Automatic Assessment of Pediatric Sleep Apnea Severity Using Overnight Oximetry and Convolutional Neural Networks

Fernando Vaquerizo-Villar*, Daniel Álvarez, Leila Kheirandish-Gozal, Gonzalo C. Gutiérrez-Tobal, *Member, IEEE*, Javier Gómez-Pilar, Andrea Crespo, Félix del Campo, David Gozal, Roberto Hornero, *Senior Member, IEEE*

Abstract. **In this study, we use the overnight blood oxygen saturation (SpO_2) signal along with convolutional neural networks (CNN) for the automatic estimation of pediatric sleep apnea-hypopnea syndrome (SAHS) severity. The few preceding studies have focused on the application of conventional feature extraction methods to obtain information from the SpO_2 signal, which may omit relevant data related to the illness. In contrast, deep learning techniques are able to automatically learn features from raw input signal. Thus, we propose to assess whether CNN, a deep learning algorithm, could automatically estimate the apnea-hypopnea index (AHÍ) from nocturnal oximetry to help establish pediatric SAHS presence and severity. A database of 746 SpO_2 recordings is involved in the study. CNN was trained using 20-min segments from the SpO_2 signal in the training set (400 subjects). Hyperparameters of the CNN architecture were tuned using a validation set (100 subjects). This model was applied to a test set (246 subjects), in which the final AHI of each patient was obtained as the average of the output of the CNN for all the segments of the corresponding SpO_2 signal. The AHI estimated by the CNN showed a promising diagnostic performance, with 74.8%, 90.7%, and 95.1% accuracies for the common AHI severity thresholds of 1, 5, and 10 events per hour (e/h), respectively. Furthermore, this model reached 28.6, 32.9, and 120.0 positive likelihood ratios for the above-mentioned AHI thresholds. This suggests that the information extracted from the oximetry signal by deep learning techniques may be useful to both establish pediatric SAHS and its severity.**

***Clinical Relevance*—This research establishes the usefulness of CNN to estimate AHI in pediatric SAHS patients using the oximetry signal.**

I. Introduction

Pediatric sleep apnea-hypopnea syndrome (SAHS) is a respiratory disorder characterized by recurrent episodes of partial and/or complete obstruction of the child's upper airway during sleep [1]. Untreated pediatric SAHS may lead to many adverse consequences for children's health and quality of life, including impairment of neuropsychological and cognitive performance, metabolic dysfunction, cardiac derangements, and systemic inflammation [1]. The prevalence of pediatric SAHS is estimated in the range 1% to 5% of children [1]. However, despite its major negative consequences, pediatric SAHS is considered an underdiagnosed condition [1], [2].

SAHS is diagnosed by means of the overnight polysomnography (PSG) test, which acts as "gold standard". This test requires an overnight stay of children in a specialized sleep unit, where multiple physiological signals are recorded. These recordings need an offline inspection to score complete breathing cessation events (apneas) and significant airflow reductions (hypopneas) to compute the apnea-hypopnea index (AHI), which is used to reach a diagnosis [3]. However, PSG is technically complex, relative unavailable, and highly intrusive, thus delaying the access for both the diagnosis and treatment [4].

These drawbacks, together with the high prevalence of the disease, have led the scientific community to explore the use of simplified screening tests. In this sense, overnight oximetry has been previously used [5]–[10]. Overnight oximetry records the blood oxygen saturation (SpO_2) signal with the only use of a pulse oximeter placed on a finger, thus being a simple, suitable, and reliable technique for children [4], [8]. SpO_2 signal is useful to detect apneic events, since decreases in blood oxygen levels, so-called oxygen desaturations, are associated to these events [3]. In this regard, the automated analysis of the SpO_2 signal has shown it usefulness as a simplified tool in the screening of pediatric SAHS [5]–[10].

Signal-processing algorithms employed in previous studies included statistical analysis, conventional clinical indices, nonlinear analysis, and frequency domain techniques, which were used to extract features from the SpO_2 signal [5]–[10]. Similarly, feature selection methods were used to obtain subsets of relevant features, whereas machine-learning algorithms were used to provide an automatic diagnosis of pediatric SAHS. However, these machine-learning based approaches require to determine which information extract

This work was supported by 'Ministerio de Ciencia, Innovación y Universidades' and 'European Regional Development Fund (FEDER)' under projects DPI2017-84280-R and RTC-2017-6516-1, by 'European Commission' and 'FEDER' under projects 'POCTEP 0378_AD_EEGWA_2_P' and 'POCTEP0702_MIGRAINEE_2_E, by CIBER-BBN (ISCIII), co-funded with FEDER funds, and by SEPAR under project 649/218.

F. Vaquerizo-Villar was in receipt of a 'Ayuda para contratos predoctorales para la Formación de Profesorado Universitario (FPU)' grant from the Ministerio de Educación, Cultura y Deporte (FPU16/02938). L. Kheirandish-Gozal and D. Gozal were supported by National Institutes of Health (NIH) grant HL130984.

F. Vaquerizo-Villar, G. C. Gutiérrez-Tobal, J. Gómez-Pilar and R. Hornero are with the Biomedical Engineering Group, Universidad de Valladolid (e-mail: fernando.vaquerizo@gib.tel.uva.es) and CIBER-BBN (ISCIII), Spain.

D. Álvarez, A. Crespo, 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) and CIBER-BBN (ISCIII), Spain.

from the physiological signals, leading to the omission of useful information from these signals that may help to detect apneic events [11].

In recent years, deep learning has emerged as a new methodology to automatically learn features and find patterns in raw physiological data [11]. Previous studies have suggested the ability of deep-learning algorithms to analyze physiological signals from PSG in adult SAHS patients [12]–[15]. In the context of pediatric SAHS, only one single preliminary study developed by our group applied convolutional neural networks (CNN) to the oximetry signal to detect apneic events [16]. However, this study neither estimated AHI nor SAHS severity, which is needed to reach a diagnosis [16]. Therefore, additional research is needed to assess the usefulness of CNNs in the diagnosis of pediatric SAHS.

We hypothesized that CNNs could help to automatically learn all the relevant information from SpO_2 signal associated with pediatric SAHS. Thus, our objective is to assess the usefulness of CNNs to estimate the AHI and hence the severity of SAHS using the oximetry signal.

II. Material and Methods

A. Subjects and signals

The nonrandomized dataset from the public multicenter Childhood Adenotonsillectomy Trial (CHAT) database was used in this study [17], [18]. The full protocol of CHAT database is provided in the supplementary material of [17]. This dataset is composed of PSG studies of 779 children ranging from 5-10 years of age. Apneas and hypopneas were scored using the 2007 American Academy of Sleep Medicine guidelines [3]. The AHI provided in the database was computed as the number of obstructive apneas, mixed apneas, and hypopneas (associated with 3% desaturation or arousal). In this study, the common AHI cutoffs of 1, 5, and 10 events per hour (e/h) were used to classify pediatric subjects into four SAHS severity degrees: no-SAHS (AHI<1), mild SAHS (1≤AHI<5 e/h), moderate SAHS (5≤AHI<10 e/h), and severe SAHS (AHI≥10 e/h) [5], [10].

SpO_2 signals from PSG were obtained at different sample frequencies: 1, 2, 10, 12, 16, 200, 256, and 512 Hz. A preprocessing stage was included to downsample the SpO_2 signals to a common sample rate of 1 Hz. Artifacts in the SpO_2 signal were also removed, following the methodology employed in previous studies [5], [10]. Finally, those subjects with a preprocessed SpO_2 recording longer than 3h were included in the study, since such duration ensures that there are enough sleep cycles [3]. A dataset of 746 preprocessed SpO_2 signals was finally obtained.

This dataset was divided into three sets: a training set (first 400 subjects, 54%), employed to train the CNNs, a validation set (100 following subjects, 13%), used to obtain the CNN model with the optimum values for the hyperparameters, and a test set (last 246 subjects, 33%), employed to assess the diagnostic ability of our proposal. SpO_2 signals were divided into 20-min segments and the AHI was estimated for each segment [15]. In the training and validation sets, SpO_2 signal segmentation was done with 95% overlap in order to increase the number of available segments.

TABLE I. Demographic and Clinical Data of the Subjects Under Study

	All	Training group	Validation group	Test group
Subjects (n)	746	400	100	246
Age (years)	7 [2]	7 [2]	7 [3]	7 [2]
Males (n)	345 (46.3%)	177 (44.3%)	46 (46.0%)	113 (45.9%)
BMI (kg/m2)	17.3 [5.2]	17.4 [5.5]	16.9 [4.4]	17.0 [4.8]
AHI (e/h)	0.8 [1.2]	0.8 [1.1]	0.9 [2.0]	0.9 [1.5]
AHI ≥ 1 (e/h)	303 (40.6%)	165 (41.3%)	41 (41.0%)	97 (39.4%)
AHI ≥ 5 (e/h)	100 (13.4%)	46 (11.5%)	17 (17.0%)	37 (15.0%)
AHI ≥ 10 (e/h)	66 (8.9%)	30 (7.5%)	11 (11.0%)	25 (10.2%)

Data are presented as median [interquartile range], n or %. BMI: Body Mass Index; AHI: Apnea-Hypopnea Index; e/h: events per hour

Thus, the training set had 197891 segments, the validation set had 49495 segments, and the test set had 6257 segments. Table I shows clinical and demographic data from the subjects under study.

B. Proposed CNN architecture

CNNs are inspired to process multidimensional arrays, such as 1D signals or 2D images, due to its multi-layer architecture with shared weights, local connections, and pooling operations [11]. Fig 1 shows the overall CNN-based architecture employed in this study. This architecture has three main sections: input, CNN, and output sections.

The input section of the network consists of the 20-min segments (1200 samples) of the SpO_2 preprocessed signals.

The CNN section is composed of λ CNN sub-blocks, each one composed of convolutional layer, batch normalization, rectified linear unit (ReLU) layer, pooling, and dropout:

- Convolutional layers. In these layers, feature maps are extracted from the data using convolution filters (kernels) [11]. In this study, each convolutional layer had 16 filters with a kernel size of 6 and a stride of 1, which are common values in these layers.
- Batch normalization (BN). BN is applied between the convolution and the nonlinearity (ReLU) in order to normalize the feature maps [11].
- ReLU layer. In this layer, a ReLU activation function is applied to perform a thresholding operation, deciding which feature maps are activated [11].
- Pooling layer. This layer is applied after ReLU layer to reduce the dimensionality, while retaining the significant information [11]. A max-pooling layer with a factor of 2, the most widely used in CNNs, was used in this study.
- Dropout layer. This is the last layer of each CNN block. Dropout is a regularization technique that randomly removes connections with a probability p in order to prevent overfitting [11].

Finally, the output section follows the last sub-block of the CNN section. A flattening layer is first applied to transform the feature maps into a 1-D column. Then, the last layer of the network is a linear activation function, which is used to obtain the estimated AHI of the 20-min SpO_2 input segment.

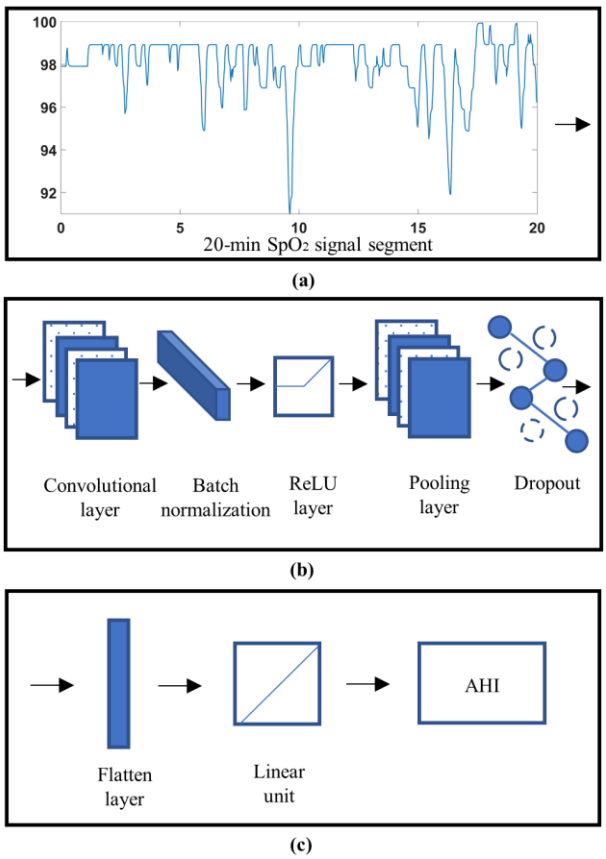

Figure 1. CNN-based architecture: (a) input block, it contains the 20-min SpO_2 segments (b) CNN block with: convolutional layer, batch normalization, ReLU layer, pooling layer, and dropout, and (c) output block, with the flattening layer, linear unit, and the estimated AHI of the segment at the output of the network.

The CNN architecture was implemented using Keras framework with TensorFlow backend. CNNs were trained on a NVIDIA GeForce RTX 2080 GPU. The estimated AHI for each patient was obtained as the average of the output values of the CNN obtained for each 20-min segment of the corresponding SpO_2 signal.

C. Statistical analysis

The diagnostic performance of the estimated AHI was assessed by means of sensitivity (Se, percentage of SAHS positive patients rightly classified), specificity (Sp, percentage of SAHS negative children rightly classified), positive predictive value (PPV, proportion of positive test results that are true positives), negative predictive value (NPV, proportion of negative test results that are true negatives), positive likelihood ratio (LR+, Se/(1-Sp)), negative likelihood ratio (LR-, (1-Se)/Sp), accuracy (Acc, percentage of subjects correctly classified), and Cohen's kappa index (kappa).

III. Results

A. CNN model optimization and training

The hyperparameters of the CNN architecture were the number of CNN sub-blocks (λ) and the dropout probability (p). Several experiments were conducted to find the optimum values of these hyperparameters. λ and p were varied from λ = 2 up to λ = 6 and p = 0 up to p = 0.5. For each λ-p pair, the corresponding CNN model was trained using the training set and kappa was computed in the validation set. He-normal initializer was used for weights initialization, mean absolute error was computed as the loss function, and the Adam algorithm with an initial learning rate of 0.001 was used for training the CNN models, with a decrease of the learning rate by a factor of 0.5 after 10 epochs and early stopping after 30 epochs of no improvement. Finally, λ = 5 and p = 0.3 were obtained as the optimum values, since this pair reached the highest kappa.

B. CNN model performance

Table II shows the confusion matrix of the CNN model in the test set. This model rightly classified 67.15% (165/246) of the subjects in the test set, with a kappa value of 0.31. Table III shows the diagnostic ability of the CNN model for the AHI cutoffs of 1 e/h, 5 e/h, and 10 e/h. These results are derived from the confusion matrix. Notice that the proposed CNN model reached high accuracies (higher than 90%) for the AHI cutoffs of 5 and 10 e/h, as well as high PPV (higher than 85%) and LR+ (higher than 25) for the three cutoffs.

IV. Discussion

In this preliminary study, we evaluated the usefulness of analyzing the SpO_2 signal by means of CNN to simplify the diagnosis of pediatric SAHS. To our knowledge, this is the first study applying deep learning techniques to establish pediatric SAHS and its severity.

Our proposal reached high diagnostic ability, with 74.8%, 90.7%, and 95.1% Acc for the AHI cutoffs of 1, 5, and 10 e/h. These AHI cutoffs are commonly employed to determine the presence of SAHS (AHI≥1 e/h), recommend surgical treatment (AHI≥5 e/h), and identify the children with a higher risk of suffering comorbidities and negative health consequences (AHI≥10 e/h) [19]. In this regard, and according to the Table II, 97.1% (198/204) of subjects predicted as no-SAHS (AHI<1 e/h) by the CNN model have an AHI<5 e/h. Furthermore, only 1 subject predicted as moderate-to-severe SAHS (AHI≥5 e/h) is no-SAHS (AHI<1 e/h). Finally, 92.9% (13/14) of subjects predicted as severe SAHS (AHI≥10 e/h) are severe SAHS. Our proposal also obtained high LR+: 28.6, 32.9, and 120 for 1, 5, and 10 e/h, respectively. A high LR+ is of the utmost importance for screening tests, since a LR+ above 10 is considered to provide strong evidence to confirm the disease [20]. According to these high LR+, our proposed

TABLE II. Confusion Matrix for the CNN Model in the Test Set

		Estimated			
		AHI<1	**1≤AHI<5**	**5≤AHI<10**	**AHI≥10**
Actual	**AHI<1**	**144**	1	1	0
	1≤AHI<5	54	**7**	1	1
	5≤AHI<10	6	6	**1**	0
	AHI≥10	0	8	3	**13**

TABLE III. Diagnostic Ability of the CNN Model in the Test Set for AHI Cutoffs= 1 e/h, 5 e/h, and 10 e/h

	Se (%)	**Sp** (%)	**Acc** (%)	**PPV** (%)	**NPV** (%)	**LR+**	**LR-**
AHI=1 e/h	40.0	98.6	74.8	95.2	70.6	28.6	0.61
AHI=5 e/h	46.0	98.6	90.7	85.0	91.2	32.9	0.55
AHI=10 e/h	54.2	99.6	95.1	92.9	95.3	120	0.46

CNN model would be especially useful to confirm the presence of positive subjects for the diagnosis of SAHS (AHI≥1), moderate-to-severe SAHS (AHI≥5), and severe SAHS (AHI≥10). On the other hand, there is a trend of the CNN to underestimate the AHI of the subjects with an AHI≥1 e/h, which results in a low sensitivity for the AHI thresholds of 1, 5, and 10 e/h.

Previous studies applied deep-learning techniques to analyze physiological signals in the context of adult SAHS diagnosis [12]–[15]. These studies reached accuracies in the range 85-96% for the detection of adult SAHS and its severity. From these studies, it is of note the work done by Nikkonen et al. [15], who applied a deep neural network to estimate AHI using 10-min SpO_2 data, reaching 90% accuracy to classify into the four adult SAHS severity degrees (AHI<5, 5≤AHI<15, 15≤AHI<30, and AHI≥30 e/h). Nonetheless, they used home polygraphy instead of PSG for the diagnosis of SAHS [15], which does not provide the total sleep time and does not quantify the hypopneas associated to arousals in the computation of AHI [3]. In contrast to these studies, our study applied CNN to estimate pediatric SAHS severity.

In the recent years, researchers have focused on the use of the oximetry signal for the screening of pediatric SAHS [5]–[10], [16]. Most of these studies have analyzed the oximetry signal by means of conventional signal processing and machine-learning techniques [5]–[10]. These works achieved varying diagnostic accuracies in the range 75-85% using the AHI cutoff of 1 e/h [5]–[7], [10], 81-85% for an AHI cutoff of 5 e/h [5]–[10], and 85-91% using the AHI cutoff of 10 e/h [5], [7], [9], [10]. Only one single preliminary study developed by our group applied deep-learning techniques to the oximetry signal to detect apneic events in pediatric SAHS patients [16], achieving 93.6% Acc (56.5% Se and 97.5% Sp) using 60-s SpO_2 segments [16]. Nonetheless, AHI estimation was not performed in [16], which is needed to establish SAHS diagnosis and severity. Our current results achieved a higher diagnostic performance for moderate-to-severe (AHI≥5 e/h) and severe SAHS (AHI≥10 e/h) groups with the use of a CNN, which automatically learns features from the SpO_2 recordings.

In spite of the promising results of our approach, some limitations should be considered. First, the number of subjects in no-SAHS group is high when compared to the other severity groups, especially the moderate and severe groups. This issue slightly contributes to the trend of the CNN to underestimate the AHI. In addition, further evaluation would be required to fairly compare our results with conventional signal processing approaches using the same dataset. Finally, future research efforts may be focused on assessing the effects of using other deep learning algorithms, as well as varying different hyperparameters, such as the activation function, the kernel size, or the loss function, in order to reduce the AHI underestimation of our proposal.

In summary, a CNN model fed with raw oximetry data achieved a promising diagnostic performance, outperforming state-of-the-art studies for AHI cutoffs of 5 and 10 e/h. This model also achieved strong evidence to confirm the presence of pediatric SAHS in its different severity degrees. This suggests that deep learning approaches could be potentially used to analyze the oximetry signal in the context of pediatric SAHS.

REFERENCES

[1] C. L. Marcus *et al.*, "Diagnosis and management of childhood obstructive sleep apnea syndrome.," *Pediatrics*, vol. 130, no. 3, pp. 576–84, 2012.

[2] L. Kheirandish-Gozal, "What is 'Abnormal' in pediatric sleep?," *Respir. Care*, vol. 55, no. 10, pp. 1366–1374, 2010.

[3] 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.

[4] G. M. Nixon, A. S. Kermack, G. M. Davis, J. J. Manoukian, A. Brown, and R. T. Brouillette, "Planning adenotonsillectomy in children with obstructive sleep apnea: the role of overnight oximetry," *Pediatrics*, vol. 113, no. 1, pp. e19–e25, 2004.

[5] R. Hornero *et al.*, "Nocturnal Oximetry-based Evaluation of Habitually Snoring Children.," *Am. J. Respir. Crit. Care Med.*, vol. 196, no. 12, pp. 1591–1598, 2017.

[6] D. Álvarez *et al.*, "Automated Screening of Children With Obstructive Sleep Apnea Using Nocturnal Oximetry : An Alternative to Respiratory Polygraphy in Unattended Settings," *J Clin Sleep Med*, vol. 13, no. 5, pp. 7–11, 2017.

[7] C. Tsai *et al.*, "Usefulness of desaturation index for the assessment of obstructive sleep apnea syndrome in children," *Int. J. Pediatr. Otorhinolaryngol.*, vol. 77, no. 8, pp. 1286–1290, 2013.

[8] A. Garde, P. Dehkordi, W. Karlen, D. Wensley, J. M. Ansermino, and G. A. Dumont, "Development of a screening tool for sleep disordered breathing in children using the phone oximeterTM," *PLoS One*, vol. 9, no. 11, 2014.

[9] F. Vaquerizo-Villar *et al.*, "Utility of bispectrum in the screening of pediatric sleep apnea-hypopnea syndrome using oximetry recordings," *Comput. Methods Programs Biomed.*, vol. 156, pp. 141–149, 2018.

[10] F. Vaquerizo-Villar *et al.*, "Detrended fluctuation analysis of the oximetry signal to assist in paediatric sleep apnoea-hypopnoea syndrome diagnosis," *Physiol. Meas.*, vol. 39, no. 11, p. 114006, 2018.

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

[12] B. Pourbabaee, M. H. Patterson, M. R. Patterson, and F. Benard, "SleepNet: automated sleep analysis via dense convolutional neural network using physiological time series," *Physiol. Meas.*, vol. 40, no. 8, p. 084005, 2019.

[13] S. Biswal, H. Sun, B. Goparaju, M. Brandon Westover, J. Sun, and M. T. Bianchi, "Expert-level sleep scoring with deep neural networks," *J. Am. Med. Informatics Assoc.*, vol. 25, no. 12, pp. 1643–1650, 2018.

[14] S. H. Choi *et al.*, "Real-time apnea-hypopnea event detection during sleep by convolutional neural networks," *Comput. Biol. Med.*, vol. 100, no. February, pp. 123–131, 2018.

[15] S. Nikkonen, I. O. Afara, T. Leppänen, and J. Töyräs, "Artificial neural network analysis of the oxygen saturation signal enables accurate diagnostics of sleep apnea," *Sci. Rep.*, vol. 9, no. 1, p. 13200, 2019.

[16] F. Vaquerizo-Villar *et al.*, "Convolutional Neural Networks to Detect Pediatric Apnea- Hypopnea Events from Oximetry," in *41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2019, pp. 3555–3558.

[17] 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.

[18] 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.

[19] G. D. Church, "The Role of Polysomnography in Diagnosing and Treating Obstructive Sleep Apnea in Pediatric Patients," *Curr. Probl. Pediatr. Adolesc. Health Care*, vol. 42, no. 1, pp. 22–25, 2012.

[20] J. J. Deeks, "Diagnostic tests 4: likelihood ratios," *Bmj*, vol. 329, no. 7458, pp. 168–169, 2004.