A deep learning model based on the combination of convolutional and recurrent neural networks to enhance pulse oximetry ability to classify sleep stages in children with sleep apnea

Fernando Vaquerizo-Villar, Daniel Álvarez*, *Member, IEEE*, Gonzalo C. Gutiérrez-Tobal, *Member, IEEE*, Félix del Campo, David Gozal, Leila Kheirandish-Gozal, Thomas Penzel, *Senior Member, IEEE*, Roberto Hornero, *Senior Member, IEEE*

***Abstract*— Characterization of sleep stages is essential in the diagnosis of sleep-related disorders but relies on manual scoring of overnight polysomnography (PSG) recordings, which is onerous and labor-intensive. Accordingly, we aimed to develop an accurate deep-learning model for sleep staging in children suffering from pediatric obstructive sleep apnea (OSA) using pulse oximetry signals. For this purpose, pulse rate (PR) and blood oxygen saturation (SpO_2) from 429 childhood OSA patients were analyzed. A CNN-RNN architecture fed with PR and SpO_2 signals was developed to automatically classify wake (W), non-Rapid Eye Movement (NREM), and REM sleep stages. This architecture was composed of: (*i*) a convolutional neural network (CNN), which learns stage-related features from raw PR and SpO_2 data; and (*ii*) a recurrent neural network (RNN), which models the temporal distribution of the sleep stages. The proposed CNN-RNN model showed a high performance for the automated detection of W/NREM/REM sleep stages (86.0% accuracy and 0.743 Cohen's kappa). Furthermore, the total sleep time estimated for each children using the CNN-RNN model showed high agreement with the manually derived from PSG (intra-class correlation coefficient = 0.747). These results were superior to previous works using CNN-based deep-learning models for automatic sleep staging in pediatric OSA patients from pulse oximetry signals. Therefore, the combination of CNN and RNN allows to obtain additional information from raw PR and SpO_2 data related to sleep stages, thus being useful to automatically score sleep stages in pulse oximetry tests for children evaluated for suspected OSA.**

***Clinical Relevance*—This research establishes the usefulness of a CNN-RNN architecture to automatically score sleep stages in pulse oximetry tests for pediatric OSA diagnosis.**

I. Introduction

Characterization of the sleep macro-structural changes (i.e., sleep stages) is essential in the diagnosis of sleep-related disorders [1]. Overnight polysomnography (PSG) is the gold standard approach, which involves the recording of a wide range of neurophysiological and cardiorespiratory signals [2]. After the test and following the rules of the American Academy of Sleep Medicine (AASM), sleep technicians visually inspect the electroencephalogram (EEG), electrooculogram (EOG), and submental electromyogram (EMG) channels to assign each 30-s non-overlapping epoch to a sleep stage: wake (W), three levels of non-Rapid Eye Movement (non-REM) sleep (N1, N2, and N3), and REM sleep [2]. However, PSG is costly, complex, highly intrusive, and scarcely available, thus delaying the diagnosis of sleep disorders [3]. Furthermore, the process of manual sleep scoring takes up to hours per sleep study and suffers from a considerable inter-rater variability [4], which may alter the accuracy of the diagnosis.

To overcome these limitations, multiple studies have proposed automated approaches for sleep scoring from a minimum number of signals [5]. A large proportion of these studies have focused on automated sleep staging in patients with obstructive sleep apnea (OSA), a highly prevalent sleep disorder that affects nearly 1 billion people around the globe [6]. OSA diagnosis is based on the apnea-hypopnea index (AHI: number of apneas and hypopneas per sleep hour), so the scoring of sleep stages and the calculation of the total sleep time (TST) are imperative in this context [2].

Among others, EEG, EOG, electrocardiogram, actigraphy, airflow and pulse oximetry signals have been employed for automatic sleep staging in OSA cohorts [5]. In this respect, pulse oximetry signals have been frequently proposed for sleep scoring and diagnosing sleep disorders as they can be recorded at patient's home with low-cost portable pulse oximeters [3], thus being an accessible and simplified alternative to PSG [7], [8]. Pulse oximeters record the photoplethysmography (PPG) signal, which is used to derive both blood oxygen saturation (SpO_2) and pulse rate (PR) signals [9].

The dynamics of PPG and PPG-derived PR and SpO_2 changes during sleep stages [7], [8], [10]. This relationship,

This work was supported by 'Ministerio de Ciencia e Innovación/Agencia Estatal de Investigación/10.13039/501100011033/', ERDF A way of making Europe, and NextGenerationEU/PRTR under projects PID2020-115468RB-I00 and PDC2021-120775-I00, and by 'CIBER -Consorcio Centro de Investigación Biomédica en Red-' (CB19/01/00012) through 'Instituto de Salud Carlos III'.

Daniel Álvarez was supported by a "Ramón y Cajal" Grant RYC2019-028566-I funded by MCIN/AEI/ 10.13039/501100011033 and by "ESF Investing in your future". L. Kheirandish-Gozal and D. Gozal were supported by the Leda J. Sears Foundation for Pediatric Research, by a Tier 2 grant from the University of Missouri and National Institutes of Health grant AG061824.

F. Vaquerizo-Villar, D. Álvarez, G. C. Gutiérrez-Tobal, F. del Campo, and R. Hornero are with the Biomedical Engineering Group, Universidad de Valladolid (e-mail: dalvgon@gmail.com) and CIBER-BBN (ISCIII), Spain.

D. Álvarez, and F. del Campo are with the Hospital Universitario Río Hortega of Valladolid, Spain (e-mail: fsas@telefonica.net) and CIBER-BBN (ISCIII), Spain.

L. Kheirandish-Gozal and D. Gozal are with the Department of Child Health, The University of Missouri School of Medicine, Columbia, Missouri, USA (email: gozald@health.missouri.edu).

T. Penzel is with the Interdisciplinary Center of Sleep Medicine, Charité-Universitätsmedizin Berlin, Germany (e-mail: thomas.penzel@charite.de).

together with the recent advances in deep-learning methodologies, has led to several studies applying deep-learning algorithms to automatically score sleep stages in adult OSA subjects from pulse oximetry signals [7], [8], [11]. Conversely, only two conference papers developed by our own group have approached sleep staging in pediatric OSA patients [10], [12], which present distinguishing etiological, diagnostic, and treatment considerations, as well as less profound and recurrent desaturations (SpO_2) and bradycardia/tachycardia (PR) patterns when compared to adult subjects [2], [13]. In these two preliminary studies, a convolutional neural network (CNN) was applied to detect sleep stages from raw PPG [12], and raw PR and SpO_2 data [10], respectively. Despite their usefulness to learn stage-related features from pulse oximetry signals, CNNs do not consider the temporal distribution of sleep stages during sleep. Instead, recurrent neural networks (RNNs) learn the temporal dependency of the data [14], which has been shown to be useful in order to learn the temporal distribution of the sleep stages [5], [7].

Based on these considerations, we hypothesized that a deep-learning architecture based on the combination of a CNN and a RNN (CNN-RNN) could extract additional information from the PR and SpO_2 signals able to improve the automated detection of sleep stages in childhood OSA patients. Consequently, our main objective is to design and assess a CNN-RNN deep-learning architecture to identify W, NREM, and REM stages from PR and SpO_2 recordings in children with suspected OSA.

II. Materials and Methods

A. Subjects and signals

The baseline dataset from the semi-public Childhood Adenotonsillectomy Trial (CHAT) database was used in this study [15], [16]. The clinical trial identifier of the CHAT database is NCT00560859 and its full research protocol can be found in the supplementary material of Marcus *et al*. [15]. The CHAT-baseline dataset is composed of PSG recordings from 453 children aged 5 to 10 years old suffering from OSA, who were randomized to a strategy of watchful waiting or early adenotonsillectomy treatment [16]. Each sleep study contains annotations of sleep stages and apnea/hypopnea events, which were done using the AASM 2007 rules [17].

This dataset provided valid PR and SpO_2 signals from 429 pediatric subjects. The data, originally recorded during PSG using sampling rates (*fs*) from 1 to 512 Hz, were resampled to a common *fs* of 1 Hz [8], [10]. Then, a subject-based standardization was performed to normalize PR and SpO_2 baseline levels among different children. PR and SpO_2 signals were finally divided into consecutive 30-second epochs, being each epoch classified as W, NREM, or REM with the annotations provided by sleep technicians [10].

The data were split into three sets: training (257 first children, 60%), used to train the CNN-RNN model; validation set (85 following children, 20%), employed to monitor the convergence of the CNN-RNN; and test set (last 87 children, 20%), used for performance assessment. Table I shows clinical and demographic data from the population under study.

TABLE I. Clinical and Demographic Data of the Children in the Study

	All	Training set	Validation set	Test set
Subjects (n)	429	257	85	87
Age (years)	6 [5, 8]	6 [5, 8]	6 [5, 7]	6 [5, 7]
Males (n)	208 (48.5%)	127 (49.4%)	35 (41.2%)	46 (52.9%)
BMI (kg/m2)	17.2 [15.4, 22.0]	17.1 [15.6, 21.7]	18.5 [15.2, 23.4]	16.5 [15.2, 22.3]
AHI (e/h)	4.7 [2.7, 8.7]	4.6 [2.6, 8.8]	4.6 [2.5, 8.5]	5.1 [3.2, 9.4]
Wake (*n*)	133891 (25.4%)	79814 (25.1%)	27685 (27.0%)	26392 (25.0%)
NREM (*n*)	319038 (60.6%)	193547 (60.9%)	61419 (59.9%)	64072 (60.6%)
REM (*n*)	73405 (14.0%)	44724 (14.1%)	13464 (13.1%)	15217 (14.4%)
TRT (min)	608 [557, 658]	617 [563, 661]	590 [539, 652]	607 [562, 640]
TST (min)	466 [429, 494]	472 [440, 497]	447 [420, 482]	461 [423, 500]

Data are presented as median [interquartile range], *n* or %. BMI: Body Mass Index; AHI: Apnea-Hypopnea Index; e/h: events per hour; REM: Rapid Eye Movement; NREM: Non-REM; TRT: Total Recording Time; TST: Total Sleep Time

B. CNN-RNN architecture

Figure 1 shows the main components of the CNN-RNN architecture employed in this study. Adapted from the CNN-RNN proposed by Korkalainen *et al.* (2019) to detect sleep stages in adults from PPG data [7], the proposed CNN-RNN receives as input a sequence of 100 consecutive epochs of 30-s of the PR and SpO_2 signals (100x30x2 samples). First, each epoch is processed separately through a time distributed layer that contains a CNN. The CNN is composed of 5 convolutional blocks (conv block), which are intended to automatically learn the features of each epoch of the PR and SpO_2 signals (30x2 samples) related with W/NREM/REM stages. Each conv block consists of: (*i*) a convolutional layer, which extracts the feature maps from PR and SpO_2 data using 32 filters of size 5x2; (*ii*) a batch normalization layer that normalizes the feature maps; (*iii*) a Rectified Linear Unit (ReLU) activation function that introduces nonlinearity to the normalized feature maps; and (*iv*) a dropout operation that minimizes overfitting by randomly removing node connections with a probability of 0.1 [14]. After the last conv block, the 3D feature maps are reshaped into 1D data using a flattening operation.

The time distributed CNN is then processed using a RNN to learn the temporal distribution of the sleep stages in the sequence. First, a dropout layer with a rate of 0.3 is used to minimize overfitting [7], [14]. Next, a bidirectional Gate Recurrent Unit (GRU) layer is applied to model the temporal dependence of the input sequence, deciding the information to be retained and the information to be forgotten from the network [14]. GRU was chosen instead of Long Short-Term Memory (LSTM) as it provided similar results with a lower computational cost [14]. This layer contains 64 units with a dropout probability of 0.3 in the forward step and 0.5 in the recurrent step [7]. Finally, a time distributed layer containing a softmax activation function is employed to obtain the probability of belonging to W (p_W^i), NREM (p_{NREM}^i), and REM (p_{REM}^i) stages for the epoch i of the input sequence.

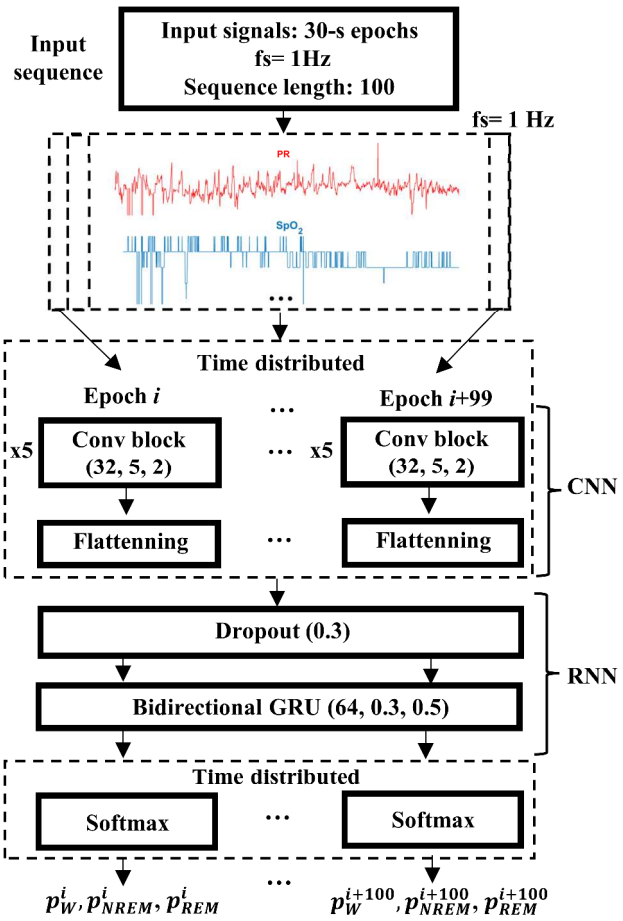

Figure 1. Overview of the proposed deep-learning architecture based on the combination of a CNN and a RNN (CNN-RNN). Each convolutional block (conv block) includes a convolutional layer, batch normalization, a ReLU activation function, and dropout.

The CNN-RNN architecture was implemented using TensorFlow library and trained with the following configuration [14]: He-normal method to initialize network weights; categorical cross-entropy as the loss function; batch size of 128 with a random data shuffling strategy; the Adam method with an initial learning rate of 0.0001 to optimize network weights; early stopping after 30 training steps of non-improvement; in the validation loss; and 500 as the maximum number of training steps.

C. *Statistical analysis*

The overall performance of the CNN-RNN for automatic sleep staging was assessed by means of confusion matrices (3-class), which were used to compute the 3-class accuracy (Acc), Cohen's kappa index (kappa), macro F1-score (MF1), and per-class F1-score (F1). Additionally, the TST was computed for each patient based on the sleep stages scored by the CNN-RNN model ($TST_{CNN\text{-}RNN}$) and compared with the TST from standard PSG (TST_{PSG}). Bland-Altman plots and the intra-class correlation coefficient (ICC) were used to assess the estimated TST agreement.

III. Results

A. *CNN-RNN model performance*

Figure 2 shows the confusion matrix of the CNN-RNN model obtained in the test set for automatic sleep staging (W/NREM/REM). Interestingly, the CNN-RNN model fed with sequences of 100 epochs of PR and SpO_2 signals rightly classified 86.1% of the 30-s epochs (91240/106029), with a kappa of 0.743, a MF1 of 0.820, and F1-scores of 0.847, 0.901, and 0.711 for W, NREM, and REM sleep stages, respectively.

B. *Estimation of the TST*

Figure 3 shows the Bland-Altman plot comparing the TST calculated from automatic CNN-RNN scoring ($TST_{CNN\text{-}RNN}$) with the TST derived from PSG (TST_{PSG}) in the test set. ICC is also shown. $TST_{CNN\text{-}RNN}$ slightly overestimated TST_{PSG}, as reported by their mean difference (16.1 min) and confidence interval (from -52.6 to 84.8 min). Additionally, $TST_{CNN\text{-}RNN}$ showed an ICC of 0.747 with TST_{PSG}.

IV. Discussion

In this work, we propose a CNN-RNN architecture to enhance the automatic scoring of wake, NREM, and REM sleep stages from pulse oximetry signals (PR and SpO_2) in childhood OSA patients. To our knowledge, the application of a deep-learning model based on the combination of a CNN and

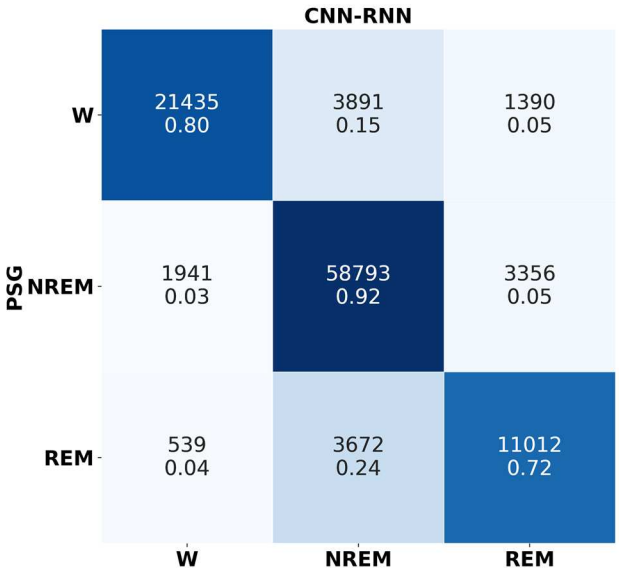

Figure 2. Confusion matrix of the CNN-RNN architecture in the test set. This matrix compares the sleep stages manually scored from PSG with the corresponding automatic assignation using the CNN-RNN model.

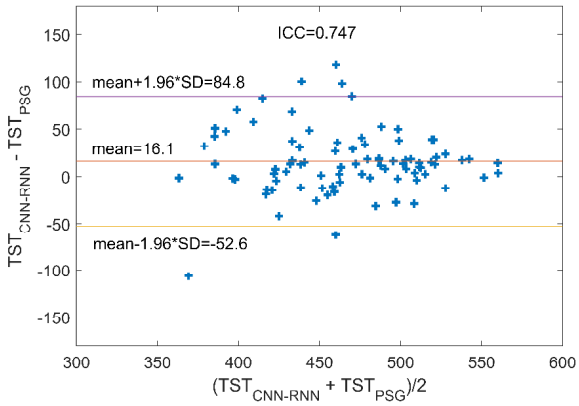

Figure 3. Bland-Altman plot comparing $TST_{CNN\text{-}RNN}$ with TST_{PSG} in the test set.

a RNN is novel in the framework of automated sleep staging in pediatric subjects.

The proposed CNN-RNN architecture reached a high performance, with 86.1% Acc and 0.743 kappa for W/NREM/REM sleep classification. Particularly, the kappa value obtained by the CNN-RNN model (in the range 0.61-0.80) indicates that there is a substantial agreement between our automatic deep-learning model and manual scoring from PSG [18]. Hence, our proposal could provide sleep stage annotations in at-home pulse oximetry tests for the screening of childhood OSA [3]. The TST derived from the CNN-RNN architecture also showed a high concordance with the TST from PSG (TST_{PSG}), with an ICC of 0.747, a mean difference of 16.1 min, and a confidence interval of -52.6 to 84.8 min. The slight overestimation of TST_{PSG} can be explained by the slight trend of the CNN-RNN to classify W epochs (15%) as NREM (see figure 2), being NREM the majority class in the data. Conversely, the obtained ICC value (in the range 0.50-0.75) indicates a moderate agreement [18], highlighting the usefulness of our proposal to derive the TST in oximetry tests [3], [7], [8].

Two preliminary studies performed by our research group have shown the usefulness of CNN-based deep-learning methodologies for pediatric sleep staging, reporting a superior performance than previous feature-based approaches [10], [12]. In Vaquerizo *et al*. [12], we reported 78.3% Acc and 0.57 kappa for the detection of W/NREM/REM from raw PPG data, and an ICC of 0.59 for the estimation of the TST. In Vaquerizo *et al*. [10], 83.1% Acc and 0.68 kappa were obtained for W/NREM/REM classification from raw PR and SpO_2 data, whereas an ICC of 0.677 was obtained for the calculation of the TST. In this work, which has used the same database as in the two previous studies [10], [12], a higher performance was obtained with a CNN-RNN fed with PR and SpO_2 data: 86.1% Acc, 0.743 kappa, and 0.747 ICC. Thus, the information about the temporal distribution of the data provided by the RNN allows to improve the detection of sleep stages.

It is important to denote some limitations of our study. First, although the sample size is considerably large (429 subjects), the database only contains children suffering from OSA (AHI≥ 1 e/h). Thus, additional pediatric datasets that include, among others, healthy control subjects would be desirable. Another limitation is the computational load of the RNN, which may hinder its implementation in portable devices. In this respect, novel deep-learning methods with a lower computational cost than RNNs (e.g., transformers), as well as novel strategies addressing imbalance between sleep stages, should be assessed in future studies. Finally, another interesting future goal could be to design and assess an automatic deep-learning model that simultaneously score sleep stages and apnea/hypopnea events, thus providing a complete diagnosis of childhood OSA from pulse oximetry signals.

V. Conclusion

In summary, a deep-learning architecture based on the combination of a CNN and a RNN has shown usefulness to automatically score wake, NREM, and REM sleep stages from raw PR and SpO_2 data in childhood OSA patients, with a higher performance than the reported by previous studies. In addition, we showed that the CNN-RNN model can provide a reliable estimation of the TST in pulse oximetry tests. Thus, we conclude that CNN-RNN architectures can be used to extract additional information on the temporal distribution of sleep stages from pulse oximetry recordings in children being evaluated for suspected OSA.

References

[1] M. J. Sateia, "International Classification of Sleep Disorders-Third Edition," *Chest*, vol. 146, no. 5, pp. 1387–1394, Nov. 2014.

[2] R. B. Berry, R. Brooks, C. E. Gamaldo, S. M. Harding, C. L. Marcus, and B. V. Vaughn, "The AASM Manual for the Scoring of Sleep and Associated Events," *Am. Acad. Sleep Med.*, vol. 53, no. 9, pp. 1689–1699, 2018.

[3] F. del Campo, A. Crespo, A. Cerezo-Hernández, G. C. Gutiérrez-Tobal, R. Hornero, and D. Álvarez, "Oximetry use in obstructive sleep apnea," *Expert Rev. Respir. Med.*, vol. 12, no. 8, pp. 665–681, 2018.

[4] A. Malhotra *et al.*, "Performance of an automated polysomnography scoring system versus computer-assisted manual scoring," *Sleep*, vol. 36, no. 4, pp. 573–582, 2013.

[5] O. Faust, H. Razaghi, R. Barika, E. J. Ciaccio, and U. R. Acharya, "A review of automated sleep stage scoring based on physiological signals for the new millennia," *Comput. Methods Programs Biomed.*, vol. 176, pp. 81–91, 2019.

[6] A. V Benjafield *et al.*, "Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis," *Lancet Respir Med*, vol. 7, no. 8, pp. 687–698, 2020.

[7] H. Korkalainen *et al.*, "Deep learning enables sleep staging from photoplethysmogram for patients with suspected sleep apnea," *Sleep*, vol. 43, no. 11, pp. 1–10, 2020.

[8] R. Casal, L. E. Di Persia, and G. Schlotthauer, "Temporal convolutional networks and transformers for classifying the sleep stage in awake or asleep using pulse oximetry signals," *J. Comput. Sci.*, vol. 59, no. December 2021, 2022.

[9] E. D. Chan, M. M. Chan, and M. M. Chan, "Pulse oximetry: Understanding its basic principles facilitates appreciation of its limitations," *Respir. Med.*, vol. 107, no. 6, pp. 789–799, 2013.

[10] F. Vaquerizo-Villar *et al.*, "A convolutional neural network to classify sleep stages in pediatric sleep apnea from pulse oximetry signals," *MELECON 2022 - IEEE Mediterr. Electrotech. Conf. Proc.*, pp. 108–113, 2022.

[11] M. Radha *et al.*, "A deep transfer learning approach for wearable sleep stage classification with photoplethysmography," *npj Digit. Med.*, vol. 4, no. 1, pp. 1–11, 2021.

[12] F. Vaquerizo-villar *et al.*, "Automatic Sleep Staging in Children with Sleep Apnea using Photoplethysmography and Convolutional Neural Networks," in *43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC 2021)*, 2021, pp. 216–219.

[13] C. L. Rosen, L. D'Andrea, and G. G. Haddad, "Adult criteria for obstructive sleep apnea do not identify children with serious obstruction," *Am Rev Respir Dis*, vol. 146, no. 5 Pt 1, pp. 1231–1234, 1992.

[14] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[15] C. L. Marcus *et al.*, "A Randomized Trial of Adenotonsillectomy for Childhood Sleep Apnea," *N. Engl. J. Med.*, vol. 368, no. 25, pp. 2366–2376, 2013.

[16] S. Redline *et al.*, "The Childhood Adenotonsillectomy Trial (CHAT): Rationale, Design, and Challenges of a Randomized Controlled Trial Evaluating a Standard Surgical Procedure in a Pediatric Population," *Sleep*, vol. 34, no. 11, pp. 1509–1517, 2011.

[17] C. Iber, S. Ancoli-Israel, A. Chesson, and S. F. Quan, "The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specification," *J. Clin. Sleep Med.*, vol. 3, no. 7, p. 752, 2007.

[18] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochem. Medica*, vol. 22, no. 3, pp. 276–282, 2012.