Review

Artificial Intelligence Diagnosis of Obstructive Sleep Apnea Using Overnight Pulse Oximetry: A Systematic Review and Bayesian Meta-Analysis

Kvan Jie Ming Yam[1, 2*], MBBS; Claire Yi Jia Lim[3*]; Esther Yanxin Gao[4, 5, 6*], MBBS; Jin Hean Koh[4, 5*], MBBS; Nicole Kye Wen Tan[4], MBBS; Adele Chin Wei Ng[4, 5, 6, 7], MBBS, MMed; Zhou Hao Leong[5, 6], MBBS, MMed; Chu Qin Phua[7, 8], MBChB, MMed; Thun How Ong[7, 9], MBBS; Leong Chai Leow[7, 9], MD; Guang-Bin Huang[10, 11, 12], PhD; Benjamin Kye Jyn Tan[4, 5, 6], MBBS (Hons), MS; Song Tar Toh[4, 5, 6, 7], MBBS, MMed,MMed, FAMS

[1]Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore, Singapore
[2]SingHealth Duke-NUS Academic Medical Centre, Singapore, Singapore
[3]School of Medicine, Faculty of Medical and Health Sciences, University of Auckland, Auckland, New Zealand
[4]Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore
[5]Surgery Academic Clinical Program, SingHealth, Singapore, Singapore
[6]Department of Otorhinolaryngology–Head & Neck Surgery, Singapore General Hospital, Singapore, Singapore
[7]Duke-NUS Sleep Centre, SingHealth, Singapore, Singapore
[8]Department of Otorhinolaryngology–Head & Neck Surgery, Sengkang General Hospital, Singapore, Singapore
[9]Department of Respiratory and Critical Care Medicine, Singapore General Hospital, Singapore, Singapore
[10]School of Automation, Southeast University, Nanjing, China
[11]Key Laboratory of Measurement and Control of Complex Systems of Engineering, Ministry of Education, Nanjing, China
[12]Mind PointEye, Singapore, Singapore
*these authors contributed equally

Corresponding Author:

Song Tar Toh, MBBS, MMed,MMed, FAMS
Department of Otorhinolaryngology–Head & Neck Surgery
Singapore General Hospital
Block 3 Outram Rd, Basement 1 Singapore General Hospital
Singapore 169608
Singapore
Phone: +65 6222 3322
Email: toh.song.tar@singhealth.com.sg

Abstract

Background: Obstructive sleep apnea (OSA) affects 38% of the population, yet over 90% of cases remain undiagnosed. The gold standard for diagnosis, polysomnography, requires specialized equipment and trained personnel, making it inaccessible in primary care and acute settings. With artificial intelligence (AI) advancements, oximetry-based AI models have emerged as potential alternatives for OSA diagnosis.

Objective: This meta-analysis aims to evaluate the diagnostic accuracy of AI models trained on pulse oximetry readings in diagnosing OSA.

Methods: A systematic search was conducted across Medline/PubMed, Embase, Scopus, Web of Science, and IEEE Xplore databases from inception to January 3, 2026. Studies that evaluated the diagnostic accuracy of AI models trained on oxygen saturation recordings, compared to the apnea-hypopnea index (AHI) as the reference standard, were included and screened by 2 blinded independent reviewers. Models were evaluated using Bayesian bivariate meta-analysis and meta-regression. Publication bias was examined using a selection model approach, while risk of bias and evidence quality were assessed with Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) and Grading of Recommendations Assessment, Development, and Evaluation (GRADE).

Results: From 13,986 screened articles, 25 studies met the inclusion criteria, encompassing 23,171 participants with a mean age of 40 (SD 10.6) to 63 (SD 13.3) years and a BMI of 25 to 37 kg/m^2. AI-oximetry models demonstrated a pooled sensitivity of 91.1% (95% credible interval [CrI] 89.7%-92.4%) and specificity of 88.4% (95% CrI 85.3%-90.8%). Neural network

classifiers achieved the highest sensitivity (92.7%) and specificity (91.3%). Deep learning feature extraction models were significantly higher in sensitivity (by 3.7%; 95% CrI 0.9%-6.9%) than domain expert-based approaches. Sensitivity decreased slightly with higher AHI cutoffs, while specificity increased by 16.6% from an AHI cutoff of ≥5 to ≥30. Sensitivity analyses showed that even with up to 40% probability of an unpublished study, changes in accuracy were modest (area under the curve: 0.902 to 0.877). QUADAS-2 and GRADE assessments found low-moderate risk of bias with high overall quality of evidence.

Conclusions: AI-oximetry models showed high diagnostic accuracy for OSA across models and AHI cutoffs, performing better than or comparably to traditional overnight oximetry and home sleep apnea tests. This review provides the first pooled quantitative synthesis of AI models trained solely on oximetry data, with additional evaluations of publication bias and methodological limitations. Prior reviews were largely narrative or used alternative AI inputs other than oximetry. This study advances the field by offering a clearer and more reliable evidence base on pooled AI oximetry performance. These findings support the potential of oximetry-based AI as a convenient and scalable tool for OSA screening and diagnosis, with potential real-world applications in both primary care and inpatient settings for early identification of high-risk patients. Prospective external validation in diverse populations and low-prevalence settings is still needed before widespread real-world use.

Trial Registration: PROSPERO CRD42025648556; https://www.crd.york.ac.uk/PROSPERO/view/CRD42025648556

J Med Internet Res 2026;28:e80349; doi: 10.2196/80349

Keywords: machine learning; neural networks; deep learning; sleep disordered breathing; sleep apnoeas; diagnostic test accuracy

Introduction

Rationale

Obstructive sleep apnea (OSA) is a highly prevalent and yet underdiagnosed condition, with an estimated 38% prevalence in the general population, yet over 90% of patients remain undiagnosed [1]. During sleep, these patients experience recurrent upper airway obstruction, which results in intermittent oxygen desaturation and sleep disruption, which increase their risk of developing devastating health complications like heart disease, stroke, chronic kidney disease, cognitive decline, depression, and cancers [2-6]. The economic ramifications of OSA extend beyond their health sequelae but also include OSA-related fatigue causing lost productivity, work-related accidents, and motor vehicle accidents [7].

A major contributor to the persistent diagnostic gap is the reliance on overnight polysomnography, the gold-standard test for OSA. Polysomnography is resource-intensive, requiring an inpatient night stay with complex equipment and skilled technicians [8]. This test is thus limited in availability, particularly in the primary care setting or in low-middle-income countries, where the majority of the world's population lives [9]. Simple and convenient screening tools such as the STOP-BANG (Snoring, Tiredness, Observed apnoea, Pressure [hypertension], BMI>35 kg/m^2, Age>50, Neck circumference>40 cm, and Gender [male]) questionnaire that have been widely used could help to address this need for OSA risk stratification in the primary care setting. However, while it has a high sensitivity of over 90%, it has a low specificity of 28%, which results in many false positives (FP) [10]. As such, many patients would be expected to screen positive with STOP-BANG, and with polysomnography as the only diagnostic tool, this would still result in a high number of undiagnosed cases. As health care systems globally are strained, this results in a long wait time for polysomnography appointments and time to treatment initiation. Such delays can negatively impact patient outcomes. For example, a center in Calgary, Canada, reported a mean time to treatment of 123 days and found that longer wait times were associated with decreased adherence to treatment and smaller improvements in daytime sleepiness assessed using the Epworth Sleepiness Scale score [11].

As the artificial intelligence (AI) sector advances rapidly, pulse oximetry readings have become a viable input for AI-guided diagnosis of OSA [12-14]. Overnight peripheral pulse oximetry for oxygen saturation (SpO_2) is a simple, noninvasive, and widely available tool already used commonly in select clinical settings to diagnose OSA when a polysomnogram may not be feasible or acceptable, such as in pediatrics. It is physiologically relevant to OSA diagnosis because apneic and hypopneic events cause characteristic episodic drops in arterial oxygen saturation [15]. Conventional quantitative indices derived from SpO_2, such as the oxygen desaturation index (ODI), correlate strongly with the apnea-hypopnea index (AHI) measured by polysomnography, the gold standard for OSA diagnosis. While ODI-based diagnosis yields high sensitivity (>90%), its specificity remains modest between 40%-60%, limiting its standalone diagnostic utility [16,17].

SpO_2 measurement is integral to the formal definition of hypopnea, underscoring its diagnostic relevance [18]. Furthermore, the pattern and severity of nocturnal hypoxemia, as measured by SpO_2, are associated with OSA-related morbidity, including neurocognitive impairment and cardiovascular risk [19]. However, oximetry alone cannot differentiate central versus obstructive events, nor can it identify sleep stages, constraining its standalone diagnostic value. An AI-driven analysis of SpO_2 data—leveraging machine learning or deep learning techniques—may overcome these limitations by extracting complex features from desaturation patterns such as central tendency, morphology, frequency, and amplitude of desaturation to more accurately classify OSA severity [20]. This approach has the potential to provide scalable, low-cost, and automated diagnostic support that could ease pressure on

sleep laboratories, facilitate earlier identification of at-risk individuals, and improve access to care in underserved regions.

Despite the growing number of studies on AI in diagnosing OSA, evidence remains fragmented. Individual studies vary widely in model type, training methods, input features, oximeter brands and specifications, and diagnostic thresholds, and no comprehensive synthesis has evaluated the pooled diagnostic performance of AI models trained on SpO_2 data. There remains a significant gap in the literature; a meta-analysis of the diagnostic capabilities of the models. Such a review is vital in assessing the accuracy, robustness, and clinical applicability of AI-driven oximetry for diagnosing OSA. To address the existing gaps, we conducted the first Bayesian meta-analysis evaluating the diagnostic accuracy of AI models trained on SpO_2 recordings for OSA detection.

Objectives

We aim to (1) quantitatively pool diagnostic performance metrics of AI models across studies and (2) use meta-regression to identify methodological and clinical factors associated with higher diagnostic accuracy. We hypothesize that AI models may be able to support risk stratification, thereby reducing the demand for polysomnography, while acknowledging that with the current capabilities of AI, it is unlikely to replace polysomnography as the definitive diagnostic modality.

Methods

Information Sources

The prespecified protocol for this review was registered on PROSPERO (International Prospective Register of Systematic Reviews; CRD42025648556). With reference to the PRISMA (Preferred Reporting Items for Systematic Review and Meta-Analyses) 2020 expanded guidelines and PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension) checklist (Checklists 1 and 2), a search was conducted on Ovid Medline/PubMed, Elsevier Embase, Elsevier Scopus, Clarivate Web of Science, and IEEE Xplore databases for studies published from inception till 3 January 2026 [21-23]. Multidatabase searching on a single platform was not performed. The PRISMA 2020 Expanded checklist is included in Checklist 3.

Search Strategy

The search strategy was adapted from a previous review on machine listening for OSA diagnosis by the same corresponding author [24], which used a combination of the following search terms: ("sleep apnea" OR "sleep apnoea" OR "nocturnal hypoxia" OR "nocturnal hypoxaemia" OR "nocturnal hypoxemia" OR "sleep disordered breathing") AND ("artificial intelligence" OR "machine learning" OR "deep learning" OR "logistic regression" OR "support vector machine" OR "neural network" OR "classification tree" OR "regression tree" OR "probability tree" OR "nearest neighbor" OR "nearest neighbor" OR "fuzzy logic" OR "naive bayes" OR "genetic algorithm" OR "multilayer perceptron" OR "random forest" OR "lasso regression" OR "kernel regression" OR "elastic net" OR "generative model" OR "generative adversarial network" OR "large language model") AND (diagnosis OR diagnose OR detect OR detection OR identify OR identification OR severity OR classify OR classification). The full search strategies for each database are available in Multimedia Appendix 1, including the description of any limits applied. Due to the extensive search strategy and large number of search results, no additional hand-searching through study registries, online browsing, citation searching, author contacts, or any other methods was performed. No search filters and no search peer review process were used.

Managing Records

A total of 13,986 records were retrieved from the database search (PubMed: 3912; Embase: 1487; Web of Science: 3424; Scopus: 2780; IEEE Xplore: 2383). Software used for deduplication includes EndNote, Rayyan (Rayyan Systems Inc), and TERA the deduplicator [25-27]. After deduplication, 5551 duplicates were removed, leaving 8435 records for title and abstract screening.

Selection Process

Records were uploaded onto Rayyan [27], an online systematic review platform that enables authors to manually screen abstracts in a blinded manner. Two blinded reviewers (KJMY and CYJL) independently screened the titles and abstracts, followed by full-text screening to check the eligibility for inclusion, with disputes being resolved by consensus from a third independent reviewer (JHK).

Eligibility Criteria

The inclusion criteria were as follows:

1. Population: adults aged at least 18 years
2. Intervention/exposure: diagnosis and classification of OSA using AI (traditional regression techniques, machine learning, etc) trained on SpO_2 recordings from overnight polysomnography or home sleep apnea tests (HSATs).
3. Comparators: diagnosis and classification of OSA using the apnea-hypopnea index (AHI) from overnight polysomnography or HSATs.
4. Outcomes: Accuracy of AI in diagnosis and classification of OSA, assessed via a random split test set or k-fold cross-validation, and measured by sensitivity, specificity, positive predictive value, negative predictive value, and/or area under the curve (AUC).
5. Study type: observational studies (eg, cohort and cross-sectional).

The exclusion criteria were as follows:

1. Case reports, reviews, letters, conference abstracts, or other records not published as full-length articles in peer-reviewed journals.
2. Studies published in languages other than English that do not have an English translation.

3. Studies assessed as having a high risk of bias across 2 or more domains.
4. Studies that did not measure the diagnostic accuracy of AI in diagnosing AHI-defined OSA

Data Collection Process, Including Data Items

Data from included articles were extracted by 2 blinded, independent reviewers (KJMY and CYJL) in duplicate onto a structured form specifically designed for the study and piloted beforehand on a sample of selected studies. Disagreement was resolved by discussion and consensus with a third reviewer (JHK). The standardized extraction spreadsheet template contained the following data: participant characteristics (percentage male, mean/median age, and OSA prevalence); study characteristics (first author, publication year, study design, study setting, country, and sample sizes for the training, validation, and test datasets where applicable); model characteristics (type of AI classifiers used, feature engineering, method of OSA diagnosis [eg, polysomnography or HSAT], reference standard used for OSA diagnosis, and AHI cutoffs); and the following outcome domains: (1) diagnostic performance of AI models for OSA, including sensitivity, specificity, accuracy, positive predictive value, negative predictive value, AUC; and (2) confusion-matrix components (true positives [TP], true negatives [TN], FP, and false negatives [FN]) where available. In line with PRISMA 2020 expanded checklist, we specified whether all results compatible with each outcome domain were sought. For studies reporting multiple eligible results (ie, several AHI thresholds, different AI model versions, or multiple test sets), we extracted all results that met our predefined inclusion criteria. When confusion matrices were reported, all relevant matrices corresponding to included analyses were extracted. This approach was used to minimize selective-outcome bias across studies.

Where data were missing or unclear, the following assumptions were applied: if sample sizes differed across sections of the manuscript without clarification, we used the number explicitly associated with the relevant dataset (training, validation, or test). When AHI cutoffs were not directly stated, we inferred thresholds based on the authors' definitions of OSA severity (mild, moderate, or severe), which clearly aligned with standard criteria. If demographic data were reported inconsistently (ie, percentages without denominators), we assumed the total study population as the denominator unless otherwise indicated. We did not impute missing accuracy metrics. No external tool was used to determine which data items to collect.

Study Risk of Bias Assessment

The quality assessment of included studies was assessed by 2 blinded, independent reviewers (KJMY and CYJL), using the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) tool to evaluate the risk of bias and applicability of diagnostic accuracy studies [28]. The QUADAS-2 tool assesses studies on the following 4 key domains, including patient selection, index test, reference standard, and flow and timing. Each domain and the overall study are graded as either low, some concerns, or high risk of bias. If any 1 domain was graded as "some concerns" or "high risk," the overall risk of bias for that study would be graded as "some concerns" or "high risk," respectively. Any disputes between reviewers were resolved by consensus from a third independent reviewer (JHK).

Statistical Analysis, Including Effect Measures, Synthesis Methods, and Reporting Bias Assessment

Binary outcome data (TP, FP, TN, and FN) were used directly from confusion matrices reported in the primary studies. When not directly reported, binary diagnostic accuracy data were derived using the formulas below:

- Sensitivity = TP / (TP + FN)
- Specificity = TN / (TN + FP)
- OSA positive sample size = TP + FN
- OSA negative sample size = TN + FP
- Total sample size = TP + TN + FP + FN
- Accuracy = (TP + TN) / (TP + TN + FP + FN)
- Recall = TP / (TP + FN)

Thus:

- TP = Sensitivity × OSA positive sample size = Recall × OSA positive sample size
- TN = Specificity × OSA negative sample size = Accuracy × Total sample size – TP

When multiple metrics were available, we prioritized directly reported confusion matrices, followed by sensitivity or specificity with prevalence, as these allowed the most accurate reconstruction. Studies for which TP, TN, FP, and FN could not be derived were excluded from the analysis. Minor discrepancies due to rounding were addressed by rounding to the nearest integer, as it is not possible to have less than a whole patient. Further checks were performed by ensuring that the calculated TP, TN, FP, and FN matched the total sample size. As there was no statistical reconstruction but rather simple calculations were performed, there was no statistical uncertainty to account for, and the authors have decided not to perform a meta-regression.

Studies with directly reported or calculated 2×2 data and which evaluated their model with a random split test set or k-fold cross-validation were then pooled in a Bayesian bivariate random effects meta-analysis, using a noninformative prior. As specificity and sensitivity are related, a Bayesian meta-analysis was the chosen method to allow for joint estimation of them. Pooled sensitivity and specificity were summarized using hierarchical summary receiver operating characteristic (HSROC) curves. Diagnostic odds ratio (DOR) and positive and negative likelihood ratios were also derived from the meta-analysis. Between-study heterogeneity was graphically visualized using 95% prediction regions on HSROC curves. Random-effects Bayesian meta-regression was performed for both continuous and categorical study-level covariates, including AI classifiers, feature engineering, AHI cutoffs, sampling frequency, sleep test reference standard, test method, prevalence, age, and gender ratio.

To assess the robustness of the synthesized results, an informative prior (where the lower bound was set as 50% sensitivity/specificity) was applied as a sensitivity analysis for the overall meta-analysis. The potential impact of 4 different mechanisms of publication bias (data, sensitivity, specificity, or DOR-driven) with varying probabilities of unpublished studies (up to 40%) was evaluated via a sensitivity analysis where the HSROC curve, AUC, sensitivity, and specificity were re-estimated for each scenario in a Bayesian hierarchical framework. Risk of bias due to missing results in the synthesis was simultaneously assessed using the same Bayesian selection-model sensitivity analyses. All analyses were conducted following statistical guidance from the Cochrane Handbook and were performed using MetaBayesDTA (1.5.2; University of Leicester) and DTAmetasa (0.9.1; Osaka University) [29-33], built using R (R Foundation for Statistical Computing) and Stan (Stan Development Team) [34-36]. The analyses were performed by one statistician and independently verified by a second analyst. Each AI model was treated as independent observations even if they originated from the same study; AI models were only grouped according to technological factors for the purposes of meta-regression.

For diagnostic accuracy outcomes, the primary effect measures used in the synthesis were sensitivity and specificity, consistent with standard practice for diagnostic test accuracy reviews and Cochrane DTA guidance. Secondary effect measures include DOR, positive likelihood ratio, negative likelihood ratio, and the area under the HSROC curve. No categorical thresholds (ie, “small,” “moderate,” or “large” effects) were applied to interpret sensitivity, specificity, or other accuracy measures; instead, statistical significance and uncertainty were assessed using 95% credible intervals (95% CrI) from the Bayesian models.

Certainty Assessment/Quality of Evidence

The quality of pooled evidence was evaluated using the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) framework [35]. The GRADE framework rates each study on the basis of study design, consistency, directness, risk of bias, precision, and publication bias. For each outcome, the level of certainty was rated as high, moderate, low, or very low, following standard GRADE decision rules. Evaluations were performed by 2 reviewers (KJMY and CYJL) independently, with disagreements resolved through discussion.

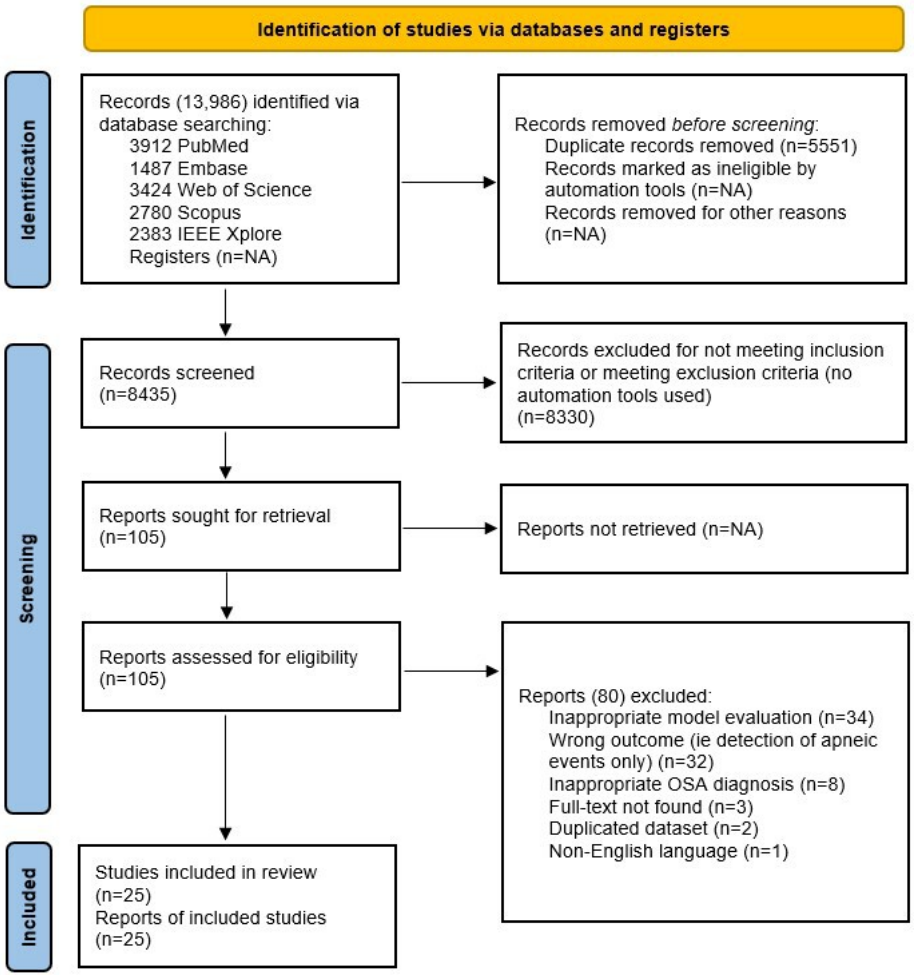
Results

Study Selection

A total of 8435 articles were included in the initial search after the removal of duplicates, of which 105 were selected for full-text review based on title and abstract screening. After full-text review, 25 [12,13,20,37-58] articles met the final inclusion criteria and were included in Table S1 in Multimedia Appendix 1. The study selection process is summarized in Figure 1.

Studies that only detected apneic events without a diagnosis of OSA or investigated sleep apnea-hypopnea syndrome but not OSA were excluded [59,60].

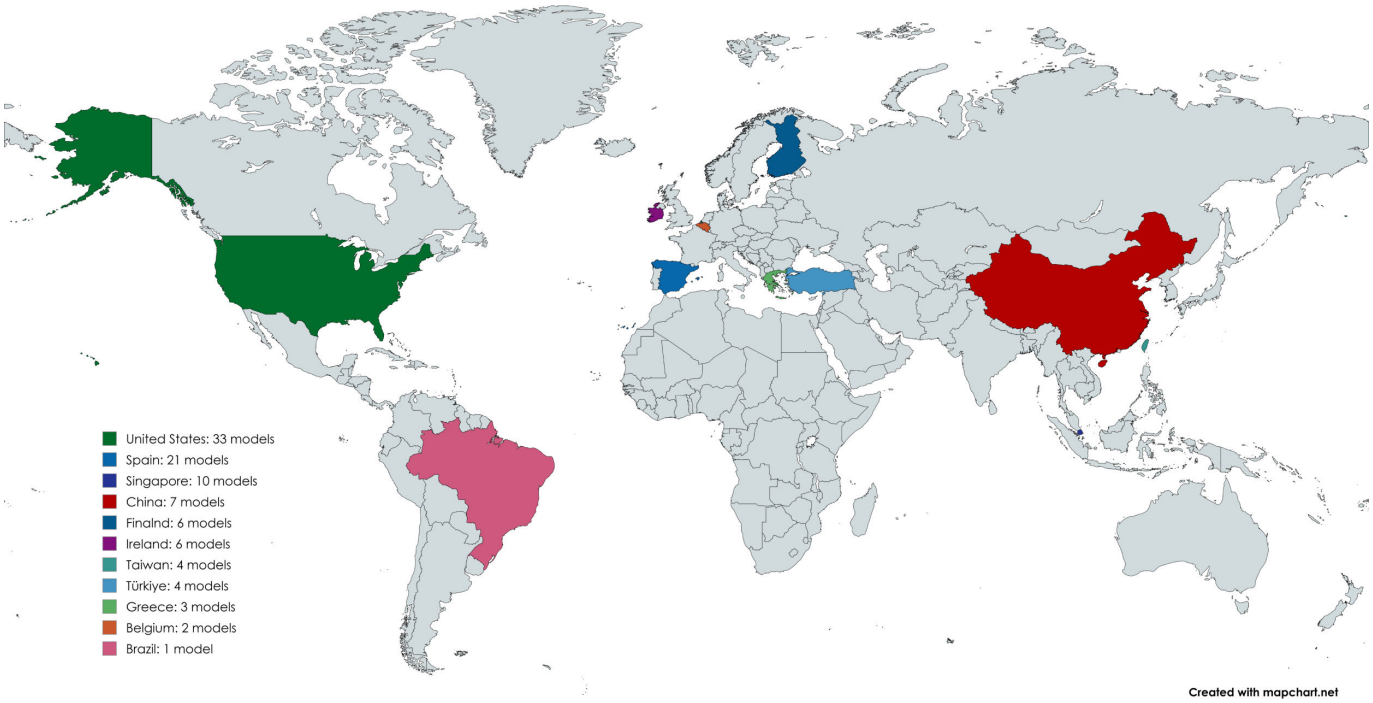
Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram to summarize the study selection process. NA: not applicable; OSA: obstructive sleep apnea.



Study Characteristics

A total of 25 [12,13,20,37-58] cross-sectional studies with 97 AI models were included. As shown in Figure 2, 42 models were tested in European populations, 1 in South America, 33 in North America, and 21 models were tested in an Asian population. The total sample (nonduplicated) consisted of 23,171 participants used to train the AI models and 15,025 participants for testing. Mean age ranged from 40.15 (SD 10.6) to 63.32 (SD 13.3) years, while the BMI of participants ranged from 25.20 to 37.08 kg/m^2. A summary of the characteristics of each study can be found in Table S1 in Multimedia Appendix 1. There were no studies with a high risk of bias based on QUADAS-2 (Table S2 in Multimedia Appendix 1).

Figure 2. Geographical locations of included studies.



Results of Individual Studies

Among the 25 studies, 12 [12,13,20,37-41,50,55-57] studies used a Nonin brand of pulse oximeter, within which there were 7 different models; 5 [45-49] studies used the Criticare 504 Oximeter, 1 study each used an Embla N7000 Polysomnography System brand [58], Grael polysomnography [42], SleepSense Adult Soft-Tip SpO_2 Sensor [51], and Masimo Oximeter [53]. The remaining 4 [43,44,52,54] studies did not specify the brand of pulse oximeter used. As there was a large variety in the brands of pulse oximeters used, the authors have decided not to perform a meta-regression for brand of pulse oximeter. Furthermore, no two studies from independent populations shared the same pulse oximeter brand and model; thus, a meta-regression may not generate any meaningful difference in accuracy. Sampling frequency varied greatly from 0.2 Hz to 500 Hz (Table S3 in Multimedia Appendix 1).

Reference Standard for OSA Diagnosis

Among the 25 studies, 22 studies [12,13,20,38,39,41-49,51-58] evaluated for OSA using overnight polysomnography, while 3 studies [37,40,50] used HSATs. All studies defined OSA and its severity using the AHI. Among the 97 AI models, 39, 10, 29, and 19 used an AHI cutoff of ≥5, ≥10, ≥15, and ≥30 events/h, respectively, to define the presence of OSA (Table S1 in Multimedia Appendix 1). The prevalence of OSA in the testing set ranged from 12.5% to 93.8% [37,40].

Artificial Intelligence Models

To extract and select SpO_2 features, 58 models used deep learning, while 39 models relied on manual feature extraction by a domain expert. To classify the severity of OSA, 15 models used decision trees (comprising 12 classification and regression trees and 3 random forest models), 6 models used linear models (comprising 4 logistic regression and 2 linear discriminant analysis models), 10 models used gradient boost, 46 models used neural networks, and 20 models used support vector machines. A total of 52 models were evaluated with a cross-validation technique (k-fold or leave-one-out) while 45 models were evaluated with a random-split test set.

Data Presentation

For each study, the raw diagnostic accuracy data (TP, FP, TN, and FN) are presented in Table S4A in Multimedia Appendix 1. These data formed the basis for the hierarchical meta-analysis; individual study-level effect estimates with precision (eg, CIs) were not generated because accuracy estimates were synthesized using the Bayesian hierarchical HSROC framework.

As shown in Table S4A in Multimedia Appendix 1, out of the 25 included studies, 13 studies presented their confusion matrix and allowed for direct extraction of the TP, TN, FP, and FN values. A total of 12 studies required a calculation of the raw accuracy data from OSA prevalence and the presented sensitivity and specificity. Raw data regarding the accuracy, sensitivity, and specificity of each AI model have also been provided.

Results of Synthesis: Meta-Analysis of Diagnostic Accuracy

All 25 included studies contributed to the primary synthesis of diagnostic accuracy. As summarized in Tables S1, S4A, and S4B in Multimedia Appendix 1, the contributing studies varied in AI classifier type, feature engineering approach, test-set construction methods, and AHI thresholds used for model evaluation. These methodological differences represent important potential sources of heterogeneity when interpreting the pooled results; therefore, subgroup analyses and

meta-regressions stratified by these study characteristics were performed. The results of each meta-regression are presented in the sections below. Overall certainty of evidence was rated using the GRADE framework, and the risk of bias was also assessed for all studies using QUADAS-2, with domain-level judgments summarized in the subsequent section.

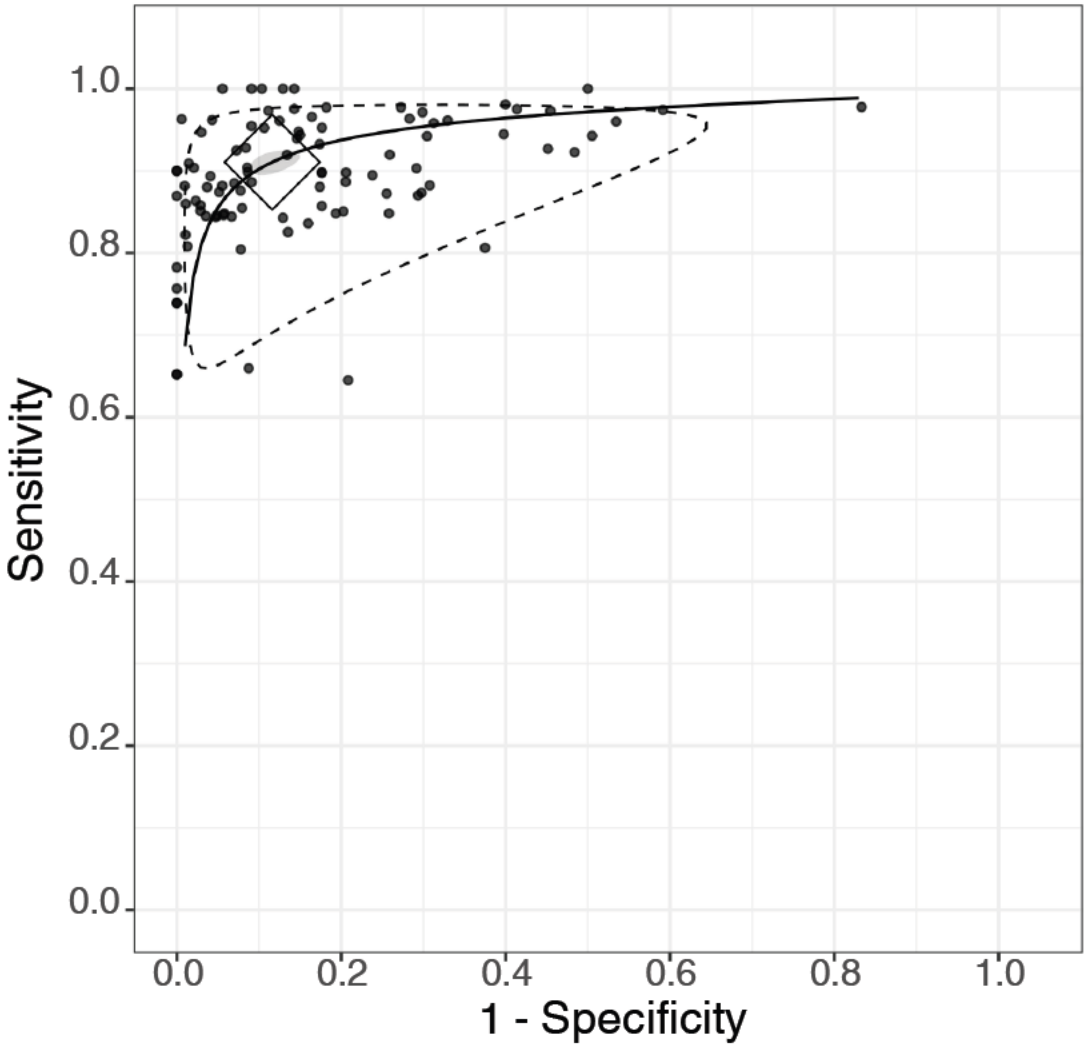
Overall Accuracy Statistics

Compared to conventional diagnostic methods, the use of AI trained on SpO_2 recordings achieved a pooled sensitivity of 91.1% (95% CrI 89.7%-92.4%) and specificity of 88.4% (95% CrI 85.3%-90.8%), with a DOR of 77.7 (95% CrI 60.2-99.6), positive likelihood ratio of 7.85 (95% CrI 6.25-9.85), and negative likelihood ratio of 0.10 (95% CrI 0.086-0.116). The summary receiver operating characteristic (SROC) curve is displayed in Figure 3.

A sensitivity analysis using an informative prior yielded identical results, with a sensitivity of 91.1% (95% CrI 89.6%-92.2%) and specificity of 88.4% (95% CrI 85.7%-90.8%), a DOR of 77.3 (95% CrI 61.7-99.0), a positive likelihood ratio of 7.83 (95% CrI 6.40-9.86), and a negative likelihood ratio of 0.10 (95% CrI 0.089-0.117). These minimal differences indicate that the pooled estimates were robust to prior specification.

In Figure 3 the solid line represents the extrapolated SROC curve. The diamond represents the summary receiver operating point. Shaded/dashed regions represent the 95% CrI. Unshaded circles/ovals are centered around individual study means; their height/width are proportionate to study weights for sensitivity/specificity.

Figure 3. Summary receiver operating characteristic (SROC) plot for the overall obstructive sleep apnea (OSA) diagnostic accuracy of artificial intelligence (AI) models trained on oxygen saturation (SpO_2) recordings.

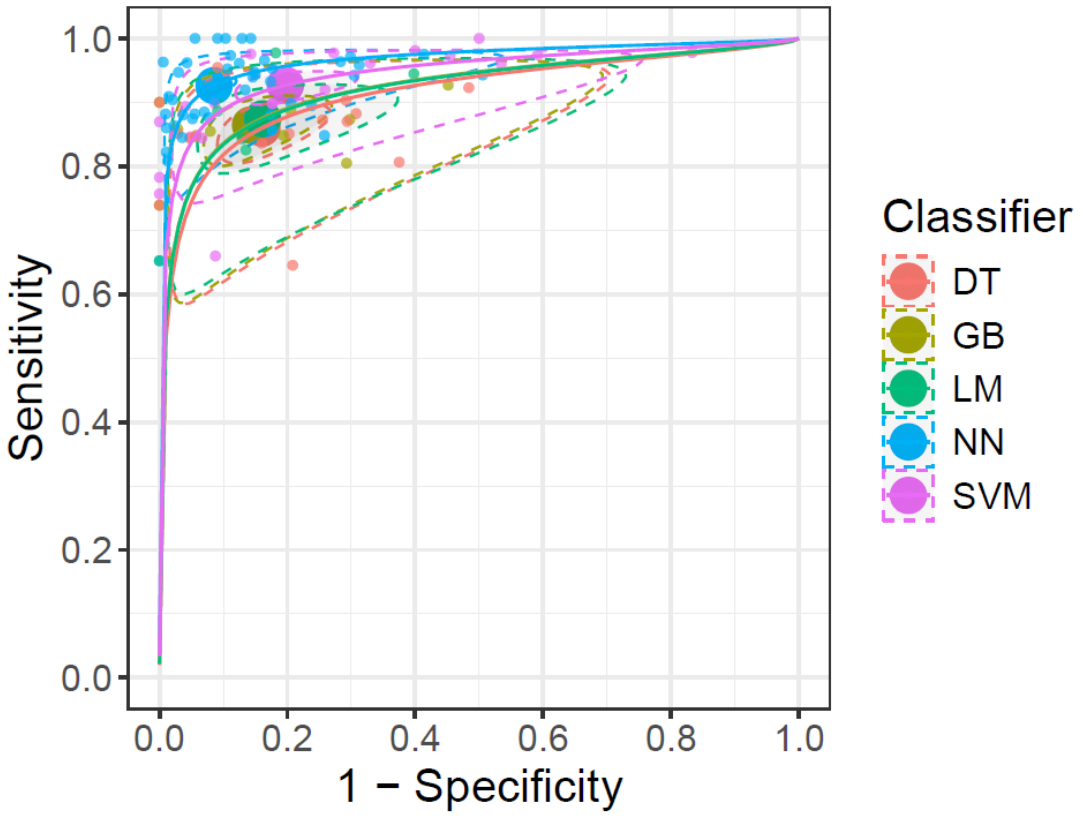


Meta-Regression of AI Classifier

As shown in Figure S1A in Multimedia Appendix 1, among the different types of AI classifiers involved in this study, neural networks consistently have the highest sensitivity and specificity at 92.7% (95% CrI 91.1%-94.0%) and 91.3% (95% CrI 87.9%-93.5%), respectively. The sensitivity of neural networks is significantly higher than that of decision trees (difference in sensitivity: –6.7%, 95% CrI –11.9% to –2.4%) and gradient boost (difference in sensitivity: –6.3%, 95% CrI –12.6% to –1.8%). There was no significant

difference in sensitivity when compared to linear models. There were no significant differences in specificities among neural networks, gradient boosts, linear models, and decision trees. The next best model is a support vector machine with a sensitivity of 92.7% (95% CrI 89.8%-94.8%) and a specificity of 80.0% (95% CrI 70.1%-87.4%). There was no significant difference in sensitivity (95% CrI –2.6 to 3.1%) between the support vector machine and the neural network. However, the neural network was more specific (difference in specificity: 11.2%, 95% CrI 3.30%-21.1%) than the support vector machine. The remaining models are comparable in sensitivity and specificity. Visual comparison of the SROC curves corroborates the above findings, with the neural network having the best performance, followed by the support vector machine, with the remaining 3 models being comparable, as shown in Figure 4.

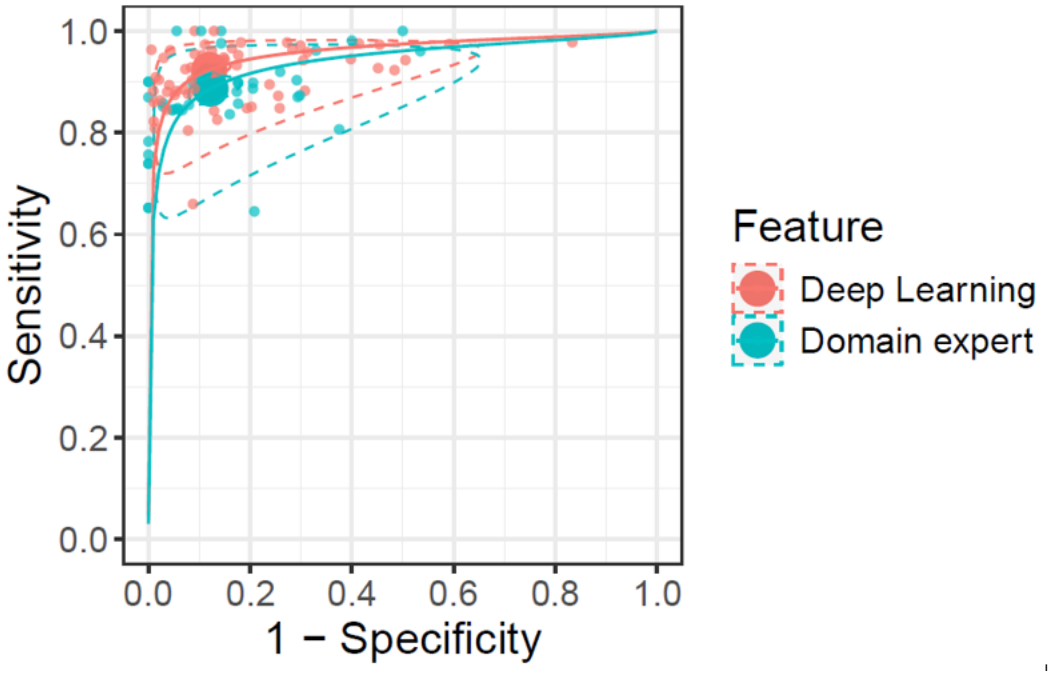
Figure 4. Summary receiver operating characteristic (SROC) plot for Bayesian meta-regression stratified by artificial intelligence (AI) classifier. DT: decision tree; GB: gradient boost; LM: linear model; NN: neural network; SVM: support vector machine.



In Figure 4 the solid lines represent the extrapolated SROC curves. Large circles represent the summary operating points. The shaded areas represent the 95% CrI, and the dotted lines represent the 95% prediction region. Small circles represent individual study estimates.

In addition, for feature extraction and selection, deep learning (92.3%, 95% CrI 90.8%-93.6%) displayed significantly higher sensitivity as compared to domain expert (88.6%, 95% CrI 85.8%-90.9%) with a difference of 3.7% (95% CrI 0.9%-6.9%; Figure S1B in Multimedia Appendix 1). However, there were no significant differences in specificities. The SROC curves can be found in Figure 5.

Figure 5. Summary receiver operating characteristic (SROC) plot for Bayesian meta-regression stratified by type of artificial intelligence feature engineering. Solid lines represent the extrapolated SROC curves. Large circles represent the summary operating points. The shaded areas represent the 95% credible regions, and the dotted lines represent the 95% prediction region. Small circles represent individual study estimates.



Meta-Regression of AHI

At the clinically relevant AHI cutoffs of ≥5, ≥10, ≥15, and ≥30, the sensitivities were 93.4% (95% CrI 91.6%-94.8%), 91.3% (95% CrI 86.4%-94.7%), 88.8% (95% CrI 85.6%-91.2%), and 88.1% (95% CrI 83.7%-91.1%), respectively, while the specificities were 79.0% (95% CrI 72.3%-85.1%), 88.6% (95% CrI 79.2%-93.9%), 87.9% (95% CrI 83.5%-91.6%), and 95.7% (95% CrI 93.1%-97.3%). These data are visualized in Figure S1C in Multimedia Appendix 1.

Sensitivity was the highest at lower AHI cutoffs, while specificity increased as the AHI cutoff increased. While an AHI cutoff of ≥5 had no significantly different sensitivity than an AHI cutoff of ≥10 (difference 2.0%, 95% CrI –1.6% to 7.0%), it had a significantly higher sensitivity than an AHI cutoff of ≥15 (difference 4.5%, 95% CrI 1.5%-8.2%) and ≥30 (difference 5.3%, 95% CrI 1.8%-9.9%).

An AHI cutoff of ≥5 has a significantly lower specificity than an AHI cutoff of ≥30 with a difference of –16.6% (95% CrI –23.9% to –10.3%), and an AHI cutoff of ≥15 has a significantly lower specificity than an AHI cutoff of ≥30 with a difference of –7.7% (95% CrI –12.1% to –3.1%).

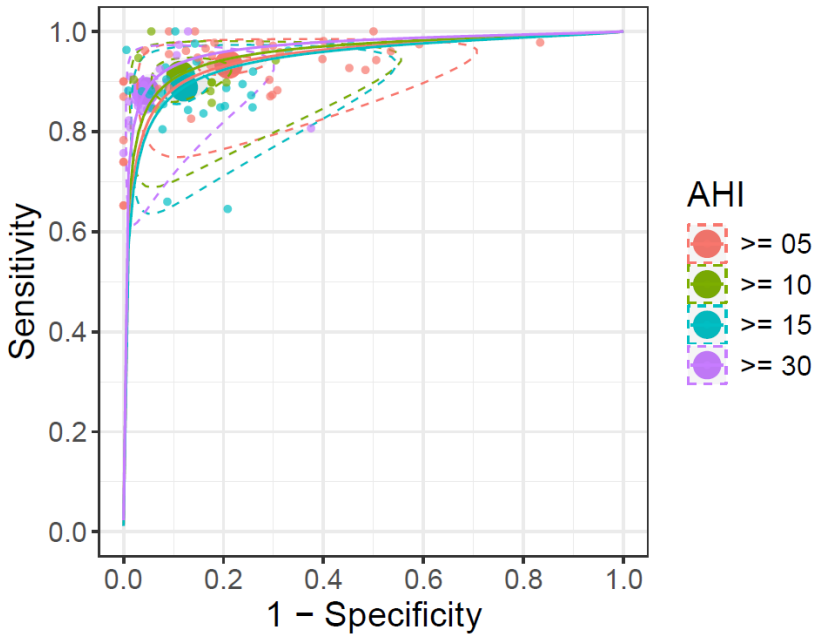
Other statistics are summarized in Table 1. Visual comparison of the SROC curves for each AHI cutoff suggested that performance was essentially similar across the various AHI cutoffs, as the curves were closely overlapping. The differences in sensitivity and specificity appeared to be mainly due to threshold shifts along the SROC curve, rather than shifts of the entire SROC curve Figure 6.

Table 1. Summary of diagnostic test accuracy statistics from Bayesian meta-analysis at clinically relevant thresholds.

Subgroup	Posterior, median (95% posterior interval)					
	Sensitivity	Specificity	FPR[a]	DOR[b]	LR+[c]	LR–[d]
Overall	91.1 (89.7-92.4)	88.4 (85.3-90.8)	11.6 (9.2-14.7)	77.7 (60.2-99.6)	7.85 (6.25-9.85)	0.10 (0.09-0.12)
AHI[e]≥5	93.4 (91.6-94.8)	79.0 (72.3-85.1)	21.0 (14.9-27.7)	52.7 (36.7-80.5)	4.44 (3.39-6.22)	0.085 (0.07-0.11)
AHI≥10	91.3 (86.4-94.7)	88.6 (79.2-93.9)	11.4 (6.1-20.8)	81.8 (39.5-173.4)	8.00 (4.47-14.81)	0.10 (0.06-0.15)
AHI≥15	88.8 (85.6-91.2)	87.9 (83.5-91.2)	12.1 (8.4-16.5)	58.2 (39.6-84.9)	7.35 (5.45-10.48)	0.13 (0.10-0.16)
AHI≥30	88.1 (83.7-91.1)	95.7 (93.1-97.3)	4.3 (2.7-6.9)	160.7 (100.5-267.7)	20.3 (12.82-31.12)	0.125 (0.094-0.169)

[a]FPR: false positive rate (1 – specificity).
[b]DOR: diagnostic odds ratio.
[c]LR+: likelihood ratio positive.
[d]LR–: likelihood ratio negative.
[e]AHI: apnea-hypopnea index.

Figure 6. Summary receiver operating characteristic (SROC) plot for Bayesian meta-regression stratified by apnea-hypopnea index (AHI) cutoffs of ≥5, ≥10, ≥15, and ≥30. Solid lines represent the extrapolated SROC curves. Large circles represent the summary operating points. The shaded areas represent the 95% credible regions, and the dotted lines represent the 95% prediction region. Small circles represent individual study estimates.

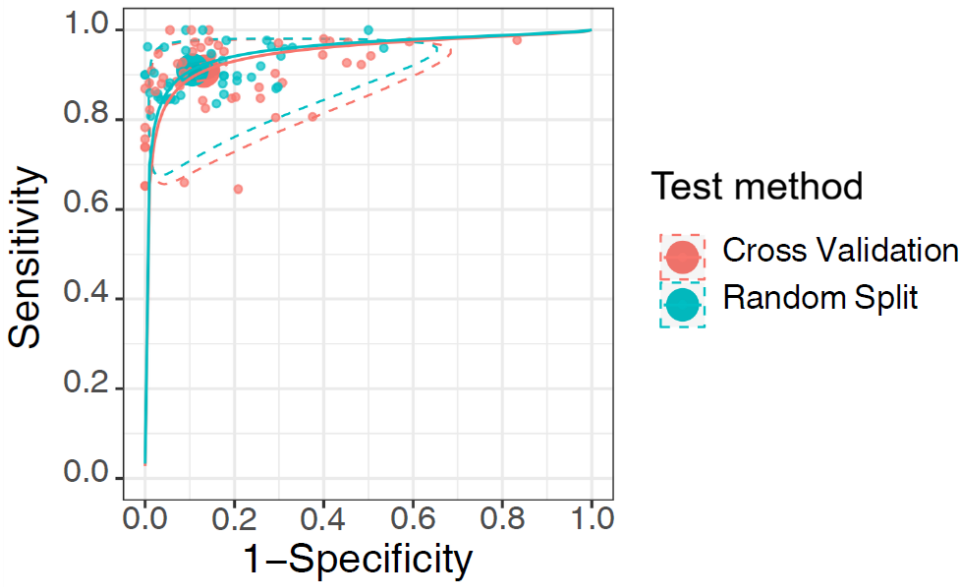


Meta-Regression of Test Method

To address the methodological heterogeneity introduced by different test methods, a meta-regression comparing studies that used random-split tests versus those that used k-fold cross-validation was conducted. Models evaluated using random-split test sets demonstrated a pooled sensitivity of 88.6% (95% CrI 85.8%-90.9%) and specificity of 87.8% (95% CrI 82.7%-92.0%), whereas models assessed with cross-validation showed a pooled sensitivity of 92.3% (95% CrI 90.8%-93.6%) and specificity of 88.0% (95% CrI 84.4%-91.2%).

The meta-regression confirmed that cross-validation was more sensitive (difference 3.7%, 95% CrI 0.9%-6.9%) but not significantly more specific (95% CrI –5.5% to 6.4%) than random split (Figure S1D in Multimedia Appendix 1). The corresponding SROC curves can be found in Figure 7. These findings indicate that differences in test method could have potentially affected diagnostic accuracy.

Figure 7. Summary receiver operating characteristic (SROC) plot for Bayesian meta-regression stratified by test method.



In Figure 7 solid lines represent the extrapolated SROC curves. Large circles represent the summary operating points. The shaded areas represent the 95% credible regions, and the dotted lines represent the 95% prediction region. Small circles represent individual study estimates.

Meta-Regression of Prevalence

Sensitivity remained consistent at 91.1% (95% CrI 89.7%-92.4%) despite changes in percentage prevalence, though specificity decreased as percentage prevalence increased, with an average of 86.7% (95% CrI 84.5%-88.7%; Figure S2A in Multimedia Appendix 1).

Meta-Regression of Age and Gender

Conversely, average age and the percentage of male participants were not associated with sensitivity or specificity (Figure S2B and S2C in Multimedia Appendix 1).

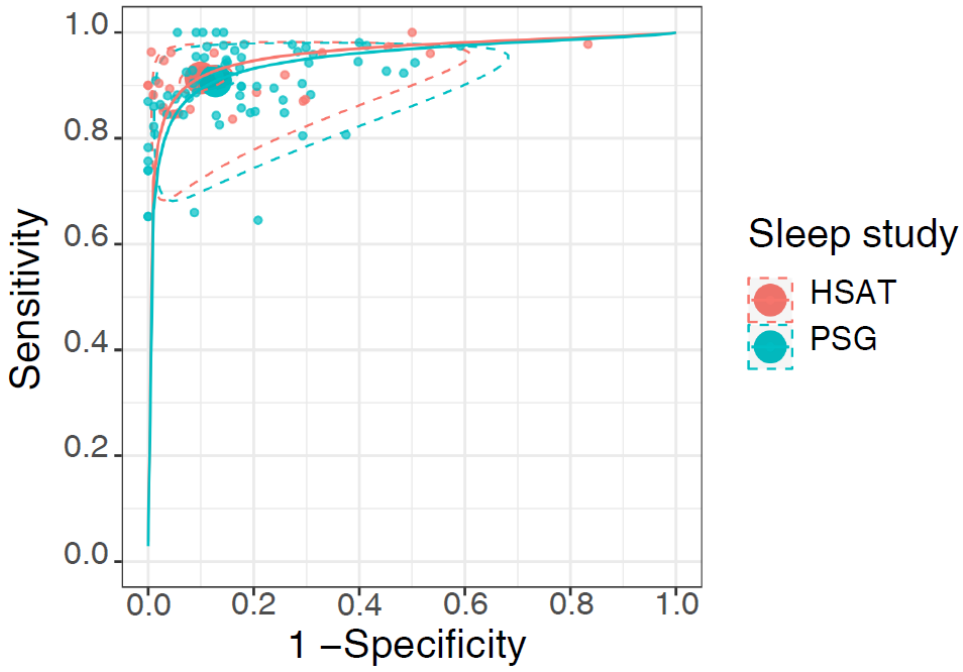
Meta-Regression of Sampling Frequency

Sensitivity remained constant at 91.1% (95% CrI 89.7%-92.4%), while specificity decreased as sampling frequency increased, with an average of 88.0% (95% CrI 85.0%-90.8%; Figure S3A in Multimedia Appendix 1).

Meta-Regression of Reference Standard

To determine if using polysomnography versus HSAT as the reference standard contributed to the heterogeneity between studies, a meta-regression was performed. Studies using polysomnography achieved a pooled sensitivity and specificity of 90.8% (95% CrI 89.2%-92.3%) and 87.4% (95% CrI 83.2%-90.5%), respectively (Figure 8). Studies using HSAT achieved a pooled sensitivity and specificity of 91.5% (95% CrI 89.0%-93.7%) and 90.0% (95% CrI 85.0%-93.5%). Meta-regression showed that this difference is not statistically significant, with the difference in sensitivity being 0.7% (95% CrI –2.4% to 3.4%) and the specificity at 2.5% (95% CrI –3.3% to 7.4%; Figure S3B in Multimedia Appendix 1).

Figure 8. Summary receiver operating characteristic (SROC) for Bayesian meta-regression stratified by reference standard. Solid lines represent the extrapolated SROC curves. Large circles represent the summary operating points. The shaded areas represent the 95% credible regions, and the dotted lines represent the 95% prediction region. Small circles represent individual study estimates. HSAT: home sleep apnea test; PSG: polysomnography.



Sensitivity Analysis and Restricting Datasets

Several studies trained and tested their AI models on the same publicly available dataset, introducing nonindependence and potentially affecting the overall diagnostic accuracy. Furthermore, this overlap in datasets could decrease study heterogeneity and generalizability. A total of 5 studies shared the Sleep Heart Health Study (SHHS 1 and 2) datasets, 4 studies shared the Clínico de Santiago de Compostela, Spain dataset, 2 studies shared the Multiethnic Study of Atherosclerosis and the Osteoporotic Fractures in Men Study datasets, while 2 studies shared the University College of Dublin dataset (Table S4B in Multimedia Appendix 1).

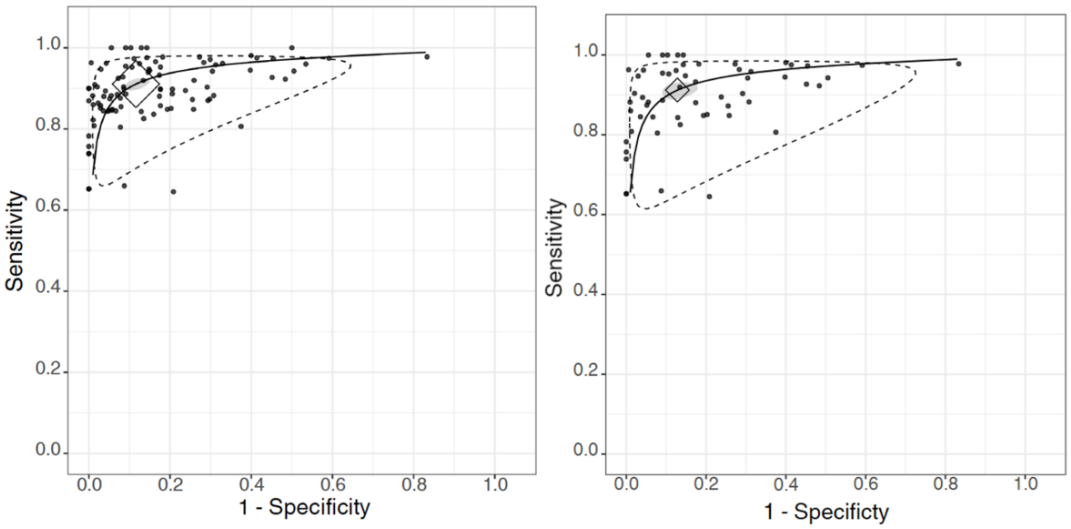
To maintain rigor and ensure validity of our meta-analytic results, a sensitivity analysis was done, ensuring that only one study per dataset was performed. To decide which study to exclude in the sensitivity analysis, we used the following criteria in decreasing order of priority:

1. Availability of an external, independent testing set (highest priority)
2. Largest test set
3. Publication date (lowest priority)

Availability of an external, independent test set was allocated the highest priority, as it directly evaluates the AI model in an independent patient population, restoring study independence. Test set size was then considered to help reduce random error. Finally, publication date was considered to ensure results reported were the latest and most up-to-date. Of the 11 studies containing shared datasets, 3 [49,54,56] of them were selected to be included in the sensitivity analysis, while the other 8 [20,40,44-48,55,57] were excluded.

After ensuring that each study had a unique dataset for training and testing, the pooled sensitivity and specificity had changed minimally. The sensitivity after exclusion was 91.2% (95% CrI 89.1%-93.1%) while specificity was 87.2% (95% CrI 82.4%-90.8%). This demonstrated that the sharing of datasets had minimal effect on the pooled sensitivity and specificity. Visual comparison of the SROC curves also depicts minimal shifts (Figure 9).

Figure 9. Original summary receiver operating characteristic (SROC) plot for the overall obstructive sleep apnea (OSA) diagnostic accuracy of artificial intelligence (AI) models trained on oxygen saturation (SpO_2) recordings (left) versus after excluding studies with duplicate datasets (right).



In Figure 9, the solid line represents the extrapolated SROC curve. The diamond represents the summary receiver operating point. Shaded/dashed regions represent the 95% CrI or prediction intervals. Unshaded circles/ovals are centered around individual study means; their height/width are proportionate to study weights for sensitivity/specificity.

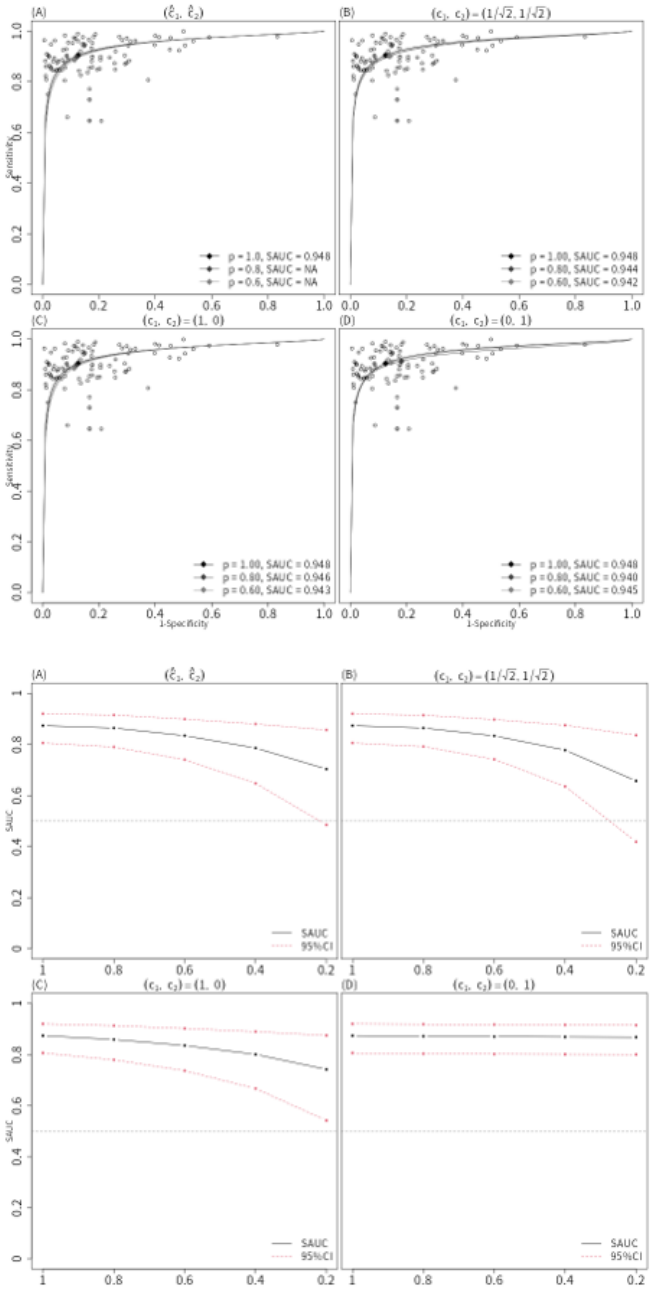
Reporting Biases

Publication Bias

Overall, the risk of bias due to missing results is judged to be low, as sensitivity analyses on the SROC curve and AUC suggested no clinically significant publication bias. When considering 4 different mechanisms of publication bias (data, sensitivity, specificity, or DOR-driven), with varying probabilities of unpublished studies (up to 40%), the SROC curve (Figure 10) and AUC (Figure 10) were almost constant, where the AUC shifted only modestly (0.902-0.877). This suggests that even if most studies remained unpublished, the conclusions of this meta-analysis would not have changed. For data-driven bias, sensitivity decreased from 81.5% to 78.7% and specificity from 86.1% to 84.1%. For sensitivity-driven bias, sensitivity declined from 81.5% to 78.6% and specificity from 86.1% to 84.2%. For specificity-driven bias, sensitivity declined from 81.5% to 74.4%, while specificity slightly increased from 86.1% to 87.3%. For DOR-driven bias, sensitivity rose slightly (81.5%-83.6%) but specificity fell (86.1%-78.2%). These shifts indicate that even under extreme assumptions of unpublished data, the summary estimates changed only modestly, and the overall conclusions of the meta-analysis remained unchanged.

In Figure 10, the potential impact of 4 different mechanisms of publication bias is shown: (A) data-driven, (B) sensitivity-driven, (C) specificity-driven, and (D) DOR-driven, with varying probabilities of unpublished studies (0%, 20%, and 40%). The solid lines and diamonds represent the SROC curves and summary operating points. The summary AUC denotes the area under the SROC curve. Black and red dots represent the mean and 95% CIs of the AUC for each probability of unpublished studies (x-axis).

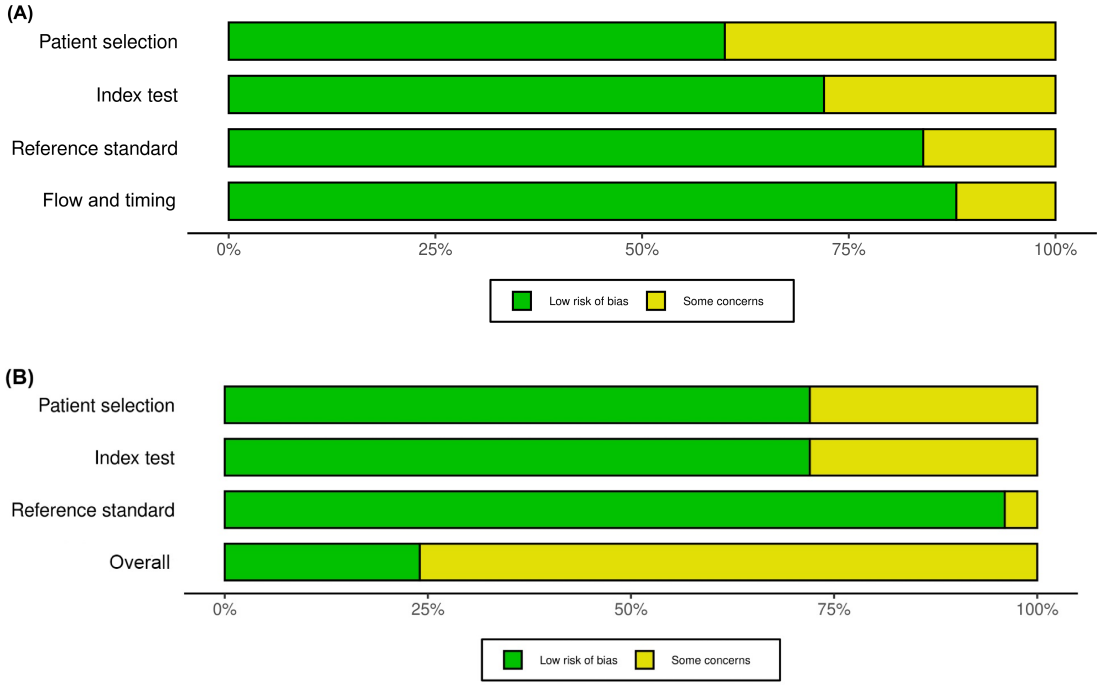
Figure 10. Effect of varying scenarios of publication bias on the (1) summary receiver operating characteristic (SROC) curve (parts A to D in top half of the figure) and (2) area under the curve (AUC) (parts A to D in the bottom half of the figure). SAUC: summary AUC.



Risk of Bias in Studies

Using QUADAS-2 to evaluate the risk of bias among our included studies demonstrated that within each domain, most studies had a low risk of bias. Details of the assessment are provided in Table S2 in Multimedia Appendix 1. The proportions of low-risk studies for patient selection, index test, reference standard, and flow and timing were 60%, 72%, 84%, and 84%, respectively (Figure 11A). Regarding applicability, 72%, 72%, and 96% of studies were scored as low risk for patient selection, index test, and reference standard domain (Figure 11B). Overall, these findings indicate that the included studies generally had low risk of bias and good applicability to the review question.

Figure 11. Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) visualization tool for (A) risk of bias assessment and (B) concerns regarding applicability.



Certainty of Evidence/Quality of Evidence

The quality of evidence at the outcome level is summarized in Table 2. The overall quality of evidence was high. There was clear evidence of a sensitivity-specificity relationship.

Overall, the results should be interpreted cautiously due to the observational nature of between-study comparisons and the potential for residual confounding from unmeasured study-level factors. Although QUADAS-2 assessments showed that most studies had low risk of bias across major domains, issues in patient selection were more common, reflecting the use of retrospective or convenience samples in several studies. Such sampling approaches may limit representativeness relative to real-world screening populations. Nonetheless, extensive publication-bias analyses demonstrated that even under extreme assumptions of missing studies, the SROC curve and AUC shifted only minimally, supporting the robustness of the main conclusions. However, as the random-split test set subgroup showed marginally lower accuracy, the quality of evidence according to the GRADE framework is only moderate, as external validation is still required to reliably evaluate model performance in new populations and settings.

Table 2. Evaluation of quality of pooled evidence using the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) framework (sensitivity: 0.91, 95% CI 0.90-0.92; specificity: 0.88, 95% CI 0.85-0.91; prevalences examined: 15%, 30%, and 60%). The question was as follows: what is the accuracy of artificial intelligence trained on overnight SpO_2[a] recordings for diagnosing OSA[b] in adults?

Outcome	Number of studies; number of patients	Study design	Factors that may decrease certainty of evidence					Effect per 1000 patients tested			Test accuracy CoE[c]
			Risk of bias	Indirectness	Inconsistency	Imprecision	Publication bias	Pretest probability of 15%, mean (95% CI)	Pretest probability of 30%, mean (range)	Pretest probability of 60%, mean (range)	
• True positives (patients with OSA) • False negatives (patients incorrectly classified as not having OSA)	25 [12,13,20,37-58] studies; 23,171 patients	Cross-sectional (cohort type accuracy study)	Serious	Not serious	Not serious	Not serious	Dose response gradient	• 137 (135-139) • 13 (11-15)	• 27 (23-31) • 273 (269-277)	• 547 (539-554) • 53 (46-61)	⊕⊕⊕⊕[d] High
• True negatives (patients without OSA) • False positives (patients incorrectly classified as having OSA)	25 [12,13,20,37-58] studies; 23,171 patients	Cross-sectional (cohort type accuracy study)	Serious	Not serious	Not serious	Not serious	Dose response gradient	• 751 (725-772) • 99 (78-125)	• 619 (597-636) • 81 (64-103)	• 354 (341-363) • 46 (37-59)	⊕⊕⊕⊕ High

[a]SpO_2: oxygen saturation.
[b]OSA: obstructive sleep apnea.
[c]CoE: certainty of evidence.
[d]Grading of Recommendations Assessment, Development, and Evaluation. The four symbols (⊕⊕⊕⊕) indicate high certainty of evidence and are a standard component of GRADE assessment tables.

Discussion

Principal Findings

This study quantitatively pooled the diagnostic performance metrics of 97 AI models in diagnosing OSA using pulse oximeter readings and explored factors influencing accuracy via meta-regression. Overall, AI models demonstrated high accuracy that could rival polysomnography alternatives. Neural network classifiers were associated with the greatest sensitivity and specificity, while models using deep learning for feature extraction outperformed those relying on domain expert features. Models had higher specificity when predicting more severe OSA, and sensitivity remained high at low AHI thresholds. There was no evidence to suggest publication bias.

To our knowledge, this is the first review to estimate the pooled diagnostic accuracy of AI models trained on oximetry data. Prior reviews were largely narrative, broader in scope, or focused on other input features such as facial features, speech, or heart rate [61-64]. This review advances the field by offering a clearer and more reliable evidence base to support the potential of oximetry-based AI as a scalable tool for OSA screening and diagnosis. Potential uses for such AI models include the rapid and convenient diagnosis of OSA not just in the primary health care setting but also in inpatients who are admitted for complications known to be associated with OSA, such as acute myocardial infarctions, strokes, or arrhythmias like atrial fibrillation [65]. This could potentially facilitate earlier diagnosis of OSA and thus earlier treatment, preventing further complications of the disease [66].

The pooled sensitivity and specificity of AI oximetry models are comparable to or exceeding existing polysomnography alternatives such as ODI and HSAT. Notably, AI approaches appear to outperform the use of ODI alone, which is often only one feature among several used to train these models. For example, Marcos et al [45] reported that a neural network achieved 31% higher sensitivity than ODI3, with only a modest 12% reduction in specificity. Commercial oximetry devices such as the Wellue O_2 Ring and Samsung Galaxy Watch 4 demonstrate lower diagnostic performance, with reported sensitivity and specificity of 87% and 78% (ODI 11) and 90% and 64% (ODI 5), respectively, at their optimal thresholds [67,68]. In contrast, this meta-analysis found higher pooled sensitivity (91.1%) and specificity (88.4%) for AI-driven oximetry. Compared with photoplethysmography-based HSAT devices such as WatchPAT (Itamar Medical Ltd), AI oximetry also performed better at lower AHI cutoffs, with higher specificity and similar sensitivity [69]. These findings suggest that AI oximetry may improve diagnostic performance over traditional ODI and HSAT metrics. However, whether this translates into real-world clinical benefit remains uncertain, and prospective studies are needed to confirm accuracy, usability, and integration into clinical workflows outside research settings.

AI models outperform ODI likely because they incorporate richer temporal and morphological information from SpO_2 signals, while deep learning approaches further enhance performance beyond traditional machine learning. The ODI may fail to detect apneas that do not lead to desaturations beyond the predefined threshold [70]. In contrast, AI models may capture additional pulse oximetry features, often using dynamic or differential desaturation thresholds across the recording to capture nonlinear patterns [71]. The slope, shape, interval, and depth of desaturation can also be incorporated, unlike in ODI. Even among the AI models, the superior sensitivity and specificity of neural network classifiers may be attributed to deep learning's ability to automatically learn intricate patterns from raw data through multiple representation layers, surpassing traditional machine learning approaches [72]. This enables extraction of higher-level features and identification of complex desaturation patterns over extended signal segments, which may even be imperceptible to humans [73-75]. Together, these mechanisms explain both the advantage of AI over ODI and the additional performance gains achieved by deep learning over conventional machine learning approaches.

The lower diagnostic accuracy of a random-split test model could perhaps be explained by the use of external validation datasets. The studies by Gutierrez-Tobal et al [40] and Nikkonen et al [50] were grouped under random-split for meta-regression purposes. However, unlike the other studies that used internal validation, they implemented an external dataset for testing that was not used for training at all. Lower accuracy is expected when models are tested on external datasets from different populations, although this is likely more representative of real-world performance across heterogeneous settings. This helps to prevent overfitting, which may be associated with machine learning models using internal validation, generating overly optimistic accuracy data [76,77]. The predominance of internal validation in included studies highlights the need for more external validation to strengthen generalizability in OSA diagnosis.

The strengths of our study include the use of state-of-the-art Bayesian bivariate random-effects meta-analytic methods, allowing joint estimation of sensitivity and specificity while accounting for between-study variability. This approach supports precise and robust pooled estimates of diagnostic accuracy. Sensitivity analyses using selection models showed that even under extreme assumptions of publication bias, changes in AUC, sensitivity, and specificity were modest, and the overall conclusions remained unchanged. Risk of bias (QUADAS-2) assessment indicated that most studies were at low risk across domains, with generally limited applicability concerns, while overall quality of evidence (GRADE) was high. Taken together, these findings support the robustness and reliability of the pooled estimates.

Nonetheless, several limitations of this study should be considered. These can be broadly grouped into issues affecting internal validity and external validity. Limitations related to internal validity include challenges inherent to the AI diagnostic literature and meta-analytic modeling. Unlike conventional clinical studies, AI models often report only the best-performing configuration among multiple internally tested variants, with selective emphasis on favorable

thresholds, architectures, or preprocessing pipelines [78,79]. Consequently, although our publication-bias sensitivity analyses using selection models did not identify substantial evidence of bias, the ability of traditional frameworks to detect selective reporting in AI diagnostic research remains limited [80]. In addition, the retrospective nature of most included studies may introduce bias in data collection and model evaluation, as model development and testing were conducted on preexisting datasets rather than prospectively collected data. Device-level variability (eg, oximeter type, sampling rate, and signal processing) also contributed to heterogeneity [81], since the included studies used 6 different brands of pulse oximeters, and one-third of studies did not report the brand. Several studies contributed multiple models or thresholds, which potentially introduced within-study correlation that could not be fully accounted for within the present modeling package. The potential for overestimated precision should thus be considered when interpreting the results. Furthermore, formal quantification of the proportion of variability due to statistical heterogeneity (ie, I^2) was not available in the current Bayesian statistical package [82], although visualization of prediction intervals and meta-regression suggested broadly consistent effects.

Limitations affecting external validity relate primarily to generalizability to routine clinical practice. Pulse oximetry is subject to known physiological and technical limitations, including motion artifact, probe displacement, and interference from factors such as skin pigmentation, nail varnish, dirt, temperature, anemia, and hemoglobinopathies [83,84]. In addition, it cannot directly detect cortical arousals or reliably distinguish obstructive from central apneas, which may lead to misclassification or underestimation of disease severity [85]. Generalizability is limited by the predominance of sleep clinic and hospital-based cohorts with high pretest probability of OSA, which reduces applicability to low-prevalence settings such as primary care. Although meta-regression suggested relatively stable sensitivity and improved specificity in lower-prevalence contexts, these findings should be interpreted cautiously given the retrospective nature of all included studies. The absence of prospective real-world validation further limits the ability to draw firm conclusions about clinical readiness, particularly in primary care or acute care environments where case mix, noise, and workflow constraints differ substantially from research settings. Additional limitations include exclusion of non-English studies, which may introduce selection bias [86], and the use of continent of study as a proxy for ethnicity due to missing individual-level data, which may mask important interethnic differences in SpO_2 accuracy [87]. Underrepresentation of certain regions, particularly Africa and South America, further limits global generalizability. Finally, reliance on publicly available datasets may introduce regional over-representation, limiting transferability across diverse health care systems.

Future studies evaluating the potential of AI for OSA diagnosis should adopt prospective study designs, with control for pulse oximeter brands and inclusion of general population cohorts encompassing both obstructive and central sleep apnea. Greater representation from underrepresented regions such as South America and Africa is also needed, alongside dedicated studies in pediatric populations where diagnosis remains challenging due to the need for hospitalization and complex equipment. Further work should also explore the role of AI in complementing polysomnography for risk stratification; for example, by leveraging high sensitivity at low AHI thresholds to rule out OSA and high specificity at higher thresholds to support case identification.

Conclusion

AI-oximetry models showed high diagnostic accuracy for OSA across models and AHI cutoffs, performing better than or comparably to traditional overnight oximetry and HSATs. This review is innovative because it provides the first pooled quantitative synthesis of AI models trained solely on oximetry data, with additional evaluation of publication bias and methodological limitations. Unlike prior reviews, which are largely narrative or broader in scope and include heterogeneous AI inputs such as polysomnography-derived features or multimodal data, this review focuses specifically on pooled diagnostic accuracy and publication bias in AI-oximetry studies. It advances the field by offering a clearer and more reliable evidence base on AI-oximetry performance. These findings support the potential of oximetry-based AI as a scalable, low-cost screening tool for OSA, with potential real-world applications in both primary care and inpatient settings for early identification of high-risk patients. Prospective external validation in diverse populations and low-prevalence settings is still needed before widespread real-world use.

Acknowledgments

We thank Dr Brian Sheng Yep Yeo and Dr Claire Jing-Wen Tan for their assistance with retrieving abstracts for a search update.

Generative artificial intelligence was not used in any portion of the manuscript.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors, and there was no funding involvement in the study design, data collection, analysis, interpretation, or writing of the manuscript.

Data Availability

The data analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: STT, BKJT, EYG
Data curation: BKJT, EYG
Formal analysis: EYG, BKJT, JHK
Funding acquisition: KJMY, CYJL
Investigation: KJMY, CYJL, EYG, BKJT, JHK
Methodology: STT, BKJT, EYG
Project administration: STT, BKJT, EYG
Resources: STT, BKJT, EYG
Software: KJMY, CYJL, EYG, BKJT, JHK
Supervision: STT, BKJT, NKWT, ACWN, ZHL, CQP, THO, LCL, GBH
Validation: KJMY, CYJL, EYG
Visualization: KJMY, CYJL, EYG
Writing – original draft: KJMY, CYJL, EYG, BKJT
Writing – review & editing: KJMY, CYJL, EYG, BKJT
STT and BKJT contributed equally as co-senior and co-corresponding authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Overnight pulse oximetry for artificial intelligence diagnosis of obstructive sleep apnea: a Bayesian meta-analysis (online supplement).
[DOCX File (Microsoft Word File), 5202 KB-Multimedia Appendix 1]

Checklist 1

PRISMA 2020 for Abstracts checklist.
[DOCX File (Microsoft Word File), 271 KB-Checklist 1]

Checklist 2

PRISMA-S checklist.
[DOCX File (Microsoft Word File), 20 KB-Checklist 2]

Checklist 3

PRISMA Expanded checklist.
[DOCX File (Microsoft Word File), 5906 KB-Checklist 3]

References

1. Senaratna CV, Perret JL, Lodge CJ, et al. Prevalence of obstructive sleep apnea in the general population: a systematic review. Sleep Med Rev. Aug 2017;34:70-81. [doi: 10.1016/j.smrv.2016.07.002] [Medline: 27568340]
2. Koh JH, Lim CYJ, Yam KJM, et al. Bidirectional association of sleep disorders with chronic kidney disease: a systematic review and meta-analysis. Clin Kidney J. Nov 2024;17(11):sfae279. [doi: 10.1093/ckj/sfae279] [Medline: 39525685]
3. Teo YH, Tan BKJ, Tan NKW, et al. Obstructive sleep apnea and the incidence and mortality of gastrointestinal cancers: a systematic review and meta-analysis of 5,120,837 participants. J Gastrointest Oncol. Dec 2022;13(6):2789-2798. [doi: 10.21037/jgo-22-153] [Medline: 36636076]
4. Tan NKW, Yap DWT, Tan BKJ, et al. The association of obstructive sleep apnea with melanoma incidence and mortality: a meta-analysis of 5,276,451 patients. Sleep Med. Dec 2021;88:213-220. [doi: 10.1016/j.sleep.2021.10.027] [Medline: 34794048]
5. Tan BKJ, Teo YH, Tan NKW, et al. Association of obstructive sleep apnea and nocturnal hypoxemia with all-cancer incidence and mortality: a systematic review and meta-analysis. J Clin Sleep Med. May 1, 2022;18(5):1427-1440. [doi: 10.5664/jcsm.9772] [Medline: 34755597]
6. Tan BKJ, Tan NKW, Teo YH, et al. Association of obstructive sleep apnea with thyroid cancer incidence: a systematic review and meta-analysis. Eur Arch Oto Rhino Laryngol. Nov 2022;279(11):5407-5414. [doi: 10.1007/s00405-022-07457-w] [Medline: 35708764]
7. Iannella G, Pace A, Bellizzi MG, et al. The global burden of obstructive sleep apnea. Diagnostics (Basel). Apr 25, 2025;15(9):1088. [doi: 10.3390/diagnostics15091088] [Medline: 40361906]
8. Goyal M, Johnson J. Obstructive sleep apnea diagnosis and management. Mo Med. 2017;114(2):120-124. [Medline: 30228558]

9. Middle income countries. World Bank Group. 2024. URL: https://www.worldbank.org/en/country/mic/overview [Accessed 2026-05-30]
10. Pivetta B, Chen L, Nagappa M, et al. Use and performance of the STOP-Bang questionnaire for obstructive sleep apnea screening across geographic regions: a systematic review and meta-analysis. JAMA Netw Open. Mar 1, 2021;4(3):e211009. [doi: 10.1001/jamanetworkopen.2021.1009] [Medline: 33683333]
11. Thornton CS, Tsai WH, Santana MJ, et al. Effects of wait times on treatment adherence and clinical outcomes in patients with severe sleep-disordered breathing: a secondary analysis of a noninferiority randomized clinical trial. JAMA Netw Open. Apr 1, 2020;3(4):e203088. [doi: 10.1001/jamanetworkopen.2020.3088] [Medline: 32310283]
12. Andrés-Blanco AM, Álvarez D, Crespo A, et al. Assessment of automated analysis of portable oximetry as a screening test for moderate-to-severe sleep apnea in patients with chronic obstructive pulmonary disease. PLoS One. 2017;12(11):e0188094. [doi: 10.1371/journal.pone.0188094] [Medline: 29176802]
13. Behar JA, Palmius N, Li Q, et al. Feasibility of single channel oximetry for mass screening of obstructive sleep apnea. EClinicalMedicine. 2019;11:81-88. [doi: 10.1016/j.eclinm.2019.05.015] [Medline: 31317133]
14. Levy J, Álvarez D, Del Campo F, Behar JA. Deep learning for obstructive sleep apnea diagnosis based on single channel oximetry. Nat Commun. Aug 12, 2023;14(1):4881. [doi: 10.1038/s41467-023-40604-3] [Medline: 37573327]
15. Álvarez D, Gutiérrez-Tobal GC, Vaquerizo-Villar F, Moreno F, Del Campo F, Hornero R. Oximetry indices in the management of sleep apnea: from overnight minimum saturation to the novel hypoxemia measures. Adv Exp Med Biol. 2022;1384:219-239. [doi: 10.1007/978-3-031-06413-5_13] [Medline: 36217087]
16. Wang Y, Fietze I, Salanitro M, Penzel T. Comparison of the value of the STOP-BANG questionnaire with oxygen desaturation index in screening obstructive sleep apnea in Germany. Sleep Breath. Aug 2023;27(4):1315-1323. [doi: 10.1007/s11325-022-02727-7] [Medline: 36269514]
17. Aaronson JA, van Bezeij T, van den Aardweg JG, van Bennekom CAM, Hofman WF. Diagnostic accuracy of nocturnal oximetry for detection of sleep apnea syndrome in stroke rehabilitation. Stroke. Sep 2012;43(9):2491-2493. [doi: 10.1161/STROKEAHA.112.665414] [Medline: 22821607]
18. Zhang Z, Qi M, Hügli G, Khatami R. The challenges and pitfalls of detecting sleep hypopnea using a wearable optical sensor: comparative study. J Med Internet Res. Jul 29, 2021;23(7):e24171. [doi: 10.2196/24171] [Medline: 34326039]
19. Thorisdottir K, Leppänen T, Hrubos-Strøm H, et al. Hypoxic load and average oxygen-saturation during sleep and wake are associated with cognitive function in obstructive sleep apnea. J Sleep Res. Feb 2026;35(1):e70136. [doi: 10.1111/jsr.70136] [Medline: 40610382]
20. Chen JW, Liu CM, Wang CY, et al. A deep neural network-based model for OSA severity classification using unsegmented peripheral oxygen saturation signals. Eng Appl Artif Intell. Jun 2023;122:106161. [doi: 10.1016/j.engappai.2023.106161]
21. Stroup DF, Berlin JA, Morton SC, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis of observational studies in epidemiology (MOOSE) group. JAMA. Apr 19, 2000;283(15):2008-2012. [doi: 10.1001/jama.283.15.2008] [Medline: 10789670]
22. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. Mar 29, 2021;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
23. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. J Med Libr Assoc. Apr 1, 2021;109(2):174-200. [doi: 10.5195/jmla.2021.962] [Medline: 34285662]
24. Tan BKJ, Gao EY, Tan NKW, et al. Machine listening for OSA diagnosis: a Bayesian meta-analysis. Chest. Aug 2025;168(2):520-530. [doi: 10.1016/j.chest.2025.04.006] [Medline: 40220991]
25. Forbes C, Greenwood H, Carter M, Clark J. Automation of duplicate record detection for systematic reviews: deduplicator. Syst Rev. Aug 2, 2024;13(1):206. [doi: 10.1186/s13643-024-02619-9] [Medline: 39095913]
26. Bramer WM, Giustini D, de Jonge GB, Holland L, Bekhuis T. De-duplication of database search results for systematic reviews in EndNote. J Med Libr Assoc. Jul 2016;104(3):240-243. [doi: 10.3163/1536-5050.104.3.014] [Medline: 27366130]
27. Harrison H, Griffin SJ, Kuhn I, Usher-Smith JA. Software tools to support title and abstract screening for systematic reviews in healthcare: an evaluation. BMC Med Res Methodol. Jan 13, 2020;20(1):7. [doi: 10.1186/s12874-020-0897-3] [Medline: 31931747]
28. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. Oct 18, 2011;155(8):529-536. [doi: 10.7326/0003-4819-155-8-201110180-00009] [Medline: 22007046]
29. Cerullo E, Sutton AJ, Jones HE, Wu O, Quinn TJ, Cooper NJ. MetaBayesDTA: codeless Bayesian meta-analysis of test accuracy, with or without a gold standard. BMC Med Res Methodol. May 25, 2023;23(1):127. [doi: 10.1186/s12874-023-01910-y] [Medline: 37231347]

30. Mizutani S, Zhou Y, Tian YS, Takagi T, Ohkubo T, Hattori S. DTAmetasa: an R shiny application for meta-analysis of diagnostic test accuracy and sensitivity analysis of publication bias. Res Synth Methods. Nov 2023;14(6):916-925. [doi: 10.1002/jrsm.1666] [Medline: 37640914]
31. Freeman SC, Kerby CR, Patel A, Cooper NJ, Quinn T, Sutton AJ. Development of an interactive web-based tool to conduct and interrogate meta-analysis of diagnostic test accuracy studies: MetaDTA. BMC Med Res Methodol. Apr 18, 2019;19(1):81. [doi: 10.1186/s12874-019-0724-x] [Medline: 30999861]
32. Patel A, Cooper N, Freeman S, Sutton A. Graphical enhancements to summary receiver operating characteristic plots to facilitate the analysis and reporting of meta-analysis of diagnostic test accuracy data. Res Synth Methods. Jan 2021;12(1):34-44. [doi: 10.1002/jrsm.1439] [Medline: 32706182]
33. Zhou Y, Huang A, Hattori S. A likelihood-based sensitivity analysis for publication bias on the summary receiver operating characteristic in meta-analysis of diagnostic test accuracy. Stat Med. Mar 15, 2023;42(6):781-798. [doi: 10.1002/sim.9643] [Medline: 36584693]
34. Carpenter B, Gelman A, Hoffman MD, et al. Stan: a probabilistic programming language. J Stat Softw. 2017;76:1. [doi: 10.18637/jss.v076.i01] [Medline: 36568334]
35. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ. Apr 26, 2008;336(7650):924-926. [doi: 10.1136/bmj.39489.470347.AD] [Medline: 18436948]
36. The R Project for Statistical Computing. R Foundation for Statistical Computing. 2021. URL: https://www.r-project.org/? [Accessed 2026-06-18]
37. Álvarez D, Cerezo-Hernández A, Crespo A, et al. A machine learning-based test for adult sleep apnoea screening at home using oximetry and airflow. Sci Rep. Mar 24, 2020;10(1):5332. [doi: 10.1038/s41598-020-62223-4] [Medline: 32210294]
38. Alvarez D, Gutierrez-Tobal GC, Vaquerizo-Villar F, et al. Automated analysis of unattended portable oximetry by means of Bayesian neural networks to assist in the diagnosis of sleep apnea. Presented at: 2016 Global Medical Engineering Physics Exchanges/Pan American Health Care Exchanges (GMEPE/PAHCE); Apr 4-9, 2016:79-82; Madrid, Spain. [doi: 10.1109/GMEPE-PAHCE.2016.7504628]
39. Cajal D, Gil E, Laguna P, et al. Obstructive sleep apnea screening by joint saturation signal analysis and PPG-derived pulse rate oscillations. IEEE J Biomed Health Inform. Nov 10, 2023;PP. [doi: 10.1109/JBHI.2023.3331947] [Medline: 37948138]
40. Gutiérrez-Tobal GC, Álvarez D, Vaquerizo-Villar F, et al. Ensemble-learning regression to estimate sleep apnea severity using at-home oximetry in adults. Appl Soft Comput. Nov 2021;111:107827. [doi: 10.1016/j.asoc.2021.107827] [Medline: 39544517]
41. Kaimakamis E, Bratsas C, Sichletidis L, Karvounis C, Maglaveras N. Screening of patients with obstructive sleep apnea syndrome using C4.5 algorithm based on non linear analysis of respiratory signals during sleep. Annu Int Conf IEEE Eng Med Biol Soc. 2009;2009:3465-3469. [doi: 10.1109/IEMBS.2009.5334605] [Medline: 19964987]
42. Leong ZH, Loh SRH, Leow LC, Ong TH, Toh ST. A machine learning approach for the diagnosis of obstructive sleep apnoea using oximetry, demographic and anthropometric data. Singapore Med J. Apr 1, 2025;66(4):195-201. [doi: 10.4103/singaporemedj.SMJ-2022-170] [Medline: 37171440]
43. Li Z, Li Y, Zhao G, Zhang X, Xu W, Han D. A model for obstructive sleep apnea detection using a multi-layer feed-forward neural network based on electrocardiogram, pulse oxygen saturation, and body mass index. Sleep Breath. Dec 2021;25(4):2065-2072. [doi: 10.1007/s11325-021-02302-6] [Medline: 33754247]
44. Ma B, Wu Z, Li S, et al. A SVM-based algorithm to diagnose sleep apnea. Presented at: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); Nov 18-21, 2019:1556-1560; San Diego, CA, USA. [doi: 10.1109/BIBM47256.2019.8983201]
45. Marcos JV, Hornero R, Alvarez D, Del Campo F, López M. Applying neural network classifiers in the diagnosis of the obstructive sleep apnea syndrome from nocturnal pulse oximetric recordings. Annu Int Conf IEEE Eng Med Biol Soc. 2007;2007:5174-5177. [doi: 10.1109/IEMBS.2007.4353507] [Medline: 18003173]
46. Marcos JV, Hornero R, Alvarez D, del Campo F, López M, Zamarrón C. Radial basis function classifiers to help in the diagnosis of the obstructive sleep apnoea syndrome from nocturnal oximetry. Med Biol Eng Comput. Apr 2008;46(4):323-332. [doi: 10.1007/s11517-007-0280-0] [Medline: 17968604]
47. Marcos JV, Hornero R, Alvarez D, Del Campo F, Zamarrón C, López M. Utility of multilayer perceptron neural network classifiers in the diagnosis of the obstructive sleep apnoea syndrome from nocturnal oximetry. Comput Methods Programs Biomed. Oct 2008;92(1):79-89. [doi: 10.1016/j.cmpb.2008.05.006] [Medline: 18672313]
48. Marcos JV, Hornero R, Alvarez D, Del Campo F, Zamarron C. A classification algorithm based on spectral features from nocturnal oximetry and support vector machines to assist in the diagnosis of obstructive sleep apnea. Presented at: 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2009); Sep 3-6, 2009:5547-5550; Minneapolis, MN. [doi: 10.1109/IEMBS.2009.5333731]

49. Marcos JV, Hornero R, Alvarez D, Nabney IT, Del Campo F, Zamarrón C. The classification of oximetry signals using Bayesian neural networks to assist in the detection of obstructive sleep apnoea syndrome. Physiol Meas. Mar 2010;31(3):375-394. [doi: 10.1088/0967-3334/31/3/007] [Medline: 20130342]
50. Nikkonen S, Afara IO, Leppänen T, Töyräs J. Artificial neural network analysis of the oxygen saturation signal enables accurate diagnostics of sleep apnea. Sci Rep. Sep 13, 2019;9(1):13200. [doi: 10.1038/s41598-019-49330-7] [Medline: 31519927]
51. Peng D, Yue H, Tan W, et al. A bimodal feature fusion convolutional neural network for detecting obstructive sleep apnea/hypopnea from nasal airflow and oximetry signals. Artif Intell Med. Apr 2024;150:102808. [doi: 10.1016/j.artmed.2024.102808] [Medline: 38553148]
52. Polat K, Yosunkaya S, Güneş S. Pairwise ANFIS approach to determining the disorder degree of obstructive sleep apnea syndrome. J Med Syst. Oct 2008;32(5):379-387. [doi: 10.1007/s10916-008-9143-y] [Medline: 18814494]
53. Wu HT, Wu JC, Huang PC, et al. Phenotype-based and self-learning inter-individual sleep apnea screening with a level IV-like monitoring system. Front Physiol. 2018;9:723. [doi: 10.3389/fphys.2018.00723] [Medline: 30013479]
54. Zhang M, Dong C, Zhang D, Tseng ML, Wei J. An intelligent classification diagnosis based on blood oxygen saturation signals for medical data security including COVID-19 in industry 5.0. IEEE Trans Ind Inf. Mar 2023;19(3):3310-3320. [doi: 10.1109/TII.2022.3152809]
55. Wu YC, Yeh CY, Lin CC. Severity prediction of obstructive sleep apnea using transformed 2D oxygen saturation signals. Sens Mater. 2025;37(12):5535. [doi: 10.18494/SAM5765]
56. Muthukumaran MP, Nnamdi MC, Tamo JB, Purnell C, Wang MD. Developing an attention-based deep learning framework for obstructive sleep apnea detection using single-channel oximetry signal. Presented at: 2025 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI); Oct 26-29, 2025:1-7; Atlanta, GA, USA. [doi: 10.1109/BHI67747.2025.11269547]
57. Li C, He S, Xu X, Wang Z. Deep model based on Mamba fusion multi-scale convolution LSTM for OSA severity grading. Appl Sci (Basel). Dec 9, 2025;15(24):12990. [doi: 10.3390/app152412990]
58. Kuo NY, Tsai HJ, Tsai SJ, Yang AC. Efficient screening in obstructive sleep apnea using sequential machine learning models, questionnaires, and pulse oximetry signals: mixed methods study. J Med Internet Res. Dec 19, 2024;26:e51615. [doi: 10.2196/51615] [Medline: 39699950]
59. Almarshad MA, Al-Ahmadi S, Islam MS, BaHammam AS, Soudani A. Adoption of transformer neural network to improve the diagnostic performance of oximetry for obstructive sleep apnea. Sensors (Basel). Sep 15, 2023;23(18):7924. [doi: 10.3390/s23187924] [Medline: 37765980]
60. Li X, Leung FHF, Su S, Ling SH. Sleep apnea detection using multi-error-reduction classification system with multiple bio-signals. Sensors (Basel). Jul 25, 2022;22(15):5560. [doi: 10.3390/s22155560] [Medline: 35898064]
61. Rani S, Gao EY, Ong JZE, et al. Multichannel machine learning for polysomnographic diagnosis of obstructive sleep apnea: a Bayesian meta-analysis. Eur Arch Otorhinolaryngol. Apr 2026;283(4):2107-2116. [doi: 10.1007/s00405-025-09808-9] [Medline: 41243013]
62. Hao Y, Tan NKW, Gao EY, et al. Electrocardiogram heart rate variability for machine learning diagnosis of obstructive sleep apnoea: a Bayesian meta-analysis. Sleep Breath. Sep 30, 2025;29(5):303. [doi: 10.1007/s11325-025-03476-z] [Medline: 41026252]
63. Gao EY, Tan BKJ, Tan NKW, et al. Artificial intelligence facial recognition of obstructive sleep apnea: a Bayesian meta-analysis. Sleep Breath. Nov 30, 2024;29(1):36. [doi: 10.1007/s11325-024-03173-3] [Medline: 39614959]
64. Gao EY, Hao Y, Tan NKW, et al. Awake speech recordings for machine learning diagnosis of obstructive sleep apnea: a Bayesian meta-analysis. J Clin Sleep Med. Dec 8, 2025;22(1):5. [doi: 10.1007/s44470-025-00013-3] [Medline: 41678061]
65. Wang Q, Zeng H, Dai J, Zhang M, Shen P. Association between obstructive sleep apnea and multiple adverse clinical outcomes: evidence from an umbrella review. Front Med (Lausanne). 2025;12:1497703. [doi: 10.3389/fmed.2025.1497703] [Medline: 40166062]
66. Li H, Pan Y, Lou Y, et al. The effects of continuous positive airway pressure therapy for secondary cardiovascular prevention in patients with obstructive sleep apnoea: a systematic review and meta-analysis. Rev Cardiovasc Med. Jun 2022;23(6):195. [doi: 10.31083/j.rcm2306195] [Medline: 39077164]
67. Tisyakorn J, Saiphoklang N, Sapankaew T, et al. Screening moderate to severe obstructive sleep apnea with wearable device. Sleep Breath. Dec 17, 2024;29(1):61. [doi: 10.1007/s11325-024-03232-9] [Medline: 39688783]
68. Jung H, Kim D, Lee W, et al. Performance evaluation of a wrist-worn reflectance pulse oximeter during sleep. Sleep Health. Oct 2022;8(5):420-428. [doi: 10.1016/j.sleh.2022.04.003] [Medline: 35817700]
69. Iftikhar IH, Finch CE, Shah AS, Augunstein CA, Ioachimescu OC. A meta-analysis of diagnostic test performance of peripheral arterial tonometry studies. J Clin Sleep Med. Apr 1, 2022;18(4):1093-1102. [doi: 10.5664/jcsm.9808] [Medline: 34879903]

70. Riha RL. Defining obstructive sleep apnoea syndrome: a failure of semantic rules. Breathe (Sheff). Sep 2021;17(3):210082. [doi: 10.1183/20734735.0082-2021] [Medline: 35035552]
71. Gutierrez-Tobal GC, Alvarez D, Crespo A, Del Campo F, Hornero R. Evaluation of machine-learning approaches to estimate sleep apnea severity from at-home oximetry recordings. IEEE J Biomed Health Inform. Mar 2019;23(2):882-892. [doi: 10.1109/JBHI.2018.2823384] [Medline: 29993673]
72. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature New Biol. May 28, 2015;521(7553):436-444. [doi: 10.1038/nature14539] [Medline: 26017442]
73. Brouillette RT, Morielli A, Leimanis A, Waters KA, Luciano R, Ducharme FM. Nocturnal pulse oximetry as an abbreviated testing modality for pediatric obstructive sleep apnea. Pediatrics. Feb 2000;105(2):405-412. [doi: 10.1542/peds.105.2.405] [Medline: 10654964]
74. Vaquerizo-Villar F, Alvarez D, Kheirandish-Gozal L, et al. A convolutional neural network architecture to enhance oximetry ability to diagnose pediatric obstructive sleep apnea. IEEE J Biomed Health Inform. Aug 2021;25(8):2906-2916. [doi: 10.1109/JBHI.2020.3048901] [Medline: 33406046]
75. Cabanas AM, Sáez N, Collao-Caiconte PO, et al. Evaluating AI methods for pulse oximetry: performance, clinical accuracy, and comprehensive bias analysis. Bioengineering (Basel). Oct 24, 2024;11(11):1061. [doi: 10.3390/bioengineering11111061] [Medline: 39593722]
76. Ho SY, Phua K, Wong L, Bin Goh WW. Extensions of the external validation for checking learned model interpretability and generalizability. Patterns (N Y). Nov 13, 2020;1(8):100129. [doi: 10.1016/j.patter.2020.100129] [Medline: 33294870]
77. Gallitto G, Englert R, Kincses B, et al. External validation of machine learning models-registered models and adaptive sample splitting. Gigascience. Jan 6, 2025;14:giaf036. [doi: 10.1093/gigascience/giaf036]
78. Ueda D, Kakinuma T, Fujita S, et al. Fairness of artificial intelligence in healthcare: review and recommendations. Jpn J Radiol. Jan 2024;42(1):3-15. [doi: 10.1007/s11604-023-01474-3] [Medline: 37540463]
79. Takita H, Kabata D, Walston SL, et al. A systematic review and meta-analysis of diagnostic performance comparison between generative AI and physicians. NPJ Digit Med. Mar 22, 2025;8(1):175. [doi: 10.1038/s41746-025-01543-z] [Medline: 40121370]
80. Rodgers MA, Pustejovsky JE. Evaluating meta-analytic methods to detect selective reporting in the presence of dependent effect sizes. Psychol Methods. Apr 2021;26(2):141-160. [doi: 10.1037/met0000300] [Medline: 32673040]
81. Harskamp RE, Bekker L, Himmelreich JCL, et al. Performance of popular pulse oximeters compared with simultaneous arterial oxygen saturation or clinical-grade pulse oximetry: a cross-sectional validation study in intensive care patients. BMJ Open Respir Res. Sep 2021;8(1):e000939. [doi: 10.1136/bmjresp-2021-000939] [Medline: 34489238]
82. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med. Jun 15, 2002;21(11):1539-1558. [doi: 10.1002/sim.1186] [Medline: 12111919]
83. Silverston P, Ferrari M, Quaresima V. Pulse oximetry in primary care: factors affecting accuracy and interpretation. Br J Gen Pract. Mar 2022;72(716):132-133. [doi: 10.3399/bjgp22X718769] [Medline: 35210248]
84. León-Valladares D, Barrio-Mateu LA, Cortés-Carmona N, et al. Determining factors of pulse oximetry accuracy: a literature review. Rev Clin Esp (Barc). May 2024;224(5):314-330. [doi: 10.1016/j.rceng.2024.04.005] [Medline: 38599519]
85. Fleetham J, Ayas N, Bradley D, et al. Canadian Thoracic Society guidelines: diagnosis and treatment of sleep disordered breathing in adults. Can Respir J. Oct 2006;13(7):387-392. [doi: 10.1155/2006/627096] [Medline: 17036094]
86. Vickers A, Goyal N, Harland R, Rees R. Do certain countries produce only positive results? A systematic review of controlled trials. Control Clin Trials. Apr 1998;19(2):159-166. [doi: 10.1016/s0197-2456(97)00150-5] [Medline: 9551280]
87. Al-Halawani R, Charlton PH, Qassem M, Kyriacou PA. A review of the effect of skin pigmentation on pulse oximeter accuracy. Physiol Meas. Jun 1, 2023;44(5):05TR01. [doi: 10.1088/1361-6579/acd51a] [Medline: 37172609]

Abbreviations

AHI: apnea-hypopnea index
AI: artificial intelligence
AUC: area under the curve
CrI: credible interval
DOR: diagnostic odds ratio
FN: false negatives
FP: false positives
GRADE: Grading of Recommendations Assessment, Development, and Evaluation
HSAT: home sleep apnea test

HSROC: hierarchical summary receiver operating characteristic
ODI: oxygen desaturation index
OSA: obstructive sleep apnea
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension
PROSPERO: International Prospective Register of Systematic Reviews
QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies-2
SHHS: Sleep Heart Health Study
SpO_2: oxygen saturation
SROC: summary receiver operating characteristic
STOP-BANG: Snoring, Tiredness, Observed apnoea, Pressure (hypertension), BMI>35 kg/m^2, Age>50, Neck circumference>40 cm, and Gender (male)
TN: true negatives
TP: true positives

Edited by Stefano Brini; peer-reviewed by Jeng-Wen Chen, Xiaolong Liang; submitted 09.Jul.2025; final revised version received 05.May.2026; accepted 06.May.2026; published 08.Jul.2026

Please cite as:
Yam KJM, Lim CYJ, Gao EY, Koh JH, Tan NKW, Ng ACW, Leong ZH, Phua CQ, Ong TH, Leow LC, Huang GB, Tan BKJ, Toh ST
Artificial Intelligence Diagnosis of Obstructive Sleep Apnea Using Overnight Pulse Oximetry: A Systematic Review and Bayesian Meta-Analysis
J Med Internet Res 2026;28:e80349
URL: https://www.jmir.org/2026/1/e80349
doi: 10.2196/80349