

Review

Accuracy of Artificial Intelligence in Diagnosing Obstructive Sleep Apnea Using Photoplethysmography: Systematic Review and Meta-Analysis

Brian Sheng Yeo^{1,2}, MBBS; Esther Yanxin Gao^{1,2}, MBBS, MRCS; Joan Ern Xin Tan¹; Jun Yuan Koo¹; Yi Siang Lee¹, MBBS; Nicole Kye Wen Tan¹, MBBS, MRCS; Adele Chin Wei Ng^{2,3}, MBBS, MRCS, MMed; Zhou Hao Leong², MBBS, MRCS, MMed; Thun How Ong^{3,4}, MBBS, MRCP; Leong Chai Leow^{3,4}, MD; Guang-Bin Huang⁵, PhD; Benjamin Kye Jyn Tan^{1,2}, MBBS (Hons), MRCS, MS; Song Tar Toh^{2,3}, MBBS, MRCS, MMed

¹Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

²Department of Otorhinolaryngology – Head and Neck Surgery, Singapore General Hospital, Singapore, Singapore

³Duke-NUS Sleep Centre, SingHealth, Singapore, Singapore

⁴Department of Respiratory and Critical Care Medicine, Singapore General Hospital, Singapore, Singapore

⁵School of Automation, Southeast University, Nanjing, China

Corresponding Author:

Song Tar Toh, MBBS, MRCS, MMed
Department of Otorhinolaryngology – Head and Neck Surgery
Singapore General Hospital
Outram Road
Singapore 169608
Phone: +6563214790
Email: toh.song.tar@singhealth.com.sg

Abstract

Background: The traditional method for diagnosing obstructive sleep apnea (OSA) through polysomnography may be expensive and inaccessible. Recent developments in AI propose the use of photoplethysmography (PPG) to aid OSA diagnosis.

Objective: This study seeks to evaluate the diagnostic accuracy of an AI-based approach in diagnosing OSA using PPG.

Methods: PubMed, Embase, Scopus, Web of Science, and IEEE Xplore were searched from inception to November 3, 2025. The inclusion criteria comprised observational studies that evaluated the accuracy of AI-based methods of OSA diagnosis using PPG compared with conventional sleep testing used in adults. We excluded case reports, case series, reviews, meta-analyses, letters, conference abstracts, pediatric studies, animal studies, foreign language studies, studies with incomplete data, and studies focusing on individual apneic events without patient-level classification. The outcome of interest was the diagnostic accuracy of PPG-based AI models for OSA, as assessed in primary studies using random split tests, cross-validation, or external validation. Independent reviewers extracted data and assessed risk of bias using the QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies-2) tool. A Bayesian bivariate meta-analysis was used to pool estimates. Further subgroup, sensitivity, and meta-regression analyses were conducted. Overall quality of evidence was assessed using the GRADE (Grading of Recommendations, Assessment, Development and Evaluations) framework.

Results: From 12,579 records, we included 13 studies, comprising 9983 participants. All studies were rated as either low or unclear for risk of bias. The overall evidence quality was moderate. AI trained on PPG achieved a pooled sensitivity of 79.6% (95% credible interval [CrI] 55.5%-93.8%) and specificity of 76.5% (95% CrI 48.2%-94.0%), compared to conventional diagnosis. The overall specificity increased (apnea-hypopnea index [AHI] ≥ 5 : 63.6%; AHI ≥ 15 : 81.8%; AHI ≥ 30 : 85.1%), but overall sensitivity decreased (AHI ≥ 5 : 87.2%; AHI ≥ 15 : 79.7%; AHI ≥ 30 : 76.7%) with greater categorical AHI severity cutoffs. Additionally, deep learning models achieved a higher specificity (82.9%) than traditional machine learning (63.6%). OSA prevalence and device type were not clearly associated with sensitivity or specificity.

Conclusions: AI models trained on PPG have reasonable accuracy and may potentially serve as a low-cost screening tool. However, limitations such as small sample sizes, potential sources of bias, underrepresentation of certain geographic regions, and the influence of potential confounding factors highlight the necessity for further research. Future work should focus on deep learning to improve the feasibility and accessibility of this approach in primary care.

Trial Registration: PROSPERO CRD42024534235; <https://www.crd.york.ac.uk/PROSPERO/view/CRD42024534235>

J Med Internet Res 2026;28:e78718; doi: [10.2196/78718](https://doi.org/10.2196/78718)

Keywords: obstructive sleep apnea; artificial intelligence; AI; photoplethysmography; polysomnography; Bayesian statistics; diagnostic test accuracy

Introduction

Obstructive sleep apnea (OSA) is a debilitating sleep-breathing disorder defined by recurring episodes of partial or complete upper airway collapse during sleep. This results in reduced or absent airflow lasting for at least 10 seconds and is associated with either cortical arousal or a decline in blood oxygen saturation [1]. It is estimated to affect 1 billion individuals globally, highlighting the need for effective and cost-efficient diagnostic and treatment approaches to reduce its adverse health sequelae [2]. OSA may be diagnosed with either home-based or laboratory-based sleep testing, with overnight polysomnography (PSG) touted as its gold standard diagnostic tool [3,4]. While it provides reliable diagnostic accuracy, it is constrained by high operating costs, limited accessibility in outpatient settings, patient inconvenience, and significant logistical and labor-intensive demands [5].

AI has emerged as a promising tool in medical diagnostics, leveraging machine learning algorithms to analyze complex clinical data and support diagnostic decision-making [6,7]. AI models are increasingly being assessed as automated alternatives to traditional PSG for OSA diagnosis. While AI-based approaches for OSA diagnosis have been explored using various data sources [8–11], the diagnostic performance of AI models trained on photoplethysmography (PPG) has not been systematically evaluated. PPG is a noninvasive, cost-effective, and portable optical technique used to provide real-time measurements of blood oxygen saturation in the microvascular tissue bed [12]. Presently, preexisting studies differ in opinion on the accuracy of AI-based diagnosis of OSA using PPG [13,14].

Given the growing presence of AI in medical diagnostics amid a digital revolution, and the expanding body of evidence, it is both clinically pertinent and timely to conduct a comprehensive synthesis of the evidence regarding the utility of AI models for OSA diagnosis, specifically those trained using PPG data [15]. The aim of this meta-analysis is to assess the diagnostic accuracy of AI models in detecting OSA using PPG.

Methods

This systematic review and meta-analysis was registered on PROSPERO (CRD42024534235) and adhered to PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. The PRISMA checklist can be found in [Checklist 1](#) [16].

Search Strategy

We conducted a comprehensive search of PubMed, Embase, Scopus, Web of Science, and IEEE Xplore from inception to

November 3, 2025, using search terms pertaining to OSA, AI, PPG, and their relevant synonyms. The full search strategy is provided in Supplementary Methods in [Multimedia Appendix 1](#).

Study Selection

The study selection process by titles and abstracts, followed by full-text articles, was conducted by 6 independent authors (BSYY, EYG, JEXT, JYK, NKWT, BKJT), with disagreements resolved through the consensus of another independent author (STT). We selected observational studies involving adults aged 18 years or older, focused on the diagnosis or classification of OSA using AI, including both traditional and deep learning models, using PPG. The comparators included the diagnosis and classification of OSA based on the apnea-hypopnea index (AHI) derived from established methods of testing, including PSG or home sleep apnea testing (HSAT). The outcome of interest was the diagnostic accuracy of AI-based OSA diagnosis, evaluated using a random split test, cross-validation, or external validation. Table S1 in [Multimedia Appendix 1](#) summarizes the inclusion criteria of the study. We excluded case reports, case series, reviews, meta-analyses, letters, conference abstracts, pediatric studies, animal studies, foreign language studies, studies with incomplete data, and studies focusing on individual apneic events without patient-level classification. Table S2 in [Multimedia Appendix 1](#) summarizes the exclusion criteria.

Data Extraction

The data of interest from the included articles were extracted by 3 independent authors (JEXT, JYK, YSL) into a structured pro forma. The details of data extraction were verified by independent authors (BSYY, BKJT). This included information pertaining to the first author, year published, country, percentage of males, mean age, mean body mass index, sample size, type of sleep study, reference AHI category thresholds, categorical AHI cutoffs, OSA prevalence, test method, AI model type, and type of sleep device used. Information regarding diagnostic accuracy, including measures of accuracy, sensitivity, specificity, and prevalence, was also extracted for analysis.

Statistical Analysis

We conducted a Bayesian bivariate random-effects meta-analysis to jointly estimate sensitivity and specificity on the logit scale, accounting for their correlation [17–20]. To incorporate multiple AI models and AHI thresholds reported within the same study, we specified a hierarchical model with 2 levels, where study-level random effects captured between-study heterogeneity in diagnostic performance, while model-level random effects (nested within studies) captured variation across models and thresholds within the same study. Random effects at both levels were assumed to follow bivariate

normal distributions, allowing sensitivity and specificity to be correlated. This approach accounted for the trade-off between sensitivity and specificity across and within studies. Pooled sensitivity and specificity were presented through summary receiver operating characteristic (SROC) curves. We performed meta-regression analyses by including study- and model-level covariates as fixed effects on the logit scale.

The following factors were examined: (1) OSA severity stratified by categorical AHI thresholds, with AHI ≥ 5 , ≥ 15 , and ≥ 30 corresponding to mild, moderate, and severe disease, respectively; (2) AI model type, distinguishing deep learning from traditional machine learning approaches; (3) type of primary validation method, including external validation using independent cohorts, cross-validation, or random data splitting; (4) OSA prevalence; and (5) type of sleep device, differentiating wearable smart devices from clinical oximeters. All results are reported as posterior means with 95% credible intervals (CrIs).

To evaluate for potential publication bias, Deeks test for funnel plot asymmetry was performed under varying probabilities for unpublished studies [21]. Further sensitivity analyses were conducted by recalculating diagnostic accuracy outcomes under each scenario using a Bayesian hierarchical framework, where an informative prior was applied with a lower bound set at 50% for both sensitivity and specificity.

Risk of Bias Assessment

The QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies-2) tool was employed to evaluate the risk of bias and the applicability of the included diagnostic accuracy

studies [22]. This tool allows for a structured assessment of potential biases and relevance of the studies using 4 domains: patient selection, index test, reference standard, and flow and timing. Three independent authors (JEXT, JYK, YSL) assessed each study, categorizing the risk of bias and applicability concerns as low, high, or unclear. Disagreements from the risk of bias assessment were resolved by independent authors (BSYY, BKJT).

Overall Quality of Evidence

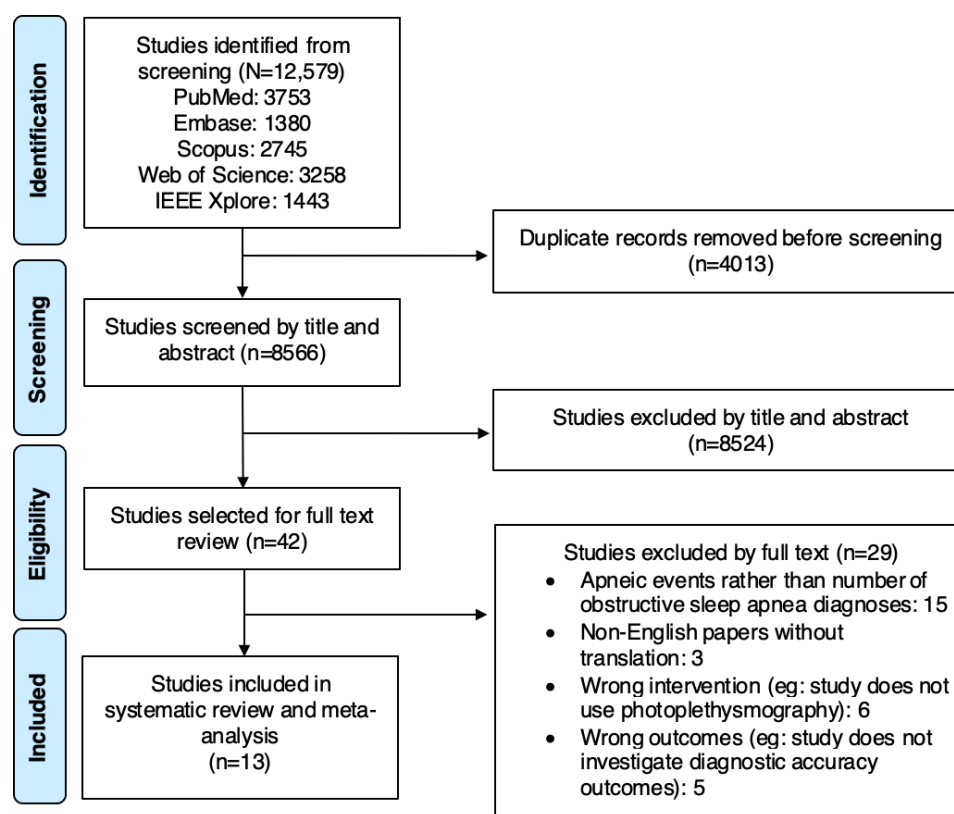
The overall quality of pooled evidence at the outcome level was assessed using the GRADE (Grading of Recommendations, Assessment, Development and Evaluations) framework [23,24]. This framework evaluates the quality of evidence by considering factors such as study design, consistency of results, directness of evidence, risk of bias, precision of estimates, and potential publication bias. The level of evidence for each outcome is categorized as high, moderate, low, or very low.

Results

Study Selection

A PRISMA flowchart documenting the study selection process is outlined in [Figure 1](#). From a comprehensive search of 12,579 articles from PubMed, Embase, Scopus, Web of Science, and IEEE Xplore, 8566 nonduplicated records were screened at the title and abstract stage. Subsequently, 42 studies were screened for full-text eligibility, and 13 studies were included and qualitatively analyzed, comprising 9983 participants [13,14,25-35].

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart for identification of studies via databases and registers.



Study Characteristics

Baseline study characteristics of included articles are found in Table S3 in [Multimedia Appendix 1](#). All studies were cross-sectional in nature, and all studies used data from patients recruited from outpatient sleep clinics. The mean age of patients ranged from 42.0 to 69.4 years, and the percentage of male participants ranged from 46.4% to 74.0%. There were 5 studies from China, 2 from the United States, and 1 each from the United Kingdom, the Netherlands, Israel, Taiwan, South Korea, and Thailand. Using the QUADAS-2 scale, the risk of bias was listed as low for 7 studies and unclear for 6 studies (Table S4 in [Multimedia Appendix 1](#)).

Diagnosis of OSA

All studies used PSG as the reference standard for diagnosing OSA, except for 1 study that used HSAT. The AHI served as the primary criterion for classification. Different studies used varying AHI thresholds, with ≥ 5 , ≥ 15 , and ≥ 30 events per hour representing mild, moderate, and severe OSA, respectively.

Evaluation Methods and AI Models

Based on the primary validation strategy, 4 studies used random split, 3 studies adopted cross-validation, 5 studies conducted external validation, and 1 study used cross-validation and external validation for separate data cohorts. There were 7 studies that used deep learning and 6 studies that used traditional machine learning. Table S3 in [Multimedia Appendix 1](#) outlines specific model subtypes.

Overall Diagnostic Accuracy

PPG-based AI models demonstrated a pooled sensitivity of 79.6% (95% CrI 55.5%-93.8%) and a specificity of 76.5% (95% CrI 48.2%-94.0%; [Figure 2](#)), compared to traditional diagnostic methods. These estimates indicated moderate diagnostic performance, although the wide CrI suggests substantial between-study variability. The SROC curve is presented in [Figure 3](#). Model diagnostics indicated good convergence (all $R^2 \leq 1.01$) with adequate effective sample sizes.

In a clinical diagnostic context, the application of Bayes theorem allows for the interpretation of findings related to OSA prevalence. When considering a pretest probability, or baseline population prevalence of OSA, of approximately 15% (Iceland, Indonesia, and the United Arab Emirates), the results indicate that a positive diagnostic test result would yield an average posttest probability of 37.4% (95% CrI 15.9%-73.4%). For a population prevalence of 30% (United States, United Kingdom, and Samoa), the average posttest probability would be 59.2% (95% CrI 31.5%-87.0%). For regions with a population prevalence of 60% (Singapore, Switzerland, and France), a positive test would yield a posttest probability of 83.6% (95% CrI 61.6%-95.9%). Conversely, a negative test result would correspond to an average posttest probability of 4.5% (95% CrI 1.2%-14.0%), 10.3% (95% CrI 2.8%-28.3%), or 28.6% (95% CrI 9.0%-58.1%) for the same prevalence categories. Furthermore, in instances where a patient's pretest probability of OSA exceeds the regional prevalence, such as when the patient

exhibits symptoms or signs indicative of OSA, the posttest probability is likely to be even higher.

Figure 2. Bayesian bivariate meta-analysis of photoplethysmography-based obstructive sleep apnea detection. Results for (A) sensitivity and (B) specificity [13,14,25-35]. CrI: credible interval.

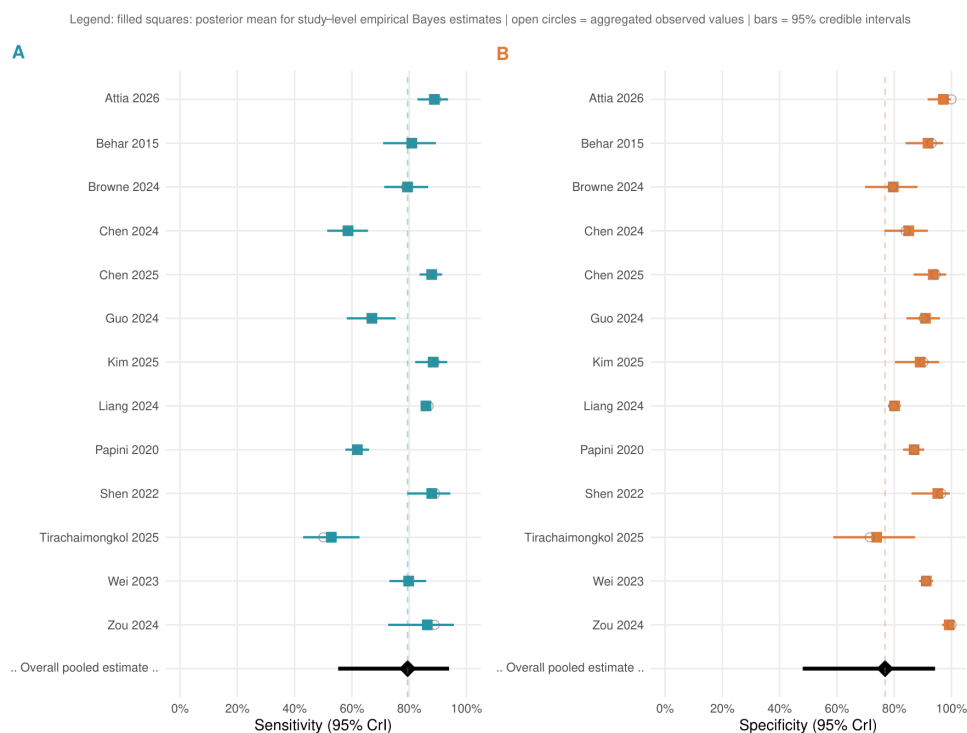
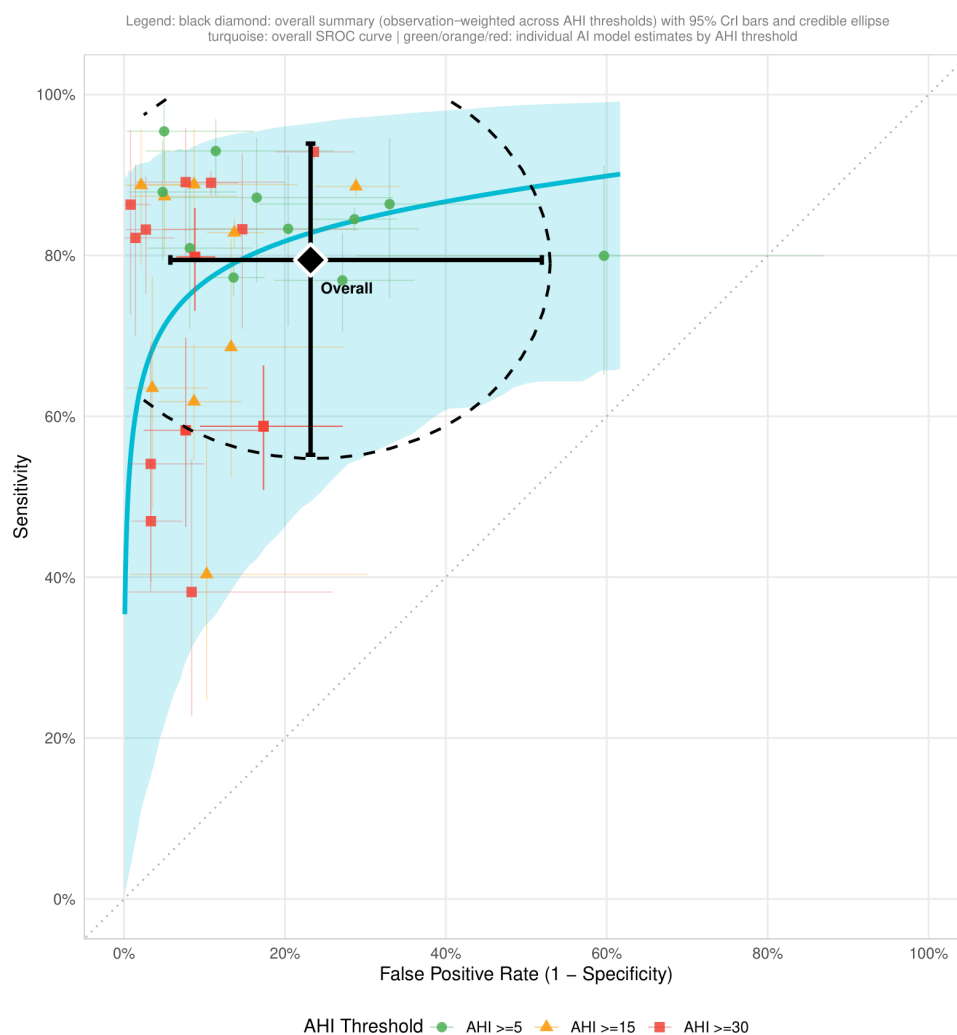


Figure 3. Summary receiver operating characteristic (SROC) curve for Bayesian 2-level bivariate model. AHI: apnea-hypopnea index; CrI: credible interval.



Meta-Regression

Meta-regression analysis was performed for OSA severity based on AHI thresholds, AI model type, type of validation method, OSA prevalence, and type of sleep device.

The specificities of detecting OSA on AI-trained PPG increased with categorical OSA severity. Pooled specificities at AHI thresholds of ≥ 5 , ≥ 15 , and ≥ 30 were 63.6%, 81.8%, and 85.1%, respectively. This trend was supported by positive logit-scale coefficients (≥ 15 vs ≥ 5 : 0.94, 95% CrI 0.02-1.83; ≥ 30 vs ≥ 5 : 1.18, 95% CrI 0.15-2.08). However, sensitivity appeared to decrease modestly with increasing categorical OSA severity. Pooled sensitivities at AHI thresholds of ≥ 5 , ≥ 15 , and ≥ 30 were 87.2%, 79.7%, and 76.7%, respectively, although the corresponding coefficients were imprecise and compatible with little or no difference (≥ 15 vs ≥ 5 : -0.55, 95% CrI -1.24 to 0.11; ≥ 30 vs ≥ 5 : -0.56, 95% CrI -1.22 to 0.12; Figure S1A in [Multimedia Appendix 1](#)).

Further meta-regression of AI model type demonstrated that deep learning models, such as convolutional neural networks, were associated with higher specificity of 82.9%, compared with other traditional machine learning models such as decision trees, hyperdimensional models, linear regression, logistic regression, local binary patterns, support

vector machines, and combinations of these methods, which achieved a pooled specificity of 63.6%. This was supported by a positive coefficient (1.02, 95% CrI 0.02-2.07). There was no clear difference in sensitivity (-0.02, 95% CrI -0.56 to 0.52; Figure S1B in [Multimedia Appendix 1](#)).

Additionally, studies that used external validation were associated with better sensitivity of 87.2% compared to studies using cross-validation (80.1%; -0.53, 95% CrI -0.80 to -0.27), whereas the comparison with studies adopting random split approaches was less certain (78.6%; -0.62, 95% CrI -1.58 to 0.41). For specificity, cross-validation was associated with higher estimates compared to external validation (0.87, 95% CrI 0.44-1.30), whereas its comparison with the estimate for random split approaches was more uncertain (0.66, 95% CrI -0.63 to 1.91; Figure S1C in [Multimedia Appendix 1](#)).

The type of device used (wearable smart devices vs clinical oximeters; Figure S1D in [Multimedia Appendix 1](#)) and the prevalence of OSA (Figure S2 in [Multimedia Appendix 1](#)) were not clearly associated with sensitivity or specificity, with all corresponding 95% CrIs including the null value.

Publication Bias

While visual inspection of funnel plots suggested possible asymmetry (Figure S3 in [Multimedia Appendix 1](#)), Deeks test did not suggest publication bias (intercept=−1.53; $P=.14$). Further sensitivity analyses on the SROC curve outlined no clinically significant publication bias. When evaluating other distinct mechanisms of publication bias, each with varying probabilities of unpublished studies of up to 60%, the SROC curve (Figure S4 in [Multimedia Appendix 1](#)) remained nearly unchanged. This suggests that even if a majority of studies had remained unpublished, the overall conclusions of the meta-analysis would not have been affected.

Quality of Evidence

The overall quality of evidence per outcome assessed using GRADE was moderate (Table S5 in [Multimedia Appendix 1](#)).

Discussion

Principal Findings

This meta-analysis found that AI models trained on PPG recordings demonstrated favorable accuracy in OSA diagnosis, achieving a good sensitivity of 79.6% and specificity of 76.5%, compared to traditional diagnostic methods. To the best of the authors' knowledge, this is the most comprehensive pooled analysis to date on the diagnostic accuracy of AI algorithms using PPG for OSA diagnosis.

Comparison to Prior Work

This study builds on preexisting literature surrounding the accuracy of using PPG-based AI models in OSA diagnosis. A prior descriptive systematic review by Osa-Sanchez et al [36] found that integrating wearable devices with AI algorithms shows strong promise for real-time, portable, and early detection of OSA. However, a pooled meta-analysis was not performed in their prior study. Separately, Abd-Alrazaq et al [37] conducted a systematic review and meta-analysis of 38 studies to evaluate the diagnostic accuracy of wearable AI in detecting apneic events in respiration and OSA. They found that AI-powered wearable devices achieved a mean accuracy of 86.9%, with good sensitivity of 93.8% but lower specificity of 75.2% in detecting OSA. The current study provides a more detailed evaluation of existing evidence by incorporating additional sensitivity, subgroup, and publication bias analyses. A Bayesian hierarchical framework was employed to evaluate diagnostic accuracy outcomes holistically under various scenarios.

Unlike traditional approaches or simple machine learning models, deep learning relies on multilayered neural networks to capture more complex patterns within data [38]. Prior research has demonstrated that deep learning significantly improves the detection of subtle physiological changes associated with OSA by enabling the automatic extraction of intricate, high-dimensional features directly from PPG recordings [39]. This ability to autonomously identify relevant patterns in large datasets allows deep

learning models to capture complex relationships that may not be immediately apparent through conventional methods.

Moreover, the automation offered by deep learning models presents significant logistical advantages over traditional diagnostic methods by reducing the need for manual intervention and expert oversight [40]. These AI-driven approaches can streamline diagnostic processes, improving efficiency and speed. Crucially, this provides an opportunity for reliable OSA diagnosis in low-resource or cost-constrained settings, where traditional methods may be prohibitively expensive or logistically challenging. WatchPAT (ZOLL Itamar) is an established HSAT modality for OSA diagnosis, particularly among patients with a high pretest probability of disease [41]. Given the encouraging diagnostic performance observed for PPG-based AI models in this review, these approaches warrant future direct comparison with established HSAT platforms to evaluate whether they offer comparable or incremental diagnostic value in clinical practice.

Interpretation of Findings

Through meta-regression analysis, it was observed that the specificity of AI models trained on PPG data for detecting OSA improves as the severity of OSA increases. With increasing severity of OSA, oxygen desaturations become more frequent, deeper, and longer lasting during apneic episodes [42]. These alterations in blood oxygen saturation within the microvascular tissue bed are more pronounced and may be more easily detected by PPG technology, thereby improving between-group separability, which contributes to higher specificity. Consequently, AI-based systems have the potential to be more effective in diagnosing severe cases of OSA correctly, where the physiological markers are clearer and more detectable. Conversely, this may also imply that for AI models to achieve accurate detection across the entire spectrum of OSA severity, further refinement is needed, particularly to improve their ability to identify mild cases.

However, the sensitivity of OSA detection appears to decrease with greater categorical OSA severity. AI-assisted PPG models may produce a higher rate of false negatives in more severe cases. They may be more accurate at distinguishing individuals with and without OSA than at distinguishing the severity of OSA. Severe OSA is inherently heterogeneous, with variable phenotypes, variable event morphology, and substantial interindividual variability in desaturation dynamics, autonomic response, vascular tone, and comorbid cardiovascular conditions, which may influence PPG detection of OSA [43,44]. The divergent trends in specificity and sensitivity across severity strata reflect the complex dynamics of harnessing AI to diagnose OSA using PPG. Future research may help clarify the relationship between OSA severity and diagnostic accuracy.

Additionally, meta-regression analysis showed that deep learning achieved superior specificity compared to traditional machine learning models. This finding suggests that the ability of deep learning to autonomously identify complex patterns from PPG signals enhances its detection of subtle physiological alterations associated with OSA [45].

Separately, traditional machine learning may depend on manually engineered features, which may limit its sensitivity. While no evidence of a difference in sensitivity was observed, higher specificity was observed for deep learning models, which highlights the potential utility of deep learning models in OSA diagnosis.

Validation strategy may influence reported diagnostic performance, with external validation associated with greater sensitivity and cross-validation with greater estimates for specificity, although comparisons involving random split approaches remained uncertain. Nonetheless, the limited number of studies using each method of validation may preclude conclusions regarding an optimal validation strategy. Future studies using consistent validation frameworks across comparable datasets may help clarify the impact of validation strategy on reported diagnostic accuracy.

Limitations

This study has several limitations. First, the underrepresentation of certain geographic groups in existing datasets may limit the generalizability of the findings. Second, the relatively small number of included studies, which typically feature small sample sizes, may introduce biases, as outliers from these small sample sizes could disproportionately affect the pooled results. Potential bias in preexisting studies may be driven by the use of convenience samples from sleep clinics, which may limit the applicability of existing studies. Third, publication bias remains a potential issue, as there is a possibility of missing relevant studies. Fourth, there was insufficient evidence to examine how BMI may influence the diagnostic accuracy outcome. High BMI is a recognized risk factor for OSA and is associated with greater disease severity [46]. The accuracy of diagnostic methods may be influenced by a patient's BMI, and a thorough understanding of the relationship between BMI and PPG-based AI models may help guide clinical decision-making. Fifth, while all studies used validation set testing, some relied on internal test sets; external test sets are preferred for evaluating AI algorithms, as they better assess performance when applied to data from different sources [47]. Future research should

focus on deep learning and incorporate prospective studies to enhance the feasibility and accessibility of this approach. Additionally, further studies are necessary to interpret these findings in a broader context and to optimize the diagnostic performance of AI-driven models in OSA diagnosis. Sixth, while the original PROSPERO registration outlined a broader review of AI modalities for OSA diagnosis, this manuscript focuses specifically on PPG due to the volume and diversity of available studies. Nonetheless, no alterations were made to the analytical methods, which remained consistent with the protocol. Seventh, heterogeneity was observed in pooled analyses, which was potentially driven by threshold effects, variation in disease severity, differences in study design, and evaluation frameworks. While additional meta-regression analyses were performed, results should be interpreted in the context of potential variability.

Future Directions

It is essential for future research to carefully evaluate the trade-offs involved in incorporating additional signal channels into AI-driven PPG systems, particularly in the context of the costs and practicalities associated with diagnosing OSA. Although this study did not include a formal cost-benefit analysis, it highlights several critical factors for consideration: balancing diagnostic accuracy against signal complexity, assessing both hardware and operational costs, promoting user compliance and comfort, and addressing the interpretability and generalizability of AI models. Future research should also concentrate on deep learning techniques to enhance the feasibility and accessibility of this approach, particularly within primary care settings. The inclusion of prospective, real-world data may also enhance the credibility of AI-assisted OSA diagnosis.

Conclusions

This meta-analysis suggests that AI models trained on PPG data may demonstrate reasonable diagnostic accuracy. Considering the logistical and economic challenges associated with current sleep testing, AI-driven models could serve as a cost-effective screening tool, pending further validation.

Acknowledgments

Generative AI was not used in this study.

Funding

BSYY gratefully acknowledges the SingHealth Medical Student Talent Development Award (Surgery Academic Clinical Program, SingHealth).

Data Availability

The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Authors' Contributions

BKJT is co-senior author.

Conceptualization: BKJT (lead), STT (equal), BSYY (supporting), EYG (supporting)

Data curation: BSYY (lead), EYG (supporting), JEXT (supporting), JYK (supporting), YSL (supporting), NKWT (supporting), BKJT (supporting)

Formal analysis: BKJT (lead), BSYY (supporting)

Funding acquisition: BSYY

Investigation: BSY (lead), EYG (supporting), JEXT (supporting), JYK (supporting), YSL (supporting), NKWT (supporting), BKJT (supporting)
Methodology: BKJT (lead), STT (equal), BSY (equal)
Project administration: BSY (lead), BKJT (equal), STT (equal)
Resources: BSY (lead), BKJT (equal), STT (equal)
Supervision: STT (lead), BKJT (equal), ACWN (supporting), ZHL (supporting), THO (supporting), LCL (supporting), GBH (supporting)
Validation: BSY (lead), BKJT (equal), STT (equal)
Visualization: BSY
Writing – original draft: BSY (lead), EYG (supporting), JEXT (supporting), JYK (supporting), YSL (supporting), NKWT (supporting)
Writing – review & editing: BSY (lead), EYG (supporting), ACWN (supporting), ZHL (supporting), THO (supporting), LCL (supporting), GBH (supporting), BKJT (supporting), STT (supporting)

Conflicts of Interest

GBH is involved in Mind PointEye, a company specializing in AI solutions, but his affiliation does not affect the current work. All other authors have no competing interests to declare that are relevant to the content of this article.

Multimedia Appendix 1

Supplementary Materials.

[\[DOCX File \(Microsoft Word File\), 4214 KB-Multimedia Appendix 1\]](#)

Checklist 1

PRISMA checklist.

[\[DOCX File \(Microsoft Word File\), 2636 KB-Checklist 1\]](#)

References

1. Gottlieb DJ, Punjabi NM. Diagnosis and management of obstructive sleep apnea: a review. *JAMA*. Apr 14, 2020;323(14):1389-1400. [doi: [10.1001/jama.2020.3514](https://doi.org/10.1001/jama.2020.3514)] [Medline: [32286648](https://pubmed.ncbi.nlm.nih.gov/32286648/)]
2. Benjafield AV, Ayas NT, Eastwood PR, et al. Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. *Lancet Respir Med*. Aug 2019;7(8):687-698. [doi: [10.1016/S2213-2600\(19\)30198-5](https://doi.org/10.1016/S2213-2600(19)30198-5)] [Medline: [31300334](https://pubmed.ncbi.nlm.nih.gov/31300334/)]
3. Kapur V, Strohl KP, Redline S, Iber C, O'Connor G, Nieto J. Underdiagnosis of sleep apnea syndrome in U.S. communities. *Sleep Breath*. Jun 2002;6(2):49-54. [doi: [10.1007/s11325-002-0049-5](https://doi.org/10.1007/s11325-002-0049-5)] [Medline: [12075479](https://pubmed.ncbi.nlm.nih.gov/12075479/)]
4. Marino M, Li Y, Rueschman MN, et al. Measuring sleep: accuracy, sensitivity, and specificity of wrist actigraphy compared to polysomnography. *Sleep*. Nov 1, 2013;36(11):1747-1755. [doi: [10.5665/sleep.3142](https://doi.org/10.5665/sleep.3142)] [Medline: [24179309](https://pubmed.ncbi.nlm.nih.gov/24179309/)]
5. Markun LC, Sampat A. Clinician-focused overview and developments in polysomnography. *Curr Sleep Med Rep*. 2020;6(4):309-321. [doi: [10.1007/s40675-020-00197-5](https://doi.org/10.1007/s40675-020-00197-5)] [Medline: [33251088](https://pubmed.ncbi.nlm.nih.gov/33251088/)]
6. Bazoukis G, Bollepalli SC, Chung CT, et al. Application of artificial intelligence in the diagnosis of sleep apnea. *J Clin Sleep Med*. Jul 1, 2023;19(7):1337-1363. [doi: [10.5664/jcsm.10532](https://doi.org/10.5664/jcsm.10532)] [Medline: [36856067](https://pubmed.ncbi.nlm.nih.gov/36856067/)]
7. 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](https://doi.org/10.4103/singaporemedj.SMJ-2022-170)] [Medline: [37171440](https://pubmed.ncbi.nlm.nih.gov/37171440/)]
8. 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](https://doi.org/10.1016/j.chest.2025.04.006)] [Medline: [40220991](https://pubmed.ncbi.nlm.nih.gov/40220991/)]
9. 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](https://doi.org/10.1007/s44470-025-00013-3)] [Medline: [41678061](https://pubmed.ncbi.nlm.nih.gov/41678061/)]
10. 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](https://doi.org/10.1007/s11325-024-03173-3)] [Medline: [39614959](https://pubmed.ncbi.nlm.nih.gov/39614959/)]
11. 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](https://doi.org/10.1007/s11325-025-03476-z)] [Medline: [41026252](https://pubmed.ncbi.nlm.nih.gov/41026252/)]
12. Allen J. Photoplethysmography and its application in clinical physiological measurement. *Physiol Meas*. Mar 2007;28(3):R1-R39. [doi: [10.1088/0967-3334/28/3/R01](https://doi.org/10.1088/0967-3334/28/3/R01)] [Medline: [17322588](https://pubmed.ncbi.nlm.nih.gov/17322588/)]
13. Wei K, Zou L, Liu G, Wang C. MS-Net: Sleep apnea detection in PPG using multi-scale block and shadow module one-dimensional convolutional neural network. *Comput Biol Med*. Mar 2023;155:106469. [doi: [10.1016/j.compbiomed.2022.106469](https://doi.org/10.1016/j.compbiomed.2022.106469)] [Medline: [36842220](https://pubmed.ncbi.nlm.nih.gov/36842220/)]

14. Papini GB, Fonseca P, van Gilst MM, Bergmans JWM, Vullings R, Overeem S. Wearable monitoring of sleep-disordered breathing: estimation of the apnea-hypopnea index using wrist-worn reflective photoplethysmography. *Sci Rep*. Aug 11, 2020;10(1):13512. [doi: [10.1038/s41598-020-69935-7](https://doi.org/10.1038/s41598-020-69935-7)] [Medline: [32782313](https://pubmed.ncbi.nlm.nih.gov/32782313/)]
15. Haug CJ, Drazen JM. Artificial intelligence and machine learning in clinical medicine, 2023. *N Engl J Med*. Mar 30, 2023;388(13):1201-1208. [doi: [10.1056/NEJMr2302038](https://doi.org/10.1056/NEJMr2302038)] [Medline: [36988595](https://pubmed.ncbi.nlm.nih.gov/36988595/)]
16. 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](https://doi.org/10.1136/bmj.n71)] [Medline: [33782057](https://pubmed.ncbi.nlm.nih.gov/33782057/)]
17. 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](https://doi.org/10.1186/s12874-023-01910-y)] [Medline: [37231347](https://pubmed.ncbi.nlm.nih.gov/37231347/)]
18. Mizutani S, Zhou Y, Tian YS, Takagi T, Ohkubo T, Hattori S. DTAmetsa: 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](https://doi.org/10.1002/jrsm.1666)] [Medline: [37640914](https://pubmed.ncbi.nlm.nih.gov/37640914/)]
19. 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](https://doi.org/10.1186/s12874-019-0724-x)] [Medline: [30999861](https://pubmed.ncbi.nlm.nih.gov/30999861/)]
20. 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](https://doi.org/10.1002/jrsm.1439)] [Medline: [32706182](https://pubmed.ncbi.nlm.nih.gov/32706182/)]
21. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *J Clin Epidemiol*. Sep 2005;58(9):882-893. [doi: [10.1016/j.jclinepi.2005.01.016](https://doi.org/10.1016/j.jclinepi.2005.01.016)] [Medline: [16085191](https://pubmed.ncbi.nlm.nih.gov/16085191/)]
22. 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](https://doi.org/10.7326/0003-4819-155-8-201110180-00009)] [Medline: [22007046](https://pubmed.ncbi.nlm.nih.gov/22007046/)]
23. 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](https://doi.org/10.1136/bmj.39489.470347.AD)] [Medline: [18436948](https://pubmed.ncbi.nlm.nih.gov/18436948/)]
24. Schünemann HJ, Mustafa RA, Brozek J, et al. GRADE guidelines: 21 part 2. test accuracy: inconsistency, imprecision, publication bias, and other domains for rating the certainty of evidence and presenting it in evidence profiles and summary of findings tables. *J Clin Epidemiol*. Jun 2020;122:142-152. [doi: [10.1016/j.jclinepi.2019.12.021](https://doi.org/10.1016/j.jclinepi.2019.12.021)] [Medline: [32058069](https://pubmed.ncbi.nlm.nih.gov/32058069/)]
25. Attia S, Oksenberg A, Levy J, et al. Clinical validation of artificial intelligence algorithms for the diagnosis of adult obstructive sleep apnea and sleep staging from oximetry and photoplethysmography-SleepAI. *J Sleep Res*. Feb 2026;35(1):e70093. [doi: [10.1111/jsr.70093](https://doi.org/10.1111/jsr.70093)] [Medline: [40346945](https://pubmed.ncbi.nlm.nih.gov/40346945/)]
26. Behar J, Roebuck A, Shahid M, et al. SleepAp: an automated obstructive sleep apnoea screening application for smartphones. *IEEE J Biomed Health Inform*. Jan 2015;19(1):325-331. [doi: [10.1109/JBHI.2014.2307913](https://doi.org/10.1109/JBHI.2014.2307913)] [Medline: [25561453](https://pubmed.ncbi.nlm.nih.gov/25561453/)]
27. Browne SH, Vaida F, Umlauf A, Kim J, DeYoung P, Owens RL. Performance of a commercial smart watch compared to polysomnography reference for overnight continuous oximetry measurement and sleep apnea evaluation. *J Clin Sleep Med*. Sep 1, 2024;20(9):1479-1488. [doi: [10.5664/jcsm.11178](https://doi.org/10.5664/jcsm.11178)] [Medline: [38652502](https://pubmed.ncbi.nlm.nih.gov/38652502/)]
28. Chen T, Zhang J, Xu Z, et al. Energy-efficient sleep apnea detection using a hyperdimensional computing framework based on wearable bracelet photoplethysmography. *IEEE Trans Biomed Eng*. Aug 2024;71(8):2483-2494. [doi: [10.1109/TBME.2024.3377270](https://doi.org/10.1109/TBME.2024.3377270)] [Medline: [38483799](https://pubmed.ncbi.nlm.nih.gov/38483799/)]
29. Chen KW, Tseng CH, Lee HC, Liu WT, Chou KT, Wu HT. Validation of a fingertip home sleep apnea testing system using deep learning AI and a temporal event localization analysis. *Sleep*. May 12, 2025;48(5):zsae317. [doi: [10.1093/sleep/zsae317](https://doi.org/10.1093/sleep/zsae317)] [Medline: [39821673](https://pubmed.ncbi.nlm.nih.gov/39821673/)]
30. Guo H, Wu H, Xia J, et al. OSAHS detection capabilities of ringconn smart ring: a feasibility study. Presented at: 2024 IEEE 6th International Conference on AI Circuits and Systems (AICAS); Apr 22-25, 2024:597-601; Abu Dhabi, United Arab Emirates. URL: <https://ieeexplore.ieee.org/document/10595903> [Accessed 2026-08-20] [doi: [10.1109/AICAS59952.2024.10595903](https://doi.org/10.1109/AICAS59952.2024.10595903)]
31. Kim D, Han JY, Jung H, et al. AI-enhanced smartwatch AHI estimation and AI-scored polysomnography for obstructive sleep apnea: real-world validation. *Nat Sci Sleep*. Sep 22, 2025;17:2297-2307. [doi: [10.2147/NSS.S540460](https://doi.org/10.2147/NSS.S540460)] [Medline: [41018852](https://pubmed.ncbi.nlm.nih.gov/41018852/)]
32. Liang Z. Developing probabilistic ensemble machine learning models for home-based sleep apnea screening using overnight SpO2 data at varying data granularity. *Sleep Breath*. Dec 2024;28(6):2409-2420. [doi: [10.1007/s11325-024-03141-x](https://doi.org/10.1007/s11325-024-03141-x)] [Medline: [39190088](https://pubmed.ncbi.nlm.nih.gov/39190088/)]

33. Shen Q, Yang X, Zou L, Wei K, Wang C, Liu G. Multitask residual shrinkage convolutional neural network for sleep apnea detection based on wearable bracelet photoplethysmography. *IEEE Internet Things J.* Dec 15, 2022;9(24):25207-25222. [doi: [10.1109/JIOT.2022.3195777](https://doi.org/10.1109/JIOT.2022.3195777)]
34. Tirachaimongkol V, Banhiran W, Chotinaiwattarakul W, et al. Belun Sleep Platform versus in-lab polysomnography for obstructive sleep apnea diagnosis. *Sleep Breath.* Aug 7, 2025;29(4):266. [doi: [10.1007/s11325-025-03433-w](https://doi.org/10.1007/s11325-025-03433-w)] [Medline: [40773145](https://pubmed.ncbi.nlm.nih.gov/40773145/)]
35. Zou L, Liu G. Multiscale bidirectional temporal convolutional network for sleep apnea detection based on wearable photoplethysmography bracelet. *IEEE J Biomed Health Inform.* Mar 2024;28(3):1331-1340. [doi: [10.1109/JBHI.2023.3335658](https://doi.org/10.1109/JBHI.2023.3335658)] [Medline: [37991905](https://pubmed.ncbi.nlm.nih.gov/37991905/)]
36. Osa-Sanchez A, Ramos-Martinez-de-Soria J, Mendez-Zorrilla A, Ruiz IO, Garcia-Zapirain B. Wearable sensors and artificial intelligence for sleep apnea detection: a systematic review. *J Med Syst.* May 19, 2025;49(1):66. [doi: [10.1007/s10916-025-02199-8](https://doi.org/10.1007/s10916-025-02199-8)] [Medline: [40387964](https://pubmed.ncbi.nlm.nih.gov/40387964/)]
37. Abd-Alrazaq A, Aslam H, AlSaad R, et al. Detection of sleep apnea using wearable AI: systematic review and meta-analysis. *J Med Internet Res.* Sep 10, 2024;26:e58187. [doi: [10.2196/58187](https://doi.org/10.2196/58187)] [Medline: [39255014](https://pubmed.ncbi.nlm.nih.gov/39255014/)]
38. Sarker IH. Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. *SN Comput Sci.* 2021;2(6):420. [doi: [10.1007/s42979-021-00815-1](https://doi.org/10.1007/s42979-021-00815-1)] [Medline: [34426802](https://pubmed.ncbi.nlm.nih.gov/34426802/)]
39. Park MJ, Choi JH, Kim SY, Ha TK. A deep learning algorithm model to automatically score and grade obstructive sleep apnea in adult polysomnography. *Digit Health.* Oct 17, 2024;10:20552076241291707. [doi: [10.1177/20552076241291707](https://doi.org/10.1177/20552076241291707)] [Medline: [39430691](https://pubmed.ncbi.nlm.nih.gov/39430691/)]
40. Krishnan G, Singh S, Pathania M, et al. Artificial intelligence in clinical medicine: catalyzing a sustainable global healthcare paradigm. *Front Artif Intell.* Aug 29, 2023;6:1227091. [doi: [10.3389/frai.2023.1227091](https://doi.org/10.3389/frai.2023.1227091)] [Medline: [37705603](https://pubmed.ncbi.nlm.nih.gov/37705603/)]
41. Ioachimescu OC, Allam JS, Samarghandi A, et al. Performance of peripheral arterial tonometry-based testing for the diagnosis of obstructive sleep apnea in a large sleep clinic cohort. *J Clin Sleep Med.* Oct 15, 2020;16(10):1663-1674. [doi: [10.5664/jcsm.8620](https://doi.org/10.5664/jcsm.8620)] [Medline: [32515348](https://pubmed.ncbi.nlm.nih.gov/32515348/)]
42. Lv R, Liu X, Zhang Y, et al. Pathophysiological mechanisms and therapeutic approaches in obstructive sleep apnea syndrome. *Signal Transduct Target Ther.* May 25, 2023;8(1):218. [doi: [10.1038/s41392-023-01496-3](https://doi.org/10.1038/s41392-023-01496-3)] [Medline: [37230968](https://pubmed.ncbi.nlm.nih.gov/37230968/)]
43. Redline S, Azarbarzin A, Peker Y. Obstructive sleep apnoea heterogeneity and cardiovascular disease. *Nat Rev Cardiol.* Aug 2023;20(8):560-573. [doi: [10.1038/s41569-023-00846-6](https://doi.org/10.1038/s41569-023-00846-6)] [Medline: [36899115](https://pubmed.ncbi.nlm.nih.gov/36899115/)]
44. Xie A, Bedekar A, Skatrud JB, Teodorescu M, Gong Y, Dempsey JA. The heterogeneity of obstructive sleep apnea (predominant obstructive vs pure obstructive apnea). *Sleep.* Jun 1, 2011;34(6):745-750. [doi: [10.5665/SLEEP.1040](https://doi.org/10.5665/SLEEP.1040)] [Medline: [21629362](https://pubmed.ncbi.nlm.nih.gov/21629362/)]
45. Janiesch C, Zschech P, Heinrich K. Machine learning and deep learning. *Electron Markets.* Sep 2021;31(3):685-695. [doi: [10.1007/s12525-021-00475-2](https://doi.org/10.1007/s12525-021-00475-2)]
46. Romero-Corral A, Caples SM, Lopez-Jimenez F, Somers VK. Interactions between obesity and obstructive sleep apnea: implications for treatment. *Chest.* Mar 2010;137(3):711-719. [doi: [10.1378/chest.09-0360](https://doi.org/10.1378/chest.09-0360)] [Medline: [20202954](https://pubmed.ncbi.nlm.nih.gov/20202954/)]
47. Cabitza F, Campagner A, Soares F, et al. The importance of being external. methodological insights for the external validation of machine learning models in medicine. *Comput Methods Programs Biomed.* Sep 2021;208:106288. [doi: [10.1016/j.cmpb.2021.106288](https://doi.org/10.1016/j.cmpb.2021.106288)] [Medline: [34352688](https://pubmed.ncbi.nlm.nih.gov/34352688/)]

Abbreviations

AHI: apnea-hypopnea index

CrI: credible interval

GRADE: Grading of Recommendations, Assessment, Development and Evaluations

HSAT: home sleep apnea testing

OSA: obstructive sleep apnea

PPG: photoplethysmography

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

PSG: polysomnography

QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies-2

SROC: summary receiver operating characteristic

Edited by Andrew Coristine; peer-reviewed by Brian Locke, Te-Huei Yeh; submitted 08.Jun.2025; final revised version received 10.Aug.2026; accepted 11.Aug.2026; published 25.Sep.2026

Please cite as:

Yeo BSY, Gao EY, Tan JEX, Koo JY, Lee YS, Tan NKW, Ng ACW, Leong ZH, Ong TH, Leow LC, Huang GB, Tan BKJ, Toh ST

Accuracy of Artificial Intelligence in Diagnosing Obstructive Sleep Apnea Using Photoplethysmography: Systematic Review and Meta-Analysis

J Med Internet Res 2026;28:e78718

URL: <https://www.jmir.org/2026/1/e78718>

doi: [10.2196/78718](https://doi.org/10.2196/78718)

© Brian Sheng Yep Yeo, Esther Yanxin Gao, Joan Ern Xin Tan, Jun Yuan Koo, Yi Siang Lee, Nicole Kye Wen Tan, Adele Chin Wei Ng, Zhou Hao Leong, Thun How Ong, Leong Chai Leow, Guang-Bin Huang, Benjamin Kye Jyn Tan, Song Tar Toh. Originally published in the Journal of Medical Internet Research (<https://www.jmir.org>), 25.Sep.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on <https://www.jmir.org/>, as well as this copyright and license information must be included.