Reliability of Machine Learning to Diagnose Pediatric Obstructive Sleep Apnea: Systematic Review and Meta-Analysis

Gonzalo C. Gutiérrez-Tobal[1,2,*], PhD, Daniel Álvarez[1,2,3], PhD, Leila Kheirandish-Gozal[4], MD, MSc, Félix del Campo[1,2,3], MD, PhD, David Gozal[4], MD, MBA, PhD (Hon), Roberto Hornero[1,2,3], PhD

[1]Biomedical Engineering Group, Universidad de Valladolid, Valladolid, Spain

[2]Centro de Investigación Biomédica en Red en Bioingeniería, Biomateriales y Nanomedicina, (CIBER-BBN), Spain

[3]Pneumology Department, Río Hortega University Hospital, Valladolid, Spain

[4]Department of Child Health, and the Child Health Research Institute, The University of Missouri School of Medicine, Columbia, Missouri, USA

[*]**corresponding author**:
e-mail: gonzalo.gutierrez@gib.tel.uva.es
Tlf: +34 983 18 47 13
Mailing address:
Escuela Técnica Superior de Ingenieros de Telecomunicación
Universidad de Valladolid
Campus Miguel Delibes
Paseo Belén 15
47011 Valladolid
España

Keywords: Pediatrics, Sleep apnea, Machine learning, Review, Meta-Analysis

Abbreviated title: Machine Learning and Pediatric Sleep Apnea Diagnosis

Funding: This work was supported by 'Ministerio de Ciencia, Innovación y Universidades - Agencia Estatal de Investigación' and 'European Regional Development Fund (FEDER)' under projects DPI2017-84280-R and RTC-2017-6516-1, by "European Commission" and "FEDER" under project 'Análisis y correlación entre la epigenética y la actividad cerebral para evaluar el riesgo de migraña crónica y episódica en mujeres' ('Cooperation Programme Interreg V-A Spain-Portugal POCTEP 2014–2020'), by Sociedad Española de Neumología y Cirugía Torácica (SEPAR) under project 649/2018, by Sociedad Española de Sueño (SES) under project "Beca de Investigación SES 2019", and by 'Centro de Investigación Biomédica en Red en Bioingeniería, Biomateriales y Nanomedicina, (CIBER-BBN), Spain' through 'Instituto de Salud Carlos III' co-funded with FEDER funds. D. Álvarez is supported by a "Ramón y Cajal" grant (RYC2019-028566-I) from the 'Ministerio de Ciencia e Innovación - Agencia Estatal de Investigación' co-funded by the European Social Fund. L. Kheirandish-Gozal is supported by NIH grant HL130984 and by the Children's Miracle Network Endowed Professorship, and D. Gozal is supported by NIH grants HL140548, and AG061824.Both LKG and DG are supported by the Leda J. Sears Foundation, and a Tier 2 grant from the University of Missouri.

Abstract:

Background: Machine-learning approaches have enabled promising results in efforts to simplify the diagnosis of pediatric obstructive sleep apnea (OSA). A comprehensive review and analysis of such studies increase the confidence level of practitioners and healthcare providers in the implementation of these methodologies in clinical practice.

Objective: To assess the reliability of machine-learning- based methods to detect pediatric OSA.

Data Sources: Two researchers conducted an electronic search on the Web of Science and Scopus using term, and studies were reviewed along with their bibliographic references.

Eligibility Criteria: Articles or reviews (year 2000 onwards) that applied machine learning to detect pediatric OSA; reported data included information enabling derivation of true positive, false negative, true negative, and false positive cases; polysomnography served as diagnostic standard.

Appraisal and Synthesis Methods: Pooled sensitivities and specificities were computed for three apnea-hypopnea index (AHI) thresholds: 1 event/hour (e/h), 5 e/h, and 10 e/h. Random-effect models were assumed. Summary receiver-operating characteristics (SROC) analyses were also conducted. Heterogeneity (I^2) was evaluated, and publication bias was corrected (trim and fill).

Results: Nineteen studies were finally retained, involving 4,767 different pediatric sleep studies. Machine learning improved diagnostic performance as OSA severity criteria increased reaching optimal values for AHI=10 e/h (0.652 sensitivity; 0.931 specificity; and 0.940 area under the SROC curve). Publication bias correction had minor effect on summary statistics, but high heterogeneity was observed among the studies.

Conclusions: Machine learning can reliably detect severe OSA. However, further steps are needed to improve diagnostic performance for less severe pediatric OSA, and thus increase the confidence levels when using these approaches.

1. INTRODUCTION

Pediatric obstructive sleep apnea (OSA) has been the focus of increasing scientific interest during the last several decades. Since it was initially described by Guilleminault and colleagues[1], the cumulative evidence regarding OSA high prevalence[2], sub-optimal diagnostic rates[3], and potential long-term cardiovascular, neurocognitive, and behavioral associated morbidities[2], have driven substantial research efforts in two major directions, namely finding effective treatments[4] and enabling simplified objective and less costly diagnostic methods[3]. In this respect, whereas both surgical and non-surgical interventions have successfully been developed and validated[4,5], the gold standard for reaching the diagnosis of OSA in children remains overnight polysomnography (PSG), and due to its complexity, costs and access delay problems, it has become obvious that PSG is far from being the ideal diagnostic solution for habitually snoring children at risk for OSA[6].

Several approaches have been proposed to overcome such PSG limitations and simplify the diagnostic methodology. For example, sleep-related questionaries[7,8], symptoms-based scores[9], and automated single-channel recordings[10–12] have been frequently assessed but have not yielded the accuracy thresholds that would be acceptable for widespread adoption[13]. In contrast, machine-learning techniques have elicited increasingly growing interest due to their prominent impact in a wide range of healthcare processes[14]. Indeed, promising results have also been reported in studies involving machine-learning approaches that facilitate automated OSA diagnosis using pediatric recordings[15–33]. However, a substantial level of skepticism remains among the sleep specialists and clinical practitioners alike, regarding the clinical use of these automatic tools[34].

There is little doubt that availability of a reliable, automated, and simplified alternative to PSG would improve OSA diagnosis in children from several different perspectives. On the one hand, less need for equipment requirements, particularly those related to the

number of sensors, would improve patient comfort. It would also open the door to home testing,[6] and consequently, reduce the long waiting lists currently in place around the world for a child to undergo a PSG. On the other hand, an automated methodology would decrease the time and effort by sleep specialists spent on the visual inspection of PSG-derived overnight physiological signals[6], thus accelerating the diagnostic process. Taken together, these advantages would facilitate earlier diagnosis and access to treatment for the affected children.

Based on the aforementioned considerations, we conducted a systematic review and meta-analysis to shed light on the reliability of machine-learning studies focused on the diagnosis of pediatric OSA. Accordingly, we have summarized the main methodological steps undertaken to systematically select the extant set of published studies and compare them with current standards and performance expectations in the field. To this effect, we assessed the type of machine-learning methods used, the validation strategy followed, and the explainability of the models obtained. Moreover, we gathered the pooled sensitivity and specificity statistics from the studies in a meta-analysis, thus providing a more accurate perspective on the clinical usefulness of machine-learning approaches in the context of the diagnosis of pediatric OSA.

2. METHODS

2.1. Eligibility criteria

Table 1 summarizes the eligibility criteria used to include studies in the systematic review and meta-analysis. These criteria were selected by consensus from all the authors. Only articles and reviews from the year 2000 onwards (until February 1st, 2021) and written in English were considered. This span embraces the boom experimented by machine-learning methods applied to health problems. Documents in both 'published' and 'early access' stages were accepted, thus accounting for the latest findings on machine-learning applied to pediatric OSA. They were required to be focused on pediatric OSA diagnosis and also that machine-learning methods were used to either directly derive an automatic diagnosis or detect the respiratory events (apneas and hypopneas) that are clinically used to reach a diagnosis. However, only those studies reporting performance metrics from automatic subject-based diagnosis were considered. Here, the term 'machine learning' was adopted in the wide sense, i.e., any classification or regression automatic method that requires a training process to derive a predictive model potentially using multiple variables. In this way, we can analyze the use and performance of the simplest models if needed. Moreover, the studies were required to report sufficient data to enable extraction or computation of the number of true positive, false negative, true negative, and false positive subjects for at least one specifically defined apnea-hypopnea index (AHI) threshold. Importantly, the performance of the methods was required to be reported in comparison with the overnight PSG-derived diagnosis.

2.2. Information sources and bibliography search

The advanced search functionality of the Web of Science (WoS) and Scopus electronic databases was used to conduct the initial literature screening. Table 2 shows the query

strings included within the searching boxes for each of the databases. The eligibility criteria related terms were searched in the title, abstract, and keywords. These terms were chosen by agreement of all the authors of the study to include suitable vocabulary on both machine learning and pediatric OSA. Those terms with different spelling in British and American English were duplicated to embrace both options. The searches were conducted by two independent researchers (GCG-T and DA), who also conducted subsequent reviews of the studies found using both electronic databases. These studies were assessed for duplicates, as well as for meeting the eligibility criteria shown in Table 1. Each researcher proposed a selection of studies to be included in the systematic review/meta-analysis. Discordances were resolved by consensus. As a secondary data source, those papers referenced in this initial set of studies were also reviewed by each researcher to check for eligibility and to form the final set. This was also obtained by consensus after this last step.

2.3. Data collection

Table 3 shows the data extracted from each of the studies selected after completion of the bibliographic search. The studies are gathered in four main categories: general information from the studies, applied methods, population, and meta-analysis data. General information was directly obtained from the search in the electronic databases (WoS and Scopus) and automatically exported to a spreadsheet. The remaining data were manually introduced in the same spreadsheet after careful review of each of the studies.

2.4. Meta-analysis

The *mada* and *meta* R packages[35,36] were used to perform meta-analysis of diagnostic performance-based studies. Total effect size for univariate sensitivity and specificity

diagnostic metrics was estimated using a random-effects model with a logit transformation of the input data, i.e., TP, FN, TN, and FP obtained from each study. The Higgins' I^2 and the *p*-value of the Cochrane Q statistics were computed to characterize heterogeneity. In addition, a bivariate diagnostic random-effects meta-analysis was performed using summarized receiver-operating characteristics (SROC) curves to account for the interdependence of sensitivity and specificity. Funnel plots were used to assess the influence of publication bias. The trim and fill method[37] from the metaphor R package was subsequently applied to adjust for this source of bias and correct the effect sizes derived from the forest plots.

3. RESULTS

3.1. Study selection and characteristics

Figure 1 displays a flowchart with the number of studies selected after each step of the bibliographic search. Sixty-three documents were found after the electronic automatic search, either using WoS (47) or Scopus (16). A total of 10 duplicated studies were identified within the search results from both databases, and duplicates were removed from subsequent analyses. The remaining 53 documents were assessed for meeting eligibility criteria, with only 17 studies deems as eligible. Up to 25 references from these studies were also reviewed, and 2 more eligible studies were added to the final set.

Table 4 summarizes the main data obtained from the 19 selected studies. All of them were original articles. For those studies reporting results from more than one machine-learning method, and without any other methodological difference, only the one highlighted by the authors as the top performing method was considered. In contrast, results derived using different data sources (e.g., different biomedical signals) but using equal machine-learning methods were all included even though they were reported in the same study. The following subsections are devoted to the data analysis included in Table 4.

3.2. Population characteristics

The studies reported a cumulative sample size of 11,200 children. Among them, 7,891 were used to obtain the metrics to evaluate the diagnostic performance of their models and are included in Table 4. However, we were able to identify that some studies analyzed totally or partially the same subjects, and that there were only 4,767 unique subjects. In addition, individual sample sizes varied greatly between studies. Overall, the studies covered the entire non-adult age range. In particular, the two studies with the largest

databases (Hornero et al.[16] 3,602 and Vaquerizo-villar et al.[24] 935 subjects) covered the ranges 2-18 and 0-18 years, respectively. However, Calderón et al. covered only children between 5-10 years[23]. There was consistency among the studies concerning gender, showing a higher prevalence of male subjects in the range of 52.0% to 65.3%, commensurate with the previous literature showing no differences in pediatric OSA prevalence between males and females or slightly increased prevalence in male subjects[2]. This same pattern appears was replicated in the selected studies herein, whereby 11 of 19 papers reported less than 60.0% of male prevalence and the remaining 8 reported 60.0% to 65.3% males. Finally, 18 out of the 19 studies involved symptomatic children, that is, showing high pre-test probability of suffering from OSA. Only the study by Skotko et al. did not recruit subjects on the basis of related symptoms[22]. However, their study focused on predicting OSA within a cohort of subjects suffering from Down syndrome, which is a group at high risk for OSA due to craniofacial and neuromuscular tone abnormalities.

3.3. Data used to train the machine-learning models

Despite not including any search term related to biomedical signal processing in the query strings of Table 2, most of the selected studies used data from overnight physiological signals to train and develop the machine-learning models. Peripheral blood oxygen saturation signal (SpO_2) was the predominant signal recorded, with 16 out of 19 studies using it alone (9)[16,19,20,23–26,29,30] or associated with another type of physiological data (8)[17,18,27–29,31–33]. Airflow signal (AF) was used in 3 studies[17,28,29], pulse rate variability (PRV) in 2[31,32], and ECG and actigraphy in one[18,21] each. Non-biomedical signal data (clinical variables, anthropometrics, and demographics) were also included in the analyses corresponding to 5 studies[15,18,22,27,33]. However, only 3 of these studies reported results of models trained exclusively with this information[15,18,22].

3.4. Machine-learning methods

A heterogeneous range of machine-learning methods was used among the selected studies. Logistic regression was the most frequent (6 studies)[17,19,20,25,32,33], whereas multi-layer perceptron (MLP) artificial neural network was used in 4 studies[16,26–28], and support vector machine (SVM)[18,30] and ensemble-learning adaptive boosting (AdaBoost)[23,29] in two each. Multivariate linear regression (MLR)[15], logic learning machine (LLM)[22], linear discriminant analysis (LDA)[31], and quadratic discriminant analysis (QDA)[21] appeared in 1 each. Interestingly, only one study followed a deep learning approach (convolutional neural network, CNN)[24], which has demonstrated superior performances in health-related problems in the last several years[38].

Efforts to explain the machine-learning predictions were also evaluated. Noticeably, only 5 out of the 19 studies reported quantitative data indicating further analysis to try to explain the decisions of their models[15,18,20,22,32]. However, no studies attempted to explain decisions of the most complex models or use the latest approaches on 'explainable artificial intelligence'[39].

3.5. Validation strategies

All the studies compared their results against the apnea-hypopnea index (AHI) derived from the full PSG, i.e., the standard method used to diagnose pediatric OSA[40,41]. AHI values were used to establish four severity categories of OSA (no OSA: AHI < 1e/h; mild: 1 e/h ≤ AHI < 5 e/h; moderate: 5 e/h ≤ AHI < 10 e/h; and severe: 10 e/h ≤ AHI)[42]. Consequently, most of the studies (n=17) reported results for one or more of these AHI thresholds. However, several studies showed results from other less frequently used thresholds, such as 2 e/h and 3 e/h, and 8 studies only reported data for a single AHI threshold[15,17,19,21,23,30,31,33].

Several approaches were used to validate the machine-learning methods. Two (Training/Test) or three (Training/Validation/Test) subgroups were used depending on whether the machine-learning method required hyperparameters to be tuned. These strategies were applied directly, i.e., with real 2 or 3 subgroups, or including subgroup simulation methods such as leave-one-out cross-validation (loo-cv), bootstrapping, or k-fold cross-validation (k-fold-cv). Only three exceptions were found to the use of a third real or simulated subgroup due to hyperparameter tunning. Two of them were the studies of Calderon et al.[23] and Bertoni et al.[18], whose corresponding AdaBoost and SVM models usually require hyperparameter tuning to reach an optimum performance (e.g., penalty parameters such as learning rate or C, respectively). In contrast, Xu et al.[26] applied the same exact MLP than the one previously internally validated by Hornero et al.[16].

3.6. Meta-analysis: Forest plots and summary ROC curves

True positive, false negative, true negative, and false positive subjects obtained from the studies were included in the meta-analysis. Only data from the AHI thresholds 1 e/h, 5 e/h, and 10 e/h were used as there was insufficient number of studies reporting data on 2 e/h and 3 e/h. This action resulted in 17 out of 19 studies being included in the meta-analysis.

Figures 3-5 show the forest plots corresponding to the analyses of sensitivity and specificity for the 3 above-mentioned AHI thresholds. Individual and composite results are provided for each statistic, including 95% confidence intervals. Results for studies involving and not involving SpO_2 are provided separately in 2 subgroups as well. Heterogeneity measures are also displayed. As can be observed, pooled sensitivity decreases as AHI threshold increases, showing values of 0.921 [0.866; 0.955], 0.762 [0.722; 0.798], and 0.682 [0.564; 0.780] for 1 e/h, 5 e/h, and 10 e/h respectively. An

increasing opposite tendency is displayed for pooled specificity, which shows 0.386 [0.232; 0.566], 0.851 [0.765; 0.909], and 0.958 [0.934; 0.973], respectively. In all cases, heterogeneity is significantly high according to the *p*-value of Cochrane Q (< 0.01) and Higgins' I^2 values, which ranges 72%-95% thus justifying the choice of the random-effects model to conduct the meta-analysis. All the pooled sensitivity and specificity values are higher when considering only those results involving SpO_2 data except in the cases of the sensitivity for the AHI thresholds 1 e/h and 5 e/h. In these 2 instances, the statistics were slightly higher for the results not involving SpO_2 at the cost of notably wider 95% confidence intervals. Still, the pooled overall performance when using SpO_2 data is clearly higher for both moderate and severe pediatric OSA. Similarly, heterogeneity decreases in all statistics when considering only results involving SpO_2 data (64%-93%). As anticipated by the pooled metrics, the top performance methods were reported for moderate and severe OSA and in studies involving SpO_2 data. Accordingly, the deep learning approach (a convolutional neural network) using only SpO_2 data, proposed by Vaquerizo-Villar et al.[24], reached the highest overall figures for moderate OSA in their 935 test subjects (73.4% sensitivity, 94.3% specificity, 88.3% accuracy), with the proposal by Garde et al.[31] (linear discriminant analysis on SpO_2 + pulse rate variability data) reaching slightly lower overall values in their 146 subjects (89.3% sensitivity, 83.3% specificity, 85.6% accuracy). On the other hand, Bertoni et al.[18] proposed a support vector machine method applied to clinical, actigraphy, and SpO_2 data that reached the highest performance for severe OSA in their 187 subjects (93.9% sensitivity, 100.0% specificity, 98.4% accuracy). Similarly, the above-mentiuoned proposal by Vaquerizo-Villar et al.[24] reached the next highest results (76.6% sensitivity, 97.3% specificity, 93.9% accuracy).

Figure 5 displays the SROC curves for bivariate analysis, which account for interdependencies between sensitivity and specificity at each AHI threshold. The shape of the curves agrees with the values of the pooled sensitivity/specificity pairs shown above for each case. Moreover, an increase in the area under the SROC curves (AUC) is observed as the AHI threshold is higher, reaching values of 0.791, 0.826, and 0.940 for 1 e/h, 5 e/h, and 10 e/h, respectively.

3.7. Publication bias

Figure 6 shows the funnel plots of sensitivities (left column) and specificities (right column) for each of the three AHI thresholds (1 e/h, 5 e/h, and 10 e/h from upper to lower rows). Filled black dots represent data from real studies, whereas blank dots represent simulated studies added by means of the trim and fill method to correct for possible publication bias. Accordingly, for AHI = 1 e/h, 5 studies were added to sensitivity and 2 to specificity; for AHI = 5 e/h, 6 studies were added to sensitivity and none to specificity; and for AHI = 10 e/h, 1 study was added to sensitivity and 5 to specificity. All the added studies represent proportions in the range 0%-26% among the sum of real and simulated results for each case.

Table 5 summarizes the corrected pooled sensitivities, specificities, and Higgins' I^2 values for each AHI threshold when considering the added studies. The number of these and the total number of results are also shown for each case. All the sensitivity and specificity values from all AHI thresholds are reduced compared to those originally reported in forest plots, except for the specificity in 1 e/h that increased by 11 decimal points. The decreased decimal points of the remaining results are in the range 2 to 7. Moreover, all the heterogeneity values were slightly higher than the original.

4. DISCUSSION

In this work, we conducted a systematic review and meta-analysis on the reliability of machine-learning methods to diagnose pediatric OSA. Nineteen studies spanning the period between 2004-2021 were included and involved 4,767 unique pediatric subjects. We found decreasing pooled sensitivities and increasing pooled specificities as OSA severity worsened, thus reflecting the well-known threshold effect of diagnostic test accuracy meta-analyses[43]. Very high pooled specificity (0.931 [0.894; 0.955]) was reached for the severe OSA AHI threshold (10 e/h), which was accompanied by a moderate sensitivity (0.652 [0.530; 0.758]). Concurrently, this moderate sensitivity and the very low number of false positives reflected in the specificity value, reveals high reliability when machine-learning methods assign a subject to the severe OSA group. This result is also supported by the SROC analysis conducted, in which the area under the curve reached 0.940 when evaluating the same severity degree. Moreover, if only results involving SpO_2 data are considered, both sensitivity and specificity of severe OSA notably rise (0.745 and 0.964, respectively), and improve the diagnostic accuracy of moderate OSA (AHI = 5 e/h) to nearly the same reliability level (0.751 sensitivity and 0.895 specificity). These are meaningful and highly encouraging findings since moderate to severe children are those who are at higher risk of cardiovascular and neurocognitive morbidity[44,45] and they benefit the most of an early diagnosis and access to treatment[46]. However, important efforts are still needed to improve the performance of these approaches to encompass less severe disease criteria, as well as enhance the level of confidence of healthcare providers and reduce their reluctance to implement the use of machine learning-derived tools in clinical practice.

4.1. Risk of within-studies biases

Several potential biases were detected among the studies included in this work. However, the roots of many of these biases reside in the sampled population. Machine-learning methods have increased their data requirements as a natural consequence of the evolution of the mathematical techniques[47]. At the same time, they have also increased their performance[47]. However, insufficient sample sizes have forced several of the published studies to use simple and relatively older and less performant machine-learning methods such as logistic regression (n=6 studies), even though these studies were published in the last 5 years (2015-2019). Noticeably, only one study used the much more powerful deep-learning approach[38]. Eventually, the use of outdated methods may be hindering the progression and utility of machine-learning algorithms in their ability to reach higher reliability and consequently adoption into the clinical realm. Of note, a similar problem concerns the methodology used to validate the machine-learning models. Scarcity of data greatly affects the number of subgroups in which the sample can be split, which should be ideally related to the degrees of freedom of the models and, in practice, being a minimum of three (Training/Validation/Test) if the machine-learning method requires hyperparameter tuning[48]. Many studies included in this review, however, needed to use techniques such as leave-one-out cross-validation, k-fold cross-validation or bootstrapping to simulate additional groups, thus biasing their results, and consequently, potentially affecting the accuracy of the estimates of disease severity. We would expect that involving more subjects would lead to the use of more precise machine-learning techniques as well as proper validation strategies, which could increase the performance and decrease the heterogeneity shown in this study. Another potential bias relates to the cohorts used, since several studies included the same subjects in more than one study (of 7,891 subjects in the studies, only 4,767 were not involved in more than one study). This

is due to the fact that two studies from different research groups[23,24] used the Children Adenotonsillectomy Trial (CHAT) public database[49]. Second, 14 out of the 19 studies shared at least one of their authors, thus potentiating the above-mentioned overlap. Although duplicates may bias the results, we surmise that 4,767 unique subjects provide sufficient statistical power to reach valid conclusions.

We would like to point out two additional sources of bias. First, none of the studies included a control group of healthy children from the general population. All but one study involved children manifesting OSA-related symptoms regardless of whether they were ultimately diagnosed as suffering from OSA. Also, another study involved a cohort of children suffering from Down syndrome. The inclusion of control subjects might affect the performance of machine-learning methods. However, it would be expected that the possible misclassification were focused on more mild OSA, thus not affecting the conclusions about the reliability to diagnose more severe OSA. Secondly, the prevalence of OSA among male and female subjects is still under discussion[2]. However, all the studies reported higher proportion of male children, and some of them remarkably higher. These two issues need to be addressed in future studies to further assess reliability of machine-learning methods.

Finally, we should remark that only a minority of studies reported further analyses to try to explain the decisions taken by their associated machine-learning models. Similarly, these were conducted only when simpler decision algorithms were adopted, and did not follow the latest standards on 'explainable artificial intelligence', such as the model-agnostic method Shapley Additive Explanations (SHAP)[39], which unifies most of its precursors. Although this issue may not bias the performance assessment of our analyses, we think that not explaining the principles operating in the context of automated processes

contributes to the traditional 'black box' perception of machine learning[50] and has an important negative impact on the confidence of healthcare providers.

4.2. Heterogeneity and risk of biases across studies

Several sources of heterogeneity among the studies were detected and may explain the high Higgings' I^2 values reached. We have detected 10 different machine-learning algorithms among the studies reviewed, and several of these have been used to implement different approaches (e.g., classification *vs*. regression; binary classification *vs*. multiclass classification). Moreover, the physiological information used to train and obtain the machine-learning models also varied among studies. SpO_2 data predominated, but there were also data from AF, actigraphy, ECG, clinical variables, anthropometrics, and demographics. The effect of these different approaches on I^2 was shown explicitly when comparing the noticeably lower heterogeneity degree reached by the studies involving SpO_2 data with those not involving it. Other potential heterogeneity precursors have been already mentioned. Issues such as the different sample sizes, validation strategies, and sex distribution may have influenced as well. However, according to the size of the 95% confidence intervals reached in our analyses, heterogeneity appears to have less effect in the results from moderate and severe OSA.

Two potential across-studies risk of biases have been also identified. First, we have already mentioned the threshold effect described in the literature for diagnostic accuracy testing using meta-analyses[43]. Univariate sensitivity/specificity analyses are common approaches, but bivariate SROC analysis should be also provided to conduct a more complete assessment of the test under study[51]. Cautions are needed with SROC interpretation when studies are not homogeneous. However, the derived area under the curve statistic has been shown to be a useful upper bound approximation even in the

presence of heterogeneity[51]. Finally, we assessed publication bias by means of the trim and fill method supported by funnel plots. In this regard, the small proportion of simulated results added, along with the minor changes in the pooled sensitivity and specificity values produced, lead us to think that the bibliography, as include, reflects a reliable sample of the results under study.

4.3. Other limitations

Other limitations need to be considered in our study. First, the eligibility criteria for the studies were chosen using consensus by the authors, which include both machine-learning engineers and sleep physicians. We followed the PICOS recommendation (participants, interventions, comparisons, outcomes, and study design)[52]. However, it is possible that other researchers may have selected different eligibility criteria. Similarly, the terms used to conduct the initial search, as well as the data collected from the studies agreed with these criteria and the purpose of our study. However, machine-learning terminology is not always homogeneous across the related fields, and it is possible that scientists from other research areas could have used different nomenclatures and collected different data. Moreover, we have used two different electronic databases to conduct the initial search (WoS and Scopus). Although these are two of the largest electronic databases, it is possible that other bibliographic sources may index more suitable studies. Finally, documents not written in English were not included.

5. CONCLUSIONS AND RECOMMENDATIONS

We found a high reliability of machine-learning methods to automatically diagnose severe pediatric OSA, thus benefiting those children at higher risk of suffering comorbidities. Pooled univariate and bivariate statistics derived from a representative sample of results strongly supported this conclusion. We have also shown that, unsurprisingly, the performance of the machine-learning models is dependent on the source of the data used to obtain them, and that overnight SpO_2 information increases its reliability. Thirdly, we consider deep-learning approaches as more advanced options with a greater potential for improved performance. However, we also identified some problems that may preclude the implementation of these techniques in real environments. In order to try to solve such constraints, we propose the following recommendations:

1. *Future studies should address the size and characteristics of the cohort*. Control groups of asymptomatic healthy children should be included in the training, validation, and test stages of the machine-learning methods. Additionally, the number of subjects involved should be large enough to let researchers use the latest data-demanding machine-learning approaches, as well as properly evaluate them. Larger databases would also help cope with different phenotypes in pediatric OSA[53].
2. *Inclusion of deep-learning techniques is needed*. Deep-learning architectures and algorithms should be considered in future studies. Currently, there exists a range of these methods that are showing remarkable performances in several healthcare issues[38]. They are particularly useful in detecting hidden patterns from temporary or spatially related data, such as biomedical signals or medical images.[38] Therefore, these techniques may improve the results reported in this meta-analysis, provided that sufficient high quality data are available to implement them.

3. *Efforts should be made to explain the outcomes of the machine-learning methods*. Beyond increasing the performance of the machine-learning methods, addressing the 'black box' issue will be crucial to boost the confidence and implementation of these diagnostic approaches in healthcare settings. A new computer science field is growing fast under the term 'explainable artificial intelligence' as a response to the demand of explainable models from science, industry, and administration due to the need to justify decisions taken based on automated algorithms[50]. New developments in these techniques not only allow for understanding automated decisions but also facilitate discovery of new knowledge in the fields of application[50], especially when combined with deep-learning methods.
4. *Sources of OSA related information other than SpO_2 should be further assessed.* Although SpO_2 has demonstrated superiority as source of patient relevant information in this study, there is a clear imbalance with the studies using other types of physiological data. Future studies should therefore explore other relevant measures and examine whether combinatorial datasets lead to further accuracy enhancements.

ACKNOWLEDGEMENTS
The authors declare no conflict of interest.

REFERENCES

1. Guilleminault C, Eldridge FL, Simmons FB, Dement WC. Sleep apnea in eight children. Pediatrics 1976.
2. Marcus CL, Brooks LJ, Ward SD, Draper KA, Gozal D, Halbower AC, Jones J, Lehmann C, Schechter MS, Sheldon S, et al. Diagnosis and management of childhood obstructive sleep apnea syndrome. Pediatrics 2012;130(3):e714--e755.
3. Kheirandish-Gozal L. What is "Abnormal" in pediatric sleep? Respir Care 2010.
4. Tan H-L, Alonso Alvarez ML, Tsaoussoglou M, Weber S, Kaditis AG. When and why to treat the child who snores? Pediatr Pulmonol 2017;52(3):399–412.
5. Gozal D, Tan H-L, Kheirandish-Gozal L. Treatment of Obstructive Sleep Apnea in Children: Handling the Unknown with Precision. J Clin Med 2020.
6. Tan HL, Kheirandish-Gozal L, Gozal D. Pediatric home sleep apnea testing slowly getting there! Chest 2015.
7. Spruyt K, Gozal D. Screening of pediatric sleep-disordered breathing: A proposed unbiased discriminative set of questions using clinical severity scales. Chest 2012.
8. Kadmon G, Shapiro CM, Chung SA, Gozal D. Validation of a pediatric obstructive sleep apnea screening tool. Int J Pediatr Otorhinolaryngol 2013.
9. Chang L, Wu J, Cao L. Combination of symptoms and oxygen desaturation index in predicting childhood obstructive sleep apnea. Int J Pediatr Otorhinolaryngol 2013.
10. Gil E, Bailón R, Vergara JM, Laguna P. PTT variability for discrimination of sleep apnea related decreases in the amplitude fluctuations of PPG signal in children. IEEE Trans Biomed Eng 2010.
11. Lázaro J, Gil E, Vergara JM, Laguna P. Pulse rate variability analysis for discrimination of sleep-apnea-related decreases in the amplitude fluctuations of pulse photoplethysmographic signal in children. IEEE J Biomed Heal Informatics 2014.
12. Nixon GM, Kermack AS, Davis GM, Manoukian JJ, Brown KA, Brouillette RT. Planning adenotonsillectomy in children with obstructive sleep apnea: the role of overnight oximetry. Pediatrics 2004.
13. Bertoni D, Isaiah A. Towards Patient-centered Diagnosis of Pediatric Obstructive Sleep Apnea—A Review of Biomedical Engineering Strategies. Expert Rev Med Devices 2019.
14. Obermeyer Z, Emanuel EJ. Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. N Engl J Med 2016.
15. Wu D, Li X, Guo X, Qin J, Li S. A simple diagnostic scale based on the analysis and screening of clinical parameters in paediatric obstructive sleep apnoea hypopnea syndrome. J Laryngol Otol 2017.
16. Hornero R, Kheirandish-Gozal L, Gutiérrez-Tobal GC, Philby MF, Alonso-Álvarez ML, Alvarez D, Dayyat EA, Xu Z, Huang YS, Kakazu MT, et al. Nocturnal oximetry-based evaluation of habitually snoring children. Am J Respir Crit Care Med 2017;196(12):1591–1598.
17. Gutiérrez-Tobal GC, Alonso-Álvarez ML, Álvarez D, Del Campo F, Terán-Santos J, Hornero R. Diagnosis of pediatric obstructive sleep apnea: Preliminary findings using automatic analysis of airflow and oximetry recordings obtained at patients' home. Biomed Signal Process Control 2015;18:401–407.
18. Bertoni D, Sterni LM, Pereira KD, Das G, Isaiah A. Predicting polysomnographic severity thresholds in children using machine learning. Pediatr Res 2020.
19. Crespo A, Álvarez D, Gutiérrez-Tobal GC, Vaquerizo-Villar F, Barroso-García V, Alonso-álvarez ML, Terán-Santos J, Hornero R, del Campo F. Multiscale entropy analysis of unattended oximetric recordings to assist in the screening of paediatric sleep apnoea at home. Entropy 2017;19(6):284.

20. Crespo A, Álvarez D, Kheirandish-Gozal L, Gutiérrez-Tobal GC, Cerezo-Hernández A, Gozal D, Hornero R, del Campo F. Assessment of oximetry-based statistical classifiers as simplified screening tools in the management of childhood obstructive sleep apnea. Sleep Breath 2018;22(4):1063–1073.
21. Shouldice RB, O'Brien LM, O'Brien C, De Chazal P, Gozal D, Heneghan C. Detection of obstructive sleep apnea in pediatric subjects using surface lead electrocardiogram features. Sleep 2004.
22. Skotko BG, Macklin EA, Muselli M, Voelz L, McDonough ME, Davidson E, Allareddy V, Jayaratne YSN, Bruun R, Ching N, et al. A predictive model for obstructive sleep apnea and Down syndrome. Am J Med Genet Part A 2017.
23. Calderón JM, Álvarez-Pitti J, Cuenca I, Ponce F, Redon P. Development of a minimally invasive screening tool to identify obese Pediatric population at risk of obstructive sleep Apnea/Hypopnea syndrome. Bioengineering 2020.
24. Vaquerizo-Villar F, Alvarez D, Kheirandish-Gozal L, Gutierrez-Tobal GC, Barroso-Garcia V, Santamaria-Vazquez E, Del Campo F, Gozal D, Hornero R. A convolutional neural network architecture to enhance oximetry ability to diagnose pediatric obstructive sleep apnea. IEEE J Biomed Heal Informatics 2021.
25. Álvarez D, Alonso-Álvarez ML, Gutiérrez-Tobal GC, Crespo A, Kheirandish-Gozal L, Hornero R, Gozal D, TerÁn-Santos J, Del Campo F. Automated screening of children with obstructive sleep apnea using nocturnal oximetry: An alternative to respiratory polygraphy in unattended settings. J Clin Sleep Med 2017;13(5):693–702.
26. Xu Z, Gutiérrez-Tobal GC, Wu Y, Kheirandish-Gozal L, Ni X, Hornero R, Gozal D. Cloud algorithm-driven oximetry-based diagnosis of obstructive sleep apnoea in symptomatic habitually snoring children. Eur Respir J 2019.
27. Vaquerizo-Villar F, Álvarez D, Kheirandish-Gozal L, Gutiérrez-Tobal GC, Barroso-García V, Crespo A, del Campo F, Gozal D, Hornero R. Utility of bispectrum in the screening of pediatric sleep apnea-hypopnea syndrome using oximetry recordings. Comput Methods Programs Biomed 2018;156:141–149.
28. Barroso-García V, Gutiérrez-Tobal GC, Kheirandish-Gozal L, Álvarez D, Vaquerizo-Villar F, Núñez P, del Campo F, Gozal D, Hornero R. Usefulness of recurrence plots from airflow recordings to aid in paediatric sleep apnoea diagnosis. Comput Methods Programs Biomed 2020.
29. Jiménez-García J, Gutiérrez-Tobal GC, García M, Kheirandish-Gozal L, Martín-Montero A, álvarez D, del Campo F, Gozal D, Hornero R. Assessment of airflow and oximetry signals to detect pediatric sleep Apnea-Hypopnea syndrome using AdaBoost. Entropy 2020.
30. Vaquerizo-Villar F, Álvarez D, Kheirandish-Gozal L, Gutiérrez-Tobal GC, Barroso-García V, Crespo A, del Campo F, Gozal D, Hornero R. Wavelet analysis of oximetry recordings to assist in the automated detection of moderate-to-severe pediatric sleep apnea-hypopnea syndrome. PLoS One 2018.
31. Garde A, Dehkordi P, Karlen W, Wensley D, Ansermino JM, Dumont GA. Development of a screening tool for sleep disordered breathing in children using the phone oximeterTM. PLoS One 2014;9(11).
32. Garde A, Hoppenbrouwer X, Dehkordi P, Zhou G, Rollinson AU, Wensley D, Dumont GA, Ansermino JM. Pediatric pulse oximetry-based OSA screening at different thresholds of the apnea-hypopnea index with an expression of uncertainty for inconclusive classifications. Sleep Med 2019.
33. Alvarez D, Crespo A, Vaquerizo-Villar F, Gutiérrez-Tobal GC, Cerezo-Hernández A, Barroso-Garcia V, Ansermino JM, Dumont GA, Hornero R, Del Campo F, et al. Symbolic dynamics to enhance diagnostic ability of portable oximetry from the Phone

Oximeter in the detection of paediatric sleep apnoea. Physiol Meas 2018.
34. Combs D, Parthasarathy S. Machines learning to detect obstructive sleep apnea in children are we there yet? Am J Respir Crit Care Med 2017.
35. Shim SR, Kim SJ, Lee J. Diagnostic test accuracy: application and practice using R software. Epidemiol Health 2019.
36. Wang J, Leeflang M. Recommended software/packages for meta-analysis of diagnostic accuracy. J Lab Precis Med 2019.
37. Duval S, Tweedie R. Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. Biometrics 2000.
38. Faust O, Hagiwara Y, Hong TJ, Lih OS, Acharya UR. Deep learning for healthcare applications based on physiological signals: A review. Comput Methods Programs Biomed 2018.
39. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems. 2017.
40. Berry RB, Budhiraja R, Gottlieb DJ, Gozal D, Iber C, Kapur VK, Marcus CL, Mehra R, Parthasarathy S, Quan SF, et al. Rules for scoring respiratory events in sleep: update of the 2007 AASM manual for the scoring of sleep and associated events. J Clin sleep Med 2012;8(05):597–619.
41. Berry RB, Brooks R, Gamaldo C, Harding SM, Lloyd RM, Quan SF, Troester MT, Vaughn B V. AASM scoring manual updates for 2017 (version 2.4). J Clin Sleep Med 2017;13(05):665–666.
42. Tsai CM, Kang CH, Su MC, Lin HC, Huang EY, Chen CC, Hung JC, Niu CK, Liao DL, Yu HR. Usefulness of desaturation index for the assessment of obstructive sleep apnea syndrome in children. Int J Pediatr Otorhinolaryngol 2013.
43. Zamora J, Abraira V, Muriel A, Khan K, Coomarasamy A. Meta-DiSc: A software for meta-analysis of test accuracy data. BMC Med Res Methodol 2006.
44. Bhattacharjee R, Kheirandish-Gozal L, Pillar G, Gozal D. Cardiovascular Complications of Obstructive Sleep Apnea Syndrome: Evidence from Children. Prog Cardiovasc Dis 2009.
45. Hunter SJ, Gozal D, Smith DL, Philby MF, Kaylegian J, Kheirandish-Gozal L, Children S, Hunter SJ, Gozal D, Smith DL, et al. Effect of sleep-disordered breathing severity on cognitive performance measures in a large community cohort of young school-aged children. Am J Respir Crit Care Med 2016;194(6):739–747.
46. Kaditis AG, Alvarez MLA, Boudewyns A, Alexopoulos EI, Ersu R, Joosten K, Larramona H, Miano S, Narang I, Trang H, et al. Obstructive sleep disordered breathing in 2- to 18-year-old children: Diagnosis and management. Eur Respir J 2016.
47. Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. Science (80-) 2015.
48. Bishop CM. Pattern recognition and machine learning. springer; 2006.
49. Marcus CL, Moore RH, Rosen CL, Giordani B, Garetz SL, Taylor HG, Mitchell RB, Amin R, Katz ES, Arens R, et al. A Randomized Trial of Adenotonsillectomy for Childhood Sleep Apnea. N Engl J Med 2013.
50. Adadi A, Berrada M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access 2018.
51. Walter SD. Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. Stat Med 2002.
52. Moher D, Liberati A, Tetzlaff J, Altman DG, Altman D, Antes G, Atkins D, Barbour V, Barrowman N, Berlin JA, et al. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. PLoS Med 2009.
53. Tan H-L, Kaditis AG. Phenotypic Variance in Pediatric Obstructive Sleep Apnea.

Pediatr Pulmonol 2021;(In press).