracy of a consumer home sleep monitoring ballistocardiograph beddit sleep tracker...." *J. of Clin. Sleep Med.*, vol. 15, no. 3, pp. 483–487, 2019.

[8] M. Amin *et al.*, "Radar Signal Processing for Elderly Fall Detection: The future for in-home monitoring," *IEEE Signal Processing Magazine*, vol. 33, pp. 71–80, mar 2016.

[9] H. Danker-hopfe *et al.*, "Interrater reliability for sleep scoring according to the rechtschaffen & kales and the new aasm standard," *Journal of sleep research*, vol. 18, no. 1, pp. 74–84, 2009.

[10] M. M. Schade *et al.*, "Sleep validity of a non-contact bedside movement and respiration-sensing device," *Journal of Clinical Sleep Medicine*, vol. 15, no. 7, pp. 1051–1061, 2019.

[11] M. de Zambotti *et al.*, "The sleep of the ring: comparison of the ōura sleep tracker against polysomnography," *Behavioral sleep medicine*, vol. 17, no. 2, pp. 124–136, 2019.